Beware of Reasoning Overconfidence: Pitfalls in the Reasoning Process for Multi-solution Tasks

Jiannan Guan[1*], Qiguang Chen[1*], Libo Qin[2†], Dengyun Peng[1], Jinhao Liu[1], Liangyu Huo[3], Jian Xie[3], Wanxiang Che[1†]

[1]Research Center for Social Computing and Interactive Robotics, Harbin Institute of Technology
[2]School of Computer Science and Engineering, Central South University
[3]Du Xiaoman (Beijing) Science Technology Co., Ltd.
{jnguan, qgchen}@ir.hit.edu.cn, lbqin@csu.edu.cn, car@ir.hit.edu.cn

Abstract

Large Language Models (LLMs) excel in reasoning tasks requiring a single correct answer, but they perform poorly in multi-solution tasks that require generating comprehensive and diverse answers. We attribute this limitation to **reasoning overconfidence**: a tendency to express undue certainty in an incomplete solution set. To examine the effect, we introduce *MuSoBench*, a benchmark of multi-solution problems. Experiments show that the conventional short chain-of-thought (Short-CoT) prompting paradigm exhibits pronounced overconfidence, whereas the emerging long chain-of-thought (Long-CoT) approach mitigates it through iterative exploration and self-reflection. We further characterise observable behaviours and influential factors. To probe the underlying cause, we propose the **cognitive-rigidity hypothesis**, which posits that overconfidence arises when the reasoning process prematurely converges on a narrow set of thought paths. An attention-entropy analysis offers preliminary support for this view. These findings provide tools for assessing the completeness of LLM reasoning and highlight the need to move evaluation beyond single-answer accuracy toward comprehensive exploration.

Code —
https://github.com/jubgjf/reasoning-overconfidence

1 Introduction

Recently, Large Language Models (LLMs) have shown strong performance on tasks requiring multiple correct answers (Achiam et al. 2023; Yang et al. 2025; Zhuang et al. 2023; Qin et al. 2024). As illustrated in Figure 1, consider planning every possible dinner from a fixed set of ingredients: success lies in listing the full menu, not a single dish. We call such problems multi-solution reasoning tasks, whose goal is completeness and diversity. Yet advanced methods like Chain-of-Thought (CoT) (Wei et al. 2022), designed for one reasoning path, often stop early. When asked to list all answers, LLMs usually produce a few options and then assert confidently that no others exist. As Figure 1 & 2 depict, this overconfidence significantly reduces possible solution exploration, exposing a mismatch between stated confidence and actual coverage. To systematically analyze this failure mode on multi-solution tasks, we introduce the concept of **Reasoning Overconfidence**: A model's subjective confidence in its solution set significantly exceeds its actual ability to recover the full set of correct answers.

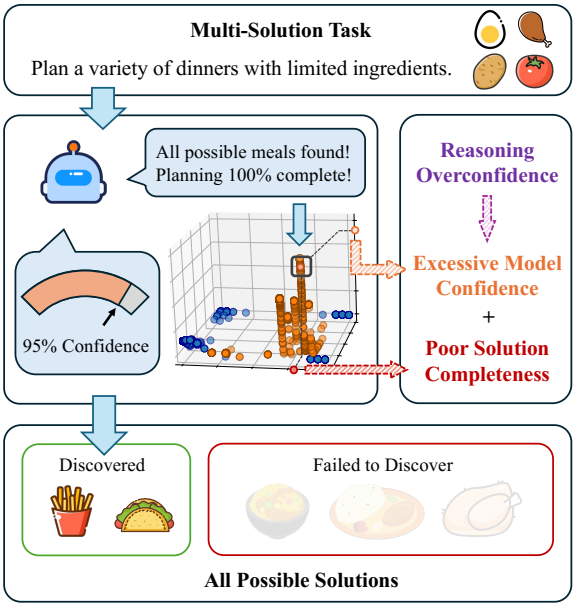


Figure 1: On multi-solution tasks, the model suffers from **reasoning overconfidence**, exhibiting excessively high confidence while exploring only a few reasoning paths. This leads to a poor completeness score for the final task.

