

# Mapping Meaning in Latin with Large Language Models: A Multi-Task Evaluation of Preverbed Motion Verbs and Spatial Relation Detection in LLMs

Andrea Farina<sup>1</sup>, Andrea Ballatore<sup>1</sup> and Barbara McGillivray<sup>1</sup>

<sup>1</sup>King's College London, Strand Campus, Strand, WC2R 2LS, London, United Kingdom

## Abstract

This paper evaluates the capabilities of Large Language Models (LLMs) on three interrelated linguistic tasks in Latin: preverbed motion verb identification, spatial relation (SR) classification, and SR type disambiguation. We evaluate GPT-4, Llama, and Mistral under zero-shot and few-shot settings, using a manually annotated dataset of Latin sentences drawn from different authors, text types, and historical periods (3rd century BCE – 2nd century CE) as our gold standard. Results show that GPT-4 consistently outperforms open-weight models, particularly in zero-shot scenarios, likely due to its substantial pretraining exposure to Latin. However, even GPT-4 struggles with syntactic disambiguation, especially in linking proper nouns to their governing verbs. SR classification performance is skewed by dataset imbalance, and SR type disambiguation errors often stem from over-reliance on salience over syntax. Qualitative analysis reveals common patterns of overgeneration and uncertainty across tasks. Our findings underscore the potential of LLMs for historical language processing while highlighting persistent challenges related to ambiguity, entity linking, and syntactic reasoning. This study represents the first evaluation of SR recognition in historical languages and lays the groundwork for future domain-adapted fine-tuning approaches in Computational Humanities.

## Keywords

Large Language Models, Latin, motion verbs, spatial relation classification, SR type disambiguation

## 1. Introduction

The central aim of this study is to evaluate the ability of Large Language Models (LLMs) to analyse spatiality in Latin texts, with a focus on motion events and their syntactic and semantic environments. In Latin, motion verbs, i.e., verbs denoting movement (cf. class 51 in [1]), often combine with *preverbs* — prefixes that attach onto verbal bases to express (among other things) nuanced spatial meanings (cf. Section 4.2). For example, the Latin motion verb *eo* ‘go’ can be prefixed with different preverbs, which deeply modify its semantics (e.g., the preverbs *ex-* ‘out of’ and *in-* ‘into’ generate *exeo* ‘exit’ and *ineo* ‘enter’). This preverbal modification is crucial for encoding spatial relations (SRs) in Latin, as directionality and argument structure are frequently expressed jointly by the verbal root and its preverb. Motion events [2] involve an entity *E* moving from a *Source* (the starting point of motion) to a *Goal* (the ending point of motion), and along a *Path* (the

set of continuous locations crossed by *E* while moving from the *Source* to the *Goal*) [3]. This usually happens both in literal and non-literal contexts [4].

This paper explores to what extent LLMs can handle such constructions in Latin, taken as an example of a historical and morphologically complex language. We focus on preverbed motion verbs as an area that demands the integration of lexical, syntactic, and spatial information. To evaluate the models’ performance, we design three linguistic tasks targeting different layers of interpretation relevant to motion events (Section 3). Preverbs often provide crucial cues to argument structure and directionality (e.g., *abeo* ‘go away’ vs. *adeo* ‘go toward/to’), which may pose significant challenges for automatic disambiguation with LLMs. This allows us to assess the extent to which LLMs are able to perform linguistic annotation on challenging verbal constructions such as preverbed motion verbs, which are structurally more complex than their non-preverbed counterparts.

## 2. Related Work

The application of NLP techniques to Latin has advanced significantly in recent years, driven by developments in neural networks, transformer-based architectures, and the increasing availability of large-scale digital corpora. A key benchmark in this area has been the Evalatin shared tasks, held annually since 2020, which provide a structured evaluation framework for a range of Latin

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 – 26, 2025, Cagliari, Italy

\*Corresponding author.

✉ andrea.farina@kcl.ac.uk (A. Farina); andrea.ballatore@kcl.ac.uk (A. Ballatore); barbara.mcgillivray@kcl.ac.uk (B. McGillivray)

🌐 <https://www.kcl.ac.uk/people/andrea-farina> (A. Farina);


<https://aballatore.space/lab> (A. Ballatore);

<https://www.kcl.ac.uk/people/barbara-mcgillivray> (B. McGillivray)

🆔 <https://orcid.org/0000-0002-1948-9008> (A. Farina);

0000-0003-3477-7654 (A. Ballatore); 0000-0003-3426-8200

(B. McGillivray)

 © 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

NLP tasks [5]. Among the most influential recent developments in Latin NLP is the introduction of contextualised language models. LatinBERT [6], a contextualised model trained on a substantial corpus comprising 642.7 million words spanning from Classical Antiquity to the contemporary period, has been shown to perform well in tasks such as lemmatisation, part-of-speech (POS) tagging, and syntactic parsing. LatinBERT has also shown promise in word sense disambiguation [7, 8] and named entity recognition [9].

Generative LLMs have demonstrated impressive performance across several NLP tasks [10, 11]. However, their success relies on vast amounts of data [12, 13], which is not typically achieved by most historical corpora. The potential of LLMs for Latin is beginning to be systematically evaluated. Volk et al. [14] showed that GPT-4-based machine translation substantially outperforms previous approaches when tested on 16th-century correspondence written in Latin and Early New High German. In addition to translation, they also evaluated GPT-4 for paragraph-level summarisation of Latin texts, with its output compared against human-generated summaries.

Parallel to these developments, efforts have been made to extract SRs from text, not only in computational linguistics, but also in information retrieval and geospatial analytics. Early approaches relied on rule-based methods and regular expressions, which have since evolved into more flexible ML methods. SRs labelling can be characterised as an ML classification task to identify combinations of *trajectors* (e.g., “ball”), *indicators* (“on”), and *landmarks* (“the ground”) [15]. More recent work leverages deep learning for this task, including convolutional neural networks for relation extraction [16].

