Do Large Language Models Know What They Don't Know?

Zhangyue Yin[◇] **Qiushi Sun**[♠] **Qipeng Guo**[◇]
Jiawen Wu[◇] **Xipeng Qiu**[◇*] **Xuanjing Huang**[◇]
[◇]School of Computer Science, Fudan University
[♠]Department of Mathematics, National University of Singapore
{yinzy21,jwwu21}@m.fudan.edu.cn qiushisun@u.nus.edu
{qpguo16,xpqiu,xjhuang}@fudan.edu.cn

Abstract

Large language models (LLMs) have a wealth of knowledge that allows them to excel in various Natural Language Processing (NLP) tasks. Current research focuses on enhancing their performance within their existing knowledge. Despite their vast knowledge, LLMs are still limited by the amount of information they can accommodate and comprehend. Therefore, the ability to understand their own limitations on the unknows, referred to as self-knowledge, is of paramount importance. This study aims to evaluate LLMs' self-knowledge by assessing their ability to identify unanswerable or unknowable questions. We introduce an automated methodology to detect uncertainty in the responses of these models, providing a novel measure of their self-knowledge. We further introduce a unique dataset, *SelfAware*, consisting of unanswerable questions from five diverse categories and their answerable counterparts. Our extensive analysis, involving 20 LLMs including GPT-3, InstructGPT, and LLaMA, discovering an intrinsic capacity for self-knowledge within these models. Moreover, we demonstrate that in-context learning and instruction tuning can further enhance this self-knowledge. Despite this promising insight, our findings also highlight a considerable gap between the capabilities of these models and human proficiency in recognizing the limits of their knowledge.

"True wisdom is knowing what you don't know."
–*Confucius*

1 Introduction

Recently, Large Language Models (LLMs) such as GPT-4 (OpenAI, 2023), PaLM 2 (Anil et al., 2023), and LLaMA (Touvron et al., 2023) have shown exceptional performance on a wide range of NLP tasks, including common sense reasoning (Wei et al., 2022; Zhou et al., 2022) and mathematical problem-solving (Lewkowycz et al., 2022; Chen et al., 2022). Despite their ability to learn from huge amounts of data, LLMs still have limitations in their capacity to retain and understand information. To ensure responsible usage, it is crucial for LLMs to have the capability of recognizing their limitations and conveying uncertainty when responding to unanswerable or unknowable questions. This acknowledgment of limitations, also known as "*knowing what you don't know*," is a crucial aspect in determining their practical applicability. In this work, we refer to this ability as model self-knowledge.

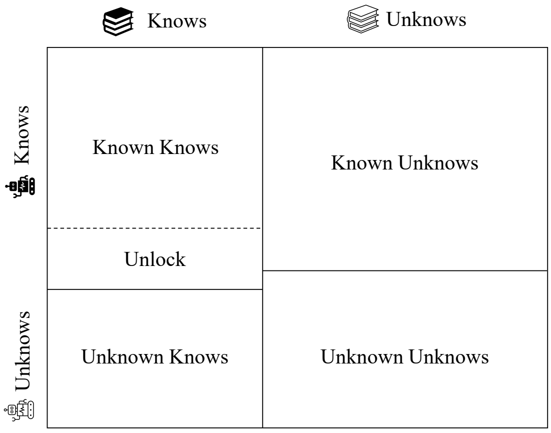

Figure 1: Know-Unknow Quadrant. The horizontal axis represents the model's memory capacity for knowledge, and the vertical axis represents the model's ability to comprehend and utilize knowledge.

The Know-Unknow quadrant in Figure 1 illustrates the relationship between the model's knowledge and comprehension. The ratio of "Known Knows" to "Unknown Knows" demonstrates the model's proficiency in understanding and applying existing knowledge. Techniques such as Chain-of-Thought (Wei et al., 2022), Self-Consistency (Wang et al., 2022), and Complex CoT (Fu et al., 2022) can be utilized to increase

[*] Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2023, pages 8653–8665
July 9-14, 2023 ©2023 Association for Computational Linguistics

this ratio, resulting in improved performance on NLP tasks. We focus on the ratio of "Known Unknows" to "Unknown Unknows", which indicates the model's self-knowledge level, specifically understanding its own limitations and deficiencies in the unknows.

