Exploring Latin WordNet synset annotation with LLMs

Daniela Santoro[1], Beatrice Marchesi[1], Silvia Zampetta[1], Erica Biagetti[2], Claudia Roberta Combei[2], Stefano Rocchi[2], Tullio Facchinetti[2], Chiara Zanchi[2]
[1]Università degli Studi di Pavia
{daniela.santoro01, beatrice.marchesi03, silvia.zampetta01}@universitadipavia.it
[2]Università degli Studi di Pavia
{erica.biagetti, claudiaroberta.combei, stefano.rocchi,
tullio.facchinetti, chiara.zanchi01}@unipv.it

Eleonora Litta[3], Riccardo Ginevra[3]
[3]Università Cattolica del Sacro Cuore
{eleonoramaria.litta, riccardo.ginevra}@unicatt.it

Marco Del Tredici
Cohere, Spain
marcodeltredici@gmail.com

Abstract

This study explores the application of Large Language Models to populate synsets in the Latin WordNet, keeping a human-in-the-loop approach. We compare zero-shot, few-shot, and fine-tuning methods against an English baseline. Quantitative analysis reveals significant improvements from zero-shot to fine-tuned approaches, with the latter outperforming the baseline. Qualitative assessment indicates better performance with verbs and polysemous lemmas. While results are encouraging, human oversight remains crucial for accuracy. Future research could focus on improving performance across different parts of speech and degrees of polysemy, potentially incorporating etymological information or cross-linguistic data.

1 Introduction

The paper explores the use of Large Language Models (LLMs) to populate synsets of the Latin WordNet (LWN) and to evaluate the extent to which these models can contribute to this task. WordNets are lexical databases that organize word meanings in a network. The original WordNet was designed for English (Miller et al., 1990) as a psycholinguistic project. Over time, it lost its psycholinguistic focus, and shifted toward computational lexical semantics, leading to the development of similar databases for other languages, including ancient ones such as Latin, Ancient Greek, Sanskrit, and Old English (Minozzi, 2009; Bizzoni et al., 2014; Hellwig, 2017; Khan et al., 2022).

The building blocks of WordNet architecture are synsets, i.e. sets of cognitive synonyms accompanied by a brief definition and an ID-number. For instance, Latin nouns such as *absentia*, *carentia*, *deliquio*, *deliquium*, *desiderium*, *defectus*, *egestas*, etc. belong to the synset n#14472871 'the state of needing something that is absent or unavailable', meaning that they are partly synonymous. Furthermore, lemmas can be assigned to multiple synsets, which indicates polysemy: this is the case of Latin *absentia* which, besides belonging to synset n#14472871 above, is also assigned to the synsets n#13984260 'the state of being absent' and n#01236910 'failure to be present'.

The Latin WordNet (LWN) was first developed in 2004 following the Expand Method (Vossen, 2002), automatically translating English and Italian data from the MultiWordNet (Bentivogli et al., 2002) into Latin through the help of bilingual dictionaries. The resulting database contained 9,378 lemmas and 8,973 synsets. However, this approach led to an over-reliance on modern English and Italian, resulting in some anachronistic and inaccurate senses, particularly in the context of technical terminology (Minozzi, 2017). Later on, Franzini et al. (2019) proposed to refine the Latin WordNet by manually removing the modern terms and adding the missing senses.[1]

[1]At Exeter University, the LWN was further expanded to 70,000 lemmas using a gloss-ranking method to assign synsets (Exeter University, 2023). This method assigns greater weight to the translation equivalents that occur more frequently across glosses in the reference dictionaries, thus reducing the impact

A further effort to clean and expand on the original LWN was started and is currently still under way in the context of the LiLa project (Passarotti et al., 2019; Mambrini et al., 2021), which consists in the construction of a Knowledge Base of interlinked resources for Latin using Linked Open Data standards. The annotated and cleaned portion of the LWN currently amounts to 18,227 synsets associated to 10,449 lemmas. The work on the Latin WordNet continues in the framework of the project *Linking WordNets for Ancient Indo-European Languages*, whose aim is to extend and harmonize three WordNets for Latin, Ancient Greek and Sanskrit (Biagetti et al., 2021).

Although several methods for automatically populating synsets have been tested, the results typically required manual evaluation. The first such method exploits morphosyntactically annotated corpora to learn syntactic patterns for automatic hypernym discovery (Snow et al. 2005). Another method uses parallel corpora, by inducing sense clusters in new languages using multilingual semantic spaces (Apidianaki & Sagot 2014). Finally, models of distributional semantics have been used to automatically identify relations missing in the WordNets (on word embeddings for Ancient Greek, see Singh et al. 2021 with references; on Sanskrit, Sandhan et al. 2021; on Latin, Mehler et al. 2020). Since manually populating all synsets is a very time-consuming process, this work aims to speed this task up by developing a human-in-the-loop pipeline aided by LLMs.

