FiRC-NLP at SemEval-2025 Task 3: Exploring Prompting Approaches for Detecting Hallucinations in LLMs

Wondimagegnhue Tufa[†◇], **Fadi Hassan**[†⋆], **Guillem Collell**[⋆], **Kuan Eeik Tan**[⋆], **Ni Sang** [⋆], **Tu Yi** [⋆], and **Tu Dandan** [⋆]

[◇]Faculty of Humanities, Vrije Universiteit Amsterdam
[⋆]Huawei Technologies Finland Research Center
{w.t.tufa}@vu.nl
{firstname.lastname}@huawei.com

Abstract

This paper presents a system description for the SemEval Mu-SHROOM task, focusing on detecting hallucination spans in the outputs of instruction-tuned Large Language Models (LLMs) across 14 languages. We compare two distinct approaches: Prompt-Based Approach (PBA), which leverages the capability of LLMs to detect hallucination spans using different prompting strategies, and the Fine-Tuning-Based Approach (FBA), which fine-tunes pre-trained Language Models (LMs) to extract hallucination spans in a supervised manner. Our experiments reveal that PBA, especially when incorporating explicit references or external knowledge, outperforms FBA. However, the effectiveness of PBA varies across languages, likely due to differences in language representation within LLMs.

1 Introduction

Large Language Models (LLMs) have brought advancements to many areas of NLP, including natural language understanding, natural language generation, and reasoning tasks (Naveed et al., 2024; Zhao et al., 2024; Minaee et al., 2024). Broadly speaking, LLMs such as GPT-4 (OpenAI et al., 2024) and LLaMA (Touvron et al., 2023) are based on transformer models and are trained on vast amounts of internet text to understand and generate human language in a coherent and contextually relevant manner (Brown et al., 2020; Chowdhery et al., 2022). The increasing scale of training data and model capacity has enabled LLMs to exhibit emergent capabilities such as chain-of-thought reasoning, instruction following, and in-context learning (Wei et al., 2023; Brown et al., 2020; Peng et al., 2023).

Currently, Natural Language Generation (NLG) faces a major challenge: Large Language Models (LLMs) can produce text that is fluent and coherent but contains factual inaccuracies or statements ungrounded in reality—a phenomenon known as hallucination (Rawte et al., 2023; Huang et al., 2025). These hallucinations are difficult to detect automatically because existing evaluation methods primarily measure fluency rather than accuracy. Detecting hallucinations is often the first step in ensuring that a model's output is consistent with known facts and in preventing the generation of misleading or false information (Chang et al., 2023). In many NLG applications, such as question-answering and translation tasks, the correctness of the model's output is crucial for its utility.

The SemEval Mu-SHROOM task focuses on detecting hallucination spans in the outputs of instruction-tuned LLMs across 14 languages (Vázquez et al., 2025). The main goal of the task is to determine which spans of a given text produced by an LLM are part of a hallucination. The organizers provide the LLM output as a string of characters, a list of tokens, and a list of logits. Participants are required to compute the probability of hallucination for each character in the LLM-generated text. Submissions are evaluated using two approaches: (1) the intersection-over-union of characters marked as hallucinations in the gold reference and the predicted output, and (2) the correlation between the probability assigned by the participants' system that a character is part of a hallucination and the empirical probabilities observed from annotators.

We explore two distinct approaches to address the Mu-SHROOM task: the Prompt-Based Approach (PBA) and the Fine-Tuning-Based Approach (FBA). In the Prompt-Based Approach, we experiment with different strategies. First, we explore prompting an instruction-tuned model without a reference by providing only the question and the model's answer. For this, we use GPT-4 (OpenAI et al., 2024) to identify hallucination spans by providing the input text and the model's output

[†]These authors contributed equally to this work.

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 2096–2102

	AR	CA	CS	DE	EN	ES	EU	FA	FI	FR	HI	IT	SV	ZH
Train	0	0	0	0	808	492	0	0	0	1850	0	0	0	210
Valid	50	0	0	50	50	50	0	0	50	50	0	50	49	50
Test	150	100	100	150	154	152	99	100	150	150	150	150	147	150

Table 1: The distribution of training, validation, and test data for different languages. Only four of the fourteen languages (English, Spanish, French, and Chinese) have both training and validation sets.

text. In this approach, since no reference context is provided, the model is expected to rely implicitly on its pre-trained knowledge to determine which parts of the output constitute hallucinations. The details of this approach are described in Section 3.1.

