

UNIVERSITY OF CAMBRIDGE

Interpreting Structured Narrative Representations in Pretrained Language Models

Hope Elizabeth McGovern



Robinson College

This thesis is submitted for the degree of Doctor of Philosophy

July 2025

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the preface and specified in the text. It is not substantially the same as any work that has already been submitted, or is being concurrently submitted, for any degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee.

Abstract

This thesis investigates whether *unstructured* embedding spaces—such as those learned by large language models (LLMs)—implicitly encode the *structured* knowledge required for narrative understanding. Storytelling is central to human cognition and social interaction, and by probing how AI systems process narrative structure, we gain insight into both the capacities of these models and the contours of human intelligence itself.

The work proceeds along two primary lines of inquiry. First, I broadly survey the research landscape, revisiting foundational claims about distributed representations in language models and asking whether modern Transformer architectures might natively support narrative understanding tasks. I describe key limitations that suggest current architectures fall short of representing the deeper structures that underpin narrative meaning.

Second, I ask whether LLMs encode *structured patterns* characteristic of human storytelling. Through a series of controlled experiments focusing on those structured patterns which are primarily evidenced and documented in classical texts like the Bible, I examine whether these models (1) detect internal rhetorical structures such as chiasmus, (2) recognise cross-textual parallels, (3) surface passages grounded in shared narrative frames rather than surface form, and (4) make analogical inferences about narratives based on deep conceptual structure rather than lexical overlap.

Together, these investigations offer a comprehensive evaluation of whether the unstructured embedding spaces learned by LLMs can support structured, hierarchical, and analogical reasoning. The results shed light on the representational limits of current language models and point toward the need for architectures that integrate narrative structure more explicitly—advancing both cognitive modelling and downstream tasks that rely on robust, explainable meaning representations.

Acknowledgements

I declared in a previous page that this dissertation is solely the product of my own work, and while this is true in a certain legalistic fashion, it is a ridiculous idea by any other metric. There have been multitudes of people shaping and forming this work, just as there have been multitudes shaping and forming me as I completed it. This is the appropriate place to shed light on those multitudes. So, with a very thankful heart, I dedicate this work:

To my father, whose hard work and sacrifice has enabled me to live the life of the mind.

To my mother, whose faithful, daily prayer I somehow suspect is the root of every good gift I have received from God.

If I am thy child, O my God, it is because thou gavest me such a mother. If I prefer the truth to all things, it is the fruit of my mother's teachings. If I did not perish long ago in sin and misery, it is because of the long and faithful years which she pleaded for me. What comparison is there between the honor I paid her and her slavery for me?

— Augustine of Hippo

To my husband, Toby, on whose most commendable qualities I could easily write another—probably longer—dissertation. Thank you for reminding me that my PhD is just a thing.

Everything will be alright in the end. And if it is not yet alright... then it is not yet the end.

— The Best Exotic Marigold Hotel

To the Condiment Bondiment—thank you for the chaos. To Stella in particular, for being the Tweedledum to my Tweedledumber.

To my loving church family, who have contended for me in prayer in the face of my despair, desperate pride, and anxiety: to Sophie, David, Robbie, Sarah, Alasdair,

Howard, Piyush and Fiona, Jonty and Linda, Ben, Michaela, Naomi, Daria, Bekah and Jon, Bethany and Trevor, Jonathan, Jono, Ariane, Seb, and countless others.

To innumerable other friends and family: the McGoverns, Brauns, Properts, Speirs, Childs', and Coates', who supported me, believed in me, and were so proud to see me accomplish this behemoth feat.

To Tom, whom I fear thinks of me far more highly than he ought. Thank you for your constant encouragement and for the refinement of my ideas. It was a pleasure to work with you, and it will always be my aim to carry on our work.

To Hale, with whom it was a joy to grow academically as we became more confident in our ability to bring ideas into reality.

To Andrew, for his constant encouragement and support over the course of my PhD.

To Simone and Paula for their early input on my PhD progress and specialization.

To Zaidi—if we had not had that random discussion about Islamic theology during an open-day tour, I would not have been at Cambridge. Thank you for being one of my fiercest advocates, and for creating opportunities for me that I worked hard, after the fact, to deserve.

To the Woolf Institute—especially Miriam and Emma—who believed in me enough to support this work financially and gave me opportunities to meet wonderful individuals of other faiths.

To the Cambridge Trust, who kindly supported this work financially and were a great source of support over the couse of my PhD.

To Fermin and David, who were kinder than I thought it possible to be about this dissertation. Thank you for your time and effort in reading it and your deep engagement with the work.

And finally, at the heart of this dissertation is the idea that all stories are some reflection of a deeper and larger one—perhaps even a true one. For me, that story is the one of the son of God who became flesh, suffered and died for the sins of the world, and rose again as the first fruit of a new and glorious creation.

Jesus did many other things as well. If every one of them were written down, I suppose that even the whole world would not have room for the books that would be written.

— John 21:25

I would not who or where I am today if I had not fallen in love with the greatest story ever told, and I thank my God for that.

Contents

1 Introduction **19**

2 Background and Motivation **25**

2.1 Narrative Typology: Variations on a Theme 26

2.1.1 What Makes a Story? 26

2.2 Language Models and Meaning 28

2.2.1 From Word Vectors to Neural Language Models 28

2.2.2 The Emergence of LLMs 30

2.2.3 Shortcuts, Cheating, and the Limits of Token-Level Reasoning . . . 32

2.2.4 The Debate Over Understanding 33

2.3 NLP for Narrative Tasks 34

2.3.1 Someone did something(s) to someone 34

2.3.2 Somewhere at some time 36

2.3.3 For some reason 36

2.4 Encoding Meaning in Narrative Form 37

2.4.1 Narrative as Information Compression 37

2.4.2 Intertextuality: Narrative as a Network 40

2.4.2.1 Biblical Intertextuality: A Special Case 41

2.4.3 Narrative as Structured Repetition 42

2.4.3.1 Figura 43

2.4.3.2 Type-Scene 44

2.4.3.3 Chiasmus 46

2.5 Summary 46

2.6 Thesis Outline 47

3 Unsupervised Discovery of Rhetorical Structure in Ancient Hebrew **49**

3.1 Poetic Devices in Classical Literature 50

3.2 Formalising Chiasmus 52

3.2.1 Task Description 52

3.2.2 Related Work . . . 54
3.2.3 Operationalising Chiastic Structure . . . 56
3.3 Data . . . 57
3.4 Sliding Window Scoring . . . 58
3.4.1 Embedding models . . . 58
3.4.2 Why Not Use an LLM? . . . 58
3.5 System Output . . . 59
3.6 Human Annotation . . . 60
3.7 Qualitative Analysis . . . 61
3.8 Conclusion . . . 63
3.8.1 From Internal Rhetoric to Intertextual Structure . . . 64

4 Maintaining Intertextuality in Translation 65
4.1 Allusive Text Reuse in the Classical World . . . 66
4.1.1 Literal Text Reuse . . . 68
4.1.2 Allusion and Stylised Reference . . . 68
4.1.3 The Missing Link: Translation . . . 68
4.2 Intertextuality Across Translation . . . 69
4.3 Data . . . 70
4.3.1 Human Translations . . . 71
4.3.2 Independent Benchmark . . . 72
4.3.3 Gold Standard for Intertextuality . . . 72
4.4 Pretrained Embeddings for Semantic Retrieval Involving Ancient Languages 73
4.5 Characterising Intertextuality between Corpora . . . 74
4.5.0.1 Benchmark Corpus . . . 76
4.6 Machine Translation Experiments . . . 76
4.6.1 Translation Quality . . . 78
4.7 Intertextuality Preservation . . . 80
4.7.1 Key Findings . . . 82
4.7.2 Qualitative Example: Overtranslation of Allusions . . . 82
4.8 Limitations and Future Work . . . 83
4.9 Conclusion . . . 84

5 Exploring Narrative Type-Scenes in the Hebrew Bible 85
5.1 Type Scene: A Fixed Constellation of Motifs . . . 86
5.2 Related Work . . . 87
5.2.1 Narrative Type Classification . . . 87
5.2.2 Information Retrieval . . . 88
5.2.3 Prior Work and Research Gap . . . 89

5.3 Curating an Evaluation Corpus . . . 90
5.3.1 Canonical Type-Scene: The Encounter at the Well . . . 91
5.3.2 Gold Standard . . . 91
5.3.3 Variants . . . 91
5.3.4 Negative sampling . . . 92
5.4 Type-Scene Identification as Information Retrieval . . . 93
5.4.1 Evaluation . . . 94
5.4.2 Naive Embedding . . . 94
5.4.3 Bag-of-Words . . . 95
5.4.4 Pretrained Models for Dense Retrieval . . . 96
5.5 Results . . . 96
5.6 Discussion and Future Work . . . 98
5.7 Conclusions . . . 99

6 Probing for Narrative and Rhetorical Structure 101
6.1 Introduction . . . 102
6.1.1 Contributions . . . 103
6.1.2 Broader Motivation . . . 103
6.2 Related Work . . . 103
6.2.1 Memory and Narrative Recall: . . . 104
6.2.2 Diagnostic Probing . . . 104
6.2.3 Analogical Reasoning for Narrative Tasks . . . 105
6.3 Task Description and Data . . . 107
6.3.1 Rhetorical Parallelism (Span-Level) . . . 107
6.3.2 Narrative Parallelism (Document-Level) . . . 108
6.3.3 Data Preprocessing . . . 111
6.3.4 Limitations . . . 111
6.4 Edge Probing for Parallelism Classification . . . 112
6.4.1 Extracting Contextual Representations . . . 113
6.4.2 Option 1: Span Representation and Pooling . . . 114
6.4.2.1 Trainable Projection and Attentive Pooling . . . 114
6.4.2.2 Scalar Mixing Over Layers . . . 114
6.4.3 Option 2: Zero-Parameter Span Pooling . . . 114
6.4.3.1 Scalar Mixing over Heads and Layers . . . 115
6.4.4 Probing Classifiers . . . 116
6.4.5 Prompting . . . 116
6.4.6 Decoding and Output Processing . . . 118
6.4.7 Model Selection . . . 118
6.5 Results . . . 119

6.5.1 Encoding and Pooling Methods 122
6.5.2 Probing vs. Prompting: Behaviour vs. Latent Knowledge 124
6.5.3 Localising Parallelism in Llama3 Models 125
6.6 Discussion 127
6.7 Conclusion 128

7 Conclusions and Future Work **131**

Bibliography **137**

List of Figures

1.1 **Propp's 31 character functions.** As presented in Jahan et al. (2021) . . 21

2.1 **Cultural Schemas.** A single narrative prompt ('Once upon a time…') activates a schematic frame containing role-slots with default fillers. These defaults vary across individuals based on prior experiences, genre expectations, and cultural background. The illustration demonstrates Minsky's concept of terminal filling defaults within frames, showing how divergent expectations can arise from a shared schematic scaffold. 38

3.1 **Experimental setup for chiasmus detection.** In a sliding window fashion, I encode verses (or half-verses) with a sentence encoder and then calculate the delta of the pair similarities and non-pair similarities to create a score distribution over all sets of verses. In this illustration, the sliding window operates right-to-left to follow Hebrew orthography. 52

3.2 **Total number of chiasmi per book in the Old Testament at verse and half-verse level.** While some books tend to have more chiasmi overall, this figure shows whether verse-level or half-verse-level chiasmi are more prevalent in each book. 60

4.1 **Verse coverage in JHUBC Human Translations.** The x-axis indicates verse position across the Old Testament (left) and New Testament (right). Darkened lines indicate a missing verse in the associated human translation. Note that Turkish and Marathi translations are particularly noisy texts. 72

5.1 **Network diagram.** A type-scene description illustrated as a graph, where nodes represent characters (blue) and events (green), while directed edges indicate semantic roles in narrative event chains, e.g. 'Maiden' is the agent of the verb 'meet' and the patient of the verb 'tell'. Intuitively, instantiations of the same type-scene should yield a similar graph structure. 95

6.1 **Experimental setup for parallelism probing tasks.** The rhetorical task is shown: a full sermon text is encoded by the model, which produces contextualised embeddings across all layers of the model. I then extract relevant span pairs and use either learned span pooling or zero-paramter pooling to reduce the embeddings to fixed size vectors, which are then concatenated and fed into a binary classifier. 113

6.2 **Narrative vs. Rhetorical Task Performance in Base and instruction-tuned Models.** Bar chart of AUC across model sizes using the classifier setup of Tenney et al. (2019a): spans encoded in one pass together, using a self-attentive pooling learned alongside a two-layer MLP classifier. 121

6.3 **The effect of pooling and encoding combinations on task performance.** Heatmaps illustrate the resulting classifier performance (AUC) on the two tasks for every combination of learned span pooling vs. zero-parameter pooling and separate encoding vs. joint encoding. The importance of pooling and encoding is much more stark in the rhetorical task, whereas the (low) performance is uniform across setups for the narrative task. For both tasks, zero-parameter pooling achieves top-scores. Values shown are using a 2-layer MLP classifier and embedded representations from a 3B parameter base model. 122

6.4 **Center-of-Gravity.** CoG represents an average over the learned layer weights of the span embeddings – effectively representing the layer of the model which is most useful to the classifier in solving a given task. Narrative and rhetorical parallelism appear to be localised in the mid-layers of the model, regardless of model depth. 126

List of Tables

2.1 The restaurant script of Schank and Abelson (1975), structured into roles, goals, and a series of scenes composed of primitive conceptual actions. Each scene involves stereotyped acts such as `PTRANS` (physical movement), `MTRANS` (mental communication), `MBUILD` (mental construction), etc., which help model how people understand and predict events in familiar contexts. 39

2.2 Two instantiations of the *Betrothal-by-Well* frame in Genesis. While the underlying frame remains constant, the slot fillers vary — highlighting differences in initiative, divine involvement, and narrative tone. 45

3.1 **Chiastic structure of Psalms 35:4-7.** Pairs (A, A') and (B, B') form an 'X' pattern, exhibiting repeated phrases and conceptual links. 51

3.2 **Comparison of Chiastic Structures.** High cosine similarity between corresponding units (A, A' and B, B') supports valid chiasm (col. 1), while antonymic inversion reflects structured opposition (col. 2). Genealogical repetition (col. 3) exhibits high similarity but lacks structural intent. Cosine similarity scores here are simulated. 54

3.3 **Summary of detected chiasmi.** 2700+ chiasmi were detected at the verse and half-verse level. The highest number of chiasmi was found in the Book of Genesis and Book of Numbers. Both the precision and the inter-annotator agreement increase for the verse-level chiasmi. 59

3.4 **English translation of a positive example of a chiasmus automatically detected by this method.** Gen 1:19-23 (ESV) 61

3.5 **The story of Jacob stealing Esau's birthright.** English translation of a chiasmus automatically detected by our method. Gen 25:30-34 (ESV) 62

3.6 **God's covenant with Noah after the flood.** English translation of a chiasmus automatically detected by our method. Gen 9:12-15 (ESV) . . . 62

4.1 **Biblical intertextuality.** The highlighted middle verse shows the intertextual reference from the New Testament to the Hebrew Bible, establishing a connection between the drying up of the Euphrates River and Moses parting the Red Sea. Both are instances of divine intervention in the context of a body of water. However, intertextuality here not only establishes a semantic parallel between two events, but it also emphasizes the difference. The passage from Exodus is moment of the divine judgment that leads to safety, whereas the drying up of the Euphrates is a preparation for the final judgment of the world. 67

4.2 **Resource Level by Language Family.** In this study, I consider translation pairs including a handful of languages for which there exists a full human translation of both the Jewish and Christian texts in the JHUBC. These languages span different resource levels; the resource levels are indicated by the number of bitext sentences readily available in online datasets, as reported by Tang et al. (2020). 71

4.3 **Distribution of validated intertextual cross-references.** I distinguish between within-testament references and across-testament references, using a vote threshold of 50. 73

4.4 Precision@1 retrieval performance using LaBSE for cross-lingual semantic search on aligned Bible verse pairs. Each source verse in Ancient Hebrew or Ancient Greek is used to retrieve its matching translation in the target language (English, Finnish, Swedish, Turkish, Marathi), based on cosine similarity of LaBSE embeddings. The metric reports the proportion of source verses whose top-scoring match in the target language has the same verse identifier. 74

4.5 **COMET scores**. The top-scoring translation for each source manuscript is in bold text. The second top-scoring translation is in italics. 78

4.6 R_{intt} **for source manuscripts and their human translations.** Ratios within, and where possible between, testaments, for the Septuagint and TAGNT (Greek), TAHOT (Hebrew), and five human translations with 95% CI. 80

4.7 **Overemphasized Intertextuality by Human Translation.** The intertextuality from Hebrews 8:12 to Isaiah 43:25 is amplified by the human translator's decision to render different words as "sin". The machine translation abstains from this and restores the original distance, but loses coherence. 82

5.1 **Common Biblical Type-Scenes and Their Scholarly Attestations.** This table summarizes frequently recurring narrative patterns ("type-scenes") in Biblical literature, along with examples and key scholarly references. . . 86

5.2 **Well-Encounter Type Scenes in Biblical Narrative.** Passages are grouped into a canonical prototype, gold standard matches, and partial variants. 90
5.3 Performance comparison (AUC) across models for the intertextuality ranking task. Best AUC values are in **bold**, second-best in *italics*. 96

6.1 **Rhetorical Parallelism in Latin:** The repeated syntactic and semantic inversion emphasizes paradoxical transformation. 108
6.2 **Dataset statistics by task type.** The table shows the number of unique source documents and the train/dev/test split for each task. 111
6.3 **AUC scores for narrative and rhetorical parallelism detection using linear (logreg) and nonlinear (MLP) probes.** Each row shows performance across two pooling strategies ('llm-attention' vs. 'learned') and model variants (base vs. instruct) for three LLaMA-3 model sizes. Probes achieve consistently higher AUCs on rhetorical tasks than narrative ones, with best results for rhetorical parallelism exceeding 0.95. Narrative parallelism remains significantly harder to detect, even with instruction-tuned models and nonlinear classifiers. Bold text indicates the highest score for a model across classifier configurations. 119
6.4 **Model performance across tasks, sizes, and tuning variants.** F1 and AUC scores are reported for both base and instruction-tuned versions of each model size. 124

1
Introduction

In 1936, Russian composer Sergei Prokofiev composed a symphony for children based upon a simple folktale – *Peter and the Wolf*. In the years that followed, a slew of animated adaptations appeared, most notably that of the Disney Corporation in 1946, whose popularity was destined to fall squarely within the remit of elementary school music teachers, like mine, who were fond of showing it for an easy Friday lesson.

In this opus, a pastoral tale unfolds in quite a peculiar way: each character in the story is represented not by words, but by a distinct musical voice. A flute chirps ebulliently for the bird; an oboe waddles chromatically for the duck; a clarinet slinks with feline grace for the cat. The grandfather's entrance is accompanied by the gruff solemnity of a bassoon, while the wolf snarls through a cacophony of French horns. Peter himself—our hero—prances in on a jaunty melody of light strings. The narrator's only task in the production is to remind us of what the music already makes clear: the story is speaking for itself.

Each of these instrumental themes is a *leitmotif*—a recurring musical idea associated with a particular character or emotion. In Prokofiev's hands, the leitmotif becomes a mechanism of comprehension: we do not *see* the grandfather admonish Peter; we hear it, through the sudden entrance of the bassoon motif against Peter's theme. There is no ambiguity about what is happening, even in the absence of language. The motifs return, recombine, and evolve, carrying the story forward.

While *Peter and the Wolf* renders its leitmotifs so clearly that even a child can follow

them, the technique is widespread across long-form musical compositions—from Wagner's operas to the soundtracks of modern film and video games. In Howard Shore's score for *The Lord of the Rings*, for example, the gentle theme of the Shire reappears unexpectedly in Mordor, evoking the emotional memory of a simpler world. This moment mirrors a character's speech about holding on to goodness in dark times. The music does not merely accompany the narrative—it deepens and transforms it.

Just as music uses recurring structures and motifs to construct meaning, so, too, do stories. My first scholarly introduction to this idea was in Robert Alter's *The Art of Biblical Narrative* (Alter, 2011). In this work, Alter embarks on a forensic exploration of literary traditions long since lost to the West but preserved latently in the library of the Biblical texts. Again and again, patriarchs meet their future wives at wells under nearly identical circumstances, barren women receive a divine promise of children, and younger sons are favoured above the firstborn. These so-called *type-scenes*—repeating narrative structures with subtle divergences—form the scaffolding of meaning.

To the ancient audience, these formulae were as familiar as a showdown between a sheriff and a vigilante at high noon is to us. They allowed storytellers to embed interpretive expectations into the story itself. Much like hearing a musical motif recur in a new context, encountering a familiar scene with a twist invites reflection, comparison, and interpretation.

This recognition of recurrent narrative form extends far beyond Biblical literature. Vladimir Propp's *Morphology of the Folktale* (Taylor et al., 1959) famously distilled Russian fairy tales into a set of 31 narrative functions—abstract roles like "the hero leaves home," "the villain is punished," or "misfortune is made known." (see Fig. 1.1). Similarly, Joseph Campbell's *The Hero with a Thousand Faces* (Campbell, 1949) outlined the archetypal "monomyth" underlying mythological narratives across many cultures. The Aarne–Thompson–Uther (ATU) index catalogues thousands of folktale types and motifs across time and geography.

Symbol	Name	Description
β	Absentation	One of the members of a family absents himself from home.
γ	Interdiction	An interdiction is addressed to the hero.
δ	Violation	The interdiction is violated.
ϵ	Reconnaissance	The villain makes an attempt at reconnaissance.
ζ	Delivery	The villain receives information about his victim.
η	Trickery	The villain attempts to deceive his victim.
θ	Complicity	Victim submits to deception and thereby unwittingly helps his enemy.
A/a	Villainy/Lack	A member of a family is harmed by the villain, or lacks something.
B	Mediation	Misfortune or lack is made known.
C	Beginning Counteraction	The hero agrees to or decides upon counteraction.
$\uparrow$	Departure	The hero leaves home.
D	Donor Encounter	The hero is tested, interrogated, attacked, etc., which prepares the way for E.
E	Hero's Reaction	The hero reacts to the actions of the future Donor.
F	Receipt of Magical Agent	The hero acquires the use of a magical agent.
G	Transference	Hero is led to the whereabouts of an object of search.
H	Struggle	The hero and the villain join in direct combat.
I	Victory	The villain is defeated.
J	Branding	The hero is branded.
K	Tension Liquidated	The initial misfortune or lack is liquidated.
$\downarrow$	Return	The hero returns.
Pr	Pursuit	The hero is pursued.
Rs	Rescue	Rescue of the hero from pursuit.
o	Unrecognized Arrival	Unrecognized, he arrives home or in another country.
L	Unfounded Claims	A false hero presents unfounded claims
M	Difficult Task	A difficult task is proposed to the hero.
N	Solution	The task is resolved.
Q	Recognition	The hero is recognised.
Ex	Exposure	The false hero or villain is exposed.
T	Transfiguration	The hero is given a new appearance
U	Punishment	The villain is punished.
W	Reward	The hero is married and ascends the throne, or receives money.

Figure 1.1: **Propp's 31 character functions.** As presented in Jahan et al. (2021)

Taken together, these ontological, empirical frameworks show that narrative structure is neither random nor infinitely varied. Instead, it is shaped by cultural inheritance, cognitive economy, and communicative function. Stories evolve in form, but they do so within limits. Certain shapes recur—not because they are clichéd, but because they are useful.

The term I will use for this phenomenon throughout the thesis is *narrative typology*. I adopt it in a dual sense: As a communicative framework, it describes how authors use recurring narrative forms—such as type-scenes, motifs, or archetypes—to signal meaning, evoke emotion, and structure expectation. As an interpretive framework, it describes how readers (ancient and modern) use these forms to make sense of stories, often linking them across time or genre.

This dual function is especially prominent in Christian typology, where earlier stories are read as prefigurations of later events. The binding of Isaac in Genesis, for instance, was understood by early Christian writers as a *type* of Christ's crucifixion—a prelude whose structure echoes across the testaments. This is not merely retrospective allegory; it is a theory of how stories *mean*.

Narrative typology, then, is not just an aesthetic device; it is a form of reasoning. Bruner (1986) argues that narrative is one of the two fundamental modes of human cognition. Schank and Abelson (1979) propose that much of our everyday reasoning is script-based: we understand events by comparing them to familiar narrative templates.

Marvin Minsky's *frames* similarly describe knowledge as structured expectations built from experience (Minsky, 1974).

These ideas, developed decades before neural networks came into fashion, have found empirical support in perceptual psychology and cognitive neuroscience. People recall events better when they are remembered as part of a narrative structure (Lee and Chen, 2022). They communicate events to others in narrative form, and change the content of that narrative based on their communicative goal (Eckardt et al., 2024; Lee et al., 2020). Incredibly, audience and storyteller show synched neural activity patterns when experiencing the same story (Lee et al., 2020). In short: narrative is an extant cognitive architecture.

The field of natural language processing (NLP) has seen tremendous progress in recent years, driven by the rise of deep learning and, in particular, Large Language Models (LLMs). Models such as GPT-4, LLaMA, and Mistral have achieved astonishing fluency on a wide range of tasks. They translate languages, summarise documents, generate poetry, and answer reading comprehension and factoid-related questions with ease. They are even making headway on previously unreachable benchmark tasks requiring commonsense reasoning and logical inference.

And yet, despite these successes, they 'hallucinate' facts. They fail basic logic puzzles. They 'cheat' on benchmarks (Balloccu et al., 2024; Geirhos et al., 2020). They are brittle in the face of small perturbations (Cai and Cui, 2023). Even more troubling are theoretical critiques: models that are trained purely on form, without access to grounded referents, may be incapable of *true* understanding (Bender et al., 2021). This is not merely a philosophical concern. It strikes at the core of whether these systems can engage with meaning in the way humans do[1].

I assert that narrative is an ideal lens through which to examine the capabilities and limits of LLMs. Unlike classification or summarisation, it demands long-range coherence, hierarchical structure, and sensitivity to both form and meaning. To understand a story is to track its agents, events, and goals—but also its deeper patterns: its symmetry, its echoes, its violations of expectation.

While there has been significant work on story generation and event extraction in NLP, there remains a gap in understanding how models process narrative form—not just who did what, but why the structure was chosen, and what it communicates.

This thesis takes up that challenge. Focusing on classical texts—especially Biblical narratives—it investigates whether, and how, LLMs encode structured representations of narrative form. Do they recognise type-scenes? Can they distinguish rhetorical from

[1]In the words of Pasquinelli and Joler (2021), "Machine learning is not bringing a new dark age but one of diffracted rationality, in which [...] an episteme of causation is replaced by one of automated correlations. More in general, AI is a new regime of truth, scientific proof, social normativity and rationality, which often does take the shape of a statistical hallucination."

literal parallelism? Do they identify analogical relationships between stories? And if so, where in their internal representations are these concepts stored?

I begin with behavioural probing, asking whether LLM-derived embeddings contain task-relevant information by testing their performance directly on new literary tasks. For instance, I frame chiasmus detection—a task involving local semantic symmetry—as a classification problem, and show that even a naive setup succeeds, suggesting that embeddings encode the relevant structure.

I then examine intertextual alignment across translations, treating it as a largely semantic task. Using multilingual embedding models, I show that cross-lingual semantic retrieval is often effective—but with clear caveats: performance degrades for low-resource languages, even in models with ostensibly universal representations.

To evaluate whether LLMs can recall higher-order narrative structures, I construct a retrieval task for type-scenes—recurring event schemas common in Biblical storytelling. Surprisingly, treating this as a purely semantic retrieval task using sentence embeddings yields the best results, but overall performance remains poor. This suggests that while models may weakly encode such structures, they do not reliably surface them in similarity-based tasks.

Finally, I test LLMs' capacity for recognising parallelism, both at the span and document level. I find that attention mechanisms capture rhetorical parallelism in local contexts, but this capacity diminishes over longer ranges. Interestingly, as model size increases, diagnostic probes trained on their embeddings do *not* reveal proportionally more structured information—yet the models perform better in prompt-based tasks. This decoupling between internal representation and behavioural output raises important questions about where and how abstract narrative knowledge resides in these models. Together, these experiments contribute to a growing understanding of how narrative structure might be embedded, obscured, or emergent within high-dimensional language model representations.

2

Background and Motivation

The central investigation of this thesis is the extent to which *unstructured* embedding spaces – such as those learned by large neural networks – implicitly encode *structured* knowledge about narratives. The ability to tell, comprehend, and recombine stories is fundamental to human intelligence (Winston, 2011), and by mapping the successes and failures in which AI as we know it processes stories, we can better hold a mirror up to human intelligence[1].

The purpose of this chapter is to acquaint the reader with three distinct yet interlocking topics: (1) the principles underlying deep learning — the dominant paradigm in artificial intelligence — with its emphasis on ground-up, data-driven representation learning; (2) an overview of current computational approaches to narrative, including both their achievements and limitations when applied to literary texts; and (3) a theory-informed account of narrative form as a vehicle for meaning, focusing on the recurrent patterns and structures that organise and communicate.

The sections that follow begin with an overview of how deep learning approaches representation, offering an intuition for why neural language models may struggle to capture the deeper semiotic structures of literary narrative. I then survey the current research landscape in narrative-focused NLP, highlighting key developments and gaps. Finally, I turn to the literary and philosophical tradition of *narrative typology*, drawing

[1]See Goldstein and Papert (1977): 'AI [is] a computational style of epistemology' and: '[a]lthough there is much practical good that can come of more intelligent machines, the fundamental theoretical goal of the discipline is understanding intelligent processes independent of their particular physical realization.'

together insights from poetics, theology, and structuralist thought to outline the forms through which stories encode and transmit meaning.

2.1 Narrative Typology: Variations on a Theme

Narrative is a cultural absolute – no human society, past or present, exists without storytelling, whether through oral traditions, pictorial representations, or written works (Barthes and Duisit, 1975). Piper et al. (2021b) argue that '[n]arrative is a core mechanism through which human beings come to understand their world, find meaning, motivate their actions and those of others, and create communities.' Increasingly, public discourse focuses on distinguishing narrative as a framing device that shapes interpretation from the raw events it recounts (Card et al., 2015).

The power of narrative lies in its ability to construct meaning through repetition and reinterpretation. To reinterpret George Santayana's adage, 'no text is an island'; all narratives exist within a web of prior texts that shape and transform them (Bamman and Crane, 2008). Latham (2017) observes that science fiction operates in precisely this manner: there exists a shared repertory of world-building elements (e.g. warp drives and AI cybernetics) common to the genre which authors draw upon to weave a backdrop against which they recount their tale. Work such as Cook and Hirst (2013) on cliché identification or Yarlott et al. (2022) on motif detection is made possible only by the implicit assumption of reusable and reused narrative elements.

In the followings sections, I explore what defines a story[2], how narrative functions as a structured channel of communication, and the myriad ways that narratologists, folklorists, theologians, and philosophers have expounded, in their own ways, the peculiar fact that humans are, fundamentally, storytelling creatures.

2.1.1 What Makes a Story?

> "Just as simple sentences can be said to have an internal structure, so too can stories be said to have an internal structure[...] in spite of the fact that no one has ever been able to specify a general structure for stories that will distinguish the strings of sentences which form stories from strings which do not."
>
> (Rumelhart, 1975)

Imagine you are sitting at a bus stop. A young man you've never met before approaches you and says, 'The name of the common wild duck is *histrionicus histrionicus histrionicus*.'

[2]For purposes of this dissertation, 'story' and 'narrative' may be taken as complete synonyms.

That is the entirety of the interaction.

Philosopher Alasdair MacIntyre[3] poses this thought experiment and concludes with the question, 'Why did this happen?' (MacIntyre, 2013). Instinctively, humans try to make sense of disconnected events by structuring them into a compelling story; perhaps the man is giving a secret codeword, mistook you for someone else, or is mentally ill. Each of these explanations imposes a narrative structure on an otherwise isolated event. Without motivations, subsequent events, or broader context, however, the interaction remains a mere fragment; it is not yet a story.

Narrative, then, is *emergent* rather than merely sequential; it is a structured whole that binds disparate elements into coherence. Paul Ricoeur (1984) articulates it this way:

> "The plot of a narrative […] 'grasps together' and integrates into one whole and complete story multiple and scattered events, thereby schematizing the intelligible signification attached to the narrative taken as a whole."

So the first way we may define a narrative is that it is a series of events and the scaffold upon which the otherwise isolated events hang. This broad, teleological definition – one that emphasises the *purpose* of narrative – is useful philosophically, but insufficiently precise for computational applications. NLP tasks require definitions that lend themselves to formalisation.

A useful analogy can be found in the realm of computer vision. Many applications depend strongly on an ability to detect objects, yet some categories defy straightforward definition. Consider chairs: chairs vary dramatically in form – some have four legs, others none; some have armrests, other do not. A three-legged stool shares very little discernible characteristics in common with a plush leather armchair, yet humans do not struggle to categorise them both as instances of the parent class 'chair'. Their defining characteristic is not a particular shape or component, but rather their *function*: a chair is that which is used for sitting (Grabner et al., 2011). Likewise, narratives cannot be defined merely by their sequence of events (their components); rather, they are structured wholes that rely on coherence, memory, and relational knowledge.

> "There's little we can find in common to all chairs – except for their intended use." (Minsky, 1988)

Piper et al. (2021b) compellingly translate key insights from narrative theory for an NLP audience. Distilling the work of Paul Ricoeur (1984) and others, they produce a minimal definition of narrativity, which I accept:

[3] I encountered this thought experiment through the work of Tim Keller in *Every Good Endeavour* (Keller, 2014), and my recounting of it closely follows his in Chapter 9 of EGE.

> "| Someone | tells | someone | somewhere that | someone | did something(s) | [to someone] | somewhere | at some time | for some reason"

This definition captures two crucial aspects of narrative. The first four elements ('Someone tells someone somewhere that') reflect postmodern literary concerns, considering the narrator and audience as integral to storytelling. The latter elements reflect classical narratology, outlining the fundamental structure of events in time and space (Piper et al., 2021b). Temporality, in particular, is central—though experimental works like *Slaughterhouse-Five* (Vonnegut, 1969) challenge linear chronology, they are the exceptions which prove the rule, underscoring the necessity of time as a narrative axis.

Returning to MacIntyre's bus stop example, we see that it meets most of these criteria but fails on one crucial point: there is no discernible *reason* for the utterance. Without a motivating force—whether psychological, causal, or teleological—the event remains incomplete as a narrative.

This definition, unlike more abstract literary conceptions, offers an operational framework for NLP. Each component—agents, actions, objects, spatiotemporal markers, and motivations—can be mapped onto computational tasks, from entity recognition to event extraction to causal inference. In Section 2.3, I explore these aspects in greater depth, assessing current progress and open challenges in narrative modelling.

Yet defining *what* constitutes a narrative is only the beginning. The more pressing question is: *why* does narrative structure matter for NLP? The answer, I argue, is that just as morphosyntactic rules enable the parsing of sentences, narrative structure enables the comprehension of stories (Rumelhart, 1975). Understanding a narrative involves more than tracking surface events—it requires the implicit recognition of character archetypes, familiar plot trajectories, and intertextual references. Narratives, in short, encode knowledge through recurrent units, offering a compact and cognitively efficient scaffold for meaning.

In the next section, I explore how these recurrent narrative structures not only shape human understanding but also present opportunities—and challenges—for computational models seeking to capture the richness of story.

2.2 Language Models and Meaning

2.2.1 From Word Vectors to Neural Language Models

A language model is an abstraction over a linguistic dataset which, given a previous word or sequence of words, predicts the next word in the sequence. The first formalised language models were statistical, relying on frequency counts as a probability estimate for a token's likelihood in a given context. However, this approach struggles with key

properties of language in use, such as polysemy (words with multiple meanings), the emergence of neologisms (new words), and sensitivity to longer-range dependencies. As a result, statistical models often fail to capture the structure and variability of natural language.