Our experiment is based on Mistral-7B (Mistral AI, 2023), which was selected for its optimal balance between performance and efficiency. The architectural features of Mistral-7B, built upon the original Transformer architecture (Vaswani et al., 2017), enable high performance with limited computational resources (Ainslie et al., 2023; Touvron et al., 2023). The demonstrated adaptability of the model, achieved through efficient fine-tuning techniques like Low-Rank Adaptation (LoRA) (Hu et al., 2021a; Zhang et al., 2024), along with its multilingual capabilities (Jiang et al., 2023), provided a solid foundation for adaptation to this task.

This paper is organized as follows. In Section 2 we describe our data and methodologies. In particular, in Section 2.1 we present the dataset used in our experiment; in Section 2.2 we discuss the zero-shot experiment, followed by the few-shot experiment in Section 2.3. The final phase of our experiment, which involved fine-tuning using the LoRA technique, is detailed in Section 2.4. In Section 3, we report the results of the experiments, providing both a quantitative (Section 3.1) and a qualitative (Section 3.2) analysis. Section 4 contains the conclusions.

2 Data and methodologies

This section outlines the data and methodologies employed in applying LLMs to automatically enrich LWN synsets through Natural Language Generation (NLG). Our experiment progressed through three methodological phases of increasing complexity:

1. Implementation of zero-shot (ZS), through the application of prompt tuning techniques on a smaller batch of lemmas, and few-shot learning (FS).

2. Development of a validation approach in English, to establish a methodological baseline (EB = English baseline).

3. Fine-tuning (FT), optimizing the model for populating the synsets of the LWN.

This progression enabled a systematic evaluation of the effectiveness of different model adaptation strategies, analyzing the contribution of each approach in improving performance in the automatic generation of Latin synsets.

Sections 2.1-2.4 detail the composition of the Latin dataset obtained from the LiLa LWN (Mambrini et al., 2021), focusing on the current state of the available training data, the selection criteria for the testing dataset, and the development of our experiment.

2.1 Datasets

The data used in our experiments were entirely extracted from the LWN. Our testing dataset was constructed by selecting 80 synsets, divided into two main categories to ensure a balanced and representative evaluation:

- 40 relatively well-populated synsets, each containing 15 chiefly polysemous lemmas, hence labelled as "polysemy dataset".

of outliers. The ranking method produced better results than those achieved by Minozzi (2017), especially expanding the scope of lemmas with precise synset assignments. However, in other cases the results have been ambiguous, requiring a careful manual review.

- 40 less populated synsets containing at least two monosemous lemmas, hence labelled as "monosemy dataset".

Thus, in spite of the assigned labels, neither subset exclusively comprises polysemous or monosemous lemmas: such datasets would have required an artificial selection of synsets, not grounded on the actual composition of the LWN. Overall, the "polysemy dataset" comprises 28 verbs and 12 nouns, while the "monosemy dataset" includes six verbs, 27 nouns, and eight adjectives.

For the subsequent fine-tuning, we employed the entire LWN updated as of May 2024, excluding the data selected for testing. This training dataset included a total of 9,345 lemmas distributed across 16,529 synsets, broken down as: 2,726 verbs in 4,601 synsets; 983 adjectives in 1,955 synsets; 5,313 common nouns in 9,463 synsets; 233 adverbs in 419 synsets; 90 proper nouns in 91 synsets.

It is important to highlight the methodological significance of using LWN itself as the source of training data. This decision creates a feedback loop where the model, initially trained on structured data, is subsequently used to generate new data of the same nature. This methodology explores not only LLMs' potential to enrich linguistic resources but also examines a potential bidirectional interaction between language models and lexical databases.

To establish a methodological benchmark, we also created an English baseline dataset. This choice aligns with established practices in NLP, where validation in a high-resourced language provides an essential reference point for assessing innovative approaches in low-resourced languages (Bender, 2011; Joshi et al., 2020; Bird et al., 2009; Navigli and Ponzetto, 2012). The English dataset was built following the same structural criteria used for the Latin dataset, ensuring a consistent evaluation framework across the two languages. Through word-by-word translation from Latin into English, we created two parallel sets of synsets.

2.2 Zero-Shot Approach

For both the English and the Latin datasets, a zero-shot experiment was first conducted. This technique, well-documented in the literature (Brown et al., 2020; Perez et al., 2021), leverages the ability of LLMs to tackle new tasks without specific training or providing any example, relying solely on the knowledge acquired during pre-training through a small set of instructions. This step also offered an opportunity to evaluate the inherent understanding and the implicit knowledge Mistral-7B has of Latin vocabulary. We developed our first set of prompts – one for English and one for Latin generation – after a series of testing on a smaller batch that comprised 10 lemmas, equally distributed from both our monosemy dataset and our polysemy one, which through trial and error and various tests led us to find the best approach to instruct our model for the task keeping in mind its limitations (see Appendix A).

2.3 Few-Shot Approach

Following the zero-shot experiment, we developed a few-shot learning strategy. This approach, as described by Brown et al. (2020), allows the model to learn from a limited number of examples provided in the prompt, potentially improving its performance on specific tasks without fine-tuning. As noted by Liu et al. (2021), the effectiveness of few-shot learning heavily depends on the quality and representativeness of the provided examples, to leverage the LLM's general linguistic knowledge, adapting it – in our case – to the specificities of the target language. Perez et al. (2021), suggested that few-shot learning can be particularly effective in specialized domains or for low-resourced languages. For this reason, we developed a set of prompts that maintained the basic structure used in the zero-shot approach but integrated a series of 15 examples with an almost equal distribution of lemmas from our monosemy dataset (7) and from the polysemy one (8). In the initial phases of this approach, we still needed to refine our prompts. Using our prompt testing dataset (see Section 2.2), we conducted a series of 10 tests. These examples provided valuable insights, allowing us to improve our instructions with each iteration and move closer to achieving the desired output. Examples extracted from the final prompt can be seen in Appendix A.

