Lemma Dilemma: On Lemma Generation Without Domain- or Language-Specific Training Data

Olia Toporkov[1], Alan Akbik[2], Rodrigo Agerri[1]
[1]HiTZ Center - Ixa, University of the Basque Country UPV/EHU
[2]Humboldt-Universität zu Berlin
{olia.toporkov, rodrigo.agerri}@ehu.eus, alan.akbik@hu-berlin.de

Abstract

Lemmatization is the task of transforming all words in a given text to their dictionary forms. While large language models (LLMs) have demonstrated their ability to achieve competitive results across a wide range of NLP tasks, there is no prior evidence of how effective they are in the contextual lemmatization task. In this paper, we empirically investigate the capacity of the latest generation of LLMs to perform in-context lemmatization, comparing it to the traditional fully supervised approach. In particular, we consider the setting in which supervised training data is not available for a target domain or language, comparing (i) encoder-only supervised approaches, fine-tuned out-of-domain, and (ii) cross-lingual methods, against direct in-context lemma generation with LLMs. Our experimental investigation across 12 languages of different morphological complexity finds that, while encoders remain competitive in out-of-domain settings when fine-tuned on gold data, current LLMs reach state-of-the-art results for most languages by directly generating lemmas in-context without prior fine-tuning, provided just with a few examples. Data and code available upon publication: https://github.com/oltoporkov/lemma-dilemma

1 Introduction

Lemmatization is one of the core NLP tasks widely used during data pre-processing in such areas as information extraction, named entity recognition, and sentiment analysis, and is of particular importance for languages with complex morphology. To lemmatize a word means to transform its inflected form (e.g. *chose, chosen*) into its dictionary-based form, also known as lemma (e.g. *choose*), according to the definition of the contextual lemmatization task in SIGMORPHON 2019 (McCarthy et al., 2019).

Most recent approaches tend to address contextual lemmatization as a supervised classification approach, which was first proposed by Chrupala et al. (2008) and became the core idea in the architecture of a variety of contextual lemmatizers (Malaviya et al., 2019; Straka et al., 2019; Yildiz and Tantuğ, 2019). This method learns to determine the minimum number of edits necessary to convert the word into its lemma. Such techniques require a large amount of annotated data, which can be especially challenging for languages with rich morphology (Straka et al., 2019; Yildiz and Tantuğ, 2019). Apart from that, the majority of lemmatization systems are evaluated mostly in-domain, while their real application is almost always out-of-domain, namely, out of the scope of the data they have been trained on. Previous work has demonstrated that lemmatizers' performance worsens substantially when deployed out-of-domain (Toporkov and Agerri, 2024).

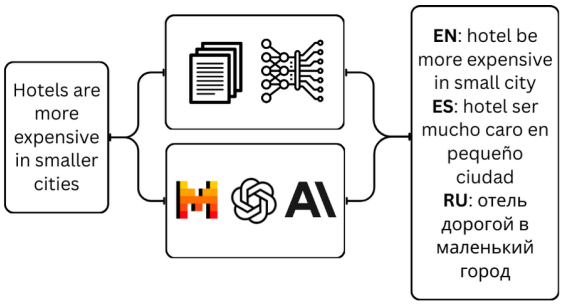

Figure 1: General overview of the lemmatization task process for in-context learning and supervised approaches.

The recent generation of large language models (LLMs) has exhibited strong capabilities on a wide variety of NLP tasks, such as reasoning, problem-solving, code generation, information extraction, and text composition (Brown et al., 2020; Clark et al., 2021; Wang et al., 2022; Bubeck et al., 2023; Sviridova et al., 2024; Sainz et al., 2024). However, as it has also been claimed, when LLMs are evaluated on languages other than English or other high-resource languages (e.g., Spanish), their per-

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 18219–18232

formance is not as good as expected (Blasi et al., 2022; Arnett and Bergen, 2025; Figueras et al., 2025). As for the contextual lemmatization task, to the best of our knowledge, there is no empirical evidence on the capacity of such models in generating the correct lemmas, especially for high-inflected and low-resource languages.

The scarcity of manually annotated data is another problem for training competitive contextual lemmatizers. The attempts to create large annotated corpora, such as Universal Dependencies (Nivre et al., 2017) and the UniMorph project (McCarthy et al., 2020), aim to bridge this resource gap, but they still cover a very limited scope of languages and domains (Joshi et al., 2020). In response to data scarcity, model- and data-based cross-lingual transfer for sequence labeling have been proposed (García-Ferrero et al., 2022; Chen et al., 2023; Yeginbergen et al., 2024). These approaches focus on overcoming the lack of data in a target language by either fine-tuning a pre-trained multilingual model on a source language (usually English) in order to make predictions for any of the target languages included in the model pre-training (*model-transfer*); or automatically generating labeled data for the target language (*data-transfer*), which is then used to train a sequence labeling model.

Considering the lemmatization challenges mentioned above, in this paper we raise the following research questions (RQ): **RQ1**: To what extent can the latest generation of large language models directly obtain lemmas in the target language without prior fine-tuning, especially for the languages with complex morphology? **RQ2**: Could in-context lemma generation, produced by LLMs, be an alternative in solving the lemmatization task out-of-domain? **RQ3**: What is the best strategy in the scenario where the data in the target language or domain is non-existent or difficult to access? The setup to investigate these RQs is illustrated in Figure 1.

Hence, our contributions are the following: (i) we empirically investigate the ability of LLMs to perform in-context lemma generation across languages of different morphological complexity; (ii) we address the out-of-domain problem by comparing the performance of LLMs against encoder-only models fine-tuned on different data distribution, and (iii) we conduct a comparative analysis of the methods to overcome the data scarcity in the target language for the lemmatization task, namely, model-transfer, data-transfer and direct lemma generation, using the data in English language as a source. Overall, our results suggest that, while fine-tuning encoders on gold data remains a competitive option for out-of-domain settings, generative LLMs reach state-of-the-art results in lemmatization by directly generating lemmas in-context without prior fine-tuning, provided just with a few examples.