In natural language processing, it is well-established that neural language models encode structured linguistic patterns through large-scale statistical pre-training (Halevy et al., 2009). Instead of assigning a fixed probability to a token based on frequency in a corpus, neural models learn internal representations that reflect both the immediate context and latent features such as syntactic roles, discourse structure, and even factual or common sense knowledge (Tenney et al., 2019a; Petroni et al., 2019). The core assumption underlying such models is that *context serves as a proxy for meaning* — 'You shall know a word by the company it keeps' (Firth, 1957). This view is formalised in distributed representation theory (Hinton et al., 1986) and popularised through models like Word2Vec (Mikolov et al., 2013).

In Word2Vec, for example, each word in the vocabulary is assigned a vector such that words appearing in similar contexts lie close together in the learned space. The model learns these embeddings by predicting nearby words within a fixed window. This allows semantic similarity to be captured as geometric proximity: words such as 'king' and 'queen' occupy similar regions of the space because they often appear in analogous contexts.

Formally, given a sequence of tokens $w = [w_1, w_2, \ldots, w_n]$, a standard neural language model defines the conditional probability of the next token as:

$$p(w_t \mid w_{<t}) = \text{softmax}(Wh_t + b)$$

where $h_t \in \mathbb{R}^k$ is the hidden state at position t, computed from a deep neural network, $W \in \mathbb{R}^{|\mathcal{V}| \times k}$ is a learned output matrix mapping to vocabulary logits, and b is a learned bias term. These probabilities are trained to reflect statistical regularities in language data.

Intuitively, the more context (previous and/or future tokens) a language model is able to consider, the more accurate its next-token predictions will be. In practice, storing and effectively utilising context across an entire sequence is computationally expensive, and earlier sequence models like Recurrent Neural Networks (RNNs) suffered from *catastrophic forgetting*, where information from earlier parts of the sequence would gradually be overwritten or lost as new tokens were processed. This happened in part because these models process tokens one at a time, passing compressed representations forward step-by-step. Not only does this limit the retention of long-range information, it also makes training inefficient and prone to error accumulation.

Transformers – the backbone architecture of modern NLP – overcome these limitations

by replacing sequential processing with *self-attention*, a mechanism that dynamically adjusts the relevance of every token in the sequence relative to every other:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

where $Q, K, V \in \mathbb{R}^{n \times d_k}$ are the query, key, and value matrices respectively, derived from the input embeddings. This allows each token to attend directly to the entire input sequence, enabling efficient modelling of long-range dependencies and complex linguistic relationships. Unlike RNNs or Convolutional Neural Networks (CNNs), Transformers process all tokens in parallel, which not only improves computational performance but also mitigates the catastrophic forgetting problem by giving equal access to the full context (Vaswani et al., 2017). This shift unlocked a new generation of eminently scalable language models. Subsequent work by Brown et al. (2020) demonstrated that that Transformers trained on large-scale text corpora could perform a broad range of tasks – machine translation, question answering, summarisation, etc. – without task-specific fine-tuning. In effect, language modelling itself had become a general-purpose pretraining objective, pushing the field beyond handcrafted features and narrow benchmarks toward expressive, task-agnostic representations of language.

There are two major paradigms for training neural language models. One is causal language modelling (CLM), where the model predicts the next word given only the preceding context. It processes tokens strictly from left to right. This is the training objective used in GPT (Generative Pre-trained Transformer), which is optimised for generation tasks. The other is masked language modelling (MLM), in which some tokens in the input are randomly replaced with a special `[MASK]` token, and the model is trained to reconstruct the original sentence. This is the objective used in BERT (Bidirectional Encoder Representations from Transformers), which learns to integrate both left and right context and is particularly suited to classification and inference tasks. CLM reflects the incremental nature of language generation, while MLM encourages the model to form deeper bidirectional representations of meaning.

2.2.2 The Emergence of LLMs

In recent years, a new pseudo-category of language models has come to dominate the field of natural language processing: *Large* Language Models, or LLMs. The term is used inconsistently across both academic and public discourse[4], but in this thesis, I use *LLM* to refer broadly to pretrained neural language models built with a Transformer architecture

[4]For conflicting examples, K et al. (2024) define a large language model as anything larger than 1B parameters, while Xu et al. (2025) define a 'small language model' as having fewer than 4B parameters.

and containing billions of parameters. I make no distinction between base models and instruction-tuned variants in the application of the term.

While earlier neural language models such as BERT and GPT-2 made architectural strides, the models that are now grouped under the LLM umbrella are distinguished by their scale — both in terms of parameter count and training data — and by the emergent behaviours that appear only at that scale. These emergent properties are abilities that do not manifest in smaller models but arise once certain thresholds of scale are crossed, including in-context learning, code generation, and multi-step reasoning (Radford et al., 2019; Brown et al., 2020).

This raises two central questions: How do LLMs achieve such broad competence? And does this competence reflect genuine understanding, or just high-dimensional statistical pattern matching?

The performance leaps observed in recent years are typically attributed to three primary factors: (1) the availability of powerful computational hardware, especially GPUs and TPUs (Krizhevsky et al., 2017); (2) the shift to transformer architectures (Vaswani et al., 2017), which allow efficient parallel computation of attention matrices; and (3) the enormous volumes of data used for pretraining (Halevy et al., 2009). LLMs are trained on internet-scale corpora using next-word prediction(as in GPT-style models) or masked token prediction (as in BERT-style models). Yet despite their generalisation abilities, this training paradigm does not incorporate explicit structural guidance: there is usually no supervision for syntax, semantics, or world knowledge beyond what is implicitly encoded in co-occurrence patterns.

For example, decoder-only models like GPT are trained to predict the next token in a sequence, then further refined using *reinforcement learning from human feedback* (RLHF) to produce outputs that better align with human preferences (Zhang et al., 2024). Instruction tuning pushes this even further, adapting models to follow natural language instructions. These techniques improve usability and alignment—but they do not introduce explicit reasoning mechanisms to the model's underlying representations.

As I will explore in the following section, this creates a tension: LLMs exhibit surprising and often useful behaviours, yet their training objective remains narrow. This raises concerns about whether their impressive performance reflects deep linguistic or conceptual understanding, or whether models are simply exploiting statistical regularities to take shortcuts on benchmark tasks.

2.2.3 Shortcuts, Cheating, and the Limits of Token-Level Reasoning

Neural language models exhibit striking capabilities in structured knowledge recall. For instance, masked language models can complete relational facts (e.g., 'C.S. Lewis was a professor at `[MASK]`' → *Cambridge University*), despite lacking an explicit world model. Petroni et al. (2019) argue that these associations emerge from statistical regularities in training corpora: models learn to associate co-occurring word patterns without needing to 'understand' their meaning in a human sense.

Despite this, recent work has demonstrated that LLMs often rely on shallow heuristics; neural networks often take 'shortcuts', defined by Geirhos et al. (2020) as 'decision rules that perform well on standard benchmarks but fail to transfer to more challenging testing conditions, such as real-world scenarios'. More provocatively, Balloccu et al. (2024) show that models may 'cheat', achieving high scores on evaluation datasets by memorising parts of them or indirectly learning from leaked examples during pretraining. Concordantly, Chang et al. (2023) catalogues the published books known to ChatGPT and GPT4 using a character name *cloze* task, determining that these large models have memorised a large number of published works still under copyright protections. This undermines the validity of benchmark performance as a proxy for generalisable reasoning.

These issues are compounded by the structural limitation that the atomic unit of representation—the token—does not, *a priori*, correspond to higher-level structures like narrative form or conceptual abstraction (Bender and Koller, 2020). One symptom of this limitation is that embedded representations tend to conflate *truth* with *plausibility*: the phrase 'Paris is the capital of France' may be encoded very similarly to 'France is the capital of Paris' in vector space (Marks and Tegmark, 2024). Both sentences are perfectly grammatical, but one is nonsensical. This failure to reliably distinguish between factual and fluent language raises critical questions about what kind of reasoning these models perform—and whether it amounts to more than high-dimensional mimicry.

The current research landscape offers several, potentially overlapping, rationales for the powerful capabilities of modern language models. One explanation, already mentioned, is that they cheat via data leakage. Another explanation points to their use of increasingly long context windows, which allow models to track discourse across thousands of tokens. On this front, the empirical data is mixed: Karpinska and Iyyer (2023) show that LLMs *do* benefit from long-range context for the task of document-level literary translation, while Sun et al. (2021) argue that providing long-range context has a negligible impact on perplexity metrics, implying that 'the evaluated models encode long-range context superficially at best'.

A final and more radical possibility is that their internal representations are converging

on faithful models of reality which thereby enable them to solve real-world tasks (Huh et al., 2024). I further expound this point below.

2.2.4 The Debate Over Understanding

This latter hypothesis—that models build internal representations of reality— sits at the heart of current philosophical and empirical debate. On one side, Huh et al. (2024) catalogues a growing body of evidence that learned embedded representations across model architectures and data modalities are converging. This indirectly implies some underlying, concrete reality which ever-more capable models are representing more faithfully. On the other side, Pasquinelli and Joler (2021) question the very premise that the patterns recognised by machines constitute valid representations of reality. If lived experience and conceptual reasoning cannot be reduced to co-occurrence statistics, then LLMs' apparent knowledge may be illusory. Similarly, Bender et al. (2021) warn against anthropomorphising these systems, arguing that their fluent output masks a fundamental absence of intent, grounding, or comprehension.

These concerns suggest a critical limitation in the reasoning abilities of LLMs. While these models can undoubtedly produce coherent and contextually appropriate outputs across a wide range of tasks, their apparent fluency and problem-solving prowess may, in fact, be the result of mimicking statistical correlations that arise from vast data exposure. Their outputs are statistically sound but semantically shallow.

This trajectory—from statistical co-occurrence to generative fluency—has produced astounding advances in NLP applications. Yet, in domains requiring deeper conceptual understanding, such as narrative reasoning, serious limitations remain. Unlike classification or translation tasks, narrative understanding demands long-range coherence, causal reasoning, and cultural or intertextual awareness. These are precisely the areas where LLMs may struggle.

In this thesis, I propose that narrative tasks serve as a particularly effective testing ground for the question of LLM understanding. Narratives require reasoning about structure, causality, and intention across long sequences—traits that challenge the limitations of token-level modelling. In subsequent chapters, I explore a variety of narrative phenomena found in Biblical narrative. These forms allow us to probe whether LLMs can go beyond surface patterning to encode reusable narrative logic. As I will show, this is not just a benchmark task, but a window into the deeper question of what, if anything, these models understand.

2.3 NLP for Narrative Tasks

In recent years, computational approaches to narrative have proliferated across NLP, literary studies, and digital humanities, tackling various aspects of narrativity from character modelling to event detection and narrative framing. These efforts often coalesce around the minimal ingredients of narrative that I outlined earlier: the presence of characters who do something, somewhere, for some reason. While many of these tasks originate in generic webtext domains such as news or social media, their adaptation to literary narratives poses distinct challenges due to longer document lengths, figurative language, complex temporal structures, and rich layers of interpretation (Bamman et al., 2019; Sims et al., 2019; Sims and Bamman, 2020; Piper et al., 2021a, 2024). While NLP has made significant progress in analysing narrative at the level of events, entities, and character roles, I argue that much of this work treats narrative meaning as *emergent* from morphosyntactic features. In contrast, narrative traditions — especially in literary and religious texts — often rely on structured, recurrent forms to encode and communicate meaning. These forms remain under-explored. In what follows, I review major strands of narrative-focused NLP to situate my own work within this landscape and to identify the gap I seek to address: the computational modelling of narrative *form* as a communicative structure.

2.3.1 Someone did something(s) to someone

Much narrative NLP research focuses on identifying actors and actions — that is, the basic semantic backbone of stories[5]. One principled lens for analysis of characters and their actions in a story is the study of *dramatis personae* – recurring character types that serve as recognisable archetypes across narratives, such as the eccentric professor, femme fatale, or dashing hero.

There are two major historical approaches to analysing character roles in narrative:

1. Joseph Campbell's Monomyth: In his comparative study of mythology, Campbell (1949) proposed a 12-stage archetypal structure for the hero's journey, arguing that classical myth and modern literature follow this repeating pattern.

2. Vladimir Propp's Morphology of the Folktale: Propp, a comparative folklorist, introduced a structuralist approach to Russian folktales, identifying 31 character func-

[5]There is some debate about whether plot or characters are the *most* integral component of stories. Fludernik (2009) states that narrative is, 'a representation of a possible world […] at whose centre there are one or several protagonists of an anthropomorphic nature who are existentially anchored in a temporal and spatial sense'. Contrast this emphasis on experientiality with Prince (1973)'s emphasis on change of state and Labov and Waletzky (1997)'s emphasis on most-reportable event (MRE). See Ouyang and McKeown (2015) for work on explicitly operationalising the MRE concept.

tions – atomic narrative units such as *interdiction* ('the hero is warned') or *struggle* ('hero and villain do battle') (Taylor et al., 1959).

Computational methods have been developed to operationalise these concepts, particularly Propp's character function taxonomy. Approaches have pursued automated tagging and annotation techniques (Finlayson, 2015; Gervás and Méndez, 2024; Yarlott and Finlayson, 2016) as well as ground-up identification of character types and motifs (Jahan et al., 2020, 2021; Jahan and Finlayson, 2019; Yarlott et al., 2022; Reiter et al., 2014). Related research has extended this effort by constructing a folktale index, which categorises folktales across multiple languages based on type (e.g. the Aarne-Thompson-Uther classification system). However, folktale classification poses inherent challenges due to the fluid nature of narrative motifs (Broadwell et al., 2017). Beyond folktales, probabilistic models have been employed to identify and cluster film characters into *persona* categories (Bamman et al., 2013, 2014).

These character-focused approaches often dovetail with work on *event detection*, which seeks to identify and structure what happens in a narrative. Chambers and Jurafsky's seminal work on narrative chains (Chambers and Jurafsky, 2008, 2009) modelled sequences of events involving a central protagonist, capturing rudimentary aspects of coherence and temporal progression. Later extensions sought to merge these chains into more generalised narrative *schemas*, approximating the structure of scripts or frames. In literary domains, however, the detection of events is significantly more complex. Sims et al. (2019) emphasise that literary events often involve imagined states, figurative expression, and subtle psychological shifts, posing problems for models trained on newswire data. Their work introduced a novel corpus of literary event annotations and demonstrated significant gains when models are trained in-domain.

Related tasks such as *entity recognition* and *coreference resolution* have likewise required domain adaptation for literature. Bamman et al. (2019) released a richly annotated dataset of literary entities that included non-named mentions ("the boy", "the cook's sister") and nested structures, showing that models trained on news data perform poorly in literary settings — particularly with regard to gendered pronouns. Similarly, long-form fiction introduces uniquely difficult cases for coreference, which standard benchmarks like OntoNotes fail to capture. The corresponding dataset (Bamman et al., 2020) allows for the study of long-distance coreference chains, often essential to narrative cohesion. Other work has modelled interactions among characters directly, treating fictional social networks as structurally distinct from real-world analogues. Piper et al. (2024) show that literary character networks tend to be more embodied, less modular, and more densely connected than social networks in other contexts.

2.3.2 Somewhere at some time

Narratives also unfold in time and space. Efforts to model these dimensions have included introducing an annotation scheme for causal and temporal events (Mostafazadeh et al., 2016) and using temporal markers as relations in a 'story intention graph' for identifying analogous stories (Elson, 2012). While these tasks have generally received less attention in the literary domain than character or event modelling, they are nonetheless crucial for any attempt to recover narrative structure, especially when stories fragment or distort chronological order. Such distortions are common in literature, where analepsis (flashbacks), prolepsis (foreshadowing), and embedded narration complicate the mapping of events onto timelines (Genette, 1997).

2.3.3 For some reason

Finally, there is the subtle question of *why* a narrative unfolds as it does. Here I describe work on *framing*, especially in news discourse. Card et al. (2016) analyse how roles and causal blame are assigned through the subtle choices of narrative framing. Similarly, Minnema et al. (2022) explore how news reports of femicide differ in their portrayal of aggressors and victims, revealing how narrative framing may encode ideology or imputed moral value. Such research shows that narratives do not merely record events — they *interpret* them, often implicitly. Other work has sought to capture narrative progression at a more global scale, such as detecting *turning points* in film plots (Papalampidi et al., 2019) or summarising screenplays using their underlying act structure (Papalampidi et al., 2020).

Across all these efforts, narrative is treated as a substrate for machine learning: a source of data, a target for prediction, a scaffold for annotation. Yet what is largely missing — and what I aim to foreground — is the idea that *narrative form itself is a communicative medium*, not merely an organisational container. If narrative is a mode of *reasoning through structure* — whether via recurring scenes, typological allusion, or rhetorical similarity — then computational models must grapple with its formalisms, not just its contents. In the next section, I turn to these macrostructural theories of narrative form, drawing on traditions from literary criticism, theology, and comparative folklore. These forms offer precisely the kind of structured knowledge that eludes current token-level models. Understanding how narratives communicate meaning through form, not just through propositions, is central to the broader project of narrative understanding.

2.4 Encoding Meaning in Narrative Form

> "[T]he mechanisms that enable humans to tell, to understand, and to recombine stories separate our intelligence from that of other primates."
>
> (Winston, 2011)

What remains notably absent in the current NLP literature is sustained attention to *structured narrative form* as a communicative device. Literary and religious narratives often rely on established rhetorical and typological forms — including figura, type-scenes, and chiasmus — to encode meaning, signal analogy, and structure memory. These macrostructural patterns function as frameworks for compressing, comparing, and transmitting stories. Recent work in cognitive neuroscience seemingly aligns with longstanding literary theories, offering evidence that narrative structure aids memory recall (Lee and Chen, 2022) and facilitates communicative goals between speakers (Eckardt et al., 2024).

While early artificial intelligence (AI) work in the 1970s–80s attempted to model narrative in a cognitively grounded way via *scripts* and *frames* (Schank and Abelson, 1975; Minsky, 1974), modern neural approaches have largely set aside these formalisms in favour of distributed representations. In what follows, I return to these structured forms, not simply as targets for detection, but as evidence of how narrative meaning is constructed through form.

2.4.1 Narrative as Information Compression

While our borrowed minimal definition of narrativity provides a structural framework, the *way* in which these elements are realised—especially the crucial 'for some reason'—is often left implicit. Not every story articulates its motivations explicitly; in many cases, cultural context or audience expectations supply the missing rationale. This phenomenon is a key aspect of narrative efficiency: instead of laboriously encoding every detail, stories rely on *defaults*—pre-existing knowledge structures shared between speaker and audience. In essence, narratives are able to say less while meaning more.

For instance, characterisation can be achieved far more efficiently through allusion than through exhaustive description. Describing a rustic hero as, 'Tarzan-esque' immediately conjures an image of a wild man swinging from vines, clad in a loincloth, and at home in the jungle[6] (Strunk, 2011). Similarly, in MacIntyre's bus stop example, if we

[6] `N.B` that this also works in the other direction. When John the Baptist is described as wearing camel's hair clothing and eating locusts and honey in the desert (Mark 1:4-6), the audience is meant to understand a reference to the prophet Elijah. This is an instance of implicit characterisation through textual allusion.

learn that the utterance is a codeword, an entire schema of espionage conventions—spy novels, James Bond films, clandestine operations—fills in the blanks, supplying context and clues about what will happen next.

Shannon's information theory defines the optimal encoding of a message relative to the receiver's prior knowledge; similarly, narrative achieves communication efficiency by activating **culturally shared schemas**. In this proposed view, stories function as *compressed encodings* of more detailed mental models. The listener reconstructs the full representation not from what is said, but from what is *activated*—from latent, previously learned frames.

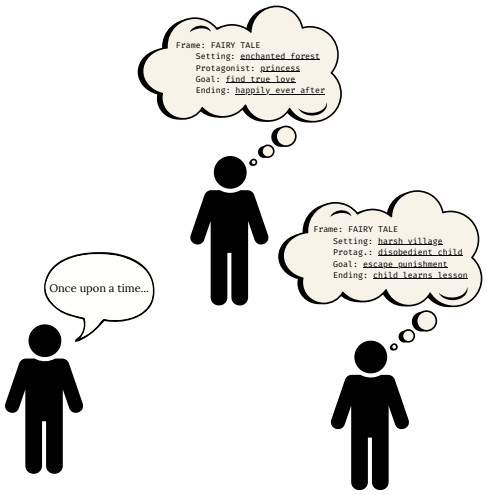


Figure 2.1: **Cultural Schemas.** A single narrative prompt ('Once upon a time…') activates a schematic frame containing role-slots with default fillers. These defaults vary across individuals based on prior experiences, genre expectations, and cultural background. The illustration demonstrates Minsky's concept of terminal filling defaults within frames, showing how divergent expectations can arise from a shared schematic scaffold.

This principle aligns with Frame Theory, first articulated by Minsky (1974) . In his formulation, cognitive understanding relies on *terminal-filling defaults*—structured knowledge templates that allow humans to infer missing information based on prior experience. Just as a well-defined 'restaurant frame' (Table 2.1) includes expectations about waiters, menus, and tipping, narrative comprehension is scaffolded by a rich repository of defaults. A visual representation of this can be found in Fig. 2.1. The key insight is that narratives, like cognitive frames, activate stored knowledge to reduce communicative burden.
A frame is a data structure representing a stereotyped situation, composed of slots (variables) and fillers (default or context-specific values). Formally, a frame F for a scene can be defined as:

$$F = \{(s_1 : v_1), (s_2 : v_2), \ldots, (s_n : v_n)\}$$

Script: *Restaurant*	
Roles	Customer, Waitress, Chef, Cashier
Goal	To get food → decrease hunger, increase pleasure
Scene	**Actions (Conceptual Act — Actor)**
Scene 1: Entering	`PTRANS` — customer enters restaurant
	`MBUILD` — customer finds table
	`PTRANS` — go to table
	`MOVE` — sit down
Scene 2: Ordering	`ATRANS` — receive menu
	`ATTEND` — look at menu
	`MBUILD` — decide on order
	`MTRANS` — tell order to waitress
Scene 3: Eating	`ATRANS` — receive food
	`INGEST` — eat food
Scene 4: Exiting	`MTRANS` — ask for check
	`ATRANS` — give tip to waitress
	`PTRANS` — go to cashier
	`ATRANS` — give money to cashier
	`PTRANS` — exit restaurant

Table 2.1: The restaurant script of Schank and Abelson (1975), structured into roles, goals, and a series of scenes composed of primitive conceptual actions. Each scene involves stereotyped acts such as `PTRANS` (physical movement), `MTRANS` (mental communication), `MBUILD` (mental construction), etc., which help model how people understand and predict events in familiar contexts.

where each s_i is a semantic slot (e.g., *agent*, *goal*, *location*) and each v_i is a default value or variable. For example, an espionage frame might include:

$$F_{\text{spy}} = \{(\text{agent} : \text{“undercover operative”}),$$
$$(\text{location} : \text{“urban safehouse”}),$$
$$(\text{goal} : \text{“extract intelligence”}),$$
$$(\text{action} : \text{“codeword exchange”})\}$$

These frames can be seen as *generative templates*:

$$x \sim P(x \mid F)$$

That is, observed stories x are sampled or constructed from an underlying frame F and some variability. The interpretive task is then:

$$\hat{F} = \arg\max_{F} P(F \mid x)$$

This aligns with the role of narrative as efficient communication: given sparse surface cues x, the listener recovers $\hat{F}$, the most probable frame. Importantly, the mapping from token sequences to frames is many-to-one; the structure reduces entropy in interpretation space.

A historical example of this efficiency in action is *commedia dell'arte*, the medieval Italian theatrical tradition featuring stock characters such as the Harlequin, Pantalone, and the Zanni. These characters are not merely familiar archetypes; they function as *narrative primitives*, allowing new plots to be understood more easily because audiences already know how these figures behave. The Harlequin is wily and mischievous; Pantalone is miserly and lecherous. Just as linguistic syntax structures sentences, *cognitive scaffolding* structures stories by invoking pre-existing schemas, streamlining the process of narrative interpretation.

This brings us to a key conclusion: **Narratives function as information compression mechanisms by activating pre-existing cognitive schemas.**

2.4.2 Intertextuality: Narrative as a Network

One mechanism by which narratives effectively populate default structures is through intertextuality, the process by which texts refer to or build upon one another, either explicitly through direct quotation or implicitly through allusion and structural mirroring (Bamman and Crane, 2008). By invoking prior works, intertextuality enables authors to engage with established themes, compress information, and guide audience expectations.

Ample examples of this phenomenon are on display in Hollywood films. Take, for example, the psychologically disturbed murderer Patrick Bateman in Bret Easton Ellis's *American Psycho*: he is plucked from Alfred Hitchcock's 1960s horror film *Psycho* and placed in corporate America to render a withering critique on the excesses of 1980s 'yuppy' capitalism. Cues from the character name (Bateman/Bates) and riffed movie title provide 'narrative shortcuts' for the audience. These intertextual cues function as interpretive signposts, layering meaning and enhancing thematic resonance.

Intertextuality is not solely a modern phenomenon; it is incredibly prevalent in the ancient and classical traditions, with the added complexity that classical writers tended to borrow phrases, motifs, and sometimes entire passages from earlier works without citation. References may be direct, as Shakespeare's line from *The Tempest* ("How beauteous mankind is! Oh *brave new world*, That has such people in't." (emphasis added)) inspiring the title of Aldous Huxley's dystopian novel. Or references may be indirect, as in the 1999 film, *10 Things I Hate About You*, a modern adaptation of Shakespeare's *The Taming of the Shrew*.

Significant computational work has been done on detecting intertextuality in the works of Homer (Konstantinidou et al., 2024), early Christian writings (Moritz et al., 2016; Lee,

2007), classical Latin poetry (Bamman and Crane, 2008; Dexter et al., 2017; Burns et al., 2021; Dexter et al., 2024), and Latin prose (Dexter et al., 2017). These works allow critical questions to be asked of ancient corpora: disentangling composite authorship (Lee, 2007; Dershowitz et al., 2015; Koppel et al., 2011), testing authorship hypotheses (Koppel et al., 2007; Lippincott, 2019; Kestemont et al., 2016), and characterising the evolution of literary traditions (Dexter et al., 2017).[7]

Most of the above work for detecting and characterising intertextuality has relied on stylometric analysis, which captures lexical and syntactic features to distinguish between authors (Mosteller and Wallace, 1963). However, some, particularly Burns et al. (2021), highlight that neural embedding spaces – specifically Word2Vec embeddings – capture latent literary features that replicate stylometric findings. This suggests that neural representations can encode intertextual structure, making them a promising tool for studying narrative structure computationally.

There has generally been growing interesting machine learning applied to ancient languages (Sommerschield et al., 2023), especially those of the classical world, for a variety of tasks like diachronic semantic analysis and dissecting polysemy in Ancient Greek (McGillivray et al., 2019; Perrone et al., 2019), building language models and toolkits for ancient languages (Fono et al., 2024; Koutsikakis et al., 2020; Coffee et al., 2013), restoring ancient texts (Assael et al., 2019) and deciphering ancient scripts (Luo et al., 2021, 2019). These studies indicate that neural representations can capture rich linguistic and stylistic patterns in historical corpora.

2.4.2.1 Biblical Intertextuality: A Special Case

Although often treated as one book, the Bible is in fact a library of texts spanning multiple genres, authors, and centuries. Many individual texts are viewed by scholars as composite works themselves, redacted over centuries by multiple authors (Friedman, 2019; Alter, 2011). Despite this diversity, there is a striking coherence with which its texts use and build upon a common set of images. For instance, the apocalyptic visions in the books of Daniel and Zechariah resurface in the Revelation of John, while theological concerns—such as divine justice and redemption—thread through the prophetic writings.

This makes Biblical intertextuality a special case: the Bible is *already* an intertextual network, referencing itself extensively. This poses a challenge when studying text reuse in later works, as distinguishing between an external reference and the Bible's own internal

[7]There are some inherent challenges in identifying non-literal textual references in situations where there are no ground truth labels, such as is the case for ancient texts. Manjavacas Arevalo et al. (2021) looks at inter-annotator agreement when labelling Biblical text reuse in the writings of Bernard of Clairvaux (writing in Latin) and Clement of Alexandria (writing in Greek). Annotator familiarity with the source text (i.e. books of the Bible) and the inherent layered nature of Biblical intertextuality proves a significant challenge.

cohesion can be difficult. It is broadly assumed that the Christian New Testament is particularly referential to the Hebrew Bible/Old Testament[8](Bamman and Crane, 2008), and modern Bible editions reinforce this through cross-reference annotations. However, translators may have amplified or diminished these intertextual links—either intentionally or inadvertently.

In Chapter 4, I investigate the extent to which Biblical intertextuality is preserved across translations in two dimensions: (1) human and machine translations, and (2) across languages with varying resource levels. This analysis sheds light on whether deep learning models can successfully retain intertextual structures across linguistic boundaries or whether translation introduces distortions that obscure original references.

By viewing narratives as networks of intertextual connections, we gain deeper insight into how texts encode, transmit, and reinterpret meaning. The ability of NLP models to detect and interpret these connections remains an open challenge—one with profound implications for both literary scholarship and machine reading comprehension.

As Minsky (1974) notes:

> "Translation affords an opportunity to observe defaults at work."

Intertextuality presents unique challenges in translation, as implicit references may be preserved, transformed, or lost depending on linguistic and cultural context. Examining these patterns computationally allows us to test whether deep learning models can successfully model intertextuality and preserve it downstream tasks like translation, summarisation, and question-answering.

2.4.3 Narrative as Structured Repetition

Narratives exhibit recurring, recognisable elements drawn from a common set of conventions. These conventions are distinct from direct intertextual references to influential works, though foundational texts are often wellsprings for the conventions themselves. For example, modern fantasy novels frequently incorporate tropes pioneered by Tolkien's *The Lord of the Rings*, not as an overt homage, but rather because the popular trilogy was foundational for the modern genre. As Genette (1997) notes, narratives are deeply informed by prior structures. Booker (2004) attempts to reduce all stories to seven basic plot types, arguing that characters and plot events ultimately revolve around a central hero[9]. While this is overly simplistic, it correctly highlights the tendency of narratives

[8]Religious canon is a divisive subject, and textual transmission is a tedious one. For the purposes of this work, I refer to the Masoretic Text (MT) as the Hebrew Bible or Old Testament, and the 27 books affirmed as authoritative by the Council of Hippo in AD 393 as the (Christian) New Testament.

[9]Hagan (2019) further extends the idea of seven basic plot to plot structures in the Bible.

to reuse familiar structures. Within specific genres or cultures, common building blocks serve as cognitive scaffolds that facilitate comprehension.

I use *typology* as a general term for these recurring narrative structures, although this obscures the differences in plot-centric templates (story structures) and character-driven patterns (*dramatis personae* and archetypal roles). In the following sections, I provide an overview of typology as it has been developed in narratology, theology, sociology, and epistemology. My goal is to demonstrate that **typology is a fundamental organising principle of human communication.**

For NLP, the key question is whether language models implicitly encode knowledge of these structures, just as they internalise implicit world knowledge from pre-training on real-world data. In this thesis, I evaluate whether neural embeddings capture these high-level narrative structures.

2.4.3.1 Figura

While modern readers may take it for granted that genres, character types, and plot structures are recognisable and describable, it is not self-evident that human culture should organise narratives in such a structured and categorical manner. In the ancient world, the relationship between model and copy—the idea that an object could be patterned on prior forms—was a subject of deep philosophical inquiry. The very notion that a *prototype* (literally, 'first type') could exist, giving rise to later instances, was explored extensively in classical and theological traditions. Auerbach (1959) traces the evolution of the term *figura* – the abstract form of a thing – from classical philosophy through to a medieval view of history. In the classical world, poets and thinkers explored 'shades of meaning between model and copy, in changing form, and in the deceptive likenesses that walk in dreams' (p.21). The concept took on new significance in early Christian thought, particularly in the writings of Tertullian, Origen, and Augustine, where *figura* became a hermeneutic tool for interpreting sacred texts. Sirin and Lippincott (2024) have used diachronic analysis to empirically track this semantic shift of *figura* from the early Classical period through the Christian era, revealing a transition from a primarily tangible sense to an increasingly procedural one.

A figural interpretation of of the Bible views characters, events, and themes present in the Old Testament as prophetic foreshadowing for characters, events, and themes in the New Testament. Particularly influential in Christian theology is the idea of the prefiguring of Christ in Old Testament patriarchs like Moses and Joshua. Crucially, '*figura* is something real and historical which announces something else that is also real and historical. The relation between the two events is revealed by an accord or similarity' (Auerbach, 1959). That is, it is not necessary to assume that secondary retellings of narratives must be metaphorical reference to the original; things may happen the same

way twice.

This structured, hierarchical relationship between events—where one narrative element prefigures another—is strikingly different from the way language models learn representations. Instead of a flat statistical mapping of token co-occurrences, figural interpretation implies a latent graph structure, where past and future events are linked by abstract narrative logic. This structure maps more closely onto known recall structures in the human brain (Lee and Chen, 2022). Understanding whether LMs can implicitly encode such structured relationships remains an open question in NLP.

2.4.3.2 Type-Scene

One way in which narratives may share an 'accord' is in the utilisation of recurring stereotyped episodes, known as *type-scenes*. Unlike archetypal characters, which focus on roles and functions, type-scenes are procedural in nature—structured templates for key narrative moments. Clark (2004) defines them as: 'recurring situations which are narrated according to a more or less fixed pattern.'

The notion of a type-scene was first proposed by Arend (1933) to explain the formulaic composition broadly observed in Homeric epics. It is now widely accepted that type-scenes function as a mnemonic device, preserving the *Odyssey* and other oral traditions until their transition to a written medium Clark (2004) .

Alter (1978) later extended this concept to Biblical narrative, observing that 'there is a series of recurrent narrative episodes attached to the careers of biblical heroes that are analogous to Homeric type-scenes in depending on the manipulation of a fixed constellation of predetermined motifs'. As an example, Alter highlights the recurring betrothal scene, which follows a recognisable procedural schema:

A future bridegroom travels to a foreign land.

He encounters a maiden by a well.

The maiden draws him water.

She brings home the news of the stranger's arrival.

The betrothal is arranged between families.

Alter argues that literary conventions provide the backdrop against which authors can either meet or subvert reader expectations to create meaning (Alter, 1983). When Jacob remains at home while his servant travels to seek a wife for his master, it underscores Jacob's passivity, a character trait which will also feature in later scenes (Alter, 1978). Viewed in this light, the presence, absence, or transformation or an expected type-scene

Slot	Rebekah (Genesis 24)	Rachel (Genesis 29)
Bridegroom	Isaac (represented by Abraham's servant)	Jacob (acting on his own behalf)
Maiden	Rebekah	Rachel
Well Encounter	Rebekah arrives first; offers water proactively to servant **and** camels (fulfilling sign).	Jacob sees Rachel; rolls away stone from well himself; waters her sheep.
Initiative	Servant prays for a sign; Rebekah fulfils it unprompted.	Jacob takes initiative; physical action (removing stone) replaces verbal request.
Hospitality	Rebekah runs to tell her family; Laban invites servant in.	Jacob kisses Rachel; Rachel runs to tell her father; Laban welcomes Jacob.
Divine Sign / Negotiation	Fulfillment of a divine test; marriage arranged by family.	No divine sign; Jacob bargains for Rachel's hand (7 years labour).
Outcome	Immediate betrothal; Rebekah leaves family promptly.	Delay and complication: Jacob is deceived and marries Leah first.

Table 2.2: Two instantiations of the *Betrothal-by-Well* frame in Genesis. While the underlying frame remains constant, the slot fillers vary — highlighting differences in initiative, divine involvement, and narrative tone.

efficiently conveys significance. in Table 2.2, I represent the sequence of betrothal events in Genesis 24 and 29 as Minsky-style frames.

Subsequent scholarship has put forward or further elucidated Biblical type-scenes, including prophetic concealment (Britt, 2002), betrothal (Fuchs, 1987), single combat (Press, 2016), the introduction of foreigner to court (Köstenberger and Goswell, 2023), and theophany (Savran, 2003). Despite extensive literary analysis, type-scenes have not yet been studied computationally. In Chapter 5 I address this gap, investigating whether language models can recognise and retrieve type-scene as structured narrative patterns.