A related task consists of detecting toponyms in text, usually as part of Named Entity Recognition (NER). A further step associates toponyms with spatial extensions, such as georeferenced points or polygons, to facilitate data integration and analysis — this process is known as geoparsing, geocoding, toponym resolution, or georeferencing. The integration of SR detection with NER has also been explored, estimating the spatial extent of expressions such as “North Milan” and “10 km from the French border” [17]. Recently, LLMs have begun to be evaluated for their effectiveness in NER for place detection and geoparsing. Initial research shows how GPT-based models can achieve high accuracy in multiple domains, including geography [18].

Toponyms exhibit strong temporal variation and require dedicated semantic resources to connect place names to appropriate spatial scopes. The World Historical Gazetteer (WHG)<sup>1</sup> gathers records from multiple sources to identify place names across temporal contexts,

such as Byzantium, Constantinople, and Istanbul, using a linked data approach.<sup>2</sup> Historical geoparsers must balance precision with historical sensitivity and domain-specific training [19]. For Latin, NER faces more challenges than for English, including orthographic and diachronic variation, as well as limited and sparse training data [20, 21]. To date, the majority of research and tools focus on contemporary languages, and no Latin evaluation exists for the extraction of SRs and geoparsing.

While the studies briefly reviewed in this section mark important progress in both Latin NLP and SR extraction, systematic evaluation of LLMs on spatial language understanding in Latin remains largely unexplored. Building on this foundation, our study investigates whether LLMs can interpret spatial constructions in Latin with a level of accuracy that approximates human linguistic analysis.

### 3. Research Questions and Evaluation Tasks

We examine whether LLMs can identify and interpret spatial constructions in Latin in ways that approximate human linguistic judgment. Specifically, we investigate three tasks that collectively test the models’ capacity to perform SR extraction and identification in Latin. This study is guided by the following research questions:

- RQ1:** To what extent can LLMs accurately identify preverbed motion verbs in Latin sentences?
- RQ2:** To what extent can LLMs detect place expressions that co-occur with preverbed motion verbs — regardless of their syntactic form — and classify them as indicating the Source, Goal, or Path?
- RQ3:** To what extent can LLMs correctly perform SR type disambiguation in Latin, especially in cases where the distinction between common nouns, proper nouns (toponyms), and adverbs is ambiguous?

These questions target key linguistic phenomena involved in spatial language understanding and test the applicability of LLMs to historical languages. Motion verbs are highly relevant for tasks involving spatial semantics and argument structure, particularly in Latin, where directional meaning is often distributed across both the verb and its preverb. Secondly, motion verbs frequently occur with locative or directional expressions (e.g., accusative or ablative prepositional phrases), providing rich ground for testing whether models can correctly associate verbs with SRs. Finally, the variability in motion verb semantics (e.g., goal-directed vs. manner-of-motion)

<sup>1</sup><https://whgazetteer.org>. Last accessed: 26 July 2025.

<sup>2</sup><https://whgazetteer.org/places/12345979/portal/>. Last accessed: 26 July 2025.

allows us to probe whether models distinguish different types of motion events. Preverbs play a central role in encoding directionality and spatial modification in Latin motion constructions. The distinction between proper and common nouns (*Roma* ‘Rome’ vs. *domus* ‘house’) is important from a cultural perspective to map how motion verbs relate to the geographical imaginary of the Roman world. Technically, it also provides more detail about the ability of LLMs to detect and interpret spatial references.

To operationalise our research questions, we define three corresponding annotation tasks:

1. **Motion Verb Identification** (RQ1): Determine whether a given Latin sentence contains a preverbed motion verb.
2. **SR Detection and Classification** (RQ2): Identify the presence of place expressions that co-occur with preverbed motion verbs and classify their semantic role in the motion event as *Source*, *Path*, or *Goal*, regardless of syntactic realization.
3. **SR Type Disambiguation** (RQ3): Perform SR type disambiguation with particular attention to expressions relevant to motion contexts, including disambiguation between common nouns, proper nouns (toponyms), and adverbs.

## 4. Corpus, Annotation, Dataset

### 4.1. The Usual Dilemma: Choosing a Representative Corpus for Latin

Given the fragmentary nature of the surviving material and the uneven transmission of texts across time, genre, and register, a fully representative corpus of Latin, as for historical languages in general, is ultimately unattainable [22]. Nevertheless, the Latin corpus used in this study is constructed specifically to address the limitations of existing resources and to meet the needs of historical corpus linguistics [23, 24]. Standard annotated corpora, such as the Latin Dependency Treebank (LDT) [25, 26], offer valuable syntactically annotated material but are limited in scope and uneven in their coverage. Many important authors — such as Plautus, Seneca, and Petronius — are entirely absent, and key texts like Caesar’s *De bello Gallico* and Virgil’s *Aeneid* are only partially included. To support quantitative and diachronic analysis, we constructed a custom corpus that is sensitive to linguistic diversity across time and genres. The corpus includes 16 Latin texts by 13 authors, and 265,707 tokens in total.<sup>3</sup> The corpus texts span from the 3rd century BCE to the 2nd century CE. This temporal range captures the major

phases of Latin’s development, across Early, Classical, and Late Latin [27]. Genre was a key consideration in corpus design. To avoid the so-called “God’s truth fallacy” [23] — the mistaken assumption that a single text type or genre can represent the full linguistic reality of a historical period — we included a range of genres that reflect different stylistic and communicative registers. The corpus contains texts from a wide range of genres: historiography, poetry, theatre, philosophy, novel, oratory.<sup>4</sup> This selection allows us to investigate genre-conditioned variation while also providing a broader basis for generalisations about Latin syntax. Texts were sourced primarily from the Perseus Digital Library<sup>5</sup> [29], except for Ennius’ *Annales*, accessed via PHI Latin Texts<sup>6</sup> [30].

Prose is more represented (61.7%) than poetry (38.3%), reflecting both textual availability and our aim to balance stylistic registers. Comedy and satire, often considered closer to spoken Latin, were included despite their underrepresentation in standard corpora. Inscriptions and epistolography were excluded due to limited data on preverbs. Text selection also accounted for varying author productivity, with prolific authors like Cicero and Seneca represented by more than one text, while preserving balance across genres.