In the second strategy, which we refer to as the dual-prompt approach, we break down the prompt into two phases. In the first phase, we prompt the model to answer the question explicitly. In the second phase, we use the output from the first phase as a reference answer and prompt the model to identify hallucinations by comparing this reference with the model's original output.

In the third approach, we incorporate external knowledge to create a reference context and prompt the model to answer the question based on this reference.

For the Fine-Tuning-Based Approach, we experiment with an encoder model by framing the task as a token classification task. We describe the details in Section 3.1.

In summary, our experiment shows that:

- Prompting LLMs to identify hallucinations without providing a reference or context results in more hallucinations. We hypothesize that this may be caused by the limitations of LLMs in implicitly recalling knowledge correctly without explicit prompting, which is crucial since no additional context is provided.
- We test this hypothesis by using dual prompts to make implicit knowledge recall explicit. We observe that providing an explicit reference from the target LLM significantly improves detection performance in most of our target languages.
- We further experiment by providing a RAG context in our prompt instead of prompting the model for a reference. We observe that providing a RAG-like context with the prompt further improves model performance in identifying hallucination spans.

2 Background

Hallucination is one of the main limitations of NLG models, where the generated text sounds fluent and coherent but contains factual inaccuracies or statements ungrounded in reality (Rawte et al., 2023; Huang et al., 2025). Hallucinations in NLG models can take two forms: intrinsic hallucinations and extrinsic hallucinations (Ji et al., 2023; Dziri et al., 2021).

In the case of intrinsic hallucinations, there is a contradiction between the source text and the generated text. Since this contradiction appears in one or more spans, it is possible to verify where the hallucination occurred. In contrast, extrinsic hallucinations do not exhibit an observable contradiction between the source and the generated text, making it impossible to pinpoint the hallucination. In the extrinsic case, there is no available evidence in the input text to determine the correctness or incorrectness of the generated text.

2.1 Hallucination Detection

Various approaches have been introduced to address hallucination in NLG, with knowledge-based methods being the most commonly used (Ji et al., 2023).

In knowledge-based approaches, a domain-specific knowledge base is used to fact-check the model-generated text. This approach is effective but is only applicable to domains where a relevant knowledge base is available. In cases where the LLM's hidden state is accessible, white-box and grey-box analyses can be employed by training an MLP classifier on the hidden state to predict truthfulness (Azaria and Mitchell, 2023).

In grey-box methods, the probability of the tokens generated by LLMs is used to detect hallucinations based on the assumption that correct text consists of high-probability tokens. In the self-evaluation approach, the LLM itself is prompted to score the likelihood of the text it generated (Kadavath et al., 2022). Similarly, any public model can be used as a proxy to assess the factuality of the