2.4 Fine-Tuning with LoRA

The final phase of our experiment involved fine-tuning using the LoRA technique (Hu et al., 2021b), which introduces low-rank matrices trained in parallel to the original model weights. This allows for targeted adaptation without modifying most of the original parameters, addressing challenges such as computational cost and "catastrophic forgetting" (McCloskey and Cohen, 1989).

We implemented LoRA using Google Colab

with access to an NVIDIA A100 GPU. The LoRA configuration was set with a low-rank matrix dimension (r) of 8 and a scale factor (lora_alpha) of 32. We targeted the query and value projections (q_proj and v_proj) within the model's attention mechanism for adaptation. A dropout rate of 10% was applied for regularization following standard practices (Srivastava et al., 2014). Performance monitoring included metrics such as accuracy, precision, recall, and F1-score (Goutte and Gaussier, 2005). An early stopping mechanism with a patience of one epoch was implemented to prevent overfitting (Prechelt, 1998).

Initially, the training was set for 10 epochs. However, we observed overfitting at the fifth epoch. In response, we recalibrated the process, empirically determining that four epochs provided an optimal balance between task-specific learning and overfitting prevention.

The training process over four epochs revealed insightful trends in both training and validation loss. The training loss showed consistent improvement, decreasing from 2.055000 in the first epoch to 1.629100 in the final epoch. This progressive reduction indicates that the model was effectively learning from the training data, refining its ability to generate Latin synonyms. The validation loss started at 2.05 and reached 1.949 in the final epoch, showing a slight increase from the previous epoch. This behavior aligns with Prechelt (1998) observations on learning dynamics and the risk of overfitting. The final divergence between training and validation loss suggests that the model reached an optimal balance point, as described by Goodfellow et al. (2016). The use of LoRA allowed us to adapt the model to the specific task efficiently, taking into account our limited computational resources and while maintaining its general language understanding capabilities.

3 Results and discussion

This section presents a comprehensive evaluation of our experiment in Latin synonym generation using various approaches of LLMs. Our analysis is twofold, combining quantitative metrics with qualitative observations to provide a bird-eye view of the models' performance and of generated synonyms. The process of annotation and validation of the model's results involved two annotators who worked on assessing the presence in the output of potential synonyms, i.e. lemmas that are semantically similar and may thus be considered for inclusion in the same LWN synset. The quantitative analysis offers a detailed examination of the performance metrics across four distinct approaches: an EB, as well as ZS, FS, and FT models for Latin. We evaluated these approaches using standard metrics such as precision, recall, and F1 score, providing insights into the models' accuracy and efficiency in relation to the final goal of our task. Complementing the statistical evaluation, our qualitative analysis focuses on the linguistic considerations regarding the potential synonyms generated.

3.1 Quantitative analysis

As discussed in section 2, our experiment encompassed different approaches. In this subsection, we will discuss the results of the model output at each stage in order to assess its weaknesses and improvements.

	Overall			**Polysemy**			**Monosemy**		
	F1	P	R	F1	P	R	F1	P	R
EB	.169	.287	.120	.196	.372	.133	.138	.208	.103
ZS	.078	.094	.066	.069	.115	.049	.096	.074	.139
FS	.175	.215	.148	.159	.254	.116	.212	.170	.280
FT	.336	.487	.256	.373	.670	.258	.221	.200	.247

Table 1: Compact performance metrics (F1: F1-score, P: Precision, R: Recall | EB: English Baseline, ZS: Zero-Shot, FS: Few-Shot, FT: Fine-Tuning)

As shown in Table 1, the EB achieved an overall F1-score of 0.169, setting our initial performance benchmark. Interestingly, it showed better performance on lemmas from the polysemy dataset, achieving an F1-score of 0.196, while for lemmas from the monosemy dataset the F1-score was 0.138. This baseline demonstrates the inherent challenges in synonym generation, even in a high-resourced language like English.

The ZS approach showed a significant drop in performance compared to the EB. It achieved an overall F1-score of 0.078, with a precision of 0.094 and a recall of 0.066. Out of 500 generated predictions, only 47 were correct against the 710 ground truth synonyms. This approach struggled particularly with the polysemy dataset (F1-score: 0.069; precision: 0.115, recall: 0.049) compared to the monosemy dataset (F1-score: 0.096; precision: 0.074, recall: 0.139).

The FS method demonstrated a marked improvement over the ZS approach, achieving an overall F1-score of 0.175 – with a precision of 0.215 and a recall of 0.148 – which is comparable to the EB.