Existing datasets such as SQuAD2.0 (Rajpurkar et al., 2018) and NewsQA (Trischler et al., 2017), widely used in question answering (QA), have been utilized to test the self-knowledge of models with unanswerable questions. However, these questions are context-specific and could become answerable when supplemented with additional information. Srivastava et al. (2022) attempted to address this by evaluating LLMs' competence in delineating their knowledge boundaries, employing a set of 23 pairs of answerable and unanswerable multiple-choice questions. They discovered that these models' performance barely surpassed that of random guessing. Kadavath et al. (2022) suggested probing the self-knowledge of LLMs through the implementation of a distinct "Value Head". Yet, this approach may encounter difficulties when applied across varied domains or tasks due to task-specific training. Consequently, we redirect our focus to the inherent abilities of LLMs, and pose the pivotal question: "*Do large language models know what they don't know?*".

In this study, we investigate the self-knowledge of LLMs using a novel approach. By gathering reference sentences with uncertain meanings, we can determine whether the model's responses reflect uncertainty using a text similarity algorithm. We quantified the model's self-knowledge using the F1 score. To address the small and idiosyncratic limitations of existing datasets, we created a new dataset called *SelfAware*. This dataset comprises 1,032 unanswerable questions, which are distributed across five distinct categories, along with an additional 2,337 questions that are classified as answerable. Experimental results on GPT-3, InstructGPT, LLaMA, and other LLMs demonstrate that in-context learning and instruction tuning can effectively enhance the self-knowledge of LLMs. However, the self-knowledge exhibited by the current state-of-the-art model, GPT-4, measures at 75.47%, signifying a notable disparity when contrasted with human self-knowledge, which is rated at 84.93%.

Our key contributions to this field are summarized as follows:

- We have developed a new dataset, *SelfAware*, that comprises a diverse range of commonly posed unanswerable questions.

- We propose an innovative evaluation technique based on text similarity to quantify the degree of uncertainty inherent in model outputs.

- Through our detailed analysis of 20 LLMs, benchmarked against human self-knowledge, we identified a significant disparity between the most advanced LLMs and humans [1].

2 Dataset Construction

To conduct a more comprehensive evaluation of the model's self-knowledge, we constructed a dataset that includes a larger number and more diverse types of unanswerable questions than Know-Unknowns dataset (Srivastava et al., 2022). To facilitate this, we collected a corpus of 2,858 unanswerable questions, sourced from online platforms like Quora and HowStuffWorks. These questions were meticulously evaluated by three seasoned annotation analysts, each operating independently. The analysts were permitted to leverage external resources, such as search engines. To ensure the validity of our dataset, we retained only the questions that all three analysts concurred were unanswerable. This rigorous process yielded a finalized collection of 1,032 unanswerable questions.

In pursuit of a comprehensive evaluation, we opted for answerable questions drawn from three datasets: SQuAD (Rajpurkar et al., 2016), HotpotQA (Yang et al., 2018), and TriviaQA (Joshi et al., 2017). Our selection was guided by SimCSE (Gao et al., 2021), which allowed us to identify and select the answerable questions semantically closest to the unanswerable ones. From these sources, we accordingly drew samples of 1,487, 182, and 668 questions respectively, amassing a total of 2,337. Given that these questions can be effectively addressed using information available on Wikipedia, the foundational corpus for the training of current LLMs, it is plausible to infer that the model possesses the requisite knowledge to generate accurate responses to these questions.

