Investigating Large Language Models for Complex Word Identification in Multilingual and Multidomain Setups

Anonymous ACL submission

Abstract

Complex Word Identification (CWI) is an important step in the lexical simplification task and has recently become a task on its own. Some variations of this binary classification task have emerged, such as lexical complexity prediction (LCP) and complexity evaluation of multi-word expressions (MWE). Large language models (LLMs) recently became popular in the Natural Language Processing community because of their versatility and capability to solve unseen tasks in zero/few-shot settings. Our work investigates LLM usage, specifically Llama 2 and ChatGPT 3.5 turbo, in the CWI, LCP, and MWE settings. We show that LLMs may struggle in certain conditions or achieve comparable results against existing methods.

1 Introduction

Complex word identification (CWI) aims to identify whether words or phrases can be difficult for a target group of readers to understand. Often, it is used in lexical simplification – a task that targets replacing complex words and expressions with simplified alternatives (North et al., 2023a). CWI represents the first step, and it was treated as part of the lexical simplification task until 2012, when it became a standalone task (Shardlow, 2013).

CWI was initially addressed as a binary classification task (Paetzold and Specia, 2016), identifying whether a word is complex in a given sentence. When the task became more popular (North et al., 2023b), it was extended to the continuous domain as Lexical Complexity Prediction (LCP, also referred to as the probabilistic classification for CWI) (Yimam et al., 2018) addressing multi-language and multi-domain settings, and then it was extended to multi-word expressions (Shardlow et al., 2021). Recently, new datasets started to emerge in various languages and domains (Ortiz Zambrano and Montejo-Ráez, 2021; Venugopal et al., 2022; Ilgen and Biemann, 2023; Zambrano et al., 2023). Previous approaches to CWI ranged from using Support Vector Machines (S.P et al., 2016) to deep neural networks based on Bidirectional Representation from Encoder Transformers (Pan et al., 2021), multi-task learning with domain adaptation (Zaharia et al., 2022), and sequence modeling (Gooding and Kochmar, 2019).

With the recent large language models (LLMs) breakthrough, OpenAI showed that Generative Pre-trained Transformer (GPT) models (Radford et al., 2019; Brown et al., 2020) are capable of improved performances on various natural language processing tasks as we scale up the model size and the amount of training data. Since ChatGPT[1] and GPT-4 (OpenAI et al., 2023) were announced, many other models (close- and open-source) emerged, such as PaLM (Anil et al., 2023), LLaMA (Touvron et al., 2023), Orca (Mitra et al., 2023), and Mistral (Jiang et al., 2023), with better performances, hence, the race to develop and fine-tune such models for various applications.

Our work shows that LLMs can address CWI and LCP and achieve comparable results with state-of-the-art approaches. We evaluate pre-trained LLaMA 2 (Touvron et al., 2023) and OpenAI's ChatGPT-3 turbo in zero-shot and fine-tuning settings. We summarize the contributions as follows:

- To the best of our knowledge, we are the first to employ LLMs for CWI and LCP.
- We evaluate LLMs in binary (discrete set of labels) and probabilistic classification (continuous space labels) on multi-domain and multi-lingual corpora.
- We show that fine-tuned LLMs can achieve comparable results or exceed other existing approaches with some limitations, and we provide some insights about the results.

[1] https://openai.com/blog/chatgpt

2 Related Work

2.1 Complex Word Identification

Aroyehun et al. (2018) compared CNN-based models with various feature engineering methods based on tree ensembles and features, achieving comparable results. Zaharia et al. (2020) employed zero- and few-shot learning techniques, along with Transformers and Recurrent Neural Networks, in a multilingual setting. CWI was also considered in a sequential task, where Gooding and Kochmar (2019) used a bidirectional LSTM with word embeddings and character-level representations and a language modeling objective to learn the complexity of words given the context. Other approaches such as graph-based (Ehara, 2019), domain adaptation (Zaharia et al., 2022), and transformer-based models (Pan et al., 2021; Cheng Sheang et al., 2022) were used to improved CWI performances.

2.2 Large Language Models

LLMs were successfully utilized in various generative tasks (Pu et al., 2023; Chen et al., 2021). The new paradigm in solving other non-generative tasks is based on prompting pre-trained language models to perform the prediction task (Liu et al., 2023a; Sun et al., 2023). Fine-tuning models on instructions showed improved results in zero-shot settings, especially on unseen tasks (Wei et al., 2022a). Prompt-based methods such as the use of demonstrations (Min et al., 2022), intermediate reasoning steps by breaking down complex tasks into simpler subtasks (also known as a chain of thought) (Wei et al., 2022b), and using LLMs to optimize their prompts (Zhou et al., 2023) made zero-shot inference much more appealing due to reduced costs and more efficient than fine-tuning LLMs.

3 Method

3.1 Problem Formulation in Pre-LLM Era

Word complexity can be defined as absolute and relative (North et al., 2023b). Absolute complexity is determined by the objective linguistic properties (e.g., semantic, morphological, phonological), while relative complexity is related to the subjective speaker's point of view (e.g., familiarity with sound and meaning). In this work, we evaluate the relative complexity of words, in general, for non-native speakers. Considering an annotated dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$ of N samples, the task can be viewed as a binary classification (known as CWI), where, given the pair $x_i = (C_i, w_i)$ of a sentence $C_i = (w_1, w_2, ...)$ and word $w_i \in C_i$ the system outputs $y_i^{CWI} \in \{0, 1\}$ (i.e., complex or non-complex) (Paetzold and Specia, 2016). A variation of the CWI task is to evaluate the complexity $y_i^{MWE} \in \{0, 1\}$ of a multi-word expression $e_i = (w_1, w_2, ...)$ containing multiple words $w_j, j = 1 : |e|$, from a given context C_i (i.e., $x_i = (C_i, e_i)$) (Shardlow et al., 2021). Later, CWI was considered in the continuous domain (known as LCP), indicating the degree of difficulty $y_i^{LCP} \in [0, 1]$, for the given word $w_i \in C_i$ in the context C_i (Yimam et al., 2018).

3.2 Problem Formulation in LLM Era

Starting from the previous formulation, we derive the formalism in the context of LLMs.

Binary classification. Given an example $x_i = (C_i, w_i)$, the model predicts if a given phrase w_i from the sentence C_i is complex. Since the access to the tokens logits is limited for closed-source models (e.g., OpenAI's ChatGPT), we consider that the model only outputs "true" or "false" (or any equivalent form) without a confidence estimation.

Probabilistic classification. The model produces a real value between 0 and 1, representing the degree of complexity for (C_i, w_i). LLMs are known to suffer from hallucination (OpenAI et al., 2023), and directly predicting real values is challenging. We abide by Liu et al. (2023b)'s solution for estimating the scoring function. We ask the model to predict on the 5-point Likert scale, in natural language, one of "very easy", "easy", "neutral", "difficult", or "very difficult". This scale is converted to a numerical representation using the following mapping: very easy - 0, easy - 0.25, neutral - 0.5, difficult - 0.75, and very difficult - 1. Since LLMs output tokens from a probability distribution, we set the temperature (in our experiments, we use 0.8) to determine how random the outputs are. The numerical representation constructed from LLM's output is denoted as $s_k \in S$ for a sampling step k, with $S = \{0, 0.25, 0.5, 0.75, 1\}$. The model's probability to output one 5-point Likert score is $p(s_k)$. The final score S is:

$$\mathbb{E}_p[S] = \sum_{s \in S} p(s) \cdot s \quad (1)$$

For experiments, we use the sample mean estimator $\bar{S} = \frac{1}{K} \sum_{k=1}^{K} s_k$ of K sampling steps.