2 Related Work

Approaches to perform lemmatization evolved from being rule-based language-dependent systems (Karttunen et al., 1992; Oflazer, 1993; Alegria et al., 1996; Segalovich, 2003; Jongejan and Dalianis, 2009) to advanced multilingual architectures, trained on a large amount of annotated data (Müller et al., 2015; Bergmanis and Goldwater, 2018; Malaviya et al., 2019; Straka et al., 2019). The idea of addressing lemmatization as a sequence labeling task, where the labels are induced as the minimum amount of edits necessary to convert an inflected word into its lemma, was first proposed by Chrupala et al. (2008) and has been adopted in a wide variety of systems such as the system by Gesmundo and Samardžić (2012), IXA pipes (Agerri et al., 2014; Agerri and Rigau, 2016), Lemming (Müller et al., 2015), the system of Malaviya et al. (2019) and Morpheus (Yildiz and Tantuğ, 2019), among others.

The advancement of supervised techniques involving deep learning algorithms, and the development of the Transformer architecture (Vaswani et al., 2017) and Transformer-based masked language models (MLMs) such as BERT (Devlin et al., 2019) and the multilingual XLM-RoBERTa (Conneau et al., 2020), have massively improved the performance of current contextual lemmatizers. Thus, in the SIGMORPHON 2019 Shared Task (McCarthy et al., 2019) on contextual lemmatization, most of the participant systems were based on MLMs (Straka et al., 2019; Kondratyuk, 2019; Shadikhodjaev and Lee, 2019).

The evaluation of contextual lemmatizers is almost always performed in-domain, namely, on the same data distribution used during the fine-tuning process. However, in practice, lemmatizers are usually deployed out-of-domain, which results in a significant performance degradation, especially for high-inflected languages (Toporkov and Agerri, 2024).

The rise and constant development of LLMs has

demonstrated remarkable abilities in dealing with a wide variety of NLP tasks such as language understanding, reasoning, language generation, code generation, and query response, especially when using a few-shot learning approach (Brown et al., 2020; Shi et al., 2022; Ahuja et al., 2023; Fernandes et al., 2025). Model families such as LLaMa (Grattafiori et al., 2024), Generative Pre-trained Transformer (GPT) (OpenAI et al., 2024), Qwen (Yang et al., 2024), Claude (Anthropic, 2025), Mistral (Jiang et al., 2023; Mistral, 2024) or Gemma (Gemma Team et al., 2024a,b), to name but a few, are experiencing continuous growth, offering LLMs with different parameter sizes and evaluated on various downstream tasks.

Nevertheless, there seems to be a noticeable difference in the performance of such models depending on the resources of the language and its complexity. Arnett and Bergen (2025) demonstrate that there is a disparity in the performance of LLMs between agglutinative and fusional languages, giving an advantage to high-resource languages like English over more morphologically complex languages such as Turkish. This is attributed to tokenization quality, stating that morphological alignment does not influence the model's performance. In one of the first systematic analyses of the morphological capabilities of LLMs for the tasks of inflection and reinflection, models such as ChatGPT are still far from achieving state-of-the-art results, performing on the level of some older supervised models (Weissweiler et al., 2023). In their multilingual version of the Wug test, Anh et al. (2024) indicate that LLMs can apply their morphological knowledge to previously unseen words and that the morphological complexity of languages is more important than their relative representation in the training data. This fact states the importance of morphology for improving low-resource language modeling.

Furthermore, it has also been demonstrated that LLMs could be used for lemma disambiguation in a dictionary-augmented approach for the endangered languages such as Erzya and Skolt Sami (Hämäläinen, 2024), reaching the level close to a human annotator. Nonetheless, to the best of our knowledge, no empirical work has studied the ability of LLMs to generate correct lemmas.

The lack of quality annotated corpora for many target languages led to the exploration of model-transfer and data-transfer techniques. Model-transfer is based on the cross-lingual capabilities of multilingual pre-trained MLMs, where the knowledge in the source data can be transferred to the target language (Wang et al., 2023). Data-transfer aims to automatically produce labeled data for the target language, traditionally based on translation and annotation approaches (Fei et al., 2020). Both techniques have demonstrated competitive results in the tasks of cross-lingual transfer for various sequence labeling tasks (García-Ferrero et al., 2022; Chen et al., 2023), although lemmatization has not been studied from a cross-lingual transfer perspective.

3 Materials and Methods

In this Section, we describe the datasets and models we use for our experiments. We also describe the prompting method we apply to perform in-context lemma generation.

3.1 Datasets

In order to address the three RQs, we use 2 types of corpora, described below.

To address RQ1 and RQ3, we use parallel corpora, namely, the PUD treebank, presented for the CoNLL 2017 Shared Task (Zeman et al., 2017). This corpus was created for 18 languages, and each PUD dataset consists of 1000 sentences extracted from online sources and Wikipedia; 750 of 1000 sentences were originally in English, while the rest came from German, Spanish, French, and Italian texts. The corpora were further translated by professional translators to the remaining languages.

Regarding RQ1, we chose a limited scope of languages from our selection in order to test the in-context lemma generation using LLMs, namely, English, Spanish, Russian, and Basque. As Basque does not have parallel corpora in the PUD treebank, we took the first 100 sentences of the BDT test corpus to establish an equal experimental setup. Such language selection was motivated by our ability to conduct detailed in-house analysis of the obtained results, as well as to determine the optimal model settings.

Concerning the experimentation to answer RQ3, we chose 12 languages of different morphological complexity, namely English, Spanish, French, German, Italian, Finnish, Icelandic, Turkish, Swedish, Czech, Polish, and Russian. In order to perform model transfer experiments, we split the PUD data into standard training, development, and test partitions, resulting in 800 sentences for training and

100 sentences for the development and test sets, respectively.

To respond to RQ2, we use datasets of different data distributions for fine-tuning and testing in Basque, Czech, English, Spanish, Russian, and Turkish. The datasets were developed for the SIGMORPHON 2019 Shared Task on contextual lemmatization (McCarthy et al., 2019). The data is annotated according to the Unimorph scheme (McCarthy et al., 2020), the only exception being the Basque Armiarma corpus, an external dataset to UD, which includes lemma annotations of literary critics (Armiarma, 2000). This selection of languages and datasets allows us to compare with previous out-of-domain lemmatization results (Toporkov and Agerri, 2024). In order to make the computational load more manageable, the number of sentences in the larger datasets is reduced to 900. Statistics regarding the token and sentence counts in the final sets are presented in the Table 1.