Our dataset, christened *SelfAware*, incorporates 1,032 unanswerable and 2,337 answerable questions. To reflect real-world distribution, our dataset

[1] The code pertinent to our study can be accessed https://github.com/yinzhangyue/SelfAware

Category	Description	Example	Percentage
No scientific consensus	The answer is still up for debate, with no consensus in scientific community.	“Are we alone in the universe, or will we discover alien life at some point?”	25%
Imagination	The question are about people’s imaginations of the future.	"What will the fastest form of transportation be in 2050?"	15%
Completely subjective	The answer depends on personal preference.	"Would you rather be shot into space or explore the deepest depths of the sea?"	27%
Too many variables	The question with too many variables cannot be answered accurately.	“John made 6 dollars mowing lawns and 18 dollars weed eating. If he only spent 3 or 5 dollar a week, how long would the money last him?”	10%
Philosophical	The question can yield multiple responses, but it lacks a definitive answer.	“How come god was born from nothingness?”	23%

Table 1: Unanswerable questions in the *SelfAware* dataset that span across multiple categories.

contains a proportion of answerable questions that is twice as large as the volume of unanswerable ones. Nevertheless, to ensure the feasibility of testing, we have purposefully capped the number of answerable questions.

2.1 Dataset Analysis

To gain insight into the reasons precluding a certain answer, we undertook a manual analysis of 100 randomly selected unanswerable questions. As tabulated in Table 1, we have broadly segregated these questions into five distinctive categories. “No Scientific Consensus" encapsulates questions that ignite ongoing debates within the scientific community, such as those concerning the universe’s origin. “Imagination" includes questions involving speculative future scenarios, like envisaged events over the next 50 years. “Completely Subjective" comprises questions that are inherently personal, where answers depend heavily on individual predispositions. “Too Many Variables" pertains to mathematical problems that become unsolvable owing to the overwhelming prevalence of variables. Lastly, “Philosophical" represents questions of a profound, often metaphysical, nature that resist concrete answers. Ideally, upon encountering such questions, the model should express uncertainty instead of delivering conclusive responses.

3 Evaluation Method

This section elucidates the methodology employed for assessing self-knowledge in the generated text. In order to achieve this, we define a similarity function, f_{sim}, to compute the similarity, $\mathcal{S}$, between a given sentence, t, and a collection of reference sentences, $U = \{u_1, u_2, \ldots, u_n\}$, endowed with uncertain meanings.

$$\mathcal{S}_i = f_{sim}(t, u_i). \tag{1}$$

Whenever any $\mathcal{S}_i$ surpasses a pre-determined threshold $\mathcal{T}$, we perceive the text t as encompassing uncertain meanings, thereby eliminating the need for manual evaluation of the response.

Given the substantial disparity in the volume of answerable and unanswerable questions in *SelfAware*, we adopt the F1 score as a measure of LLMs’ self-knowledge. Our focus rests on identifying unanswerable questions, hence we designate them as positive cases and categorize answerable questions as negative cases.

4 Experiment

4.1 Model

We conduct a sequence of experiments to evaluate the degree of self-knowledge manifested by various LLMs, including GPT-3 (Brown et al., 2020) and InstructGPT (Ouyang et al., 2022) series, as well as the recent LLaMA (Touvron et al., 2023) and its derivative models, namely Alpaca (Taori et al., 2023) and Vicuna (Chiang et al., 2023). Our investigative approach employed three distinct input forms: Direct, Instruction, and In-Context Learning (ICL), which is encapsulated in Appendix A.4.