3.3 Prompting LLMs

We provide the instructions to the model regarding the task and how the output should be formatted, and then we ask the model to predict an example. All prompt templates are listed in Appendix A, which were obtained after multiple trial-and-error interactions until we reached the wanted behavior. The prompts for German and Spanish are translations of the English prompts. The same prompts are used across different models.

4 Experimental Setup

4.1 Models

We employ two families of LLMs: Llama 2 (Touvron et al., 2023) and ChatGPT-3.5-turbo (OpenAI et al., 2023). For Llama 2, we use the pre-trained 7 and 13 billion parameter variants, both base (pre-trained on 2 trillion tokens) and chat models (fined-tuned using reinforcement learning with human feedback). The chat model is used in the zero-shot setting. In addition, we fine-tune the base model on the training set for CWI/LCP. In the multi-lingual settings, because the base Llama 2 models were not trained to handle languages other than English, we use instead checkpoints found on Huggingface for German[2] and Spanish[3]. These checkpoints are used for Llama 2 in the same way as the English checkpoint but in multilingual settings. For ChatGPT-3.5-turbo (175 billion parameters), we use the latest available checkpoint `gpt-3.5-turbo-1106` to accomplish all experiments. For inference and fine-tuning, we use OpenAI's API[4].

4.2 Datasets

CompLex LCP 2021. Proposed at SemEval 2021 Task 1 (Shardlow et al., 2021), CompLex LCP 2021 comprises around 10,000 sentences in English from three domains: European Parliament proceedings, the Bible, and biomedical literature. The data is split across two tasks: single-word (Single) and multi-word expressions (MWE). The complexity is provided as continuous values between 0 and 1, addressed as the probabilistic classification task. The average complexity is 0.3 for single and 0.42 for MWE. For evaluation, the test set has 907 samples for single words and 185 for multi-word expressions.

CWI Shared Dataset. It was proposed at the CWI Shared Task in 2018 (Yimam et al., 2018) and addresses English multi-domain and multi-lingual settings. The English split contains samples from three sources (News, WikiNews, and Wikipedia) totaling approx. 35,000 samples. In the multi-lingual setting, the dataset features German and Spanish with approx. 8,000 and 17,600 samples, respectively, and a French test set containing 2,251 samples. The dataset was developed to address binary and probabilistic classification tasks by offering probabilities and labels such that samples with 0% probability are non-complex and others as complex. We consider only the binary classification tasks (see Limitations 7). The English News dataset has 2,095 samples for the test sets, English WikiNews has 1,287 samples, English Wikipedia has 870 samples, German has 961 samples, and Spanish has 2,233 samples.

4.3 Baselines

We compare against top-performing methods at CWI Shared task and LCP 2021. Camb (Gooding and Kochmar, 2018) employs heterogeneous features combined with an ensemble of AdaBoost classifiers. TMU system (Kajiwara and Komachi, 2018) uses a random forest classifier on multiple hand-crafted features. ITEC (De Hertog and Tack, 2018) combines CNN and LSTM layers. SB@GU (Alfter and Pilán, 2018) employs Random Forest and Extra Tree on top of multiple hand-crafted features. In addition, we include the XLM-RoBERTa-based approach combined with text simplification and domain adaptation (Zaharia et al., 2022), the MLP combined with Sent2Vec solution Almeida et al. (2021), and $RoBERTa_{LARGE}$ with an ensemble of RoBERTa-based models (LR-Ensemble) (Pan et al., 2021).

4.4 Evaluation

We adopt the same evaluation methodology as in Shardlow et al. (2021) for CompLex and Yimam et al. (2018) for CWI datasets. Therefore, we use Pearson correlation (P) and Mean Average Error (MAE) on the CompLex dataset and F1-score (F1) for the CWI dataset. We also include accuracy (Acc) on the CWI dataset. We report all results on a single run for CWI and multiple runs (described by N) for LCP.

[2] https://huggingface.co/LeoLM/leo-hessianai-7b

[3] https://huggingface.co/clibrain/Llama-2-7b-ft-instruct-es

[4] https://platform.openai.com/docs/guides/fine-tuning

Model	News		WikiNews		Wikipedia	
	F1↑	Acc↑	F1↑	Acc↑	F1↑	Acc↑
Camb	**87.3**	-	**84.0**	-	**81.2**	-
ITEC	86.4	-	81.1	-	78.1	-
TMU	86.3	-	78.7	-	76.1	-
			Zero-shot			
Llama-2-7b-chat	54.8	59.5	46.9	48.7	63.3	60.3
Llama-2-13b-chat	47.6	61.6	41.0	57.6	51.4	53.4
ChatGPT-3.5-turbo	47.3	68.6	47.0	64.3	52.3	61.6
			Fine-tuned			
Llama-2-7b-ft	78.0	82.9	78.2	81.1	77.4	76.7
Llama-2-13b-ft	77.6	83.3	77.7	81.3	73.1	74.6
ChatGPT-3.5-turbo-ft	80.7	83.9	80.9	83.1	80.2	79.4

Table 1: Results on the multi-domian English test set from CWI 2018 Shared Dataset. In bold, we denote the best score and underlined are the second-best results for zero-shot and fine-tuned settings.

Model	German		Spanish	
	F1↑	Acc↑	F1↑	Acc↑
TMU	**74.5**	-	77.0	-
ITEC	-	-	76.3	-
SB@GU	74.2	-	72.8	-
Llama-2-7b-ft	66.9	75.1	66.3	72.3
Llama-2-13b-ft	70.8	76.6	75.3	81.0
ChatGPT-3.5-turbo-ft	66.6	78.0	**78.1**	74.4
ChatGPT-3.5-turbo	61.5	67.6	66.3	63.7

Table 2: Results on the multi-lingual test sets from CWI 2018 Shared Dataset. In bold we denote the best score, and underlined are the second-best results.

Model	Single-Word		Multi-Word	
	P↑	MAE↓	P↑	MAE↓
MLP+Sent2Vec	.4598	.0866	.3941	.1145
XLM-RoBERTa-based	.7744	.0652	.8285	.0708
RoBERTa$_{\text{LARGE}}$	**.7903**	**.0648**	.7900	.0753
LR-Ensemble	-	-	**.8612**	**.0616**
		Zero-shot		
Llama-2-7b-chat	.3302	.1977	.4979	.1797
Llama-2-13b-chat	.4429	.1355	.5794	.1186
ChatGPT-3.5-turbo	.5231	.2307	.6665	.1952
		Fine-tuned		
Llama-2-7b-ft	.7732	.0670	.7919	.0766
Llama-2-13b-ft	.7815	.0797	.8317	.0717
ChatGPT-3.5-turbo-ft	.7372	.1379	.7493	.1834

Table 3: Results on the CompLex LCP 2021 dataset. In bold we denote the best score, and underlined are the second-best results.

5 Results

5.1 English Multi-Domain Setup

We notice that LLM-based methods fall behind these classifiers on the CWI Shared dataset (see Table 1). The top-performing LLM is ChatGPT-3.5-turbo, which generally achieves higher scores, especially when fined-tuned, over 80% F1-score. Llama-2-7b variants achieve higher scores than the larger 13b variant. In addition, we noticed that fine-tuned models obtain consistent results across datasets.