Language	Corpus	Tokens	Sentences
Basque	BDT	11901	900
Basque	Armiarma	17172	900
Spanish	AnCora	26917	900
Spanish	GSD	24412	900
English	EWT	13690	900
English	GUM	8189	440
Turkish	IMST	5734	564
Turkish	PUD	1795	100
Czech	CAC	17855	900
Czech	PUD	1930	100
Russian	GSD	9874	503
Russian	SynTagRus	16594	900

Table 1: Datasets for the experiments on corpora across different distributions.

3.2 Models

We evaluate the performance of several state-of-the-art multilingual instruction-tuned generative large language models, specifically, Mistral-Large-Instruct-2407 (Mistral, 2024), LLaMA-3.3-70B-Instruct (Grattafiori et al., 2024), Qwen-2.5-72B-Instruct (Yang et al., 2024), and Claude-3.7-Sonnet (Anthropic, 2025) (100B+ paremeters). All models are evaluated using zero-shot and few-shot (1, 2, 3, 4 and 5) in-context learning. We use the default configuration for all LLMs. Every model except Claude has publicly released their weights.

For the contrastive supervised approach, we apply the large version of XLM-RoBERTa (Conneau et al., 2020). This encoder model is based on the RoBERTa architecture and was pre-trained on 2.5TB of filtered CommonCrawl data for 100 languages and has obtained state-of-the-art results for many discriminative tasks (García-Ferrero et al., 2022), such as cross-lingual sequence labeling.

3.3 Prompt Design

To establish which prompt could provide us with the optimal results in the lemma generation task, we try 4 different prompt settings to perform zero-shot and few-shot experiments. As mentioned earlier, we experiment with 4 languages of different morphological complexity, namely English, Russian, Spanish, and Basque.

For each prompt type, we introduce two different inputs to the model: the whole sentence as one string and the sentence introduced as a list of words. We aim to receive the output as a whole sentence presented in the form of *word-lemma* pairs. The instructions provided for the models are always given in English. The prompts are designed in the following fashion:

- *basic prompt*: simple description of the task;
- *full prompt*: simple description of the task + lemmatization instructions;
- *basic prompt + k-shot examples [1:5]*;
- *full prompt + k-shot examples [1:5]*.

Simple description of the task (basic prompt). The basic prompt consists of a brief description of the lemmatization task without specifying any particular instructions. The example of such a prompt is presented below:

> *"Your task is to lemmatize a sentence in Spanish. You will be given a sentence, where each word starts from the new line. You need to provide for each word in the given sentence its dictionary form (lemma).*
> *Provide the output in **TSV format** (Tab-Separated Values) with the format:*
> *'initial word lemma'*
> *Sentence: "El Parque Golden Gate ofrece un jardín botánico , un planetario , y un jardín japonés ."*
> *Answer with the required output only, without extra spaces, quotation marks, or comments."*

Simple description of the task + explicit lemmatization instructions (full prompt). The preliminary output analysis using the basic prompt demonstrated that the model was skipping particular words, introducing amendments to the existing words (e.g., if the word was misspelled in the original corpus), or struggling with specific particles such as the articles in Spanish. Therefore, to improve the LLM's performance, we accompanied

the prompt with a detailed set of instructions to perform the lemmatization task, presented below:

> *Instructions:*
> *1. Copy the word exactly as it is, and provide its lemma.*
> *2. **Process Every Word**: Lemmatize **each word** in the sentence. Do not omit, change, or remove any word.*
> *3. **Handle Spelling Errors**: If a word is misspelled, retain the original spelling as the initial word, but lemmatize it to the closest dictionary form.*
> *4. **Proper Nouns**: Proper nouns should retain their capitalization.*
> *5. **Punctuation**: Include punctuation marks in the output, using the mark itself as the lemma.*
> *6. **Part-of-Speech**: Lemmatize words based on their part of speech (POS) (e.g., verbs to their infinitive form, nouns to singular form).*
> *7. **Articles**: Use the masculine singular form for articles.*
> *8. **Multi-Word Expressions**: If an input contains multiple words, process each word separately.*

Basic prompt and full prompt + k-shot examples. The third and fourth prompts include examples. In order to choose the examples for the few-shot experiments, we experimented with the development set of the PUD corpus using Mistral-Large-Instruct-2407. We tried manual example selection, random example selection, and choosing the examples with the highest number of errors committed, such as skipping one of the words in the sentence, wrong lemma generation, and generation of additional elements. We then introduce examples in a range of 1 to 5 to each prompt configuration, namely, basic and full (an example of the prompt is given in Appendix A). The prompt configuration with the best results on this particular experiment is then applied to the rest of our experiments.

4 Experiments

In this paper, we aim to answer the following questions: (RQ1) What are the capabilities of the latest generation of LLMs to perform in-context learning for lemmatization? (RQ2) Can we use direct lemma generation to address the problem of poor out-of-domain performance, namely, when evaluated on a data distribution different from the one seen during training? (RQ3) What is the optimal strategy in a scenario in which annotated training data in the target domain or language is not available?

To address these questions, we conduct three sets of experiments. Each experimental setup and its corresponding results are described below in a separate subsection. Evaluation is performed using average word and sentence accuracy metrics.[1] We conduct 3 runs and report the average results for all experiments and models, except for Claude-3.7-Sonnet, where only 1 run was performed due to computational costs. To assess whether the observed differences in model performance are statistically significant, we apply McNemar's test (Dietterich, 1998).

4.1 In-context Lemma Generation

To respond to RQ1, we experiment with direct lemmatization using LLMs, employing Mistral-Large-Instruct-2407 and the prompt types described in Section 3.3. This model was selected for its strong performance-to-cost ratio, fast inference times, and accessibility within our computational environment. For each prompt, 3 different runs are performed using the datasets described in Section 3.1 on English, Spanish, Russian, and Basque. The most effective prompt setting will be identified based on word and sentence accuracy results, performance on highly inflected languages such as Russian and Basque, and the number of hallucinations and errors.

Table 2 reports the results on 0-shot and 4-shot (best overall configuration in terms of the number of examples) experiments combining different prompt strategies. It can be observed that when given the basic prompt without specific instructions and in a zero-shot setting, the model fails to generate a lot of input words across all languages.