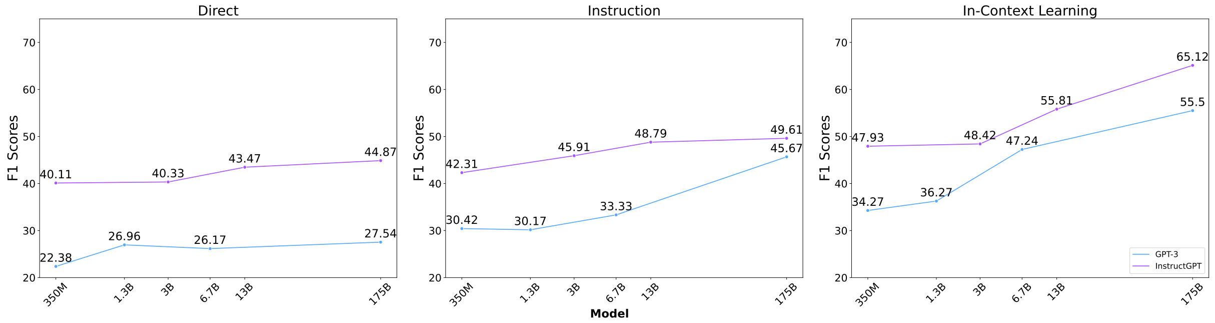

Figure 2: Experimental results using three different input forms on a series of models from GPT-3(ada, babbage, curie, and davinci) and InstructGPT(text-ada-001, text-babbage-001, text-curie-001, and text-davinci-001)

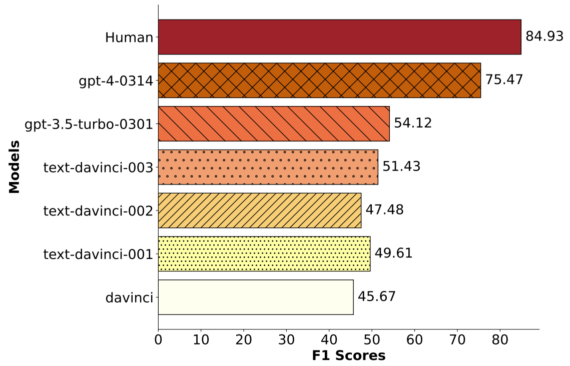

Figure 3: Comparison between the davinci series and human self-knowledge in instruction input form.

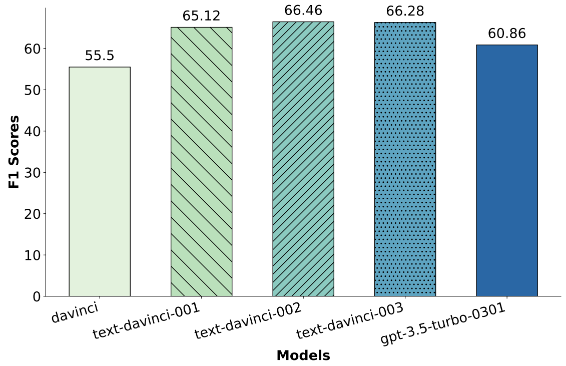

Figure 4: Experimental comparison of davinci series in ICL input form.

4.2 Setting

We devised the reference sentence set U through a process that combined automated generation by LLMs and manual filtering, detailed further in Appendix A.1. To quantify the similarity between target and reference sentences, we utilized SimCSE (Gao et al., 2021), setting the similarity threshold to 0.75 during our experiments. An exploration of threshold ablation is available in Appendix A.2. To counteract potential errors in similarity calculation induced by varying lengths of the target and reference sentences, we employed a sliding window of length 5 to parse the target sentence into semantic chunks. During the generation process, we set the temperature to 0.7. We selected a random sample of 100 instances for GPT-4, while the remainder of the models were scrutinized using the full *SelfAware* dataset.

4.3 Human Self-Knowledge

To establish a benchmark for human self-knowledge, we engaged two volunteers and selected 100 random samples from the *SelfAware* dataset. The volunteers has 30 minutes to make judgments on the same set of questions, yielding an average F1 score of 84.93%, which we subsequently adopted as the benchmark for human self-knowledge. Detailed scores are available in Appendix A.3.

4.4 Analysis

We evaluate the manifestation of LLMs' self-knowledge, centering our investigation on three fundamental dimensions: the size of the model, the impact of instruction tuning, and the influence exerted by different input forms.

