Decomposed Prompting: Probing Multilingual Linguistic Structure Knowledge in Large Language Models

Anonymous ACL submission

Abstract

Probing the multilingual knowledge of linguistic structure in LLMs, often characterized as sequence labeling, faces challenges with maintaining output templates in current text-to-text prompting strategies. To solve this, we introduce a *decomposed prompting* approach for sequence labeling tasks. Diverging from the single text-to-text prompt, our prompt method generates for each token of the input sentence an individual prompt which asks for its linguistic label. We test our method on the Universal Dependencies part-of-speech tagging dataset for 38 languages, using both English-centric and multilingual LLMs. Our findings show that *decomposed prompting* surpasses the *iterative prompting* baseline in efficacy and efficiency under zero- and few-shot settings. Moreover, our analysis of multilingual performance of English-centric LLMs yields insights into the transferability of linguistic knowledge via multilingual prompting.

1 Introduction

Current Large Language Models (LLMs), such as GPT-3, GPT-4, PaLM, and LLaMA (Brown et al., 2020; Chowdhery et al., 2023; Touvron et al., 2023a), have demonstrated remarkable capabilities in in-context learning across a broad spectrum of language understanding and generation tasks (Zhao et al., 2023; Zhang et al., 2023; Ziyu et al., 2023). These models are predominantly trained on massive amounts of English text data, with some limited exposure to other languages. For instance, LLaMA2's pretraining corpus comprises over 89% English content, with the rest in other languages or code (Touvron et al., 2023b). Yet, these English-centric LLMs [1] exhibit effective performance in complex multilingual language understanding tasks (Deng et al., 2023; Wang et al., 2023). In multilingual evaluation with prompting, a model performs tasks by directly generating outputs based on a task description and/or a few examples provided in a pivot language (typically English), along with input in a different target language (Ahuja et al., 2023). Despite the remarkable multilingual performance of LLMs, the extent and nature of their cross-lingual capabilities remain underexplored (Ye et al., 2023).

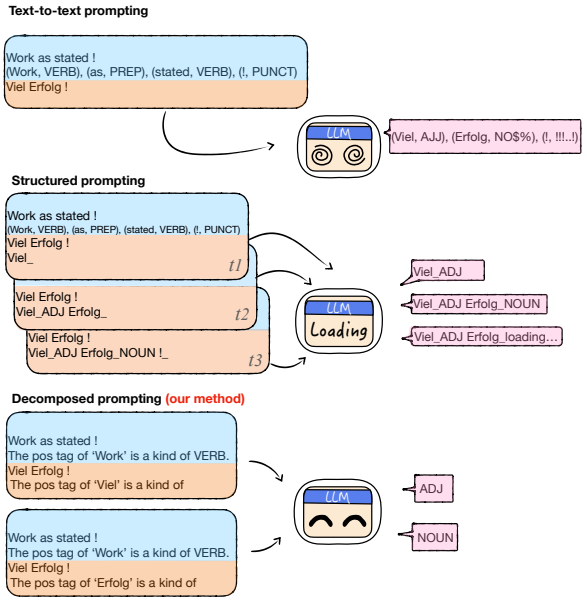

Figure 1: Comparison of different prompting methods for sequence labeling.

We hypothesize that these models harbor substantial multilingual knowledge. This knowledge, particularly relating to linguistic structure, is commonly conceptualized through sequence tagging tasks (Jurafsky, 2000). However, the current prompting strategies designed for sequence labeling in LLMs are not well suited to test. For instance, behavioral probing methods (Belinkov et al., 2020), aimed at measuring knowledge stored in language models, struggle to adapt to tasks predicting more complex structures. To overcome the challenges in

[1] In this paper, we regard a model pretrained primarily on English text as English-centric.

probing the multilingual knowledge of linguistic structure in LLMs characterized as sequence labeling, drawing inspiration from the token-level prompt-based fine-tuning method by Ma et al. (2024), we introduce the ***decomposed prompting*** strategy, aiming to probe English-centric LLMs for their understanding of linguistic structure framed as sequence labeling tasks. As shown in Figure 1, instead of employing a single text-to-text prompt for labeling an entire sequence in one step, our method decomposes this process into multiple discrete prompts. More precisely, we first split the input sentence into tokens. Subsequently, we generate an individual prompt for each token which inquires about its linguistic label.

We evaluate our approach on the Universal Dependency (UD) part-of-speech (POS) tagging dataset (Nivre et al., 2020) covering 38 languages with 3 English-centric LLMs and 2 multilingual LLMs. Our approach outperforms the iterative prompting baseline in both zero- and few-shot settings in terms of accuracy and efficiency. Furthermore, our investigation into the multilingual performance of English-centric LLMs offers valuable insights into their capabilities of transferring linguistic knowledge through multilingual prompting.

2 Background and Related Work

Multilinguality of English-Centric LLMs English-centric LLMs are primarily pretrained on large English text data, with a limited exposion to multilingual data. LLaMA (Touvron et al., 2023a), for example, is pretrained on an extensive scale of corpora comprising over 1.4 trillion tokens, of which less than 4.5% constitute multilingual data from 20 different languages. LLaMA 2 (Touvron et al., 2023b) expands this linguistic diversity, featuring 27 languages each representing more than 0.005% of the pertaining data. Mistral 7B (Jiang et al., 2023) achieves superior performance and efficiency through the adoption of advanced attention techniques such as Sliding Window Attention (SWA) (Child et al., 2019), facilitating faster inference. To enhance the robustness of multilingual processing, the Byte-level Byte-Pair-Encoding (BBPE) algorithm (Sennrich et al., 2016; Wang et al., 2020) is commonly used for tokenization in LLMs. This approach is able to decompose UTF-8 characters, which are outside the scope of the model vocabulary, into their constituent bytes. Thus, BBPE tokenization equips LLMs with the versatility to handle scripts from any language, theoretically, even those not encountered during training. In summary, *limited exposure to non-English data* and *byte-level encoding capability*, these two factors discussed above, jointly contribute to the robust multilingual abilities observed in English-centric LLMs.

Prompting for Sequence Labeling Prompting LLMs for sequence labeling tasks remains a challenge (Ahuja et al., 2023). While *text-to-text prompting* is widely adopted across various benchmarking tasks for LLMs (Lai et al., 2023), their application to sequence labeling is hindered by the challenges in maintaining the output templates (Asai et al., 2023). In response, a decent *iterative prompting* strategy for structured prediction has been introduced (Blevins et al., 2023) (Figure 1). In this approach, the model decodes in step t_i a label for the word at position t_i of the sequence. This predicted label, along with the next word, is then input back into the model to predict the next label. However, the dependency of each token's prediction on the preceding one substantially slows down the inference process. In contrast, our proposed *decomposed prompting* method offers improvements in both efficacy and efficiency. Our method is similar to Ma et al. (2024) in that both methods decompose an input sentence into a series of prompts; however, their method is used for fine-tuning, while our method is in an in-context learning paradigm without training.