Adding some examples to the prompt significantly improves the results, especially for more complex languages such as Russian and Basque. The input format also plays a role: with the sentence represented as separated tokens, the model hallucinates less and obtains higher accuracies. As stated by Arnett and Bergen (2025), this could be directly connected to the tokenization quality. Surprisingly, introducing example selection (described in Section 3.3) into the prompt makes the basic prompt outperform the full one when the input is a list of tokens.

In addition to word and sentence accuracy, the optimal prompt design is chosen based on the amount of produced errors, hallucinations, and missing words. Thus, Table 2 shows that the best setting corresponds to the ***basic prompt + 4-shot examples***, where the input is a sentence as a list of

[1] Sentence accuracy allows for better discrimination between models' performance (Toporkov and Agerri, 2024).

			input: sentence as one string				*input: sentence as a list of words*			
			WAcc	*SentAcc*	*missing words*	*wrong words*	*WAcc*	*SentAcc*	*missing words*	*wrong words*
en	*basic prompt*	*0-shot*	0.93	0.28	30	0	0.95	0.38	0	0
	full prompt	*0-shot*	0.95	0.43	4	0	0.95	0.42	2	0
	basic prompt	*4-shot*	0.96	0.42	0	0	0.96	0.44	0	0
	full prompt	*4-shot*	**0.96**	**0.46**	0	0	0.95	0.43	0	0
es	*basic prompt*	*0-shot*	0.92	0.12	32	2	0.93	0.22	0	0
	full prompt	*0-shot*	0.93	0.22	4	3	0.92	0.17	1	0
	basic prompt	*4-shot*	0.96	0.41	2	0	**0.97**	**0.45**	1	0
	full prompt	*4-shot*	0.96	0.42	2	0	0.96	0.43	0	0
ru	*basic prompt*	*0-shot*	0.91	0.20	32	5	0.92	0.29	0	2
	full prompt	*0-shot*	0.94	0.33	4	3	0.94	0.34	0	4
	basic prompt	*4-shot*	0.93	0.41	42	0	**0.95**	**0.40**	2	0
	full prompt	*4-shot*	0.93	0.37	42	0	0.94	0.37	1	0
eu	*basic prompt*	*0-shot*	0.76	0.00	78	3	0.82	0.07	3	0
	full prompt	*0-shot*	0.78	0.03	44	3	0.81	0.05	1	0
	basic prompt	*4-shot*	0.86	0.23	38	0	**0.89**	**0.29**	7	0
	full prompt	*4-shot*	0.86	0.25	34	0	0.89	0.27	1	0

Table 2: Word and sentence accuracy using different prompting strategies with Mistral-Large-Instruct-2407. In **bold**: best overall accuracy results per language.

words, and the example selection is based on ranking sentences by the number of errors the model produced during the inference on the development set. The explicit word and sentence accuracy results for each of the prompt types are detailed in Appendix A.

4.2 Experiments with Corpora of Different Distributions

In the second part of our experiments, we aim to determine the best strategy for performing the lemmatization task on data from a distribution different from the one seen during training. We take the experiments by Toporkov and Agerri (2024) as a starting point, where they show that lemmatizers (based on fine-tuned encoder-only models) substantially worsen when evaluated out-of-domain, their most common use case. Therefore, in this setting, we compare the performance of encoder models fine-tuned on gold-annotated corpora and applied out-of-domain against the direct lemma generation based on in-context learning using LLMs. For this purpose, we employ datasets and models described in Section 3. We fine-tune XLM-RoBERTa large for each of the 6 languages of our selection in a token classification task, where the model learns to predict labels corresponding to the minimum number of edits required to transform the target word into its lemma. We use 16 as a batch size, 0.01 weight decay, 5e-5 learning rate, and 20 epochs as hyperparameters. For the in-context lemma generation, we choose the best prompt setting from the previous experimental step and apply it using the models described in Section 3.2.

Table 3 reports the results. We mark in bold the best overall result for each language and with an asterisk those results that are statistically significant according to McNemar's test. As previously mentioned in Section 4, each result corresponds to the average over 3 runs. The standard deviation for both word and sentence accuracy across these runs is consistently low (≤ 0.02), indicating stable behavior of the models.

We could see that directly generating the lemmas with *Claude* and *Mistral* outperforms the fine-tuned encoder in Turkish, Czech, and Russian. For English, the results across models are very close, although they are only statistically significant for EWT corpus. For Basque and Spanish XLM-RoBERTa large evaluated out-of-domain is the superior option. Among LLMs, the highest accuracy for 6 out of 12 corpora is achieved using Claude-3.7-Sonnet, while Mistral-Large-Instruct-2407 ranks a close second for Spanish, English, and Russian. It is worth noticing that the open-weights Mistral significantly outperforms XLM-RoBERTA large in 6 of the 11 evaluation settings.

4.3 Experiments with Parallel Corpora

To address RQ3, we perform a set of experiments using cross-lingual transfer and in-context lemma generation. For both model- and data-transfer, we assume our source language to be English. For *model-transfer*, we train a contextual lemmatizer

	Mistral-LI-2407		Llama-3.3-70B		Qwen-2.5-72B		Claude-3.7		XLM-R large	
language_corpus	*Wacc*	*SentAcc*	*Wacc*	*SentAcc*	*Wacc*	*SentAcc*	*Wacc*	*SentAcc*	*Wacc*	*SentAcc*
eu_bdt [2]	0.84	0.12	0.78	0.05	0.71	0.02	**0.89**	**0.20**	-	-
eu_armiarma	0.83	0.08	0.75	0.04	0.70	0.02	0.88	0.15	**0.89***	**0.18***
es_ancora	0.93	0.23	0.87	0.16	0.92	0.24	0.94	0.32	**0.97***	**0.51***
es_gsd	0.93	0.25	0.87	0.15	0.91	0.21	0.94	0.32	**0.96***	**0.43***
en_ewt	**0.93***	**0.44***	0.92	0.39	0.92	0.42	0.93	0.33	0.92	0.39
en_gum	**0.94**	**0.47**	0.92	0.40	0.93	0.46	0.94	0.37	0.94	0.43
tr_imst	0.90	0.44	0.86	0.31	0.84	0.31	**0.94***	**0.60***	0.81	0.19
tr_pud (sigm'19)	0.81	0.06	0.80	0.04	0.77	0.01	0.84	0.08	**0.85**	**0.08**
cs_cac	0.94	0.39	0.87	0.21	0.89	0.19	**0.97***	**0.55***	0.93	0.32
cs_pud (sigm'19)	0.95	0.34	0.91	0.24	0.89	0.15	**0.97***	**0.55***	0.95	0.43
ru_gsd	0.94	0.41	0.87	0.26	0.92	0.29	**0.96***	**0.51***	0.94	0.39
ru_syntagrus	0.95	0.43	0.91	0.34	0.93	0.35	**0.96***	**0.48***	0.94	0.41
average	0.91	0.32	0.87	0.23	0.87	0.24	**0.93**	**0.39**	0.92	0.34