Out of 530 predictions, 114 were correct against the 771 ground truth synonyms. Unlike the EB, this approach performed better on the monosemy dataset (F1-score: 0.212) compared to the polysemy one (F1-score: 0.159). This suggests that even a small number of examples can significantly enhance the model's ability to generate Latin potential synonyms, bringing its performance closer to that of the EB.

The FT approach using LoRA showed the most substantial improvement, surpassing both the EB and the former FS approach to Latin with an overall F1-score of 0.336. It achieved a precision of 0.487 and a recall of 0.256. Out of 464 predictions, 226 were correct against the 882 ground truth synonyms. Notably, this approach demonstrated a significant boost in performance for the polysemy dataset (F1-score: 0.373; precision: 0.669, recall: 0.258) compared to the monosemy one (F1-score: 0.221; precision: 0.200, recall: 0.247).

Across all approaches, we observed a general trend of lower recall compared to precision, suggesting that the models were more conservative in their predictions but relatively accurate when they did generate potential synonyms. The fine-tuned model showed the most balanced precision-recall trade-off, particularly for the polysemy dataset (precision: 0.669, recall: 0.258).

The progression from the EB through the various approaches to Latin reveals several interesting trends in synonym generation performance. The ZS generated a similar number of predictions (500) compared to the EB (499), but it experienced a significant drop in accuracy, with precision (0.094) and recall (0.066) both falling well below the baseline. This indicates the difficulty of transferring general language knowledge to a specialized task in an ancient language without task-specific adaptation. The FS method marked a substantial improvement over the ZS approach, bringing the performance close to, and in some aspects surpassing, the EB. With 530 predictions and 114 potential synonyms, it demonstrated that even a small number of examples could enhance the model's ability to generate Latin synonyms. The performance on the monosemy dataset (F1: 0.212) surpassed the one on the polysemy dataset (F1: 0.159), contrasting with the baseline's trend. The fine-tuned model, however, demonstrated the most significant improvement. Despite generating fewer predictions (464) than the other approaches, it produced the highest number of potential synonyms (226). This efficiency is reflected in its greater precision (0.487) and recall (0.256), both outperforming the EB and the previous approaches to Latin (ZS and FS). The fine-tuned model's performance on the polysemy dataset was particularly impressive, with an F1-score (0.373) nearly doubling the performance on the monosemy dataset (0.221), indicating a nuanced understanding of multi-meaning Latin lemmas.

In addition, the disparity in performance between the polysemy and the monosemy datasets is particularly interesting from a linguistic perspective, as it gives insights into the model's ability to navigate semantic complexity. The superior performance on the polysemy dataset (F1-score 0.373 vs 0.221 for the monosemy one) suggests that the model effectively leverages the broader semantic variation associated with polysemous words to generate more potential synonyms; also, as we will further discuss in Section 3.2, it is worth noting that in the polysemy dataset rather common and more frequent lemmas (which we can assume the model has already encountered in its pre-training) are more likely to occur. On the other hand, most of the monosemous terms are rare and less frequent, but at the same time their limited semantic variation seems to constrain the model's ability to generate diverse potential synonyms, resulting in lower recall. These findings align with previous research on the challenges of word sense disambiguation and synonym generation in NLP (Navigli, 2009; Ide and Véronis, 1998). The task of accurately identifying and generating synonyms requires not only a deep understanding of semantic relationships but also the ability to discern subtle nuances in word meanings and usage contexts, which remains a significant challenge for current language models (Camacho-Collados and Pilehvar, 2018).

3.2 Qualitative analysis

The results described in the previous section will now be discussed from a qualitative perspective, in order to better understand how the model actually performed in each phase, highlighting challenges such as hallucinations, task misalignment, and differences between the monosemy and polysemy datasets.

Table 2 provides two examples, illustrating the progression of word generation across the ZS, FS, and FT approaches for both monosemy and poly-

Dataset	Word	ZS	FS	FT
Monosemy	pinna	auris, pinne, pinnulae, pilae, pars alaris externa, caudum, crus	pinnula, remigium	penna, ala, vexillum
Polysemy	abundantia	abundare, copiosus, affluere, plenitudo, primoresprimos, summosmaximioptimom	copiositas, plenitudine, proceritate, opulentas, abundans, affluentias	copiositas, opulentia, uber, ubertas, sufficitentia

Table 2: Examples of Latin Word Generation

semy datasets. This progression demonstrates how the model's performance changes with different levels of task-specific training.

3.2.1 Zero-Shot and Few-Shot

Starting from the ZS approach, the model performed well below the baseline (see Section 3.1). Despite the explicit instructions given in the prompts (see Appendix A), the generations include inadequate results such as English words, multi-word expressions, inflected Latin forms and pseudo-words. These pseudo-words are a byproduct of hallucinations. Hallucinations in LLMs are a well-documented phenomenon whereby the model generates content that is factually incorrect or nonsensical, yet presented confidently (Ji et al., 2022). In the context of this task, hallucinations manifest as the generation of plausible-sounding but non-existent Latin words. This behavior aligns with observations in other studies where LLMs produce fluent but inaccurate responses, especially in specialized domains or low-resourced languages (Maynez et al., 2020). An example of a zero-shot generation for lemmas of the monosemy dataset is given in (1), while (2) exemplifies the results for lemmas of the polysemy dataset:[2]