3 Decomposed Prompting for LLMs

Given a test sequence set $\mathcal{X}_{test}$, a label set L, and an LLM M, we approach the task of sequence labeling as follows: for an input sequence $X \in \mathcal{X}_{test}$ of length n, $X = x_1, \cdots, x_n$, the model M is expected to produce a corresponding sequence of labels $\hat{Y} = \hat{y}_1, \cdots, \hat{y}_n$, where each label $\hat{y}_i \in L$ is associated with the linguistic feature of the token x_i.

In decomposed prompting, we design a prompt template function $T(\cdot, \cdot)$ which generates a specific prompt for each token. T takes the input sequence X and an individual token x_i as arguments and returns a prompt for predicting the label of the token. The true label y_i can be optionally included as an argument to T; if included, T utilizes y_i to provide a demonstration.

$C = c_1, \cdots, c_m$ is a sample from the training set. In the few-shot learning scenario, k examples

in the tuple format (C_j, c_j, l_j) are given along with the input sequence X, where c_j is a token in C_j, and $l_j \in L$ is the label for c_j. The demonstration D of an input sequence X is formulated as:

$$D = I \circ T(C_1, c_1, l_1) \circ \cdots \circ T(C_k, c_k, l_k) \quad (1)$$

where I denotes an optional instruction in natural language, $\circ$ denotes the string concatenation operation. Finally, we use a prompt generator function $G(\cdot, \cdot)$ to create the set of decomposed prompts for an input sequence X:

$$G(X, D) = \{D \circ T(X, x_1), \cdots, D \circ T(X, x_m)\} \quad (2)$$

The label $\hat{y}_i$ of token x_i is predicted as follows:

$$\hat{y}_i = \underset{y \in L}{\operatorname{argmax}} P_M(l | D \circ T(X, x_i)) \quad (3)$$

For each possible label y, we obtain the probability that the model predicts this label as the next token and select the most likely label as the predicted label.

4 Experiment and Results

4.1 Experimental Setup

Dataset and Language We use a subset of the Universal Dependency treebanks (UDPOS) (Nivre et al., 2020) to probe the multilingual linguistic knowledge of LLMs. The UDPOS dataset adopts a universal POS tag set consisting of 17 tags (Appendix A.1.1). Our chosen subset, derived from the XTREME multilingual benchmark (Hu et al., 2020), comprises 38 languages from diverse language families distributions (Appendix A.1.2). We randomly sample 200 instances of each language for the evaluation.

Model and Setup We experiment on three English-centric LLMs: LLaMA2-7B, LLaMA2-13B (Touvron et al., 2023b), and Mistral-7B (Jiang et al., 2023), as well as two multilingual LLMs: BLOOMZ-7B (Muennighoff et al., 2023) and mTk-Instruct (Wang et al., 2022). All LLMs in our experiment are instruction-tuned versions accessible through the HuggingFace framework (Wolf et al., 2020). We use the weighted average F1 scores for different tags as our evaluation metric. All experiments were conducted on a server with 4 `A100-SXM4-80GB` GPUs. More details of experimental settings are described in Appendix A.2.

Model	Method	Zero-shot		Few-shot		Avg.
		en	mult.	en	mult.	
`LLaMA2-7B`	*Iter*	33.1	27.2	68.0	48.6	44.2
	Decom	58.2	43.2	74.7	50.5	56.7
`LLaMA2-13B`	*Iter*	47.6	37.4	68.0	52.6	51.4
	Decom	**67.3**	54.7	77.3	54.5	63.5
`Mistral-7B`	*Iter*	65.2	54.3	80.2	58.9	64.7
	Decom	63.6	**61.8**	**85.0**	**64.4**	**68.7**
`BLOOMZ-7B`	*Decom*	20.6	17.6	44.1	36.2	29.6
`mTK-Instruct`	*Decom*	47.6	43.1	57.3	44.7	48.2

Table 1: Overall results of iterative and decomposed prompting methods on POS tagging tasks in zero- and few-shot settings, with F1 score reported. **en** indicates the results for English, and **mult.** represents the average F1 score across other 37 languages. The best performance of each column is highlighted in **bold**.

4.2 Overall Results

We evaluate the performance of *iterative prompting*, the baseline method, and *decomposed prompting*, our proposed method, for English and multilingual POS tag labeling tasks under zero- and few-shot settings. The few-shot examples and the prompts employed in our experiment are presented in Appendix B.2 for reference. Our preliminary experiment to explore the influence of the number of few-shot samples (k) reveals a mild impact on performance once k increases to around 10. More details are provided in Appendix C.

Superiority in Efficacy The overall results for English-centric LLMs, as detailed in Table 1, demonstrate that our proposed decomposed prompting obviously outperforms the iterative prompting baseline across both zero- and few-shot settings, in both English and multilingual evaluations. This trend holds true for all three English-centric models tested, with the sole exception in the zero-shot setting for the English evaluation with the Mistral-7B model, where *Decom* slightly lags behind *Iter* (63.6 vs. 65.2). In addition, English-centric LLMs outperform multilingual LLMs by a considerable margin. The complete experimental results are displayed in Appendix D.

	BLOOMZ-7B	LLaMA2-7B	Mistral-7B	Avg.
zero-shot	3.2×	2.5×	1.4×	2.4×
few-shot	9.2×	7.9×	3.1×	6.7×

Table 2: The ratio by which the inference is accelerated for *Decom* promoting compared to *Iter* prompting. The inference speed was measured over the entire test set.

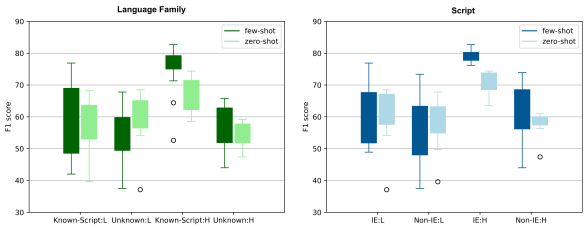

Figure 2: Analysis of decomposed promoting performance grouped by language family (a) and script type (b) under zero- and few-shot settings on Mistral. "IE" refers to the Indo-European language family. "L" (Low) represents languages that constitute less than 0.005% of the pretraining corpus, while "H" (High) denotes all other languages.

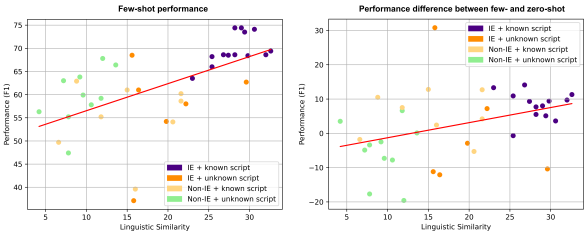

Figure 3: Panorama of Mistral model's per-language performance. Each node symbolizes a distinct language. (a) shows the few-shot performance and (b) shows the difference between few- and zero-shot performance for each language.

Superiority in Efficiency In addition to superior performance, *decomposed prompting* offers enhanced efficiency during inference, especially in few-shot prompting. As demonstrated in Table 2, our proposed method achieves, on average, a 2.4-fold increase in speed compared to the baseline in the zero-shot prompting setting and a 6.7-fold increase in the few-shot setting. The efficiency advantage is less obvious with Mistral, owing to Mistral's implementation of a modified attention mechanism designed to enhance inference efficiency.