Prior work has examined LLM performance on multi-solution tasks. Some work focuses on reasoning under structured constraints. For example, the 24-point game requires enumerating all valid arithmetic expressions (Yao et al. 2023), while CalibratedMath use problems with multiple correct answers to assess uncertainty calibration (Lin, Hilton, and Evans 2022). Others center on open-ended generation. Creative tasks such as story generation evaluate pro-

*These authors contributed equally.
†Corresponding Author

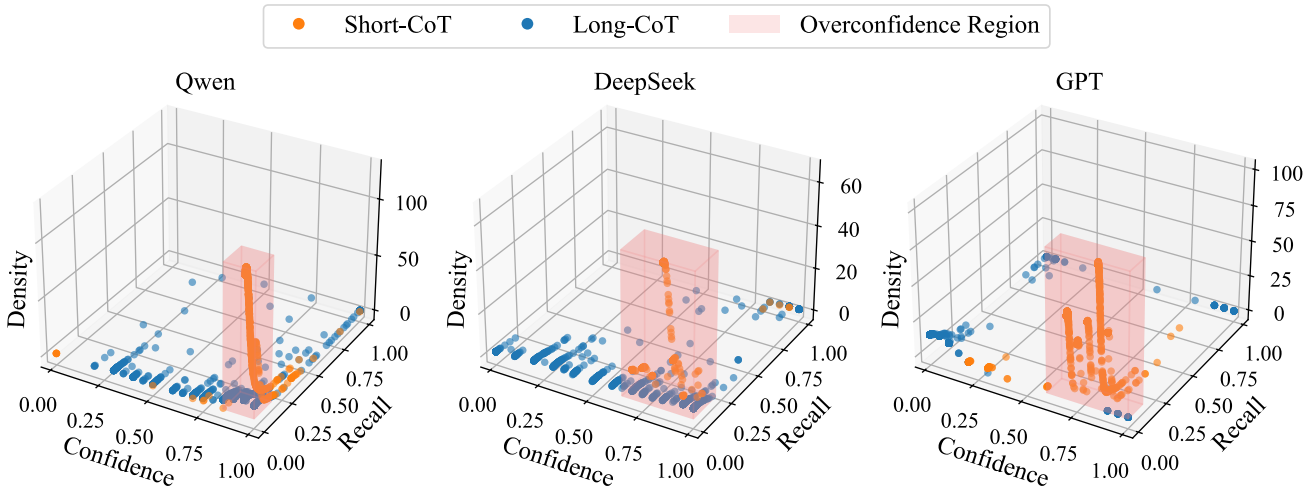


Figure 2: Distribution plots of recall vs. confidence on the TimeTabling dataset. The plots clearly show Short-CoT results clustering in the low-recall, high-confidence corner (red). For SubsetSum results, see Figure 8 in Appendix.

ducing diverse content from an effectively unbounded solution space (Xu et al. 2024; Wang and Kreminski 2024). However, the datasets used in existing work share a fundamental limitation: Their solution spaces are either tightly constrained or nearly unbounded, which hinders reliable estimation of completeness.

To enable more precise empirical study, we developed the **Multi-solution Benchmark (MuSoBench)**, a task suite designed to evaluate reasoning overconfidence under controlled solution spaces. When applied to MuSoBench, the conventional short chain-of-thought (Short-CoT) approach exhibits persistent overconfidence: as shown in Figure 2, its outputs cluster in the high-confidence, low-recall region, providing direct empirical evidence of this behavior. Behavioral analysis indicates that Short-CoT performs a shallow search, seldom revising its initial reasoning path, which largely explains its inflated confidence. By contrast, the Long-CoT paradigm, which promotes iterative exploration and self-reflection (Chen et al. 2025a,b; Zeng et al. 2024; Li et al. 2025), substantially improves both recall and precision, thereby reducing reasoning overconfidence. Finally, analysis of internal activations supports the cognitive rigidity hypothesis, which attributes overconfidence to premature convergence on a narrow set of reasoning paths.

The main contributions are summarized as follows:

- We first introduce the reasoning overconfidence concept as a critical failure mode of LLMs on multi-solution tasks and present MuSoBench, a new benchmark that documents this phenomenon through evidence on solution diversity, stability, and calibration.
- We analyze factors influencing overconfidence and its mitigation. Our results show that the extent of overconfidence is governed chiefly by the length of the reasoning trace, the presence of reflective steps, and the breadth of exploration, thereby linking the phenomenon to both reasoning paradigms and task properties.
- We advance the cognitive-rigidity hypothesis to explain this behavior, examining internal model states that give rise to overconfidence and offering a fresh perspective on the fundamental multi-solution reasoning.