5.2 Multi-Lingual Setup

The results are presented in Table 2 for the German and Spanish languages. On the German dataset, the best LLM result is achieved by Llama2-13b-ft, which achieves 3.7% lower than the random-forest-based classifier. ChatGPT-3.5-turbo showed lower performances than Llama-based models. However, it outperforms all other approaches on the Spanish dataset. The reason could be the imbalance and the text quality across languages in the pre-training stage since Llama models reveal this case. The German checkpoint was pre-trained on text translated by ChatGPT and generated by GPT-4.

5.3 Lexical Complexity Prediction Setup

On the CompLex LCP dataset, Pan et al. (2021) achieved the highest scores. For Llama 2, we set the number of inference steps $N = 20$, while for ChatGPT-3-5 turbo, we evaluated on $N = 10$ inferences (see Appendix G). Refer to Table 3 for the results. Fine-tuned LLM-based models outperform RoBERTa-based models, the best-performing model being Llama-2-13b-ft. Llama-2-7b-ft performs similarly to the XLM-RoBERTa-based model. We notice a considerable performance drop in the zero-shot settings, where the models tend to predict and consider the phrases easier (see Appendix F). The ensemble of RoBERTa models outperforms LLMs.

6 Conclusions

In conclusion, we addressed CWI and LCP using LLMs, specifically Llama 2 and OpenAI's ChatGPT-3.5 turbo. We observed that these models can determine the word difficulty level in multiple domains and languages. But simultaneously, these models struggle to label very difficult phrases correctly. Future directions imply investigating multiple models in more languages, including the state-of-the-art GPT-4 model. Also, as we noticed that the prompts and example selection greatly influence the models' performance, other future work should rely on reducing hallucination and determining which adversarial examples affect the model's capabilities most in the context of CWI.

7 Limitations

Our approach has some limitations regarding prompt design. During experiments, we noticed that prompt design can highly influence the results, especially in the case of zero-shot settings. Using the same prompt across all models is not optimal, but we tried to find those instructions that benefit all models. Providing the model with specific instructions helps the model to better focus on the task and reduce hallucination. One way to mitigate hallucinations was to use a specific JSON format (see Appendix A), which the model required to confirm the task. We do not provide results for the zero-shot multi-lingual setup using Llama 2 since the model could not output the requested format, which made evaluation very difficult.

Also, we know that random sampling is not the optimal solution for choosing fine-tuning examples for ChatGPT–3.5-turbo. The size and quality of data greatly impact the prediction performance. To reduce this effect, we created a balanced dataset among label difficulties, such that the model equally sees easy and difficult words. We also kept a uniform distribution among complexity probabilities strictly greater than zero for both tasks (CWI and LCP).

Another limitation during experiments was access to hardware and pricing. We trained and ran inferences on NVIDIA RTX 4080 (consumer-class GPU) and NVIDIA A100 40GB-PCIe (server-class GPU), depending on the minimal requirements to run the model. For using OpenAI's API, we tried to keep the budget for all experiments under $50 while achieving good performances (with the pricing at the time of writing this paper: $0.0005 per 1k input tokens and $0.0015 per 1k output tokens for chatgpt-3.5-turbo; and $0.0080 per training tokens, $0.003 per 1k input tokens, and $0.006 per 1k output tokens). For this reason, we limited our experiments to only classification on large test sets.

8 Ethical Considerations

Since we used pre-trained LLMs, all their limitations apply to our work. Developing CWI and LCP systems can be beneficial for new language learners (e.g., chat-based applications in which LLMs help new language learners to understand difficult words and even provide alternatives), but at the same time, because of hallucination and inaccuracies that such models may provide, these systems can violate codes of ethics and harm or address attacks to such individuals. We are aware of the fast-paced development in the LLM area, and we think this area of research needs some attention. Therefore, we will make the fine-tuned models publicly available for transparency and fair comparison with feature works. These models should only be used for research. All the data we used is already publicly available, and the pre-trained LLaMA models are available on HuggingFace[5], under the LLaMA 2 License Agreement[6]. We did not use the resources for other purposes than the ones allowed.

References

David Alfter and Ildikó Pilán. 2018. SB@GU at the complex word identification 2018 shared task. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 315–321, New Orleans, Louisiana. Association for Computational Linguistics.

Raul Almeida, Hegler Tissot, and Marcos Didonet Del Fabro. 2021. C3SL at SemEval-2021 task 1: Predicting lexical complexity of words in specific contexts with sentence embeddings. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 683–687, Online. Association for Computational Linguistics.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek,

[5] https://huggingface.co/meta-llama

[6] https://github.com/facebookresearch/llama/blob/main/LICENSE

Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report.

Segun Taofeek Aroyehun, Jason Angel, Daniel Alejandro Pérez Alvarez, and Alexander Gelbukh. 2018. Complex word identification: Convolutional neural network vs. feature engineering. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 322–327, New Orleans, Louisiana. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, Jared Kaplan, Harrison Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code. *CoRR*, abs/2107.03374.

Kim Cheng Sheang, Anaïs Koptient, Natalia Grabar, and Horacio Saggion. 2022. Identification of complex words and passages in medical documents in French. In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, pages 116–125, Avignon, France. ATALA.

Dirk De Hertog and Anaïs Tack. 2018. Deep learning architecture for complex word identification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 328–334, New Orleans, Louisiana. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms.

Yo Ehara. 2019. Graph-based analysis of similarities between word frequency distributions of various corpora for complex word identification. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 1982–1986.

Sian Gooding and Ekaterina Kochmar. 2018. CAMB at CWI shared task 2018: Complex word identification with ensemble-based voting. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 184–194, New Orleans, Louisiana. Association for Computational Linguistics.

Sian Gooding and Ekaterina Kochmar. 2019. Complex word identification as a sequence labelling task. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1148–1153, Florence, Italy. Association for Computational Linguistics.

Bahar Ilgen and Chris Biemann. 2023. Cwitr: A corpus for automatic complex word identification in turkish texts. In *Proceedings of the 2022 6th International Conference on Natural Language Processing and Information Retrieval*, NLPIR '22, page 157–163, New York, NY, USA. Association for Computing Machinery.

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.

Tomoyuki Kajiwara and Mamoru Komachi. 2018. Complex word identification based on frequency in a learner corpus. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 195–199, New Orleans, Louisiana. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023a. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on*

Empirical Methods in Natural Language Processing, pages 2511–2522, Singapore. Association for Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andres Codas, Clarisse Simoes, Sahaj Agarwal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, Hamid Palangi, Guoqing Zheng, Corby Rosset, Hamed Khanpour, and Ahmed Awadallah. 2023. Orca 2: Teaching small language models how to reason.

Kai North, Tharindu Ranasinghe, Matthew Shardlow, and Marcos Zampieri. 2023a. Deep learning approaches to lexical simplification: A survey.

Kai North, Marcos Zampieri, and Matthew Shardlow. 2023b. Lexical complexity prediction: An overview. *ACM Comput. Surv.*, 55(9).

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, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report.

Jenny A. Ortiz Zambrano and Arturo Montejo-Ráez. 2021. CLexIS2: A new corpus for complex word identification research in computing studies. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1075–1083, Held Online. INCOMA Ltd.