5 Multilingual Analysis

Figure 2 provides a stratified view of decomposed prompting performance by language family and script, under both zero- and few-shot settings on the Mistral model. The results indicate that Indo-European languages generally achieve higher F1 scores compared to their non-Indo-European counterparts. Notably, the presence of few-shot examples consistently improves the overall performance across all categories, but the box plot also shows that some languages are negatively impacted by the use of English demonstrations. As discussed in §2, English-centric LLMs are adept at tokenizing words from Latin or Cyrillic scripts into subtokens. For scripts less familiar to these models, they often default to breaking down the text into UTF-8 encodings, which may lead to suboptimal representations for languages using these less common scripts. Thus, to capture a more nuanced understanding of LLM performance across linguistic varieties, we categorize languages not only by family but also by script type. Figure 2(b) illustrates that, in both few-shot and zero-shot settings, languages with known scripts tend to yield better performance than unknown scripts. An exception to this trend is observed among the language group with smaller corpora in the zero-shot setting.

To further understand the impact of English demonstrations on languages with varied properties in multilingual prompting, we delve deeper into the cross-lingual transferability of English-centric LLMs and conduct a detailed analysis of individual language performance. We begin by quantifying the linguistic proximity of each tested language to English. This was achieved by calculating the cosine similarity between language vectors (Littell et al., 2017) that incorporate syntactic, phylogenetic, and geographic attributes, among others, following Nie et al. (2023) and Ma et al. (2023). Further information on the computation of language similarity is available in Appendix A.3. From Figure 3, we observe that the performance gain from few-shot prompting is more substantial for languages that are linguistically closer to English, as indicated by the upward trend on the right side of the plot. Remarkably, languages distant from English may even experience a decline in performance when using English demonstrations.

6 Conclusion

In conclusion, we propose *decomposed prompting*, a simple yet effective prompting method specially designed for sequence labeling tasks, addressing the difficulties of LLM benchmarking on sequence labeling tasks. Our method outperforms iterative prompting techniques in terms of accuracy and efficiency in different experimental settings. By applying *decomposed prompting* to UDPOS dataset, we probe the multilingual linguistic structure knowledge of English-centric LLMs. Our multilingual investigation reveals that gain from few-shot decomposed prompting is generally more pronounced for languages closer to English.

Limitations

Although our proposed *decomposed prompting* method achieves overall remarkable performance in terms of both accuracy and efficiency, it has limitations for some special cases, for example, it can not well handle the case where the same word occurs twice in a sentence with different POS tags. Besides, the efficiency of decomposed prompting suffers as the length of the input sequence and the complexity of the task increase. Our study uses decomposed prompting methods for part-of-speech (POS) tagging as a means to evaluate the multilingual structural knowledge of English-centric Large Language Models (LLMs). This provides a foundational assessment of the models' capabilities. Nevertheless, extending the application scope of this methodology to probe more intricate aspects of linguistic structure is necessary. Future research could beneficially apply decomposed prompting to the analysis of complex linguistic phenomena, including sentence chunking and named entity recognition, to gain a deeper understanding of the nuanced capabilities of LLMs in processing and understanding language.

References

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.

Akari Asai, Sneha Kudugunta, Xinyan Velocity Yu, Terra Blevins, Hila Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi. 2023. Buffet: Benchmarking large language models for few-shot cross-lingual transfer. *arXiv preprint arXiv:2305.14857*.

Yonatan Belinkov, Sebastian Gehrmann, and Ellie Pavlick. 2020. Interpretability and analysis in neural NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 1–5, Online. Association for Computational Linguistics.

Terra Blevins, Hila Gonen, and Luke Zettlemoyer. 2023. Prompting language models for linguistic structure. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6649–6663, Toronto, Canada. 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, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. 2023. Multilingual jailbreak challenges in large language models. *arXiv preprint arXiv:2310.06474*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Dan Jurafsky. 2000. *Speech & language processing*. Pearson Education India.

Viet Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Nguyen. 2023. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

Bolei Ma, Ercong Nie, Helmut Schmid, and Hinrich Schuetze. 2023. Is prompt-based finetuning always better than vanilla finetuning? insights from cross-lingual language understanding. In *Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023)*, pages 1–16, Ingolstadt, Germany. Association for Computational Lingustics.

Bolei Ma, Ercong Nie, Shuzhou Yuan, Helmut Schmid, Michael Färber, Frauke Kreuter, and Hinrich Schuetze. 2024. ToPro: Token-level prompt decomposition for cross-lingual sequence labeling tasks. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2685–2702, St. Julian's, Malta. Association for Computational Linguistics.

Chaitanya Malaviya, Graham Neubig, and Patrick Littell. 2017. Learning language representations for typology prediction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2529–2535, Copenhagen, Denmark. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Ercong Nie, Sheng Liang, Helmut Schmid, and Hinrich Schütze. 2023. Cross-lingual retrieval augmented prompt for low-resource languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8320–8340, Toronto, Canada. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Changhan Wang, Kyunghyun Cho, and Jiatao Gu. 2020. Neural machine translation with byte-level subwords.

Wenxuan Wang, Zhaopeng Tu, Chang Chen, Youliang Yuan, Jen-tse Huang, Wenxiang Jiao, and Michael R Lyu. 2023. All languages matter: On the multilingual safety of large language models. *arXiv preprint arXiv:2310.00905*.

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, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jiacheng Ye, Xijia Tao, and Lingpeng Kong. 2023. Language versatilists vs. specialists: An empirical revisiting on multilingual transfer ability. *arXiv preprint arXiv:2306.06688*.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Zhuang Ziyu, Chen Qiguang, Ma Longxuan, Li Mingda, Han Yi, Qian Yushan, Bai Haopeng, Zhang Weinan,

and Ting Liu. 2023. Through the lens of core competency: Survey on evaluation of large language models. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 2: Frontier Forum)*, pages 88–109, Harbin, China. Chinese Information Processing Society of China.

A Experimental Setup Details

Details of the experimental setup are introduced in this section.

A.1 Dataset and Languages

A.1.1 POS Tag Set

Figure 4 shows the pos tag set in UD. We also use the text in the box as the task instruction in our experiments.

```
POS tag set: ADJ ADP ADV AUX CCONJ DET INTJ NOUN
NUM PART PRON PROPN PUNCT SCONJ SYM VERB X
```

Figure 4: UD POS tag set.

A.1.2 Profile of Languages

As Figure 5 shows, our experiment involves 38 languages with diverse language family distributions.

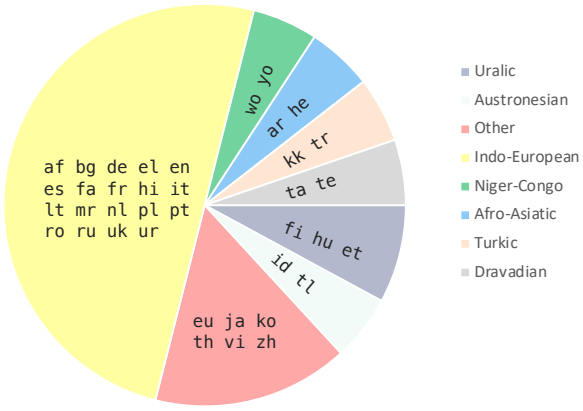

Figure 5: Distribution of languages by language family in the dataset.