Table 3: Word and sentence accuracy results comparing in-context lemma generation against XLM-RoBERTa large performance on lemmatization fine-tuned on a different data distribution. In **bold**: best overall accuracy results across models for each language. *:statistically significant results at $\alpha = .05$.

on English data using XLM-RoBERTa large in a token classification task, applying the same hyper-parameters as in the experiments of the previous section. The *data-transfer* method, implemented as translate-train, requires the automatic generation of the training data for each of the target languages. We apply Claude-3.7-Sonnet to translate the PUD training set to the 11 target languages and to obtain the lemmas directly from the translations. We then fine-tune XLM-RoBERTa large using the generated training set.

Table 4 displays the results of the experiments conducted with the parallel corpora. We mark in bold the best overall result per language and underline the best model for the language per each subset of models (decoder-only vs. encoder-only models). We observe that the latest generation of LLMs demonstrates strong results in the lemma generation task, and the models' performance remains stable across consecutive runs, as in the previous experimental setup (standard deviation ≤ 0.02). It should be emphasized that we explore the performance of these models using in-context learning, without them being specifically fine-tuned for lemmatization. Among the 4 models of our choice, Claude-3.7-Sonnet exhibits the highest accuracy for all the languages except English, outperforming Mistral-Large-Instruct-2407, LLaMA-3.3-70B-Instruct, and Qwen-2.5-72B-Instruct. Still, the open-weights Mistral-Large-Instruct-2407 stays a close second in this setting, while the other two models demonstrate lower capabilities for the task.

Cross-lingual transfer obtains less competitive results, suggesting that, in data scarcity scenarios, directly generating lemmas with LLMs is a very effective method, far superior to using traditional model-transfer or data-transfer approaches (at least for the set of languages considered).

Surprisingly, direct lemma generation using *Claude* or *Mistral* is superior to even fine-tuning XLM-RoBERTA large in-domain (e.g., monolingual results) for 7 out of the 12 languages, although for English, Swedish, Italian, French, and Turkish XLM-RoBERTA remains the best option. It can also be observed that significant differences in word and sentence accuracy are demonstrated only for Turkish, Swedish, and English, while for French and Italian the distinction could be perceived only by comparing sentence accuracy results. This highlights the importance of using sentence accuracy as an alternative metric to word accuracy (Manning, 2011; Toporkov and Agerri, 2024).

Finally, results obtained with the data-transfer approach exceed those of direct in-context lemma generation using *LLaMa* or *Qwen* for 6 out of 11 languages, indicating the high quality of translations produced by *Claude*.

To perform all the experiments we utilized several NVIDIA A100 80GB GPUs and the total computational resources amounted to approximately 300 hours of processing time and 75 kWh of energy consumption, resulting in 32.4 kg of CO2 emissions.

[2] Since the Armiarma dataset lacks training data, we could not perform out-of-domain experiments on the BDT test set, and thus results for XLM-R large model are not reported.

	Mistral-LI-2407		Llama-3.3-70B		Qwen-2.5-72B		Claude-3-7		monolingual		model-transfer		data-transfer	
language	*Wacc*	*SentAcc*	*Wacc*	*SentAcc*	*Wacc*	*SentAcc*	*Wacc*	*SentAcc*	*Wacc*	*SentAcc*	*Wacc*	*SentAcc*	*Wacc*	*SentAcc*
en	0.96	0.44	0.94	0.35	0.95	0.42	0.94	0.30	**0.97**	**0.58**	-	-	-	-
de	0.96	0.51	0.93	0.25	0.93	0.31	**0.97**	**0.64**	0.95	0.37	0.64	0.00	0.95	0.39
is	0.90	0.13	0.84	0.10	0.80	0.01	**0.94**	**0.30**	0.89	0.09	0.73	0.00	0.87	0.03
sv	0.94	0.34	0.91	0.21	0.88	0.12	0.95	0.36	**0.96**	**0.48**	0.76	0.00	0.92	0.26
ru	0.95	0.40	0.93	0.36	0.93	0.30	**0.96**	**0.49**	0.94	0.27	0.76	0.00	0.92	0.21
cs	0.95	0.37	0.91	0.23	0.95	0.24	**0.97**	**0.65**	0.95	0.45	0.80	0.00	0.94	0.35
pl	0.95	0.42	0.92	0.22	0.90	0.17	**0.95**	**0.46**	0.94	0.36	0.75	0.00	0.91	0.18
es	0.97	0.45	0.94	0.31	0.96	0.40	**0.98**	**0.65**	0.97	0.55	0.78	0.00	0.96	0.38
it	0.96	0.37	0.92	0.25	0.94	0.26	0.97	0.44	**0.97**	**0.55**	0.78	0.00	0.95	0.34
fr	0.97	0.51	0.94	0.26	0.95	0.24	0.98	0.42	**0.98**	**0.59**	0.78	0.00	0.96	0.36
fi	0.89	0.15	0.84	0.09	0.80	0.03	**0.93**	**0.35**	0.86	0.06	0.77	0.00	0.82	0.05
tr	0.81	0.03	0.79	0.02	0.78	0.01	0.85	0.06	**0.88**	**0.16**	0.75	0.00	0.82	0.03
average	0.93	0.34**	0.90	0.22	0.90	0.21	**0.95**	**0.43**	**0.94**	**0.38**	0.75	0.00	0.91	0.23

Table 4: Word and sentence accuracy across various approaches to lemmatization task using parallel corpora. In **bold**: best overall accuracy per language. Underlined: best model for the language for each subset of models.