2 Problem Formulation & Benchmark

2.1 Multi-solution Tasks

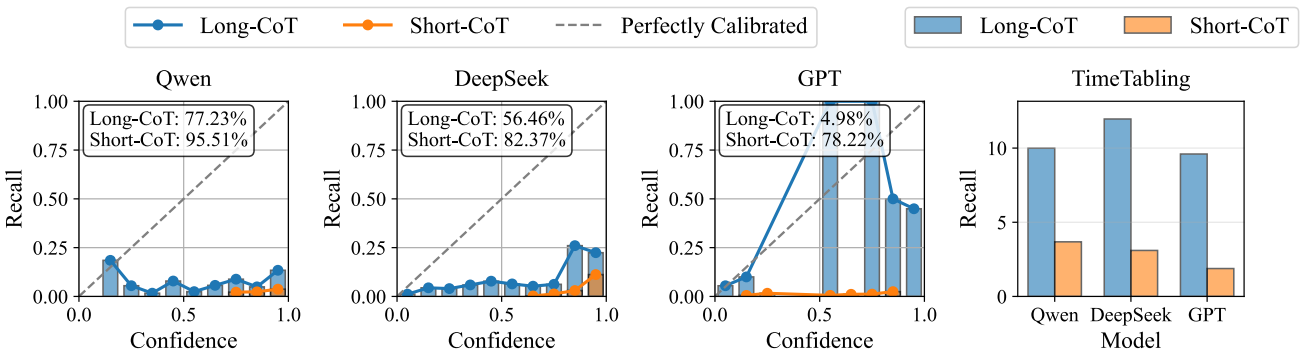
Multi-solution tasks require a model to enumerate all valid answers to a single problem rather than to return just one. Formally, such a task is characterised by a dataset $\mathcal{T} = \{(x_i, \hat{\mathcal{Y}}_i) \mid |\hat{\mathcal{Y}}_i| \geq 1\}_{i=1}^{N}$, where x_i denotes the i-th problem instance and $\hat{\mathcal{Y}}_i$ is the corresponding ground-truth solution set. Given x_i, a model $\mathcal{M}$ outputs its own solution set $\mathcal{Y}_i = \mathcal{M}(x_i)$. The ideal outcome is that $\mathcal{Y}_i$ matches $\hat{\mathcal{Y}}_i$ exactly, covering every valid solution and omitting none.

2.2 Reasoning Overconfidence

Reasoning **O**ver**C**onfidence (ROC) refers to a model's tendency to report confidence levels that exceed its actual performance. In multi-solution tasks, this mismatch often drives the model to return only a subset of correct answers and to stop searching prematurely. To characterise the phenomenon quantitatively, we use the *Expected Calibration Error* (ECE), which compares expected and realised performance. Let $\mathcal{C}(\mathcal{Y}_i \mid x_i, \mathcal{M})$ denote the confidence that model $\mathcal{M}$ assigns to its proposed solution set $\mathcal{Y}_i$. A model is *reasoning-overconfident* when its reported confidence exceeds its observed performance $\mathrm{Perf}(\cdot)$:

$$\mathcal{C}(\mathcal{Y}_i|x_i, \mathcal{M}) > \mathrm{Perf}(\mathcal{Y}_i, \hat{\mathcal{Y}}_i) \tag{1}$$

For tasks with multiple valid answers, the relevant performance dimension is the **completeness** of the returned set. Because our focus is the model's failure to enumerate *all* correct solutions, not the precision of each individual answer, we measure performance with **Recall**. Metrics that



(a) Reliability Diagrams of Short-CoT vs. Long-CoT on TimeTabling across different models.

(b) Performance of Short-CoT vs. Long-CoT on TimeTabling.

Figure 3: Calibration and performance of Short-CoT vs. Long-CoT on TimeTabling dataset. As shown in (a), the diagonal line represents perfect calibration. Long-CoT models (blue) are better calibrated than Short-CoT models (orange). As shown in (b), Long-CoT models achieve significantly higher recall than Short-CoT models. For SubsetSum results, see Figure 9 in Appendix.

combine precision and recall (e.g., F1-score) would blur the exploration shortfall we seek to isolate; hence, we instantiate $\mathrm{Perf}(\cdot)$ as Recall throughout this work.

2.3 MuSoBench Construction

To systematically and controllably evaluate models on multi-solution tasks, we introduce the **Multi-Solution Benchmark (MuSoBench)**, comprising two subtasks: *TimeTabling* and *SubsetSum*. The *TimeTabling* subtask is to construct conflict-free course schedules subject to constraints on course overlap, instructor availability, and classroom capacity. The *SubsetSum* subtask requires enumerating all subsets of a given integer set summing to target number.