A.2 Baselines and Settings

Iterative Prompting (*Iter*) Blevins et al. (2023) introduced a structured prompting approach that *iteratively* labels an entire sentence by appending each predicted label to the context along with the subsequent word. This method is employed as a strong baseline in our study.

Decomposed Prompting (*Decom*) To evaluate our proposed approach, we employ the prompt template outlined in §3 to decompose the entire sequence into a set of individual prompts for prediction. In our experiments, we use the 17 POS tags themselves as the label words, i.e., we expect the model to directly predict a tag from the tagset shown in tagset by selecting the tag with the highest logit.

Zero- and Few-Shot Prompting We devised two experimental scenarios for multilingual prompting—zero-shot and few-shot—to evaluate the performance of both approaches under different conditions. In the zero-shot setting, only an English *task instruction* is provided alongside the input in the target language. The text in Figure 4, which outlines the tag set information, serves as the instruction in our experiments. In few-shot prompting, we supplement the prompt with a few English demonstrations, structured according to the prompt template of each method. For *Decom*, we randomly select an example for each tag type from the English training set to create a demonstration. For a fair comparison, the same number of demonstrations are used for the *Iter* baseline.

A.3 Language Similarity Computation

Malaviya et al. (2017) and Littell et al. (2017) proposed LANG2VEC, language vectors to represent various linguistic features for languages. A language can be represented by five vectors, containing syntactic, phonological, phonetic, phylogenetic, and geographical features, respectively. Linguistic similarities among different languages with respect to these linguistic features can be calculated through the cosine similarity. In our study, we utilized the language vectors provided by LANG2VEC to calculate the cosine similarity between target languages and English. We used a rank-based similarity score to average the rank of languages in each feature dimension. Table 3 illustrates the computation details.

B More Details of Decomposed Prompting Method

B.1 Intuition

This method draws inspiration from the step-by-step thinking process humans employ when annotating linguistic features within a sentence. Typically, humans approach such tasks incrementally, addressing each token individually. Mirroring this

intuitive strategy, our method first decomposes an input sentence into tokens. Subsequently, we generate a distinct prompt for each token, thereby transforming the sequence labeling task into a series of focused, manageable prompts. Figure 6 illustrates the generation of sequence labeling prompts for the German sentence *"Viel Erfolg!"* via *decomposed prompting*.

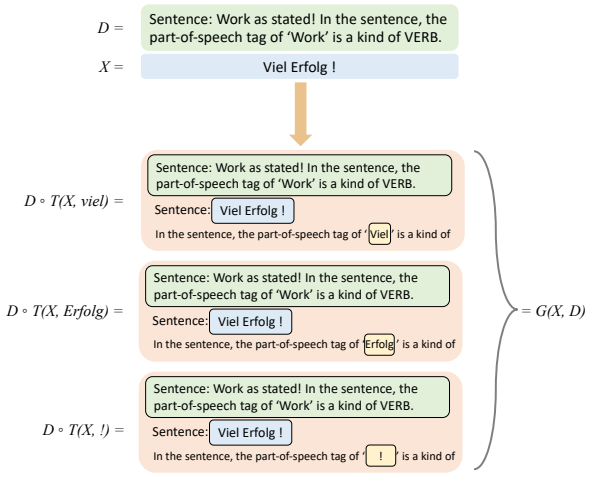

Figure 6: An example of how *decomposed prompting* is implemented for sequence labeling.

Figure 6 illustrates the generation of sequence labeling prompts for the German sentence *"Viel Erfolg!"* via *decomposed prompting*.

An example of a template function is illustrated as follows.

> $T(X, x_i)$ = "Sentence: X. In the sentence, the part-of-speech tag of 'x_i' is a kind of"
> $T(X, x_i, y_i)$ = "Sentence: X. In the sentence, the part-of-speech tag of 'x_i' is a kind of y_i."

B.2 Prompt Details

Zero- and few-shot prompts used in this work are shown in Figure 8 (decomposed prompting) and Figure 9 (iterative prompting).

C Few-Shot Ablation Study

we investigate the impact of the number of few-shot examples on the performance in the decomposed prompting. We randomly select 8 languages (en, de, el, fa, hi, hl, ru, zh) and explore their performance dynamics with the increasing of the few-shot samples. Figure 7 shows that overall, when k is small, increasing the number of samples bring performance improvement. As k continues to increase, the performance tends to be stable and even gets worse when samples are too many.

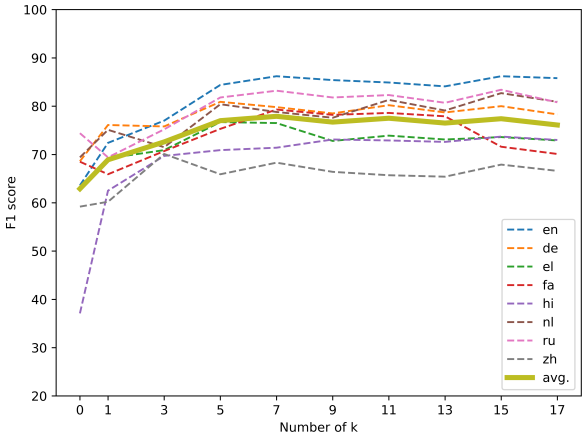

Figure 7: Performance dynamics with different numbers of few-shot samples. Experimental results of decomposed prompting with Mistral-7B.

D Full Results

Full experimental results are displayed in Table 4 (Mistral 7B), Table 5 (LLaMA2 7B), Table 6 (LLaMA 13B), Table 7 (BLOOMZ 7B), Table 8 (mTk 13B), and Table 9 (few-shot ablation study).

Zero-shot prompt

POS tag set: ADJ ADP ADV AUX CCONJ DET INTJ NOUN NUM PART PRON PROPN PUNCT SCONJ SYM VERB X

Sentence: Viel Erfolg !

In the sentence, the part-of-speech tag of ‘Viel’ is a kind of

Few-shot prompt (w/o Instruction)

Sentence: And if you send me a story , that would be great !

In the sentence, the part-of-speech tag of ‘if’ is a kind of SCONJ.

Sentence: I ‘ll admit I was n’t expecting much from this place , but they really did do a good job .

In the sentence, the part-of-speech tag of ‘good’ is a kind of ADJ.

Sentence: I do n’t know . The girl shrugged once again . In the sentence, the part-of-speech tag of ‘girl’ is a kind of NOUN.

Sentence: The dancers were falling back round a Polish agriculturalist who was teaching a gangling Englishman and two young Africans an Eastern European peasant dance .

In the sentence, the part-of-speech tag of ‘around’ is a kind of ADP.

Sentence: Antigua was awesome .

In the sentence, the part-of-speech tag of ‘was’ is a kind of AUX.

Sentence: The food is fresh and taste great .

In the sentence, the part-of-speech tag of ‘the’ is a kind of DET. Sentence: Now I have wife and son .