Type-scenes share a marked similarity with the *scripts* of Schank and Abelson (1975) : 'a script is a predetermined, stereotyped sequence of actions that define a well-known situation.' However, in contrast to Schank and Abelson's 'very boring little [stories]', type-scenes mark deeply significant structured events, rather than commonplace scenes. Related work on learning narrative event chains has not considered pre-existing significance to the event structure (Chambers and Jurafsky, 2008, 2009; Chambers, 2013). One fascinating strain of research on modern corpora, inspired by screenplay writing theory, models *turning points*, key transitions in a film's plot, using an annotated film script

dataset (Papalampidi et al., 2019, 2020, 2021). These efforts aim to develop extractive screenplay summarisation techniques and structured models for plot analysis. Type-scene detection also bears a distant similarity with work on aspect-based document similarity (Ostendorff et al., 2020; Ginzburg et al., 2021), which uses structural similarity to inform document-level similarity ranking.

2.4.3.3 Chiasmus

An important discourse-level repetition structure common in ancient text is *chiasmus*, a rhetorical structure whose key characteristic is inverse repetition, as in the phrase, 'Ask not what your country can do for you, but what you can do for your country'. Bothwell et al. (2023) consider it as a restricted version of a more broad phenomenon of rhetorical parallelism. Most work on chiasmus detection has been done on modern English corpora (Dubremetz and Nivre, 2015, 2017), with Schneider et al. (2021) examining line-level chiasmus in German text. Modern chiasmus detection is a needle-in-a-haystack problem (Dubremetz and Nivre, 2015), whereas ancient rhetoric frequently employs more structured parallelism, including chiasmus, much more frequently (Bothwell et al., 2023).

Yet despite its prominence in ancient rhetoric, chiasmus remains under-explored in computational studies of classical and Biblical texts. In this thesis, I begin by investigating whether embedded representations can surface smaller-scale discourse phenomena in classical and Biblical texts, using chiasmus as a test case. As a structured form of rhetorical repetition, chiasmus allows us to assess models' sensitivity to latent narrative structure beyond surface-level token patterns. This sets the stage for subsequent chapters, where I examine how models handle larger-scale intertextual alignments across translations, and ultimately probe their capacity for analogical and conceptual reasoning over entire narratives.

2.5 Summary

This chapter has laid the historical and conceptual groundwork for understanding narrative as a form of structured knowledge representation, and for evaluating the extent to which current AI systems — particularly large language models — approximate this representational capacity. While some work (e.g. Bender and Koller (2020)) expresses justified skepticism regarding the semantic competence of LMs, there remains strong foundational and practical value in probing whether these models *implicitly* encode hierarchical, compositional structures akin to those humans use when processing narrative.

This thesis investigates that question through two primary lines of inquiry. First, I revisit longstanding assumptions about distributed representations — that word or sentence embeddings encode systemic conceptual mappings — by applying modern decoder-only

architectures with extended context windows to narrative-level tasks. While such models show improved performance on document-level benchmarks, particularly in narrative understanding, I identify several persistent limitations suggesting the need for more explicitly structured representations.

Second, I empirically test whether LMs encode structured patterns characteristic of human storytelling. Specifically, I ask:

1. Do LMs learn to identify internal, localised rhetorical structures (e.g., chiasmus, type-scenes)?

2. Can they detect *cross-textual* structural parallels — that is, intertextuality grounded in shared frames rather than surface features?

3. Most critically, do LMs make *analogical decisions* based on deep conceptual structure, or are such judgments driven primarily by shallow lexical or syntactic similarity?

These questions form the backbone of the thesis. Taken together, they constitute a comprehensive evaluation of whether unstructured embedding spaces — as learned by large neural networks — can support the kind of structured, hierarchical, and analogical reasoning required for narrative understanding. The results have implications not only for cognitive modelling but also for downstream tasks that depend on robust, explainable representations of meaning.

2.6 Thesis Outline

This thesis is structured as follows:

Chapter 2 discusses theoretical and technical context for this work. I expound the relevant insights from narrative theory, deep learning, and classical studies which underpin and motivate this work.

In Chapter 3, I present the first empirical case study, exploring whether pretrained language models can identify chiasmus—a mirrored, anti-parallel rhetorical structure commonly found in ancient texts, particularly the Hebrew Bible. Chiasmus poses a challenge for NLP systems because it is often subtle and hierarchical, rather than overtly lexical. We test whether statistical models operating over unstructured embedding spaces can detect these non-linear patterns. Our findings suggest that pretrained models do encode some sensitivity to rhetorical structure, raising broader questions about how deep models internalise literary form.

Chapter 4: this chapter turns to *intertextuality*—the ways in which texts echo, allude to, or revise one another over time. We examine whether multilingual embedding models

capture intertextual relationships between verses across translations, using a parallel Bible corpus as a testbed. Intertextual links play a central role in meaning-making across religious and classical traditions, especially through typology. By evaluating whether semantically related verses remain close in embedding space across languages, we probe the limits of multilingual representation and ask whether such models can track meaning-in-relation, not just meaning-in-isolation.

I investigate another facet of classical narrative in Chapter 5, wherein I explore a case study in *type-scenes*—recurring narrative templates in Biblical literature—as a test of higher-level structural recall. Unlike tasks focused on sentence-level meaning, type-scene recognition requires procedural and relational understanding across multiple events. We curate a small test set of type-scenes based on Biblical scholarship and evaluate the performance of various pretrained models on a retrieval task. Surprisingly, models perform best when the task is treated as pure semantic retrieval, without additional structural supervision—but even then, performance remains objectively weak. These results underscore both the representational limitations of LMs and the difficulty of recovering abstract narrative form from unstructured embeddings.

Finally in Chapter 6, I move beyond behavioural evaluation, and adopt a probing-based methodology to investigate how and where narrative concepts are stored in pretrained models. Using diagnostic classifiers, we test whether abstract features such as rhetorical parallelism and narrative analogy are linearly recoverable from LLM embeddings. While some concepts are robustly encoded at span level, especially in smaller contexts, performance drops at document scale. Notably, we observe a paradox: as model size increases, downstream prompting performance improves, but the information retrievable via probing does not scale proportionally. This raises questions about the internal representation of narrative knowledge and its accessibility to external tools.

Chapter 7, then, offers reflections on the findings across all experimental studies, summarising what they reveal about the narrative competence of large language models. It also outlines key directions for future research: from structure-aware pretraining and improved probing techniques, to alternative approaches for modelling typology, analogy, and cultural memory in machine learning systems. Ultimately, the chapter argues that understanding how machines process narrative is not only an NLP challenge, but a way of illuminating the deeper cognitive structures of human storytelling itself.

3

Unsupervised Discovery of Rhetorical Structure in Ancient Hebrew

This chapter[1] presents the first empirical case study in my investigation of whether pre-trained language models encode structured patterns characteristic of human storytelling. I focus here on chiasmus, a rhetorical structure in which elements appear in a mirrored, anti-parallel pattern. In ancient texts, chiasmus often serves to foreground contrast, create symmetry, or provide rhetorical emphasis. It is a hallmark of ancient religious literature, particularly in the Hebrew Bible and, to a lesser degree, early Christian texts[2].

From a linguistic perspective, chiasmus is an instance of recursive parallelism: a formal structure that arises not from local co-occurrence but from global symmetry and conceptual alignment. As such, it presents a useful test for the limits of modern language models. Can models trained on token-level prediction tasks detect these deeper rhetorical structures? Or do their representations remain confined to surface-level features like lexical similarity and word order?

[1]This work was originally published as a main conference short paper at NAACL 2025 in the Computational Social Sciences and Cultural Analytics track (McGovern et al., 2025b). I have modified the content and framing to fit the thesis and made other minor revisions. My co-authors on the paper (Drs Hale Sirin and Tom Lippincott) acted in an advisory role for this project.

[2]While there are a few recognised instances of quasi-chiastic structures in the Christian New Testament (see, for example, 'Markan Sandwiches' (Edwards, 1989)), these are often less strict and happen on a broad, organisational level such as the structure of Paul's letter to Philemon (Heil, 2001).

In this chapter, I test the hypothesis that chiasmus, despite its abstract and often subtle structure, can be discovered by statistical models operating over unstructured embedding spaces. If successful, this would provide evidence that pretrained language models learn not just to associate co-occurring words, but to encode the hierarchical, non-linear structures that underlie narrative and argumentation.

3.1 Poetic Devices in Classical Literature

Structured rhetorical forms are central to the literary traditions of the ancient world. These devices offer not just aesthetic value but also forensic clues about authorship, tradition, and cultural memory. For example, Proverbs chapter 30 features with a distinctive numerical refrain:

> '… Three things are never satisfied; four never say, "Enough": …' (Proverbs 30:15)
>
> '… Three things are too wonderful for me; four I do not understand: …' (Proverbs 30:18)
>
> 'Under three things the earth trembles; under four it cannot bear up: …' (Proverbs 30:21)

This kind of poetic device exemplifies a patterned form of expression used to encode emphasis, rhythm, and memory[3]. It stands in contrast to modern poetic traditions — such as Walt Whitman's blank verse — which often favour free-form and unstructured expression. Classical poetry, by contrast, tends to be highly formal and pattern-driven, making it particularly amenable to statistical analysis. However, this same formalism poses practical challenges: these patterns are rarely annotated, and their detection often relies on expert interpretation.

Chiasmus — the arrangement of textual units in a mirror-like structure — has long captured the attention of Biblical scholars. The term comes from the Greek letter *chi* (χ), evoking the 'X'-shaped symmetry of classic forms (e.g., ABB'A'). One widely cited modern example is John F. Kennedy's rhetorical reversal[4]:

> Ask not what *your country can do for you,*
> but what *you can do for your country.*

[3]This specific numerical format occurs only in Proverbs 30, a chapter attributed to Agur son of Jakeh, suggesting that it may represent a stylistic deviation from broader Hebrew poetic norms.

[4]Technically, this example is an instance of a subset of chiasmus called antimetabole, discussed in Section 3.2.1. However, this phrase stands as a guide for intuition.

A	[...] Let them be turned back and disappointed who devise evil against me!
B	Let them be like chaff before the wind, with the angel of the LORD driving them away!
B'	Let their way be dark and slippery, with the angel of the LORD pursuing them!
A'	For without cause they hid their net for me; without cause they dug a pit for my life.

Table 3.1: **Chiastic structure of Psalms 35:4-7.** Pairs (A, A') and (B, B') form an 'X' pattern, exhibiting repeated phrases and conceptual links.

This syntactic and conceptual reversal exemplifies chiastic structure. While such devices are relatively rare in modern English prose — and often confined to elevated oratory (Bothwell et al., 2023) — they are pervasive in ancient literature. For instance, Dubremetz and Nivre (2015) found only seven clear chiastic structures across nearly 200 speeches by Winston Churchill. In contrast, Biblical texts, especially Hebrew poetry, are replete with examples of this form (Lowth, 1989; Welch, 1981).

As illustrated in Table 3.1, Psalms 35:4-7 exhibits inverted parallelism: lines A and A' contain conceptual and lexical parallels (e.g., 'devising evil' vs. 'digging a pit'), while B and B' feature symmetrical imagery involving the angel of the Lord. Colours highlight these analogies.

Yet despite wide recognition of chiasmus as a literary and rhetorical phenomenon, scholars remain divided on its precise definition, function, and prevalence. Some propose that it aids characterisation in narrative passages (Assis, 2002), while others emphasise its poetic role in the Psalms (Martin, 2018). These debates reveal a deeper issue: chiasmus detection remains largely qualitative, relying on expert judgment and subjective interpretation.

In this chapter, I introduce a computational method for detecting chiasmus in Biblical texts that operationalises a central question of this dissertation: whether structured rhetorical patterns are encoded in unstructured vector representations of the kind learned by large, pretrained language models. Unlike rule-based approaches, this method avoids hand-crafted templates. Instead, it leverages statistical similarity among sentence embeddings, enabling it to capture conceptual and semantic symmetry.

This approach builds on prior work in computational rhetoric (Dubremetz and Nivre, 2017; Schneider et al., 2021), while addressing several limitations. Earlier systems often relied on predefined syntactic templates, were language-specific (i.e. English or German), or focused on shallow symmetry at the phrase or clause level. In contrast, my method is:

- Agnostic to language structure, operating in multilingual embedding spaces,
- Capable of handling arbitrary-length passages rather than intraline symmetry, and

- Adaptable to ancient languages, where syntactic parsing may be unreliable.

More broadly, this chapter offers a case study in how statistical models might recover structures that were historically understood as meaningful. If successful, it would support the broader thesis that modern neural models, despite lacking explicit knowledge of syntax or rhetoric, may implicitly encode higher-order narrative structures. Chiasmus provides a stringent and culturally significant test of this claim — one that bridges computational modelling with ancient literary analysis.

3.2 Formalising Chiasmus

3.2.1 Task Description

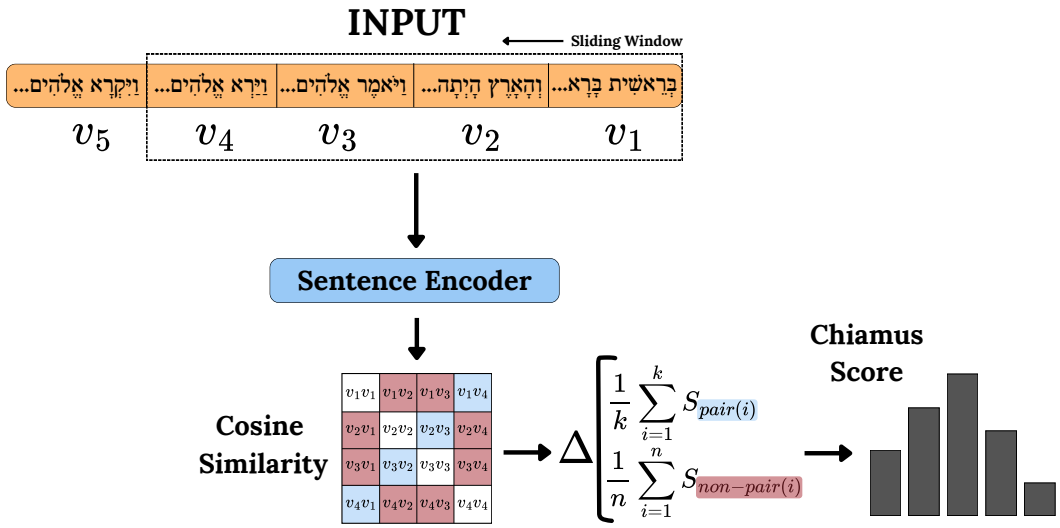


Figure 3.1: **Experimental setup for chiasmus detection.** In a sliding window fashion, I encode verses (or half-verses) with a sentence encoder and then calculate the delta of the pair similarities and non-pair similarities to create a score distribution over all sets of verses. In this illustration, the sliding window operates right-to-left to follow Hebrew orthography.

Chiasmus, as a literary device, exhibits properties similar to those used in computational models of textual allusion. In particular, the features identified by Bamman and Crane (2008) for the detection of literary references—lexical overlap, word order, syntactic alignment, semantic similarity, and even phonetic or metrical resonance—are also relevant here. Conceptually, a chiasmus can be understood as a kind of localised textual echo: a structured pattern that relies on parallelism and mirroring.

This resemblance is especially pronounced in ancient, morphologically rich languages like Biblical Hebrew and Latin. In such languages, the free word order amplifies the significance of syntactic and positional repetition, since similarity in structure is not merely a by-product of grammatical constraint. Chiastic forms may therefore exploit word order as a deliberate poetic signal rather than a grammatical necessity.

However, chiasmus presents a more nuanced structural challenge than straightforward repetition or paraphrase. Unlike standard parallelism—which typically aligns similar ideas in sequence—a chiasmus reverses the order of elements, often pairing not only similar units but also antithetical ones. That is, the mirrored elements (e.g., A $\sim$ A', B $\sim$ B') may reflect either semantic similarity (e.g., repetition, synonymy, or paraphrase) or semantic contrast (e.g., antonymy, opposition of agency, or moral inversion). This rhetorical juxtaposition is especially characteristic of Biblical and classical texts, where chiasmus can signal thematic reversal or ironic contrast.

A robust formalisation must therefore accommodate both similarity and structured contrast between corresponding elements. Importantly, this alignment must be selective: elevated similarity or contrast should occur *only* between mirrored elements in the chiastic pattern. Non-corresponding lines (e.g., A $\sim$ B or B $\sim$ A') should not exhibit strong mutual similarity. This structural constraint is essential to distinguish meaningful chiasmi from looser repetitions such as genealogical lists, formulaic poetry, or catalogues, where broad similarity is distributed evenly across units with no internal mirroring.

Chiasmi may occur in even or odd forms. Even chiasmi (e.g., `ABBA`) symmetrically balance their mirrored halves. Odd chiasmi (e.g., `ABCBA`) include a central, unmatched unit—often interpreted as a semantic focal point. In narrative or poetic settings, this central element frequently carries the thematic core or interpretive hinge of the passage. Such structures reflect a *circular* rather than *linear* logic: the 'point' lies not at the end (as in Western narrative convention), but in the centre (Douglas, 2007). This circular structure can serve theological, poetic, or ritual functions—framing the heart of a covenant or embedding emphasis within the narrative ring.

It is also important to distinguish chiasmus from *antimetabole*, a narrower rhetorical figure. Whereas chiasmus involves the reversal of elements at multiple linguistic or conceptual levels (e.g., structure, meaning, grammar), antimetabole refers more strictly to the inversion of exact words or phrases (e.g., "eat to live, not live to eat"). Several of the works discussed below focus on antimetabole, which can be considered a degenerate or lexical case of chiasmus. The formalisation must be general enough to capture both surface-level and conceptual symmetry, without reducing chiastic structure to mere word reversal.

We can now formalise the objective: I aim to detect passages that satisfy the following criteria:

1. **Mirrored Pairing**: The passage must contain a sequence of units that exhibit high similarity or structured contrast in a symmetrical pattern (e.g., A $\sim$ A', B $\sim$ B').

2. **Selective Alignment**: Non-paired elements within the same window must not exhibit comparable levels of similarity, ensuring that the mirroring is non-trivial.

3. **Structural Boundedness**: Candidate chiasmi should be self-contained, not spanning across major textual divisions (e.g., books or chapters).

4. **Optional Central Focus**: In odd-length candidates, the central unit (C in `ABCBA`) is left unmatched and may exhibit elevated salience or semantic weight.

The chiasm scoring pipeline is illustrated in Fig. 3.1.

(1) Valid Chiastic Similarity	**(2) Valid Chiastic Opposition**	**(3) Invalid Repetition**
A "The LORD is my shepherd; I shall not want."	**A** "The wise will inherit honor,"	**A** "Reu fathered Serug,"
B "He makes me lie down in green pastures."	**B** "but fools get disgrace."	**B** "Serug fathered Nahor,"
B' "He leads me beside still waters."	**B'** "but shame is the promotion of fools."	**B'** "Nahor fathered Terah,"
A' "You prepare a table before me..."	**A'** "The fool says in his heart, 'There is no God.'"	**A'** "Terah fathered Abram."
Sim(A, A') = 0.81↑ Sim(B, B') = 0.74↑	Sim(A, A') = 0.67↑ Sim(B, B') = 0.54↓	Sim(A, A') = 0.96↓ Sim(B, B') = 0.92↓

Table 3.2: **Comparison of Chiastic Structures.** High cosine similarity between corresponding units (A, A' and B, B') supports valid chiasm (col. 1), while antonymic inversion reflects structured opposition (col. 2). Genealogical repetition (col. 3) exhibits high similarity but lacks structural intent. Cosine similarity scores here are simulated.

3.2.2 Related Work

Existing approaches to chiasmus detection are few but methodologically instructive. Most prior work applies heuristic or feature-based methods to shallow chiastic structures—typically flat, two-part inversions (AA'). In contrast, my approach leverages modern contextual embeddings to enable detection of nested, semantically rich chiasmi. However, I retain key aspects of these earlier pipelines: namely, an unsupervised scoring function applied over a sliding window. This section provides a detailed account of four relevant works and situates this model within this methodological lineage.

Dubremetz and Nivre (2015): This foundational work on rhetorical figure detection targets chiasmus in the Europarl corpus. Their approach combines linguistically motivated heuristics with a sliding window of fixed length. Candidate windows are scored

based on additive features drawn from four classes: (1) stopword and punctuation density, (2) relative size of the window, (3) repeated or inverted n-grams, and (4) hand-curated lexical clues. Each feature class is weighted based on manual tuning on a development set.

To identify inversion candidates, they apply TreeTagger to extract pairs of repeated or reversed lemmata within a given token span. For each candidate, a scoring function computes the degree of rhetorical inversion. They manually annotate the top 200 examples from each feature class, yielding a final corpus of 1200 excerpts—of which only 31 contain true chiasmi. This low prevalence highlights a key challenge of English data: chiasmus is a needle-in-a-haystack problem. By contrast, in curated Biblical or classical texts, chiasmus is a known, culturally embedded pattern, which increases the base rate of true positives and justifies my focus on higher-precision modelling.

Dubremetz and Nivre (2017): In a follow-up paper, the authors apply supervised learning to the same 1200-instance corpus. Using logistic regression, they train a classifier over the same manually engineered features and achieve an overall F1-score of 78.3 despite the scarcity of positive examples. Interestingly, the learned weights of the model closely track their original hand-tuned heuristics, reinforcing the value of linguistically informed features.

However, a key limitation remains: their approach only models flat chiasmi with simple two-part mirroring. There is no support for multi-part or nested patterns (e.g.,`ABBA`, `ABCBA`). My method generalises the structural search space by allowing arbitrary nesting within the sliding window, scored via symmetric pairwise alignments.

Schneider et al. (2021): More recently, Schneider and colleagues introduce *semantic* features into the chiasmus detection pipeline. Like Dubremetz and Nivre, they apply a 30-token sliding window, but identify candidates using part-of-speech tag inversions rather than lemmata. They retain similar lexical and structural features, but crucially add two more types: (1) lexical n-grams to counteract the loss of surface content from PoS-only filtering, and (2) embedding features, derived from SpaCy models[5].

Their corpus consists of Friedrich Schiller's German dramas, manually annotated to include 45 general chiasmi and 9 antimetaboles. From over a million syntactic inversions, they sample 4000 negative examples. Their results confirm the added value of semantic representations: embedding features produce the largest improvement in detecting general chiasmus. Lexical features, by contrast, are more important for the stricter task of antimetabole detection.

They also validate generalisability by applying their scoring model to unseen Schiller texts. Of the top 100 candidates, the full feature set with embeddings doubles the number of correct detections—10 compared to 5 using Dubremetz-style features. This motivates

[5] `en_core_web_lg` and `de_core_news_lg`

my own use of semantic similarity via contextual embeddings, but instead of SpaCy static vectors, I use pretrained transformer models to capture richer contextual nuance.

Bothwell et al. (2023): In a broader shift toward rhetorical structure, Bothwell et al. propose a general framework for detecting rhetorical parallelism, including chiasmus, across languages and genres. Their annotation scheme accommodates nested parallel structures, although their trained models focus only on flat alignments. They construct a tagging scheme for parallel units (BIO-style) and apply it to both Latin sermons (e.g., from Augustine of Hippo) and Chinese student essays.

A key finding from their model comparisons is that BERT-style pretrained models outperform traditional feature-based methods. Specifically, LatinBERT outperforms embeddings learned from scratch and a Word2Vec+BiLSTM baseline on rhetorical parallelism detection. This strongly supports my design decision to use a pretrained transformer (e.g., BERT or a similar model) as the backbone for the embedding-based similarity scoring—in lieu of explicitly modelling syntactic, lexical, or n-gram features.

In summary, my approach adapts the unsupervised detection pipeline of Dubremetz and Nivre (2017) and Schneider et al. (2021)—based on a sliding window and scoring function—but replaces engineered features with contextual embeddings from a BERT-like model. This allows us to scale up to nested, semantically diverse chiasmi while retaining interpretability and structural precision. I further justify this design choice by the demonstrated effectiveness of pretrained Transformers in related rhetorical detection tasks.

3.2.3 Operationalising Chiastic Structure

To detect candidate chiasmi, I begin by computing a cosine similarity matrix (S), where each entry (S_{ij}) reflects the semantic similarity between textual units (i) and (j), often a single verse. A candidate chiasmus consists of a sequence of mirrored pairs—for example, ($A \dots A'$), ($B \dots B'$), and so on, that define a symmetric pattern within a bounded window.

For each such candidate, I define the *chiasmus score*, (μ_{chiasmus}) as the average similarity between these hypothesised mirror pairs:

$$\mu_{\text{chiasmus}} = \frac{1}{k} \sum_{i=1}^{k} S_{\text{pair}(i)}$$

where pair(i) indexes the i-th matching pair in the hypothesised chiastic structure.

To assess distinctiveness, I also compute the mean similarity between all non-paired lines within the same window (e.g., ($S_{A,B}, S_{B,C'}$), etc.):

$$\mu_{\text{non-pair}} = \frac{1}{n} \sum_{i,j \in \text{non-pair}} S_{ij}$$

The final score for a candidate structure is then defined as the difference between the average similarity of the mirror pairs and that of the non-pairs:

$$\text{Final Score} = \mu_{\text{chiasmus}} - \mu_{\text{non-pair}}$$

This differential score favours candidate chiasmi that exhibit a salient internal structure distinct from the surrounding textual background.

To search across a text corpus, this scoring procedure is applied using a sliding window approach. Each window spans a fixed number of textual units N (e.g., verses), and within each window I search for potential chiastic pairings of varying depths (e.g., `ABBA`, `ABCBA`, etc.). I experiment with multiple values of $N \in [4, 8]$. To ensure interpretability and preserve narrative coherence, candidate windows are prohibited from crossing book or chapter boundaries.

After computing scores for all candidate structures, I standardise them by calculating their z-scores:

$$z_i = \frac{\mu_{\text{chiasmus},i} - \mu_{\text{chiasmus,mean}}}{\sigma_{\text{chiasmus}}}$$

where $\mu_{\text{chiasmus,mean}}$ and σ_{chiasmus} are the mean and standard deviation of all scores, respectively. I keep only those candidates which have a threshold of three standard deviations above the mean score as the model's predictions.

3.3 Data

My primary data source is the Translator's Amalgamated Hebrew Old Testament (TAHOT), based on the Leningrad Codex—the oldest complete manuscript of the Hebrew Bible. Due to the complex history of Biblical translation, modern English translations often follow a versification system that diverges from the earliest Hebrew manuscripts. For example, what appears as Genesis 31:55 in most English Bibles corresponds to Genesis 32:1 in the Leningrad Codex. I adhere to the Hebrew versification throughout, as this alignment is essential for recovering chiasmi as they may have been structured in the original text.

I analyse the text at two levels: verses and half-verses. In the Masoretic tradition, the cantillation symbol *atnach* (֣) typically marks a semantic and syntactic midpoint. I consider all words up to and including the word marked with the *atnach* as the first half, and the remainder as the second. This division reflects traditional pauses in liturgical reading and often corresponds to internal poetic or rhetorical balance.

Genesis 1:4 with semantic division at the atnach

Before atnach (֣):

וַיַּ֤רְא אֱלֹהִים֙ אֶת־הָאֹ֔ור כִּי־טֹ֑וב

And God saw the light, that it was good.

After atnach:

וַיַּבְדֵּ֣ל אֱלֹהִ֔ים בֵּ֥ין הָאֹ֖ור וּבֵ֣ין הַחֹֽשֶׁךְ׃

And God separated the light from the darkness.

Before embedding, I remove all vocalisations (*niqqud*) and cantillation symbols. In total, I analyse 23,213 verses, yielding 44,779 half-verses. For clarity and accessibility, I present English translations in all tables and figures, while all computational analysis is performed on the Hebrew source text.

3.4 Sliding Window Scoring

Building on the formalism described above, I apply the model across the Hebrew Bible, treating each verse or half-verse as a potential anchor point. For every such unit, I test candidate chiastic structures of length $N \in [4, 8]$ using a fixed sliding window approach.

Rather than attempt exhaustive annotation, I follow established practice in the literature (e.g. Dubremetz and Nivre (2017); Schneider et al. (2021)), and focus human effort on the most promising candidates. Specifically, I extract the top 50 scoring outputs for both the half-verse and full-verse settings. These outputs undergo manual annotation, detailed in Section 3.6.

3.4.1 Embedding models

I embed each verse or half-verse using the E5 multilingual sentence encoder (Wang et al., 2024). I adopt the 'small' variant (118M parameters), balancing speed and quality. This approach follows recent work by Burns et al. (2021), who use cosine similarity to surface intertextual echoes in Latin literature.

3.4.2 Why Not Use an LLM?

While large language models (LLMs), specifically instruction-tuned, decoder-only LLMs, dominate many NLP benchmarks, several factors make them ill-suited for chiasmus de-

		Half-Verse	Verse
Full Output	Num. Found	1896	879
	Top Book	Genesis	Numbers
	Avg. Length	5.93 ± 1.34	6.01 ± 1.38
	Avg. Score	0.32 ± 0.1	0.29 ± 0.08
Annotated	System Precision (Top-50)	0.60	0.80
	Cohen Kappa (κ)	0.76	0.89
	Top Genre	Narrative	Narrative

Table 3.3: **Summary of detected chiasmi.** 2700+ chiasmi were detected at the verse and half-verse level. The highest number of chiasmi was found in the Book of Genesis and Book of Numbers. Both the precision and the inter-annotator agreement increase for the verse-level chiasmi.

tection in the Hebrew Bible.

First, in preliminary tests, I found that LLMs often regurgitated copyrighted English translations (e.g., the ESV) when prompted with Hebrew verses. This behaviour likely stems from data contamination via online commentaries—many of which explicitly discuss chiastic structure. As a result, LLMs may unwittingly echo human analyses, undermining the goal of independent discovery (Balloccu et al., 2024).

Second, and more importantly, LLMs remain difficult to interpret. Biblical scholars—the primary audience for this tool—require methods they can trace, verify, and discuss. My model provides a mathematically grounded framework that surfaces candidate chiasmi without the opaqueness of black-box generation. By foregrounding structural criteria, my system invites scholarly critique rather than obscuring it.

The next section presents the system's outputs and evaluates their quality through both expert annotation and quantitative metrics.

3.5 System Output

Table 3.3 provides an overview of the system's output. At the half-verse level, the model identifies 1,896 chiastic structures, with an average length of 5.93 textual units ($\pm$1.34) and an average score of 0.32 ($\pm$0.1). At the verse level, 879 chiastic structures are identified, with an average length of 6.01 lines ($\pm$1.38) and an average score of 0.29 ($\pm$0.08). The book of Genesis contains the highest number of half-verse chiasmi, while Numbers yields the most verse-level chiasmi.

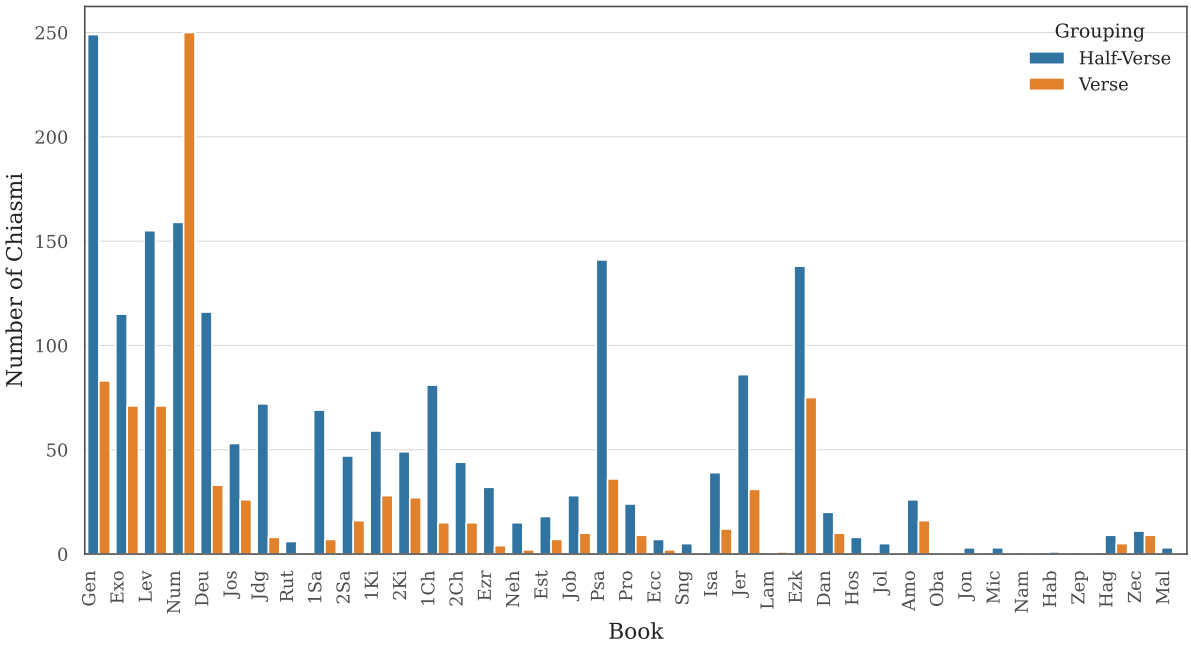


Figure 3.2: **Total number of chiasmi per book in the Old Testament at verse and half-verse level.** While some books tend to have more chiasmi overall, this figure shows whether verse-level or half-verse-level chiasmi are more prevalent in each book.

As visualised in Fig. 3.2, chiastic density is not evenly distributed. Half-verse chiasmi consistently outnumber their verse-level counterparts, though their relative frequency varies by book. Genesis, 1 Samuel, Judges, 1 Chronicles, Psalms, Jeremiah, and Ezekiel exhibit especially high rates of half-verse chiasmus. This pattern maps onto known literary and genre features: the Psalms, Jeremiah, and Ezekiel contain dense poetic material where chiastic parallelism is both expected and aesthetically integral; Genesis interweaves narrative with genealogical and covenantal formulae, fertile ground for structural mirroring.

The elevated counts in historical books such as 1 Samuel, Judges, and 1 Chronicles may reflect the system's sensitivity to recurring narrative templates—e.g., formulaic successions in royal accounts: 'X became king, reigned for Y years, and did evil in the sight of the Lord.'

3.6 Human Annotation

To assess precision, the 50 highest-scoring outputs at each level were annotated by myself and a co-author on the original publication, both of whom have graduate-level training in ancient languages and/or literature. While the identification process is fully in Hebrew, annotators used a literal English translation preserving Hebrew word order for ease of inspection. Each candidate was labeled using a three-class rubric:

Chiastic repetition clear structural reversal, lexical or semantic.

Non-chiastic repetition repetition without chiastic order.

No repetition no repetition.

Annotators reached strong agreement, with Cohen's κ at 0.76 for half-verses and 0.89 for verses. Additionally, ambiguities arose rarely: annotators disagree on only two half-verse examples when distinguishing between 'no repetition' and 'chiastic repetition'. Six additional cases sat on the borderline between chiastic and non-chiastic repetition, underscoring the fuzziness of the boundary between formal inversion and loose recurrence when only incorporating semantic features.

Among the model's top 50 predictions, 60% of half-verse outputs and 80% of verse-level outputs were judged chiastic by both annotators, reflecting strong alignment between model scores and human judgments.

A notable secondary finding is the frequent emergence of non-chiastic repetition in top-scoring outputs: these often involve ritual, legal, or genealogical language with formulaic structure. This category may be of separate interest to scholars studying the formal features of Biblical Hebrew. I find 29 such cases among the top 100 candidates. Only 3% of top passages fell into the 'no repetition' class, indicating that the model is highly successful at filtering out completely irrelevant matches.

3.7 Qualitative Analysis

A	And there was evening and there was morning, the fourth day.
B	And God said, "Let the waters swarm with swarms of living creatures, and let birds fly above the earth across the expanse of the heavens."
C	So God created the great sea creatures and every living creature that moves, with which the waters swarm, according to their kinds, and every winged bird according to its kind. And God saw that it was good.
B'	And God blessed them, saying, "Be fruitful and multiply and fill the waters in the seas, and let birds multiply on the earth."
A'	And there was evening and there was morning, the fifth day.

Table 3.4: **English translation of a positive example of a chiasmus automatically detected by this method.** Gen 1:19-23 (ESV)

Several high-quality examples recovered by this model align with traditional literary analysis. One appears in Genesis 1:19–23 (Table 3.4), where a five-part chiasmus highlights thematic symmetry between divine command and evaluation ('And God saw that it was

good'). The inversion of motifs—creation of sea and sky life nested within divine speech—emphasises order and divine approval, echoing interpretations of Genesis as structured poetic theology.