Problem complexity is measured by the size of each instance's solution space. The TimeTabling corpus spans ten complexity levels and the SubsetSum corpus seven, with 100 instances per level. For every instance, we algorithmically enumerated all feasible solutions and manually verified them to ensure correctness and completeness. A detailed description of the dataset construction procedure, together with illustrative examples, is provided in Appendix A.1.

To quantitatively assess model behavior in multi-solution scenarios, we utilize the following metrics:

- **Model Performance Metric:** (1) *Precision (P)* (↑): Proportion of correct model response. (2) *Recall (R)* (↑): Proportion of correct answers the model recovers. This is the primary metric for multi-solution tasks.
- **Overconfidence Metric:** *Expected Calibration Error (ECE)* (↓): Average gap between confidence and realized performance. Lower ECE indicates better calibration.
- **Model Behavior Metric:** (1) *Correct Solution Retention Rate (CSR)* (↑): Capability to maintain previously correct solutions. (2) *Error Solution Correction Rate (ESC)* (↑): Capability to correct earlier error solutions. (3) *New Solution Discovery Rate (NSD)* (↑): Capability to discover additional correct solutions.

All detailed formulas are in Appendix A.2.

3 Experimental Setup

Our experiments are conducted on the Qwen, DeepSeek, and GPT series of models. We mainly compare the following two CoT reasoning paradigms:

- **Short-CoT:** Appends zero-CoT prompt (Kojima et al. 2022) to a base instruction to elicit CoT from standard LLMs, including Qwen3-8B (non-thinking mode) (Yang et al. 2025), DeepSeek-V3 (Liu et al. 2024), and gpt-4o-mini (Achiam et al. 2023).
- **Long-CoT:** Uses models trained for extended, iterative reasoning and reflection; we use Qwen3-8B (thinking mode) (Yang et al. 2025), DeepSeek-R1 (Guo et al. 2025), and o4-mini (Jaech et al. 2024).

We contend this is a valid comparison of the entire paradigm as deployed in practice, rather than comparing different prompts or training methods.

To estimate LLM confidence, we use a verbal-elicitation strategy (Xiong et al. 2023). This approach applies to both open-source and API-only models and is supported by evidence that verbalized confidence closely tracks internal probabilities (Kumar et al. 2024). In our setup, the model first generates a full answer and then reports its confidence on a 0–100 scale. The prompts are given in Appendix A.3.

4 Analysis of Reasoning Overconfidence

4.1 Existence Verification

Short-CoT displays substantial reasoning overconfidence across all model series. To assess the existence of ROC, we quantify this phenomenon using recall–confidence reliability diagrams, in which a perfectly calibrated model aligns with the main diagonal. In Figure 3 (a), the Short-CoT bars aggregate in the lower-right quadrant, indicating high confidence yet low recall, and therefore fall well below the diagonal. Consistently, Short-CoT produces expected calibration error (ECE) values above 78.22% for every model series, corroborating its marked reasoning overconfidence.

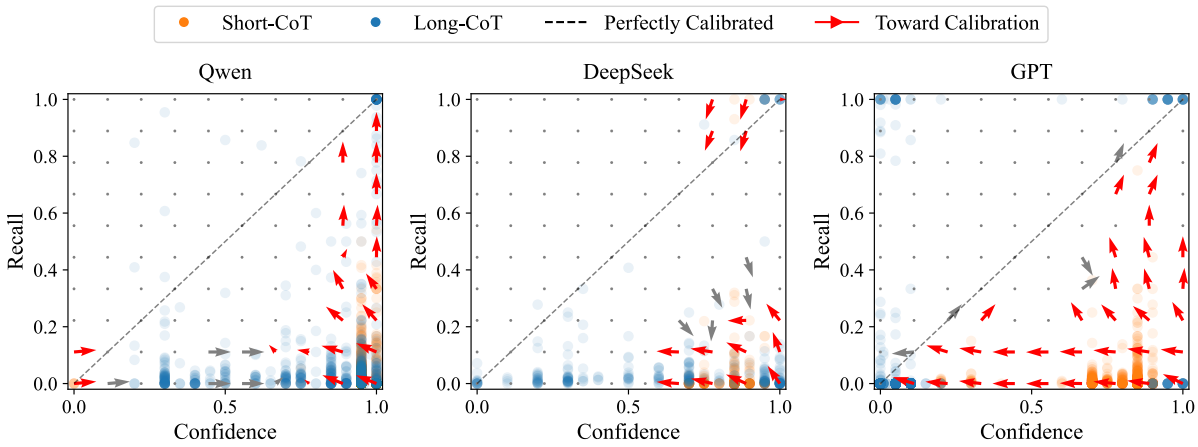


Figure 4: The arrows indicate the movement of model confidence and performance from Short-CoT to Long-CoT. The results show that adopting Long-CoT causes most data points to shift toward the diagonal, indicating improved calibration (red). Results for SubsetSum are shown in Figure 10 in Appendix.