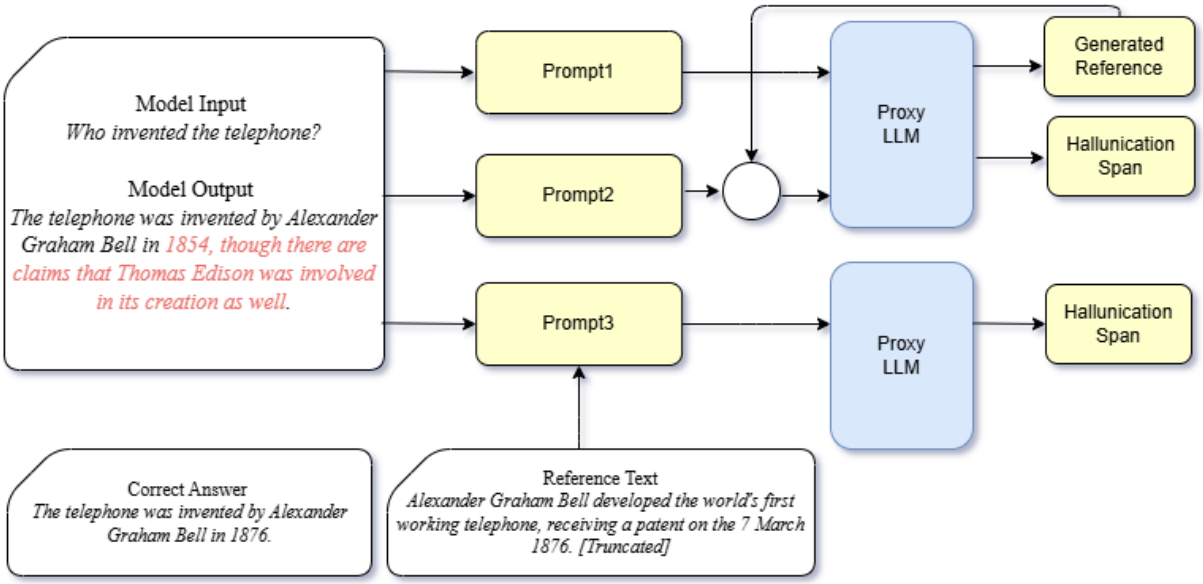

Figure 1: Overview of the prompt-based approach. In the first approach (Prompt-1), we directly prompt the proxy model to identify hallucination spans by providing the model input and output. In the second approach (Prompt-2), we first prompt the model to generate an answer based on the model input, and then use this generated reference to prompt the proxy model to identify hallucinations. In the third approach (Prompt-3), we use an external reference text along with the model input and output to prompt the proxy model to identify hallucinations.

generated text by estimating the token probability of a black-box model.

SelfCheckGPT (Manakul et al., 2023) is another black-box, sampling-based approach. It relies on the hypothesis that if an LLM has correct knowledge of a particular topic, sampled responses on that topic will have high similarity, whereas hallucinated text will diverge significantly.

2.2 Task Description

The SemEval Mu-SHROOM task focuses on the detection of hallucination spans in the outputs of instruction-tuned LLMs across 14 languages: Arabic, Basque, Catalan, Chinese (Mandarin), Czech, English, Farsi, Finnish, French, German, Hindi, Italian, Spanish, and Swedish (Vázquez et al., 2025). The data distribution across the splits is provided in Table 1.

The following data points are provided as part of the challenge:

- **Model Input**: A prompt provided to the model to generate text.
- **Model Id**: The name of the models used to produce each output. Two models are used: TheBloke/Mistral-7B-Instruct-v0.2-GGUF and TheBloke/SauerkrautLM-7B-v1-GGUF.
- **Model Output**: A string of characters, a list of tokens, and a list of logits.
- **Hard Labels**: A label of 1 is assigned when the corresponding span contains a hallucination. We determine hard labels using majority voting among the annotators.
- **Soft Labels**: The confidence-based judgments of the annotators. Calculated as the proportion of annotators who marked the span as part of a hallucination out of the total number of annotators.
- **Evaluation**: Submissions are evaluated using intersection-over-union (IOU) of characters marked as hallucinations in the gold reference and predicted, and the probability assigned by the participants' system that a character is part of a hallucination correlates with the empirical probabilities observed in the annotation. For hard labels, intersection-over-union (IoU) is used and the Spearman correlation between predicted and reference soft labels is used for soft labels.

2.3 Models

We employ GPT-4o-mini from OpenAI as a proxy model for all prompt-based experiments. For fine-tuning, we use XLM-R (Conneau et al., 2019) as

the base model and fine-tune it by framing the task as a token classification task. Additionally, we utilize Perplexity AI with search capability to generate more accurate reference text.

3 System Description

In this section, we describe our proposed system. We employ two distinct approaches: the Prompt-Based Approach (PBA) and the Fine-Tuning-Based Approach (FBA). In the prompt-based method, we use an LLM as a proxy model and apply a standard prompt to identify hallucination spans by providing pairs of input text and model output. In the fine-tuning approach, we fine-tune an encoder model by framing the task as a token classification problem.

3.1 Prompt-Based Approach

Figure 1 shows an overview of our PBA approach. We experiment with three prompt strategies: prompting without a reference, dual prompting, and prompting with an external reference.

Prompt without Reference (PWR) In this approach, we design a simple prompt and request a proxy model to identify hallucination spans by providing the model input and output. Since no reference text is provided, the proxy model implicitly relies on its pre-trained knowledge to answer the question correctly and compare this answer with the provided output to determine which parts of the text contain hallucinations. We test the hypothesis that a proxy model can reliably identify hallucinations in text generated by another LLM.

Dual Prompting (DP) In this approach, we modify the first method by splitting the prompt into two parts. In the first part, we prompt the proxy model to generate an answer by providing the original model input. In the second part, we prompt the model to identify hallucinations by comparing the generated answer with the model output. By explicitly prompting for the answer, we can assess whether the proxy model relies on correct knowledge or introduces errors when comparing the reference with the model output.

Prompting with External Reference (PEXT) In this approach, we use external knowledge to create a reference text. We utilize an API from Perplexity, which has search capabilities, to generate the reference text. We then use this reference to prompt the proxy model to identify hallucination spans.