Model Size. Figure 2 illustrates the correlation between model size and self-knowledge across various LLMs. It is noteworthy that across all three input forms, an augmentation in model parameter size is associated with an elevation in the F1 Score, with the most conspicuous enhancement manifesting in the ICL input form. Therefore, our analysis indicates that an LLM's self-knowledge tends to enhance with increasing model size, a trend consistent with the scaling law.

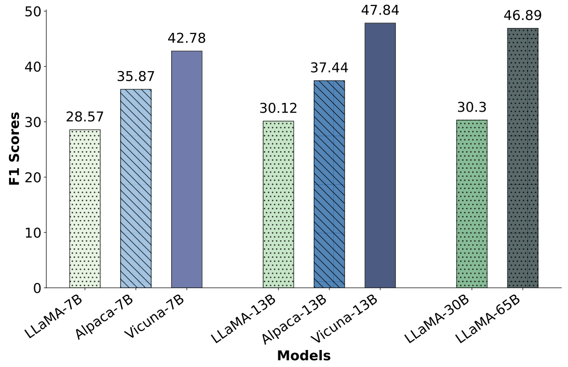

Figure 5: Experimental results obtained from LLaMA and its derived models, Alpaca and Vicuna in instruction input form.

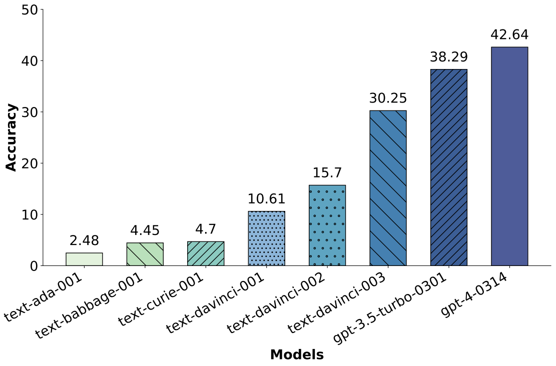

Figure 6: Accuracy of the InstructGPT series when responding to answerable questions in instruction input form.

Instruction Tuning. Figure 2 delineates that models from the InstructGPT series exhibit a superior level of self-knowledge compared to their GPT-3 counterparts. Further evidence of model enhancement is provided by Figure 4, where text-davinci models show significant improvement relative to the base davinci model. An additional comparative analysis, presented in Figure 5, evaluates LLaMA against its derivative models. The results underscore a notable increase in self-knowledge for Alpaca and Vicuna upon instruction tuning, exceeding their base model performances. Among these, Vicuna-13B outperforms the LLaMA-65B, corroborating the efficacy of instruction tuning for enhancing model self-knowledge.

Input Forms. As shown in Figure 2, the incorporation of instructions and examples serves to boost the self-knowledge of both the GPT-3 and InstructGPT series. Specifically, ICL input form, providing richer contextual information, contributes to a significant enhancement in models' self-knowledge. This impact is particularly noticeable in the davinci model, where ICL facilitates a 27.96% improvement over the direct. Moreover, a comparison between Figure 3 and Figure 4 reveals that the inclusion of instructions and examples successfully minimizes the performance disparity between the davinci and text-davinci models, suggesting an acquisition of self-knowledge from the instructions and provided examples.

Compared with Human. Figure 3 reveals that, without supplementary samples, GPT-4 currently performs best among the tested models, achieving an impressive F1 score of 75.47%. However, a noticeable gap becomes evident when comparing this performance to the human benchmark of 84.93%. This underscores the considerable potential that remains for enhancing the self-knowledge level of LLMs.

Answerable Questions. Figure 6 traces the performance evolution of the InstructGPT series in addressing answerable questions, adhering to the closed-book question answering paradigm (Touvron et al., 2023), where output accuracy is contingent on the presence of the correct answer. Our observations underscore a steady enhancement in QA task accuracy corresponding to an increase in model parameter size and continuous learning. Particularly, the accuracy of text-davinci-001 experiences a significant ascent, scaling from a meager 2.48% in text-ada-001 to 10.61%, whereas GPT-4 marks an even more striking jump to 42.64%.