1. ELOCUTILIS: **eloquens**, *eloquentior*, *etiam loquax*, **verbosus**, *atque fluentissimus*, *fluidores*. Score: 2/6

2. ADHORTATIO: *exhortationem*, **admonitio**, *monitus esto*, *instigare*, *provocatum esse*. Score: 1/5

Another factor that negatively affects the results is the tendency of the model to generate words with a different part of speech (POS) than the input, which does not meet the target of the task. This phenomenon, often referred to as "task misalignment" or "goal misgeneralization", occurs when the model fails to fully grasp or adhere to the specific requirements of the given task (Shah et al., 2022). In our case, the model's propensity to generate words with incorrect POS suggests a failure to maintain consistent morphosyntactic constraints across the input-output pairs.

Comparing these results to the English baseline, some common issues emerge, specifically, the mismatch of POS between the input and the generations and the inclusion of multi-word expressions, pseudo-words and inflected forms in the results. An example from the baseline monosemy dataset is shown in (3) and one from the baseline polysemy dataset in (4):

3. PRUNING: **trimming**, *cutting back*, prune, *a plant*, *cut off branches from*. Score: 1/5.

4. CHAINING: strolling, marching, treading, wandering, ambling, *saunters*, striding, joyfully. Score: 0/8

These examples highlight the complexity of the task of populating synsets, even in a high-resourced language like English. The model struggles with consistently returning one-word items, often providing verbal phrases (e.g., "cutting back") or definitions (e.g., "cut off branches from"). Moreover, the inclusion of inflected forms (e.g., "saunters") and words that are completely misaligned to the POS of the target word (e.g., "joyfully" for CHAINING) further illustrates the difficulty of the task. This complexity is additionally evidenced by our quantitative analysis of the English baseline, discussed in 3.1. These results highlight the inherent challenges in automated synonym generation. However, such issues are much more frequent in the Latin synonym generation task, thus having a greater impact on the results.

It is interesting to note that the results of the zero-shot approach sometimes include portions of text that are unrelated to the task, such as instructions given in the prompts, as in (5), and texts probably retrieved from the data used for the training of the model, as in (6), in which what seems to be glosses and morphological tags are reported in the generation:

5. ACCIO: *exigere*, *esse*, *your response should*

[2]The examples present the input lemma in capital letters, while the generations are thus marked: potential synonyms are bolded (e.g. **admonitio**), existing Latin lemmas that are not valid synonyms are rounded (e.g. actio), and generations corresponding to English words, multi-word expressions, inflected Latin forms or pseudo-words are reported in italics (e.g. *to mix, etiam loquax, eloquentior, fluidores*).

be a json object containing an array of strings. Score: 0/3

6. AGO: *person singular indicative active present tense neuter nominative case*, *agit*, *declension noun*, actio, *accusative plural feminine case*, *actiones*. Score: 0/6

Outputs such as (6) suggest the presence of Latin linguistic data in the model's pre-training corpus. The inclusion of grammatical terms and inflected forms suggests that the model has been exposed to Latin grammatical descriptions and dictionary entries during its training phase. It should be mentioned that our use of the zero-shot method served as a litmus test, allowing us to gauge the extent of Latin knowledge already embedded in the model. The model's tendency to reproduce fragments of linguistic descriptions rather than providing simple synonyms indicates that, while the model has some knowledge of Latin, it struggles to apply this knowledge appropriately within the specific context of our task.

Coming now to the few-shot approach, this methodological turn seems to have a positive effect on the results, which is confirmed by the quantitative analysis. In particular, the presence of English lemmas and multi-word expressions in the generations decreases significantly, while pseudo-words and inflected forms remain frequent. Examples are offered in (7), presenting the generations for a lemma from the monosemy dataset, and (8), which contains the results for a lemma from the polysemy dataset:

7. SUAVILOQUENS: *dilatabiliter*, *loquebatur*, *modestius*. Score: 0/3

8. ACCUMULATIO: **acervus**, *agglomeratum*, **caterva**, *congestion*, **copia**, **grex**, **massa**, mulier, *monte*, *pecuie*, *plebe*. Score: 5/11

The mismatch between the POS of the input and that of the results is an issue that also affects the generations obtained through this approach. Overall, the performance of the model increases compared to the zero-shot approach, showing a greater adherence to the task and to the prompts. This means that the few-shot approach proves useful in adapting the model's performance to a low-resourced language such as Latin (as highlighted in 3.1).

3.2.2 LoRA Fine-Tuning

Moving on to the results of fine-tuning, one can acknowledge great improvements, both for monosemous lemmas, as (9), and for polysemous lemmas, like (10):

9. COGITAMENTUM: **meditatio**, cognomen, *cognomentum*, nomen. Score: 1/4.

10. ABSCONDO: **abscido**, delego, delicio, **detraho**, effingo, extinguo, **pello**, **occulto**, **praeservo**, **servo**, **tego**. Score: 7/11

The occurrence of English lemmas and multi-word expressions in fine-tuning results is close to none, while inflected forms and pseudo-words are still present, though less frequently than in the results of the previous approaches. Furthermore, the correspondence between the POS of the input and that of the results is higher.