### 4.2. Selecting Motion Verbs and Preverbs

The study requires a representative sample of motion verbs exhibiting diverse syntactic behaviour and frequently co-occurring with place expressions in Latin texts. We select eight verbal bases denoting different motion domains, and 16 preverbs. This results in a combinatorial space of 128 verb–preverb combinations (though not all are attested). The selection is based on the PRE-MOVE dataset (cf. Section 4.3), which provides gold-standard annotations for these verbs and preverbs, ensuring both linguistic coverage and empirical grounding. The verbal bases are: *eo* ‘go’, *venio* ‘go, come’ (all referring to generic motion), *fugio* ‘flee’, *gradior* ‘walk’, *curro* ‘run’, *volo* ‘fly’, *no* ‘swim’ (manner-of-motion verbs denoting specific types of movements along different media: ground, sky, water), and *navigo* ‘sail’ (motion by water via vehicle). These bases are selected to ensure coverage of different spatial event types and to test model performance across varying lexical, morphological, and syntactic profiles. Apart from the comitative preverb *cum*- ‘together’, denoting accompaniment, all preverbs possess an inherent spatial meaning. They can be categorised into four classes, based on the SR they inherently focus on:

- **Source-preverbs:** *ab*- ‘away, away from’, *de*-

<sup>3</sup>Since punctuation is not present in the original Latin texts, punctuation marks are excluded from the token count.

<sup>4</sup>Labels from [28].

<sup>5</sup><https://www.perseus.tufts.edu>. Last accessed: 26 July 2025.

<sup>6</sup><https://latin.packhum.org>. Last accessed: 26 July 2025.

**Table 1**

Overview of Latin Texts in the Corpus.

Author	Text	Century	Genre	Token Count
Ennius	Annales	3rd cent. BCE	Poetry, epic	1,194
Plautus	Amphitruo	3rd cent. BCE	Theatre, comedy	9,988
	Mostellaria	3rd cent. BCE	Theatre, comedy	9,988
Caesar	De bello Gallico 1-4	1st cent. BCE	Historiography	20,498
Cicero	In Catilinam 1-3	1st cent. BCE	Oratory	11,625
	De amicitia	1st cent. BCE	Philosophy, dialogue	9,471
Sallust	Bellum Catilinae	1st cent. BCE	Historiography	10,655
Livy	Ab Urbe condita 1-2	1st cent. BCE	Historiography	39,913
Virgil	Aeneid	1st cent. BCE	Poetry, epic	63,719
Propertius	Elegies 1.1-1.22	1st cent. BCE	Poetry, elegy	4,384
Horace	Satires	1st cent. BCE	Poetry, satire	7,048
Seneca	De ira	1st cent. CE	Philosophy, treatise	22,614
	Medea	1st cent. CE	Theatre, tragedy	5,639
Tacitus	Historiae 1	1st-2nd cent. CE	Historiography	11,852
Suetonius	Life of August	2nd cent. CE	Historiography, biography	13,915
Apuleius	Metamorphoses 1-5	2nd cent. CE	Novel	23,358

‘down from’, *ex*- ‘out, out of’;

- **Goal**-preverbs: *ad*- ‘to, towards’, *in*- ‘into’ (in contexts entailing motion), *intro*- ‘within, inside of’, *pro*- ‘forward’, *sub*- ‘under’ (in contexts entailing motion);
- **Path**-preverbs: *per*- ‘through’, *trans*- ‘across’;
- **Location**-preverbs: *circum*- ‘around’, *inter*- ‘between, among’;

### 4.3. Gold Standard

To create the gold standard for evaluation, we manually annotate occurrences of motion verb constructions in the Latin corpus described above. The annotation is carried out using the INCEPTION platform [31, 32, 33, 34]. INCEPTION’s user-friendly interface and extensible architecture proves essential for this study. All annotations are carried out by a single expert annotator (the first author). To verify task clarity, we conducted an Inter-Annotator Agreement (IAA) test on a random sample of 10 sentences, independently annotated by two additional historical linguists. The test yielded perfect agreement (IAA = 1.0), confirming that the task is sufficiently clear and unambiguous to justify relying on a single expert annotator for the full dataset. The annotation follows the guidelines described in [35]. Each sentence containing a preverbed motion verb is analysed to determine the presence of SRs, following a multi-layered annotation scheme (cf. Section 3):

1. **Motion Verb Identification (Task 1):** Identify whether the sentence contains a target motion verb.

2. **SR Detection and Classification (Task 2):** If a motion verb is present, determine whether it co-occurs with a SR. When a SR is present, classify its type as *Source*, *Goal*, or *Path*. Prepositions, case morphology, and preverb semantics are used to guide this decision, making the task unambiguous (e.g., *ex urbe* ‘from the city’ = *Source*; *in urbem* ‘to the city’ = *Goal*; *per urbem* ‘through the city’ = *Path*).
3. **SR Type Disambiguation (Task 3):** Annotate the SR type of spatial expressions, i.e. distinguish between proper nouns (e.g., *Roma* ‘Rome’), common nouns (e.g., *domus* ‘house’), and adverbs (e.g., *hinc* ‘from here’).

These annotations form part of the PREMOVE dataset [36], which also contains additional annotation layers as it is developed within the context of a broader research project [37].

## 5. Experimental Setup

### 5.1. Dataset and Models

**Dataset.** The experiments are conducted on the dataset described in Section 4.1, which consists of 1,483 Latin sentences. Since our focus is on spatial semantics, we filter out sentences that lacked SR annotations. The resulting dataset used for experimentation comprises 649 sentences (cf. Section 4.1).