Model	Paradigm	TimeTabling			SubsetSum		
		CSR (%) ↑	ESC (%) ↑	NSD (%) ↑	CSR (%) ↑	ESC (%) ↑	NSD (%) ↑
Qwen	Short-CoT	**43.15**	65.22	0.19	14.89	15.63	0.20
	Long-CoT	53.13	**81.83**	**2.50**	**75.65**	**93.93**	**0.49**
DeepSeek	Short-CoT	6.77	**98.86**	0.01	**70.44**	95.14	0.27
	Long-CoT	**47.56**	97.51	**0.86**	63.95	**96.58**	**0.70**
GPT	Short-CoT	**34.34**	71.05	0.26	30.79	33.52	0.16
	Long-CoT	10.41	**92.12**	**1.28**	**45.05**	**99.23**	**3.03**

Table 1: Long-CoT models demonstrate significantly higher rates of Error Correction and New Solution Discovery, indicating a more flexible and reflective reasoning process compared to the cognitive rigidity of Short-CoT.

Long-CoT lowers reasoning over-confidence relative to Short-CoT, yet open-source models still need further improvement. Figure 3 (a) shows that the Long-CoT reliability curve lies closer to the diagonal compared to Short-CoT, indicating better calibration and a lower rate of ROC. Quantitatively, Long-CoT decreases the ECE by at least 18.28% across all model families and even delivers single-digit ECE for closed-source models. Nevertheless, open-source models continue to exhibit pronounced ROC, exceeding 56.46%, even after Long-CoT prompting. In sum, although Long-CoT markedly mitigates ROC compared with Short-CoT, additional advances are required to enhance calibration in open-source models.

4.2 Behavioral Diagnostics

We demonstrate the cause of ROC under the Short-CoT paradigm. We discover that models meeting with ROC can trigger the following unexpected behavior:

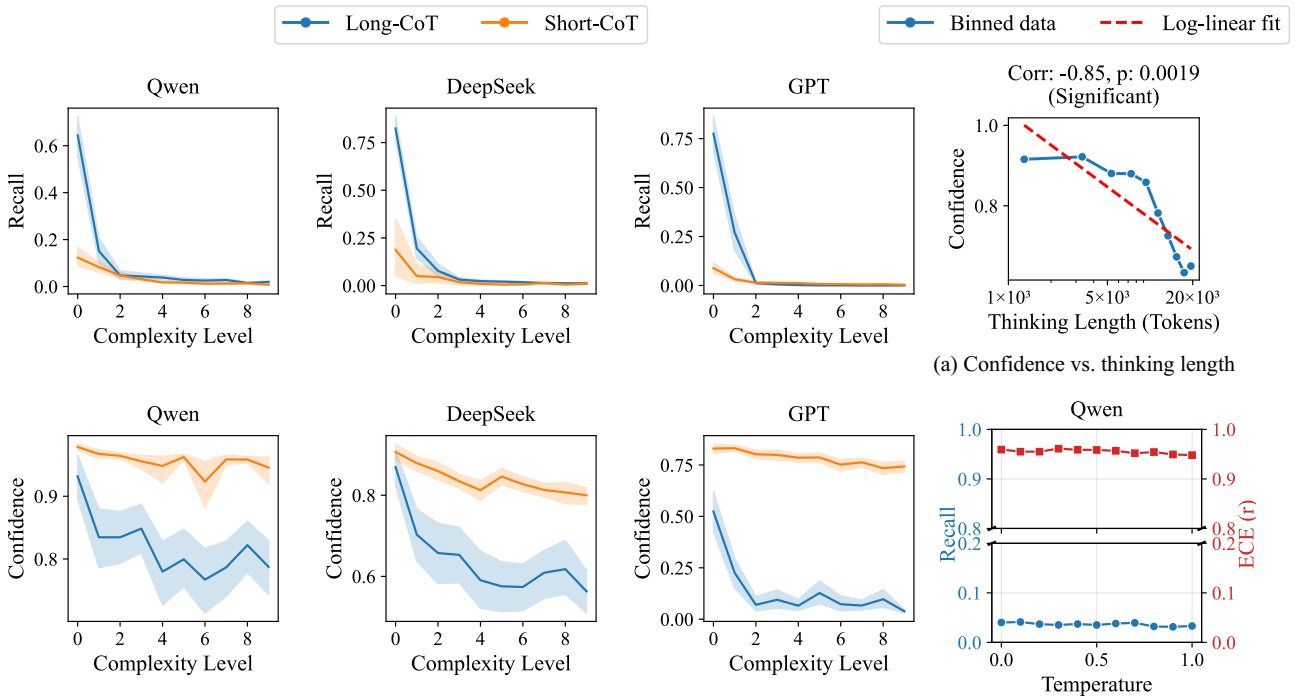
Calibrating ROC is coupled with actual performance improvement. Short-CoT instances concentrate in the **"low-recall, high-confidence"** quadrant (Figure 2). By halting early, as shown in Figure 3 (b), the model retrieves few correct solutions yet remains overly sure of their completeness, constraining answer diversity. Figure 4 plots, for each problem, the vector from the Short-CoT point to its Long-CoT counterpart. The consistent upward shift indicates that Long-CoT searches more exhaustively and recovers solutions missed by Short-CoT, while the concurrent leftward shift toward lower confidence reveals better self-calibration. Thus, ROC calibration aligns with performance gains.

