Guiding Large Language Models with Their Own Linguistic Knowledge for Morphological Inflection

Yifei Zhou[1], Yves Lepage[2]

[1] Graduate School of Information, Production, and Systems, Waseda University
2-7 Hibikino, Wakamatsu-ku, Kitakyushu-shi, 808-0135 Fukuoka-ken, Japan
yifei.zhou@ruri.waseda.jp
[2] Graduate School of Information, Production, and Systems, Waseda University
2-7 Hibikino, Wakamatsu-ku, Kitakyushu-shi, 808-0135 Fukuoka-ken, Japan
yves.lepage@waseda.jp

Abstract

Large language models excel at many natural language processing tasks but often struggle with morphological inflection, especially for low-resource languages with morphologically complex systems. We propose a two-stage prompting approach that integrates explicit linguistic knowledge into the generation process. In the first stage, the model produces inflectional rules for a target language and part of speech. In the second stage, these rules are used as linguistic knowledge to guide the production of inflected forms. Evaluations on Polish and Lower Sorbian from the CoNLL–SIGMORPHON 2017 shared task show that our method improves zero-shot accuracy across model sizes, Gemma 3 12B gains 23% for Polish and 8% for Lower Sorbian, and Gemma 3 27B gains 14% and 13%, respectively, narrowing the gap with supervised baselines. Additional experiments reveal that the benefit of linguistic knowledge decreases as the number of in-context examples increases, combining it with 10-shot prompting of Gemma 3 27B achieves the best performance, matching supervised baselines trained on much larger labeled datasets. These results highlight the effectiveness of incorporating explicit linguistic knowledge into LLM prompting, particularly for morphologically rich, low-resource languages.

Keywords: large language models, linguistic knowledge, morphological inflection

1. Introduction

Large language models (LLMs) have shown strong performance in generating fluent, human-like text, making them powerful tools for tasks such as summarization, dialogue generation, and machine translation (Bang et al., 2023; Chang et al., 2024). However, their performance in morphological inflection often lags behind that of task-specific models, particularly for languages with rich inflectional systems or those underrepresented in training data (Weissweiler et al., 2023).

As morphology interacts with syntax, semantics, and phonology (Haspelmath and Sims, 2010), it is difficult for models to learn from raw text alone. While large-scale training enables LLMs to acquire certain morphological capabilities, such as following frequent or regular patterns in high-resource languages (Weissweiler et al., 2023; Mortensen et al., 2024), extending these abilities to previously unseen languages remains difficult. This challenge becomes more pronounced in morphologically complex languages, where the range of possible inflected forms far exceeds what LLMs are exposed to during training.

Our hypothesis is that LLMs already encapsulate implicit linguistic knowledge, which can be made explicit and used to improve morphological inflection. Rather than treating LLMs as uninterpretable models, we prompt them to first produce general morphological descriptions and then apply these rules to generate inflected forms. Figure 1 illustrates this two-stage prompting process. By making the models' implicit knowledge explicit, our goal is to enhance their morphological ability to produce accurate inflected forms.

We evaluate this approach on two Slavic languages, Polish (pol) and Lower Sorbian (dsb), both of which have highly complex morphological systems. Using Gemma 3 (Gemma Team, 2025), we assess whether prompting with explicit linguistic knowledge improves inflection accuracy compared to standard prompting strategies. We also examine the effect of in-context learning by varying the number of examples included in the prompt.

Our experiments reveal two findings. First, incorporating explicit linguistic knowledge significantly improves accuracy over standard prompting in the zero-shot setting for both languages. However, its relative advantage diminishes as more in-context examples are added. Second, combining linguistic knowledge with in-context examples allows LLMs to match the performance of supervised baselines trained on much larger labeled data. For Gemma 3 27B, 10-shot prompting with linguistic knowledge reaches the accuracy of baselines trained on 10,000 examples for Polish and 1,000 for Lower Sorbian. These results underscore the value of integrating explicit linguistic knowledge into LLM prompting for morphological inflection.