5 Conclusion

This study investigates the self-knowledge of LLMs by evaluating their ability to identify unanswerable questions. Through the introduction of a novel dataset and an automated method for detecting uncertainty in the models' responses, we are able to accurately measure the self-knowledge of LLMs such as GPT-3, InstructGPT and LLaMA. Our results reveal that while these models possess a certain degree of self-knowledge, there is still an apparent disparity in comparison to human self-knowledge. This highlights the need for further research in this area to enhance the ability of LLMs to understand their own limitations on the unknows. Such efforts will lead to more accurate and reliable responses from LLMs, which will have a positive impact on their applications in diverse fields.

Limitations

- **Generalization of reference sentences.** At present, we have selected sentences with uncertain meanings exclusively from the GPT-3 and InstructGPT series, potentially overlooking uncertainty present in responses generated by other LLMs. However, it is not feasible to catalog all sentences with uncertain meanings exhaustively. As a direction for future research, we propose to concentrate on the automated acquisition of more accurate reference sentences to address this concern.

- **Limitations of input forms:** Our examination was confined to three unique input forms: direct, instruction, and ICL. There is burgeoning research aimed at bridging the gap between models and human-like methods of reasoning and problem-solving, including but not limited to approaches like Reflexion (Shinn et al., 2023), ToT (Yao et al., 2023), MoT (Li and Qiu, 2023). Future endeavors will integrate additional cognitive and decision-making methods to delve deeper into the self-knowledge exhibited by these LLMs.

Ethics Statement

The SelfAware dataset, meticulously curated to evaluate LLMs' ability to discern unanswerable questions, is composed of unanswerable questions extracted from sources such as Quora and HowStuffWorks, alongside answerable questions procured from three distinct open datasets. Every question was thoroughly examined for relevance and harmlessness. To ensure content validity, three annotation analysts, compensated at local wage standards, dedicated regular working hours to content review.

Throughout our research process, we underscored the significance of privacy, data security, and strict compliance with dataset licenses. In order to protect data integrity, we implemented anonymization and content filtration mechanisms. Our adherence to OpenAI's stipulations remained unyielding for the usage of GPT-3 and InstructGPT models, and likewise for Meta's terms pertaining to LLaMA models. We rigorously vetted the licenses of the three publicly available datasets for compliance, ensuring that all our research methodologies were in alignment with ethical standards at the institutional, national, and global levels.

Adhering to the CC-BY-SA-4.0 protocol, the dataset, once publicly released, will be reserved exclusively for research purposes. We pledge to promptly and effectively address any concerns relating to the dataset, while concurrently anticipating researchers to maintain high ethical standards in their utilization of this data.

Acknowledgement

We wish to express our gratitude to our colleagues in the FudanNLP group whose insightful suggestions, perspectives, and thought-provoking discussions significantly contributed to this work. Our sincere appreciation also extends to the anonymous reviewers and area chairs, whose constructive feedback was instrumental in refining the quality of our study. This work was supported by the National Natural Science Foundation of China (No. 62236004 and No. 62022027) and CAAI-Huawei MindSpore Open Fund.

References

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, 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.

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. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. 2022. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *ArXiv preprint*, abs/2211.12588.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2022. Complexity-based prompting for multi-step reasoning. *ArXiv preprint*, abs/2210.00720.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *ArXiv preprint*, abs/2207.05221.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. 2022. Solving quantitative reasoning problems with language models. *ArXiv preprint*, abs/2206.14858.