In the sentence, the part-of-speech tag of ‘Now’ is a kind of ADV.

Sentence: However , this fruitful period was short-lived , as Greece suffered badly under the Ottoman Empire , only to recover in the 19th century as the capital of independent Greece .

In the sentence, the part-of-speech tag of ‘suffered’ is a kind of VERB.

Sentence: I survived it without a problem .

In the sentence, the part-of-speech tag of ‘.’ is a kind of PUNCT. Sentence: The food is fresh and taste great .

In the sentence, the part-of-speech tag of ‘and’ is a kind of CCONJ.

Sentence: you can view at dresscod.com

In the sentence, the part-of-speech tag of ‘dresscod.com’ is a kind of X.

Sentence: I do n’t know . The girl shrugged once again .

In the sentence, the part-of-speech tag of ‘I’ is a kind of PRON.

Sentence: I ‘ll admit I was n’t expecting much from this place , but they really did do a good job .

In the sentence, the part-of-speech tag of ‘n’t’ is a kind of PART.

Sentence: Antigua was awesome .

In the sentence, the part-of-speech tag of ‘Antigua’ is a kind of PROPN.

Sentence: The dancers were falling back round a Polish agriculturalist who was teaching a gangling Englishman and two young Africans an Eastern European peasant dance .

In the sentence, the part-of-speech tag of ‘two’ is a kind of NUM. Sentence: Yes , the Cyclone is almost certain to lose strength as it surges over land .

In the sentence, the part-of-speech tag of ‘Yes’ is a kind of INTJ.

Sentence: −−== Posted via Newsfeed.Com - Unlimited - Uncensored - Secure Usenet News ==−−

In the sentence, the part-of-speech tag of ‘−−== ‘ is a kind of SYM.

Sentence: Viel Erfolg ! In the sentence, the part-of-speech tag of ‘Viel’ is a kind of

Figure 8: Prompt design of decomposed prompting.

Zero-shot prompt

POS tag set: ADJ ADP ADV AUX CCONJ DET INTJ NOUN NUM PART PRON PROPN PUNCT SCONJ SYM VERB X

Sentence: Viel Erfolg !

Viel_

Few-shot prompt (w/o Instruction)

Context: Chahine said her immediate family spent about $ 20,000 to return to Detroit via Syria and Jordan .

Tagged: Chahine_PROPN said_VERB her_PRON immediate_ADJ family_NOUN spent_VERB about_ADV $_SYM 20,000_NUM to_PART return_VERB to_ADP Detroit_PROPN via_ADP Syria_PROPN and_CCONJ Jordan_PROPN ._PUNCT

Context: Welcome Darin !

Tagged: Welcome_INTJ Darin_PROPN !_PUNCT

Context: you can view at dresscod.com

Tagged: you_PRON can_AUX view_VERB at_ADP dresscod.com_X

...

Context: They work on Wall Street , after all , so when they hear a company who's stated goals include " Do n't be evil , " they imagine a company who's eventually history will be " Do n't be profitable . "

Tagged: They_PRON work_VERB on_ADP Wall_PROPN Street_PROPN ,_PUNCT after_ADV all_ADV ,_PUNCT so_ADV when_ADV they_PRON hear_VERB a_DET company_NOUN who's_PRON stated_VERB goals_NOUN include_VERB "_PUNCT Do_AUX n't_PART be_AUX evil_ADJ ,_PUNCT "_PUNCT they_PRON imagine_VERB a_DET company_NOUN who's_PRON eventually_ADJ history_NOUN will_AUX be_VERB "_PUNCT Do_AUX n't_PART be_AUX profitable_ADJ ._PUNCT "_PUNCT

Context: It 's not quite as freewheeling an environment as you 'd imagine : Sergey Brin has actually created a mathematical ' proof ' that the company 's self - driven research strategy , which gives employees one day a week to do research projects on their own , is a good , respectable idea .

Tagged: It_PRON 's_AUX not_PART quite_ADV as_ADV freewheeling_ADJ an_DET environment_NOUN as_SCONJ you_PRON 'd_AUX imagine_VERB :_PUNCT Sergey_PROPN Brin_PROPN has_AUX actually_ADV created_VERB a_DET mathematical_ADJ '_PUNCT proof_NOUN '_PUNCT that_SCONJ the_DET company_NOUN 's_PART self_NOUN -_PUNCT driven_VERB research_NOUN strategy_NOUN ,_PUNCT which_PRON gives_VERB employees_NOUN one_NUM day_NOUN a_DET week_NOUN to_PART do_VERB research_NOUN projects_NOUN on_ADP their_PRON own_ADJ ,_PUNCT is_AUX a_DET good_ADJ ,_PUNCT respectable_ADJ idea_NOUN ._PUNCT

Context: Read the entire article ; there 's a punchline , too .

Tagged: Read_VERB the_DET entire_ADJ article_NOUN ;_PUNCT there_PRON 's_VERB a_DET punchline_NOUN ,_PUNCT too_ADV ._PUNCT

Context: My opinion piece on the implications of Arafat 's passing for al - Qaeda has appeared at Newsday .

Tagged: My_PRON opinion_NOUN piece_NOUN on_ADP the_DET implications_NOUN of_ADP Arafat_PROPN 's_PART passing_NOUN for_ADP al_PROPN -_PUNCT Qaeda_PROPN has_AUX appeared_VERB at_ADP Newsday_PROPN ._PUNCT

Context: Viel Erfolg ! Tagged: Viel_

Figure 9: Prompt design of iterative prompting.