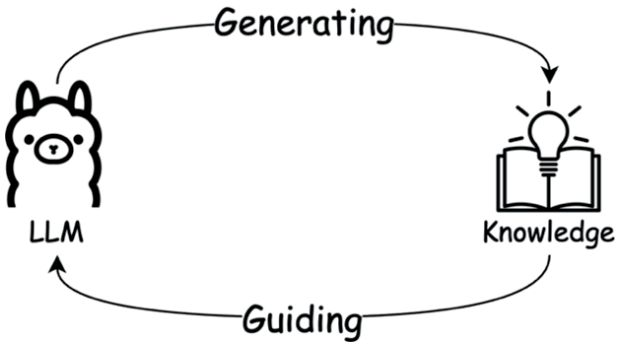

Fig. 1: Overview of our two-stage prompting process.

Yifei Zhou, Yves Lepage, *Guiding Large Language Models with Their Own Linguistic Knowledge for Morphological Inflection*. In: Z. Vetulani, P. Paroubek, I. Kardava (Eds.), Human Language Technologies as a Challenge for Computer Science and Linguistics – 2025, pp. 105–109. Poznań: Adam Mickiewicz University Press, 2025.
© Authors, 2025. https://doi.org/10.14746/amup.9788323245414.19. Open Access chapter, distributed under the terms of the CC licence (BY-NC-ND, https://creativecommons.org/licenses/by-nc-nd/4.0/).

Lemma	MSFs	Target form
robić	Verb; Present; First person; Singular	robię
robić	Verb; Past; First person; Singular; Masculine	robiłem
robić	Verb; Conditional; Second person; Singular; Feminine	robiłabyś
robić	Verb; Imperative-Jussive; First person; Plural	róbmy
robić	Participle; Past; Active; Singular; Neuter	robiące
robić	Participle; Past; Passive; Singular; Feminine	robiona

Table 1. Example of the morphological inflection task, illustrated with the lemma *robić* ('to do/make') in Polish. In this task, the lemma *robić* together with MSFs serve as the input, and the inflected forms (Target form) are the output.

2. Related Work

2.1. Morphological Inflection

Morphological inflection is the process of generating the correct word form from a lemma given its morphosyntactic features (MSFs), such as case, number, gender, or tense. Table 1 illustrates this task using the Polish lemma *robić*. A single lemma can yield dozens of distinct forms in morphologically rich languages. For example, a Polish verb may have nearly 100 forms. The SIGMORPHON shared tasks [1] provide benchmarks for evaluating inflection systems across hundreds of languages.

Early approaches relied on rule-based methods or finite-state transducers, which required human effort to manually create rules (Forsberg and Hulden, 2016). With the shift to neural methods, inflection is solved as a character-level sequence-to-sequence problem, mapping a lemma and its features to the correct inflected form. LSTM-based or Transformer-based models have achieved over 80% accuracy across many languages (Canby et al., 2020), leading some to consider the task as largely solved. These models are trained specifically for inflection on curated datasets, leaving the capabilities of LLMs in this domain underexplored.

2.2. Morphological Abilities in LLMs

Recent studies have begun to examine how LLMs handle morphological phenomena. One line of work uses the Wug Test (Jean, 1958) to probe their ability to generate inflected or derived forms for novel or nonce words. In these tests, the model is given a sentence with a blank and must produce the correct word form based on context (Weissweiler et al., 2023; Anh et al., 2024). Results suggest that LLMs can follow some regular inflectional patterns, especially in high-resource languages.

Another line of research investigates morphological segmentation, which involves splitting words into their constituent morphemes. Unlike common subword methods in LLM training, such as Byte Pair Encoding (BPE), morphological segmentation aligns more closely with linguistic boundaries (Marco and Fraser, 2024). Studies show that prompting strategies can enhance morphological segmentation abilities in LLMs (Truong et al., 2024).