Perplexity AI integrates an LLM with internet search capabilities to retrieve reference text from external sources.

Fine-Tuning-Based Approach (FBA) For the fine-tuning-based approach, we fine-tune a multilingual encoder model by framing the task as a token classification problem. Specifically, we use XLM-R as the base model. We combine the training sets for four languages—English, Spanish, French, and German—and fine-tune the model on this multilingual dataset. The input consists of tokenized model outputs, and the objective is to predict, for each token, whether it is part of a hallucination span.

4 Analysis and Conclusion

In this section, we compare the strengths and limitations of the four proposed approaches. Table 2 presents the performance of these approaches in detecting hallucinations across 14 languages, evaluated using two metrics: Intersection over Union (IoU) and Correlation (Cor). We analyze the performance differences between the approaches and examine variations across languages, providing possible explanations for these differences.

4.1 Prompting Approaches

PWR The PWR approach exhibits variable performance across languages. In terms of IoU, it achieves the highest scores in languages such as French and Hindi but performs notably worse in languages like Chinese. Similarly, the correlation scores align with the IoU results, showing strong performance in French and Hindi but weaker performance in Chinese. Overall, while PWR demonstrates strong performance with high correlation in certain languages, its effectiveness remains inconsistent across languages.

Dual Prompting The DP method consistently outperforms PWR across both metrics, achieving high IoU scores in languages such as Hindi and French. These improvements can be attributed to explicitly prompting the model for a reference text, which reduces ambiguity and minimizes potential hallucinations. The correlation scores follow a similar trend, demonstrating relatively stable performance across languages.

PEXT The PEXT approach further improves upon the DP approach. One possible explanation for this improvement is that when the proxy model lacks the correct answer, incorporating reliable external knowledge helps bridge the gap. This pre-

Method	Metric	AR	CA	CS	DE	EN	ES	EU	FA	FI	FR	HI	IT	SV	ZH
PWR	IoU	38.58	49.61	29.95	39.69	38.77	34.15	47.06	58.93	45.15	40.94	65.76	65.12	51.45	27.82
	Cor	34.49	54.18	31.75	40.37	40.97	41.83	44.01	56.89	44.24	45.18	67.68	69.73	43.20	19.38
DP	IoU	**49.07**	57.74	38.83	51.14	49.15	39.61	53.44	63.10	60.38	55.06	67.58	70.29	61.54	42.09
	Cor	**42.29**	64.09	42.94	51.34	51.73	53.53	49.71	66.31	53.90	55.03	71.45	71.00	38.42	**29.18**
PEXT	IoU	48.92	**63.67**	**45.58**	**53.26**	**49.24**	**43.11**	**55.66**	**65.87**	**61.71**	**60.88**	**69.97**	**74.11**	**62.38**	41.41
	Cor	42.07	**68.55**	**50.08**	**54.13**	**52.37**	**54.34**	**54.34**	**65.48**	**54.30**	**60.22**	**74.79**	**75.80**	**44.19**	28.27
FBA	IoU	33.54	20.51	23.80	29.98	4.06	9.88	21.13	19.62	34.38	33.06	20.73	26.47	43.46	**43.81**
	Cor	9.62	14.02	23.44	22.48	-0.08	3.58	3.75	5.85	13.30	9.60	5.68	12.93	6.73	23.89

Table 2: Performance of the four approaches for hallucination span detection on the test set across 14 languages. PWR refers to prompting without a reference text, DP denotes dual prompting, PEXT indicates prompting with external context, and FBA corresponds to the fine-tuning-based approach.

vents the model from relying on incorrect or hallucinated information when identifying hallucinations. PEXT performs similarly to DP in most target languages, with IoU and correlation scores often comparable to those of DP. However, like DP, it struggles with Chinese (41.41 IoU, 28.27 Cor).

FBA The FBA approach shows the lowest scores across all languages in both IoU and correlation metrics. IoU values are particularly low, especially for English (4.06) and French (20.73). Similarly, correlation scores are weak, with negative values in languages such as English (-0.08), further indicating that FBA is not well-suited for hallucination detection. Despite being fine-tuned specifically for this task, the poor performance suggests that fine-tuning encoder models may not be the most effective strategy for hallucination detection, at least within the current setup.