An interesting observation stems from example (9), in which *meditatio* is validated as a potential synonym of *cogitamentum*. Currently, in the LWN, *meditatio* and *cogitamentum* are not considered synonyms as they do not share any common synsets. However, these two lemmas both appear in Latin dictionaries with the definition 'a thought' (Glare 1968, s.v. *meditatio*; Lewis and Short 1879, s.v. *cogitamentum*). This instance thus proves the potential of the approach adopted in this work in assisting humans in the annotation process by identifying synonymy relations which might not have been encoded in the WordNet.

It should be mentioned that the model produced empty outputs on three occasions during the synonym generation task: once for a monosemous lemma (*commisereor*) and twice for polysemous lemmas (*carpo*, *circumscriptio*). This phenomenon has otherwise been observed only once, specifically with the zero-shot approach on the monosemy dataset (*actutum*). While the generation of an empty output is inconclusive for the task at hand, at the same time it might be a sign of improvement and adaptation of the model, showing a preference for generating an empty output instead of unrelated results.

Interestingly, the fine-tuning approach shows more encouraging results in generating potential synonyms for verbs as opposed to other POS. The model also performs better with polysemous rather than with monosemous lemmas. This improved performance with polysemous lemmas can be partially attributed to the nature of the generation process

itself. As the model's output is based on stochastic prediction, polysemous lemmas offer a broader semantic space from which to generate potential synonyms, increasing the likelihood of producing correct responses. This phenomenon aligns with several studies in the field of NLP and cognitive science. Pilehvar and Camacho-Collados (2019) discuss how word sense disambiguation benefits from the rich semantic space of polysemous lemmas in vector space models, which is analogous to our observation in synonym generation. Similarly, Ethayarajh (2019) demonstrates that contextual word embeddings capture more information for polysemous lemmas due to their varied usage contexts. These studies collectively support the idea that the richer semantic variation of polysemous lemmas can lead to improved performance in various language tasks, including, as we observe, synonym generation with language models.

While this trend toward better performance with polysemous lemmas neatly emerges from our results, a closer examination reveals an interesting pattern specific to our dataset. The model's performance appears to be particularly better with verbs. This pattern persists despite the makeup of our training data (cf. 2.1): even though nouns nearly double the number of verbs in the training data, the model still performs better with verbs. This observation may be partially explained by the composition of our dataset: verbs are much more represented in the polysemy than in the monosemy dataset (28 vs 6 verbs). This unbalance is probably due to the fact that verbs are inherently more polysemous than other POS. For instance, Gentner and France (1988) demonstrated that verbs are more likely to extend their meanings in novel contexts compared to nouns. Similarly, Fellbaum (1990) showed that verbs have a higher degree of polysemy in the Princeton WordNet compared to nouns. On the other hand, the difference in performance might be explained considering the lemmas that constitute the monosemy dataset: the majority seems to be rare words, often associated with a very specific meaning and few synonyms. The issues with monosemous lemmas can thus be explained by the fact that they are under-represented in the pre-training dataset on account of their low frequency and also by the fact that they have few synonyms because of the specificity of their meaning. This hypothesis is backed up by the obtaining a satisfactory performance with frequent monosemous lemmas that have numerous synonyms, such as (11) and (12):

11. ASPORTATIO: *abductionem*, **captura**, carnificina, *furta*, **rapina**, *stulcium*, *ueneficiam*, **latrocinium**, *strage*, pugna, bellum, *luparium*, *saeculariua*, nex, *mordebatio*, *praedae*, **spoliatio**. Score: 4/17

12. POLLICITATIO: **votum**, **fides**, **foedus**, **pactum**, sancimentum, **testamentum**. Score: 5/6.

In conclusion, the experiment – and particularly the fine-tuning approach – has revealed complex patterns that go beyond simple performance differences based on the monosemy-polysemy opposition. Furthermore, the challenges encountered with monosemous lemmas, and especially with particularly rare terms with highly specific meanings, highlight the importance of considering word frequency and semantic specificity in model training and evaluation.

4 Conclusions

This study investigated the use of LLMs to enrich the LWN through automated synonym generation, specifically by comparing ZS, FS and FT approaches. The results provide several important insights and suggest potential paths for advancing the use of LLMs in enriching lexical resources for ancient and low-resourced languages such as Latin. First, we found that the zero-shot approach offers an initial baseline for Latin synonym generation, but it lacks accuracy, showing the difficulty of directly applying LLMs to ancient languages without task-specific adaptation. The few-shot approach shows a significant improvement in the synsets population, suggesting that even a small number of task-specific examples can significantly improve the model's performance. The most important results were achieved by the FT approach using the LoRA technique. This approach produced better results than ZS and FS approaches, particularly in the generation of potential synonyms for polysemous lemmas. Overall, this study not only advances our understanding of automatic synonym generation for Latin, but also provides insights into the broader challenges of processing ancient languages and dealing with semantic complexity in NLP. Furthermore, the results obtained with our fine-tuned model can be used to partially automate the synset

annotation process, providing substantial support to annotators.