SRs are unevenly distributed across the data: *Goal* relations appears in 68.4% of the occurrences, while *Source* and *Path* occur in only 19.6% and 12.0%, respectively. This is in line with the Goal-over-Source principle, according



to which languages express the Goal more frequently because it plays a more central role in the conceptualisation of motion events, making the event appear complete and cognitively salient [38]. Moreover, Goal-oriented motion is often perceived as more intentional and purposeful, while Source expressions suggest less human agency [39, 40]. To mitigate this imbalance and ensure a fairer evaluation of model behaviour across relation types, we also construct three distinct, balanced subsets of the dataset (cf. Sections 5.2, 6.1). Each subset isolates a single SR and balances positive and negative examples for that relation. The resulting subset sizes are as follows:

- *Goal* subset: 394 sentences
- *Source* subset: 256 sentences
- *Path* subset: 150 sentences

The total number of sentences across the subsets exceeds the total number of sentences in the dataset (649), as individual sentences can encode more than one type of SR.

**Models.** We choose two open-weight LLMs (Mistral and Llama) and one proprietary model (OpenAI’s GPT) to compare performance across different architectures and accessibility levels. Open-weight models are LLMs whose trained parameters (weights) are publicly released, allowing researchers and developers to run, fine-tune, and deploy them independently. In contrast, proprietary models like GPT are closed-source and accessible only via API or controlled platforms. We use Mistral-7B-Instruct-v0.1, Meta’s Llama-3.2-3B-Instruct, and OpenAI’s GPT-4. We did not perform any fine-tuning on the open-weight models. We used the pre-trained versions of the models as provided on Hugging Face, without further adaptation or training. The prompts are described in section 5.2. In few-shot settings, manually annotated examples from our corpus (section 5.1) are randomly added to the prompts. We evaluate model performance under zero-shot, one-shot, and five-shot conditions. In the zero-shot setting, the model is given only the task instruction without any examples. In the one-shot and five-shot settings, we include respectively one or five manually annotated examples from our corpus (Section 5.1) to the prompt. These examples are selected at random and aim to reflect typical structures found in the corpus. This design allows us to test how much model performance improves with limited supervision. We intentionally selected models that were not specifically fine-tuned for Latin to ensure a fair comparison across general-purpose architectures. Our aim is to evaluate how LLMs trained primarily on large multilingual or general corpora perform out of the box on Latin. All experiments are performed locally, with a machine comprising 8 CPU cores and 8 GB of RAM. The

experiments are implemented in Python 3.9.13, using the PyTorch and Hugging Face Transformers libraries. To run the Mistral and Llama models, we use an A100 GPU (purchased) and a T4 GPU via Google Colab. Our code is freely available on GitHub<sup>7</sup>.

## 5.2. Prompt Engineering

**Task 1.** Task 1 consists in identifying all inflected forms of a given Latin verb in one or more input sentences. The core prompt includes the verb lemma, a linguistic framing, and clear task constraints. Importantly, the input to the models consists of individual sentences rather than full passages. These are extracted directly from PREMOVE (cf. 4.3), in order to isolate sentence-level syntactic and semantic behaviour and reduce computational cost during inference. The prompt is given below:

This is a task of Latin linguistics. Given the following Latin sentences, identify all the forms of the verb ‘{verb}’ across all sentences. Note that verbs may occur more than once and in more than one sentence, so PROVIDE ALL THE FORMS YOU DETECT.

This task is designed to evaluate models’ ability to identify all inflected forms of a given Latin verb, not to test their recognition of motion semantics per se. While the target lemmas are motion verbs, they are explicitly provided in the prompt to ensure clarity and task focus. This approach also avoids ambiguity in cases where multiple motion verbs may occur in the same sentence, some of which fall outside the scope of annotation. Testing the models’ ability to detect motion verbs without guidance would indeed be a valuable direction for future work, but lies beyond the controlled objectives of this task.

**Task 2.** The base prompt includes a task explanation and binary labels for each SR. A representative zero-shot version is shown below:

This is a task of Latin linguistics. Given the following Latin sentence, identify all the forms of the verb ‘{verb}’. Then, additionally answer: Does the sentence contain a **source expression**? True or False; Does the sentence contain a **goal expression**? True or False; Does the sentence contain a **path expression**? True or False

**Task 3.** This task consists of classifying a spatial token linked to a motion verb as either an adverb, a common noun, or a proper noun. Initial prompts list classification labels and provide a target token. As early outputs show

<sup>7</sup><https://github.com/farina-andrea/latin-spatial-relations-llms>. Last accessed: 26 July 2025.

that the models confuses proper nouns and their association with the target verbs (cf. 6.2), we implement the prompt to increase specificity:

This is a task of Latin linguistics. Given the Latin sentence below, and focusing specifically on the verb ‘{verb}’, identify the noun or adverb in the sentence governed by ‘{verb}’ and expressing the spatial relation ‘{relation type}’ (Source, Goal, or Path). Classify this token as one of the following:

- An adverb (e.g., ‘hinc’)
- A common noun referring to a place (e.g., ‘domus’, ‘forum’)
- A proper noun referring to a place name (e.g., ‘Roma’, ‘Carthago’).

Sentence: ‘{sentence}’  
 Answer with exactly two lines, no extra text:  
 Token: <token>  
 adverb | common noun | proper noun

## 6. Results

### 6.1. Quantitative Evaluation

**Task 1.** The results of Task 1 are given in Table 2.

Model	Setting	Precision	Recall	F1-score
<b>Mistral-7B</b>	Zero-shot	0.09	0.23	0.13
	One-shot	0.08	0.19	0.11
	Five-shots	0.04	0.10	0.06
<b>Llama-3.2B</b>	Zero-shot	0.33	0.12	0.05
	One-shot	0.03	0.10	0.05
	Five-shots	0.01	0.06	0.02
<b>GPT-4</b>	Zero-shot	<b>0.95</b>	<b>0.98</b>	<b>0.96</b>
	One-shot	<b>0.91</b>	<b>0.98</b>	<b>0.94</b>
	Five-shots	<b>0.85</b>	<b>0.97</b>	<b>0.91</b>

**Table 2**

Task 1. Model performances across different shot settings on all 649 sentences. Highest scores per shot setting are highlighted in bold.

GPT-4 strongly outperforms both Llama-3.2-3B-Instruct and Mistral-7B-Instruct on all 649 sentences. Its precision, recall, and F1-scores remain consistently high across all prompt settings, indicating robust zero- and few-shot generalisation. The open-weight models perform poorly and also degrade in performance as shots increase, suggesting that additional examples may introduce noise rather than aid in disambiguation.