A	And Esau said to Jacob, "Let me eat some of that red stew, for I am exhausted!" (Therefore his name was called Edom.)
B	Jacob said, "Sell me your birthright now."
C	Esau said, "I am about to die; of what use is a birthright to me?"
B'	Jacob said, "Swear to me now." So he swore to him and sold his birthright to Jacob.
A'	Then Jacob gave Esau bread and lentil stew, and he ate and drank and rose and went his way. Thus Esau despised his birthright.

Table 3.5: **The story of Jacob stealing Esau's birthright.** English translation of a chiasmus automatically detected by our method. Gen 25:30-34 (ESV)

Another salient example is the story of Jacob and Esau (Table 3.5). The structure mirrors the reversal of birthright exchange, underscoring themes of deception and covenantal destiny.

A	And God said, "This is the sign of the covenant that I make between me and you and every living creature that is with you, for all future generations:
B	I have set my bow in the cloud, and it shall be a sign of the covenant between me and the earth.
B'	When I bring clouds over the earth and the bow is seen in the clouds,
A'	I will remember my covenant that is between me and you and every living creature of all flesh. And the waters shall never again become a flood to destroy all flesh.

Table 3.6: **God's covenant with Noah after the flood.** English translation of a chiasmus automatically detected by our method. Gen 9:12-15 (ESV)

A third example from Genesis 9:12–15 (Table 3.6) frames God's post-diluvian covenant with repeated references to the 'bow' and divine remembrance, embedding the symbol of restoration within a structured theological promise.

More broadly, many instances of divine speech exhibit chiastic or poetic structure. This pattern likely reflects stylistic conventions in Biblical Hebrew, where symmetry signals rhetorical solemnity or theological emphasis. These findings support existing scholarship linking chiasmus to theological intent and formal artistry in ancient narrative and law.

3.8 Conclusion

This chapter has explored whether embedding spaces from pretrained, deep contextualised language models can support the discovery of structured rhetorical patterns, using the formal device of chiasmus in ancient religious texts as a diagnostic case. My findings suggest that even in the absence of explicit structural supervision, computational methods are capable of identifying these mirrored, hierarchical forms with surprising fidelity. This observation carries dual significance: it offers scholars in theology, literature, and digital humanities a promising tool for uncovering subtle rhetorical patterns, and it provides early empirical evidence that LLMs encode aspects of internal narrative structure—patterns that go beyond surface-level co-occurrence or shallow syntactic symmetry.

By recovering chiastic symmetry from pretrained representations alone, I begin to answer one of this thesis's central questions: To what extent can large language models represent the recursive, hierarchical, and formal properties of human storytelling? Chiasmus, with its rigid symmetry and theological weight, functions here as a probe—a litmus test for the model's grasp not just of semantics, but of form.

This approach is not without its limitations; the detected chiastic structures operate at the verse and half-verse levels. Many of the most significant theorised chiasmi in Biblical literature—such as the flood narrative in Genesis or the apocalyptic visions in Daniel—unfold over entire episodes or chapters. These longer-range structures often depend on thematic inversion, symbolic parallelism, or structural echoes, rather than lexical repetition. As such, they fall outside the reach of this current method, which relies on surface-level alignment within bounded windows of text.

This limitation reflects a broader challenge: scaling from local rhetorical structures to global narrative architectures requires models with longer memory, more abstract conceptual inference, and deeper integration of world knowledge. The detection of high-level chiasmi—those involving motifs, plot points, or theological reversals—remains an open and enticing problem, likely demanding hybrid approaches that blend neural representations with symbolic or discourse-based modelling.

Furthermore, I note that my work does not yet engage directly with prior efforts in rhetorical structure detection that focus on intra-sentential patterns (e.g., rhetorical role labelling or fine-grained discourse parsing). Such methods often operate at a much smaller granularity than the unit of analysis presented here and are geared toward prose rather than poetic structure. A rigorous comparison would require adaptation to a shared segmentation scheme—a task I leave to future work.

One especially promising direction would be the creation of a labeled corpus of Ancient Hebrew chiastic structures at multiple levels of granularity, especially in poetic books like the Psalms. These texts are dense with parallelism and refrain, making them fertile

ground for both automated detection and theological interpretation. Such a resource would enable systematic evaluation of model performance and open new possibilities for studying the function of rhetorical form across translations, traditions, and historical contexts.

3.8.1 From Internal Rhetoric to Intertextual Structure

While this chapter has focused on *internal* rhetorical structures—formal symmetries within a single passage—some of the most profound literary patterns in religious texts operate across books, traditions, and centuries. These are not merely repetitions, but reverberations: intertextual echoes, typological foreshadowings, and theological correspondences that connect disparate moments into a coherent interpretive arc.

These cross-textual structures often work through shared frames—conceptual scaffolds that recur across stories. For example, the language of John 1:1 ('In the beginning was the Word…') unmistakably invokes Genesis 1:1 ('In the beginning God created…') through salient lexical overlap and positional prominence. Recognising such echoes requires models to align texts based not on words, but on roles, functions, and thematic positions—a capacity that pushes beyond rhetorical pattern recognition toward conceptual generalisation.

The next chapter shifts our focus from rhetorical form to metatextual reference. I ask whether large language models can detect these frame-level correspondences across texts and translations—whether they can track the *idea* of creation, covenant, sacrifice, or exile as it reappears under new guises. In doing so, I move from detecting internal symmetry to tracing intertextual continuity, and from rhetorical structure towards narrative typology.

If embedding spaces encode such structure, then models may not only predict the next word or classify surface patterns—they may possess, however faintly, the architecture required for deep narrative understanding.

4

Maintaining Intertextuality in Translation

If the previous chapter explored internal symmetries—formal structures within a single passage—this chapter[1] turns outward, toward the metatextual phenomenon of *intertextuality*: the way texts speak to, echo, and revise one another across time. Intertextuality is not simply a literary flourish; it is the mechanism by which religious and classical texts construct meaning across generations, through typological allusion, shared narrative frames, and the reactivation of prior texts in new contexts.

Our central question here is whether large pretrained embedding models encode sensitivity to these intertextual structures—and crucially, whether this sensitivity persists in translation. Using a massively parallel Bible corpus, I ask whether embeddings maintain the semantic relationship between verses that are intertextually linked in the original language.

Intertextuality serves not just as an interpretive lens but as a thread of continuity: it binds cultural memory, reinforces theological motifs, and constructs narrative coherence across books and traditions. This makes it an especially revealing test case for evaluating whether multilingual embeddings can track not just meaning, but meaning-in-relation—to past texts, source traditions, and canonical structures.

[1]This work was originally published as a main conference short paper at NAACL 2025 in the Computational Social Sciences and Cultural Analytics track (McGovern et al., 2025a). My collaborators and co-authors on the paper (Drs Hale Sirin and Tom Lippincott) operated in an advisory role. I have modified the content and framing to fit the thesis and made other minor revisions.

4.1 Allusive Text Reuse in the Classical World

Coined by literary theorist Julia Kristeva, *intertextuality* describes the idea that every text is shaped by, and in dialogue with, other texts (Kristeva, 1986). This dialogue can be overt—as in quotation or paraphrase—or subtle, involving thematic or structural allusions. For instance, James Joyce's *Ulysses* (1922) retells Homer's *Odyssey* through a dense web of character mappings (Molly/Penelope, Stephen/Telemachus, Leopold/Odysseus), setting ancient narrative frames within a modernist Dublin (Currie, 2019).

Intertextual references often serve two purposes. First, they invite the reader to interpret a passage in light of another—projecting a shared frame or echo. But they can also introduce contrast: subverting or recontextualising the original. As Bamman and Crane (2008) note, this dual movement—similarity and divergence—makes intertextuality a powerful tool of reinterpretation.

Wilfred Owen's poem *Dulce et Decorum Est* provides a striking example[2]. In it, he recontextualises a line from the Latin poet Horace that had long served as patriotic shorthand for martial valour by juxtaposing it with appalling imagery of battle in the first world war:

> "If you could hear, at every jolt, the blood Come gargling from the froth-corrupted lungs,
> Obscene as cancer, bitter as the cud
> Of vile, incurable sores on innocent tongues,—
> My friend, you would not tell with such high zest
> To children ardent for some desperate glory,
> The old Lie: *Dulce et decorum est Pro patria mori.*"[3]

The classical maxim is here portrayed as grim irony in the face of modern atrocities. This technique—using intertextual reference to recontextualise prior texts—is very common in Biblical literature. Prophetic texts frequently depict God as a shepherd-king, guiding his people with strength and care. But when Israel's first king, Saul, is introduced in 1 Samuel 9, he is wandering aimlessly in search of lost donkeys—a figure so far removed from the royal ideal that the scene borders on satire. The intertextual echo of shepherd imagery thus becomes a foil, revealing Saul's inadequacy by contrast rather than similarity (Alter, 2011).

This is a form of typological inversion—common in theological texts, particularly where a New Testament scene may echo a Hebrew Bible narrative while simultaneously reconfiguring its meaning (Hays, 1989).

[2] https://www.poetryfoundation.org/poems/46560/dulce-et-decorum-est

[3] "It is sweet and fitting to die for one's country"

Exodus 14:21 *Hebrew Bible*	Revelation 16:12 *New Testament*
20 [...] and it was a cloud and darkness to them, but it gave light by night to these: so that the one came not near the other all the night.	11 And blasphemed the God of heaven because of their pains and their sores, and repented not of their deeds.
21 And Moses stretched out his hand over the sea; and the LORD caused the sea to go back by a strong east wind all that night, and made the sea dry land, and the waters were divided.	12 And the sixth angel poured his vial upon the great river Euphrates; and the water thereof was dried up, that the way of the kings of the east might be prepared.
22 And the children of Israel went into the midst of the sea upon the dry ground [...]	13 And I saw three unclean spirits like frogs come out of the mouth of the dragon [...]

Table 4.1: **Biblical intertextuality.** The highlighted middle verse shows the intertextual reference from the New Testament to the Hebrew Bible, establishing a connection between the drying up of the Euphrates River and Moses parting the Red Sea. Both are instances of divine intervention in the context of a body of water. However, intertextuality here not only establishes a semantic parallel between two events, but it also emphasizes the difference. The passage from Exodus is moment of the divine judgment that leads to safety, whereas the drying up of the Euphrates is a preparation for the final judgment of the world.

Biblical texts, in particular, are dense with intertextual structure. From direct quotations to thematic echoes and formulaic patterns, scholars have catalogued hundreds of cross-references between the Hebrew Bible and New Testament (Torrey and Canne, 1982). Table 4.1 illustrates one such allusion. The challenge in translation is not only to preserve the local meaning of each verse, but to retain these intertextual connections across languages—a task that requires models to recognise not just what a passage says, but where and how it echoes.

Recent computational studies have made significant progress in detecting intertextuality in classical texts, particularly within Latin and Koine Greek corpora. These works have focused primarily on monolingual reuse—tracing how authors quote, echo, or adapt earlier sources in the same language tradition.

In the following sections, I distinguish between text reuse, which refers to an instance of borrowing—whether verbatim, truncated, or modified—and allusion, a subset of text reuse involving implicit or indirect reference. Both literal and allusive text reuse contribute to intertextuality as a document-level phenomenon: a coherent network of such reuses and references forming a larger metatextual structure.

4.1.1 Literal Text Reuse

Much early work focused on fairly literal reuse. Lee (2007) compares nine scholarly hypotheses of textual reuse across the Synoptic Gospels, particularly between Mark and Luke, believed to share material from the hypothesised *Q* source. Using hand-crafted lexical features and cosine similarity scores on Koine Greek, the model learns to replicate individual scholars' theories of reuse, demonstrating that even subtle patterns of influence can be computationally modelled.

Manjavacas et al. (2019) similarly treat reuse as primarily literal, using n-gram and string-matching techniques to detect overlap in patristic texts. Drawing on the Biblindex corpus—writings of Clement of Alexandria (Greek) and Bernard of Clairvaux (Latin)—they show that such methods are highly effective when the reuse is direct.

Moritz et al. (2016) also work with Biblindex, but manually analyse reuse across Ancient Greek and Latin texts, categorising transformations such as synonym substitution, changes in case or capitalisation, and shifts in part-of-speech—still within a monolingual scope.

4.1.2 Allusion and Stylised Reference

Other work turns toward more allusive reference. Bamman and Crane (2008) examine Latin poetry, identifying allusion through TF-IDF-weighted lexical, syntactic, and positional features. Working across 14 texts by five Latin poets, they construct a typology of features that signal literary echo, though they do not evaluate against a gold-standard dataset. Burns et al. (2021) take a vector-based approach, training Word2Vec embeddings on CLTK-parsed Latin to detect the literary influence of Livy's work. They validate embeddings on synonym retrieval tasks and apply anomaly detection to identify stylistic deviation, using a support vector machine (SVM) to classify 'Livianness' across passages.

Konstantinidou et al. (2024) use character-level statistical language models trained on Homeric Greek to detect low-surprisal passages across the *Iliad* and *Odyssey*, reframing intertextuality as a function of probabilistic familiarity between texts.

4.1.3 The Missing Link: Translation

What these studies share, however, is a monolingual frame. While they successfully model intertextual structure *within* Greek or Latin, they do not address how that structure is preserved—or transformed—*across translation*. This omission is critical, particularly in religious literature, where multilingual intertextuality is not the exception but the norm.

The most direct treatment of this challenge comes from Bamman and Crane (2009), who investigate allusion between Milton's *Paradise Lost* (English) and Virgil's *Aeneid*

(Latin). They construct an evaluation set of known phrasal correspondences and propose a multi-step pipeline: aligning possible translation candidates, clustering them into phrasal spans, and applying similarity metrics between source and derived phrases. Their work demonstrates that cross-lingual allusion is tractable, though technically complex.

Yet even this work remains relatively isolated. The broader field has not systematically asked whether intertextual structures can persist through translation—despite strong historical motivation to do so. Ancient authors often operated in multilingual environments where translations, not original texts, shaped intertextual reuse. For instance, New Testament writers frequently quote the Septuagint—the Koine Greek translation of the Hebrew Bible—rather than the Hebrew source itself. These citations often reflect distinctive features of the Greek translation. For example, Hebrews 10:5 quotes Psalms 40:6 as saying, 'a body you prepared for me', which is a reading found in the Septuagint but not in the Masoretic Text, which reads 'my ears you have opened'. The author is clearly drawing on the Greek version, demonstrating that scriptural reuse can be mediated through a translation.

4.2 Intertextuality Across Translation

Intertextuality is a mode of interpretation as well as a literary device. As such, its preservation (or transformation) across translation plays a crucial role in how texts are understood and received in new linguistic and cultural contexts. Yet rhetorical devices such as intertextual reference are among the most difficult elements to preserve in translation. While semantic content may transfer with minimal distortion, intertextual resonance often depends on specific phrasing, echo, and lexical or syntactic alignment with source material. Identifying intertextuality and the degree to which it is preserved in translation is crucial for the interpretation and appreciation of literary and historical texts. Due to its significance, computational methods for identifying intertextuality have become an expanding field of research, and it is closely connected to other NLP tasks that are grouped under narrative reasoning and comprehension (Sang et al., 2022; Piper et al., 2021b).

This problem is especially pressing in literary and historical corpora. Frequently treated as a single book, the Bible is in fact a multilingual library of texts authored across more than a millennium, consisting of passages in Hebrew, Aramaic, and Koine Greek, with additional loanwords from Egyptian (Noonan, 2016), Babylonian, and Persian (Price, 1895). Many of its most significant rhetorical structures—particularly in the New Testament—are intertextual in nature, referencing earlier Hebrew texts through quotation, paraphrase, or allusion.

In this work, I propose a novel computational method for analysing how intertextuality is preserved across both human and machine translation. I use Language-Agnostic BERT

Sentence Embeddings (LaBSE) (Feng et al., 2022) —a model trained for semantic retrieval across 112 languages—to embed Biblical passages and identify intertextual links across translations. I evaluate LaBSE's capacity to recover known intertextual pairs in multiple languages and assess how translation affects the strength and clarity of these connections.

To this end, I:

- Present a new method for characterising intertextuality within and across translations using multilingual embedding spaces.
- Develop a corpus-level metric for evaluating the degree to which intertextual links are preserved.
- Compare human and machine translations across a range of languages and translation traditions.
- Provide a qualitative analysis of an illustrative case where human translations amplify or suppress intertextuality relative to a machine-generated baseline.

These results offer empirical support for long-standing hypotheses in Biblical scholarship—that human translators, intentionally or not, tend to emphasise continuity between the Old and New Testaments by making intertextual links more explicit than in the original source texts. By treating multilingual embeddings not merely as tools for semantic retrieval, but as instruments for measuring literary interpretation, I argue for a new methodological synthesis between NLP and historical-literary criticism.

4.3 Data

I use three primary sources for textual analysis of Biblical intertextuality:

1. The Translator's Amalgamated Hebrew Old Testament (TAHOT)
2. The Translator's Amalgamated Greek New Testament (TAGNT)
3. The Septuagint (LXX)

The TAHOT is based on the Leningrad Codex, the oldest complete extant manuscript of the Hebrew Bible. The TAGNT consolidates the Greek New Testament from multiple early editions. Both are curated by Biblical scholars at Tyndale House (Cambridge, UK) and distributed through the STEP Bible project[4]. I release cleaned and structured versions of these datasets on the Hugging Face Hub[5].

The Septuagint (LXX) is the earliest known Greek translation of the Hebrew Bible, produced by Jewish scribes several centuries before the composition of the New Testament.

[4]www.STEPBible.org

[5]TAHOT Huggingface site: 10.57967/hf/4174, TAGNT Huggingface site: 10.57967/hf/4184

Language	Family	Bitext pairs
English	West Germanic	$> 10M$
Finnish	Uralic	$> 1M$
Turkish	Turkic	$> 100K$
Swedish	North Germanic	$> 10K$
Marathi	Indo-Aryan	Small

Table 4.2: **Resource Level by Language Family.** In this study, I consider translation pairs including a handful of languages for which there exists a full human translation of both the Jewish and Christian texts in the JHUBC. These languages span different resource levels; the resource levels are indicated by the number of bitext sentences readily available in online datasets, as reported by Tang et al. (2020).

Its inclusion is crucial for studying intertextual links that cross not only canonical divisions but also linguistic boundaries.

As discussed in the previous chapter (see Chapter 3), differences in versification pose challenges when aligning Biblical texts. In this work, however, I standardise all source texts to follow the English versification system, to better reflect how modern translations typically order verses and to improve comparability across languages.

4.3.1 Human Translations

To analyse how intertextuality is preserved across modern translations, I use the Johns Hopkins University Bible Corpus (JHUBC) (McCarthy et al., 2020), which provides complete parallel translations for both testaments in a variety of languages. I select five languages to represent a range of language families and resource levels, as shown in Table 4.2.

However, I find that the JHUBC harbours extensive noise for translations of low-resource languages. I visualise the verse-level coverage of each translation in Fig. 4.1. Translations in low-resource languages (e.g., Marathi and Turkish) exhibit more missing verses and occasional misalignments, verified through manual inspection.

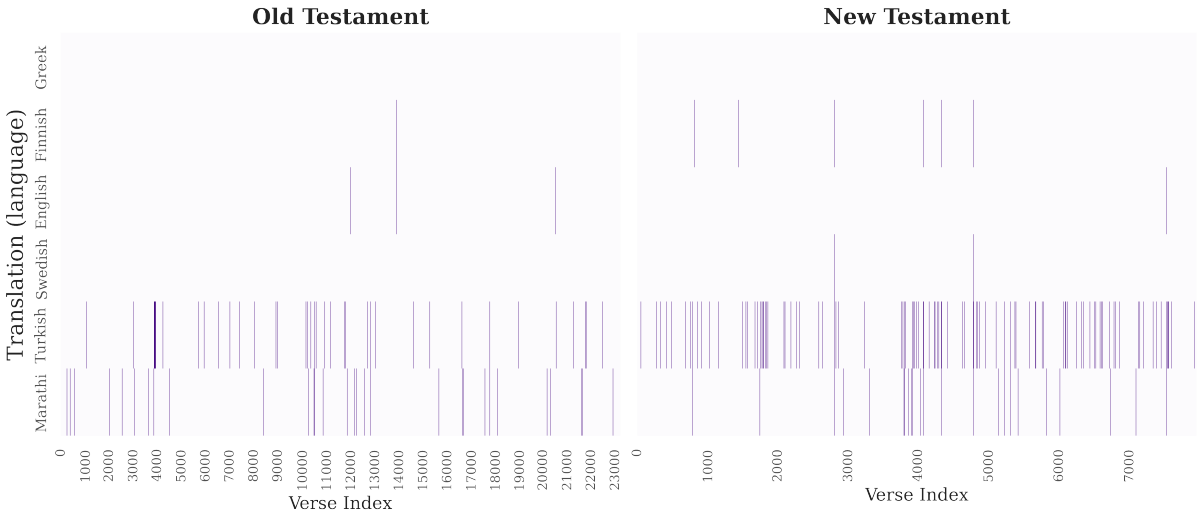


Figure 4.1: **Verse coverage in JHUBC Human Translations.** The x-axis indicates verse position across the Old Testament (left) and New Testament (right). Darkened lines indicate a missing verse in the associated human translation. Note that Turkish and Marathi translations are particularly noisy texts.

4.3.2 Independent Benchmark

To independently evaluate the generalisability of this approach, I also apply it to a benchmark corpus of Classical Latin intertextuality curated by Burns et al. (2021). This dataset contains 945 expert-annotated references connecting Book I of Valerius Flaccus' *Argonautica* to four earlier Roman works: Vergil's *Aeneid*, Ovid's *Metamorphoses*, Lucan's *Bellum Civile*, and Statius' *Thebaid*.

The references typically span two-word phrases, following conventions from the Tesserae Project (Coffee et al., 2013), which established methods for lexical and allusive matching in Latin literature.

4.3.3 Gold Standard for Intertextuality

To establish a ground truth set of intertextual connections, I use a curated dataset of Bible cross-references (Owens, 2010). The dataset originates from the *Treasury of Scripture Knowledge* (TSK) (Torrey and Canne, 1982), a historical compilation of interpretive cross-references created by Biblical scholars over several centuries. This initial collection was: deduplicated, merged for adjacent verse ranges, and subsequently extended via crowdsourced annotation, where members of the public could vote on the relevance of links.

I retain only verse-to-verse links that:

Connect passages from different books (to exclude local cohesion links)

Can be resolved across all manuscripts (to ensure alignment robustness)

Have a combined vote count $\geq$ 50, regardless of link direction.

That is, a reference is considered valid if

$$\texttt{Votes}_{A \to B} + \texttt{Votes}_{B \to A} \geq 50$$

To validate this threshold, I manually confirm that 96.0% of such references are attested in an online version of the *Treasury of Scripture Knowledge*, further affirming the reliability of this filtering method.

This results in a total of 2,183 cross-references, distributed as follows: 548 within the *Jewish* (Old) Testament, 961 within the *Christian* (New) Testament, and 674 *across* the two testaments.

Reference Type	Testament(s) Involved	Count
Within Jewish Testament	Old $\rightarrow$ Old	548
Within Christian Testament	New $\rightarrow$ New	961
Across Testaments	Old $\leftrightarrow$ New	674
Total	—	**2183**

Table 4.3: **Distribution of validated intertextual cross-references.** I distinguish between within-testament references and across-testament references, using a vote threshold of 50.

Throughout this work, I distinguish between *within*-testament references (Old $\leftrightarrow$ Old, New $\leftrightarrow$ New) and *across*-testament references (Old $\leftrightarrow$ New), which often reflect theological or typological relationships fundamental to Christian interpretation.

4.4 Pretrained Embeddings for Semantic Retrieval Involving Ancient Languages

To assess whether LaBSE is suitable for capturing cross-lingual semantic similarity for the languages of interest, I evaluate it out-of-the-box on a multilingual Bible verse retrieval task. This evaluation helps validate its utility for subsequent experiments on intertextuality preservation in translation.

I follow the standard sentence-level retrieval task as in Feng et al. (2022): for each verse in a source language, I embed all possible candidate verses in the target language using LaBSE, and identify the top-1 nearest neighbour based on cosine similarity. Precision@1 is then calculated as the proportion of verses for which the correct translation is ranked first among all candidates.

	English	Finnish	Swedish	Turkish	Marathi
Ancient Hebrew →	0.8526	0.7254	0.7171	0.6163	0.4691
Ancient Greek →	0.6116	0.5001	0.4623	0.3829	0.3531

Table 4.4: Precision@1 retrieval performance using LaBSE for cross-lingual semantic search on aligned Bible verse pairs. Each source verse in Ancient Hebrew or Ancient Greek is used to retrieve its matching translation in the target language (English, Finnish, Swedish, Turkish, Marathi), based on cosine similarity of LaBSE embeddings. The metric reports the proportion of source verses whose top-scoring match in the target language has the same verse identifier.

As Table 4.4 shows, retrieval performance is notably higher for Ancient Hebrew than for Ancient Greek, and highest for English targets across both source languages. While performance is not on par with LaBSE's best results on high-resource, modern language pairs, the results remain surprisingly strong given the domain specificity and antiquity of the source languages. This supports LaBSE's use in the downstream intertextuality experiments, with the caveat that performance drops significantly for low-resource languages such as Turkish and Marathi.

These findings also help calibrate expectations: when evaluating the preservation of intertextuality in human and machine translations, we might expect higher consistency in languages like English or Finnish compared to more linguistically distant or morphologically complex languages. Interestingly, Swedish yields stronger semantic retrieval performance than might be expected given its comparatively low resource level. This may be due to the presence of closely related languages (e.g. Danish, Norwegian) in LaBSE's multilingual training corpora, which might enhance transfer learning. By contrast, morphologically rich or orthographically distant languages like Turkish and Marathi may pose additional challenges for both tokenisation and translation. These patterns suggest that factors beyond raw resource size—such as linguistic similarity, shared vocabulary, or structural distance from the source language—also play a role in MT performance. The lower performance may also point to misalignments in the JHU Bible Corpus or inconsistencies in verse tokenisation, especially for the more fragmented or idiomatic translations.

4.5 Characterising Intertextuality between Corpora

To measure the degree of intertextuality between corpora, I use cosine similarity between multilingual sentence embeddings produced by LaBSE. This allows us to quantify semantic overlap between known intertextual passages and compare it to a baseline of non-intertextual pairs[6]. I apply this method to evaluate how faithfully different translations

[6]Note that I am making an assumption that while intertextuality is borne out in phrasing, echo, structure, and linguistic choice, semantic overlap is a sufficient proxy for preservation across translation.

preserve the intertextual structure of the source text.

Let $\mathcal{P}_{\text{intt}} = \{(v_i, v_j)\}$ denote the set of verse pairs annotated as intertextual references (i.e., the gold standard). For each verse v, I compute a sentence embedding using LaBSE. For all pairs $(v_i, v_j) \in \mathcal{P}_{\text{intt}}$, compute:

$$\text{cosine}(v_i, v_j) = \frac{\mathbf{e}_i \cdot \mathbf{e}_j}{\|\mathbf{e}_i\|\|\mathbf{e}_j\|}$$

Where $e_i = \text{LaBSE}(v_i)$ and $e_j = \text{LaBSE}(v_j)$. This yields a distribution of semantic similarity scores for the intertextual pairs, denoted D_{intt}.

To interpret this distribution, I compare it against a control distribution D_{random} constructed by replacing one verse in each gold pair with a randomly selected verse from the *same chapter*. This preserves general topical consistency while removing the intentional intertextual connection, offering a meaningful baseline:

$$\mathcal{P}_{\text{rand}} = \{(v_i, v_k) \mid v_k \neq v_j \wedge v_k \in \text{same chapter as } v_j\}$$

I define the intertextuality ratio (R_{intt}) as the ratio of the mean similarity for gold references to that of the random baseline:

$$R_{intt} = \frac{\mathbb{E}[D_{\text{intt}}]}{\mathbb{E}[D_{\text{random}}]}$$

A ratio $R_{intt} \gg 1$ indicates strong intertextual coherence, while a ratio $R \leq 1$ suggests that the supposedly intertextual verses are no more similar than might arise from chance.

To assess the stability of this ratio, I compute 95% confidence intervals using bootstrapping. Specifically, I resample the gold pairs in $\mathcal{P}_{\text{intt}}$ with replacement 10,000 times. For each resample, I recompute the intertextuality and random distributions, yielding a bootstrapped estimate of the ratio between the two means. This process results in an empirical distribution over ratio values, from which I extract the 2.5th and 97.5th percentiles to form the 95% confidence interval.

I compute intertextuality ratios across all corpora—original, human, and machine-translated—splitting the reference set into those that occur within a single testament (*within*) and those that span between testaments (*across*). This allows us to test whether translation influences the degree of semantic alignment, particularly in how the New Testament refers to the Old.

Throughout Christian theological history, the continuity between the Jewish and Christian scriptures has been strongly emphasised (van der Waal, 1980). It is plausible that human translators, aware of such theological framing, may unconsciously reinforce intertextual connections. By contrast, machine translation models lack such priors. Measuring the shift in intertextuality ratios between human and machine translations thus

offers a novel empirical test of whether theological intertextuality is a product of the source text or introduced in translation.

4.5.0.1 Benchmark Corpus

To validate this method, I apply the same intertextuality ratio to a well-studied corpus: Valerius Flaccus' *Argonautica I* and its known literary echoes from earlier Roman authors. I compute a ratio of 1.55 (95% CI: [1.53, 1.56]), confirming that this method can reliably detect corpus-level intertextual coherence in established literary datasets. The full results across translation settings and reference types are reported in Table 4.6.

4.6 Machine Translation Experiments

To assess how machine translation affects intertextual structure, I translate all source manuscripts into five typologically diverse target languages using Cohere's Aya23 multilingual model[7] (Aryabumi et al., 2024). Aya23 outperforms many multilingual models of similar or larger size on machine translation benchmarks, while remaining practical for academic use on limited compute resource. I employ a few-shot prompting setup: for each target language, I construct a static prompt containing four example verse translations drawn from the human-translated JHUBC[8]. The full prompt for GRC → English is:

[7] I use the 8B parameter version, accessed through the Huggingface Hub (Wolf et al., 2020).

[8] Ideally, these translations would be drawn from other parallel sources, but for most of the translation pairs (e.g. Ancient Hebrew → Marathi), the Biblical texts are the only parallel data available.

Ancient Greek Translations

1. Ancient Greek:
"Ο δε αποκριθεις ειπεν αυτοις· οψιας γενομενης λεγετε· ευδια, πυρραζει γαρ ο ουρανος."
English: “He answered and said unto them, When it is evening, ye say, It will be fair weather: for the sky is red.”

2. Ancient Greek:
"φανερωθεντος δε νυν δια τε γραφων προφητικων κατ᾽ επιταγην του αιωνιου θεου εις υπακοην πιστεως εις παντα τα εθνη γνωρισθεντος,”
English: “But now is made manifest, and by the scriptures of the prophets, according to the commandment of the everlasting God, made known to all nations for the obedience of faith:”

3. Ancient Greek:
"Ταυτα δε, αδελφοι, μετεσχηματισα εις εμαυτον και Απολλων δι᾽ υμας, ινα εν ημιν μαθητε το μη υπερ α γεγραπται φρονειν, ινα μη εις υπερ του ενος φυσιουσθε κατα του ετερου.”
English: “And these things, brethren, I have in a figure transferred to myself and to Apollos for your sakes; that ye might learn in us not to think of men above that which is written, that no one of you be puffed up for one against another.”

4. Ancient Greek:
"λεγω δε υμιν οτι του ιερου μειζον εστιν ωδε.”
English: “But I say unto you, That in this place is one greater than the temple.”

5. Ancient Greek:
“<INPUT_TEXT>”
English:

Preprocessing: To minimise orthographic noise, I also remove all *niqqud* (vowel pointings) and diacritics from the ancient source text. Empirically, I find that the inclusion of diacritics dramatically decreases LaBSE's efficacy on semantic retrieval from Ancient Greek and Hebrew.

Generation: At inference time, I use a maximum output length of 100 new tokens. I use the default BPE tokeniser with all of the default settings.

Post-Processing: I find that I need to post-process the outputs: I use regular expressions to isolate the text generated inside the first set of quotation marks after the prompt and exclude the rest. This is necessary to prevent nonsensical continued generations

Target	Source Manuscript Hebrew OT	Greek OT	Greek NT
English	**69.5**	**61.2**	**72.6**
Finnish	47.6	43.9	48.8
Turkish	*66.7*	*65.4*	*68.2*
Swedish	54.0	53.7	56.0
Marathi	27.6	26.5	29.8

Table 4.5: **COMET scores**. The top-scoring translation for each source manuscript is in bold text. The second top-scoring translation is in italics.

(often hallucinated urls to online Bible resources).

`N.B.` Although the Greek Septuagint is itself a translation of the Hebrew Bible, I treat it as a source text. This choice is historically justified: New Testament authors often quote the Septuagint directly rather than the Hebrew. Thus, the Septuagint acts as a de facto canonical text within the Christian tradition.

I evaluate two aspects of translation:

- **Translation Quality**: Measured using COMET (Rei et al., 2020), a neural evaluation metric which assess adequacy and fluency against human references. Within the range $0 \rightarrow 100$, a higher COMET score indicate greater adequacy and fluency.
- **Intertextuality Preservation**: This is my novel R_{intt} value, calculated using LaBSE-based semantic similarity between aligned verse pairs (intertexts), evaluating how well translation preserves allusions across texts. The range of a semantic similarity score is in $[0, 1]$, where a scores of 1 indicates perfect similarity and 0 indicates no similarity. R_{intt} is a ratio of the semantic similarity of intertexts compared to that of unrelated verses. A ratio greater than 1 indicates strong intertextual signal, while a ratio of less than 1 indicates that the intertextual signal is not distinguishable from noise.

4.6.1 Translation Quality

Table 4.5 reports COMET scores for machine translations from three Biblical source corpora—the Hebrew Old Testament (OT), Greek OT, and Greek New Testament (NT)—into five target languages. Across all source texts, English translations receive the highest overall COMET scores, with values of 69.5 (Hebrew OT), 61.2 (Greek OT), and 72.6 (Greek NT). These results are consistent with expectations: English is the primary working language of most large-scale pretrained translation models and benefits from an abundance of high-quality Biblical and general-domain parallel corpora. Additionally, many source editions of these texts are already translated into English with considerable care, contributing further to the model's ability to generate natural outputs.

Turkish translation consistently ranks second, with scores close to those of English: 66.7, 65.4, and 68.2 for the Hebrew OT, Greek OT, and Greek NT, respectively. These relatively strong results are somewhat surprising given that Turkish is not typically considered a high-resource language in the Biblical domain. However, Turkish benefits from multiple modern Bible translations and is often included in multilingual training corpora, which may contribute to its high performance. Notably, Turkish outperforms English for the Greek OT and achieves the second-highest score across all corpora, suggesting robust generalisation in the model's treatment of Turkish Biblical language.

Swedish performs in the mid-range, with scores between 53 and 56 across all sources. As a Germanic language like English, Swedish likely shares some structural and lexical similarity that aids the translation model. Moreover, Sweden has a long history of Bible translation, which may support the model's performance despite a smaller amount of available training data compared to English or Turkish.

Finnish follows a similar pattern, though with slightly lower scores ranging from 43.9 to 48.8. Finnish's greater typological distance from the source languages and from other Indo-European target languages likely makes accurate translation more difficult. Its complex morphology and agglutinative structure pose additional challenges for models trained primarily on fusional or analytic languages.

At the bottom of the performance range is Marathi, with scores of 27.6 (Hebrew OT), 26.5 (Greek OT), and 29.8 (Greek NT). Marathi is a low-resource Indic language with relatively limited availability of high-quality Biblical parallel data. Its low scores illustrate the substantial gap in translation quality for underrepresented languages, particularly when the source material is complex, ancient, and semantically dense. These results highlight the limitations of current multilingual translation models in reaching equitable performance across languages, especially in low-resource religious or historical domains.