Gustavo Paetzold and Lucia Specia. 2016. SemEval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Se-*

mantic Evaluation (SemEval-2016), pages 560–569, San Diego, California. Association for Computational Linguistics.

Chunguang Pan, Bingyan Song, Shengguang Wang, and Zhipeng Luo. 2021. DeepBlueAI at SemEval-2021 task 1: Lexical complexity prediction with a deep ensemble approach. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 578–584, Online. Association for Computational Linguistics.

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Matthew Shardlow. 2013. A comparison of techniques to automatically identify complex words. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, pages 103–109, Sofia, Bulgaria. Association for Computational Linguistics.

Matthew Shardlow, Richard Evans, Gustavo Henrique Paetzold, and Marcos Zampieri. 2021. SemEval-2021 task 1: Lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1–16, Online. Association for Computational Linguistics.

Sanjay S.P, Anand Kumar M, and Soman K P. 2016. AmritaCEN at SemEval-2016 task 11: Complex word identification using word embedding. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1022–1027, San Diego, California. Association for Computational Linguistics.

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023. Text classification via large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8990–9005, Singapore. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Gayatri Venugopal, Dhanya Pramod, and Ravi Shekhar. 2022. CWID-hi: A dataset for complex word identification in Hindi text. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5627–5636, Marseille, France. European Language Resources Association.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022a. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022b. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Brandon T Willard and Rémi Louf. 2023. Efficient guided generation for llms. *arXiv preprint arXiv:2307.09702*.

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018. A report on the complex word identification shared task 2018. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 66–78, New Orleans, Louisiana. Association for Computational Linguistics.

G. Zaharia, D. Cercel, and M. Dascalu. 2020. Cross-lingual transfer learning for complex word identification. In *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 384–390, Los Alamitos, CA, USA. IEEE Computer Society.

George-Eduard Zaharia, Răzvan-Alexandru Smădu, Dumitru Cercel, and Mihai Dascalu. 2022. Domain adaptation in multilingual and multi-domain monolingual settings for complex word identification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 70–80, Dublin, Ireland. Association for Computational Linguistics.

Jenny Alexandra Ortiz Zambrano, César Espin-Riofrio, and Arturo Montejo-Ráez. 2023. Legalec: A new corpus for complex word identification research in

law studies in ecuatorian spanish. *Proces. del Leng. Natural*, 71:247–259.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023. Large language models are human-level prompt engineers. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

A Prompting and Fine-Tuning

For zero-shot settings, we employed chain of thoughts (Wei et al., 2022b) to reduce hallucination and keep the model focused on the task. The model was asked to confirm the sentence and the word and then, before the final answer, to provide a short demonstration about the reason for the response. After we fine-tune the model, we follow a similar procedure, but we do not ask the model to produce a demonstration – only to confirm the task and directly provide the answer.

For fine-tuning, we prepare the dataset as follows. First, we discretize the probabilities as follows, similar to Shardlow et al. (2021): scores between 0 and 0.2 are very easy, between 0.2 and 0.4 are easy, between 0.4 and 0.6 are neutral, between 0.6 and 0.8 are difficult, and between 0.8 and 1 are very difficult. Next, we prepare the dataset following the prompt template specific to the model. For Llama 2 chat models, we followed the inference instructions specific to the model.

Llama 2 models were fine-tuned using QLoRA (Dettmers et al., 2023) with 4-bit quantization. R was set to 16, α to 32 and dropout to 0.05. The batch size was set between 10 and 32, and the learning rate using a linear scheduler with 10% warmup and a maximum value of 1e-4. Fine-tuning OpenAI's ChatGPT models involved uploading the training and validation files and starting the training job. No hyper-parameter could be changed. Fine-tuning defaulted to three epochs. We limited the training size to 250 samples uniformly sampled among labels from the train set specific to the dataset task and language.

A.1 Zero-shot Prompts

A.1.1 LCP English Prompt

> You are a helpful, honest, and respectful assistant for identifying the word complexity for non-native English speakers. You are given one sentence in English and a word from that sentence. Your task is to evaluate the complexity of the word. Answer with one of the following: very easy, easy, neutral, difficult, very difficult. Be concise. Please, answer using the following JSON format:

```
{
    "sentence": "the sentence you were provided",
    "word": "the word or words you have to analyze",
    "proof": "explain your response in maximum 50 words",
    complex": "either very easy, easy, neutral, difficult, or very difficult",
}
```

> What is the difficulty of ‘{token}‘ from ‘{sentence}‘?

A.1.2 CWI English Prompt

> You are a helpful, honest, and respectful assistant for identifying the words complexity for beginner English learners. You are given one sentence in English and a phrase from that sentence. Your task is to say whether the phrase is complex. Assess the answer for the phrase, given the context from the sentence. Be concise. Please, use the following JSON schema:

```
{
    "sentence": "the sentence you were provided",
    "word": "the word or words you have to analyze",
    "proof": "explain your response in maximum 50 words",
   "complex": "either false (for simple) or true (for complex)",
}
```

> Is ‘{token}‘ complex in ‘{sentence}‘?

A.1.3 CWI German Prompt

> Sie sind ein hilfsbereiter, ehrlicher und respektvoller Assistent, um die Wortkomplexität für Anfänger im Deutschen zu identifizieren. Sie erhalten einen Satz auf Deutsch und eine Phrase aus diesem Satz. Ihre Aufgabe ist es zu sagen, ob die Phrase komplex ist. Bewerten Sie die Antwort für die Phrase, anhand des Kontexts aus dem Satz. Seien Sie kurz. Bitte verwenden Sie das folgende JSON-Schema:

```
{
  "sentence": "der Satz, den Sie erhalten haben",
```

```
    "word": "das Wort oder die Wörter, die Sie analysieren müssen",
    "proof": "erklären Sie Ihre Antwort in maximal 50 Wörtern",
    "complex": "entweder false (für einfach) oder true (für komplex)",
}
```

Ist ‘{token}‘ von ‘{sentence}‘ complex?

A.1.4 CWI Spanish Prompt

Eres un asistente útil, honesto y respetuoso para identificar la complejidad de las palabras para los principiantes que aprenden español. Se te da una oración en español y una frase de esa oración. Tu tarea es decir si la frase es compleja. Evalúa la respuesta para la frase, dada el contexto de la oración. Sé conciso. Por favor, usa el siguiente esquema JSON:

```
{
    "sentence": "la oración que se te proporcionó",
    "word": "la palabra o palabras que tienes que analizar",
    "proof": "explica tu respuesta en máximo 50 palabras",
   "complex": "false (para simple) o true (para complejo)"
}
```

¿Es ‘{token}‘ complejo en ‘{sentence}‘?

A.2 Fine-Tune Prompts

A.2.1 LCP English Prompt

You are a helpful, honest, and respectful assistant for identifying the word difficulty for non-native English speakers. You are given one sentence in English and a word from that sentence. Your task is to evaluate the difficulty of the word. Answer only with one of the following: very easy, easy, neutral, difficult, very difficult.

sentence: ‘{sentence}‘
word: ‘{token}‘

A.2.2 CWI English Prompt

You are a helpful, honest, and respectful assistant for identifying the word complexity for non-native English speakers. You are given one sentence in English and a word from that sentence. Your task is to say whether a word is complex or not. Answer only with one of the following: yes, no.

sentence: ‘{sentence}‘
word: ‘{token}‘

A.2.3 CWI German Prompt

Du bist ein hilfsbereiter, ehrlicher und respektvoller Assistent für die Identifizierung der Wortkomplexität für nichtdeutsche Muttersprachler. Dir wird ein Satz auf Deutsch und ein Wort aus diesem Satz gegeben. Deine Aufgabe ist es zu sagen, ob ein Wort komplex ist oder nicht. Antworten nur mit einem der Folgenden: ja, nein.