5 Discussion

Although LLMs show promising results in in-context lemma generation for the selected languages, their performance depends heavily on the provided examples. In prompt development across 4 languages of varied morphological complexity, the differences between the basic in a zero-shot setting and the 4-shot prompt were quite significant. We identified a number of problems when using LLMs for direct lemma generation, such as: (i) *randomly generated output.* LLMs such as Mistral-Large-Instruct-2407 and Claude-3.7-Sonnet tend to randomly generate quotation marks, but there are also cases, exhibited for English (EWT corpus), where Claude-3.7-Sonnet provided the output for the same sentence more than once. In the case of LLaMA-3.3-70B-Instruct, the model gives additional explanations, for instance, in the case of ambiguous Basque auxiliary verbs such as *edun* and *izan*. (ii) *modified wordforms.* Despite the explicit instructions on not performing any changes to the initial word, even if it is misspelled, LLMs tend to ignore them (this was the case during the experimental prompting phase with Mistral-Large-Instruct-2407). Apart from that, it is common to lowercase the input words at the beginning of the sentence. (iii) *arbitrarily skipping words.* This was the most interesting observation, as there was no perceivable pattern of which words the model was skipping. For instance, for Basque Mistral-Large-Instruct-2407 ignored verbal forms such as 'egingo' (*'will do'* in English); (iv) *struggling with certain lemmas that do not appear in the few-shot examples*. This case was observed for articles in Spanish for which, instead of providing the lemma for definite articles such as '*el*', LLMs were returning the same word form of the determiner given in the input text (e.g., *la*, *los* or *las*). This behaviour could be changed by adding an example in the prompt to deal with these particular cases.

Model scale also plays a crucial role in the quality of in-context lemmatization. Table 5 presents comparative results for the smaller and larger versions of the *Qwen* and *Mistral* models, indicating that larger models demonstrate significantly stronger performance on the lemmatization task. Additionally, we experimented with LLaMa-3.1-8B, but the results were difficult to report, as the model failed to generate a coherent output for 10 out of 12 languages, performing a lot of hallucinations, incorrect output format and numerous sentence repetitions.

These are only some of the most common observations obtained during the analysis of LLMs' performance on contextual lemmatization. The quality of examples used in the prompt design plays an important role, and careful and elaborated example selection may be specifically beneficial for low-resource languages. A deeper analysis should be conducted regarding other potential pitfalls of LLMs for this task, especially for languages with more complex morphology.

6 Conclusions

In this paper, we present the first empirical analysis of the ability of the latest-generation LLMs to perform in-context lemma generation. Our results suggest that, although fine-tuning encoders such as XLM-RoBERTa large on gold data remains a competitive option for its use out-of-domain, large size LLMs still reach results close to the state-of-the-art by directly generating lemmas in-context

language	Qwen-2.5-7B		Qwen-2.5-32B		Qwen-2.5-72B		Ministral-8B		Mistral-LI-2407	
	Wacc	*SentAcc*	*Wacc*	*SentAcc*	*Wacc*	*SentAcc*	*Wacc*	*SentAcc*	*Wacc*	*SentAcc*
en	0.81	0.08	0.90	0.26	0.95	0.42	0.85	0.09	0.96	0.44
de	0.72	0.00	0.90	0.16	0.93	0.31	0.77	0.01	0.96	0.51
is	0.62	0.00	0.77	0.01	0.80	0.01	0.71	0.01	0.90	0.13
sv	0.75	0.00	0.86	0.06	0.88	0.12	0.80	0.02	0.94	0.34
ru	0.89	0.13	0.93	0.25	0.93	0.30	0.82	0.16	0.95	0.40
cs	0.78	0.03	0.87	0.07	0.95	0.24	0.81	0.03	0.95	0.37
pl	0.75	0.02	0.88	0.11	0.90	0.17	0.80	0.00	0.95	0.42
es	0.78	0.01	0.93	0.23	0.96	0.40	0.81	0.06	0.97	0.45
it	0.76	0.00	0.92	0.17	0.94	0.26	0.82	0.04	0.96	0.37
fr	0.70	0.00	0.92	0.19	0.95	0.24	0.83	0.03	0.97	0.51
fi	0.65	0.00	0.78	0.03	0.80	0.03	0.72	0.01	0.89	0.15
tr	0.66	0.00	0.76	0.03	0.78	0.01	0.71	0.00	0.81	0.03
average	0.74	0.02	0.87	0.13	0.90	0.21	0.79	0.04	0.93	0.34

Table 5: Word and sentence accuracy for in-context lemma generation across different model sizes using parallel corpora.

in a few-shot setting. We also investigate the scenario in which no training data is available for a given language. Comparing model-transfer, data-transfer, and direct lemma generation with LLMs, we conclude that the best lemmatization approach in such a case would also be direct in-context lemma generation, which would remain predominantly achievable with large-scale language models. Finally, future work should include model comparison, broader linguistic sampling, and comprehensive prompt optimization.

Lemmatization is a task generally studied and evaluated in-domain, which is rather surprising as the use of lemmatizers is predominantly out-of-domain (Toporkov and Agerri, 2024). Our work demonstrates, for the first time, the potential to perform direct lemmatization directly without any training data by applying direct in-context learning with large size LLMs, even for high-inflected relatively low-resource languages.

Limitations

Several limitations constrain this investigation. First, the comprehensive evaluation across the full spectrum of large-scale language models remains unexplored, potentially excluding architectures that may demonstrate superior lemmatization capabilities. Second, the scarcity of available evaluation datasets, particularly for low-resource and morphologically complex languages, limits our ability to conduct extensive cross-linguistic validation. Third, we did not systematically explore alternative prompt variations or instructional formats.

Acknowledgments

This work has been partially supported by the Basque Government (Research group funding IT-1805-22). We are also thankful to the following MCIN/AEI/10.13039/501100011033 projects: (i) DeepKnowledge (PID2021-127777OB-C21) and by FEDER, EU; (ii) DeepMinor (CNS2023-144375) and European Union NextGenerationEU/PRTR.

References

Rodrigo Agerri, Josu Bermudez, and German Rigau. 2014. IXA pipeline: Efficient and ready to use multilingual NLP tools. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3823–3828, Reykjavik, Iceland. European Language Resources Association (ELRA).

Rodrigo Agerri and German Rigau. 2016. Robust multilingual named entity recognition with shallow semi-supervised features. *Artificial Intelligence*, 238(2):63–82.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

Iñaki Alegria, Xabier Artola, Kepa Sarasola, and Miriam Urkia. 1996. Automatic morphological analysis of basque. *Literary and Linguistic Computing*, 11:193–203.