A few trends emerge from comparing scores across source corpora. The Greek NT source text consistently produces the highest COMET scores for each target language. This could be due to the Greek NT's generally simpler literary register, as it is in large part comprised of short epistolary writings and narrative passages composed by non-native Greek speakers. The Greek OT produces lower scores than the Hebrew OT for all languages except Turkish, suggesting that the Septuagint may be more challenging for the model due to its hybrid translation style and textual variability.

In summary, these results demonstrate that translation quality is not only a function of the source and target language pair but also the characteristics of the source corpus. High-resource languages like English and moderately resourced ones like Turkish perform best, while low-resource languages like Marathi significantly lag behind. These disparities underline the importance of continued efforts to improve translation quality for underrepresented languages, particularly in domains of cultural and religious significance.

4.7 Intertextuality Preservation

		Within		Across
		Jewish (OT)	Christian (NT)	
Orig.	Ancient Hebrew	0.98± 0.14	–	–
	Ancient Greek	1.27± 0.21	1.30±0.19	1.31± 0.20
Human	English	1.66± 0.21	1.70± 0.30	1.69± 0.27
	Finnish	1.42± 0.41	1.36± 0.53	1.48± 0.22
	Turkish	1.50± 0.18	1.43± 0.29	1.51± 0.12
	Swedish	1.33± 0.15	1.39± 0.12	1.37± 0.09
	Marathi	1.35± 0.12	1.42± 0.12	1.44± 0.10
NMT	English	1.32± 0.20	1.50± 0.17	1.48± 0.20
	Finnish	1.24± 0.22	1.28± 0.18	1.26± 0.11
	Turkish	1.60± 0.15	1.71 ± 0.12	1.52± 0.32
	Swedish	1.31± 0.22	1.29± 0.31	1.36± 0.30
	Marathi	1.02 ± 0.10	1.22± 0.25	1.30± 0.18

Table 4.6: R_{intt} **for source manuscripts and their human translations.** Ratios within, and where possible between, testaments, for the Septuagint and TAGNT (Greek), TAHOT (Hebrew), and five human translations with 95% CI.

Table 4.6 presents intertextuality ratios (R_{intt}) for the original source manuscripts in Ancient Hebrew and Ancient Greek, as well as for their human and machine translations into five modern languages. These ratios, defined in section Section 4.5, quantify the degree to which semantic similarity between intertextually linked passages is preserved, with higher values indicating stronger intertextual cohesion. Importantly, we include the associated 95% confidence intervals, which help to allay the possibility of spurious analysis.

Starting with the original texts, the Ancient Hebrew corpus, representing the Jewish Old Testament, shows an intertextuality ratio of 0.98 ± 0.14. This value, close to 1.0, suggests a modest but detectable degree of thematic and linguistic reuse within the Hebrew Bible. In contrast, the Ancient Greek corpora show consistently higher ratios: 1.27 ± 0.21 within Jewish (OT) texts, 1.30 ± 0.19 within Christian (NT) texts, and 1.31 ± 0.20 across testaments. Although the Greek means exceed the Hebrew baseline, the confidence interval for Ancient Hebrew overlaps partially with those of the Greek measures, indicating that this difference should be interpreted as a strong trend rather than a strictly significant separation. Nevertheless, the consistent elevation across all Greek conditions suggests

systematically stronger intertextual cohesion in the Greek source tradition. This pattern plausibly reflects both translational harmonisation in the Septuagint and the deliberate reuse of Old Testament language and motifs within the New Testament.

Human translations show further amplification of R_{intt}, particularly in English, though the degree to which languages differ from one another varies once uncertainty is taken into account. English attains the highest mean ratios across all axes (1.66 ± 0.21 for the Jewish OT, 1.70 ± 0.30 for the Christian NT, and 1.69 ± 0.27 across testaments). While these values are clearly higher than those observed for the source languages, their confidence intervals overlap with those of Turkish and, to a lesser extent, Swedish. Consequently, English cannot be said to significantly outperform all other human translations, although it does form part of a high-performing group characterised by consistently elevated intertextuality ratios.

Turkish human translations likewise show relatively high R_{intt} values with moderate uncertainty, overlapping with English across conditions. In contrast, Finnish and Marathi exhibit lower means accompanied by wider confidence intervals. Marathi, in particular, occupies the lower end of the human translation range, and its confidence intervals show limited overlap with those of English and Turkish, indicating a more substantial difference in intertextual preservation. Swedish occupies an intermediate position, with ratios that overlap with both higher- and lower-performing languages, suggesting moderate preservation without a clear separation from either group.

Machine translation outputs reveal a similar but more compressed pattern. Turkish NMT achieves the highest mean ratios among machine-translated languages, and in some cases these values exceed those of the corresponding human Turkish translations. However, the associated confidence intervals overlap substantially, indicating that these apparent gains should be interpreted cautiously rather than as definitive improvements. English and Swedish NMT outputs yield ratios comparable to their human counterparts, with overlapping confidence intervals that suggest broadly similar levels of intertextual preservation. Finnish and Marathi NMT results are notably lower, and for Marathi in particular the confidence intervals show minimal overlap with those of higher-resource languages, indicating a meaningful degradation in intertextual coherence under machine translation.

Overall, once standard error is taken into account, the results indicate fewer sharp distinctions between high-resource languages than might be inferred from mea values alone. Clear separations emerge primarily between source languages and translations, and between high-resource and low-resource target languages. The Greek originals consistently trend higher than the Hebrew corpus, human translations tend to amplify intertextual cohesion relative to source texts, and machine translation preserves these patterns unevenly, with the most pronounced failures occurring in under-resourced languages. These

Original	Heb	ὅτι ἵλεως ἔσομαι ταῖς ἀδικίαις αὐτῶν, καὶ τῶν ἁμαρτιῶν αὐτῶν οὐ μὴ μνησθῶ ἔτι.
	Isa	ἐγώ εἰμι ἐγώ εἰμι ὁ ἐξαλείφων τὰς ἀνομίας σου καὶ οὐ μὴ μνησθήσομαι.
Human	Heb	For I will be merciful to their unrighteousness, and their sins and their iniquities will I remember no more.
	Isa	I, even I, am he that blotteth out thy transgressions for mine own sake, and will not remember thy sins.
NMT	Heb	For I will beware of their iniquity, and their sinner's iniquity; for I will not abhor them:
	Isa	I am the last of thy iniquitous acts, and I hate not myself.

Table 4.7: **Overemphasized Intertextuality by Human Translation.** The intertextuality from Hebrews 8:12 to Isaiah 43:25 is amplified by the human translator's decision to render different words as "sin". The machine translation abstains from this and restores the original distance, but loses coherence.

findings underscore the importance of interpreting intertextuality metrics through their associated uncertainty and highlight the need for targeted improvements in low-resource machine translation settings where intertextual fidelity remains fragile.

4.7.1 Key Findings

- Human translations tend to *amplify* intertextuality, producing similarity scores consistently higher than those of the original texts.
- Machine translations generally preserve the *relative* patterns of intertextuality across books and testaments, but this effect is entangled with subpar translation quality.
- The English human translation preserves intertextuality most strongly, while Marathi shows the weakest preservation, both in human and machine settings.

4.7.2 Qualitative Example: Overtranslation of Allusions

To better understand how human and machine translations differ in handling intertextuality, I compare one intertextual pair from Hebrews and Isaiah in Table 4.7, showing the original Greek, the human English translation, and the unconstrained machine translation for one such pair, between the Epistle to the Hebrews and the Book of Isaiah. The pair of verses has strong similarity in the original Greek (0.332), but this is nearly doubled by the human English translation (0.656). The highlighted Greek word, *hamartion*, typically translated as *sin*, occurs in Hebrews but not Isaiah, yet the latter's human translation

makes a point of using the term. Surface-level lexical decisions like this, and presumably many less direct choices, lead to uncalibrated translations that reinforce the received interpretation.

This example highlights a broader trend: human translators often harmonize phrasing to strengthen theological continuity, while machine translations retain more semantic distance between verses. This supports the use of machine translation as a tool for analysing historical patterns of intertextuality with fewer doctrinal or stylistic biases.

4.8 Limitations and Future Work

The current evaluation of intertextual translation operates under a deliberately constrained setup: I translate isolated verses with no preceding or surrounding context. This is a stark departure from how human translators work, who rely heavily on both immediate narrative context and broader intertextual resonances. While this setup allows us to control for direct influence and assess learned translation behaviour at the verse level, it underrepresents the full range of cues available in actual human translation.

In future work, I propose a retrieval-augmented generation setup that more faithfully models the translator's cognitive environment. Using multilingual semantic encoders such as LaBSE, I could retrieve the k most semantically similar verses from the source text, surface their human-translated equivalents in the target language, and feed these into the translation model. Unlike the current fixed, randomly sampled context (identical for all target languages), this dynamic, semantically motivated context would be specific to each verse and language pair. Such a setup would allow us to test the hypothesis that *intertextuality preservation is a learnable behaviour* for LLMs: that, when provided with the right contextual evidence, language models can learn to preserve higher-order narrative and theological linkages across languages.

Another limitation lies in the quality and consistency of parallel Bible corpora. Misalignments between source and target verses are unfortunately common in the JHUBC, and these misalignments introduce noise into both the training and evaluation metrics. Future work should apply robust multilingual alignment algorithms, such as those proposed by Bamman and Crane (2009) to achieve more consistent verse-level alignment across languages. This would increase the reliability of both contrastive and token-level evaluations of intertextual translation.

Finally, this study has focused on verse-level phenomena. Yet many of the intertextual connections I aim to evaluate—particularly typological or figural correspondences—manifest at the level of narrative episodes, not isolated verses. A natural extension of this work is to explore models' ability to handle *extended narrative context*, perhaps by comparing episode-level embeddings or analysing translation consistency across canonical

type-scenes.

4.9 Conclusion

This chapter explored whether pretrained multilingual models can preserve intertextual structure across translations, specifically in the domain of Biblical texts. I introduced a set of test cases where source verses exhibit strong allusive or typological links to earlier texts, and examined whether these links are preserved across multiple target languages. I also proposed a novel evaluation metric, R_{intt}, designed to capture deviations from a baseline level of intertextuality across translational shifts, and I used this metric to evaluate human and machine translations of Biblical texts. I found suggestive evidence that pretrained models sometimes—but not consistently—capture these structured relations. While the models demonstrate some sensitivity to shared theological or narrative frames (e.g., when translating Genesis 1:1 and John 1:1), their ability to preserve structured intertextuality is limited by context size, alignment quality, and the absence of explicit narrative modelling.

These findings point to a broader conclusion about the capacities and limits of pretrained language models. While they are remarkably effective at capturing surface-level semantic correspondences, their grasp of structured phenomena—such as intertextual allusion, typological foreshadowing, and narrative recurrence—remains incomplete. Structure, in this case, is not purely syntactic or distributional but conceptual: it requires modelling texts as networks of meaning and reference, often across vast spans of discourse.

The next chapter extends this investigation into structure by turning from intertextual translation to retrieval of type-scenes, a narrative-level phenomenon involving recurring story structures across texts. There, I shift from translation to retrieval, and from verses to episodes, but the core question remains the same: to what extent do pretrained models implicitly encode structured knowledge about narrative?

5

Exploring Narrative Type-Scenes in the Hebrew Bible

Our two previous chapters investigated tasks in which pretrained embeddings successfully captured local rhetorical structures, such as chiasmus, and meta-textual phenomena including intertextuality. These tasks align closely with language models' strengths in semantic representation. In contrast, this chapter[1] addresses a more structurally complex challenge: the identification of narrative *type-scenes*—recurring, prototypical story patterns that demand reasoning over events, participant roles, and character relationships. As recent work highlights, most pretrained language models struggle with structured data tasks in the absence of explicit structure-aware pretraining, limiting their ability to generate effective representations for retrieval and reasoning (Li et al., 2023a). This presents a significant challenge for type-scene identification, which inherently involves complex, hierarchical narrative knowledge.

This task, largely unexplored within natural language processing, shifts the emphasis from lexical semantics toward procedural and relational understanding. To ground our study, I focus on a canonical example: the betrothal type-scene in Biblical narrative, curating a evaluation dataset from existing scholarship, and comparing a number of

[1]The seed of this work was originally published as a long paper at the first workshop on Machine Learning for Ancient Languages (ML4AL) at ACL 2024 (McGovern et al., 2024) and subsequently presented as a dissemination paper at the Text2Story 2025 Workshop. My coauthors on the paper functioned in an advisory role. The contents of this chapter reflect a substantially modified version of the original work.

pretrained models for type-scene retrieval.

5.1 Type Scene: A Fixed Constellation of Motifs

The term *type-scene* originates in Homeric scholarship, where it denotes a 'repetitive compositional pattern' that recurs across Greek epics, marking key narrative waypoints in a protagonist's journey (Arend, 1933). Biblical scholars, notably Robert Alter, have adapted this concept to characterise the repeated use of analogous narrative patterns in Biblical texts, where similar story structures unfold either around different characters or the same character under varying circumstances (Alter, 1978).

I conceptualise type-scene extraction as a structured information retrieval task. Given a query passage or narrative scene, the goal is to retrieve thematically similar passages (candidates) from a corpus. For example, a query involving one rivalry between a barren but favoured wife would surface narrative passages surrounding Sarah and Hagar, Rachel and Bilhah, Leah and Zilpah (Alter, 1978).

Type-scenes align with a broader literary tradition concerned with recurrent narrative structures—fixed constellations of motifs that can manifest across varied characters and settings. This encompasses well-established concepts such as Propp's *dramatis personae* and their narrative functions in folktales (Taylor et al., 1959), as well as the notion of *figura* in Biblical hermeneutics, which captures similarities in event structures or accompanying circumstances across different stories (Auerbach, 1959; Sirin and Lippincott, 2024).

Alter (1983), Fuchs (1987), Britt (2002), and Press (2016) build on Alter (1978) to expand the catalogue of type-scenes present in the Hebrew Bible to those seen in Table 5.1.

Type-Scene Description	**Example**	**Scholarly Support**
1. A man meets a woman by a well	Genesis 24:11	Alter (1978), Fuchs (1987)
2. An angel announces a barren woman will bear a child	Judges 13:3	Alter (1983)
3. The younger son is preferred to the older son	Malachi 1:2–3	Greenspahn (1994)
4. God sends a prophet to speak to a rebellious people	Jonah 1:1–3	Long (1973)
5. A man has an epiphany in the field	Ezekiel 1:1	Alter (1978)
6. A prophet veils his face	Exodus 34:33	Britt (2002)
7. A well is found in the desert	Genesis 16:7	Alter (1978)
8. The words of a dying hero	Genesis 50:25	Alter (1978)
9. A hero is tempted	Judges 16:6	Alter (1978)

Table 5.1: **Common Biblical Type-Scenes and Their Scholarly Attestations.** This table summarizes frequently recurring narrative patterns ("type-scenes") in Biblical literature, along with examples and key scholarly references.

These constellations are not static repetitions; they undergo adaptive modifications in each instantiation, producing nuanced meanings that contribute to the interpretive richness of the text. This interplay between fixed pattern and variation will be explored through detailed examples in this chapter.

5.2 Related Work

Recent scholarship increasingly emphasises the necessity of integrating theoretical frameworks from the humanities with computational methodologies—such as information retrieval, summarisation, commonsense reasoning, and event detection—to deepen narrative understanding and unlock novel practical applications (Piper et al., 2021b). This convergence highlights that tasks demanding comprehension of complex correspondence networks, and their mapping onto well-defined theoretical constructs, remain beyond the capabilities of current pretrained language models. This is particularly true when applied to specialised domains outside their training distribution. Despite rapid advances, generalised reasoning abilities necessary for such tasks have yet to be fully realised.

5.2.1 Narrative Type Classification

The automatic classification of themes, motifs, and character functions within narrative texts presents a formidable challenge, often beset by scarce annotated corpora and the absence of coherent ontological frameworks. One prominent line of inquiry concerns the computational modelling of folktales through the lens of Propp's taxonomy. Pioneering efforts such as those by Finlayson (2015) and Yarlott and Finlayson (2016) assembled the *ProppLearner* corpus, featuring a rigorously annotated dataset of Russian folktales explicitly designed to train models for the classification of folktale narrative functions.

More recent investigations have explored few-shot learning paradigms leveraging LLMs to annotate stories with Proppian functions directly from textual synopses. For example, Gervás and Méndez (2024) examine the efficacy of prompting models like ChatGPT and Gemini for this task, noting their *apparent* linguistic fluency but identifying persistent issues with hallucinated or contextually inaccurate annotations. This exposes the challenges LLMs face in reliably grounding narrative function tagging in domain-specific knowledge.

Other efforts, such as Yarlott et al. (2022), apply off-the-shelf metaphor detectors to motif detection tasks on annotated narrative data, albeit with limited performance (F1 score around 0.35), further highlighting the complexity inherent in automated motif recognition.

Character identification and classification have also been explored through an extended ProppLearner corpus. For instance, Jahan and Finlayson (2019) analyse annotated folktales as well as a curated subset of 94 stories from a corpus of stereotypical activity 'scripts'

focused on the activity of 'taking a bath.' Employing features such as coreference chain length, semantic subject roles, named entity recognition, and WordNet-based personhood tagging, their support vector machine classifier achieved impressive F1 scores (0.81 and 0.99) on different datasets. These results indicate that with sufficiently fine-grained annotations, character classification tasks become tractable for conventional supervised learning methods.

Extending this foundation, Jahan et al. (2020) emphasise the importance of narratological grounding, particularly focusing on 'animate' characters critical to plot progression, and advocate for coreference-based feature engineering. In subsequent work, Jahan et al. (2021) employ unsupervised k-means clustering over coreference chains representing animate entities, successfully inducing six out of seven prototypical *dramatis personae* aligned with Proppian categories solely from structural plot information.

Collectively, these studies suggest that when high-quality, task-specific annotations are available, narrative classification can be effectively learned. Conversely, LLMs currently struggle with hallucination and produce superficially fluent yet often incorrect annotations in the absence of such data. In domains lacking grounded ontological frameworks or large annotated datasets, supervision signals must therefore be indirect or weakly supervised.

Bamman et al. (2013) curate a corpus of movie plot summaries and extract typed dependency relations (agent, patient, modifier) to cluster characters into latent personas, under the premise that characters sharing a persona exhibit common attributes and participate in similar actions. They augment this approach with crowdsourced annotations and validate clusters via indirect metrics—such as consistency of character clusters across movie sequels (e.g., Indiana Jones consistently classified as a 'hero' persona). Extending this work, Bamman et al. (2014) propose Bayesian mixed-effects models incorporating author metadata to better capture authorial influence on persona distributions, comparing several model variants to isolate the effects of persona and authorial style.

5.2.2 Information Retrieval

Information retrieval (IR) broadly aims to identify and rank relevant documents or passages in response to queries; dense retrieval approaches, which embed queries and documents into shared vector spaces, have demonstrated superior flexibility and recall compared to sparse keyword-based methods.

The retrieval task bears resemblance to aspect-based document similarity paradigms. For example, Ostendorff et al. (2020) leverage citation contexts within academic literature to infer thematic similarity between documents. The Bible's canonical chapter and verse divisions afford a structurally analogous framework, enabling us to treat verses or verse clusters as retrieval units annotated by narrative features.

Particularly relevant is the approach of Thai et al. (2022), who develop a model for re-

trieving primary textual evidence from literary sources in support of claims articulated in scholarly commentary. Their dual encoder architecture jointly embeds source quotations and scholarly passages which cite them. The encoder is then trained with a contrastive loss objective and in-batch negative sampling. This enables retrieval of textual evidence for paraphrased or out-of-domain queries—e.g., retrieving passages from *Pride and Prejudice* supporting the statement 'Elizabeth is frustrated with her mother.' Their work illustrates the promise of retrieval models capable of nuanced recognition beyond exact lexical matches.

5.2.3 Prior Work and Research Gap

My first version of this work, (McGovern et al., 2024), adapted the setup of Thai et al. (2022) to Biblical text retrieval. In it, I used a RoBERTa-based dual encoder with contrastive triplet learning to encourage Bible verses connected by a cross-reference (the same data used in Chapter 4) to exhibit closer co-location in latent space. I then used these embeddings to rank all possible verses by similarity to a type-scene query – a one-line type-scene description such as, 'An angel announces a barren woman will bear a child'[2]. This approach provided qualitatively decent results, but faced several limitations: cross-references indicate a connection between verses for many reasons, only some of which include thematic or typological similarity. They may also indicate lexical overlap, mention of the same historical event (as in the military campaigns dually recorded in the books of Kings and Chronicles), or a direct quotation of earlier scripture. So the informational signal provided the cross-references is not guaranteed to aid the type-scene identification task. Furthermore, my previous work posed type-scene retrieval as a verse-level task, when it is more naturally formulated as a scene-level task. I also lacked an evaluation corpus, which limited my analysis to qualitative exploration rather than metric-based evaluation.

This chapter aims to address these challenges in former work by developing a (small) evaluation set to measure retrieval of specific instances of a canonical type-scene—the betrothal by a well. This set includes manually chosen hard negatives, and, crucially, all examples are included as a passage/scene rather than as individual verses. I also remove the contrastive learning setup to instead directly compare different pretrained models (sentence encoders and structure-aware encoder-decoder models) to linguistic baselines.

[2]This analysis was done in the original text languages, so the queries were likewise in Ancient Greek and Ancient Hebrew.

5.3 Curating an Evaluation Corpus

I introduce a small but carefully curated evaluation corpus designed to test narrative pattern recognition in scriptural texts – specifically, seven instances of the betrothal type-scene (Table 5.2). Unlike synthetic or randomly sampled benchmarks, the dataset is grounded in long-standing scholarly work. These patterns encode cultural norms, character development, and theological meaning, making them especially rich targets for evaluating a model's capacity for analogical reasoning and narrative understanding.

Canonical	**Description**	**Support**
Genesis 24:10–61	Abraham's servant meets Rebekah by a well.	Alter (1978)
Gold Standard		
Genesis 29:1–20	Jacob meets Rachel by a well.	Alter (1978)
Exodus 2:15b–21	Moses meets Zipporah by a well.	Alter (1978)
Variants		
1 Sam 9:11–12	Saul meets women by a well while in search of his lost donkeys.	Alter (1978)
Ruth 2	Boaz meets Ruth in a field.	Alter (1978)
Judges 14	Samson seeks a Philistine woman without formalities.	Alter (1978)
John 4:4–42	Jesus meets a Samaritan woman by Jacob's well.	Arterbury (2010)

Table 5.2: **Well-Encounter Type Scenes in Biblical Narrative.** Passages are grouped into a canonical prototype, gold standard matches, and partial variants.

Crucially, scene boundaries in this dataset are not defined by fixed-length context windows but are determined by scholarly exegesis. This reflects the underlying narrative logic of the text rather than arbitrary token limits. Future work may extend this effort by developing automatic scene segmentation tools. However, I note that the task of identifying what constitutes a 'scene' or a meaningful narrative unit is itself a complex problem in both literary studies and NLP.

For the corpus, I use the English translation provided in the Johns Hopkins University Biblical Corpus (JHUBC) (McCarthy et al., 2020). While this exploratory work uses English text, I mark as future work the extension to the original languages (Biblical Hebrew and Koine Greek). As I explored in Chapter 4, translations may obfuscate literary and rhetorical features of the original text, and lexical links between type-scenes may be more readily apparent in the ancient sources. This extension would likely require advances in parsing tools for Ancient Hebrew in particular, including semantic role labelling (SRL) and dependency parsing.

The evaluation corpus distinguishes between the prototypical instances (attested in scholarship as archetypal), gold-standard analogues, and variants, which include subver-

sions or partial instantiations of the motif.

5.3.1 Canonical Type-Scene: The Encounter at the Well

I focus our prototype evaluation on one of the most widely attested type-scenes in Biblical narrative: the encounter between a man and a woman at a well. According to Robert Alter's authoritative work on Biblical type-scenes the 'well-meeting' motif functions both as a courtship narrative and a literary signal of divine or covenantal selection. Alter outlines a basic structure for the well-encounter scene (repeated from Chapter 2):

A future bridegroom travels to a foreign land.

He encounters a maiden by a well.

The maiden draws him water.

She brings home the news of the stranger's arrival.

The betrothal is arranged between families.

Later episodes (Jacob-Rachel, Moses-Zipporah) follow this pattern closely but with variations that Alter (1983) argues provide the audience a glimpse into the psychological profile of characters in an otherwise incredibly laconic text.

5.3.2 Gold Standard

Episodes widely accepted in the scholarly literature as clear instantiations of the well-encounter type-scene include:

Genesis 29:1–20: Jacob meets Rachel at a well and demonstrates strength by rolling away the stone covering the well's opening.

Exodus 2:15b–21: Moses aids Zipporah and her sisters, the daughters of Jethro, in drawing water, and shortly after will marry Zipporah.

These instantiations follow the procedural schema almost exactly and function as benchmarks in the corpus. Note that even in these clear type-scenes, there are slight variations (Zipporah is with her sisters, Jacob is more proactive in seeking a wife than Isaac, etc.).

5.3.3 Variants

Some episodes subvert or abbreviate the type-scene. These 'aborted' or allusive forms are crucial for evaluating nuanced recognition (Alter, 1983). Their absence is often as meaningful as their presence, shaping character perception in these highly laconic texts:

1 Samuel 9: Saul meets maidens near a well but fails to initiate a courtship or dialogue—his passivity foreshadows his failure as king: '…the deflection of the anticipated type-scene somehow isolates Saul, sounds a faintly ominous note that begins to prepare us for the story of the king who loses his kingship, who will not be a conduit for the future rulers of Israel, and who ends skewered on his own sword.' (Alter, 1978).

Ruth 2: Boaz meets Ruth in a field, not at a well, but the language of *naᶜărah* ('maiden') and the themes of provision and marriage evoke the type-scene indirectly. The application of the term 'maiden' to Ruth is particularly salient, as Ruth is a widow at the beginning of the story rather than a young virgin. There are other surprises as well: it is Ruth, not Boaz, who is the one travelling in a foreign land, and Naomi, Ruth's mother-in-law, who will (discreetly) arrange the match.

Judges 14: Samson seeks a Philistine wife without the hospitality and narrative buildup typical of the scene, signalling his outsider status and willingness to disobey the ritual and religious law of his people.

John 4: While type-scenes are less common in the New Testament, Arterbury (2010) suggests a clear link between a narrative episode in John 4 with the well-encounter type-scene. In this episode, Jesus, travelling through a foreign land (Samaria), meets a Samaritan woman at Jacob's well (referencing Genesis 29). Their exchange echoes the conventional type-scene but in deliberately subversive ways. First, instead of drawing water from the well, Jesus offers 'living water' that quenches all thirst. Second, he identifies the woman's checkered marital history, subverting the marital expectations of the schema. Finally, the woman runs to tell her community about the arrival of this foreign man. Whereas Old Testament well-encounters typically advance the covenantal lineage of Abraham (Alter, 1978), Jesus' encounter reconfigures the type-scene into a moment of revelation—one that affirms his identity as the awaited Jewish Messiah and ultimate fulfilment of Abraham's line.

5.3.4 Negative sampling

To evaluate retrieval or classification systems, I require negative examples: narrative episodes that do *not* instantiate the type-scene, even if they contain semantically similar elements. These were manually curated using keyword queries from StepBible[3], using their cross-linguistic corpus search tool. I searched for occurrences of:

- אִשָּׁה - (*ishah*, "woman" or "wife")
- עַלְמָה - (*ᶜalmāh*, "virgin")
- בְּאֵר - (*be'er*, "well")

[3]https://www.stepbible.org/

I intentionally excluded the search term אִישׁ (*ish*, 'man'), due to its ubiquity and therefore low saliency. I also exclude any verses that overlap with type-scene instances.

For each candidate verse, I extracted a 13-verse context window (6 verses before and after), matching the average length of gold standard scenes. While relatively naive, this manual negative sampling provides sufficient contrast to test retrieval models. I flag as future work the use of automatic scene boundary taggers and clustering methods to identify false positives or overlooked variants. In total, I provide a dataset of one canonical instance, two gold-standard positive instances, four noted variants, and 45 negative instances.

5.4 Type-Scene Identification as Information Retrieval

This study focuses on the task of identifying the most relevant candidate passage within a set of Biblical scenes given a query passage.

Formally, given a query q and a structured data document (candidate) c, I apply an embedding function f to generate vector representations of both the query and the candidate: $h_q = f(q)$ and $h_c = f(c)$. I then compute a similarity score between these representations, defined as

$$f(q, c) = sim(h_q , h_c)$$

where $\text{sim}(\cdot, \cdot)$ denotes the dot product function, measuring the relevance between the query q and candidate c.

I evaluate several embedding strategies: a baseline naive approach using bag-of-words (BoW) combined with TF-IDF weighting applied to structured input, alongside two Transformer-based models leveraging T5 architectures (Raffel et al., 2023) —a pretrained base model and a variant fine-tuned specifically for structured input and retrieval tasks.

All methods receive as input either the canonical query or a synthetic English template query designed to capture the narrative essence of the scene. While canonical queries might suffice for many type-scenes, the passage in Genesis 24 presents particular complexity. As Alter (1978) observes, this scene contains numerous details absent from subsequent retellings—such as the gifting of gold rings—that serve as potential distractors from the core features of the type-scene. `N.B.` This introduces a degree of circularity in query design. Future work aims to overcome this limitation by pursuing fully unsupervised discovery of salient steps in type-scenes through ground-up analysis.

The synthetic query is illustrated below:

```
1 """
2 A man traveled to a foreign land. There, he came upon a young
```

```
    woman at a well.
3 She had come to draw water, and he asked her for a drink.
4 After speaking with him, the girl ran home to tell her family
    about the stranger.
5 He was then invited to a meal. After the meal, arrangements were
     made for her to be given to him in marriage.
6 """
```

The candidate passages comprise Biblical texts, ranging from just a few verses to whole chapters. The objective is to retrieve both canonical and variant instances of the targeted type-scene.

5.4.1 Evaluation

I evaluate retrieval performance quantitatively using Mean Reciprocal Rank (MRR), complemented by qualitative case studies. The MRR is defined as

$$\text{MRR} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\text{rank}_i},$$

where rank_i denotes the position of the first relevant result for the i-th query. It measures the average rank of the first relevant result over all queries. I report the MRR of the two gold standard examples.

To complement MRR, I report report average rank of the gold standard, variants, and negative examples. The intuition is that an ideal model would rank the gold standard examples very highly (therefore a very high average rank) compared with the query, while the variants should be slightly less highly ranked. The unrelated passages should exhibit a lower average rank that that of the gold standard and variant passages.

5.4.2 Naive Embedding

I employ LaBSE (Language-agnostic BERT Sentence Embedding) (Feng et al., 2022), a model recognised for its robust semantic retrieval capabilities across languages. Utilising the `sentence-transformers` Python package (Reimers and Gurevych, 2019), I generate fixed-length embeddings of passages, leveraging LaBSE's capacity to capture nuanced semantic relationships beyond surface lexical overlap.

5.4.3 Bag-of-Words

The bag-of-words (BoW) approach leverages BookNLP[4], a comprehensive suite of literary text processing tools. BookNLP incorporates models trained on literary corpora produced in Bamman et al. (2019), Bamman et al. (2020), and Sims et al. (2019). Literary texts exhibit distinct lexical registers and syntactic structures compared to typical web text, complicating tasks such as coreference resolution. For instance, embedded narrative layers (e.g., nested speech in Dostoevsky's novels) pose challenges that BookNLP tackles more effectively than general purpose tools like SpaCy for this domain.

Specifically, BookNLP facilitates the extraction of verb predicates and character mentions. To accommodate the sparse, often laconic nature of Biblical text, I implement a hierarchical canonicalisation strategy: when named mentions are absent, I default to the most frequent common noun mention (e.g., 'a maiden'). If no explicit referent exists, I assign gendered placeholders such as 'unnamed man' or 'unnamed woman' based on inferred pronouns from coreference clusters. I utilise the 'large' pre-trained BookNLP model to maximise accuracy.

This pipeline transforms raw text into structured *scene descriptors*—lists of dependency relation tuples representing characters and events. Conceptually, these can be visualised as graphs, where nodes represent canonical characters and verb-lemma events (see Fig. 5.1). Directed edges encode semantic roles: from characters to events for agents, and from events to characters for patients.

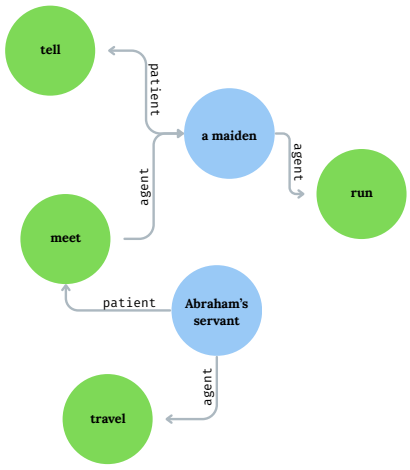


Figure 5.1: **Network diagram.** A type-scene description illustrated as a graph, where nodes represent characters (blue) and events (green), while directed edges indicate semantic roles in narrative event chains, e.g. 'Maiden' is the agent of the verb 'meet' and the patient of the verb 'tell'. Intuitively, instantiations of the same type-scene should yield a similar graph structure.

[4]https://github.com/booknlp/booknlp

To quantify similarity between scene descriptors, I employ a TF-IDF model, a classical sparse retrieval technique that represents text as high-dimensional, sparse vectors encoding exact token matches. While efficient and interpretable, this approach inherently lacks the ability to capture synonyms, paraphrases, or deeper semantic relationships, limiting its representational capacity in nuanced narrative domains.

5.4.4 Pretrained Models for Dense Retrieval

I explore pretrained Transformer models fine-tuned for text-to-text tasks as dense retrievers. The T5 model, originally trained to follow diverse textual instructions (e.g., summarisation prompts), serves as a foundational baseline. Variants such as CodeT5 (Wang et al., 2021), derived from RoBERTa and optimised for code-related tasks, and the enhanced CodeT5+ model (Wang et al., 2023), which offers architectural and training improvements, further expand the dense retrieval toolkit.

Complementing these, I incorporate SANTA (Structure-Aware Transformer pretraining) (Li et al., 2023a), a model explicitly designed to encode structural information within text. SANTA demonstrates moderate performance gains in dense retrieval tasks compared to vanilla T5 and CodeT5, highlighting the value of integrating structural inductive biases into pretrained models for narrative retrieval. Following Li et al. (2023a), I take the first token from the decoder as the dense representation of the input with which to compute cosine similarity for ranking and retrieval.

5.5 Results

Table 5.3 summarises the retrieval performance of different methods on the type-scene retrieval task. Across all models, synthetic queries substantially outperform canonical queries—this discrepancy likely reflects the canonical query's length and complexity, which introduces distracting details not shared by later variant passages.

Model	Num Params	MRR (↑)	Avg. Rank (Gold) (↓)	Avg. Rank (Variant) (↓)	Avg. Rank (Unrelated) (↓)
BoW+TfIdf	–	0.04	25.5	19.75	20.89
LaBSE	471M	**0.306**	**5.5**	6.75	23.51
Flan-T5-base	248M	0.062	17.0	15.75	21.83
Flan-T5-XL	2.85B	0.116	18.0	21.26	20.25
CodeT5	220M	*0.149*	*12.5*	15.25	22.14
CodeT5+	2B	0.047	21.5	17.75	21.34
SANTA	249M	0.054	20.5	21.49	17.0

Table 5.3: Performance comparison (AUC) across models for the intertextuality ranking task. Best AUC values are in **bold**, second-best in *italics*.

While LaBSE leads by a significant margin on the retrieval task, it is notable that even the largest Transformer model, Flan-T5-XL (2.85B parameters), underperforms compared to LaBSE on key metrics such as MRR (0.116 vs. 0.306) and average rank for gold passages

(18.0 vs. 5.5). This suggests that model size alone does not guarantee better retrieval of intertextual relationships in this domain. Flan-T5-XL's comparatively lower performance may be attributed to its more general pretraining objectives, which do not explicitly target semantic similarity or intertextuality, unlike LaBSE's sentence embedding focus.

CodeT5, specialised in code and structured data, achieves moderate performance (MRR = 0.149), surpassing Flan-T5-base and all other baselines except LaBSE, indicating that its training on structured inputs might aid in recognising certain narrative structures or patterns. However, the enhanced variant CodeT5+ with significantly more parameters (2B) surprisingly performs worse (MRR = 0.047), which could reflect overfitting or a mismatch between its training data and the literary narrative task.