**Task 2.** Results for Task 2 on all 649 sentences are shown in Table 3. Performance varies significantly between GPT on the one hand, and Mistral and Llama on

the other. Mistral and Llama achieve near-identical results across almost all task conditions. This suggests that both are relying on similar simplistic prediction strategies, as seen in the uniformly perfect (1.00) or null (0.00) recall, and very low precision values across categories. The deceptively strong F1 scores (0.82) for *Goal* likely reflect an overgeneralisation strategy: the models tend to label nearly all inputs as positive, which inflates recall and leads to misleadingly moderate F1 scores, especially when the positive class *Goal* is frequent (cf. Section 5.1).

GPT-4 demonstrates a more balanced performance, with better alignment between precision and recall. It shows consistently strong results for *Source* and *Path*, with F1 scores stable across prompting conditions. In contrast, *Goal* shows unexpectedly low performance in one- and three-shot settings, likely due to example sampling variability — none of the randomly selected few-shot prompts included a *Goal* instance, which may have misled the model (cf. 6.1 below).

**Literal motion.** We evaluate Task 2 on a subset annotated exclusively for literal motion verbs, focusing on physical movement and excluding figurative uses. This dataset includes *Source*, *Goal*, and *Path*, but is unbalanced across SRs. Mistral, Llama, and GPT are tested under zero-, one-, and six-shot settings, with the latter including one positive and one negative example per relation.

As shown in Table 4, Llama’s and Mistral’s performances remain identical and unreliable, marked by low precision and F1-scores, particularly for *Path*, which is never correctly identified. While slight improvements can be seen for *Source* under six-shot prompting (F1 = 0.67 for Mistral), overall performance remains inconsistent and largely unchanged compared to the mixed dataset (cf. Table 3). For this reason, both models were excluded from further experiments on Task 2 and the entirety of Task 3, as it builds upon SR classification performed in Task 2.

GPT-4 performs considerably better. The *Goal* relation continues to be the most robust, reaching an F1-score of 0.83 in the six-shot setting. Performance for *Source* and *Path*, however, remains more variable and consistently lower, with best F1-scores of 0.61 and 0.54 respectively. This suggests that even in literal motion contexts, *Source* and *Path* relations are harder to detect reliably — possibly because *Goal* is more commonly and overtly expressed in motion events, giving the model stronger and more consistent lexical or structural cues to rely on.

**Controlled SRs.** To check whether the imbalance between *Goal*, *Source*, and *Path* is contributing to GPT-4’s lower performance on the *Goal* class, we test the model on a three separate subsets of the data. The Task was split into three separate sub-tasks, each focused on a sin-

Model	Metric	Source			Goal			Path		
		0-shot	1-shot	3-shots	0-shot	1-shot	3-shots	0-shot	1-shot	3-shots
<i>Mistral-7B</i>	Precision	0.19	0.00	0.19	0.69	0.69	0.69	0.12	0.00	0.00
	Recall	1.00	0.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00
	F1-score	0.33	0.00	0.33	<b>0.82</b>	<b>0.82</b>	<b>0.82</b>	0.21	0.00	0.00
<i>Llama-3.2B</i>	Precision	0.22	0.00	0.19	0.69	0.69	0.69	0.12	0.00	0.00
	Recall	1.00	0.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00
	F1-score	0.36	0.00	0.33	<b>0.82</b>	<b>0.82</b>	<b>0.82</b>	0.21	0.00	0.00
<i>GPT-4</i>	Precision	0.79	0.80	0.80	0.75	0.30	0.30	0.69	0.88	0.88
	Recall	0.85	0.80	0.80	0.76	0.30	0.30	0.69	0.88	0.88
	F1-score	<b>0.82</b>	<b>0.80</b>	<b>0.80</b>	0.75	0.30	0.30	<b>0.69</b>	<b>0.88</b>	<b>0.88</b>

**Table 3**

Task 2. Model performance across tasks and shot settings. Highest F1-score values per shot setting are highlighted in bold.

Model	Metric	Source			Goal			Path		
		0-shot	1-shot	6-shots	0-shot	1-shot	6-shots	0-shot	1-shot	6-shots
<i>Mistral-7B</i>	Precision	0.26	0.26	0.50	0.65	0.00	0.50	0.14	0.00	0.00
	Recall	1.00	1.00	1.00	1.00	0.00	1.00	1.00	0.00	0.00
	F1-score	0.41	0.41	0.67	0.79	0.00	0.67	0.24	0.00	0.00
<i>Llama-3.2B</i>	Precision	0.26	0.26	0.26	0.65	0.00	0.00	0.14	0.00	0.00
	Recall	1.00	1.00	1.00	1.00	0.00	0.00	1.00	0.00	0.00
	F1-score	0.41	0.41	0.41	0.79	0.00	0.00	0.24	0.00	0.00
<i>GPT-4</i>	Precision	0.59	0.37	0.40	0.71	0.70	0.73	0.35	0.22	0.41
	Recall	0.62	0.87	0.78	0.98	1.00	0.96	0.84	0.95	0.78
	F1-score	<b>0.61</b>	<b>0.52</b>	<b>0.53</b>	<b>0.82</b>	<b>0.82</b>	<b>0.83</b>	<b>0.49</b>	<b>0.35</b>	<b>0.54</b>

**Table 4**

Task 2. SR classification results on literal motion verb subset (unbalanced). Highest F1-score values per shot setting are highlighted in bold.

gle SR, with corresponding dataset subsets (cf. 5.1). We restrict this analysis to GPT-4, as it seems to be the only model to produce SR predictions that are not effectively random (cf. 6.1 above).