Dang Anh, Limor Raviv, and Lukas Galke. 2024. Morphology matters: Probing the cross-linguistic morphological generalization abilities of large language models through a wug test. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 177–188, Bangkok, Thailand. Association for Computational Linguistics.

Anthropic. 2025. Claude 3.7 sonnet and claude code. https://www.anthropic.com/news/claude-3-7-sonnet/.

Armiarma. 2000. Armiarma corpus. https://armiarma.eus/. SUSA-LITERATURA, PID.

Catherine Arnett and Benjamin Bergen. 2025. Why do language models perform worse for morphologically complex languages? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6607–6623, Abu Dhabi, UAE. Association for Computational Linguistics.

Toms Bergmanis and Sharon Goldwater. 2018. Context sensitive neural lemmatization with Lematus. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1391–1400, New Orleans, Louisiana. Association for Computational Linguistics.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *Preprint*, arXiv:2303.12712.

Yang Chen, Chao Jiang, Alan Ritter, and Wei Xu. 2023. Frustratingly easy label projection for cross-lingual transfer. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5775–5796, Toronto, Canada. Association for Computational Linguistics.

Grzegorz Chrupala, Georgiana Dinu, and Josef van Genabith. 2008. Learning morphology with Morfette. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Thomas G. Dietterich. 1998. Approximate statistical test for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923.

Hao Fei, Meishan Zhang, and Donghong Ji. 2020. Cross-lingual semantic role labeling with high-quality translated training corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7014–7026, Online. Association for Computational Linguistics.

Daniel Fernandes, João P. Matos-Carvalho, Carlos M. Fernandes, and Nuno Fachada. 2025. Deepseek-v3, gpt-4, phi-4, and llama-3.3 generate correct code for lorawan-related engineering tasks. *Electronics*, 14(7):1428.

Blanca Calvo Figueras, Eneko Sagarzazu, Julen Etxaniz, Jeremy Barnes, Pablo Gamallo, Iria De Dios Flores, and Rodrigo Agerri. 2025. Truth knows no language: Evaluating truthfulness beyond english. *Preprint*, arXiv:2502.09387.

Iker García-Ferrero, Rodrigo Agerri, and German Rigau. 2022. Model and data transfer for cross-lingual sequence labelling in zero-resource settings. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6403–6416, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, and 89 others. 2024a. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024b. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Andrea Gesmundo and Tanja Samardžić. 2012. Lemmatisation as a tagging task. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 368–372, Jeju Island, Korea. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Mika Hämäläinen. 2024. Dag: Dictionary-augmented generation for disambiguation of sentences in endangered uralic languages using chatgpt. *Preprint*, arXiv:2411.01531.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Bart Jongejan and Hercules Dalianis. 2009. Automatic training of lemmatization rules that handle morphological changes in pre-, in-and suffixes alike. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 145–153.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Lauri Karttunen, Ronald M. Kaplan, and Annie Zaenen. 1992. Two-level morphology with composition. In *COLING 1992 Volume 1: The 14th International Conference on Computational Linguistics*.

Dan Kondratyuk. 2019. Cross-lingual lemmatization and morphology tagging with two-stage multilingual BERT fine-tuning. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 12–18, Florence, Italy. Association for Computational Linguistics.

Chaitanya Malaviya, Shijie Wu, and Ryan Cotterell. 2019. A simple joint model for improved contextual neural lemmatization. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1517–1528, Minneapolis, Minnesota. Association for Computational Linguistics.

Christopher D. Manning. 2011. Part-of-speech tagging from 97% to 100%: Is it time for some linguistics? In *Computational Linguistics and Intelligent Text Processing*, pages 171–189, Berlin, Heidelberg. Springer.

Arya D. McCarthy, Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, Timofey Arkhangelskiy, Nataly Krizhanovsky, Andrew Krizhanovsky, Elena Klyachko, Alexey Sorokin, John Mansfield, Valts Ernštreits, Yuval Pinter, Cassandra L. Jacobs, and 3 others. 2020. UniMorph 3.0: Universal Morphology. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3922–3931, Marseille, France. European Language Resources Association.

Arya D. McCarthy, Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Christo Kirov, Miikka Silfverberg, Sabrina J. Mielke, Jeffrey Heinz, Ryan Cotterell, and Mans Hulden. 2019. The SIGMORPHON 2019 shared task: Morphological analysis in context and cross-lingual transfer for inflection. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–244, Florence, Italy. Association for Computational Linguistics.

AI Mistral. 2024. Mistral-large-instruct-2407. Accessed: 2025-05-17.

Thomas Müller, Ryan Cotterell, Alexander Fraser, and Hinrich Schütze. 2015. Joint lemmatization and morphological tagging with lemming. In *Proceedings of*

the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2268–2274, Lisbon, Portugal. Association for Computational Linguistics.

Joakim Nivre, Daniel Zeman, Filip Ginter, and Francis Tyers. 2017. Universal Dependencies. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, Valencia, Spain. Association for Computational Linguistics.

Kemal Oflazer. 1993. Two-level description of Turkish morphology. In *Sixth Conference of the European Chapter of the Association for Computational Linguistics*, Utrecht, The Netherlands. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. 2024. Gollie: Annotation guidelines improve zero-shot information-extraction. In *ICLR*.

Ilya Segalovich. 2003. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. In *MLMTA*.

Uygun Shadikhodjaev and Jae Sung Lee. 2019. CBNU system for SIGMORPHON 2019 shared task 2: a pipeline model. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 19–24, Florence, Italy. Association for Computational Linguistics.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2022. Language models are multilingual chain-of-thought reasoners. *Preprint*, arXiv:2210.03057.

Milan Straka, Jana Straková, and Jan Hajic. 2019. UDPipe at SIGMORPHON 2019: Contextualized embeddings, regularization with morphological categories, corpora merging. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 95–103, Florence, Italy. Association for Computational Linguistics.

Ekaterina Sviridova, Anar Yeginbergen, Ainara Estarrona, Elena Cabrio, Serena Villata, and Rodrigo Agerri. 2024. CasiMedicos-arg: A medical question answering dataset annotated with explanatory argumentative structures. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18463–18475, Miami, Florida, USA. Association for Computational Linguistics.

