Mitigating Overconfidence in Large Language Models: A Behavioral Lens on Confidence Estimation and Calibration

Bingbing Wen[*1] **Chenjun Xu**[*1] **Bin Han**[1] **Robert Wolfe**[1] **Lucy Lu Wang**[12] **Bill Howe**[1]
[1]University of Washington [2] Allen Institute for AI
{bingbw,chenjux,bh193,rwolfe3,lucylw,billhowe}@uw.edu

Abstract

Confidence estimation is a crucial area in machine learning, particularly with large language models (LLMs), which are prone to overconfidence, leading to inaccurate predictions, hallucinations, and impaired decision-making. As LLMs are increasingly integrated into real-world applications, overconfidence poses challenges for effective human-machine collaboration. We examine LLM overconfidence through the lens of human behavior, proposing a mechanism for understanding of how models exhibit overconfidence and how to mitigate its effects to improve LLM interpretability and calibration. Drawing on models of human overconfidence in cognitive and psychological research, we consider whether LLMs mirror human overconfidence patterns related to perceived task difficulty and comparisons with others. Our findings indicate that LLMs exhibit varied confidence patterns. Larger models, similar to humans, tend to overestimate their performance on challenging tasks and underestimate it on simpler ones, while small models display consistent overconfidence across all task levels. However, LLMs' self-assessments are generally less sensitive to task difficulty than human estimates. We propose Answer-Free Confidence Estimation (AFCE), a method that reduces overconfidence by asking models for confidence scores on question sets without providing answers. This approach decouples confidence estimation from answer generation, significantly lowering overconfidence, particularly on challenging tasks. We then consider how LLMs' self-assessment compares to their assessment of experts and laymen, providing insight into how LLMs place their own abilities, even though the actual accuracies between the two groups remains comparable. We aim to motivate psychology-grounded research for better confidence calibration in LLMs.

1 Introduction

Reliable confidence (uncertainty) estimates are essential for effective human-machine collaboration [7]. Large language models (LLMs), however, are prone to overconfidence [24], which can result in inaccurate predictions when they should abstain [23]. As these models are increasingly deployed in real-world tasks such as medical diagnosis [19], legal analysis [5], and decision support systems [25], their performance directly impacts outcomes that affect human lives. Overconfidence in LLMs can lead to significant errors [26], reduced trust [10], and potentially harmful downstream consequences [13]. Therefore, understanding whether LLMs exhibit overconfidence in ways that parallel or exceed human behavior is critical to improving their reliability and safety in real-world applications.

Human overconfidence is recognized as a significant cognitive bias [11]. Moore and Healy [17] reconcile experimental findings that 1) individuals tend to overestimate their abilities on difficult tasks

[*]equal contribution

NeurIPS 2024 Workshop on Behavioral Machine Learning.

and underestimate them on easy tasks, and 2) they misjudge others' abilities, often underplacing themselves on challenging tasks and overplacing themselves on simpler ones. The authors explain these phenomena using an information theoretic model demonstrating individuals' regressive estimates of their performance and even more regressive estimates of others' performance. When performance is exceptionally high (e.g., on easy tasks), individuals underestimate their own performance and underestimate others' performance even more so; and when performance is exceptionally low (e.g., on hard tasks), they overestimate their own performance and overestimate others' performance even more so. In this paper, we designed experiments to consider whether these results hold for LLMs.

To our knowledge, this is the first study to explore overconfidence in LLMs from a cognitive and psychological perspective. We address two key research questions: (RQ1) Is model confidence sensitive to task difficulty, and does it exhibit the same over-confidence and under-confidence patterns as previously observed in human subjects? (RQ2) Do models exhibit overplacement (underplacement) behavior when estimating their performance relative to humans with varying levels of expertise?