The SANTA model, despite having a parameter count comparable to smaller Transformers (249M), exhibits relatively weak retrieval capability (MRR = 0.054), highlighting potential limitations in its architecture or pretraining for capturing complex intertextual connections in Biblical narrative.

Examining average rank metrics for variant passages further reveals that the models' abilities to generalise beyond exact matches vary considerably. LaBSE again ranks highest with 6.75, while others cluster around 15–21, underscoring its robustness in identifying paraphrased or semantically related but lexically divergent passages.

Most models rank unrelated passages, on average, lower than variant or gold passages, but not all. The BoW+TfIdf, CodeT5+, and SANTA are the exceptions, ranking unrelated passages higher on average than gold or variant passages. LaBSE ranks unrelated passages lowest (23.51), while SANTA gives them an average rank of 17.0, higher than its average rank for gold or variant.

Overall, these results emphasise the challenge of the type-scene retrieval task, where surface lexical matching is insufficient, and semantic, contextual, and structural nuances must be captured. LaBSE's strong showing validates the effectiveness of sentence embedding models trained specifically on multilingual semantic similarity for this task. At the same time, the relatively low absolute MRR values across the board point to ample room for improvement, motivating future research into architectures and training regimes better suited to narrative and intertextual understanding in ancient texts.

Additionally, an analysis of average cosine similarity scores provides further insight into model behaviour. The BoW+TfIdf baseline again performs near chance, with minimal differences in similarity between gold, variant, and unrelated passages, reinforcing that lexical overlap alone is insufficient for capturing type-scene relationships. For the strongest models, LaBSE and CodeT5, similarity scores for variant passages closely approximate those for gold passages, tentatively supporting the notion that annotated variants instantiate the same underlying type scene.

5.6 Discussion and Future Work

This study addresses a novel and challenging retrieval task that requires identifying passages based on complex *narrative structures*, rather than relying on simple lexical similarity or paraphrase detection. This intrinsic complexity substantially raises the difficulty of the task. Additionally, the dataset exhibits significant class imbalance, with only two gold passages and four partial (variant) passages among 45 false candidates. Such imbalance makes ranking metrics like MRR and average rank especially important for meaningful evaluation.

Although semantic embedding models like LaBSE demonstrate promising performance, especially when paired with synthetic queries, the overall retrieval effectiveness remains limited. Analysis of average cosine similarity scores indicates that top-performing models generate similarity distributions for variant passages closely approximating those for gold passages. This suggests these models capture semantic generalisations that align with annotated variants representing the same type scene. However, these preliminary findings require further validation through expanded annotation and normalisation of similarity measures.

Current approaches do not incorporate symbol frequency distributions or information content in narrative elements, concepts which could potentially improve retrieval by weighting rarer and more informative features more heavily Gombrich (1969). TF-IDF naturally incorporates this saliency weighting, but only through direct lexical overlap. Future work should explore integrating such probabilistic weighting schemes.

Looking ahead, advances in neurosymbolic methods—such as SCALLOP (Li et al., 2023b)—which combine algorithmic templates with probabilistic modelling, may reduce dependence on extensive human annotation by enabling structured, logic-based retrieval frameworks. Ultimately, fully data-driven discovery methods (e.g., Jahan et al. (2021)) for inducing type-scene structures without presupposed annotations offer the greatest potential for advancing the field. The creation of a larger, manually annotated corpus will be an essential step toward enabling such approaches.

Additional improvements may include training on substantially larger, domain-relevant datasets and incorporating structure-aware fine-tuning objectives, such as contrastive learning tailored to narrative retrieval. For example, fine-tuning with adapters on corpora like the Analogical Reasoning over Narratives (ARN) dataset (Sourati et al., 2024), which shares conceptual parallels to type scenes, might improve model generalisation, though domain transfer challenges would be likely.

The LaBSE model with synthetic queries achieves the strongest performance, achieving an MRR of 0.306 and an average rank of 5.5 for gold passages, outperforming all baselines by a wide margin. When including variant passages as correct matches, MRR rises further

to 0.356. This demonstrates LaBSE's superior ability to capture semantic and structural similarity inherent in type-scenes—remarkable given that it operates purely on semantic similarity rather than explicit structural encoding. Objectively, this MRR is still low in real-terms, but provides the first benchmark for automatic type-scene classification in Biblical narrative.

The BoW+TfIdf baseline exhibits very weak performance (MRR = 0.04, average rank 25.5) compared to the best model, LaBSE (MRR=0.31, average rank 5.5), reflecting its inability to discriminate between relevant and irrelevant passages based on surface lexical overlap alone. This underlines the inadequacy of sparse lexical representations for capturing complex narrative structures.

5.7 Conclusions

I present the first evaluation corpus for type-scene retrieval in Biblical texts, focused on the betrothal at the well type-scene, and use various pretrained embedding methods to benchmark the task, improving upon methodological weaknesses in prior work. My findings demonstrate that semantic embedding models such as LaBSE, when paired with carefully crafted synthetic queries, outperform purely lexical and typed-tuple features, although the task remains challenging with so few positive instances for retrieval.

This work lays the foundation for deeper investigations into narrative parallelism and rhetorical correspondence. In the subsequent chapter, I employ diagnostic probing techniques to analyse networks of correspondences between type scenes more broadly and systematically.

6

Probing for Narrative and Rhetorical Structure

This final chapter[1] addresses a central question that has thus far been approached only obliquely: to what extent do large language models (LLMs) encode and exploit narrative structure and rhetorical typology in their learned representations? Previous chapters have shown that off-the-shelf Transformer embeddings perform well on tasks that are largely semantic in nature—such as detecting intertextual allusions or approximating rhetorical symmetry in Biblical chiasmus. These tasks, while impressive, do not require deep structural understanding; they can often be solved through surface-level semantic pattern matching.

In this chapter, I adopt a more systematic approach to evaluating what narrative and rhetorical information is actually captured in pretrained embedding spaces. Drawing on the framework of *diagnostic probing* (Belinkov, 2022), I test whether decoder-only LLMs encode the kinds of abstract, discourse-level structures that typological reasoning requires.

[1] This chapter reflects completely novel work which has not yet been submitted for peer review. Similarly to other chapters, the work was developed in collaboration with Drs Hale Sirin and Tom Lippincott, where their roles were advisory in nature.

6.1 Introduction

As debates around the reasoning capabilities of LLMs intensify, narrative understanding emerges as a critical litmus test. Unlike sentence-level fluency or local coherence, narrative comprehension requires the integration of causal chains, temporal dependencies, and thematic abstractions across long contexts. This makes narrative an ideal diagnostic space for evaluating whether LLMs construct hierarchical, structured representations, or simply leverage surface-level statistical regularities.

Prior work has shown that pretrained models such as BART and T5 can encode latent information about entities, states, and relations without explicit supervision (Li et al., 2021). However, such results leave open the deeper question: do LLMs implicitly learn typological structures—the abstract narrative schemas that underpin coherent storytelling and rhetorical design?

I argue that narrative typology provides a theoretically rich and cognitively grounded framework for evaluating representational efficiency. Human brains, despite operating with far fewer parameters than state-of-the-art LLMs, appear to project experience onto reusable narrative types as one strategy for compression. If LLMs truly 'understand' texts, we should expect similar efficiencies to emerge: not rote memorisation, but the abstraction and reuse of functional narrative templates.

Yet, there are reasons for skepticism. LLMs are known to be vastly over-parameterised for text embeddings tasks (K et al., 2024), which may reduce the evolutionary pressure to learn efficient encodings. If typological abstractions are not encoded in the representation space, this would support the view that LLMs, while deep in architecture, remain shallow in representation and comprehension.

To test this, I apply diagnostic probing to state-of-the-art decoder-only LLMs from the LLaMA family. I focus on two challenging literary tasks:

1. **Narrative parallelism**, where structurally distinct stories must be aligned by underlying functions or schema.

2. **Rhetorical symmetry**, where the presence of lexical and semantic parallel structures must be inferred.

Diagnostic probing allows us to isolate the extent to which specific layers encode information relevant to these abstract tasks. Previous research has found that layers in encoder-only Transformers recapitulate classical NLP pipelines (Tenney et al., 2019a): early layers emphasise lexical knowledge, middle layers handle syntactic and semantic patterns, and later layers often capture discourse-level information. Similar patterns have been observed in decoder-only architectures on tasks like BliMP (Warstadt et al., 2020), where grammatical acceptability is modelled via differences in predicted perplexity.

6.1.1 Contributions

This chapter makes the following contributions:

1. I benchmark decoder-only LLMs on challenging literary tasks, highlighting their limitations in recognising rhetorical and narrative parallelism—a task that demands deep structural alignment.

2. I provide an interpretability analysis using diagnostic probes, identifying which layers encode narrative and rhetorical information, and to what extent.

3. I offer a framework for evaluating narrative typology as a signal of representational efficiency, proposing that successful narrative inference implies abstraction and generalisation, not just memorisation.

6.1.2 Broader Motivation

Discourse-level NLP has historically lagged behind sentence-level modelling, due in part to the computational demands of modelling long contexts. However, recent advances in Transformer architectures and context length have reawakened interest in long-form reasoning. This opens up the possibility of evaluating narrative analogies not as paraphrastic or lexical similarity tasks, but as tests of schema alignment and conceptual mapping.

Recognising that two texts instantiate the same narrative function—despite divergent surface forms—is a higher-order cognitive skill. It plays a central role in typological exegesis, literary analysis, and even theological interpretation. As such, it provides an essential benchmark for assessing whether LLMs have learned anything akin to narrative understanding.

6.2 Related Work

I situate this work at the intersection of interpretability methods for large language models (LLMs), cognitive science, and the study of analogical reasoning in narrative understanding. I draw methodological inspiration from diagnostic probing, particularly edge probing as introduced by Tenney et al. (2019b), but pursue it within a distinct cognitive and task-specific context: analogical narrative recall. This setting departs from typical NLP benchmarks by introducing richer and more structurally complex stimuli, motivated by findings in human memory research.

6.2.1 Memory and Narrative Recall:

> "...[T]he brain actively reshapes our understanding of past events in light of new incoming information"
>
> Zadbood et al. (2022)

There is increasing evidence from cognitive neuroscience that human memory operates not as a sequence of discrete snapshots, but through structured narratives; the brain organises past experiences as networks of semantically or causally connected events, and this structure critically shapes memory performance.(Lee and Chen, 2022; Zadbood et al., 2022). Lee and Chen (2022) show that in realistic audiovisual narratives, central events—those with stronger connections to other events—are more reliably remembered. These central events elicit stronger hippocampal boundary responses during encoding and more consistent activation patterns during recall, especially across individuals. This suggests that memory formation and retrieval are modulated by event centrality within a narrative graph.

Further work shows that memory is not static but dynamically updated: neural representations of prior events are reshaped as new information becomes available (e.g., a narrative twist). This was demonstrated in fMRI studies where participants reinterpreted earlier scenes from *The Sixth Sense* after learning the ending (Zadbood et al., 2022). These updates were localised in the default mode network, which has been shown to underlie abstract situation model construction.

Narratives are ideal ecological tools for studying memory: they involve temporally unfolding, semantically rich events that mirror how humans encode and communicate experience. Crucially, storytelling itself can alter the content and perceived value of memories, modulating which details are retained and how they are framed for others (Eckardt et al., 2024; Lee et al., 2020). These findings offer a pseudo-cognitive foundation for studying model behaviour in narrative tasks: if human memory is shaped by narrative structure, might I find evidence of similar organisation (or lack thereof) in large language models?

6.2.2 Diagnostic Probing

My method of analysing model internals builds on diagnostic probing, particularly the *edge probing* framework introduced by Tenney et al. (2019b). This approach decomposes various linguistic tasks—ranging from POS tagging to semantic role labelling—into a set of graph edges that can be independently predicted from a model's hidden representations. They find that pretrained models like BERT capture strong syntactic representations and

progressively deeper semantic features across layers, culminating in high-level semantic phenomena like coreference and role assignment.

In follow-up work, Tenney et al. (2019a) showed that different linguistic tasks are best represented at different layers in BERT: syntactic information emerges early and is sharply localised, while semantic features are more diffuse and appear at higher layers. This 'layerwise progression' suggested a developmental-like hierarchy of representation in pretrained models.

More recent work, however, has complicated this picture. He et al. (2024) use probing on the BLiMP dataset and find that grammatical features are distributed throughout GPT-2's layers and vary with sentence complexity. Morphological and interface-level semantic features remain especially difficult to isolate. Moreover, probing results can differ markedly depending on architecture and task complexity.

Critically, several papers challenge the foundational assumptions of probing. Niu et al. (2022) introduce the GridLoc framework, showing that previously reported layer effects may reflect artefacts of position and training dynamics rather than principled encoding. Belinkov (2022) similarly warns that information discovered by a probe is not necessarily used by the model—i.e., it may be decodable from hidden states without being functional at inference time.

This motivates a core aspect of my analysis: I compare a model's probed representations to its *prompted* performance, showing that for complex narrative analogies, what is achievable with prompting alone is not always decodable from internal states. In this way, my work contributes to ongoing discussions about the interpretability and functional validity of probing, especially for abstract tasks.

Finally, probing remains under-explored in literary or narrative contexts. Existing probing benchmarks overwhelmingly focus on linguistic phenomena at the sentence level, while narrative understanding involves event structure, temporal coherence, and causal reasoning across long-form inputs. Moreover, recent work (Li et al., 2025) demonstrates that probing results vary significantly across languages, especially in multilingual models—underscoring the limits of generalisability from standard English NLP tasks.

6.2.3 Analogical Reasoning for Narrative Tasks

Analogical reasoning—understanding that X is like Y—is central to human cognition and plays a particularly salient role in how we interpret and connect narratives. Yet the precise mechanisms and representational structures underlying narrative analogy remain underexplored, especially in computational settings.

Piper et al. (2021b) approaches this question through the lenses of *narrative economies* (how one narrative influences the understanding of others), *narrative beliefs* (the persuasive efficacy of narratives in public discourse), and *narrative responses* (the ways narratives

may inform reasoning in the human mind). While this framework captures how narratives resonate across contexts, it stops short of offering an explicit taxonomy of cross-narrative similarity.

More targeted computational work has attempted to model analogical narrative similarity. Elson (2012) introduced a *story intention graph* approach that uses propositional generalisation over discourse relations. For instance, the propositions 'A Lion watched a fat Bull' and 'A Fox observed a Crow' are abstracted to a shared form like 'An animal perceives another animal.' Their system relies on dependency graphs and logic-based pattern-matching (via Prolog), incorporating both hypernym generalisation and temporal sequencing to detect analogies. This method contrasts bottom-up statistical matching with top-down structural isomorphisms, offering one of the earliest computational treatments of analogical narrative structure.

A key insight here is that narrative analogy often involves category-level parallelism: mapping characters, goals, and events by type or function rather than surface similarity. This notion underpins more recent empirical evaluations of LLMs' analogical capacity. Sourati et al. (2024) introduce a triplet-based benchmark dataset, Analogical Reasoning over Narratives (ARN), that tests whether models can distinguish deep analogies from superficial resemblance. Their dataset, drawn from the ePiC corpus, includes human-annotated dimensions such as character roles, relationships, settings, goals, actions, and overarching 'proverb' meanings (e.g., 'the early bird gets the worm'). While LLMs perform well when surface and structural cues align, they struggle in cases where an analogy is not apparent on the surface but requires latent information—suggesting limitations in abstraction and generalisation.

Outside narrative domains, analogical reasoning has also been probed through prompting techniques. Wicke et al. (2024) show that Chain-of-Thought (CoT) explanations for spatial analogies yield modest improvements in alignment with human judgments. More compellingly, Yasunaga et al. (2024) demonstrate that *analogical prompting*—asking models to first generate a relevant exemplar before solving a target problem—outperforms both zero- and few-shot CoT baselines, particularly in larger models like GPT-4 and PaLM-2. This capacity to invent structurally relevant but novel examples suggests emerging analogical abilities, though the role of training data memorisation remains an open question.

Taken together, these findings suggest that LLMs *can* sometimes exhibit analogical reasoning, especially under structured prompting regimes and at larger scales. However, two central challenges remain. First, it is unclear whether their apparent analogical inferences reflect genuine conceptual abstraction or sophisticated pattern-matching. Second, the absence of explicit structural priors—such as event schemas, narrative roles, or discourse-level relations—may constrain both the generalisability and interpretability of these analogies. More systematic modelling of narrative structures is likely required to

bridge this gap.

6.3 Task Description and Data

I investigate analogical and rhetorical reasoning through two parallelism classification tasks, each cast as a binary decision problem over span- or document-level inputs. Both tasks follow a general formalism in which textual units—either spans or documents—are encoded as fixed-length vectors and compared to determine relational similarity.

Let a sentence or document be represented as a list of contextual embeddings $E = [\mathbf{e}_0, \mathbf{e}_1, \ldots, \mathbf{e}_n]$, obtained from a pretrained language model. A *span* is defined as a half-open interval $s = [i, j)$ denoting the sublist $[\mathbf{e}_i, \mathbf{e}_{i+1}, \ldots, \mathbf{e}_{j-1}]$. Following Tenney et al. (2019b), I compute fixed-length span representations by applying a learned projection to each $\mathbf{e}_k$ in the span, followed by self-attention pooling over the span window.

Formally, each classification instance consists of either:

1. A pair of spans (s_1, s_2) within a single discourse context, or
2. A pair of documents (d_1, d_2) represented by sentence-level embeddings or pooled document-level features.

The classifier then predicts a binary label $y \in 0, 1$ indicating whether the input pair exhibits a structural or semantic parallelism. I apply this general framework to two specific tasks: rhetorical parallelism (span-level) and narrative parallelism (document-level), each accompanied by task-specific data preprocessing and sampling strategies.

6.3.1 Rhetorical Parallelism (Span-Level)

I introduce a rhetorical parallelism classification task derived from the Augustinian Sermon Parallelism (ASP) dataset introduced by Bothwell et al. (2023). This dataset consists of 80 Latin sermons by Augustine of Hippo, annotated by hand for rhetorical structure. Each annotation captures sets of parallel spans (or 'branches') that collectively instantiate a rhetorical pattern—either synchystic (parallel order) or chiastic (inverted order). One particularly complex annotation is visualised in Table 6.1 (the same illustrative example highlighted in their paper).

To construct binary classification instances, I extract all span pairs within a given sermon. Positive pairs consist of spans that are annotated as belonging to the same parallelism structure. Negative pairs are sampled from spans, still within the same sermon, but drawn from distinct parallelism annotations. I restrict negative sampling to within-document pairs to ensure the model learns intra-textual rhetorical similarity rather than document-level distributional cues.

Latin	English
panis esuriret	so that **bread might hunger**,
satietas sitiret	**satiety might thirst**,
uirtus infirmaretur	**strength might be weakened**,
sanitas uulneraretur	**health might be wounded**,
uita moreretur	**life might die**.

Table 6.1: **Rhetorical Parallelism in Latin:** The repeated syntactic and semantic inversion emphasizes paradoxical transformation.

Formally, each instance consists of two spans $s_1 = [i_1, j_1)$ and $s_2 = [i_2, j_2)$, and a label y indicating whether they belong to the same rhetorical parallelism.

Each of these clauses forms a span in the annotation, with shared syntactic and stylistic structure. While parallelisms in the annotation scheme are not limited in the number of branches they may have, I only construct binary pairs, taking every permutation of the spans within a parallelism as independent examples.

6.3.2 Narrative Parallelism (Document-Level)

To probe analogical reasoning at the document level, I construct a narrative parallelism classification task using the Analogical Reasoning over Narratives (ARN) dataset introduced by Sourati et al. (2024). The narratives in the story were curated from the ePiC stories dataset (Ghosh and Srivastava, 2022), which contains 2500 narratives written by humans crowd-sourced on Amazon Mechanical Turk to illustrate a given English proverb (e.g., 'Hindsight is 20/20', 'Slow and steady wins the race'). Sourati et al. (2024) then use an LLM to extract characters, relations, actions, goals, and locations (which they call 'surface mappings'), as well as the labelled proverb (a 'system mapping'), in each narrative to create triplets of: an anchor story, an analogous story, and a distractor story. Analogous stories share a system mapping, i.e. proverb, to an anchor story, regardless of how similar their surface mappings are. An analogous story which has high overlap in surface mappings (e.g. similar locations, character goals, etc.) to an anchor is a 'near' analogy. On the other hand, distractor stories *may* share surface mappings to the anchor story, but do not share the system mapping (proverb). A 'near' distractor story shares many surface features with the anchor story but has a different overall message, while a 'far' distractor is completely different in surface features as well as proverb. The ARN dataset comprises 1096 such triplets.

I convert this into a binary classification task over story pairs (d_1, d_2), where the label indicates whether both narratives instantiate the same proverb. Anecdotally, I find that the language fluency and grammatical quality of the narratives varies considerably, which may introduce unwanted noise into the analogical detection task. To ensure fluency, I use

a model from Krishna et al. (2020) which was trained on the BliMP dataset to produce 'grammatical acceptability' scores; I retain only narratives in the ARN dataset with a grammatical acceptability score ≥ 0.9 . This filters out ill-formed generations, e.g.:

Example with Annotated Errors and Corrections

"I like my manager a lot as he is the only one in our management who always ~~take~~ takes the initiatives on anything he ~~purposes~~ proposes for the team, and we always follow him. Recently, we were under pressure to deliver many projects in time, so ~~he recommended us~~ he recommended that we spend extra time on the projects, and he himself started spending extra time out of hours. When our ~~tea~~ team saw that he was himself spending extra time, they also followed him."

Note: Strikeouts indicate original text; red text shows corrections or insertions. An example from the ARN dataset.

The resulting dataset is pared down from 1315 unique narratives to 872 unique fluent narratives. Rather than relying on the preconstructed triplet pairs, I construct all possible permutations of narratives to upsample the dataset. I create positive pairs by matching anchor–analogy pairs from the same proverb, and negative pairs by matching anchor–distractor or analogy–distractor pairs drawn from different proverbs. Each pair is labeled as analogous (1) or non-analogous (0).

Two illustrative examples which are designed to probe whether models align on abstract narrative structure rather than surface lexical cues:

Illustrative Examples from Extremes of the Analogy Dataset

Category	Example Content
Easiest Analogy (Close Analogy, Low Distractor Similarity)	**Proverb:** *That which does not kill us makes us stronger* **Anchor:** When I remember the challenges I went through when I was starting my business, I break into tears. But I do not regret a thing. I think that the most precious gold goes through the hottest furnace. There are great and unforgettable lessons that I learned during that period that I will always cherish. It made me better. **Analogy:** Once upon a time, in a small village, there lived a talented young presenter named Lily. She was faced with countless challenges, from technical difficulties and stage fright to harsh criticism and rejection. However, with each obstacle she overcame, Lily grew stronger, honing her skills, building resilience, and gaining the respect and admiration of her audience. Ultimately, her unwavering determination and ability to thrive in the face of adversity transformed her into a renowned and influential figure, inspiring others to embrace their own inner strength. **Distractor:** I stained my cedar house this last summer in advance thinking about the consequences of not doing so. If I don't, the sun fades and cracks the wood, and could let drafts in.
Hardest Analogy (Far Analogy, High Distractor Similarity)	**Proverb:** *Hindsight is always twenty-twenty* **Anchor:** Mark was the new CEO of the company. Under adrenaline rush he decided to go after a small startup that he thought would be profitable for the company. However, months later, it was discovered that the startup would not benefit them much but instead it was costing them a fortune to make a bid for the startup. **Analogy:** Looking at the past decisions, I put too much oil in the fryer and burned the turkey last year, and the year previous to that and also a couple of months ago. That's also true that at those moments I didn't know what I'm doing. **Distractor:** After becoming the new CEO of the company John decided to change the microchip in the laptop being produced by his company. However he understood that they need to design an entirely new laptop instead of just changing the chip as the new chip won't be compatible with the old hardware setup.

6.3.3 Data Preprocessing

For both tasks, I partition the data into training, validation, and test sets using an 80/10/10 split. To prevent data leakage and preserve the coherence of the underlying narrative or rhetorical units, I apply stratified sampling at the level of documents—sermons for rhetorical parallelism and individual proverbs for narrative parallelism—rather than at the level of individual examples.

There are far more negative examples in the dataset than positive examples due to our sampling method, so to avoid such a severe class imbalance, I downsample the negative examples to match the number of positive examples, using a fixed random seed (`seed=42`) for reproducibility.

Task	# Unique Docs	# Train Examples	# Dev Examples	# Test Examples
Narrative	872 narratives	3744	270	446
Rhetorical	80 sermons	5604	616	564

Table 6.2: **Dataset statistics by task type.** The table shows the number of unique source documents and the train/dev/test split for each task.

Together, these two tasks probe neural models' capacity to detect structural and conceptual similarity at different granularities—within-discourse rhetorical patterns and across-discourse analogical narratives—while carefully controlling for superficial textual overlap.

6.3.4 Limitations

The validity of probing as an interpretive tool remains contested. Ravichander et al. (2021) highlight that the accuracy of a probe does not necessarily indicate whether a model *uses* the probed information in downstream reasoning. Their case study in Natural Language Inference shows that models can encode linguistic features even when they are not functionally required for the task at hand.

Relatedly, Sharkey et al. (2025) argue that much probing work lacks the rigour of *causal* interpretation, since most methods rely on correlational evidence. Without interventions that manipulate model internals or inputs, it's difficult to establish whether a given feature genuinely contributes to model behaviour.

Another ongoing debate concerns the complexity of probes themselves. Early work (e.g., Hewitt and Liang (2019)) cautioned against complex probes, warning they might 'do the work' of the task rather than expose information already present in the representations. They advocate for *selectivity*—a good probe should perform well on the linguistic task and poorly on a control task. However, Pimentel et al. (2020) challenge this simplicity bias, proposing an information-theoretic view in which mutual information between representation R and target variable T should be maximised. From this perspective, more

complex probes yield tighter bounds on the information present in the model, even if they blur the line between decoding and representation.

Finally, my own experiments are exploratory and currently limited in scope. I do not implement control tasks to evaluate selectivity, nor do I apply causal interventions to the embedded representations. These remain important directions for future work.

6.4 Edge Probing for Parallelism Classification

To evaluate whether large language models (LLMs) encode information about rhetorical and narrative parallelism, I adapt the *edge probing* framework introduced by Tenney et al. (2019b), with modifications to support these tasks and accommodate decoder-only models such as those in the LLaMA3 family.

Edge probing is a method for analysing what kinds of linguistic information are encoded in pretrained language model representations at specific token positions. It treats various structured prediction tasks—such as part-of-speech tagging, syntactic parsing, semantic role labelling, and coreference—as sets of edges between spans of text within a sentence. A simple classifier is trained on top of the fixed contextual embeddings to predict these edges independently, allowing researchers to isolate and evaluate the extent to which models capture both local and long-range syntactic and semantic relationships. This approach provides a fine-grained lens into the types of linguistic and structural knowledge encoded at different layers and positions in the model.

The experimental pipeline is illustrated in Fig. 6.1. The process involves (1) passing documents through an LLM, (2) extracting contextual embeddings for spans or documents of interest, (3) computing fixed-length representations of annotated spans using a pooling method (I compare two different methods, described below) and concatenation, and (4) feeding these representations into lightweight probing classifiers. Below, I describe each stage in detail.

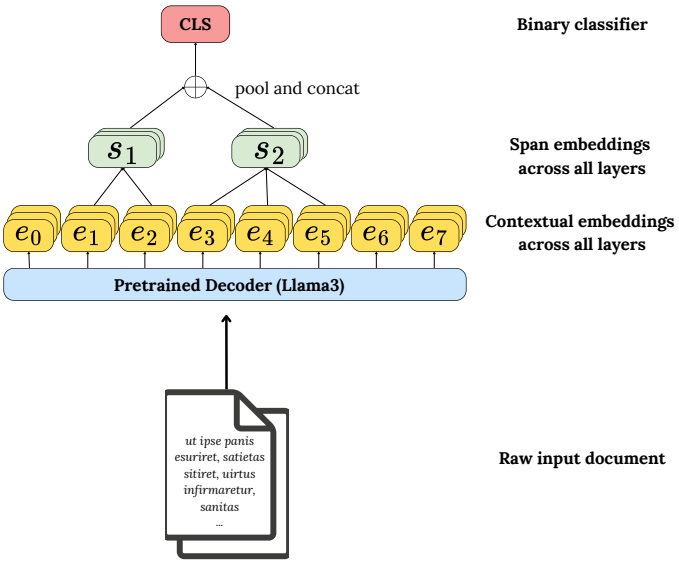


Figure 6.1: **Experimental setup for parallelism probing tasks.** The rhetorical task is shown: a full sermon text is encoded by the model, which produces contextualised embeddings across all layers of the model. I then extract relevant span pairs and use either learned span pooling or zero-paramter pooling to reduce the embeddings to fixed size vectors, which are then concatenated and fed into a binary classifier.

6.4.1 Extracting Contextual Representations

I use decoder-only LLaMA3 models of various sizes (1.1B, 3.8B, and 8B parameters)[2], including both base and instruction-tuned variants, as feature-extractors. For each input document (e.g., a sermon or a pair of narratives), I use a model in a feedforward, non-generative manner to extract contextual token representations from all Transformer layers. I record the hidden states at every layer of the model, capturing the contextual embeddings $\mathbf{h}_i^{(\ell)} \in \mathbb{R}^d$ at each token position i for each layer $\ell \in 0, 1, \ldots, L$.

To manage memory usage on limited university compute resources, I implement a custom chunking strategy: inputs are split into overlapping chunks of 1024 tokens with a stride of 512, ensuring that no span annotation is split across chunk boundaries. This facilitates extraction of complete span representations without truncation while maintaining efficient memory usage during inference.

[2]Models were accessed through the Huggingface Hub (Wolf et al., 2020).

6.4.2 Option 1: Span Representation and Pooling

In this experimental setup, I then extract the annotated spans of interest—pairs of spans labelled as rhetorically or narratively parallel or not. Each span is a sequence of token-level embeddings from the hidden states of each layer, $\mathbf{h}_{i_1}, \mathbf{h}_{i_1+1}, \ldots, \mathbf{h}_{j_1-1}$ or $\mathbf{h}_{i_2}, \mathbf{h}_{i_2+1}, \ldots, \mathbf{h}_{j_2-1}$.

6.4.2.1 Trainable Projection and Attentive Pooling

Following Tenney et al. (2019b), I compute a fixed-length span representation via a trainable projection layer and a *self-attentive pooling* mechanism:

1. **Projection**: Each token embedding is first projected to a 256-dimensional space:

$$\tilde{\mathbf{h}}_i = \mathbf{W}_{\text{proj}} \cdot \mathbf{h}_i + \mathbf{b}_{\text{proj}}, \quad \tilde{\mathbf{h}}_i \in \mathbb{R}^{256}$$

2. **Attentive Pooling**: Attention weights α_i are computed over the span tokens, and the weighted sum produces the span embedding:

$$\alpha_i = \frac{\exp(\mathbf{v}^\top \tanh(\mathbf{W}_{\text{attn}} \tilde{\mathbf{h}}_i))}{\sum_j \exp(\mathbf{v}^\top \tanh(\mathbf{W}_{\text{attn}} \tilde{\mathbf{h}}_j))}, \quad \mathbf{s} = \sum_i \alpha_i \tilde{\mathbf{h}}_i$$

This mechanism is applied separately to each of the two spans. The resulting span embeddings from the two branches are concatenated to form the input to a probing classifier.

6.4.2.2 Scalar Mixing Over Layers

Since contextual information may be distributed unevenly across model layers, I apply a *scalar mix* (Peters et al., 2018) over the layer-wise span representations. For each span, I learn a softmax-normalised set of weights s_ℓ and a scaling parameter γ, computing:

$$\mathbf{h}_i^{\text{mix}} = \gamma \sum_{\ell=0}^{L} s_\ell \cdot \mathbf{h}_i^{(\ell)}$$

This allows the probe to learn which layers are most informative for parallelism detection.

6.4.3 Option 2: Zero-Parameter Span Pooling

In addition to the learned self-attentive pooling mechanism, I introduce a novel **zero-parameter pooling method** that leverages the model's own attention weights, without learning any new parameters. This method is based on the hypothesis that if a model has

already learned to attend to rhetorical or narrative parallelism during pretraining, then its internal attention distributions may already encode useful signals for this task.

I experiment with two variants of this approach:

1. **Joint encoding**: I encode a single input document that contains both spans s_1 and s_2. I then extract the attention weights from the final token of s_2, and use these weights to pool the token embeddings across the entire span pair. The intuition is that if the model recognises rhetorical parallelism between s_1 and s_2, this may be reflected in the attention patterns at the end of the second span.

2. **Separate encoding**: As a contrastive baseline, I encode s_1 and s_2 independently. For each span, I extract attention weights from its final token and use them to pool token embeddings *within that span only*. This setup removes any explicit modelling of inter-span relationships and tests whether the model assigns similar attention distributions to parallel spans even when processed in isolation. If so, this would suggest that the model has learned invariant representations of features relevant to rhetorical structure.

In both setups, I compute a weighted average of token embeddings within each span using attention distributions from the final token. Let $S = \{i, \ldots, j\}$ denote a span, and let T be its final token. For each layer ℓ and head h, I extract the attention weight $A_{T,i}^{(\ell,h)}$ from token T to token i, and compute:

$$\mathbf{s}^{(\ell,h)} = \sum_{i \in S} A_{T,i}^{(\ell,h)} \cdot \mathbf{h}_i^{(\ell)}$$

This yields a representation tensor of shape (L, H, d).

6.4.3.1 Scalar Mixing over Heads and Layers

To condense the attention-weighted embeddings into a fixed span representation, I apply a double scalar mix:

- A scalar mix over layers (as in Section 6.4.2.2)
- A scalar mix over attention heads

Let s_ℓ and s_h be softmax-normalised weights over layers and heads, respectively. The final span embedding is:

$$\mathbf{s} = \sum_{\ell=0}^{L} \sum_{h=0}^{H} s_\ell s_h \cdot \mathbf{s}^{(\ell,h)}$$

These span embeddings are concatenated and passed to a classifier.

6.4.4 Probing Classifiers

I evaluate multiple lightweight probing models to avoid overfitting and preserve the interpretability of the results:

1. **Logistic Regression**: A linear classifier operating on the concatenated span representations.

2. **Two-Layer MLP**: A feed-forward neural network with a hidden dimension of 256, ReLU activations, and a final sigmoid output layer.

Each model is evaluated using F1 score and Area under the receiver operating characteristic curve (AUC). F1 score is a harmonic mean of the precision (how many true positives were predicted out of all predicted positives) and recall (how many true positives there were out of all predictions):

$$F_1 = \frac{2}{\text{recall}^{-1} + \text{precision}^{-1}}$$

AUC measures the area under the plot of the curve of the recall (also called the sensitivity) and the proportion of true negatives correctly identified by the model (also known as the specificity) at different threshold values – this is the receiver operating characteristic curve. Crucially, AUC provides an estimation of how well a model can distinguish between binary classes without relying on a hardcoded threshold, but rather, considers all possible thresholds.

I train the probes using the Adam optimiser Kingma and Ba (2014) with batch size of 32 and initial learning rate of $1e^{-4}$. I use L2 norm gradient clipping, and I implement early stopping if there has been no AUC improvement on the validation set observed after 5 epochs.

6.4.5 Prompting

To compare the representational knowledge captured by probing models against what LLMs use at inference time, I implement a direct prompting baseline. This allows us to evaluate whether decoder-only language models can identify rhetorical or narrative parallelism *without* additional classifiers or gradient-based learning—i.e., by relying solely on their forward generation behaviour given an appropriately crafted instruction. This experiment is designed to address the central question: do LLMs have more latent information encoded than what they use at inference time?

I frame both rhetorical and narrative parallelism detection as binary classification tasks via prompting. Specifically, the model is asked to generate either 'Yes' or 'No' as a one-token answer to a natural language question.

For each labeled example, I insert the relevant context and spans into a templated instruction prompt. The spans are always marked explicitly as `[SPAN1]...[/SPAN1]` and `[SPAN2]...[/SPAN2]` in the rhetorical case, or as `[DOC1]...[/DOC1]` and `[DOC2]...[/DOC2]` for narrative examples. The goal is to elicit a judgment of structural parallelism using only the language model's internal representations and next-token prediction.