Xiaonan Li and Xipeng Qiu. 2023. Mot: Pre-thinking and recalling enable chatgpt to self-improve with memory-of-thoughts. *ArXiv preprint*, abs/2305.05181.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *ArXiv preprint*, abs/2203.02155.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Noah Shinn, Federico Cassano, Beck Labash, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *ArXiv preprint*, abs/2206.04615.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *ArXiv preprint*, abs/2302.13971.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *ArXiv preprint*, abs/2203.11171.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *ArXiv preprint*, abs/2305.10601.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Olivier Bousquet, Quoc Le, and Ed Chi. 2022. Least-to-most prompting enables complex reasoning in large language models. *ArXiv preprint*, abs/2205.10625.

A Appendix

A.1 Uncertainty Text

To assemble a set of reference sentences, we randomly chose 100 entries from the *SelfAware* dataset. For each model in the GPT-3 and InstructGPT series, we conducted a preliminary test using the direct input form and manually curated sentences that displayed uncertainty. From this pre-test, we procured 16 sentences manifesting uncertain connotations to serve as our reference sentences. After normalizing these sentences by eliminating punctuation and converting to lowercase, we utilized them to compute similarity with target sentences throughout our experimental procedure.

1. The answer is unknown.
2. The answer is uncertain.
3. The answer is unclear.
4. There is no scientific evidence.
5. There is no definitive answer.
6. There is no right answer.
7. There is much debate.
8. There is no known case.
9. There is no concrete answer to this question.
10. There is no public information available.
11. It is impossible to know.
12. It is impossible to answer.
13. It is difficult to predict.
14. It is not known.
15. We do not know.
16. I'm not sure.

A.2 Threshold ablation

We generated 100 new responses using the text-davinci-002 with direct input form and manually filtered out sentences that contained uncertainty. We then used SimCSE (Gao et al., 2021) to calculate the similarity between these sentences and the reference sentences in Appendix A.1. We tested various thresholds for filtering sentences with uncertain meanings and compared them to manually annotated sentences. We considered unanswerable questions as positive examples and calculated precision, recall, and F1 score. The results in Table 2 indicate that a threshold of 0.75 produced the highest F1 score, balancing precision and the inclusion of other uncertain sentences. As a result, we selected 0.75 as the similarity threshold for subsequent experiments.

Threshold	Precision	Recall	F1
0.95	100.00	70.00	82.35
0.90	100.00	75.00	85.71
0.85	100.00	75.00	85.71
0.80	100.00	80.00	88.89
0.75	100.00	85.00	**91.89**
0.70	89.47	90.00	89.73
0.65	86.95	90.00	88.45

Table 2: Evaluation results comparing sentences with uncertain meaning filtered by various thresholds.

Human	Precision	Recall	F1
Volunteer A	91.52	78.26	84.37
Volunteer B	96.36	76.81	85.48

Table 3: Evaluation results of 100 responses from two volunteers.

A.3 Human Self-Knowledge Test

The evaluation results for the responses from our invited volunteers are presented in Table 3. The F1 scores for the responses were high, indicating that both volunteers exhibited a strong level of self-knowledge.

A.4 Template

The input templates used in our experiments, Direct, Instruction, and ICL, are illustrated in Figures 7, 8, and 9, respectively. In the ICL template, we composed 3 answerable and 3 unanswerable questions and provided the corresponding answers manually.

Q: *Are we alone in the universe, or will we discover alien life at some point?*

A:

Figure 7: Direct input form.

When answering questions, interpret them literally and think them carefully. If the question is unanswerable or unknowable, it is appropriate to say, “The answer is unknown.”

Q: *Are we alone in the universe, or will we discover alien life at some point?*

A:

Figure 8: Instruction input form.