	syn.	syn_rank	pho.	pho_rank	inv.	inv_rank	fam.	fam_rank	geo.	geo_rank	rank_score
eng-nld	92.43	37	81.83	18	76.28	36	44.51	35	99.96	37	32.6
eng-deu	90.26	36	80.60	15	78.68	37	54.49	37	99.76	35	32.0
eng-ukr	84.73	32	85.83	32	74.91	33	15.03	30	99.28	26	30.6
eng-por	84.24	31	90.46	35	74.03	28	10.14	22	99.68	33	29.8
eng-ell	78.31	25	95.35	37	74.74	32	15.03	32	98.96	22	29.6
eng-pol	78.64	26	85.83	29	74.09	29	15.03	31	99.63	32	29.4
eng-bul	85.78	35	85.83	30	74.38	30	13.73	27	99.01	23	29.0
eng-ita	85.78	34	85.83	28	72.94	26	11.21	23	99.53	30	28.2
eng-rus	81.18	29	85.83	31	74.63	31	16.80	33	95.81	17	28.2
eng-ron	79.60	27	90.46	34	73.42	27	11.89	24	99.22	25	27.4
eng-spa	82.16	30	85.83	27	72.83	25	9.71	21	99.59	31	26.8
eng-lit	69.33	18	80.42	14	75.58	34	19.39	34	99.44	27	25.4
eng-afr	84.94	33	81.83	17	75.91	35	50.46	36	86.84	6	25.4
eng-fra	81.18	28	75.28	7	72.24	24	9.71	20	99.93	36	23.0
eng-est	77.35	24	85.83	25	70.81	19	0.23	15	99.45	28	22.2
eng-hun	69.40	19	85.83	24	70.66	18	0.33	18	99.46	29	21.6
eng-fin	71.08	21	87.05	33	70.00	17	0.19	13	99.19	24	21.6
eng-eus	62.36	13	85.29	21	70.00	16	3.33	19	99.76	34	20.6
eng-urd	61.63	12	85.83	26	71.98	23	12.71	25	92.54	13	19.8
eng-mar	56.50	8	80.42	13	71.57	22	13.73	28	89.80	11	16.4
eng-wol	63.92	14	85.83	23	69.73	15	0.17	10	96.24	18	16.0
eng-hin	61.63	11	78.35	10	70.91	20	12.71	26	91.10	12	15.8
eng-fas	50.03	3	78.35	11	70.94	21	13.73	29	94.23	14	15.6
eng-ind	72.66	22	90.92	36	67.09	12	0.12	4	79.16	1	15.0
eng-heb	75.15	23	72.55	5	69.10	14	0.13	6	97.16	20	13.6
eng-ara	65.11	16	70.09	3	68.38	13	0.15	9	97.04	19	12.0
eng-tur	50.68	4	81.83	16	67.09	11	0.14	7	98.25	21	11.8
eng-zho	71.08	20	72.55	4	66.94	10	0.33	16	88.42	9	11.8
eng-kaz	44.77	1	83.64	19	66.59	9	0.14	8	95.22	16	10.6
eng-vie	66.04	17	78.35	9	65.81	8	0.19	11	85.25	3	9.6
eng-tel	52.07	6	80.42	12	64.76	4	0.19	14	89.18	10	9.2
eng-tgl	60.89	10	85.83	22	64.76	5	0.13	5	82.15	2	8.8
eng-tam	51.36	5	85.29	20	64.37	3	0.11	3	87.95	8	7.8
eng-kor	55.29	7	74.65	6	63.83	2	0.33	17	86.93	7	7.8
eng-tha	63.95	15	78.35	8	65.40	7	0.11	2	85.25	4	7.2
eng-yor	60.04	9	66.77	2	65.29	6	0.10	1	94.98	15	6.6
eng-jpn	50.03	2	66.77	1	56.88	1	0.19	12	85.65	5	4.2

Table 3: Details of language similarity computation.

	language	en	af	ar	bg	de	el	es	et	eu	fa	fi	fr	he
zero-shot	Iter	65.2	67.8	57.2	68.6	65.0	55.0	64.8	49.4	35.6	58.3	50.2	65.4	51.5
	Decom (prob.)	63.6	66.0	67.8	74.4	68.6	62.7	68.6	58.0	54.1	68.5	60.2	63.5	66.4
	Decom (gen.)	45.3	43.8	49.6	50.5	49.0	50.7	43.3	53.6	50.7	56.0	55.5	40.5	55.6
few-shot	Iter	80.2	66.4	65.0	77.3	66.9	56.4	70.8	53.7	50.7	57.4	63.9	67.7	66.4
	Decom (prob.)	85.0	76.9	48.1	82.4	78.3	52.3	82.7	65.2	48.8	57.3	64.4	76.9	66.6
	Decom (gen.)	81.4	74.8	44.3	80.4	77.0	46.3	82.0	64.0	48.1	54.1	63.6	76.4	64.9
	Decom (prob.) + I	83.4	77.9	42.4	76.9	77.8	33.6	77.6	64.6	57.4	42.9	67.6	74.8	58.5
	Decom (gen.) + I	78.7	75.8	34.0	74.9	76.6	24.7	76.4	62.6	56.8	34.4	64.5	73.4	54.5

	language	hi	hu	id	it	ja	kk	ko	lt	mr	nl	pl	pt	ro
zero-shot	Iter	61.3	50.6	54.7	64.0	42.2	36.7	39.9	52.8	39.1	60.4	66.5	63.9	66.2
	Decom (prob.)	37.1	58.6	61.0	68.6	56.3	57.8	47.4	68.2	61.0	69.4	73.5	68.4	68.5
	Decom (gen.)	35.6	46.7	41.8	45.1	48.9	50.2	42.2	60.3	56.7	46.8	59.5	43.1	44.6
few-shot	Iter	65.7	50.4	70.0	67.2	42.0	43.8	42.6	63.2	54.4	66.6	70.9	75.1	65.9
	Decom (prob.)	67.8	71.3	73.9	76.2	59.8	50.0	44.0	67.5	48.9	80.6	78.6	77.8	77.8
	Decom (gen.)	66.2	70.8	73.0	76.0	57.1	50.2	43.4	67.1	48.9	77.2	78.3	76.9	77.0
	Decom (prob.) + I	57.6	66.5	70.4	72.2	54.2	58.4	49.2	69.9	53.1	78.5	76.7	75.0	76.4
	Decom (gen.) + I	55.3	63.9	68.2	70.3	53.1	57.9	48.2	69.5	52.7	76.9	75.7	74.2	75.1

	language	ru	ta	te	th	tl	tr	uk	ur	vi	wo	yo	zh	avg.
zero-shot	Iter	68.2	39.2	51.1	54.1	65.0	47.7	67.0	56.0	41.7	31.5	41.3	58.8	54.3
	Decom (prob.)	74.4	55.2	63.8	63.0	62.9	55.2	74.1	54.2	59.9	39.6	49.7	59.2	61.8
	Decom (gen.)	54.7	52.2	57.4	50.1	51.3	43.2	57.4	40.3	45.9	29.2	43.3	55.7	48.7
few-shot	Iter	74.0	52.0	62.4	57.1	37.3	62.0	68.2	59.6	41.0	25.2	39.0	62.3	58.9
	Decom (prob.)	79.9	37.5	61.4	58.2	73.4	62.7	77.7	51.3	52.6	42.0	47.8	65.8	64.4
	Decom (gen.)	78.0	33.9	61.3	56.9	73.4	62.6	76.2	45.7	52.8	42.0	47.6	64.5	63.0
	Decom (prob.) + I	76.8	35.7	67.0	45.8	74.9	63.7	75.1	40.5	59.4	43.1	49.2	62.9	62.3
	Decom (gen.) + I	73.9	28.0	66.6	42.9	74.9	62.6	73.4	32.9	59.7	43.2	48.6	61.4	59.9

Table 4: Full results on Mistral 7b.