Our three contributions directly address the research questions posed above:

1) We evaluate LLMs' confidence estimations across tasks of varying difficulty and find that different models display distinct confidence patterns. Large models mirror trends seen in human subjects, which tend to be underconfident on easier tasks and overconfident on more challenging ones. While smaller models consistently exhibit overconfidence across all levels of task difficulty. Additionally, LLMs' confidence estimates are generally less influenced by task difficulty compared to human confidence estimates.
2) We propose a confidence calibration measure called Answer-Free Confidence Estimation (AFCE), which reduces overconfidence and achieves promising results in confidence calibration, particularly outperforms baseline verbalized confidence elicitation techniques on challenging tasks.
3) We investigate LLMs' ability to estimate the confidence of experts and laymen in accomplishing the tasks. We find that LLMs consistently estimate higher performance among experts and lower performance among laymen, despite the actual accuracy remaining comparable, suggesting a superficiality to the estimates.

2 Related Work

We review the related work on human overconfidence and confidence elicitation methods for LLMs.

Human Overconfidence Overconfidence refers to an unjustified belief in one's knowledge and abilities [11], concerning for its prediction of undesirable outcomes in consequential domains such as medicine [3], politics [21], and science [14]. Models to explain overconfidence have been broadly considered (e.g., Dunning-Kruger [11], or recent contrasting results from Sanchez and Dunning [20] that found that those with intermediate knowledge tend to exhibit the most overconfidence). In this paper, we focus on the experiments of Moore and Healy [17], whose influential unifying model explained a variety of previous findings.

Confidence Elicitation in Language Models Previous methods for eliciting confidence have primarily relied on white-box approaches, which have estimated confidence using token likelihoods [22] and internal state-based methods [9, 12]. While effective, these techniques require internal access to the model, making them less applicable to models served over closed APIs, like GPT-4 [1]. Verbalized confidence approaches appropriate to such models (*i.e.,* prompting the model to write out its confidence in text) tend to produce uniformly high estimations of model confidence, usually between 80% and 100% [16, 24]. To address this, some studies have introduced consistency-based methods [15, 24] that calibrate LLM confidence and mitigate overconfidence. In this study, we adopt these widely-used, prompt-based confidence elicitation methods as baselines, and we develop a novel prompt-based strategy that consistently outperforms baseline methods on hard tasks.

3 Data & Models

Datasets To approximate the research design of Moore and Healy [17], who considered six subject domains and questions across a range of difficulty levels, we use (1) **MMLU** [8], a collection of domain-specific multiple-choice questions across 57 subjects and multiple difficulties corresponding to education level—we use questions from the High School, College, and expert difficulty levels in the subject domains of Physics, Chemistry, Biology, Math, Computer Science, and Medicine; and (2) **GPQA** [18], a dataset of multiple-choice questions crafted by experts (*i.e.*, individuals holding or pursuing a Ph.D.) in the subject of Physics, Chemistry, and Biology. We treat a subject at a difficulty level as a subtask. Following the methodology of Moore and Healy [17], we randomly group 10 questions into a single prompt for each subtask. More details about datasets are in the Appendix.

Models We present results from three models, ranging from small to large, and spanning different model families: google/gemma-2-9b-it (Gemma2-9B), meta-llama/Meta-Llama-3-70B-Instruct (Llama3-70B), claude-3-sonnet-20240229 (Claude-3-sonnet). We set temperature to 0 and top-p sampling to 1 in the interest of reproducibility to reduce the variability of model output.