For ASP (Augustinian Sermon Parallelism) data, I include a local context window surrounding the marked spans. Specifically, I include a window of `left_window = 10` words before the start of `SPAN1` and `right_window = 10` words after the end of `SPAN2`. This context is tokenised and formatted such that both spans are visible in a continuous excerpt, simulating the original sermon text.

```
1 "The following is an excerpt from a Latin sermon. Two spans in
     the sermon are marked as [SPAN1]...[/SPAN1] and
     [SPAN2]...[/SPAN2]. Determine whether the two spans are
     rhetorically parallel.
2 Parallel structures should have some linguistic features in
     common, which could include features of phonology, morphology
     , syntax, semantics, or even style. The features could be the
      same (e.g., the repeated sounds in 'ueni, uidi, uici') or
     diametrically opposed (e.g., the opposite meanings in 'a time
      to be born' and 'a time to die'). These features typically
     occur in the same order or the opposite order.
3
4 <marked_text>
5
6 Are the two marked spans rhetorically parallel?
7 Answer with a single word: Yes or No.
8 Answer: "
```

The information in the prompt about parallel structures is taken verbatim from the task description of Bothwell et al. (2023).

In the second prompting condition, I assess models on a task of analogical narrative parallelism. Here, I present two short narrative passages, each marked with its own delimiter (`[DOC1]` and `[DOC2]`). The prompt instructs the model to judge whether the *relational structure* between entities and events is analogically parallel, rather than relying on surface similarity.

```
1 "Below are two short narratives. Determine whether they are
     analogically parallel by evaluating not just surface
     similarities (such as shared characters, actions, or
     locations), but whether the relations between elements in one
```

```
      narrative correspond systematically to relations in the
      other. Focus on whether the two narratives share the same
      underlying structure of events or roles, even if the surface
      features differ.\n\n"

"[DOC1]"
f"{text1}"
"[/DOC1]"

"[DOC2]"
f"{text2}"
"[/DOC2]"

"Are these two narratives analogically parallel based on
   systematic relational correspondences?"

"Answer with one of: 'Yes' or 'No'.
Answer: "
```

The instructions to the model are based on the task description of Sourati et al. (2024).

6.4.6 Decoding and Output Processing

I use a simple and deterministic decoding configuration, as my interest is not in creative generation but in extracting the model's most probable classification decision. I use greedy decoding, a generation temperature of zero, and restrict the number of new tokens generated to be one. Following best practices, I extract the log-probabilities for the next-token prediction over the tokens ' Yes' and ' No'[3] rather than parsing raw text output.

I apply a softmax function over these two logits and treat the probability of 'Yes' as the output score. If this exceeds a fixed threshold of 0.5, I predict the model's answer to be 'Yes'. I additionally save raw text responses for manual inspection.

6.4.7 Model Selection

I run all prompting experiments on the same model family used for probing—namely, the LLaMA3 series. I evaluate three parameter scales: 1B, 3B, and 8B, using their corresponding tokenisers. All models are accessed via the Hugging Face Hub. One key axis of comparison is the effect of instruction tuning (Zhang et al., 2024), a process in which pretrained language models are further finetuned on collections of input–output

[3]In practice, the model always returns a token prepended by a leading space, so I take care to use the tokens ' Yes' and ' No', rather than 'Yes' and 'No'.

Task	Model	Variant	AUC (Log Reg) ↑		AUC (MLP) ↑	
			llm-attention	learned	llm-attention	learned
Narrative	Llama-3.2-1B	Base	**0.585**	0.527	0.563	0.548
		Instruct	**0.555**	0.517	0.547	0.528
	Llama-3.2-3B	Base	**0.595**	0.558	0.578	0.541
		Instruct	**0.587**	0.577	0.532	0.569
	Llama-3.1-8B	Base	0.561	0.534	**0.581**	0.503
		Instruct	0.507	0.547	0.555	**0.559**
Rhetorical	Llama-3.2-1B	Base	0.949	0.950	0.953	**0.954**
		Instruct	0.929	0.934	0.934	**0.936**
	Llama-3.2-3B	Base	**0.943**	0.928	0.937	0.925
		Instruct	**0.948**	0.929	0.945	0.936
	Llama-3.1-8B	Base	0.911	0.819	**0.923**	0.829
		Instruct	0.902	0.845	**0.931**	0.840

Table 6.3: **AUC scores for narrative and rhetorical parallelism detection using linear (logreg) and nonlinear (MLP) probes.** Each row shows performance across two pooling strategies ('llm-attention' vs. 'learned') and model variants (base vs. instruct) for three LLaMA-3 model sizes. Probes achieve consistently higher AUCs on rhetorical tasks than narrative ones, with best results for rhetorical parallelism exceeding 0.95. Narrative parallelism remains significantly harder to detect, even with instruction-tuned models and nonlinear classifiers. Bold text indicates the highest score for a model across classifier configurations.

pairs framed as natural language instructions, explicitly optimising them for task-following behaviour.

6.5 Results

I use area under the ROC curve (AUC) as the primary evaluation metric because it is robust to class imbalance and provides a threshold-independent measure of the classifier's ability to distinguish between positive and negative examples. This is crucial in this setting, where binary classification (is a pair parallel or not?) is evaluated across multiple probe configurations and model scales, each of which may produce highly skewed prediction distributions.

Across all experiments, I observe stark differences in task difficulty and probe effectiveness between the narrative and rhetorical parallelism detection tasks (see Table 6.3 and Fig. 6.2). Rhetorical parallelism is reliably learnable by probes, with AUCs exceeding 0.95 in the best setup. Narrative parallelism, by contrast, proves far more elusive to linear and nonlinear probes alike. Even the best configuration achieves a modest AUC of 0.595, suggesting that these relationships are either more weakly represented in the model's embedding space or encoded in a format poorly matched to linear separation.

Overall, MLP classifiers outperform logistic regression in the majority of settings,

particularly on the rhetorical task. This suggests that rhetorical relationships, while often linearly separable, can still benefit from additional nonlinearity.

On the rhetorical task, all best-performing configurations use an MLP classifier.

- LLaMA3-1B (Base) reaches 0.954 AUC with an MLP and learned pooling.
- LLaMA3-8B (Instruct) achieves 0.931 AUC with an MLP and LLM-attention pooling.

On the narrative task, gains from MLP are more modest and inconsistent.

- LLaMA3-3B (Base) reaches **0.578** AUC with an MLP and LLM-attention pooling, slightly ahead of its logistic regression counterpart.
- Some setups (e.g. LLaMA3-8B Instruct) show only marginal improvement from MLP over logreg.

Despite these gains, MLPs do not fundamentally change the task difficulty: narrative AUCs remain low regardless of classifier type.

To test whether larger LLMs show signs of emergent structure relevant to parallelism detection, I compare model scales (1B, 3B, 8B) across base and instruction-tuned variants, holding the probe architecture fixed.

Fig. 6.2 visualises AUC scores from this comparison. The bar chart plots results from a consistent probing setup: spans are encoded jointly using the LLM, pooled via a self-attentive mechanism learned by the probe, and classified using a 2-layer MLP. Each group of bars corresponds to one model scale (1B, 3B, 8B), with colour indicating whether the model is a base or instruction-tuned variant. This setup mirrors that of Tenney et al. (2019a) and provides a strong baseline for analysing scaling effects under joint encoding.

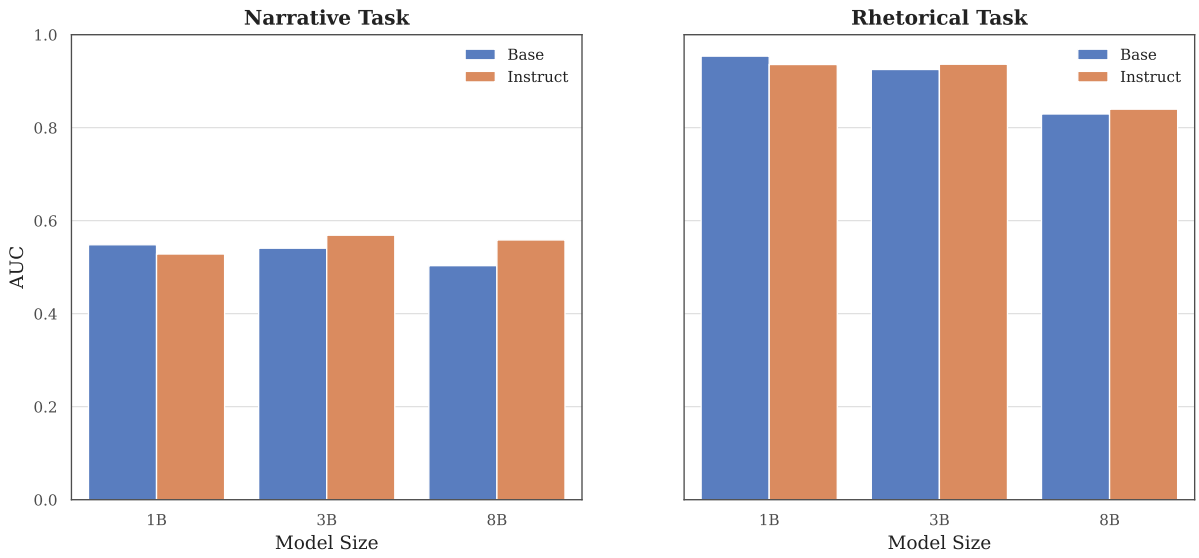


Figure 6.2: **Narrative vs. Rhetorical Task Performance in Base and instruction-tuned Models.** Bar chart of AUC across model sizes using the classifier setup of Tenney et al. (2019a): spans encoded in one pass together, using a self-attentive pooling learned alongside a two-layer MLP classifier.

Effect of Model Size: Surprisingly, larger models do not consistently outperform smaller ones. In several cases, the 1B model performs as well as or better than its 3B and 8B counterparts:

- On the rhetorical task, LLaMA3-1B Base with an MLP and learned pooling achieves the highest AUC of 0.954, outperforming both the 3B (0.937) and 8B (0.929) models under the same configuration.
- On the narrative task, LLaMA3-3B Base yields the best result overall (**0.595**) with logistic regression and LLM-attention pooling—but even here, the 1B variant is competitive.

Effect of Instruction-Tuning: The impact of instruction tuning is subtle and not reliably beneficial for probing performance. In some settings, instruction-tuned models outperform their base counterparts:

- On the narrative task, LLaMA3-8B Instruct slightly outperforms Base with learned pooling and an MLP (0.559 vs. 0.503).
- On the rhetorical task, LLaMA3-8B Instruct also outperforms Base with an MLP and LLM-attention pooling (0.931 vs. 0.923).

However, in other settings, instruction tuning either has no effect or performs worse than the base model. For example: LLaMA3-3B Base achieves 0.578 on narrative with MLP,

while Instruct drops to 0.532 under the same setup. On rhetorical with learned pooling and MLP, LLaMA3-3B Base slightly outperforms Instruct (0.925 vs. 0.936).

The lack of consistent gains implies that while instruction tuning might improve usability or task generalisation in generative settings, it does not reliably enhance the representational accessibility of relational features such as rhetorical or narrative structure—at least not in a way that makes them more amenable to linear probing.

6.5.1 Encoding and Pooling Methods

To understand how relational signals are distributed in pretrained representations, I tested two encoding and two pooling strategies. A heatmap visualising AUC for each combination of pooling and encoding method is in Fig. 6.3.

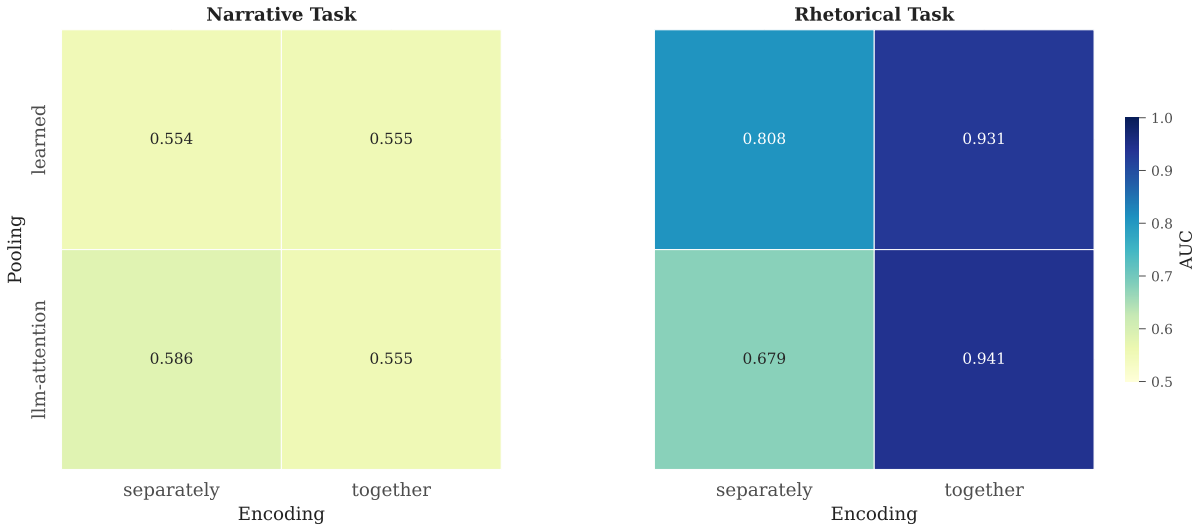


Figure 6.3: **The effect of pooling and encoding combinations on task performance.** Heatmaps illustrate the resulting classifier performance (AUC) on the two tasks for every combination of learned span pooling vs. zero-parameter pooling and separate encoding vs. joint encoding. The importance of pooling and encoding is much more stark in the rhetorical task, whereas the (low) performance is uniform across setups for the narrative task. For both tasks, zero-parameter pooling achieves top-scores. Values shown are using a 2-layer MLP classifier and embedded representations from a 3B parameter base model.

Encoding Methods:

Together: both spans are input jointly to the LLM.

Separately: spans are encoded independently, this is a cross-document approach.

Joint encoding allows the model to directly compare and align the spans in context. Separate encoding simulates scenarios where spans come from different documents or are processed in isolation.

Pooling Methods:

Learned: span tokens are pooled via trainable weights (learned by the probe).

LLM-attention: pooling uses the model's own attention distribution, taken at the final token(s) of the input span(s).

Learned pooling reflects the probe's ability to discover where information is encoded. LLM-attention pooling uses the model's internal attention distributions as a heuristic for relevance, without learning additional parameters.

Results for rhetorical parallelism show strong effects of both encoding and pooling. The best configuration overall uses joint encoding with LLM-attention pooling, achieving AUC scores of: 0.945 for LLaMA3-3B Instruct and 0.943 for LLaMA3-3B Base.

This suggests that rhetorical cues are most accessible when spans are presented together, allowing the model to contextualise them jointly and attend to corresponding structure. LLM-attention pooling effectively identifies those features without needing to learn where to look.

By contrast, the same LLM-attention strategy performs worse under separate encoding, dropping AUC by 1–2 points in both Base and Instruct variants. Learned pooling under separate encoding performs even worse (e.g. 0.928 for Base, 0.929 for Instruct), indicating that rhetorical relations are not robustly encoded span-by-span, but emerge during joint contextualisation.

Learned pooling under joint encoding performs well (e.g. 0.936 AUC for Instruct, 0.925 for Base), but does not surpass LLM-attention pooling—suggesting that the model's native attentional structure is already aligned with rhetorical similarity in jointly encoded spans.

The picture for narrative parallelism is markedly different. Across all encoding and pooling configurations, performance remains close to chance: The best configuration uses joint encoding with LLM-attention pooling, reaching 0.578 AUC (LLaMA3-3B Base). Other configurations (including separate encoding and learned pooling) perform similarly, ranging from 0.532 to 0.569.

Unlike the rhetorical task, there is no consistent benefit from joint encoding, nor from either pooling method. This suggests that narrative parallelism is not strongly encoded in the token-level representations, regardless of context.

In fact, the flat performance across pooling strategies indicates that neither the model's internal attention nor probe-learned pooling can reliably extract narrative structure. This

aligns with the hypothesis that narrative parallels require higher-order, globally integrated representations, which may not be encoded in a linear or local fashion at the span level.

These results suggest that rhetorical structure is linearly accessible, especially when spans are processed together and pooled using the model's own attention. In contrast, narrative structure is either more abstract or more sparsely represented, making it less accessible to simple probing.

The fact that LLM-attention pooling, a zero-parameter method, consistently performs on par with or better than learned pooling also highlights its value for probing tasks. It outperforms the self-attentive span pooling of Tenney et al. (2019a) in this setting, and may be especially useful in tasks where probe complexity must be minimised.

6.5.2 Probing vs. Prompting: Behaviour vs. Latent Knowledge

Task	Model Size	Variant	F1 Score	AUC
Narrative	1B	Base	0.67	0.50
		Instruction-Tuned	0.00	0.57
	3B	Base	0.67	0.59
		Instruction-Tuned	0.28	0.59
	8B	Base	0.14	0.67
		Instruction-Tuned	0.20	0.71
Rhetorical	1B	Base	0.67	0.14
		Instruction-Tuned	0.01	0.57
	3B	Base	0.00	0.05
		Instruction-Tuned	0.04	0.22
	8B	Base	0.08	0.38
		Instruction-Tuned	0.00	0.30

Table 6.4: **Model performance across tasks, sizes, and tuning variants.** F1 and AUC scores are reported for both base and instruction-tuned versions of each model size.

To evaluate how well models *express* vs. *encode* relational knowledge, I compare direct prompting (behavioural performance) with probing (latent information accessible from embeddings). Table 6.4 summarises the results of prompting across model sizes and variants, using a consistent classification prompt and measuring both F1 and AUC.

Prompted performance on the narrative parallelism task improves consistently with model size, especially in the instruction-tuned variants. AUC rises from 0.57 (1B Instruct) to 0.71 (8B Instruct), the highest prompted performance across all conditions and tasks. This suggests that emergent narrative alignment behaviour arises at larger scales and is amplified by instruction tuning.

Base models also improve with size (AUC: 0.50 $\rightarrow$ 0.67), though F1 scores remain erratic. For instance, the 1B and 3B Base models achieve F1 = 0.67 despite low AUCs,

likely due to overconfident single-class predictions.

Notably, even the best prompting results fall short of probe performance on rhetorical parallelism, but exceed probe performance on ARN. This highlights a key contrast: narrative knowledge appears behaviourally accessible with the right prompt and model scale, even if probing struggles to extract it from embeddings.

Prompting results for the rhetorical task are uniformly poor. AUCs remain low across all sizes and variants, ranging from 0.05 (3B Base) to 0.57 (1B Instruct). F1 scores are typically near zero — even when AUC is modest (e.g. 1B Instruct: F1 = 0.01, AUC = 0.57).

These results stand in stark contrast to the probing setup, where models achieve AUCs 0.94 under the best encoding/pooling configurations. This suggests that rhetorical structure is richly encoded in the latent space, but difficult to elicit directly via prompting. In other words, the model 'knows' the information, but does not reliably express it in generation without stronger cues. This supports Belinkov (2022)'s findings.

Across both tasks, F1 scores are low even when AUC is respectable. This reflects poor probability calibration in the model outputs: the classification head (based on prompt-completion likelihood) tends to favour a single class disproportionately, leading to sharp decision boundaries and poor thresholded predictions.

Manual inspection confirms that models frequently generate fluent, reasonable-sounding completions, but misclassify span relationships — especially on the rhetorical task. This suggests the model is capable of linguistic and stylistic fluency even when it misunderstands the core relational task.

These findings reinforce a key distinction: Prompting tests what the model *does*, while probing reveals what the model *knows*. For narrative parallelism, prompting outperforms probing at large scales, hinting at emergent behaviour not readily accessible to linear probes. For rhetorical parallelism, prompting fails, while probing succeeds — showing that latent representations encode structure even when behavioural outputs do not reflect it. This distinction is essential when evaluating LLMs: high latent capacity does not guarantee good behavioural performance, and vice versa. Unlocking this latent knowledge may require better prompting strategies, specialised decoding, or interactive protocols beyond standard few-shot prompting.

6.5.3 Localising Parallelism in Llama3 Models

To better understand where information about parallelism is represented within the LLaMA3 architecture, I apply the **Center-of-Gravity (CoG)** analysis introduced by Tenney et al. (2019a). This metric calculates the average layer depth at which the learned probe places the most weight during span pooling, offering a proxy for the location of task-relevant features within the model stack. Given learned scalar weights over hidden layers

during span pooling, I compute the CoG as:

$$\bar{E}_s = \sum_{\ell=0}^{L} \ell \cdot s(\ell)$$

where $s(\ell)$ is the softmax-normalised weight for layer ℓ.

I visualise CoG for the best classifier setup (i.e., the configuration achieving the highest AUC) for each task and model size.

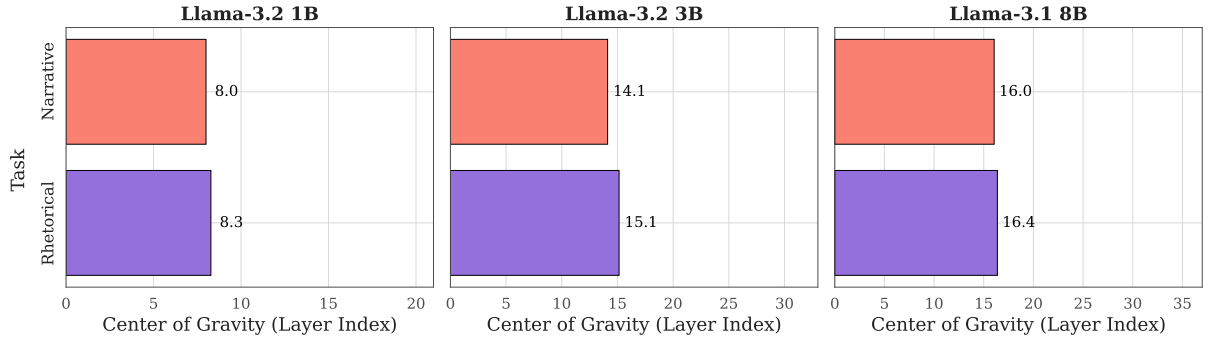


Figure 6.4: **Center-of-Gravity.** CoG represents an average over the learned layer weights of the span embeddings – effectively representing the layer of the model which is most useful to the classifier in solving a given task. Narrative and rhetorical parallelism appear to be localised in the mid-layers of the model, regardless of model depth.

For both tasks, the CoG shifts deeper in absolute layer index as model size increases: For narrative, CoG moves from 8.0 (1B) → 14.1 (3B) → 16.0 (8B). For rhetorical, from 8.3 → 15.1 → 16.4. This replicates a key finding from Tenney et al. (2019a): CoG increases linearly with layer depth, but its relative position remains stable.

Across all sizes and tasks, the relative CoG remains around the midpoint of the model (47–52% of total depth), suggesting that mid-to-late layers are where both types of parallelism are most encoded. There is a slight tendency for rhetorical CoG values to be higher than narrative ones at each size (e.g., 16.4 vs 16.0 in the 8B model), but the difference is too small to draw strong conclusions. These values are close enough that I interpret the variation as noise or model-specific fluctuation rather than a stable pattern.

The fact that both tasks peak in the middle layers aligns with expectations from prior work on layer specialisation in Transformer models: shallow layers tend to capture local and positional information, while middle layers are associated with syntactic and semantic features (e.g., grammaticality, NER, dependency structure), and late layers consolidate task-specific or interactional features.

That the signals for both the narrative and rhetorical parallelism probes peak in the mid-layers is consistent with these tasks requiring semantic and discourse-level representations.

6.6 Discussion

These results offer a nuanced view of how LLaMA 3 models encode parallel structures. Rhetorical parallelism is robustly and consistently encoded, yielding high AUCs across probing methods and model sizes, despite failing under prompting. It is accessible through both native attention-based pooling and learned span representations, with especially strong performance when spans are encoded jointly—suggesting that LLMs attend to rhetorical structure when contextual information is explicitly shared. In contrast, narrative parallelism proves more elusive: probing performance remains low across all configurations, though modest improvements emerge through direct prompting at larger model scales. For instance, rhetorical probing with joint encoding and LLM-attention pooling reached AUCs above 0.95, while the best narrative probing configuration remained below 0.60. In contrast, narrative prompting with the 8B Instruct model reached 0.71 AUC, outperforming all probe-based approaches.

This asymmetry points to a difference not just in difficulty but in representational encoding: rhetorical parallelism benefits from shallow, surface-level cues (e.g., syntactic or lexical repetition), whereas narrative parallelism demands deeper integration across context, structure, and semantics—something LLaMA does not reliably learn unless prompted. While model size and instruction tuning do not consistently improve probing performance, prompting reveals emergent capabilities at scale. The Center-of-Gravity analysis supports this view, showing that both tasks rely on information stores in the middle layers of the embedding model—consistent with the representation of higher-order abstractions—but rhetorical signals are more reliably localised.

For future work, I plan to conduct a detailed linguistic error analysis using tools such as spaCy, Stanza, and CLTK to extract features like part-of-speech overlap, semantic similarity, and length metrics. This will help categorise where different methods fail and whether certain linguistic cues are more or less predictive of successful detection. I also aim to analyse how performance in the analogy selection probing (ASP) task varies as a function of branch distance—i.e., how far apart the span pairs are in the original sermon—to test whether embeddings capture structural separation. If performance degrades with increasing span distance, it would suggest that LLM embeddings preserve some notion of structural proximity—useful for downstream tasks like discourse parsing or narrative segmentation. Finally, I intend to extend this work into a broader suite of literary probing benchmarks, drawing on annotated datasets for coreference, character relations, and event detection from Bamman et al. (2020), Bamman et al. (2019), and Sims et al. (2019) to evaluate how well large language models encode narrative and literary reasoning.

These findings suggest that benchmarking model capabilities via probing alone may understate their full potential—especially for high-level reasoning tasks, where emergent

behaviour arises only through prompting or interaction.

6.7 Conclusion

These experiments underscore a fundamental difference in the learnability of rhetorical versus narrative parallelism. Rhetorical structure is consistently recoverable from model representations, with strong performance across model scales and classifier types. In contrast, narrative parallelism proves far more elusive: even nonlinear probes offer only marginal improvements, and performance remains low across the board. This suggests that narrative structure is either less robustly encoded or requires representational formats not well captured by standard probing methods.

Surprisingly, neither increased model size nor instruction tuning consistently improves probe performance. In several cases, smaller models rival or even outperform their larger counterparts, and instruction tuning shows no systematic benefit. These findings suggest that neither scale nor supervised alignment to human preferences reliably enhances the encoding of relational features like parallelism in a way that makes them more accessible to probes. The lack of clear emergent capabilities highlights the limits of current LLM representations for tasks requiring deeper structural abstraction.

Encoding and pooling strategies play a crucial role in the rhetorical task but have little effect on narrative performance. Rhetorical relationships are most accessible when spans are jointly encoded and pooled using the model's own attention distributions, indicating that these features emerge during contextual comparison and are already aligned with the model's internal attentional structure. In contrast, narrative parallelism shows no consistent gains from encoding or pooling choices, suggesting it relies on more abstract or globally integrated representations that are not easily accessible from token-level embeddings.

I introduced a zero-parameter pooling method which uses the attention weights from the pretrained language model to combine an arbitrary number of token embeddings into a fixed-length vector. This approach demonstrated comparable or, often, stronger performance than learned token weighting and introduces no extra parameters for a classifier to learn. I believe this approach is promising and warrants further use and exploration.

Prompting and probing yield divergent pictures of model competence. For narrative parallelism, prompting outperforms probing at larger model sizes, suggesting that behavioural signs of narrative understanding can emerge even when the underlying representations are not linearly accessible. In direct contrast, rhetorical structure is richly encoded and easily extracted via probing, but models fail to express it reliably through prompting. This dissociation underscores a key distinction: what models *know* may not align with what they *do*. Latent relational knowledge can exist without being readily

expressed in generation, and high behavioural fluency does not guarantee structural understanding.

Center-of-Gravity analysis reveals that both narrative and rhetorical parallelism are predominantly encoded in the middle layers of LLaMA3 models, regardless of size. As model depth increases, the absolute CoG shifts upward, but its relative position remains near the architectural midpoint—echoing trends found in earlier BERT-based analyses. This suggests that discourse-level relational features, like parallelism, are not concentrated in early perceptual layers or late task-specialised ones, but emerge most saliently in intermediate layers where semantic abstractions are integrated.

Taken together, these results paint a somewhat messy picture. Unlike in human cognition, narrative does not appear to be a dominant or deeply integrated mode of representation in large language models. While LLMs can often mimic narrative coherence in their outputs–especially at larger scales and with instruction tuning–this capacity seems to emerge from statistical pattern learning rather than from structured, internalised narrative reasoning. In contrast, rhetorical parallelism shows clearer signs of being systematically encoded within the model's representations. Probing reveals that rhetorical structure is accessible in the mid layers of the model and can be reliably extracted using simple classifiers, even when it fails to surface in prompted outputs. In summary: LLMs may know rhetorical relations without being able to say them, and may say narrative-like things without clearly knowing their structure.

7

Conclusions and Future Work

This thesis has investigated how large language models represent structured literary and rhetorical patterns, with a particular focus on classical texts such as Biblical narratives. Across four chapters, I examined whether LLMs encode the abstract structures that organise meaning at the discourse level—such as rhetorical parallelism, narrative analogy, and intertextual symmetry— by applying models to literary tasks. I further explored how this information is surfaced or obscured by different probing and prompting methods.

In Chapter 3, I began by framing chiasmus detection as a constrained semantic similarity task, proposing a detection algorithm which required high semantic similarity of aligned pairs (e.g. A & A') and low semantic similarity of unaligned pairs (e.g. A & B). I showed that this simple and lightweight algorithm, applied in a sliding window fashion to every verse and half-verse in the Hebrew Bible, identified thousands of chiastic structures, primarily in dense narrative and poetic sections of text. Human annotation on the top-50 scoring chiasmi illuminated that the system achieved high precision (0.8 for full-verse, and 0.6 for half-verse). Future work on this problem includes integrating lexical and syntactic features, which may involved improving or creating parsing tools for Ancient Hebrew.

Then, in Chapter 4, I looked at how well human and machine translators are able to preserve intertextuality – echos of former texts – across a language shift. I found that human translators often amplify intertextual links, sometimes introducing or strengthening allusions that are only implicit in the source text. By contrast, machine translations tended to preserve the *relative structure* of intertextuality more faithfully, though with

reduced translation quality, especially for low resource translation pairs like Ancient Hebrew → Marathi. This pattern suggests that while human translators are influenced by theological or literary expectations, pretrained multilingual models exhibit a kind of neutral fidelity that, in some cases, better preserves the original semantic distances. However, these models still fall short in consistently maintaining structured links like typological echoes or allusive motifs, especially when operating on isolated verses. This motivates future work in retrieval-augmented translation, where providing models with semantically relevant context might help them learn to preserve higher-order narrative and theological structure across languages.

In Chapter 5, I examined type-scenes—recurring narrative schemas that structure episodes such as betrothals, battles, or journeys. I constructed a small evaluation set comprised of seven examples the 'betrothal well-encounter' from Biblical scholarship to evaluate whether models could identify analogical story patterns across non-overlapping texts. Surprisingly, a simple embedding-based approach using sentence Transformers (i.e. LaBSE) outperformed more structured retrieval strategies. Yet even the best models performed well below human ceiling, suggesting that while some coarse schema information is captured, LLMs struggle to recover the deeper causal, temporal, and thematic scaffolding that defines type-scenes. This finding echoes earlier concerns: that although embeddings are semantically rich, they may remain structurally shallow.

Chapter 6 synthesised and extended these themes by contrasting rhetorical and narrative parallelism in a unified framework. Using both probing and prompting techniques, I tested whether LLaMA3 models could identify rhetorical parallelisms in Latin text (e.g. 'veni, vidi, vici') or English narrative parallelisms in short narrative passages (i.e. a similarity of overall message/meaning despite surface differences, such as character names). Here, a striking asymmetry emerged: rhetorical parallelism was robustly encoded and easily surfaced via probing, while narrative parallelism remained elusive. Probing models consistently failed to recover narrative analogies, even at larger model scales. However, instruction-tuned variants exhibited modest success when directly prompted, suggesting that some capabilities may emerge only in behavioural settings. Further analysis revealed that probing classifiers relied most heavily on the mid-layers of an LLM's embedding—those typically associated with abstraction and composition—but only rhetorical signals were reliably localised.

Taken together, these chapters support a broader conclusion: current language models excel at recognising surface-level patterns that correlate with rhetorical or stylistic regularities, but struggle to generalise structural abstractions across distance, modality, or genre. Those tasks which may be somewhat approximated with semantic similarity – chiasmus detection, intertextuality preservation, and rhetorical parallelism detection — are far more easily recoverable than tasks which rely on structure (like type-scene retrieval) or

coherence and abstraction over longer distances (surfacing narrative parallelism). While prompting can sometimes elicit emergent behaviours suggestive of narrative reasoning, these behaviours are brittle and poorly grounded in the underlying representation space, indicating that they do not reflect stable internal abstractions.

This distinction has important implications for how narrative is represented (or not represented) within current architectures. It suggests that existing language models are not well suited to encoding narrative form as a first-class object. Transformers offer powerful local representations and scale remarkably well, but they lack inductive biases for tracking structured analogies, plot arcs, character roles, or thematic correspondences across documents. As such, even when narrative patterns are present in the training data, models do not reliably learn or apply them in a principled way.

Crucially, this limitation does not imply that rhetorical or narrative structure is inaccessible to machine learning systems, but rather that **it is poorly aligned with the dominant architectural assumptions of current LLMs**. One promising direction is therefore to move beyond purely data-driven pattern induction toward models that explicitly represent and manipulate structure. Neurosymbolic approaches offer a particularly compelling path forward. By combining learned representations with probabilistic or symbolic rules, such systems can encode constraints and abstractions—such as parallelism, analogy, or narrative role structure—that are difficult to infer implicitly from distributional data alone. Frameworks such as SCALLOP (Li et al., 2023b)

suggest how neural representations might be integrated with rule-based reasoning over structured forms, allowing models to reason about narrative relations rather than merely correlating surface features.

Relatedly, renewed attention to hierarchical and modular architectures may be warranted. Earlier work on hierarchical neural networks sought to capture multi-level structure in language but fell out of favour as large, flat transformers proved empirically dominant. However, recent developments—particularly mixture-of-experts (MoE) models—suggest new ways of reintroducing hierarchy and modularity at scale. In this setting, different experts or modules could specialise in rhetorical functions, discourse relations, or narrative roles, enabling models to reason across levels of abstraction without sacrificing scalability. Such architectures may provide the inductive bias needed to represent narrative structure explicitly, rather than treating it as an emergent by-product of token prediction.

Making language models more sensitive to rhetorical and narrative structure would offer tangible advantages beyond improved performance on literary benchmarks. Structurally aware models would be better equipped for long-range planning and generation, enabling more coherent extended writing and more consistent argumentation. They would also be better positioned to analyse previously unseen texts, reasoning about structure,

analogy, and thematic organisation rather than relying on memorised content or pretraining artifacts. In interpretive contexts, such models could support more robust forms of textual analysis, comparison, and explanation, grounded in abstract narrative relations rather than surface similarity alone.

These considerations underscore why narrative remains a critical stress test for claims that large language models *understand* language in a human-like sense. Narrative is not merely one genre among many; it is a fundamental mode of human cognition, shaping how we organise experience, infer causality, and transmit meaning across time. Our capacity to track characters, recognise structural echoes, and interpret events within broader narrative schemas is central to human intelligence. The persistent difficulty models have with these tasks – despite their fluency and scale – reveals a gap between linguistic competence and narrative understanding that cannot be closed by scale alone.