	language	en	af	ar	bg	de	el	es	et	eu	fa	fi	fr	he
zero-shot	Iter	33.1	38.8	30.2	33.2	34.5	38.1	38.9	19.7	11.8	17.7	26.0	37.5	21.3
	Decom (prob.)	58.2	45.1	49.6	55.9	53.3	50.4	44.7	37.7	36.4	40.5	41.3	46.8	39.5
	Decom (gen.)	53.8	46.8	38.5	45.8	57.1	54.3	52.4	28.6	20.2	35.9	39.8	53.1	37.5
few-shot	Iter	68.0	56.1	58.0	63.4	56.9	48.7	55.3	46.5	41.3	51.1	50.5	54.2	54.0
	Decom (prob.)	74.7	60.0	29.9	64.7	63.0	30.6	55.7	53.0	44.4	29.7	62.9	54.4	42.8
	Decom (gen.)	62.1	51.0	25.7	60.3	52.4	23.9	50.3	48.3	42.9	26.0	56.8	49.5	37.5
	Decom (prob.) + I	68.2	55.9	23.7	61.6	61.0	20.2	52.5	43.2	40.8	22.7	49.4	54.8	35.4
	Decom (gen.) + I	63.4	53.2	19.0	57.9	56.2	12.0	47.8	39.3	40.0	15.5	46.4	51.2	30.1

	language	hi	hu	id	it	ja	kk	ko	lt	mr	nl	pl	pt	ro
zero-shot	Iter	35.2	29.3	31.1	35.1	28.7	13.6	19.8	24.9	13.2	37.5	37.7	38.4	32.0
	Decom (prob.)	36.9	47.0	46.9	46.7	32.4	39.0	29.0	34.9	45.3	54.9	54.0	48.6	43.6
	Decom (gen.)	34.8	47.4	39.1	45.2	30.9	33.0	33.2	37.7	42.0	51.1	44.1	48.5	42.6
few-shot	Iter	54.0	41.0	51.3	49.6	40.0	43.2	25.0	52.5	50.3	52.2	52.4	52.0	53.8
	Decom (prob.)	45.8	62.6	60.9	56.4	40.2	51.4	48.2	56.3	47.3	58.9	67.2	60.3	63.6
	Decom (gen.)	42.4	57.0	56.5	51.6	34.1	47.5	44.7	51.7	43.5	51.3	64.2	54.5	55.5
	Decom (prob.) + I	30.6	52.3	54.1	51.3	37.3	46.6	41.9	46.5	45.7	64.2	65.4	55.2	56.4
	Decom (gen.) + I	24.1	50.6	49.5	44.1	32.9	46.0	40.7	45.3	34.5	60.2	62.0	51.2	51.8

	language	ru	ta	te	th	tl	tr	uk	ur	vi	wo	yo	zh	avg.
zero-shot	Iter	29.8	19.2	13.8	29.2	28.6	22.2	30.3	20.7	29.7	13.3	13.7	32.2	27.2
	Decom (prob.)	55.8	38.0	34.0	37.5	57.3	48.3	57.4	31.6	39.5	27.6	29.1	42.9	43.2
	Decom (gen.)	48.7	25.5	36.9	34.6	66.3	45.9	48.8	28.4	35.3	18.7	21.8	44.0	40.4
few-shot	Iter	58.2	30.9	54.3	49.4	37.3	34.4	57.7	44.0	46.5	40.7	39.3	52.0	48.6
	Decom (prob.)	67.2	31.7	44.7	36.5	46.8	58.1	62.9	27.1	41.4	39.9	37.1	64.8	50.5
	Decom (gen.)	62.3	25.3	43.5	34.7	45.4	55.9	59.4	23.7	40.7	36.2	35.5	50.9	45.8
	Decom (prob.) + I	59.6	20.3	38.4	20.9	63.1	54.1	59.9	19.3	49.7	32.2	33.8	48.2	45.1
	Decom (gen.) + I	56.9	12.5	34.5	16.7	58.8	52.7	57.5	13.0	47.8	29.7	31.7	44.2	41.0

Table 5: Full results on LLaMA2 7b.

	language	en	af	ar	bg	de	el	es	et	eu	fa	fi	fr	he
zero-shot	Iter	47.6	37.4	43.2	44.5	45.7	38.4	46.8	37.0	26.5	42.0	40.7	45.5	40.0
	Decom (prob.)	67.3	60.1	54.4	62.7	63.6	60.5	55.9	49.9	37.4	59.8	62.6	53.4	55.4
	Decom (gen.)	59.2	54.1	45.0	52.5	57.5	51.3	56.3	37.6	36.7	49.7	50.2	54.7	44.3
few-shot	Iter	68.0	62.3	57.4	69.9	60.3	57.9	66.7	44.8	41.0	49.1	54.2	63.2	59.8
	Decom (prob.)	77.3	67.8	33.2	67.6	67.5	35.0	62.6	58.5	46.9	34.7	62.8	64.8	48.4
	Decom (gen.)	65.3	59.1	25.1	61.3	58.6	24.6	53.5	51.8	45.8	27.4	55.4	55.9	43.9
	Decom (prob.) + I	74.3	67.6	25.9	60.7	70.5	21.5	59.1	51.4	44.1	21.8	59.1	63.1	40.3
	Decom (gen.) + I	68.7	64.4	19.2	58.7	66.2	12.4	53.9	47.9	42.2	15.5	54.0	59.7	35.0
	language	hi	hu	id	it	ja	kk	ko	lt	mr	nl	pl	pt	ro
zero-shot	Iter	45.0	38.8	40.9	41.8	42.8	24.1	29.8	41.2	30.5	36.6	42.2	43.3	43.1
	Decom (prob.)	53.8	57.6	57.4	54.8	48.3	51.8	45.1	54.3	50.2	62.0	66.4	56.6	57.9
	Decom (gen.)	45.4	47.9	48.2	51.3	35.9	48.7	35.3	43.2	48.7	56.9	58.2	51.3	51.4
few-shot	Iter	51.6	46.1	60.8	62.7	46.5	32.0	26.6	50.8	52.7	61.0	64.4	68.9	58.9
	Decom (prob.)	45.4	69.8	62.2	61.2	44.6	52.3	46.1	63.0	49.6	65.4	68.1	62.3	63.6
	Decom (gen.)	37.3	60.5	55.8	54.5	40.7	49.4	42.6	58.4	46.9	54.9	61.4	54.3	54.9
	Decom (prob.) + I	31.4	64.2	55.3	55.3	38.1	51.7	47.1	58.9	52.5	65.4	60.2	56.3	60.4
	Decom (gen.) + I	23.4	60.0	50.2	52.4	35.5	49.0	45.3	56.9	50.8	61.1	58.2	54.1	56.1
	language	ru	ta	te	th	tl	tr	uk	ur	vi	wo	yo	zh	avg.
zero-shot	Iter	42.6	21.8	22.5	45.6	29.3	29.9	39.8	35.1	36.0	24.4	24.1	45.2	37.4
	Decom (prob.)	66.5	49.1	50.8	44.6	66.5	56.9	65.7	47.2	45.3	34.5	47.7	58.7	54.7
	Decom (gen.)	55.2	46.2	54.1	44.2	73.1	52.8	57.3	40.2	45.4	29.9	39.6	52.5	48.7
few-shot	Iter	64.9	33.5	51.5	51.5	60.2	46.3	61.6	45.4	41.8	36.3	31.6	52.1	52.6
	Decom (prob.)	71.0	30.4	54.4	40.1	74.0	54.1	69.0	30.1	47.5	39.4	36.2	66.6	54.5
	Decom (gen.)	63.3	21.9	51.3	33.9	70.9	52.2	61.4	22.1	45.2	38.1	34.8	56.5	48.3
	Decom (prob.) + I	63.3	22.3	52.2	23.5	70.7	53.9	62.4	19.0	48.4	36.9	36.4	56.7	49.4
	Decom (gen.) + I	59.8	14.1	48.4	18.5	70.2	53.2	59.1	12.0	47.1	34.5	34.5	52.7	45.6