4 Self-Estimation & Task Hardness

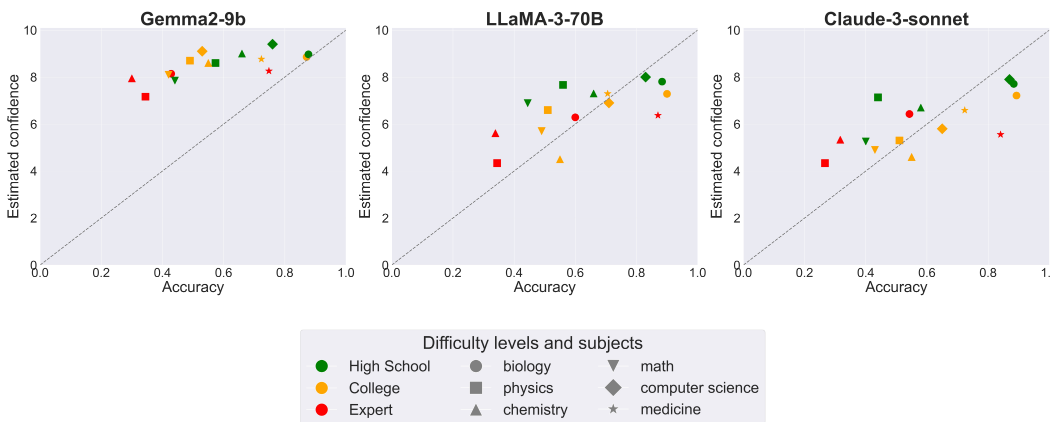

Figure 1: Comparison of confidence estimation patterns across models using the AFCE method on tasks with varying difficulty levels. Each dot represents the performance of a subject at a specific difficulty level (e.g., college biology).

In this section, we study the relationship between LLMs' confidence estimation and task difficulty. We also compare our novel Answer-Free Confidence Estimation (AFCE) method with other widely-used confidence estimation methods for calibrating confidence against actual model performance.

4.1 Experiment Setup

Baselines. We compare our method against several widely-used prompt-based confidence elicitation baselines. These include Vanilla Verbalized Confidence [24], which prompts the model with "Read the question, provide your answer, and report your confidence in this answer"; Top-k Prompting Verbalized Confidence [24], which prompts the model to provide "your K best guesses and the probability that each is correct (0% to 100%) for the following question"; and Quiz-Like Prompting, which prompts the model to "Answer the following 10 questions and estimate how many were answered correctly" [17]. We employ Expected Calibration Error (ECE) [6] as a metric to evaluate confidence calibration, which quantifies the difference between a model's predicted confidence and its actual accuracy.

Our Method: Answer-Free Confidence Estimation

We propose *Answer-Free Confidence Estimation*, which employs two discrete processes to evaluate task performance and elicit confidence estimation. To evaluate performance, we prompt the model with "Please answer the following 10 questions by selecting only the option letter," and we use the model's responses to compute its accuracy. We separately obtain the model's confidence by prompting the model to "Read the questions and estimate how many you can answer correctly (choose a number from 0-10)." This method more closely adheres to the psychological instruments provided to human subjects in confidence elicitation experiments [17] than strategies that combine confidence estimation and task performance into a single step such as vanilla prompting.

Method	High School			College			Expert		
	Acc	AvC	ECE	Acc	AvC	ECE	Acc	AvC	ECE
Gemma2-9B									
Vanilla	<u>56.0</u>	98.3	43.1	<u>48.0</u>	99.6	51.6	<u>34.4</u>	99.0	64.5
Top-K	49.3	91.4	42.8	47.0	93.3	48.7	30.6	91.9	62.1
Quiz-like	**58.7**	96.0	<u>37.3</u>	<u>48.0</u>	93.0	<u>45.0</u>	**36.7**	92.8	<u>56.1</u>
AFCE	57.3	86.0	**28.7**	**49.0**	87.0	**38.0**	<u>34.4</u>	71.7	**37.2**
LLaMA-3-70B									
Vanilla	**60.0**	88.4	28.4	<u>53.0</u>	85.8	32.8	35.0	82.0	47.0
Top-K	<u>59.3</u>	71.8	**12.7**	**56.0**	68.9	<u>16.8</u>	**36.1**	62.6	<u>26.4</u>
Quiz-like	56.7	80.0	23.3	<u>53.0</u>	80.0	27.0	35.6	81.1	45.6
AFCE	56.0	76.7	<u>20.7</u>	51.0	66.0	**15.0**	<u>34.4</u>	43.3	**16.7**
Claude-3-sonnet									
Vanilla	**48.7**	89.6	40.9	**52.0**	88.9	37.0	<u>28.9</u>	83.9	55.6
Top-K	42.0	67.1	**25.2**	46.0	66.9	<u>20.9</u>	**31.7**	51.2	<u>19.5</u>
Quiz-like	42.7	98.7	56.0	50.0	95.0	45.0	24.4	89.4	65.0
AFCE	<u>44.0</u>	71.3	<u>27.3</u>	<u>51.0</u>	53.0	**2.0**	26.7	43.3	**16.7**