	Relation	Setting	Precision	Recall	F1-score
<i>GPT-4</i>	Source	Zero-shot	0.85	0.57	0.68
		One-shot	0.70	0.85	<b>0.77</b>
		Two-shots	0.64	0.79	0.71
	Goal	Zero-shot	0.58	0.95	<b>0.72</b>
		One-shot	0.57	0.97	<b>0.72</b>
		Two-shots	0.57	0.94	0.71
	Path	Zero-shot	0.70	0.84	0.76
		One-shot	0.59	0.92	0.72
		Two-shots	0.70	0.91	<b>0.79</b>

**Table 5**

Task 2 (GPT-4). SR classification results after the dataset splitting (balanced). Highest F1-score per shot setting is highlighted in bold.

The results on the split dataset show more stable performance across relations (Table 5). For *Source*, the best F1 is 0.77 with one-shot prompting; for *Goal*, recall remains high (0.95) with moderate precision (0.57); and for *Path*, the best F1 (0.79) is achieved with two-shot prompting.

**Task 3.** Table 6 summarises the performance of GPT-4 in classifying parts of speech in sentences related to motion. We exclude the other two models because of their poor performance on the previous two tasks, on which Task 3 relies on (cf. 6.1). Zero- and one-shot prompting achieve the highest F1 score for common nouns, followed by adverbs. For proper nouns, recall is high, while precision is low. This discrepancy between high recall and low precision for proper nouns suggests that while GPT-4 reliably detects their presence, it often overpredicts and misattributes them within the sentence structure (cf. 6.2).

	Setting	SR Type	Precision	Recall	F1-score
GPT-4	Zero-shot	adverb	0.90	0.68	0.77
		common noun	0.91	0.83	<b>0.87</b>
		proper noun	0.47	0.92	0.62
	One-shot*	adverb	0.87	0.76	0.81
		common noun	0.92	0.79	<b>0.85</b>
		proper noun	0.42	0.83	0.55

**Table 6**

Task 3 (GPT-4). SR type disambiguation: adverbs, common nouns, proper nouns, under zero-shot and one-shot prompting. The one-shot (\*) is given on a proper noun instance. Highest F1-score per shot setting is highlighted in bold.

## 6.2. Qualitative Evaluation

**Task 1.** Mistral and Llama show high confusion for verb identification, with an overgeneration of predictions that do not include the correct value. They often include forms that are morphologically or semantically related to the correct one (e.g., *conveniens* instead of *conveniunt*, *subeo* instead of *subit*), though in some cases the forms are entirely unrelated (e.g., *advena*, *adgredior*, *excolui* instead of *aggressus*). A qualitative inspection of the (few) mismatches for GPT-4 reveals that the model occasionally produces multiple verb forms within its output for a single sentence. Examples include cases such as *transierat*, *traduxisse* and *evolo*, *evigila*, where multiple words are listed. In these cases, the words are not different inflected forms of the same lemma, but rather distinct verbs or nouns. Nonetheless, the correct verb form is always present among these outputs (*evolo*, *transierat*), indicating that these are instances of overgeneration or model uncertainty. This behaviour persists despite prompt engineering efforts to constrain the output format, suggesting a tendency of the model to hedge its predictions in ambiguous cases. Interestingly, increasing the number of shots does not improve performance, suggesting that additional examples for verb identification may introduce noise or ambiguity rather than reinforcing the model’s task-specific behaviour [41].

**Task 2.** Mistral’s and Llama’s predictions show that the models randomly assign a positive or negative value to a specific SR. For *Goal*, F1 is high as *Goal* is mostly present in the examples, due to the Goal-over-Source principle [38]. GPT-4 has a different performance depending on the relation type and prompt format. For the *Goal*, performance drastically drops under the one-shot and three-shot settings with an unbalanced dataset. In these cases, the prompt examples possibly do not include a representative positive instance of *Goal*, causing a steep drop in its recognition. Balancing the dataset improves consistency across SRs, but qualitative errors remain. For instance, the model often confuses *Source* and *Path* when

the contextual cues are subtle or ambiguous. On the subset limited to literal motion verbs, the model demonstrates relatively strong recognition of *Goal*, but struggles more with *Source* and *Path*.

**Task 3.** The SR type disambiguation task (GPT-4 only) displays different levels of the models’ accuracy across parts of speech. While common nouns are identified with high confidence and accuracy, proper nouns pose some challenges, as reflected in lower precision and F1 scores. This finding reinforces the need to treat them separately. Even after prompt engineering (which yielded a slight performance improvement), a consistent pattern of error persists: whenever a proper noun appears in the sentence but is not governed by the target motion verb, the model still annotates it as the relevant argument. Although this is technically a correct identification of a proper noun, it is incorrect in the context of the task. For instance, in the sentence:

*Nam, ut scis optime, secundum quaestum Macedoniam profectus, [...] per transitum spectaculum obiturus, in quadam avia et lacunosa convalli a vastissimis latronibus obsessus atque omnibus privatus tandem evado*

‘So, as you well know, I had set out for Macedonia to earn a living. On the way, planning to take in some sights, I was ambushed in a remote and marshy valley by a band of enormous robbers. Stripped of everything, I finally managed to escape.’ (Apul.Met.1.7)

the model correctly identifies *Macedoniam* as a proper noun but incorrectly links it to the motion verb *obeo* (in the form *obiturus*), instead of recognising that it belongs to a different motion verb (*profectus*, from *proficiscor*), which is not among the verbs considered for annotation. This may suggest that in the context of proper nouns, the model relies heavily on their salience and tends to overlook verb-governance constraints. In other words, the model appears to prioritise SR type recognition and semantic prominence over syntactic dependencies when proper nouns are involved. In other cases, the model occasionally misclassifies common nouns as proper nouns. Examples include words like *finis* ‘borders’ or *urbs* ‘city’, which are common nouns, but are mistakenly labeled as proper nouns.

## 7. Discussion and Conclusion

This study evaluates LLMs across three interconnected tasks in Latin linguistic analysis: motion verb identifica-



tion, SR classification, and SR type disambiguation. Our results are encouraging, but they also highlight the significant differences in performance between models — particularly the stark contrast between GPT-4 and open-weight models such as Llama and Mistral.