Despite these insights, prior work treats LLMs as uninterpretable models, assessing performance without using explicit linguistic knowledge. Our approach differs in that we prompt LLMs to first generate morphological rules and then apply these rules to produce inflected forms. This is conceptually similar to chain-of-thought prompting (Wei et al., 2022), which improves performance on complex tasks through step-by-step reasoning. By externalizing implicit linguistic knowledge, we provide explicit guidance for morphological inflection.

3. Proposed Method

We propose a two-stage prompting approach that integrates explicit linguistic knowledge to improve LLM performance in morphological inflection. In the first stage, the model generates general morphological rules for a given language and part of speech. In the second stage, these rules guide the production of correctly inflected forms.

3.1. Knowledge Elicitation Prompt

The first stage prompts the LLM to generate inflectional rules for a specific language and part of speech, such as declension patterns for nouns and adjectives or conjugation patterns for verbs. For example, when eliciting conjugation rules for Polish verbs, the prompt is constructed as follows:

> *Generate the {conjugation} rules for {Polish} {verbs}.*

The prompt asks the model to describe relevant patterns without providing specific lemmas or MSFs. The model might respond with a pattern such as:

> *Regular Polish verbs form the present tense singular with endings that vary by verb type: 1st person -ę, 2nd person -esz/-isz, and 3rd person -e/-i.*

These rules are not manually edited for correctness and are generated independently of the input data used in the next stage. Each rule serves as a reusable piece of linguistic knowledge, explicitly guiding the model during morphological inflection.

3.2. Morphological Inflection Prompt

The second stage prompts the model to perform morphological inflection. The prompt combines three elements: the linguistic knowledge elicited in the first stage, a small set of in-context examples for few-shot settings, and the test lemma with its MSFs. Figure 2 illustrates this prompting format. All prompts are written in English and

[1] https://sigmorphon.github.io/

avoid language-specific formatting, making the method applicable to different LLMs and target languages.