Satz: ‘{sentence}‘
Wort: ‘{token}‘

A.2.4 CWI Spanish Prompt

Eres un asistente útil, honesto y respetuoso para identificar la complejidad de las palabras para hablantes no nativos de inglés. Se te da una oración en inglés y una palabra de esa oración. Tu tarea es decir si una palabra es compleja o no. Responde solo con una de las siguientes opciones: sí, no.

oracion: ‘{sentence}‘
palabra: ‘{token}‘

B LLMs' Task Understanding

In this section, we investigate what is the LLMs' level to understand the task firsthand before generating any output. That is, we check if the model, before providing the answer, can reproduce what it has to solve (the sentence and word pair). We report the sentence error count (S) and the word error count (W). In this setting, we mainly focus on the chat models. In our experiments, the fine-tuned models follow the provided instructions. Note that we do not include results for the multi-lingual datasets (i.e., German, Spanish) since these models could not produce a meaningful output.

In the CWI setting, we obtained an output using various packages such as Outlines (Willard and Louf, 2023), but it was not correlated with the examples, and the overall performance was not better than random. The results for the chat models on

English domains are presented in Table 4. When the model is larger, in general, the error rates decrease. The ChatGPT model obtains the lowest error counts, while the Llama 2 7b model obtains the highest. In general, the models struggle to understand what is the word they need to evaluate. Investigating the errors, we mostly see that the model considers more words than the target, for example, "America" (ground truth) vs "South America" (extracted by LLM). Other error cases we identified were completely different words to evaluate. For example, the target “years” was replaced by Llama-2-13b-chat with “Aegyptosaurus”. Text locality is not always the main reason; in the previous examples, in the first one, we have locality; in the second one, the words were in different parts of the sentence.

In the LCP setting, we considered all the sampling runs, and thus, we reported the average and standard deviation across those runs. We report lower absolute error counts. Similar to the previous setting, we note that the sentence error count is lower than the word error count, in most cases being closer to 0. In addition, ChatGPT achieves error counts very close to 0, meaning that the models understand the task it needs to solve. In the case of Llama-2-7b, the models struggle to recall the word.

Model	News		Wikinews		Wikipedia	
	S	W	S	W	S	W
Llama-2-7b-chat	50	245	120	190	61	85
Llama-2-13b-chat	36	225	44	173	93	125
chatgpt-3.5-turbo	2	47	4	17	0	10

Table 4: LLMs’ task understanding capabilities on the CWI English multi-domain dataset. The S column indicates wrong sentences, and the W column indicates wrong words.

Model	LCP-single		LCP-multi	
	S	W	S	W
Llama-2-7b-chat	$2.4_{\pm 0.7}$	$0.1_{\pm 0.2}$	$1.0_{\pm 0.2}$	$6.5_{\pm 0.5}$
Llama-2-13b-chat	$0_{\pm 0}$	$0.1_{\pm 0.3}$	$0_{\pm 0}$	$3.9_{\pm 1.2}$
chatgpt-3.5-turbo	$0_{\pm 0}$	$0_{\pm 0}$	$0.1_{\pm 0.3}$	$0_{\pm 0}$

Table 5: LLMs’ task understanding capabilities on the LCP English datasets. The S column indicates wrong sentences, and the W column indicates wrong words.

C LLMs’ Difficulty Understanding

In the zero-shot learning stage, before letting the model output the answer, we ask it to provide a brief proof regarding the choice. We ask first about the proof and then ask for the answer to enforce the model to "think before answer". If we let the model answer and then provide proof, the proof would have been influenced by the initial answer, which would have influenced the model’s internal bias. In Table 6, we show some examples of reasoning regarding the answer provided by Llama-2-13b-chat on the CWI English dataset. The proof motivates the answer in our setting, but we notice some flaws in the reasoning. For example, the model says that "ft" (i.e., feet as a unit of measurement) is common in English, but at the same time, it tends to contradict that being an abbreviation makes it difficult to understand. We notice this pattern quite often in the Llama models.

D Confusion Matrices on CWI

To investigate how the predictions are affected by the domain, language, and LLM, we generate the confusion matrices, which are shown in Figures 1 and 2. The general tendency is that chat models have higher false-positive or false-negative rates. The same model checkpoints have the same bias towards one false rate in the multi-domain setting. For example, Llama-2-7b-chat has a high false-positive rate, while Llama-2-13b-chat has a high false rate. Correlated with the proofs generated by the LLMs, this is motivated by the fact that LLMs tend to either overestimate or underestimate the difficulty of a word. This is especially true if the model finds a synonym for the target word. Also, the high false rates correlate with the model’s incapacity to understand the task in the zero-shot setting.

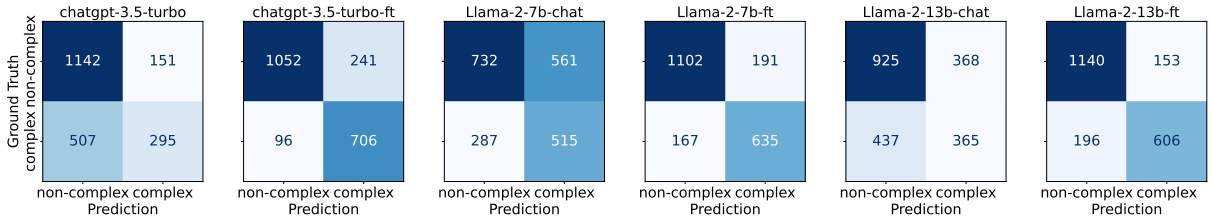
On the other hand, fine-tuned models show lower false-positive/negative rates, meaning that fine-tuning makes the model focus better on the task and learn latent instructions directly from the data.

E Fine-tuned Predictions on LCP

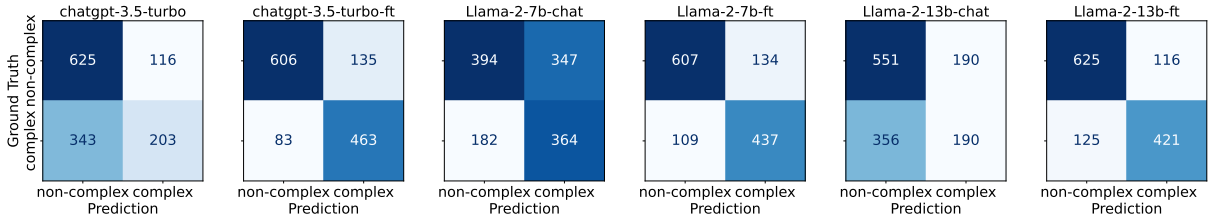
We analyzed the complexity probability distribution outputted by the LLMs in Figures 3, 6, 4, 5, 6. This is constructed by binning the models’ real-valued estimates (on the x-axis) and generating a histogram (on the y-axis). The discrete labels were mapped equidistantly in the range 0-1, i.e., very easy (VE) in 0-0.2, easy (E) in 0.2-0.4, neutral (N) in 0.4-0.6, difficult (D) in 0.6-0.8, and very difficult (VE) in 0.8-1. In gray, we indicate the outside of the expected label (i.e., wrong labels); in the white stripe, we indicate the correctly predicted labels.