GPT-4 achieves high performance across all tasks, already in zero-shot settings. This is likely due to the substantial presence of Latin data in its pretraining corpus. While the precise contents of GPT-4’s training data remain undisclosed, estimates based on GPT-3 suggest at least 339 million Latin tokens were included [42], and GPT-4 was trained on significantly more data. This makes it plausible that GPT-4 has substantial exposure to Latin, unlike models such as Llama and Mistral, which likely lack such training data and perform accordingly worse — often failing completely in zero-shot settings.

For preverbed motion verb identification, GPT-4 achieves strong performance, particularly under zero-shot settings [41]. SR classification exposes challenges due to data imbalance, with *Goal* relations dominating the dataset. Creating balanced subsets helps obtain more reliable and interpretable results. SR type disambiguation proves the most difficult task, with the model frequently misclassifying proper nouns and failing to correctly link them to relevant motion verbs. This highlights a gap in the way the models can use contextual reasoning to disambiguate entities. This may be mitigated by expanding the length of the input text so to offer more context to the models. Error analysis suggests that the model’s dependence on lexical familiarity and world knowledge, which may not perfectly align with classical contexts, limits its accuracy.

These findings demonstrate that while LLMs show promising semantic understanding in Latin, syntactic and contextual challenges persist. Balancing datasets and employing few-shot prompting improve performance, but do not fully resolve issues related to ambiguity and entity linking.

Future work should focus on domain-specific fine-tuning with classical corpora, possibly integrating external knowledge sources to enhance disambiguation and semantic grounding. This combined approach can better support the complex linguistic features of Latin and ultimately advance computational tools for classical language research. In parallel, similar experiments should be conducted on other languages to assess how especially open-weight models handle spatial relations in languages for which they have broader coverage. Such comparisons can clarify whether the poor performance observed in Latin stems from language-specific limitations or from more general architectural and training differences. Additionally, future studies could isolate prose texts to control for syntactic regularity, as poetic language often introduces greater structural variability and long-distance dependencies that may challenge model

performance.

Our study — the first on LLMs’ SR recognition in historical languages — clarifies their performance and limits in this area. It lays the groundwork for more specialised computational methods in Computational Humanities and Historical Linguistics, with potential applications to other historical languages where preverbs are vastly employed, such as Ancient Greek [43].

## Author contributions

AF was responsible for conceptualisation, methodology, formal analysis, software implementation (including all code used for analysis), and manual annotation of the dataset; he wrote the original draft for Sections 1, 3-7, and edited the final manuscript. AB and BMcG contributed to the conceptualisation and methodology of the project, drafted Section 2, and participated in review, editing, and supervision of the research.

## References

- [1] B. Levin, English verb classes and alternation, A preliminary investigation, Chicago: The University of Chicago Press, 1993.
- [2] L. Talmy, Toward a Cognitive Semantics. Vol. 1: Concept Structuring Systems, Cambridge (MA): MIT Press, 2000.
- [3] G. Lakoff, Women, Fire and Dangerous Things. What Categories Reveal about the Mind., Chicago: The University of Chicago Press, 1987.
- [4] G. Lakoff, M. Johnson, Metaphors we live by, Chicago: The University of Chicago Press, 1980.
- [5] R. Sprugnoli, F. Iurescia, M. Passarotti, Overview of the evalatin 2024 evaluation campaign, in: Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA), Language Resources and Evaluation Conference (LREC 2024), 2024, pp. 190–197.
- [6] D. Bamman, P. J. Burns, Latin bert: A contextual language model for classical philology, arXiv preprint arXiv:2009.10053 (2020). URL: <https://arxiv.org/abs/2009.10053>.
- [7] P. Lendvai, C. Wick, Finetuning latin bert for word sense disambiguation on the thesaurus linguae latinae, in: Proceedings of the Workshop on Cognitive Aspects of the Lexicon, Association for Computational Linguistics, Taipei, Taiwan, 2022, pp. 37–41.
- [8] I. Ghinassi, S. Tedeschi, P. Marongiu, R. Navigli, B. McGillivray, Language pivoting from parallel corpora for word sense disambiguation of historical languages: A case study on latin, in: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources