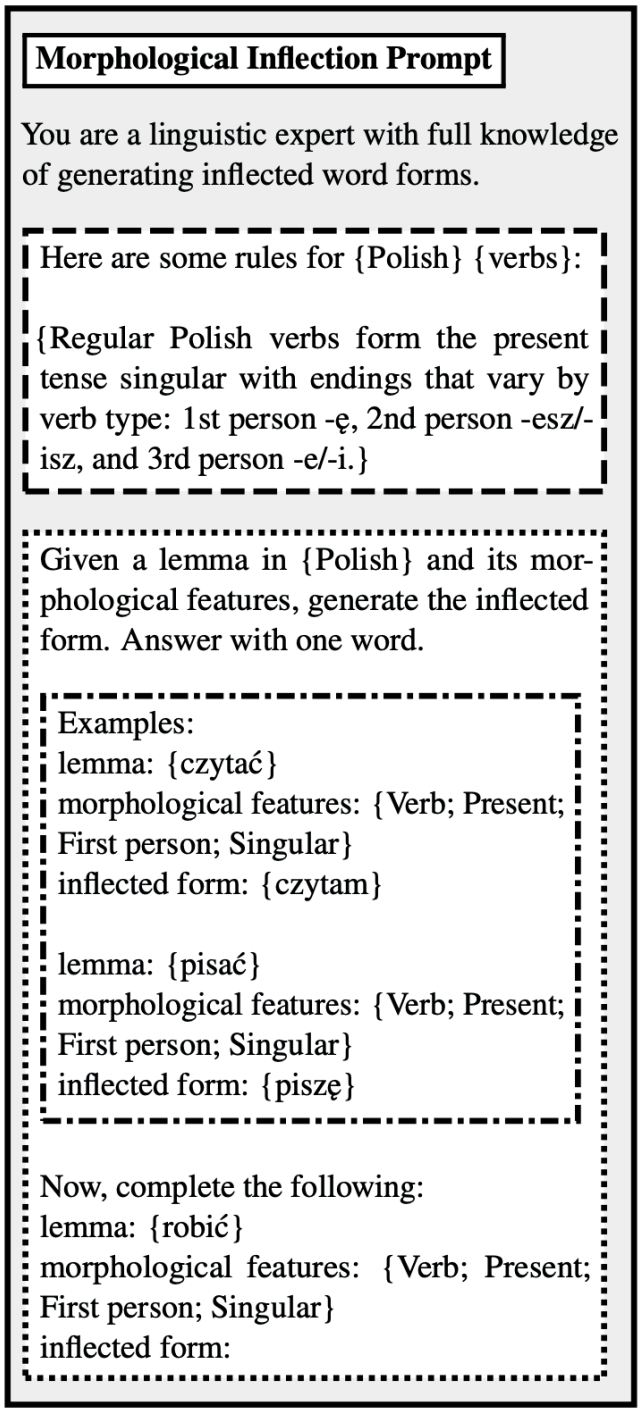

Fig. 2: Example of a morphological inflection prompt for the Polish lemma *robić* with features *Verb; Present; First person; Singular*. The dashed box contains the elicited linguistic knowledge, the dotted box shows the task description, and the dash-dotted box presents the in-context examples.

4. Experiments

4.1. Datasets

We use data from the CoNLL–SIGMORPHON 2017 shared task 1: universal morphological (re)inflection[2] (Cotterell et al., 2017), a widely adopted benchmark for morphological inflection. We evaluate our approach on two morphologically rich Slavic languages: Polish and Lower Sorbian. For each language, the dataset is split into training (three regimes: *high*—10,000 instances, *medium*—1,000, *low*—100), development (1,000), and test (1,000) sets. Each instance consists of a lemma, a set of MSFs, and the corresponding inflected form. Importantly, we do not use the training data to fine-tune the LLM. All LLM evaluations are based on prompting.

4.2. Models

We include the official CoNLL–SIGMORPHON 2017 system[3] as a supervised baseline, trained under three data regimes (*low*, *medium*, *high*) to provide a reference for comparing LLM performance with task-specific models. We use the open-source Gemma 3 as the base LLM, in 12B and 27B parameter configurations, implemented via Ollama[4]. To evaluate the effect of incorporating linguistic knowledge, we compare two prompting conditions: standard prompting (*w/o knowledge*) and prompting with linguistic knowledge (*w/ knowledge*). All prompts and hyper-parameters are applied consistently across both languages and conditions.

4.3. Evaluation Metric

Performance is measured using accuracy, defined as the percentage of predicted forms that exactly match the gold-standard inflections. Higher accuracy indicates better performance.

5. Results

5.1. Effect of Linguistic Knowledge

We first compare Gemma 3 prompted with and without linguistic knowledge in the zero-shot setting, alongside the supervised baseline trained under three data regimes (Table 2). When prompted without linguistic knowledge, LLMs with larger model sizes achieve stronger zero-shot performance but still fall short of the supervised baseline. For instance, zero-shot Gemma 3 27B achieves 58% accuracy on Polish, surpassing the supervised baseline trained on 100 examples, while its accuracy on Lower Sorbian is only 26%, which is below the 100-example baseline. In contrast, zero-shot Gemma 3 12B reaches only 40% on Polish and 20% on Lower Sorbian, both falling short of the 100-example supervised baselines. This suggests that pretraining captures some patterns in relatively high-resource languages like Polish but offers limited benefits for truly low-resource languages like Lower Sorbian, where exposure during training is minimal.

[2] https://github.com/sigmorphon/conll2017

[3] https://github.com/sigmorphon/conll2017/tree/master/baseline

[4] https://ollama.com/

Integrating linguistic knowledge into zero-shot prompting yields gains. For Gemma 3 12B, zero-shot accuracy increases by 23% for Polish and by 8% for Lower Sorbian. For Gemma 3 27B, zero-shot accuracy rises by 14 percentage for Polish, nearly matching the supervised baseline trained on 1,000 examples, and by 13% for Lower Sorbian, surpassing the 100-example baseline. These results show that explicitly providing linguistic knowledge can narrow the gap between zero-shot LLM performance and supervised systems. Without fine-tuning LLMs on task-specific labeled data, incorporating such knowledge into prompts can substantially improve morphological performance in both languages.