Table 1: Confidence calibration performances of AFCE with baselines methods across models in the Physics at varying difficulty levels. Results for the Chemistry and Biology are in the Appendix.

4.2 Results

LLMs exhibit varied confidence estimation patterns but all models are less responsive to changes in task difficulty. Figure 1 illustrates the relationship between LLM confidence estimation and task difficulty. LlaMA-3-70B and Claude-3-Sonnet exhibit underconfidence on easier tasks ($accuracy > 0.8$) and overconfidence on harder tasks ($accuracy < 0.4$), in accordance with established findings in human subjects [17]. For tasks of medium difficulty ($accuracy \sim 0.5$), models confidence aligns more closely with performance. In contrast, Gemma2-9b consistently demonstrates overconfidence across all tasks, with this overconfidence increasing as task difficulty rises. Our findings suggest that larger models align more closely with human behavior. However, LLM confidence is more uniform than human answers and less sensitive to task difficulty, as LLMs exhibit a tendency to report a "standard" confidence answer, possibly limiting the utility of verbalized confidence elicitation strategies. Moreover, while LLMs exhibit lower accuracy on expert-level tasks, they sometimes exhibit stronger performance on college-level subjects than on high school-level subjects, breaking with typical human judgments of task difficulty. We speculate that the estimation of performance on college-level tasks may be more influenced by parameterized knowledge learned from sources like college-level textbooks.

The Answer-Free Confidence Estimation method outperforms baseline verbalized confidence elicitation methods on hard tasks across models. As shown in Table 1, AFCE mitigates overconfidence significantly, especially for difficult tasks, such that for Claude-3-sonnet ECE is reduced to 2.0 for College Physics and 16.7 for Expert Physics subject domains, outperforming other baseline methods. Quiz-Like prompting achieves the most comparable performance, likely due to its similar construction to AFCE, and it outperforms AFCE for High-School Physics. Both AFCE and Quiz-Like prompting may be sensitive to the size of the question set. Though it is not the intention of the method, we note that AFCE does not improve accuracy over baseline methods. We speculate that AFCE reduces overconfidence by limiting engagement with the subject domain, constraining the model's reasoning to an assessment of its confidence rather than simultaneously handling the epistemically intensive process of generating factual information. Indeed, it is possible that engaging both processes simultaneously could be part of the cause of overconfidence. We intend to explore underlying mechanisms like these and their relationship to human cognition in future work that expands the utility of AFCE.

5 Overplacement: Estimating Others' Performance

In this section, we investigate whether LLMs exhibit overplacement when estimating their own performance relative to that of others. In previous work with human subjects [17], participants estimated the performance of a randomly chosen peer. We adapt this experiment for LLMs by prompting language models to adopt the personas of other individuals and estimate their confidence.