- and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 10073–10084.
- [9] M. Beersmans, E. de Graaf, T. V. de Cruys, M. Fantoli, Training and evaluation of named entity recognition models for classical latin, in: A. Anderson, S. Gordin, S. Klein, B. Li, Y. Liu, M. C. Passarotti (Eds.), *Proceedings of the Ancient Language Processing Workshop (ALP 2023) associated with The 14th International Conference on Recent Advances in Natural Language Processing (RANLP 2023)*, 2023.
- [10] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, W. Ye, Y. Zhang, Y. Chang, P. S. Yu, Q. Yang, X. Xie, A survey on evaluation of large language models, *ACM Trans. Intell. Syst. Technol.* 15 (2024). URL: <https://doi.org/10.1145/3641289>. doi:10.1145/3641289.
- [11] Q. Xue, Unlocking the potential: A comprehensive exploration of large language models in natural language processing, *Applied and Computational Engineering* 57 (2024) 247–252. URL: <https://doi.org/10.54254/2755-2721/57/20241341>. doi:10.54254/2755-2721/57/20241341.
- [12] Z. Wang, W. Zhong, Y. Wang, Q. Zhu, F. Mi, B. Wang, L. Shang, X. Jiang, Q. Liu, Data management for training large language models: A survey, 2024. URL: <https://arxiv.org/abs/2312.01700>. arXiv:2312.01700.
- [13] I. Vieira, W. Allred, S. Lankford, S. Castilho, A. Way, How much data is enough data? fine-tuning large language models for in-house translation: Performance evaluation across multiple dataset sizes, in: R. Knowles, A. Eriguchi, S. Goel (Eds.), *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, Association for Machine Translation in the Americas, Chicago, USA, 2024, pp. 236–249. URL: <https://aclanthology.org/2024.amta-research.20/>.
- [14] M. Volk, D. P. Fischer, L. Fischer, P. Scheurer, P. B. Ströbel, Llm-based machine translation and summarization for latin, in: *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024, ELRA and ICCL, Torino, Italia, 2024*, pp. 122–128.
- [15] P. Kordjamshidi, M. Van Otterlo, M.-F. Moens, Spatial role labeling: Towards extraction of spatial relations from natural language, *ACM Transactions on Speech and Language Processing (TSLP)* 8 (2011) 1–36.
- [16] Q. Qiu, Z. Xie, K. Ma, Z. Chen, L. Tao, Spatially oriented convolutional neural network for spatial relation extraction from natural language texts, *Transactions in GIS* 26 (2022) 839–866.
- [17] M. A. Syed, E. Arsevska, M. Roche, M. Teisseire, Geospatre: extraction and geocoding of spatial relation entities in textual documents, *Cartography and Geographic Information Science* 52 (2025) 221–236.
- [18] S. Wang, X. Sun, X. Li, R. Ouyang, F. Wu, T. Zhang, J. Li, G. Wang, GPT-NER: Named Entity Recognition via Large Language Models, *arXiv preprint arXiv:2304.10428* (2023).
- [19] J. Kenyon, J. W. Karl, B. Godfrey, Evaluation of placename geoparsers, *Journal of Map & Geography Libraries* 19 (2023) 185–197.
- [20] A. Erdmann, C. Brown, B. Joseph, M. Janse, P. Ajaka, M. Elsner, M.-C. de Marneffe, Challenges and solutions for Latin named entity recognition, in: *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, 2016, pp. 85–93.
- [21] M. Beersmans, E. de Graaf, T. Van de Cruys, M. Fantoli, Training and evaluation of named entity recognition models for classical Latin, in: *Proceedings of the Ancient Language Processing Workshop*, 2023, pp. 1–12.
- [22] T. McEnery, A. Wilson, *Corpus Linguistics. An Introduction*. Second edition, Edinburgh: Edinburgh University Press, 2001.
- [23] M. Rissanen, Three problems connected with the use of diachronic corpora, *ICAME Journal* 13 (1989) 16–19.
- [24] G. B. Jensen, B. McGillivray, *Quantitative Historical Linguistics. A Corpus framework*, Oxford University Press, Oxford, 2017.
- [25] D. Bamman, G. Crane, The Latin Dependency Treebank in a cultural heritage digital library, *Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007)*. Prague (Czech Republic) (2007) 33–40.
- [26] D. Bamman, G. Crane, The Ancient Greek and Latin Dependency Treebanks, in: *Language Technology for Cultural Heritage*, Springer, Berlin/Heidelberg, 2011, pp. 79–98.
- [27] P. Cuzzolin, G. V. M. Haverling, Syntax, sociolinguistics, and literary genres, in: P. Baldi, P. Cuzzolin (Eds.), *New perspectives on historical Latin syntax*, 2009, pp. 16–63.
- [28] E. Biagetti, C. Zanchi, W. M. Short, Toward the creation of WordNets for ancient Indo-European languages, in: *Proceedings of the 11th Global Wordnet Conference, University of South Africa (UNISA)*, volume 13, 2021, pp. 258–266.
- [29] G. Crane, Building a Digital Library: The Perseus Project as a Case Study in the Humanities, in: *DL '96: Proceedings of the First ACM International Conference on Digital Libraries*, 1996, pp. 3–10.
- [30] P. H. Institute, Classical latin texts. a resource prepared by the packard humanities institute (phi),

2015.

- [31] J.-C. Klie, INCEpTION: Interactive Machine-assisted Annotation, in: Proceedings of the First Biennial Conference on Design of Experimental Search & Information Retrieval Systems (DESIRES), Bertinoro, Italy, 2018.
- [32] B. Boullosa, R. E. de Castilho, N. Kumar, J.-C. Klie, I. Gurevych, Integrating knowledge-supported search into the inception annotation platform, Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2018) 127–132.
- [33] R. E. de Castilho, J.-C. Klie, N. Kumar, B. Boullosa, I. Gurevych, Inception - corpus-based data science from scratch, Digital Infrastructures for Research (DI4R) 2018, 9-11 October 2018, Lisbon, Portugal (2018a) 1. URL: <https://inception-project.github.io/publications/DI4R-2018.pdf>.
- [34] R. E. de Castilho, J.-C. Klie, N. Kumar, B. Boullosa, I. Gurevych, Linking text and knowledge using the inception annotation platform, Proceedings of the 14th eScience IEEE International Conference, Amsterdam, Netherlands (2018b) 1. URL: <https://inception-project.github.io/publications/ESCIENCE-2018.pdf>.
- [35] A. Farina, Guidelines for a linguistic annotation of preverbed verbs of motion, Figshare (2024). URL: <https://doi.org/10.18742/25055573>.
- [36] A. Farina, PREMOVE – a diachronic dataset of ancient greek and latin annotated PREverbed MOtion Verbs, Oxford Text Archive (2025). URL: <http://hdl.handle.net/20.500.14106/2579>.
- [37] A. Farina, The differences in Ancient Greek and Latin motion verbs as a way to understand the conceptualisation of reality in the two cultures, UK Research and Innovation (ref. number: 2749398), 2022-2026.
- [38] Y. Ikegami, ‘source’ vs. ‘goal’: A case of linguistic dissymmetry, in: R. Driven, G. Radden (Eds.), Concepts of Case, Narr., Tübingen, 1987, pp. 122–146.
- [39] F. Ungerer, H.-J. Schmidt, An Introduction to Cognitive Linguistics, London: Longman, 1996.
- [40] R. Dirven, M. Verspoor, Cognitive Exploration of Language and Linguistics, Amsterdam/Philadelphia: John Benjamins, 2004.
- [41] B. McGillivray, A. Farina, Are large language models able to grasp latin semantics? a study on motion verbs, International Colloquium on Latin Linguistics 2025. 9-13 June, Udine (Italy) (2025).
- [42] P. J. Burns, Research recap: How much latin does chatgpt "know"?, Blogpost at NYU ISAW (2023). URL: <https://isaw.nyu.edu/library/blog/research-recap-how-much-latin-does-chatgpt-know>.
- [43] A. Farina, Aquamotion Verbs in Ancient Greek: A Study on pléō and Its Compounds, University of

Pavia: MA Thesis, 2021.