Model	pol	dsb
Baseline		
low	41.30	33.80
medium	74.00	70.80
high	88.00	86.40
Gemma 3 12B		
w/o knowledge	39.70	20.29
w/ knowledge	62.60	28.05
Gemma 3 27B		
w/o knowledge	58.39	26.23
w/ knowledge	72.03	39.33

Table 2. Zero-shot accuracy (%) of Gemma 3, with and without linguistic knowledge, compared to supervised baselines trained under three data regimes (*low*, *medium*, *high*).

5.2. Effect of In-Context Examples

We next investigate the impact of increasing the number of in-context examples from 0 to 10. Table 3 shows Gemma 3 performance grouped by model size, while Figure 3 illustrates the trends grouped by language and including supervised baselines. For both prompting strategies, with and without linguistic knowledge, accuracy improves as more examples are provided, but the relative advantage becomes smaller. This reflects the diminishing returns of in-context learning.

Across all example counts, incorporating linguistic knowledge achieves higher accuracy than using in-context examples alone. The advantage is most pronounced in the zero-shot and one-shot settings, where explicit rules provide guidance in the absence of contextual demonstrations. As the number of examples approaches 10, this gap narrows, suggesting that sufficient examples can compensate for the lack of linguistic knowledge.

In all settings, the best performance is achieved by combining linguistic knowledge with 10-shot prompting of Gemma 3 27B, matching the accuracy of supervised baselines trained on much larger datasets: 10,000 labeled examples for Polish and 1,000 for Lower Sorbian. This demonstrates that prompting LLMs with explicit linguistic knowledge and few-shot examples can achieve performance comparable to task-specific supervised baselines, highlighting the potential of applying this approach across different languages and models.

6. Conclusion

We introduced a two-stage prompting strategy that incorporates explicit linguistic knowledge into LLMs for morphological inflection. By first eliciting general inflectional rules and then using them as guidance, our method enables LLMs to generate more accurate and interpretable inflected forms.

Experiments on Polish and Lower Sorbian show that linguistic knowledge improves zero-shot and few-shot performance. Increasing the number of in-context examples further enhances accuracy. The best results, obtained by combining linguistic knowledge with 10-shot prompting of Gemma 3 27B, surpass the performance of supervised baselines trained on much larger datasets. This shows that prompting LLMs with explicit linguistic knowledge and few-shot examples can compensate for the lack of labeled data, even in morphologically complex or low-resource languages.

These findings highlight the potential of integrating explicit linguistic knowledge into LLM prompting, particularly for morphologically complex, low-resource languages. Future work could extend the approach to additional languages and combine linguistic knowledge with retrieval-augmented generation or chain-of-thought prompting.

Model	Lang	Linguistic knowledge	Number of examples				
			0-shot	1-shot	3-shot	5-shot	10-shot
Gemma 3 12B	pol	*w/o*	39.70±3.10	70.00±3.15	76.80±2.68	78.86±2.68	81.96±2.53
		w/	62.60±3.10	77.88±2.73	81.90±2.32	82.45±2.53	84.26±2.37
	dsb	*w/o*	20.29±2.45	41.28±3.10	54.25±2.95	57.45±3.05	60.87±2.85
		w/	28.05±2.75	48.41±3.20	56.83±3.05	60.21±3.15	63.56±2.90
Gemma 3 27B	pol	*w/o*	58.39±3.15	78.81±2.58	84.88±2.17	86.76±2.12	86.94±2.17
		w/	72.03±2.89	82.36±2.27	87.34±2.12	88.09±2.17	88.68±2.01
	dsb	*w/o*	26.23±2.75	54.93±3.00	64.75±2.85	67.19±3.00	71.71±2.65
		w/	39.33±2.95	57.77±2.95	67.86±2.85	68.41±3.00	73.16±2.80

Table 3. Accuracy (%) of Gemma 3 as the number of in-context examples increases from 0 to 10, with and without linguistic knowledge, grouped by model size.