Sentence	Word	Answer	Proof	Ground Truth
Toronto traded Stewart to Chicago early in the 1947-48 season in a deal that brought Max Bentley to the Leafs.	brought	False	The verb 'brought' is not complex as it is a commonly used verb in everyday language with a straightforward meaning	True
His use of Russian language formed the basis of the style of novelists Ivan Turgenev, Ivan Goncharov, and Leo Tolstoy, as well as that of subsequent lyric poets such as Mikhail Lermontov.	basis	False	The word 'basis' is not particularly complex, but the concept it represents (i.e., the foundation or starting point of something) may be unfamiliar to some beginner English learners	True
Vertical distance measurements in the "down" direction are commonly referred to as depth.	depth	True	The word 'depth' has multiple meanings, including a vertical extent or height, making it potentially complex for beginning English learners to understand without proper context or explanation	True
The lack of oxygen above 2,400 meters (8,000 ft) can cause serious illnesses such as altitude sickness, high altitude pulmonary edema, and high altitude cerebral edema.	ft	True	The abbreviation 'ft' is commonly used in English to represent feet, which is a unit of measurement. However, in this context, it may be challenging for beginners to understand because they might not be familiar with the abbreviation.	False

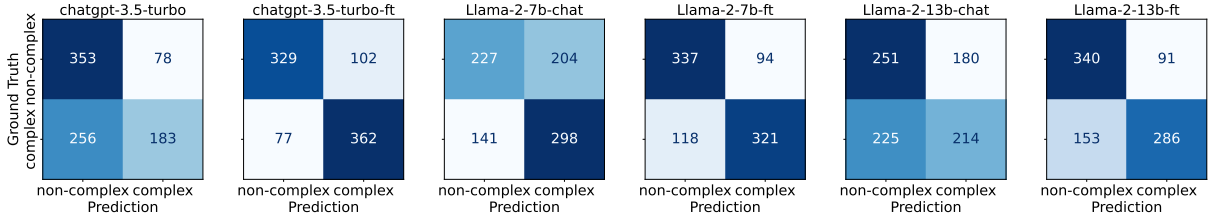
Table 6: Examples of predictions and proofs for the Llama-2-13b-chat model on the CWI English Wikipedia dataset.



(a) CWI 2018 EN News

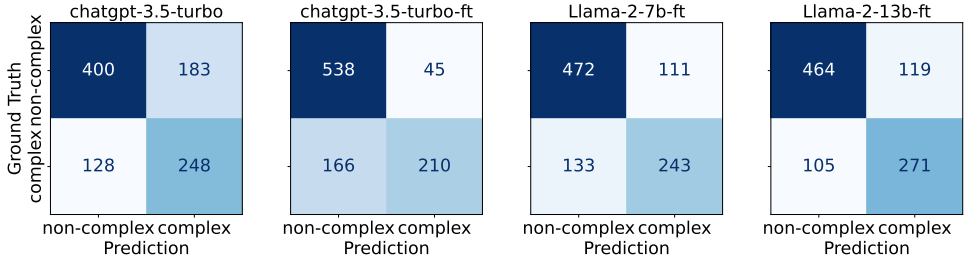


(b) CWI 2018 EN WikiNews

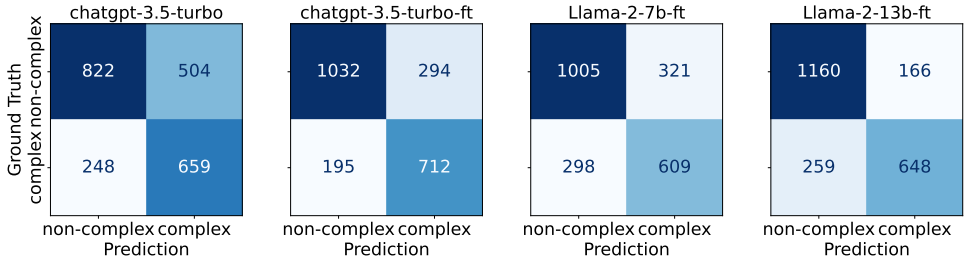


(c) CWI 2018 EN Wikipedia

Figure 1: Confusion matrices computed on the English CWI datasets for News, WikiNews, and Wikipedia domains.



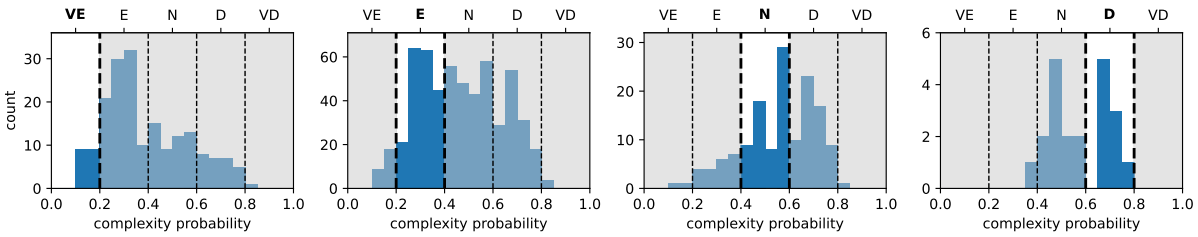
(a) CWI 2018 DE



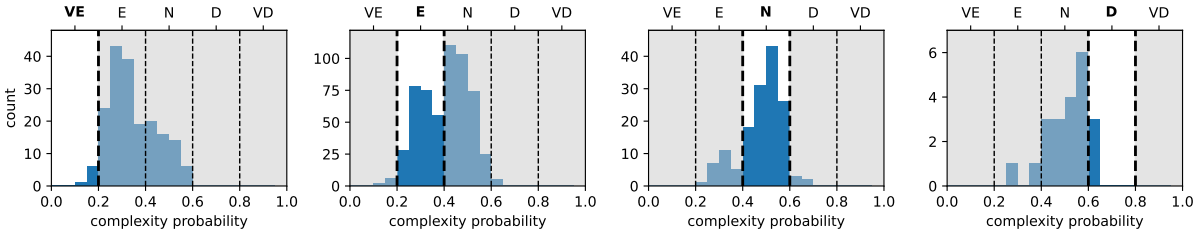
(b) CWI 2018 ES

Figure 2: Confusion matrices computed on the German and Spanish CWI datasets.

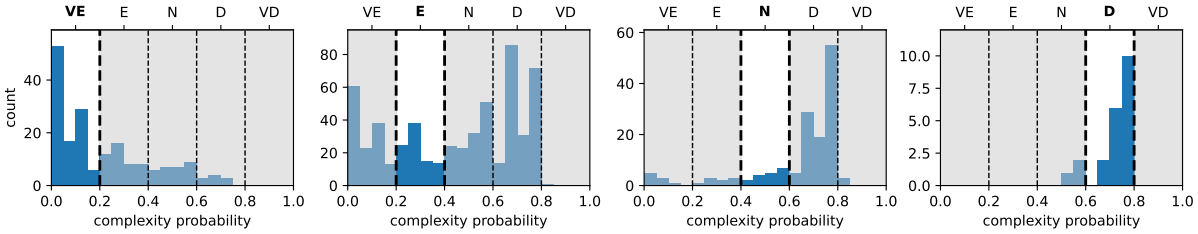
In the case of chat models, we notice a more uniform distribution among models' predictions, especially for the low-complexity words. The absolute error is more than one step in the difficulty scale. We notice that the models struggle to identify the very difficult label, regardless of whether the model was fine-tuned or not. In the fine-tuned setting, we notice that Llama-based models tend to misclassify neutral and difficult words, generally considering the words easier than the ground truth. Also, there is a tendency to label very easy words as easy. In the case of ChatGPT-3.5-turbo-ft, we notice that the outputs tend to be more deterministic – the majority of labels lie on the class scores.