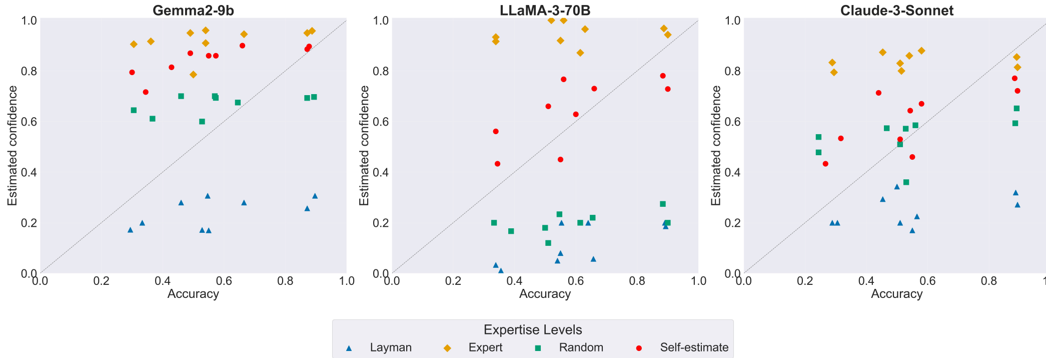

Figure 2: Confidence estimates across models prompted to adopt random person, expert, and layman personas. Each dot represents the performance of a subject at a specific difficulty level (e.g., college biology). Since RQ2 focuses on overplacement(underplacement), we do not differentiate between individual subjects or difficulty levels.

5.1 Experiment Setup

Drawing closely on prior work [17] to inform experimental design, we prompt the model to adopt the persona of another person and 1) answer the questions and 2) estimate its confidence. Aside from instructing the model to adopt the persona, we utilize AFCE as described in the previous section. We specifically instruct the model to adopt the persona of a "random" person, an "expert" in the subject under consideration, and a "layman" with regard to the subject. In prompting language models to adopt personas, we build on much recent work on using LLMs for simulation in computational social science [2], as well as assessments of model bias and fairness [4]. Full prompts are in the appendix.

5.2 Results

Estimated Confidence towards Random Chosen Person, Expert and Layman. Figure 2 illustrates that LLMs exhibit little variance in confidence estimation for a given persona, consistently providing high confidence estimates when prompted to adopt the persona of an expert and low confidence estimates when prompted to adopt the persona of a layman or a random individual. LLM self-estimations fall between these two extremes, such that the model's default confidence estimate exceeds that when prompted as a layman, but falls short of that when prompted as an expert. Despite these differences, though, accuracy remains roughly uniform across different personas, such that models overestimate confidence when prompted as an expert, and underestimate it when prompted as a layman. However, we hesitate to make a definitive claim about overplacement, as the results also suggest that the model's confidence estimation is disconnected from its actual capabilities when adopting a persona. We expect that this mismatch could prove problematic for the range of research that now utilizes persona-prompted LLMs in social scientific simulations [2, 27].

6 Discussion & Conclusion

In this study, we investigated overconfidence in LLMs by drawing on previous work in experimental psychology [17]. Our findings reveal that LLMs exhibit different confidence patterns. Some, like LLaMA-3-70B and Claude-3-Sonnet, display overconfidence patterns similar to those of humans, while others, like Gemma2-9b, show consistent overconfidence across all tasks. But, all of models are less sensitive to task difficulty. We introduced the "Answer-Free Confidence Estimation" method, which improves LLM calibration by disentangling task performance from confidence estimation. Additionally, our analysis of a persona-prompted LLM demonstrates that while a model prompted as an expert produces a higher confidence estimate than the model prompted as a layman, actual task performance remains similar across both groups.

While these findings offer important insights, they also reveal several limitations that should guide future work. First, our analysis was limited to multiple-choice datasets. It remains to be seen whether the observed patterns hold across a broader range of models and tasks. Our study focused on prompting confidence elicitation without incorporating more advanced techniques, such as sampling or aggregation strategies, which could further refine model calibration.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, pages 337–371. PMLR, 2023.

[3] Eta S. Berner and Mark L Graber. Overconfidence as a cause of diagnostic error in medicine. *The American journal of medicine*, 121 5 Suppl:S2–23, 2008. URL `https://api.semanticscholar.org/CorpusID:1659486`.