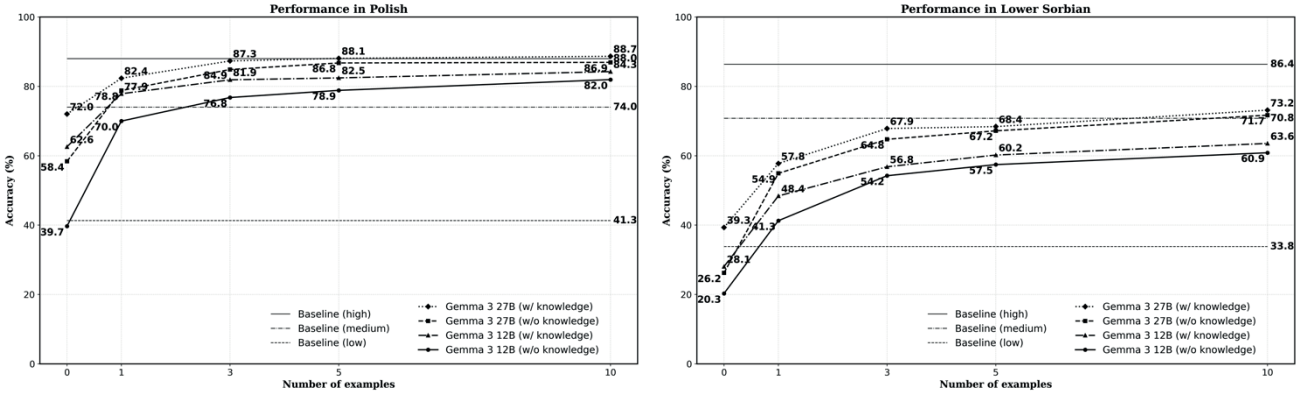

Fig. 3: Gemma 3 performance as the number of in-context examples increases from 0 to 10, with and without linguistic knowledge, grouped by language and including supervised baselines. Left panel: Polish; right panel: Lower Sorbian.

References

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.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wen- liang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In Proceedings of the 13th International Joint Conference on Natural Lan- guage Processing and the 3rd Conference of the Asia- Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pages 675–718, Nusa Dua, Bali. Association for Computational Lin- guistics.

Marc Canby, Aidana Karipbayeva, Bryan Lunt, Sahand Mozaffari, Charlotte Yoder, and Julia Hockenmaier. 2020. University of Illinois submission to the SIGMORPHON 2020 shared task 0: Typologically di- verse morphological inflection. In Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 137–145, Online. Association for Computa- tional Linguistics.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language mod- els. ACM Trans. Intell. Syst. Technol., 15(3).

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2017. CoNLL-SIGMORPHON 2017 shared task: Universal morphological reinflection in 52 languages. In Proceed- ings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection, pages 1–30, Vancouver. Association for Computational Linguis- tics.

Markus Forsberg and Mans Hulden. 2016. Learning transducer models for morphological analysis from example inflections. In Proceedings of the SIGFSM Workshop on Statistical NLP and Weighted Automata, pages 42–50, Berlin, Germany. Association for Computational Linguistics.

Gemma Team. 2025. Gemma 3 technical report. Martin Haspelmath and Andrea Sims. 2010. Understanding Morphology. Routledge, Oxford, UK.

Berko Jean. 1958. The child's learning of english morphology. WORD, 14(2-3):150–177.

Marion Di Marco and Alexander Fraser. 2024. Subword segmentation in LLMs: Looking at inflection and consistency. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 12050–12060, Miami, Florida, USA. Association for Computational Linguistics.

David R. Mortensen, Valentina Izrailevitch, Yunze Xiao, Hinrich Schütze, and Leonie Weissweiler. 2024. Verbing weirds language (models): Evaluation of English zero-derivation in five LLMs. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 17359–17364, Torino, Italia. ELRA and ICCL.

Thinh Truong, Yulia Otmakhova, Karin Verspoor, Trevor Cohn, and Timothy Baldwin. 2024. Revisiting subword tokenization: A case study on affixal negation in large language models. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 5082–5095, Mexico City, Mexico. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

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.