(a) Llama-2-7b-chat



(b) Llama-2-13b-chat



(c) ChatGPT-3.5-turbo-chat

Figure 3: Predictive probability distribution of zero-shot LLMs on LCP single-word test set. Highlighted is the ground truth interval. Neither model predicts in the VD interval. VE – very easy, E – easy, N – neutral, D – difficult, VD – very difficult.

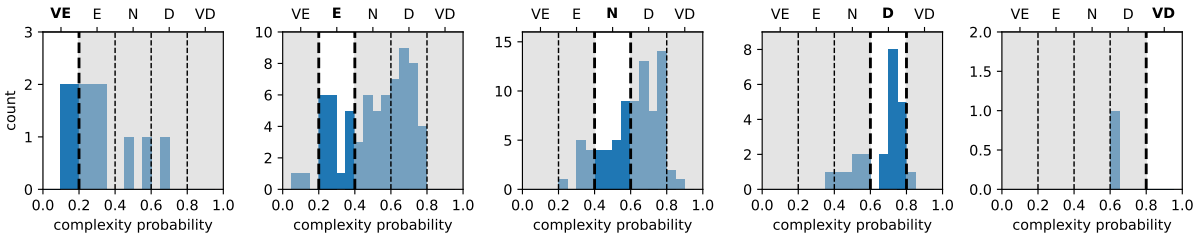
F Results Discussions

LLMs can grasp word complexity, depending on the model's capabilities. We observed that performances across domains, language, and whether we deal with a word or a phrase, are similar if the model is fine-tuned. In the zero-shot setting, the input prompt and prediction temperature yield a high variance across the results. Also, we noticed that sometimes the models (especially Llama-2-13b-chat, in the zero-shot setting) refused to answer some examples (especially in the Biblical domain) because of racial discrimination, despite that not being the case. Models tend to consider words easier than they are, mainly because if asked to explain the choice, they could find another synonym that is not necessarily simpler.

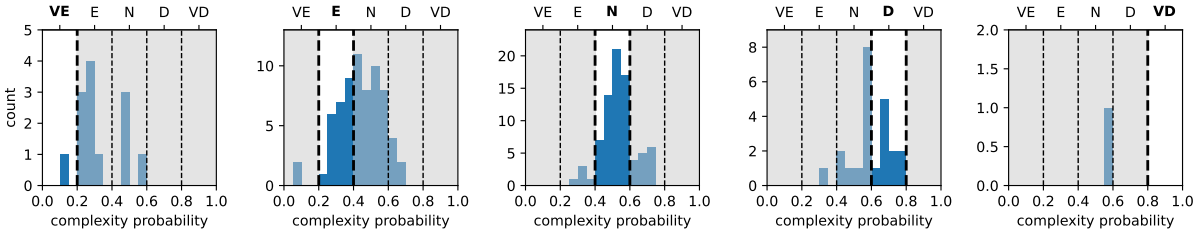
In addition, zero-shot prompting is achieved every time poor performances, and the main effect is that models tend to have a high false positive rate in the CWI task. This can be changed during fine-tuning when we notice that imbalanced datasets towards a class lead to the model being biased and producing more often the predominant label from the fine-tuning set.

G Choice for Number of Inference Steps

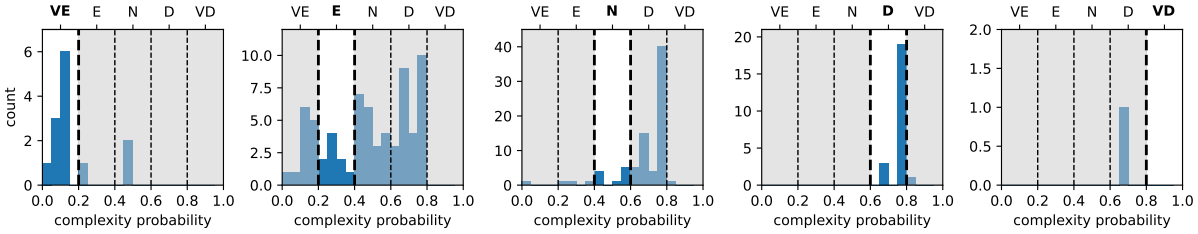
As presented in Section 3, the estimated score in the LCP setting was an average of scores obtained after N inference steps. We wanted to know what is the minimum number of inference steps required until the results do not change significantly anymore. Therefore, we set $N = 25$ for Llama-2-7b-ft, $N = 20$ for Llama-2-13b-ft, and $N = 10$ for ChatGPT-3.5-turbo-ft, and then estimated the average score per number of iterations using bootstrapping, with 100 samples. The plots are shown in Figure 7. We obtained that at least 10 to 15 runs are required, after which the scores do not change significantly.



(a) Llama-2-7b-chat

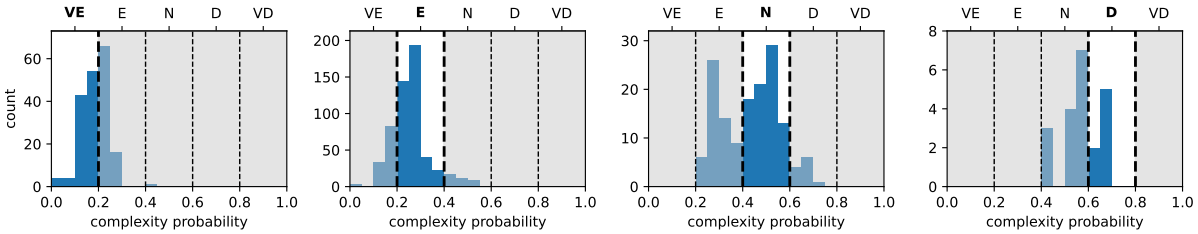


(b) Llama-2-13b-chat

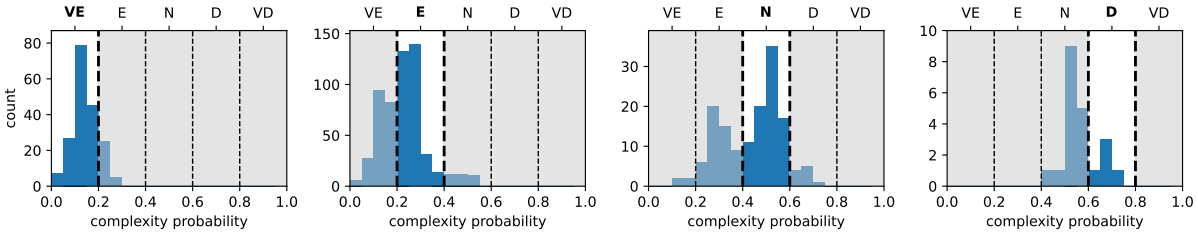


(c) ChatGPT-3.5-turbo

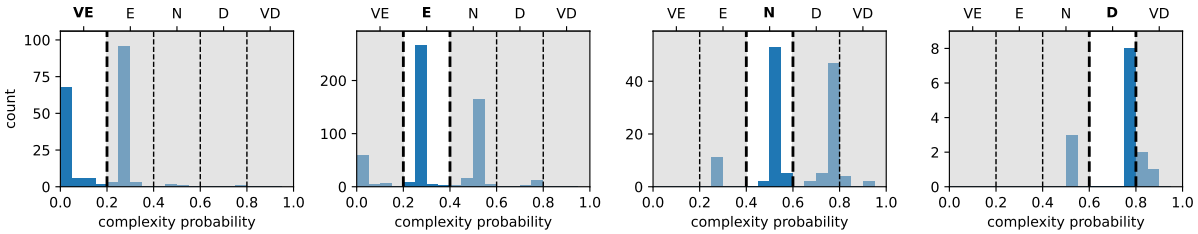
Figure 4: Predictive probability distribution of zero-shot LLMs on LCP multi-word test set. Highlighted is the ground truth interval. Neither model predicts in the VD interval. VE – very easy, E – easy, N – neutral, D – difficult, VD – very difficult.