Cognitive Rigidity and Resistance to Guidance. As shown in Table 1, models using Short-CoT exhibit extremely low error correction and new solution discovery rates. Even when prompted to reconsider, the model is largely unable to identify its previous errors or explore new, correct paths. This cognitive rigidity indicates that the model stops exploring alternative reasoning paths once it settles on an initial set of answers. Long-CoT models, in contrast, are far more capable of self-correction and discovery. Experiments on dataset generalizability are available in Appendix A.4.

4.3 Influencing Factors

We now investigate key factors that mitigate ROC behavior.

Inference-Time Scaling Law also holds for the phenomenon of reasoning overconfidence. Building on the



(a) Confidence vs. thinking length

(b) Performance (Recall, top row) and confidence (bottom row) as a function of task complexity on the TimeTabling dataset.

(c) The effect of decoding temperature on Recall and ECE.

Figure 5: Factors that influence reasoning overconfidence. (a) A strong negative correlation shows that Long-CoT has moderate confidence. (b) As task complexity rises, Short-CoT keeps unjustifiably high confidence despite falling recall, indicating poor self-monitoring, whereas Long-CoT lowers its confidence in line with the harder setting, demonstrating better calibration. (c) Decoding temperature has little effect on recall or expected calibration error. More results see Figure 11 in Appendix.

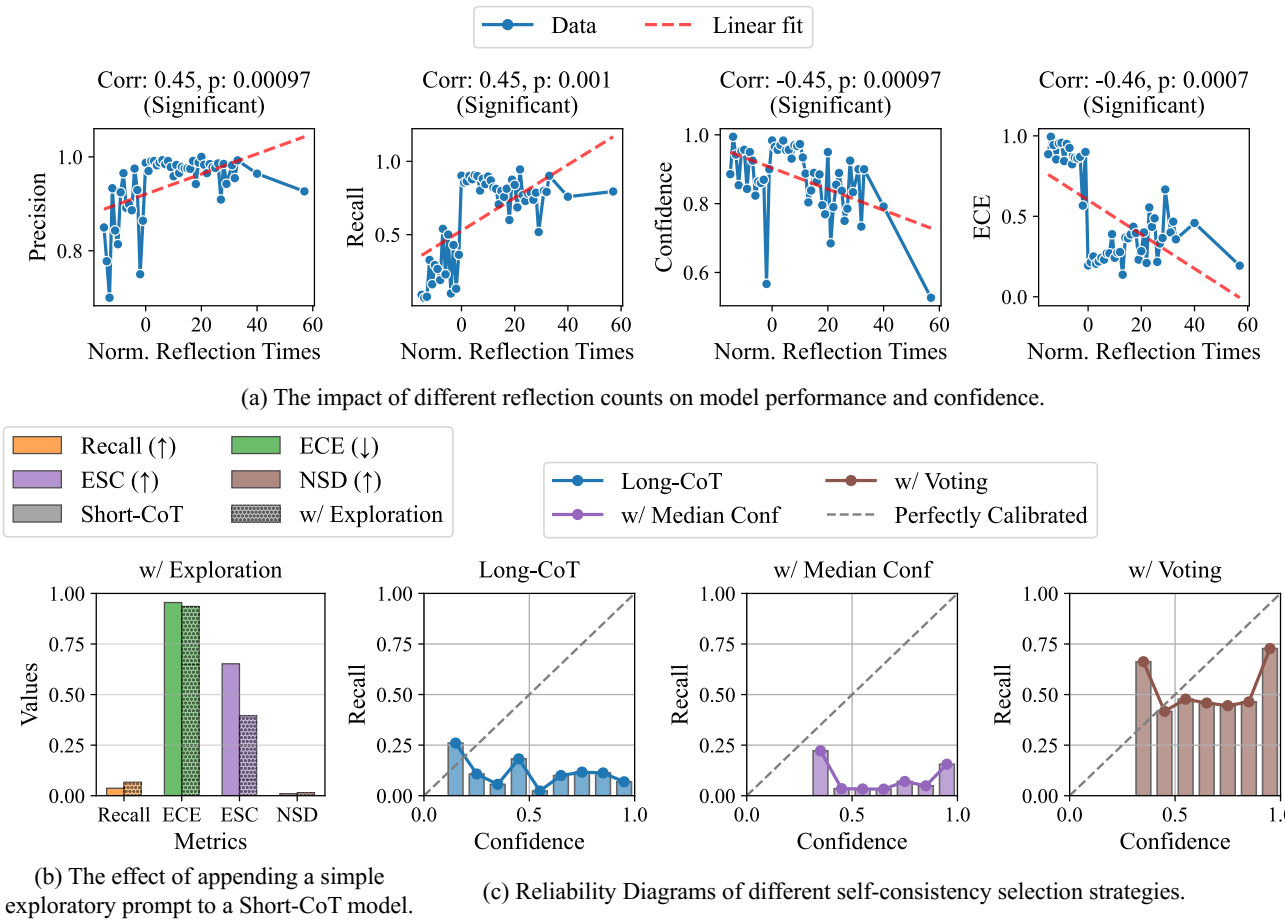
inference-time scaling law, which raises accuracy by increasing inference calculations (Wu et al. 2024), we investigate whether longer chains also temper ROC. Figure 5 (a) shows a strong negative correlation: extended reasoning chains yield lower confidence estimates. Thus, extra computation not only improves accuracy but also calibrates predictions by curbing ROC; longer reasoning promotes a more cautious self-assessment.