4.2 Cross-lingual Analysis

The performance variation of prompt-based methods across languages reflects differences in the proxy LLM's ability to analyze text in different languages. We hypothesize two possible explanations for this variation. First, the type of questions used in the prompt may vary across languages, leading to discrepancies in generating accurate reference texts. For instance, the distribution of simpler or easier questions might favor certain languages. Second, LLMs do not perform equally well across all languages, favoring high-resource languages that are better represented in the model's training data. For example, the PWR approach relies solely on the prompt without additional context or external references. The variation in hallucination detection performance suggests that languages with higher representation in the training data tend to achieve better results, as the proxy LLM is more effective at understanding the task even with a simplified prompt.

5 Conclusion

In this work, we investigate the efficacy of prompt-based and fine-tuning-based approaches for detecting hallucinations in instruction-tuned LLMs, using the SemEval Mu-SHROOM task across 14 languages as a benchmark. Our findings indicate that prompt-based approaches (PBAs), particularly those leveraging explicit references or external knowledge, outperform the fine-tuning-based approach (FBA). Providing explicit references enhances a model's ability to pinpoint hallucination spans, while prompting without references leads to a higher incidence of hallucinations. Furthermore, incorporating external knowledge improves the identification of hallucination spans.

As future work, exploring hybrid approaches could be highly beneficial. Combining the strengths of both prompt-based and fine-tuning-based methods might lead to improved performance. For instance, a fine-tuned model could be integrated with a knowledge base system, where the knowledge base generates reference answers and the fine-tuned model uses both the LLM-generated output and the reference to identify hallucinations.

Limitations

Our study has several limitations. First, the prompt-based approaches heavily rely on a single proxy model (GPT-4o-mini), making the system's effectiveness dependent on the proxy's multilingual capabilities and potential biases. Second, the fine-tuning-based approach was implemented with a relatively simple setup using XLM-R, without explor-

ing more advanced strategies. Third, while external references in the PEXT approach were retrieved via Perplexity AI, no rigorous filtering was applied, introducing the possibility of noisy or irrelevant knowledge negatively affecting performance. Additionally, our system exhibited variability across languages, particularly for lower-resource or typologically distinct languages like Chinese, highlighting challenges in cross-lingual generalization. Finally, our evaluation was limited to the Mu-SHROOM dataset, and further validation on broader hallucination detection benchmarks or real-world outputs remains an important direction for future work.

Acknowledgments

This work was conducted as part of an internship by Wondimagegnhue Tufa at Huawei Technologies Finland Research Center. We thank the Huawei Research team for providing computational resources and technical support throughout the project. We also express our gratitude to the SemEval-2025 Task 3 organizers for creating the multilingual hallucination detection benchmark and offering valuable guidelines. Finally, we appreciate the feedback from anonymous reviewers, which helped improve the quality of this paper.

References

Amos Azaria and Tom Mitchell. 2023. The internal state of an llm knows when it's lying. *Preprint*, arXiv:2304.13734.

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. *Preprint*, arXiv:2005.14165.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2023. A survey on evaluation of large language models. *Preprint*, arXiv:2307.03109.

Aakanksha Chowdhery et al. 2022. Palm: Scaling language modeling with pathways. *Preprint*, arXiv:2204.02311.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Nouha Dziri, Andrea Madotto, Osmar Zaïane, and Avishek Joey Bose. 2021. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2197–2214, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Saurav Kadavath, Tom Conerly, et al. 2022. Language models (mostly) know what they know. *Preprint*, arXiv:2207.05221.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *Preprint*, arXiv:2303.08896.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *Preprint*, arXiv:2402.06196.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2024. A comprehensive overview of large language models. *Preprint*, arXiv:2307.06435.

OpenAI, Josh Achiam, et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *Preprint*, arXiv:2304.03277.

Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models. *Preprint*, arXiv:2309.05922.

Hugo Touvron, , et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Raúl Vázquez, Timothee Mickus, Elaine Zosa, Teemu Vahtola, Jörg Tiedemann, Aman Sinha, Vincent Segonne, Fernando Sánchez-Vega, Alessandro Raganato, Jindřich Libovický, Jussi Karlgren, Shaoxiong Ji, Jindřich Helcl, Liane Guillou, Ona de Gibert, Jaione Bengoetxea, Joseph Attieh, and Marianna Apidianaki. 2025. SemEval-2025 Task 3: Mu-SHROOM, the multilingual shared-task on hallucinations and related observable overgeneration mistakes.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2024. A survey of large language models. *Preprint*, arXiv:2303.18223.