Table 6: Full results on LLaMA2 13b.

	language	en	af	ar	bg	de	el	es	et	eu	fa	fi	fr	he
zero-shot	Iter	6.4	7.2	10.9	7.6	9.5	8.4	8.2	12.4	7.5	7.3	9.3	9.0	9.6
	Decom (prob.)	20.6	20.5	14.5	19.7	26.2	18.3	18.2	22.3	19.0	12.8	19.2	19.4	15.2
	Decom (gen.)	28.7	18.3	16.4	22.6	26.8	22.7	24.9	21.2	25.0	11.3	20.9	20.9	21.8
few-shot	Iter	30.9	6.4	14.4	23.8	19.3	7.7	23.2	16.6	28.4	11.1	22.3	25.1	7.5
	Decom (prob.)	44.1	33.1	28.7	35.9	44.0	39.2	33.6	39.0	38.4	25.6	38.5	35.6	34.3
	Decom (gen.)	40.6	31.0	25.5	31.4	39.5	35.8	30.5	36.9	33.8	21.6	36.8	31.0	33.6
	Decom (prob.) + I	33.3	24.7	27.2	35.2	30.0	31.0	30.1	36.5	37.4	24.7	34.4	29.0	29.2
	Decom (gen.) + I	33.3	24.5	27.1	35.0	29.7	30.4	30.0	36.4	37.1	24.5	34.5	28.9	29.1
	language	hi	hu	id	it	ja	kk	ko	lt	mr	nl	pl	pt	ro
zero-shot	Iter	3.9	13.0	10.0	9.1	2.8	4.5	8.5	7.8	0.4	9.1	9.9	8.6	8.8
	Decom (prob.)	12.0	27.0	17.7	23.1	13.5	17.7	19.5	23.6	12.4	18.6	23.6	19.5	19.6
	Decom (gen.)	15.2	21.9	17.3	26.2	26.2	16.8	21.3	23.4	25.8	14.7	23.2	27.8	24.3
few-shot	Iter	20.5	13.4	30.5	19.0	6.3	17.0	5.9	15.0	35.2	20.8	17.9	27.4	13.4
	Decom (prob.)	27.0	38.2	43.8	33.9	25.9	45.6	35.0	40.3	39.6	39.8	39.7	34.4	33.3
	Decom (gen.)	24.8	36.9	41.2	31.1	22.5	43.8	32.7	39.5	28.0	36.5	36.5	31.7	32.0
	Decom (prob.) + I	25.6	32.3	36.0	30.7	25.3	45.2	27.7	41.0	44.5	29.0	34.7	30.4	32.5
	Decom (gen.) + I	25.6	32.2	35.9	30.6	25.1	45.1	27.7	41.0	43.7	28.6	34.6	30.3	32.5
	language	ru	ta	te	th	tl	tr	uk	ur	vi	wo	yo	zh	avg.
zero-shot	Iter	6.8	5.0	5.1	6.8	3.9	9.0	5.2	6.6	4.2	1.4	7.2	7.6	7.4
	Decom (prob.)	26.1	15.0	7.9	8.7	7.8	15.5	23.7	8.1	14.4	11.0	18.9	21.7	17.6
	Decom (gen.)	27.9	20.7	12.8	2.7	1.9	17.4	28.1	12.8	25.7	21.1	28.3	26.0	20.6
few-shot	Iter	20.3	24.3	47.0	3.1	22.5	20.9	20.9	15.5	18.3	16.5	16.9	20.7	18.8
	Decom (prob.)	41.9	36.5	48.2	25.0	41.9	37.9	39.6	26.2	26.9	34.1	39.2	40.8	36.2
	Decom (gen.)	36.8	33.5	41.7	23.1	41.9	36.4	37.0	24.7	24.5	33.2	36.5	35.7	33.2
	Decom (prob.) + I	37.0	34.1	39.0	13.7	57.8	38.0	35.8	26.4	34.0	30.3	33.3	32.8	32.9
	Decom (gen.) + I	36.9	33.9	38.8	13.6	57.8	38.0	35.4	26.4	33.9	30.3	33.3	32.6	32.7

Table 7: Full results on BLOOMZ 7b.

	language	en	af	ar	bg	de	el	es	et	eu	fa	fi	fr	he
zero-shot	Decom (gen.)	47.6	45.7	37.8	48.9	48.9	45.8	40.0	45.3	41.5	44.2	46.8	42.6	42.6
few-shot	Decom (gen.)	49.0	41.0	16.2	37.6	43.9	31.0	37.2	34.8	33.9	33.4	32.1	38.5	34.1
	Decom (gen.) + I	57.3	51.9	27.4	47.2	55.4	40.1	50.1	41.2	43.6	48.1	42.4	49.9	45.6
	language	hi	hu	id	it	ja	kk	ko	lt	mr	nl	pl	pt	ro
zero-shot	Decom (gen.)	40.6	38.7	39.3	39.3	32.9	46.1	29.2	47.4	47.5	42.8	46.1	40.6	49.4
few-shot	Decom (gen.)	23.8	33.5	39.9	36.5	14.3	32.4	17.7	37.5	34.9	42.7	36.1	37.1	35.6
	Decom (gen.) + I	44.7	36.2	51.9	45.7	44.6	45.7	26.7	45.7	48.8	55.3	46.2	48.9	51.5
	language	ru	ta	te	th	tl	tr	uk	ur	vi	wo	yo	zh	avg.
zero-shot	Decom (gen.)	45.9	39.4	51.3	47.1	59.3	46.9	47.4	37.9	48.4	22.3	37.5	42.8	43.1
few-shot	Decom (gen.)	33.5	28.1	50.9	21.9	65.7	34.7	31.2	17.7	33.9	10.5	22.4	17.2	32.5
	Decom (gen.) + I	43.8	38.0	55.3	46.6	70.5	46.0	41.5	36.0	49.0	19.8	38.6	34.5	44.7

Table 8: Full results on mTk 13b.

k	en	de	el	fa	hi	nl	ru	zh	avg.
0	63.6	68.6	62.7	68.5	37.1	69.4	74.4	59.2	62.9
1	72.4	76.1	69.2	65.9	62.5	75.1	69.4	60.2	68.9
3	77.0	75.8	70.9	70.7	69.7	71.5	75.2	70.1	72.6
5	84.4	80.9	76.7	75.3	70.9	80.4	81.8	65.9	77.0
7	86.2	79.8	76.5	79.3	71.4	78.8	83.2	68.3	77.9
9	85.4	78.5	72.8	78.2	73.1	77.6	81.8	66.4	76.7
11	84.9	80.2	73.9	78.6	72.9	81.3	82.3	65.7	77.5
13	84.1	78.7	73.1	77.9	72.6	79.1	80.7	65.4	76.5
15	86.2	80.0	73.6	71.6	73.7	82.7	83.4	67.9	77.4
17	85.8	78.3	72.9	70.1	73.0	80.9	80.8	66.6	76.1

Table 9: Full results of few-shot ablation study.