Future research could explore the development of models that result in a better performance across different parts of speech and degrees of polysemy, potentially incorporating etymological information or using cross-linguistic data from related languages. Also, it could be interesting to further evaluate the results related to the addition of new data – such as other dictionaries – and a possible revision of the current dataset – taking into account the findings of this experiment on rare lemmas – to fine-tune and ground the model even more, with the ultimate goal to improve overall performance and reduce hallucination. In addition, investigating whether and how the approaches we employed apply to other ancient languages could contribute to understanding the universality of these semantic processing patterns in computational linguistics.

A Prompts Used in the Experiment

This appendix contains the full prompts used in our experiment for both Latin and English.

A.1 Latin Prompt

```
latin_prompt = f"""You are a powerful AI
    ↪  assistant trained in semantics.
You are a Latin native speaker. The only
    ↪  language you speak is Latin.
Your task is to provide a bullet list of
    ↪  Latin synonyms for a user-chosen
    ↪  word.
Observe the following instructions very
    ↪ closely:
[INST]
- Generate only Latin synonyms.
- Provide single-word expressions only.
- Do NOT generate long phrases.
- ABSOLUTELY AVOID including any
    ↪ additional explanations or
    ↪ comments in your output.
- VERY IMPORTANT: DO NOT translate the
    ↪ words.
- VERY IMPORTANT: Use LATIN exclusively.
- For NOUNS generate only the NOMINATIVE
    ↪  CASE, as shown in the examples
    ↪ below.
- For VERBS generate only the FIRST-
    ↪ PERSON SINGULAR of the INDICATIVE
    ↪ , as shown in the examples below.
- List each Latin word separately with
    ↪ proper formatting.
## Note
Note that the examples provided may
    ↪ predominantly feature words
    ↪ starting with specific letters by
    ↪  chance and should not influence
    ↪ the generation process to favor
    ↪ those letters.
Ensure that the generated Latin synonyms
    ↪  start with a wide range of
    ↪ letters from the alphabet.
### Examples
(...)
[/INST]
'{word}':
Synonyms:
"""
```

A.2 English Prompt

```
english_prompt = f"""You are a powerful
    ↪ AI assistant trained in semantics
    ↪ .
Your task is to provide a bullet list of
    ↪  English synonyms for a user-
    ↪ chosen word.
Observe the following instructions very
    ↪ closely:
[INST]
- Generate only English synonyms.
- Provide single-word expressions only.
- Do NOT generate long phrases.
- IMPORTANT: Do NOT any additional
    ↪ explanations or comments in your
    ↪ output.
- List each English word separately with
    ↪  proper formatting.
### Examples
(...)
[/INST]
'{word}':
Synonyms:
"""
```

A.3 Examples from the Final Prompt

```
word: 'asparagus'
synonyms: ['bracchium', 'cacumen', '
    ↪ flagellum', 'frutex', 'pertica',
    ↪ 'planta',
            'propago', 'sagitta', '
                ↪ sarmentum', 'semen', '
                ↪ stirps', 'suboles',
            'suffrago', 'uirga', 'uitis']

word: 'ordo'
synonyms: ['protelum', 'series', 'uersus
    ↪ ']
```

References

Ainslie, J., Aneja, J., Cowan, B., Eltanbouly, A., Gillick, D., Goldberg, Y., Gopalakrishnan, K., Jiang, A., King, M., Martens, J., et al. (2023). Grouped query attention for long context large language models. *arXiv preprint arXiv:2305.13245*.

Apidianaki, M. and Sagot, B. (2014). Data-driven synset induction and disambiguation for wordnet development. *Language Resources and Evaluation*, 48:655–677.

Bender, E. M. (2011). *Linguistic fundamentals for natural language processing: 100 essentials from morphology and syntax*. Morgan & Claypool Publishers.

Bentivogli, L., Forner, P., Magnini, B., and Pianta, E. (2002). Revising the wordnet domains hierarchy: semantics, coverage and balancing. In *Proceedings of the Workshop on Multilingual Linguistic Resources*, pages 101–108.

Biagetti, E., Zanchi, C., and Short, W. M. (2021). Towards a gold standard for a latin wordnet: Setting evaluation standards from ancient to modern languages. In *Proceedings of the 11th Global Wordnet Conference*, pages 47–57.

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Bizzoni, Y., Boschetti, F., Del Gratta, R., Diakoff, H., Monachini, M., and Crane, G. (2014). The making of ancient greek wordnet. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland*, pages 1140–1147. European Language Resources Association (ELRA).

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Camacho-Collados, J. and Pilehvar, M. T. (2018). From word to sense embeddings: A survey on vector representations of meaning. *Journal of Artificial Intelligence Research*, 63:743–788.

Ethayarajh, K. (2019). How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. 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 55–65.

Exeter University (2023). Expansion of the latin wordnet at exeter university. Unpublished work.

Fellbaum, C. (1990). English verbs as a semantic net. *International Journal of Lexicography*, 3(4):278–301.

Franzini, G., Peverelli, A., Ruffolo, P., Passarotti, M., Sanna, H., Signoroni, E., Ventura, V., and Zampedri, F. (2019). Refining the latin wordnet. In *Proceedings of the Sixth Italian Conference on Computational Linguistics*.