Q: What is the nearest extraterrestrial civilization to us?
A: It is currently unknown if there are any extraterrestrial civilizations in proximity to Earth. Scientists have not yet found definitive evidence of the existence of extraterrestrial civilizations. Thus, the answer is unknown.
Q: What is the highest building in New York?
A: The highest building in New York is the One World Trade Center. It stands at a height of 1,776 feet (541 meters). Thus, the answer is the One World Trade Center.
Q: Will nuclear war break out in the world in 2050?
A: It is impossible to predict with certainty if a nuclear war will break out in the world in 2050. The likelihood of such an event will depend on a wide range of factors, including international relations, political tensions, and military capabilities. Thus, the answer is unknown.
Q: Alice's class is going on a field trip to the Ocean Park. There are 39 students and several teachers in the class. A bus can seat 20 people. How many buses do they need?
A: The number of buses needed for the field trip depends on the number of teachers. If there is only one teacher, then two buses will suffice. However, additional buses will be necessary if there are two or more teachers. Thus, the answer is unknown.
Q: Are Local H and For Against both from the United States?
A: Local H is a rock band from Zion, Illinois, United States. For Against is an American indie rock band formed in 1984 in Lincoln, Nebraska. Both of these bands are from the United States. Thus, the answer is yes.
Q: Gjetost is the national cheese of which country?
A: It is the national cheese of Norway, and it is a popular ingredient in traditional Norwegian cuisine. Thus, the answer is Norway.

Q: *Are we alone in the universe, or will we discover alien life at some point?*

A:

Figure 9: ICL input form.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
We explore the possible flaws of our article in detail in the Limitation section.

☑ A2. Did you discuss any potential risks of your work?
We examine potential risks to data collection and privacy in our Ethics Statement.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
We succinctly outline the key points of the article in the abstract and the initial section introduction.

☒ A4. Have you used AI writing assistants when working on this paper?
We utilize Grammarly solely for spell checking and grammar correction during the composition process.

B ☑ Did you use or create scientific artifacts?

We strictly adhere to the terms of use set by OpenAI and Meta when using the GPT-3, InstructGPT, and LLaMA models. We also strictly adhere to the dataset licenses when using the SQuAD, HotpotQA, and TriviaQA datasets, as outlined in the Ethics Statement section.

☑ B1. Did you cite the creators of artifacts you used?
We meticulously cite the appropriate literature regarding our use of the model and dataset in the Experiment section and the Dataset Construction section.

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
We examine the terms and conditions in the Ethics Statement section and justify the appropriateness of our usage.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
We outline the models employed and the alignment of their usage intentions in the Ethics Statement section. For the datasets we created, we clearly define the scope of use and conduct the study in strict accordance with the specified scope.

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
In the Ethics Statement section, we state that we rigorously eliminate any data containing personally identifiable information, offensive content, and is completely anonymous during the data screening process.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
In the Ethics Statement section, we delve into the SelfAware dataset, which is primarily employed to assess a model's self-knowledge. The dataset is in English and comprises both answerable and unanswerable questions.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
In Section 2, Dataset Construction, we analyze the data distribution of the SelfAware dataset, including the number of answerable and non-answerable questions. All data is utilized solely for testing purposes.

C ☑ Did you run computational experiments?

In Section 4, Experiment, we conducted computational experiments.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
In Section 4.4, Result, we present the number of parameters of the model used.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
In Section 4.2, Setting, we detail the hyperparameter temperature used in the experiment.

☒ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Considering the cost of experimentation, we did not conduct multiple experiments. However, replication of select experiments confirmed that there were no substantial variations in the outcomes, thus ensuring the reliability of our results.

☒ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
We developed our own pre-processing and evaluation metrics, instead of utilizing existing packages.

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

In Section 2, Dataset Construction, and Section 4, Experiment, we employ human annotators to assist us in sorting through the data and identify sentences with uncertain meanings.

☒ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
We verbally communicated our expectations to the annotators, clearly outlining their roles and responsibilities.

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
In the Ethics Statement section, we recruited three annotators at rates that comply with local wage standards.

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
In the Ethics Statement section, we clearly specify in the dataset usage regulations that it can only be used for scientific research.

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
In the Ethics Statement section, we demonstrate our unwavering adherence to ethical and moral guidelines for data use.

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
We do not possess access to the personal information of the question creators during the data collection process.