This limitation underscores why narrative remains a critical stress test for any claim that large language models *understand* in a human-like way. Narrative is not just one genre among many—it is a fundamental mode of human cognition, tightly bound to how we structure memories, interpret events, and communicate meaning. Our ability to track characters, infer motives, perceive causal links, and recognise deep structural echoes across time is a central function of natural (i.e. human) intelligence. The difficulty models have with capturing these patterns despite their fluency in language reveals a persistent gap between linguistic competence and narrative understanding. This work highlights that, while LLMs demonstrate remarkable sensitivity to surface features and can mimic aspects of interpretive behaviour, they still fall short of encoding the deeper conceptual structures that make narrative such a powerful and universal framework for human thought.

In parallel with architectural innovation, there is also a need for richer evaluation frameworks that foreground global structure rather than local coherence. One direction is to expand the type-scene evaluation set, potentially extending it to Homeric texts, in order to benchmark models on culturally and historically grounded narrative abstractions. More ambitiously, future work could move beyond supervised retrieval against established scholarly hypotheses toward the *de novo* discovery of recurring narrative tropes and structured scenes. Such an approach would open the door to studying narrative structure across genres and media not traditionally associated with type-scene theory, including Shakespearean drama, film, or serialised television.

In sum, this thesis delineates the current limits of large language models in representing rhetorical and narrative structure. While contemporary models capture surface regularities with surprising competence, deeper structural abstractions—such as analogy, typology, and narrative schema—remain poorly represented. Addressing these limitations will likely require a shift toward architectures and benchmarks that treat structure not as an incidental pattern to be inferred, but as a central object of representation and

reasoning.

Bibliography

Robert Alter. 1978. Biblical Type-Scenes and the Uses of Convention. *Critical Inquiry*, 5(2):355–368. Cited on pages 44, 86, 90, 92, and 93.

Robert Alter. 1983. How Convention Helps Us Read: The Case of the Bible's Annunciation Type-Scene. *Prooftexts*, 3(2):115–130. Cited on pages 44, 86, and 91.

Robert Alter. 2011. *The Art of Biblical Narrative*. Basic Books. Cited on pages 20, 41, and 66.

Walter Arend. 1933. *Die typischen Scenen bei Homer.* Forschungen zur klassischen Philologie. Problemata ; Hft. 7. Weidmann, Berlin. Cited on pages 44 and 86.

Andrew E. Arterbury. 2010. Breaking the Betrothal Bonds: Hospitality in John 4. *The Catholic Biblical Quarterly*, 72(1):63–83. Cited on pages 90 and 92.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open Weight Releases to Further Multilingual Progress. *Preprint*, arXiv:2405.15032. Cited on page 76.

Yannis Assael, Thea Sommerschield, and Jonathan Prag. 2019. Restoring ancient text using deep learning: A case study on Greek epigraphy. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6368–6375, Hong Kong, China. Association for Computational Linguistics. Cited on page 41.

Elie Assis. 2002. Chiasmus in Biblical Narrative: Rhetoric of Characterization. *Prooftexts*, 22(3):273–304. Cited on page 51.

Erich Auerbach. 1959. *Scenes from the Drama of European Literature; Six Essays.* New York, Meridian Books. Cited on pages 43 and 86.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. 2024. Leak, Cheat, Repeat: Data Contamination and Evaluation Malpractices in Closed-Source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93, St. Julian's, Malta. Association for Computational Linguistics. Cited on pages 22, 32, and 59.

David Bamman and Gregory Crane. 2008. The Logic and Discovery of Textual Allusion. Cited on pages 26, 40, 41, 42, 52, 66, and 68.

David Bamman and Gregory Crane. 2009. Discovering Multilingual Text Reuse in Literary Texts. Cited on pages 68 and 83.

David Bamman, Olivia Lewke, and Anya Mansoor. 2020. An Annotated Dataset of Coreference in English Literature. *Preprint*, arXiv:1912.01140. Cited on pages 35, 95, and 127.

David Bamman, Brendan O'Connor, and Noah A. Smith. 2013. Learning Latent Personas of Film Characters. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 352–361, Sofia, Bulgaria. Association for Computational Linguistics. Cited on pages 35 and 88.

David Bamman, Sejal Popat, and Sheng Shen. 2019. An annotated dataset of literary entities. In *Proceedings of the 2019 Conference of the North*, pages 2138–2144, Minneapolis, Minnesota. Association for Computational Linguistics. Cited on pages 34, 35, 95, and 127.

David Bamman, Ted Underwood, and Noah A. Smith. 2014. A Bayesian Mixed Effects Model of Literary Character. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 370–379, Baltimore, Maryland. Association for Computational Linguistics. Cited on pages 35 and 88.

Roland Barthes and Lionel Duisit. 1975. An Introduction to the Structural Analysis of Narrative. *New Literary History*, 6(2):237–272. Cited on page 26.

Yonatan Belinkov. 2022. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*, 48(1):207–219. Cited on pages 101, 105, and 125.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, Virtual Event Canada. ACM. Cited on pages 22 and 33.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics. Cited on pages 32 and 46.

Christopher Booker. 2004. *The Seven Basic Plots: Why We Tell Stories*. A&C Black. Cited on page 42.

Stephen Bothwell, Justin DeBenedetto, Theresa Crnkovich, Hildegund Müller, and David Chiang. 2023. Introducing Rhetorical Parallelism Detection: A New Task with Datasets, Metrics, and Baselines. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5007–5039, Singapore. Association for Computational Linguistics. Cited on pages 46, 51, 56, 107, and 117.

Brian Britt. 2002. Prophetic Concealment in a Biblical Type Scene. *The Catholic Biblical Quarterly*, 64(1):37–58. Cited on pages 45 and 86.

Peter Broadwell, David Mimno, and Timothy R. Tangherlini. 2017. The Tell-Tale Hat: Surfacing the Uncertainty in Folklore Classification. In *Journal of Cultural Analytics*. Cited on page 35.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *Preprint*, arXiv:2005.14165. Cited on pages 30 and 31.

Jerome Bruner. 1986. *Actual Minds, Possible Worlds*. Actual Minds, Possible Worlds. Harvard University Press, Cambridge, MA, US. Cited on page 21.

Patrick J. Burns, James A. Brofos, Kyle Li, Pramit Chaudhuri, and Joseph P. Dexter. 2021. Profiling of Intertextuality in Latin Literature Using Word Embeddings. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4900–4907, Online. Association for Computational Linguistics. Cited on pages 41, 58, 68, and 72.

Shuyang Cai and Wanyun Cui. 2023. Evade ChatGPT Detectors via A Single Space. *ArXiv*. Cited on page 22.

Joseph Campbell. 1949. *The Hero with a Thousand Faces*. Pantheon Books. Cited on pages 20 and 34.

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The Media Frames Corpus: Annotations of Frames Across Issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444, Beijing, China. Association for Computational Linguistics. Cited on page 26.

Dallas Card, Justin Gross, Amber Boydstun, and Noah A. Smith. 2016. Analyzing Framing through the Casts of Characters in the News. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1410–1420, Austin, Texas. Association for Computational Linguistics. Cited on page 36.

Nathanael Chambers. 2013. Event Schema Induction with a Probabilistic Entity-Driven Model. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Seattle, Washington, USA. Association for Computational Linguistics. Cited on page 45.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised Learning of Narrative Event Chains. In *Annual Meeting of the Association for Computational Linguistics*. Cited on pages 35 and 45.

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL '09, pages 602–610, USA. Association for Computational Linguistics. Cited on pages 35 and 45.

Kent Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. 2023. Speak, Memory: An Archaeology of Books Known to ChatGPT/GPT-4. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7312–7327, Singapore. Association for Computational Linguistics. Cited on page 32.

Matthew Clark. 2004. Formulas, metre and type-scenes. In Robert Fowler, editor, *The Cambridge Companion to Homer*, Cambridge Companions to Literature, pages 117–138. Cambridge University Press, Cambridge. Cited on page 44.

Neil Coffee, Jean-Pierre Koenig, Shakthi Poornima, Christopher W. Forstall, Roelant Ossewaarde, and Sarah L. Jacobson. 2013. The Tesserae Project: Intertextual analysis of Latin poetry. *Literary and Linguistic Computing*, 28(2):221–228. Cited on pages 41 and 72.

Paul Cook and Graeme Hirst. 2013. Automatically Assessing Whether a Text Is Clichéd, with Applications to Literary Analysis. Technical report, Association for Computational Linguistics. Cited on page 26.

Bruno Currie. 2019. The Iliad, the Odyssey, and Narratological Intertextuality*. *Symbolae Osloenses*, 93(1):157–188. Cited on page 66.

Dershowitz, Akiva, Koppel, and Dershowitz. 2015. Computerized Source Criticism of Biblical Texts. *Journal of Biblical Literature*, 134(2):253. Cited on page 41.

Joseph P. Dexter, Pramit Chaudhuri, Patrick J. Burns, Elizabeth D. Adams, Thomas J. Bolt, Adriana Cásarez, Jeffrey H. Flynt, Kyle Li, James F. Patterson, Ariane Schwartz, and Scott Shumway. 2024. A Database of Intertexts in Valerius Flaccus' Argonautica 1: A Benchmarking Resource for the Evaluation of Computational Intertextual Search of Latin Corpora. *Journal of Open Humanities Data*, 10(1). Cited on page 41.

Joseph P. Dexter, Theodore Katz, Nilesh Tripuraneni, Tathagata Dasgupta, Ajay Kannan, James A. Brofos, Jorge A. Bonilla Lopez, Lea A. Schroeder, Adriana Casarez, Maxim Rabinovich, Ayelet Haimson Lushkov, and Pramit Chaudhuri. 2017. Quantitative criticism of literary relationships. *Proceedings of the National Academy of Sciences of the United States of America*, 114(16):E3195–E3204. Cited on page 41.

Mary Douglas. 2007. *Thinking in Circles: An Essay on Ring Composition*. Yale University Press, New Haven London. Cited on page 53.

Marie Dubremetz and Joakim Nivre. 2015. Rhetorical Figure Detection: The Case of Chiasmus. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 23–31, Denver, Colorado, USA. Association for Computational Linguistics. Cited on pages 46, 51, and 54.

Marie Dubremetz and Joakim Nivre. 2017. Machine Learning for Rhetorical Figure Detection: More Chiasmus with Less Annotation. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 37–45, Gothenburg, Sweden. Association for Computational Linguistics. Cited on pages 46, 51, 55, 56, and 58.

Devlin Eckardt, Chelsea Helion, Helen Schmidt, Janice Chen, and Vishnu P. Murty. 2024. Storytelling changes the content and perceived value of event memories. *Cognition*, 251:105884. Cited on pages 22, 37, and 104.

James R. Edwards. 1989. Markan Sandwiches. The Significance of Interpolations in Markan Narratives. *Novum Testamentum*, 31(3):193–216. Cited on page 49.

David K. Elson. 2012. Detecting Story Analogies from Annotations of Time, Action and Agency. Cited on pages 36 and 106.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT Sentence Embedding. *Preprint*, arXiv:2007.01852. Cited on pages 70, 73, and 94.

Mark A. Finlayson. 2015. *ProppLearner* : Deeply Annotating a Corpus of Russian Folktales to Enable the Machine Learning of a Russian Formalist Theory. *Digital Scholarship in the Humanities*, page fqv067. Cited on pages 35 and 87.

John Rupert Firth. 1957. *A Synopsis of Linguistic Theory*. Cited on page 29.

Monika Fludernik. 2009. *An Introduction to Narratology*. Routledge. Cited on page 34.

Niv Fono, Harel Moshayof, Eldar Karol, Itai Assraf, and Mark Last. 2024. Embible: Reconstruction of Ancient Hebrew and Aramaic Texts Using Transformers. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 846–852, St. Julian's, Malta. Association for Computational Linguistics. Cited on page 41.

Richard Friedman. 2019. *Who Wrote the Bible?* Simon & Schuster. Cited on page 41.

Esther Fuchs. 1987. Structure and Patriarchal Functions in the Biblical Betrothal Type-Scene: Some Preliminary Notes. *Journal of Feminist Studies in Religion*, 3(1):7–13. Cited on pages 45 and 86.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673. Cited on pages 22 and 32.

Gärard Genette. 1997. *Palimpsests: Literature in the Second Degree*. U of Nebraska Press. Cited on pages 36 and 42.

Pablo Gervás and Gonzalo Méndez. 2024. Tagging Narrative with Propp's Character Functions Using Large Language Models. In *Text2Story@ECIR*. Cited on pages 35 and 87.

Sayan Ghosh and Shashank Srivastava. 2022. ePiC: Employing Proverbs in Context as a Benchmark for Abstract Language Understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3989–4004, Dublin, Ireland. Association for Computational Linguistics. Cited on page 108.

Dvir Ginzburg, Itzik Malkiel, Oren Barkan, Avi Caciularu, and Noam Koenigstein. 2021. Self-Supervised Document Similarity Ranking via Contextualized Language Models and

Hierarchical Inference. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3088–3098, Online. Association for Computational Linguistics. Cited on page 46.

Ira Goldstein and Seymour Papert. 1977. Artificial intelligence, language, and the study of knowledge. *Cognitive Science*, 1(1):84–123. Cited on page 25.

Ernst H. Gombrich. 1969. *Art and Illusion: A Study in the Psychology of Pictorial Representation - Millennium Edition*, volume 5. Princeton University Press. Cited on page 98.

Helmut Grabner, Juergen Gall, and Luc Van Gool. 2011. What makes a chair a chair? In CVPR. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1529–1536. Cited on page 27.

Frederick E. Greenspahn. 1994. *When Brothers Dwell Together: The Preeminence of Younger Siblings in the Hebrew Bible*. Oxford University Press. Cited on page 86.

Harry Hagan. 2019. Basic Plots in the Bible: A Literary Approach to Genre. *BiBlical Theology BulleTin*, 49:198–213. Cited on page 42.

Alon Halevy, Peter Norvig, and Fernando Pereira. 2009. The Unreasonable Effectiveness of Data. *IEEE Intelligent Systems*, 24(2):8–12. Cited on pages 29 and 31.

Richard B. Hays. 1989. *Echoes of Scripture in the Letters of Paul*. Yale University Press. Cited on page 66.

Linyang He, Peili Chen, Ercong Nie, Yuanning Li, and Jonathan R. Brennan. 2024. Decoding Probing: Revealing Internal Linguistic Structures in Neural Language Models using Minimal Pairs. *Preprint*, arXiv:2403.17299. Cited on page 105.

John Paul Heil. 2001. The Chiastic Structure and Meaning of Paul's Letter to Philemon. *Biblica*, 82(2):178–206. Cited on page 49.

John Hewitt and Percy Liang. 2019. Designing and Interpreting Probes with Control Tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics. Cited on page 111.

Geoffrey E. Hinton, Jay L. McClelland, and David E. Rumelhart. 1986. Distributed representations. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*, pages 77–109. MIT Press, Cambridge, MA, USA. Cited on page 29.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. 2024. The Platonic Representation Hypothesis. Cited on page 33.

Labiba Jahan and Mark Finlayson. 2019. Character Identification Refined: A Proposal. In *Proceedings of the First Workshop on Narrative Understanding*, pages 12–18, Minneapolis, Minnesota. Association for Computational Linguistics. Cited on pages 35 and 87.

Labiba Jahan, Rahul Mittal, and Mark Finlayson. 2021. Inducing Stereotypical Character Roles from Plot Structure. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 492–497, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. Cited on pages 13, 21, 35, 88, and 98.

Labiba Jahan, Rahul Mittal, W. Victor H. Yarlott, and Mark Finlayson. 2020. A Straightforward Approach to Narratologically Grounded Character Identification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6089–6100, Barcelona, Spain (Online). International Committee on Computational Linguistics. Cited on pages 35 and 88.

Thennal D. K, Tim Fischer, and Chris Biemann. 2024. Large Language Models Are Overparameterized Text Encoders. *Preprint*, arXiv:2410.14578. Cited on pages 30 and 102.

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. *Preprint*, arXiv:2304.03245. Cited on page 32.

Timothy Keller. 2014. *Every Good Endeavor: Connecting Your Work to God's Work*. Penguin Publishing Group. Cited on page 27.

Mike Kestemont, Justin Stover, Moshe Koppel, Folgert Karsdorp, and Walter Daelemans. 2016. Authenticating the writings of Julius Caesar. *Expert Systems With Applications*, 63:86–96. Cited on page 41.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *CoRR*. Cited on page 116.

Maria Konstantinidou, John Pavlopoulos, and Elton Barker. 2024. Exploring intertextuality across the Homeric poems through language models. In *Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024)*, pages 260–268, Hybrid in Bangkok, Thailand and online. Association for Computational Linguistics. Cited on pages 40 and 68.

Moshe Koppel, Navot Akiva, Idan Dershowitz, and Nachum Dershowitz. 2011. Unsupervised Decomposition of a Document into Authorial Components. In *Proceedings*

of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 1356–1364, Portland, Oregon, USA. Association for Computational Linguistics. Cited on page 41.

Moshe Koppel, Jonathan Schler, and Elisheva Bonchek-Dokow. 2007. Measuring Differentiability: Unmasking Pseudonymous Authors. *J. Mach. Learn. Res.*, 8:1261–1276. Cited on page 41.

Andreas J. Köstenberger and Gregory Goswell. 2023. *Biblical Theology: A Canonical, Thematic, and Ethical Approach*. Crossway. Cited on page 45.

John Koutsikakis, Ilias Chalkidis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2020. GREEK-BERT: The Greeks visiting Sesame Street. In *11th Hellenic Conference on Artificial Intelligence*, SETN 2020, pages 110–117, New York, NY, USA. Association for Computing Machinery. Cited on page 41.

Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating Unsupervised Style Transfer as Paraphrase Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 737–762, Online. Association for Computational Linguistics. Cited on page 109.

Julia Kristeva. 1986. *The Kristeva Reader*. Columbia University Press. Cited on page 66.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2017. ImageNet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90. Cited on page 31.

William Labov and Joshua Waletzky. 1997. Narrative analysis: Oral versions of personal experience. *Journal of Narrative & Life History*, 7(1-4):3–38. Cited on page 34.

Rob Latham. 2017. *Science Fiction Criticism: An Anthology of Essential Writings*. Bloomsbury Publishing. Cited on page 26.

Hongmi Lee, Buddhika Bellana, and Janice Chen. 2020. What can narratives tell us about the neural bases of human memory? *Current Opinion in Behavioral Sciences*, 32:111–119. Cited on pages 22 and 104.

Hongmi Lee and Janice Chen. 2022. Predicting memory from the network structure of naturalistic events. *Nature Communications*, 13(1):4235. Cited on pages 22, 37, 44, and 104.

John Lee. 2007. A Computational Model of Text Reuse in Ancient Literary Texts. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 472–479, Prague, Czech Republic. Association for Computational Linguistics. Cited on pages 40, 41, and 68.

Belinda Z. Li, Maxwell Nye, and Jacob Andreas. 2021. Implicit Representations of Meaning in Neural Language Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1813–1827, Online. Association for Computational Linguistics. Cited on page 102.

Daoyang Li, Haiyan Zhao, Qingcheng Zeng, and Mengnan Du. 2025. Exploring Multilingual Probing in Large Language Models: A Cross-Language Analysis. *Preprint*, arXiv:2409.14459. Cited on page 105.

Xinze Li, Zhenghao Liu, Chenyan Xiong, Shi Yu, Yu Gu, Zhiyuan Liu, and Ge Yu. 2023a. Structure-Aware Language Model Pretraining Improves Dense Retrieval on Structured Data. *Preprint*, arXiv:2305.19912. Cited on pages 85 and 96.

Ziyang Li, Jiani Huang, and Mayur Naik. 2023b. Scallop: A Language for Neurosymbolic Programming. *Preprint*, arXiv:2304.04812. Cited on pages 98 and 133.

Tom Lippincott. 2019. Graph convolutional networks for exploring authorship hypotheses. In *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 76–81. Cited on page 41.

Burke O. Long. 1973. 2 Kings III and Genres of Prophetic Narrative. *Vetus Testamentum*, 23(3):337–348. Cited on page 86.

Robert Lowth. 1989. Lectures on the Sacred Poetry of the Hebrews (1753, translated from Latin into English by Richard Gregory in 1787), Lectures IV (extracted), V (less the first and last paragraphs), XIV, XVII (less the first two paragraphs and the second half of the lecture, viz. pp. 377–387 of the 1787 edition), XIX. In John Drury, editor, *Critics of the Bible, 1724–1873*, Cambridge English Prose Texts, pages 69–102. Cambridge University Press, Cambridge. Cited on page 51.

Jiaming Luo, Yuan Cao, and Regina Barzilay. 2019. Neural Decipherment via Minimum-Cost Flow: From Ugaritic to Linear B. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3146–3155, Florence, Italy. Association for Computational Linguistics. Cited on page 41.

Jiaming Luo, Frederik Hartmann, Enrico Santus, Regina Barzilay, and Yuan Cao. 2021. Deciphering undersegmented ancient scripts using phonetic prior. *Transactions of the Association for Computational Linguistics*, 9:69–81. Cited on page 41.

Alasdair MacIntyre. 2013. *After Virtue*. A&C Black. Cited on page 27.

Enrique Manjavacas, Brian Long, and Mike Kestemont. 2019. On the Feasibility of Automated Detection of Allusive Text Reuse. In *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 104–114, Minneapolis, USA. Association for Computational Linguistics. Cited on page 68.

Enrique Manjavacas Arevalo, Laurence Mellerin, and Mike Kestemont. 2021. Quantifying Contextual Aspects of Inter-annotator Agreement in Intertextuality Research. In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 31–42, Punta Cana, Dominican Republic (online). Association for Computational Linguistics. Cited on page 41.

Samuel Marks and Max Tegmark. 2024. The Geometry of Truth: Emergent Linear Structure in LLM Representations of True/False Datasets. Cited on page 32.

Lee Roy Martin. 2018. The Chiastic Structure of Psalm 106. *Old Testament Essays*, 31(3):506–521. Cited on page 51.

Arya D. McCarthy, Rachel Wicks, Dylan Lewis, Aaron Mueller, Winston Wu, Oliver Adams, Garrett Nicolai, Matt Post, and David Yarowsky. 2020. The Johns Hopkins University Bible Corpus: 1600+ Tongues for Typological Exploration. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2884–2892, Marseille, France. European Language Resources Association. Cited on pages 71 and 90.

Barbara McGillivray, Simon Hengchen, Viivi Lähteenoja, Marco Palma, and Alessandro Vatri. 2019. A computational approach to lexical polysemy in Ancient Greek. *Digital Scholarship in the Humanities*, 34(4):893–907. Cited on page 41.

Hope McGovern, Hale Sirin, and Tom Lippincott. 2025a. Characterizing the Effects of Translation on Intertextuality using Multilingual Embedding Spaces. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 161–167, Albuquerque, New Mexico. Association for Computational Linguistics. Cited on page 65.

Hope McGovern, Hale Sirin, and Tom Lippincott. 2025b. Computational Discovery of Chiasmus in Ancient Religious Text. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 154–160, Albuquerque, New Mexico. Association for Computational Linguistics. Cited on page 49.

Hope McGovern, Hale Sirin, Tom Lippincott, and Andrew Caines. 2024. Detecting Narrative Patterns in Biblical Hebrew and Greek. In *Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024)*, pages 269–279, Hybrid in Bangkok, Thailand and online. Association for Computational Linguistics. Cited on pages 85 and 89.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *Preprint*, arXiv:1301.3781. Cited on page 29.

Gosse Minnema, Sara Gemelli, Chiara Zanchi, Tommaso Caselli, and Malvina Nissim. 2022. Dead or Murdered? Predicting Responsibility Perception in Femicide News Reports. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1078–1090, Online only. Association for Computational Linguistics. Cited on page 36.

Marvin Minsky. 1974. A Framework for Representing Knowledge. Cited on pages 22, 37, 38, and 42.

Marvin Minsky. 1988. *Society Of Mind*. Simon and Schuster. Cited on page 27.

Maria Moritz, Andreas Wiederhold, Barbara Pavlek, Yuri Bizzoni, and Marco Büchler. 2016. Non-Literal Text Reuse in Historical Texts: An Approach to Identify Reuse Transformations and its Application to Bible Reuse. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1849–1859, Austin, Texas. Association for Computational Linguistics. Cited on pages 40 and 68.

Nasrin Mostafazadeh, Alyson Grealish, Nathanael Chambers, James Allen, and Lucy Vanderwende. 2016. CaTeRS: Causal and Temporal Relation Scheme for Semantic Annotation of Event Structures. In *Proceedings of the Fourth Workshop on Events*, pages 51–61, San Diego, California. Association for Computational Linguistics. Cited on page 36.

Frederick Mosteller and David L Wallace. 1963. Inference in an Authorship Problem. *Source: Journal of the American Statistical Association*, 58(302):275–309. Cited on page 41.

Jingcheng Niu, Wenjie Lu, and Gerald Penn. 2022. Does BERT Rediscover a Classical NLP Pipeline? In *International Conference on Computational Linguistics*. Cited on page 105.

Benjamin J. Noonan. 2016. Egyptian Loanwords as Evidence for the Authenticity of the Exodus and Wilderness Traditions. In James K. Hoffmeier, Alan R. Millard, and

Gary A. Rendsburg, editors, *?Did I Not Bring Israel Out of Egypt?"*, volume 13 of *Biblical, Archaeological, and Egyptological Perspectives on the Exodus Narratives*, pages 49–68. Penn State University Press. Cited on page 69.

Malte Ostendorff, Terry Ruas, Till Blume, Bela Gipp, and Georg Rehm. 2020. Aspect-based Document Similarity for Research Papers. *Preprint*, arXiv:2010.06395. Cited on pages 46 and 88.

Jessica Ouyang and Kathleen McKeown. 2015. Modeling Reportable Events as Turning Points in Narrative. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2149–2158, Lisbon, Portugal. Association for Computational Linguistics. Cited on page 34.

Conley Owens. 2010. Bible Cross References. https://www.openbible.info/labs/cross-references/. Cited on page 72.

Pinelopi Papalampidi, Frank Keller, Lea Frermann, and Mirella Lapata. 2020. Screenplay Summarization Using Latent Narrative Structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1920–1933, Online. Association for Computational Linguistics. Cited on pages 36 and 46.

Pinelopi Papalampidi, Frank Keller, and Mirella Lapata. 2019. Movie Plot Analysis via Turning Point Identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1707–1717, Hong Kong, China. Association for Computational Linguistics. Cited on pages 36 and 46.

Pinelopi Papalampidi, Frank Keller, and Mirella Lapata. 2021. Movie Summarization via Sparse Graph Construction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13631–13639. Cited on page 46.

Matteo Pasquinelli and Vladan Joler. 2021. The Nooscope manifested: AI as instrument of knowledge extractivism. *AI & SOCIETY*, 36(4):1263–1280. Cited on pages 22 and 33.

Paul Ricoeur. 1984. *Time and Narrative, Volume 1.* University of Chicago Press, Chicago. Cited on page 27.

Valerio Perrone, Marco Palma, Simon Hengchen, Alessandro Vatri, Jim Q. Smith, and Barbara McGillivray. 2019. GASC: Genre-Aware Semantic Change for Ancient Greek. pages 56–66. Cited on page 41.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language Models as Knowledge Bases? In *Proceedings*

of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2463–2473, Hong Kong, China. Association for Computational Linguistics. Cited on pages 29 and 32.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-Theoretic Probing for Linguistic Structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622, Online. Association for Computational Linguistics. Cited on page 111.

Andrew Piper, Sunyam Bagga, Laura Monteiro, Andrew Yang, Marie Labrosse, and Yu Lu Liu. 2021a. Detecting Narrativity Across Long Time Scales. In *Workshop on Computational Humanities Research*. Cited on page 34.

Andrew Piper, Richard Jean So, and David Bamman. 2021b. Narrative Theory for Computational Narrative Understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 298–311, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. Cited on pages 26, 27, 28, 69, 87, and 105.

Andrew Piper, Michael Xu, and Derek Ruths. 2024. The Social Lives of Literary Characters: Combining citizen science and language models to understand narrative social networks. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 472–482, Miami, USA. Association for Computational Linguistics. Cited on pages 34 and 35.

Michael D. Press. 2016. A Single Combat Type-Scene in the Hebrew Bible? *Hebrew Studies*, 57:93–115. Cited on pages 45 and 86.

Ira M. Price. 1895. Some Queries about the Book of Daniel. *The Biblical World*, 6(4):264–269. Cited on page 69.

Gerald Prince. 1973. *A Grammar of Stories: An Introduction*. Mouton. Cited on page 34.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. Cited on page 31.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Preprint*, arXiv:1910.10683. Cited on page 93.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. 2021. Probing the Probing Paradigm: Does Probing Accuracy Entail Task Relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3363–3377, Online. Association for Computational Linguistics. Cited on page 111.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics. Cited on page 78.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Preprint*, arXiv:1908.10084. Cited on page 94.

Nils Reiter, Anette Frank, and Oliver Hellwig. 2014. An NLP-based cross-document approach to narrative structure discovery. *Literary and Linguistic Computing*, 29(4):583–605. Cited on page 35.

David E. Rumelhart. 1975. Notes on a Schema for Stories. In DANIEL G. Bobrow and ALLAN Collins, editors, *Representation and Understanding*, pages 211–236. Morgan Kaufmann, San Diego. Cited on pages 26 and 28.

Yisi Sang, Xiangyang Mou, Jing Li, Jeffrey Stanton, and Mo Yu. 2022. A Survey of Machine Narrative Reading Comprehension Assessments. arXiv. Cited on page 69.

George Savran. 2003. Theophany as Type Scene. *Prooftexts*, 23(2):119–149. Cited on page 45.

Roger C. Schank and Robert P. Abelson. 1975. Scripts, Plans and Knowledge. In *International Joint Conference on Artificial Intelligence*. Cited on pages 15, 37, 39, and 45.

Roger C. Schank and Robert P. Abelson. 1979. Scripts, plans, goals and understanding, an inquiry into human knowledge structures. *Journal of Pragmatics*, 3(2):211–217. Cited on page 21.

Felix Schneider, Phillip Brandes, Björn Barz, Sophie Marshall, and Joachim Denzler. 2021. Data-Driven Detection of General Chiasmi Using Lexical and Semantic Features. In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 96–100, Punta Cana, Dominican Republic (online). Association for Computational Linguistics. Cited on pages 46, 51, 55, 56, and 58.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, Stella Biderman, Adria Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, Eric J. Michaud, Stephen Casper, Max Tegmark, William Saunders, David Bau, Eric Todd, Atticus Geiger, Mor Geva, Jesse Hoogland, Daniel Murfet, and Tom McGrath. 2025. Open Problems in Mechanistic Interpretability. *Preprint*, arXiv:2501.16496. Cited on page 111.

Matthew Sims and David Bamman. 2020. Measuring Information Propagation in Literary Social Networks. *Preprint*, arXiv:2004.13980. Cited on page 34.

Matthew Sims, Jong Ho Park, and David Bamman. 2019. Literary Event Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3623–3634, Florence, Italy. Association for Computational Linguistics. Cited on pages 34, 35, 95, and 127.

Hale Sirin and Thomas Lippincott. 2024. Dynamic embedded topic models and change-point detection for exploring literary-historical hypotheses. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 231–236, St. Julians, Malta. Association for Computational Linguistics. Cited on pages 43 and 86.

Thea Sommerschield, Yannis Assael, John Pavlopoulos, Vanessa Stefanak, Andrew Senior, Chris Dyer, John Bodel, Jonathan Prag, Ion Androutsopoulos, and Nando De Freitas. 2023. Machine Learning for Ancient Languages: A Survey. *Computational Linguistics*, pages 1–44. Cited on page 41.

Zhivar Sourati, Filip Ilievski, Pia Sommerauer, and Yifan Jiang. 2024. ARN: Analogical Reasoning on Narratives. *Transactions of the Association for Computational Linguistics*, 12:1063–1086. Cited on pages 98, 106, 108, and 118.

William Strunk. 2011. *The Elements of Style*. Cited on page 37.

Simeng Sun, Kalpesh Krishna, Andrew Mattarella-Micke, and Mohit Iyyer. 2021. Do Long-Range Language Models Actually Use Long-Range Context? *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 807–822. Cited on page 32.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual Translation with Extensible Multilingual Pretraining and Finetuning. *Preprint*, arXiv:2008.00401. Cited on pages 16 and 71.

Archer Taylor, Vladimir Propp, Svatava Pirkova-Jakobson, and Laurence Scott. 1959. *Morphology of the Folktale*, volume 3. Cited on pages 20, 35, and 86.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019a. BERT Rediscovers the Classical NLP Pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics. Cited on pages 14, 29, 102, 105, 120, 121, 124, 125, and 126.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019b. What do you learn from context? Probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*. Cited on pages 103, 104, 107, 112, and 114.

Katherine Thai, Yapei Chang, Kalpesh Krishna, and Mohit Iyyer. 2022. RELIC: Retrieving Evidence for Literary Claims. *Preprint*, arXiv:2203.10053. Cited on pages 88 and 89.

R.A. Torrey and J. Canne. 1982. *The Treasury of Scripture Knowledge: Five-hundred Thousand Scripture References and Parallel Passages*. Hendrickson Publishers Marketing, LLC. Cited on pages 67 and 72.

C. van der Waal. 1980. The Continuity Between the Old and New Testaments. *Neotestamentica*, 14:1–20. Cited on page 75.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Neural Information Processing Systems*. Cited on pages 30 and 31.

Kurt Vonnegut. 1969. *Slaughterhouse-Five*. Random House Publishing Group. Cited on page 28.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual E5 Text Embeddings: A Technical Report. Cited on page 58.

Yue Wang, Hung Le, Akhilesh Deepak Gotmare, Nghi D. Q. Bui, Junnan Li, and Steven C. H. Hoi. 2023. CodeT5+: Open Code Large Language Models for Code Understanding and Generation. *Preprint*, arXiv:2305.07922. Cited on page 96.

Yue Wang, Weishi Wang, Shafiq Joty, and Steven C. H. Hoi. 2021. CodeT5: Identifier-aware Unified Pre-trained Encoder-Decoder Models for Code Understanding and Generation. *Preprint*, arXiv:2109.00859. Cited on page 96.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The Benchmark of Linguistic Minimal Pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392. Cited on page 102.

John W. Welch. 1981. *Chiasmus in Antiquity: Structures, Analyses, Exegesis*. Gerstenberg. Cited on page 51.

Philipp Wicke, Lea Hirlimann, and Joao Miguel Cunha. 2024. Using Analogical Reasoning to Prompt LLMs for their Intuitions of Abstract Spatial Schemas. *The First Workshop on Analogical Abstraction in Cognition, Perception, and Language (Analogy-ANGLE)*. Cited on page 106.

Patrick Henry Winston. 2011. The Strong Story Hypothesis and the Directed Perception Hypothesis. In *AAAI Fall Symposium: Advances in Cognitive Systems*. Cited on pages 25 and 37.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics. Cited on pages 76 and 113.

Borui Xu, Yao Chen, Zeyi Wen, Weiguo Liu, and Bingsheng He. 2025. Evaluating Small Language Models for News Summarization: Implications and Factors Influencing Performance. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4909–4922, Albuquerque, New Mexico. Association for Computational Linguistics. Cited on page 30.

W. Victor H. Yarlott and Mark A. Finlayson. 2016. ProppML: A Complete Annotation Scheme for Proppian Morphologies. In *OASIcs, Volume 53, CMN 2016*, volume 53, pages 8:1–8:19. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. Cited on pages 35 and 87.

W. Victor H. Yarlott, Armando Ochoa, Anurag Acharya, Laurel Bobrow, Diego Castro Estrada, Diana Gomez, Joan Zheng, David McDonald, Chris Miller, and Mark A. Finlayson. 2022. Finding Trolls Under Bridges: Preliminary Work on a Motif Detector. Cited on pages 26, 35, and 87.

Michihiro Yasunaga, Xinyun Chen, Yujia Li, Panupong Pasupat, Jure Leskovec, Percy Liang, Ed H. Chi, and Denny Zhou. 2024. Large Language Models as Analogical Reasoners. In *ICLR 2024*. arXiv. Cited on page 106.

Asieh Zadbood, Samuel Nastase, Janice Chen, Kenneth A Norman, and Uri Hasson. 2022. Neural representations of naturalistic events are updated as our understanding of the past changes. *eLife*, 11:e79045. Cited on page 104.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2024. Instruction Tuning for Large Language Models: A Survey. *Preprint*, arXiv:2308.10792. Cited on pages 31 and 118.