Gentner, D. and France, I. M. (1988). The verb mutability effect: Studies of the combinatorial semantics of nouns and verbs. In *Lexical ambiguity resolution*, pages 343–382. Morgan Kaufmann.

Glare, P. G. W., editor (1968). *Oxford Latin Dictionary (OLD)*. Oxford University Press, Oxford.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

Goutte, C. and Gaussier, E. (2005). A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In *European conference on information retrieval*, pages 345–359. Springer.

Hellwig, O. (2017). Coarse semantic classification of rare nouns using cross-lingual data and recurrent neural networks. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS)*, volume 137, pages 3934—-3941. European Language Resources Association (ELRA).

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., and Chen, W. (2021a). Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Chen, W., et al. (2021b). Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Ide, N. and Véronis, J. (1998). Introduction to the special issue on word sense disambiguation: the state of the art. *Computational linguistics*, 24(1):2–40.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., and Fung, P. (2022). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Jiang, Z., Chen, Y., Tu, K., Wang, W., Qin, B., and Li, T. (2023). Multilingual language models are better zero-shot learners. *arXiv preprint arXiv:2309.07445*.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., and Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv:2004.09095*.

Khan, F., Minaya Gómez, F. J., Cruz González, R., Diakoff, H., Diaz Vera, J. E., McCrae, J. P., O'Loughlin, C., Short, W. M., and Stolk, S. (2022). Towards the construction of a wordnet for old english. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, Marseille, France*, volume 137, pages 3934–3941. European Language Resources Association (ELRA).

Lewis, C. T. and Short, C. (1879). *A Latin Dictionary (LS)*. Clarendon Press, Oxford.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2021). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Mambrini, F., Passarotti, M., Litta, E., and Moretti, G. (2021). Interlinking Valency Frames and WordNet Synsets in the LiLa Knowledge Base of Linguistic Resources for Latin. In *Further with Knowledge Graphs*, volume 53 of *Studies on the Semantic Web*, pages 16–28.

Maynez, J., Narayan, S., Bohnet, B., and McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.

McCloskey, M. and Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of learning and motivation*, 24:109–165.

Mehler, A., Jussen, B., Geelhaar, T., Trautmann, W., Sacha, D., Schwandt, S., Glądalski, B., Lücke, D., and Gleim, R. (2020). The frankfurt latin lexicon: From morphological expansion and word embeddings to semiographs. *Studi e Saggi Linguistici*, 58(1):121–155.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J. (1990). Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244.

Minozzi, S. (2009). The latin wordnet project. In *Latin Linguistics Today. Akten des 15. Internationalem Kolloquiums zur Lateinischen Linguistik*, volume 137, pages 707–716. Innsbrucker Beiträge zur Sprachwissenschaft.

Minozzi, S. (2017). Latin wordnet, una rete di conoscenza semantica per il latino e alcune ipotesi di utilizzo nel campo dell'information retrieval. *Umanistica Digitale*, 1(1).

Mistral AI (2023). Mistral 7b. https://github.com/mistralai/mistral-src. Accessed: 2023-10-11.

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*, 41(2):1–69.

Navigli, R. and Ponzetto, S. P. (2012). Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. In *Artificial Intelligence*, volume 193, pages 217–250. Elsevier.

Passarotti, M., Cecchini, F. M., Franzini, G., Litta, E., Mambrini, F., and Ruffolo, P. (2019). Lila: Linking latin. risorse linguistiche per il latino nel semantic web. In *Umanistica Digitale*, volume 3(5).

Perez, E., Kiela, D., and Cho, K. (2021). True few-shot learning with language models. *Advances in Neural Information Processing Systems*, 34:11054–11070.

Pilehvar, M. T. and Camacho-Collados, J. (2019). Wic: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of NAACL-HLT 2019*, pages 1267–1273.

Prechelt, L. (1998). Early stopping-but when? *Neural Networks: Tricks of the trade*, pages 55–69.

Sandhan, J. K., Adideva, O., Komal, D., Modani, N., Naik, A., Muthiah, S. K., and Kulkarni, M. (2021). Evaluating neural word embeddings for sanskrit. https://arxiv.org/pdf/2104.00270.pdf. Accessed: [Insert access date here].

Shah, R., Al-Shedivat, M., Carbonell, J., and Gu, A. (2022). On the pitfalls of goal misgeneralization in learning diverse action sequences. *arXiv preprint arXiv:2206.01222*.

Singh, P., Rutten, G., and Lefever, E. (2021). Pilot study for bert language modelling and morphological analysis for ancient and medieval greek. In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 129–135, Punta Cana, Dominican Republic (online). Association for Computational Linguistics.

Snow, R., Jurafsky, D., and Ng, A. Y. (2005). Learning syntactic patterns for automatic hypernym discovery. In *Advances in Neural Information Processing Systems*, volume 17, pages 1297–1304. MIT Press.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Vossen, P. (2002). Eurowordnet: general document. In *EuroWordNet: A multilingual database with lexical semantic networks*, pages 1–39. Springer.

Zhang, K., Luo, Y., Qin, Y., Zhang, S., Wu, Y., Xu, R., and Fu, Q. (2024). Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.