[4] Myra Cheng, Esin Durmus, and Dan Jurafsky. Marked personas: Using natural language prompts to measure stereotypes in language models. *ArXiv*, abs/2305.18189, 2023. URL `https://api.semanticscholar.org/CorpusID:258960243`.

[5] Aniket Deroy, Kripabandhu Ghosh, and Saptarshi Ghosh. Applicability of large language models and generative models for legal case judgement summarization. *ArXiv*, abs/2407.12848, 2024. URL `https://api.semanticscholar.org/CorpusID:271222741`.

[6] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 2017. URL `http://proceedings.mlr.press/v70/guo17a.html`.

[7] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

[8] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=d7KBjmI3GmQ`.

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

[10] Sunnie S. Y. Kim, Q. Vera Liao, Mihaela Vorvoreanu, Steph Ballard, and Jennifer Wortman Vaughan. "i'm not sure, but...": Examining the impact of large language models' uncertainty expression on user reliance and trust. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2024. URL `https://api.semanticscholar.org/CorpusID:269484145`.

[11] Justin Kruger and David Dunning. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology*, 77(6):1121, 1999.

[12] Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *ArXiv preprint*, abs/2302.09664, 2023. URL `https://arxiv.org/abs/2302.09664`.

[13] Zihao (Michael) Li. The dark side of chatgpt: Legal and ethical challenges from stochastic parrots and hallucination. *ArXiv*, abs/2304.14347, 2023. URL `https://api.semanticscholar.org/CorpusID:258352281`.

[14] Nicholas Light, Philip M Fernbach, Nathaniel Rabb, Mugur V Geana, and Steven A Sloman. Knowledge overconfidence is associated with anti-consensus views on controversial scientific issues. *Science Advances*, 8(29):eabo0038, 2022.

[15] Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *ArXiv preprint*, abs/2305.19187, 2023. URL `https://arxiv.org/abs/2305.19187`.

[16] Sabrina J. Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872, 2022. doi: 10.1162/tacl_a_00494. URL `https://aclanthology.org/2022.tacl-1.50`.

[17] Don A Moore and Paul J Healy. The trouble with overconfidence. *Psychological review*, 115 (2):502, 2008.

[18] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

[19] Alejandro Ríos-Hoyo, Naing Lin Shan, Anran Li, Alexander T. Pearson, Lajos Pusztai, and Frederick M. Howard. Evaluation of large language models as a diagnostic aid for complex medical cases. *Frontiers in Medicine*, 11, 2024. URL `https://api.semanticscholar.org/CorpusID:270656792`.

[20] Carmen Sanchez and David Dunning. Intermediate science knowledge predicts overconfidence. *Trends in Cognitive Sciences*, 28(4):284–285, 2024.

[21] Jan-Willem van Prooijen. Overconfidence in radical politics. *The Psychology of Populism*, 2021. URL `https://api.semanticscholar.org/CorpusID:233884520`.

[22] Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. " my answer is c": First-token probabilities do not match text answers in instruction-tuned language models. *arXiv preprint arXiv:2402.14499*, 2024.

[23] Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. The art of refusal: A survey of abstention in large language models. *arXiv preprint arXiv:2407.18418*, 2024.

[24] Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=gjeQKFxFpZ`.

[25] Zeqiu Xu, Lingfeng Guo, Shuwen Zhou, Runze Song, and Kaiyi Niu. Enterprise supply chain risk management and decision support driven by large language models. *Applied Science and Engineering Journal for Advanced Research*, 3(4):1–7, Jul. 2024. doi: 10.5281/zenodo.12670581. URL `https://asejar.singhpublication.com/index.php/ojs/article/view/103`.

[26] Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. Navigating the grey area: Expressions of overconfidence and uncertainty in language models. *ArXiv*, abs/2302.13439, 2023. URL `https://api.semanticscholar.org/CorpusID:257220189`.

[27] Caleb Ziems, William B. Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. Can large language models transform computational social science? *Computational Linguistics*, 50: 237–291, 2023. URL `https://api.semanticscholar.org/CorpusID:258547324`.