(a) Llama-2-7b-ft

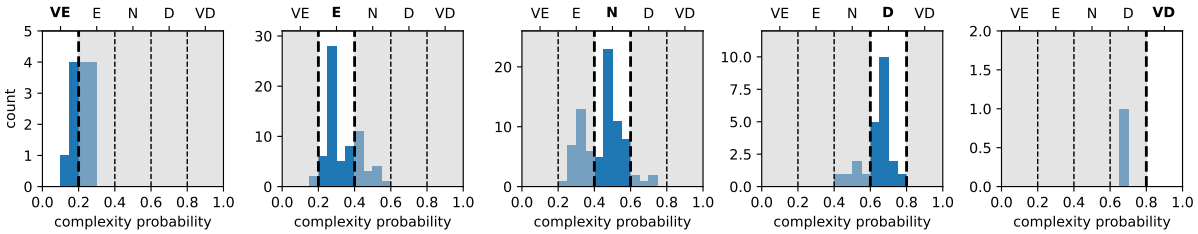


(b) Llama-2-13b-ft

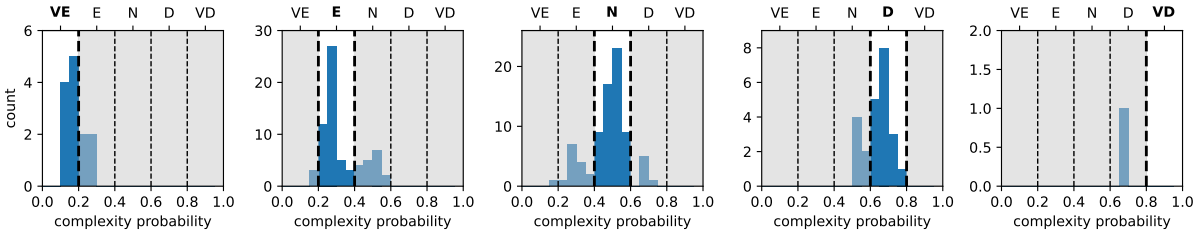


(c) ChatGPT-3.5-turbo-ft

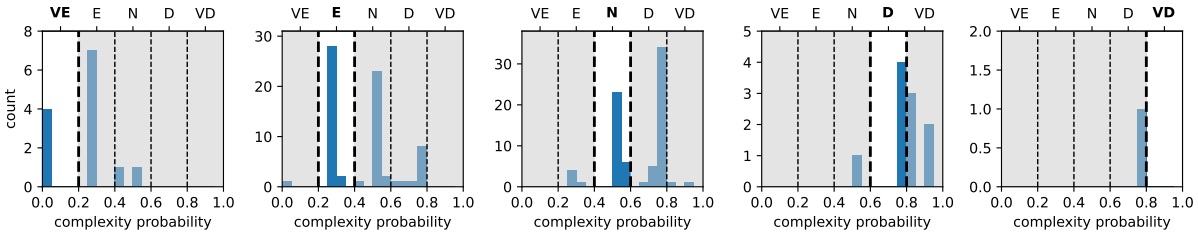
Figure 5: Predictive probability distribution of fine-tuned LLMs on LCP single-word test set. Highlighted is the ground truth interval. Neither model predicts in the VD interval. VE - very easy, E - easy, N - neutral, D - difficult, VD - very difficult.



(a) Llama-2-7b-ft

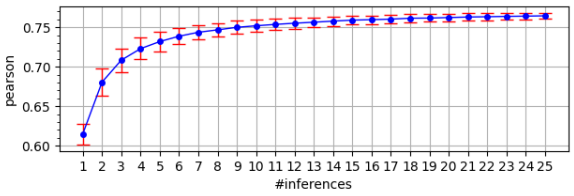


(b) Llama-2-13b-ft

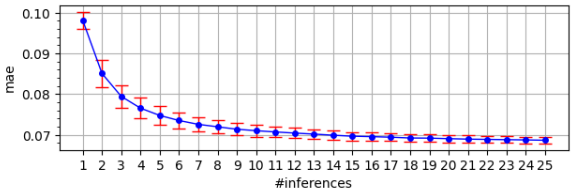


(c) ChatGPT-3.5-turbo-ft

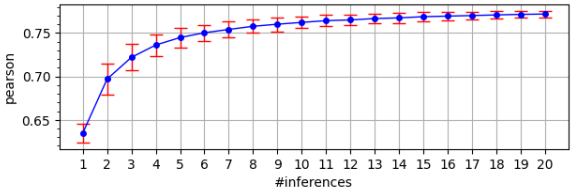
Figure 6: Predictive probability distribution of fine-tuned LLMs on LCP multi-word test set. Highlighted is the ground truth interval. Neither model predicts in the VD interval. VE - very easy, E - easy, N - neutral, D - difficult, VD - very difficult.



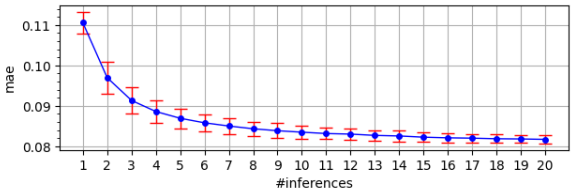
(a) Pearson score on Llama-2-7b-ft



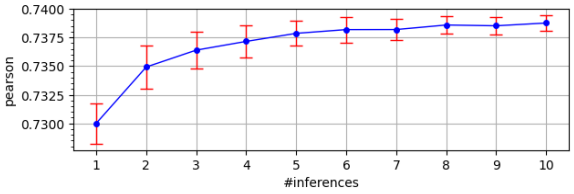
(b) MAE score on Llama-2-7b-ft



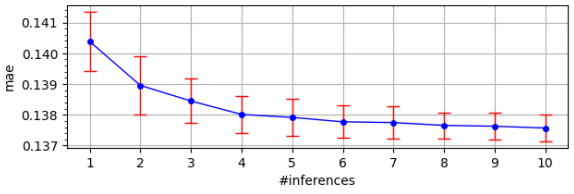
(c) Pearson score on Llama-2-13b-ft



(d) MAE score on Llama-2-13b-ft



(e) Pearson score on ChatGPT-3.5-turbo-ft



(f) MAE score on ChatGPT-3.5-turbo-ft

Figure 7: Estimated Pearson and MAE scores against the number of LLM inference steps.