Long-CoT strategies substantially reduce ROC as task complexity increases, whereas Short-CoT is insensitive to task complexity. Following previous work (Xu et al. 2025), we posit that greater complexity, reflected by a bigger solution space, should dampen a model's confidence. Figure 5(b) confirms this expectation for Long-CoT: its confidence decreases as complexity rises, indicating proper calibration. Short-CoT, however, maintains high confidence while recall drops sharply, revealing persistent overconfidence. Hence, Long-CoT acknowledges growing difficulty, whereas Short-CoT remains blind to task complexity.

Decoding temperature increases token-level diversity rather than alleviating ROC or expanding solution exploration. Figure 5(c) demonstrates that adjusting the decoding temperature scarcely affects LLM confidence and fails to reduce ROC. Higher temperatures likewise do not encourage a broader search across alternative reasoning paths. Instead, temperature primarily injects stochastic variation at the token level, enlarging lexical diversity. These observations indicate that ROC originates in the model's intrinsic reasoning mechanism rather than in tunable decoding heuristics.

Method	P (%) ↑	R (%) ↑	ECE (r) (↓)
TimeTabling			
Long-CoT	54.88	9.99	77.23
w/ Median Conf	59.10	19.81	81.13
w/ Voting	**74.04**	**49.55**	**64.32**
SubsetSum			
Long-CoT	85.98	21.24	68.31
w/ Median Conf	**92.13**	33.33	71.68
w/ Voting	72.75	**51.40**	**65.43**

Table 2: Performance of self-consistency strategies on the Long-CoT Qwen3-8B model, using n=32 parallel reasoning paths. Results highlight that the choice of aggregation strategy involves significant trade-offs between performance and model calibration.



(a) The impact of different reflection counts on model performance and confidence.

(b) The effect of appending a simple exploratory prompt to a Short-CoT model.

(c) Reliability Diagrams of different self-consistency selection strategies.

Figure 6: Mitigate reasoning overconfidence through various strategies. (a) As reflection time increases, recall improves, and overconfidence decreases significantly. (b) A simple exploratory prompt boosts both precision and recall, demonstrating its potential at breaking cognitive rigidity. (c) *w/ Voting* aggregation strategy significantly improves calibration. More results are shown in Figure 12 in Appendix.

5 Mitigation Strategies

Building on Chen et al. (2025b), we separate the reflection and exploration mechanisms of Long-CoT, to quantify their individual effects and derive practical guidance.