A Appendix / supplemental material

Dataset	**Hardness**	**Subject**	**Test Size**
MMLU	High School	Physics	173
MMLU	High School	Chemistry	230
MMLU	High School	Biology	347
MMLU	High School	Math	304
MMLU	High School	Computer Science	114
MMLU	College	Physics	118
MMLU	College	Chemistry	113
MMLU	College	Biology	165
MMLU	College	Math	116
MMLU	College	Computer Science	116
MMLU	College	Medicine	200
MMLU	Expert	Medicine	308
GPQA	Expert	Physics	227
GPQA	Expert	Chemistry	214
GPQA	Expert	Biology	105

Table 2: Dataset statistics.

// *High school*
Question: The plates of a capacitor are charged to a potential difference of 5 V. If the capacitance is 2 mF, what is the charge on the positive plate?
A. 0.005 C B. 0.01 C C. 0.02 C D. 0.5 C
Answer: B

// *College*
Question: The quantum efficiency of a photon detector is 0.1. If 100 photons are sent into the detector, one after the other, the detector will detect photons?
A. an average of 10 times, with an rms deviation of about 4
B. an average of 10 times, with an rms deviation of about 3
C. an average of 10 times, with an rms deviation of about 1
D. an average of 10 times, with an rms deviation of about 0.1
Answer: B

// *Expert*
Astronomers are studying a star with a Teff of approximately 6000 K. They are interested in spectroscopically determining the surface gravity of the star using spectral lines (EW < 100 mA) of two chemical elements, El1 and El2. Given the atmospheric temperature of the star, El1 is mostly in the neutral phase, while El2 is mostly ionized. Which lines are the most sensitive to surface gravity for the astronomers to consider?
A. El2 I (neutral) B. El1 II (singly ionized) C. El2 II (singly ionized) D. El1 I (neutral)
Answer: C

Table 3: Examples from dataset in three task hardness level.

// Prompt for accuracy calculation
prompt = “‘Please answer the following questions by selecting only the option letter. Do not provide any explanation.
Questions:
Q1: <>
Q2: <>
Q3: <>
Q4: <>
Q5: <>
Q6: <>
Q7: <>
Q8: <>
Q9: <>
Q10: <>
Now, please answer these questions with the following format:
1. <Only Option Letter >
2. <Only Option Letter >
3. <Only Option Letter >
4. <Only Option Letter >
5. <Only Option Letter >
6. <Only Option Letter >
7. <Only Option Letter >
8. <Only Option Letter >
9. <Only Option Letter >
10. <Only Option Letter >
’’’

// Prompt for confidence calculation
prompt = f"""Read the question, and estimate how many questions you can answer correctly (choose a number from 0-10). Do not provide any explanation.
Questions:
Q1: <>
Q2: <>
Q3: <>
Q4: <>
Q5: <>
Q6: <>
Q7: <>
Q8: <>
Q9: <>
Q10: <>
Estimate how many questions you can answer correctly: <choose a number from 0-10>."""

Table 4: Prompts for self-estimation experiments.

// Prompt for accuracy calculation
prompt = “‘You are a {randomly chosen person}/{expert}/{laymen}. Please answer the following questions by selecting only the option letter. Do not provide any explanation.
Questions:
Q1: <>
Q2: <>
Q3: <>
Q4: <>
Q5: <>
Q6: <>
Q7: <>
Q8: <>
Q9: <>
Q10: <>
Now, please answer these questions with the following format:
1. <Only Option Letter >
2. <Only Option Letter >
3. <Only Option Letter >
4. <Only Option Letter >
5. <Only Option Letter >
6. <Only Option Letter >
7. <Only Option Letter >
8. <Only Option Letter >
9. <Only Option Letter >
10. <Only Option Letter >
’’’

// Prompt for confidence calculation
prompt = f"""A is an self.expertise in self.subject. Read the question, and after considering A’s ability, estimate how many questions A can answer correctly (choose a number from 0-10). Do not provide any explanation.
Questions:
Q1: <>
Q2: <>
Q3: <>
Q4: <>
Q5: <>
Q6: <>
Q7: <>
Q8: <>
Q9: <>
Q10: <>
Estimate how many questions you think A can answer correctly: <choose a number from 0-10>."""