Olia Toporkov and Rodrigo Agerri. 2024. On the role of morphological information for contextual lemmatization. *Computational Linguistics*, 50(1):157–191.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30. Curran Associates, Inc.

Haifeng Wang, Jiwei Li, Hua Wu, Eduard Hovy, and Yu Sun. 2023. Pre-trained language models and their applications. *Engineering*, 25:51–65.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, and 16 others. 2022. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Leonie Weissweiler, Valentin Hofmann, Anjali Kantharuban, Anna Cai, Ritam Dutt, Amey Hengle, Anubha Kabra, Atharva Kulkarni, Abhishek Vijayakumar, Haofei Yu, Hinrich Schuetze, Kemal Oflazer, and David Mortensen. 2023. Counting the bugs in ChatGPT's wugs: A multilingual investigation into the morphological capabilities of a large language model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6508–6524, Singapore. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Anar Yeginbergen, Maite Oronoz, and Rodrigo Agerri. 2024. Argument mining in data scarce settings: Cross-lingual transfer and few-shot techniques. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11687–11699, Bangkok, Thailand. Association for Computational Linguistics.

Eray Yildiz and A. Cüneyd Tantuğ. 2019. Morpheus: A neural network for jointly learning contextual lemmatization and morphological tagging. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 25–34, Florence, Italy. Association for Computational Linguistics.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, and 43 others. 2017. CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.

A Appendix

Example of the full prompt with a 1-shot example, where the input sentence is introduced as a list of words.

Your task is to lemmatize a sentence in Spanish. You will be given a sentence, where each word starts from the new line. You need to provide for each word in the given sentence its dictionary form (lemma). For example, for the sentence:
Tina
Anselmi
se
ocupo
sobre
todo
de
los
derechos
de
los
trabajadores
textiles
y
los
profesores
.
The desired output is:
Tina Tina
Anselmi Anselmi
se el
ocupó ocupar
sobre sobre
todo todo
de de
los el
derechos derecho
de de
los el
trabajadores trabajador
textiles textil
y y
los el
profesores profesor
. .
*Provide the output in **TSV format** (Tab-Separated Values) with the format: 'initial word lemma'*
Sentence:
['El', 'festival', 'de', 'Venecia', 'cerró', 'hoy', 'con', 'la', 'entrega', 'de', 'los', 'premios', 'que', 'coronaron', 'a', 'el', 'realizador', 'Alexander', 'Sokurov', 'y', 'a', 'el' 'actor', 'Michael', 'Fassbender', '.']
Answer with the required output only, without extra spaces, quotation marks, or comments.

			Sentence input					Wordform input				
Language	Prompt Type	Shots	WAcc	SentAcc	Missing	Wrong word	Random	WAcc	SentAcc	Missing	Wrong word	Random
Spanish (PUD)	basic prompt	0-shot	0.92	0.12	32	2	2	0.93	0.22	0	0	9
	full prompt	0-shot	0.93	0.22	4	3	6	0.92	0.17	1	0	6
	basic prompt + worst examples	4-shot	0.96	0.37	3	0	3	0.96	0.38	1	0	0
	basic prompt + random examples	4-shot	0.97	0.50	5	0	0	0.94	0.33	0	0	0
	basic prompt + most errors	4-shot	0.96	0.41	2	0	0	0.97	0.45	1	0	0
	full prompt + worst examples	4-shot	0.97	0.41	3	0	3	0.97	0.40	0	0	0
	full prompt + random examples	4-shot	0.96	0.49	1	0	0	0.95	0.41	0	0	0
	full prompt + most errors	4-shot	0.96	0.49	1	0	0	0.95	0.41	0	0	0
English (PUD)	basic prompt	0-shot	0.93	0.28	30	0	0	0.95	0.38	0	0	6
	full prompt	0-shot	0.95	0.43	4	0	1	0.95	0.42	2	0	12
	basic prompt + worst examples	4-shot	0.96	0.50	0	0	0	0.95	0.36	0	0	0
	basic prompt + random examples	4-shot	0.95	0.47	0	0	0	0.96	0.45	0	0	0
	basic prompt + most errors	4-shot	0.96	0.42	0	0	0	0.96	0.44	0	0	4
	full prompt + worst examples	4-shot	0.96	0.46	0	0	0	0.95	0.41	0	0	0
	full prompt + random examples	4-shot	0.96	0.50	0	0	0	0.96	0.42	0	0	0
	full prompt + most errors	4-shot	0.96	0.46	0	0	0	0.95	0.43	0	0	0
Russian (PUD)	basic prompt	0-shot	0.91	0.20	32	5	8	0.92	0.29	0	2	6
	full prompt	0-shot	0.94	0.33	4	3	3	0.94	0.34	0	4	4
	basic prompt + worst examples	4-shot	0.94	0.39	3	0	0	0.95	0.39	2	0	0
	basic prompt + random examples	4-shot	0.94	0.39	4	0	1	0.95	0.40	0	0	0
	basic prompt + most errors	4-shot	0.93	0.41	42	0	6	0.95	0.40	2	0	0
	full prompt + worst examples	4-shot	0.94	0.40	3	0	0	0.95	0.36	2	0	0
	full prompt + random examples	4-shot	0.94	0.38	5	0	1	0.95	0.39	2	0	0
	full prompt + most errors	4-shot	0.93	0.37	42	0	0	0.94	0.37	1	0	0
Basque (BDT, 100 sentences)	basic prompt	0-shot	0.76	0.00	78	3	58	0.82	0.07	3	0	72
	full prompt	0-shot	0.78	0.03	44	3	44	0.81	0.05	1	0	68
	basic prompt + worst wacc	4-shot	0.85	0.11	23	0	2	0.89	0.24	0	0	2
	basic prompt + random examples	4-shot	0.87	0.19	38	0	2	0.89	0.24	0	0	2
	basic prompt + most errors	4-shot	0.86	0.23	38	0	2	0.89	0.29	7	0	2
	full prompt + worst examples	4-shot	0.84	0.12	20	0	2	0.89	0.25	1	0	2
	full prompt + random examples	4-shot	0.87	0.22	39	0	2	0.88	0.22	0	0	2
	full prompt + most errors	4-shot	0.86	0.25	34	0	0	0.89	0.27	1	0	2

Table 6: Word and sentence accuracy using different prompting strategies with Mistral-Large-Instruct-2407.