Table 5: Prompts for overplacement experiments.

Method	High School			College			Expert		
	Acc	AvC	ECE	Acc	AvC	ECE	Acc	AvC	ECE
Gemma2-9B									
Vanilla	64.0	98.8	35.1	49.0	99.0	50.0	**30.0**	98.0	68.3
Top-K	56.5	91.3	35.9	49.0	94.0	45.0	28.3	89.3	62.7
Quiz-like	**67.5**	95.0	27.5	**55.0**	96.0	41.0	29.4	96.7	67.2
Ours	66.0	90.0	**24.0**	55.0	86.0	**31.0**	**30.0**	79.4	**49.4**
LLaMA-3-70B									
Vanilla	**67.0**	87.7	20.7	**56.0**	85.9	29.9	30.6	81.2	50.7
Top-K	63.0	73.7	12.2	53.0	71.6	19.1	31.7	61.9	32.0
Quiz-like	65.0	80.0	15.0	53.0	80.0	27.0	**35.6**	82.8	47.2
Ours	66.0	73.0	**11.0**	55.0	45.0	**6.0**	33.9	56.1	**22.2**
Claude-3-sonnet									
Vanilla	**59.5**	90.4	30.9	54.0	88.3	35.9	32.2	84.1	51.8
Top-K	52.5	68.0	15.5	51.0	64.1	13.2	28.3	51.5	24.4
Quiz-like	57.5	96.0	38.5	52.0	91.0	39.0	**33.3**	86.1	52.8
Ours	58.0	67.0	**9.0**	55.0	46.0	**9.0**	31.7	53.3	**21.7**

Table 6: A comparison of confidence elicitation and performance for Gemma2-9B, LLaMA-3-70B, and Claude-3-Sonnet in the **Chemistry** domain across three difficulty levels.

Method	High School			College			Expert		
	Acc	AvC	ECE	Acc	AvC	ECE	Acc	AvC	ECE
Gemma2-9B									
Vanilla	**91.3**	99.2	7.9	**87.1**	99.2	12.1	**50.0**	96.6	46.6
Top-K	84.5	93.3	10.4	85.0	93.5	**9.9**	42.9	88.8	46.0
Quiz-like	87.7	99.4	12.3	86.4	99.3	12.9	41.4	97.1	55.7
Ours	87.7	89.7	**7.7**	87.1	88.6	14.3	42.9	81.4	**38.6**
LLaMA-3-70B									
Vanilla	**90.0**	89.8	**2.1**	90.0	88.6	**1.4**	54.3	84.1	29.9
Top-K	88.7	79.6	9.3	87.9	77.6	10.8	54.3	70.0	15.7
Quiz-like	87.7	96.8	9.0	**90.7**	90.0	5.0	**60.0**	82.9	22.9
Ours	88.4	78.1	10.3	90.0	72.9	17.1	**60.0**	62.9	**11.4**
Claude-3-sonnet									
Vanilla	**89.7**	90.5	**2.3**	87.1	90.5	**4.2**	47.1	87.1	39.9
Top-K	84.8	78.8	6.9	81.4	74.4	8.1	47.1	55.2	12.6
Quiz-like	88.1	99.0	11.0	**90.7**	97.1	7.9	**54.3**	87.1	32.9
Ours	88.4	77.1	11.3	89.3	72.1	17.1	**54.3**	64.3	**10.0**

Table 7: A comparison of confidence elicitation and performance for Gemma2-9B, LLaMA-3-70B, and Claude-3-Sonnet in the **Biology** domain across three difficulty levels.