Reflection steps enhance solution diversity and reduce ROC. We evaluated the reflection's effect on ROC by pausing Long-CoT at scheduled checkpoints. After a fixed number of iterations, and before the final answer, we inserted the control token `</think>` to record the intermediate output. Figure 6 (a) reveals that additional reflection rounds raise recall while lowering confidence, thereby lessening ROC. This evidence indicates that reflection exposes otherwise overlooked solutions and thus decreases ROC.

Sequential exploration–scaling prompts unlock rigid thinking and reduce ROC in Short-CoT. We tested whether ROC of Short-CoT can be eased by an external exploratory cue. After the model produced its initial answer, we appended the prompt "Wait, there may be other solutions." Figure 6 (b) shows that this cue markedly improves performance. The added exploration frees the model from cognitive rigidity and yields more correct answers. ROC also drops slightly, as shown in Figure 13 in Appendix.

Parallel exploration–scaling prompting mitigates ROC and boosts recall. We use a self-consistency paradigm that generates multiple reasoning paths in parallel and aggregates them by two strategies: (1) *w/ Median Conf*: select the path with the median confidence score; (2) *w/ Voting*: unite all unique answers and weight their confidence by frequency. As shown in Table 2, aggregation choice is decisive. *w/ Voting* markedly increases recall on both datasets and improves calibration, indicating stronger ROC mitigation and broader solution coverage. By contrast, as shown in Figure 6 (c), *Median Conf* raises precision and recall over the baseline but degrades calibration, worsening ECE.

6 Investigating the Internal Mechanism

To explain the rationale of reasoning overconfidence, we introduce the **cognitive-rigidity hypothesis**. It posits that

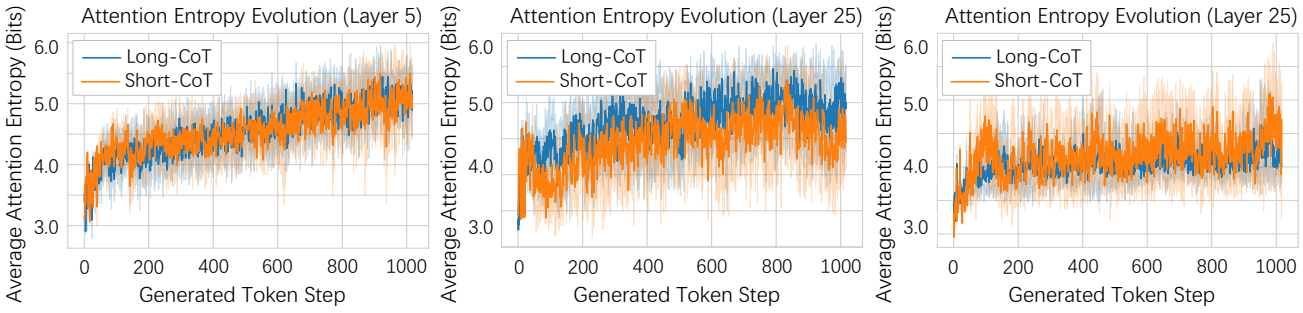


Figure 7: Attention entropy of Short-CoT vs. Long-CoT on Qwen3-8B. We present layers 5, 25, and 35 as representative examples of three phases of the reasoning process. The model's entropy evolution follows a three-phase pattern of convergence, divergence, and reversal: the Long-CoT paradigm shows higher entropy in core layers, promoting exploration, whereas the Short-CoT paradigm ends with higher entropy in the deepest layers. Detailed results are provided in Figure 14 in the Appendix.

ROC arises when the model's core reasoning layers lock too early onto a single trajectory, exhibiting cognitive rigidity. Following Cui et al. (2025), we treat attention entropy as a proxy for internal diversity. Accordingly, we compute layer-wise attention entropy for Qwen3-8B under the Short-CoT and the lower-ROC Long-CoT settings.

Paradigms with reduced ROC show low attention entropy in core reasoning layers, indicating cognitive rigidity. Figure 7 exhibits three phases consistent with Cui et al. (2025): (1) Shallow layers (0–10): Initial entropies differ but rapidly converge. (2) Core reasoning layers (15–30): Reduced-ROC paradigms sustain higher entropy than high-ROC counterparts. (3) Deep layers (≈35): The trend reverses; high-ROC paradigms terminate with greater entropy.

These patterns support the cognitive-rigidity hypothesis. In core layers, the low entropy of high-ROC paradigms signals narrowly focused attention that restricts exploratory reasoning, whereas Long-CoT's higher entropy reflects flexibility to pursue alternative paths. The final reversal strengthens this interpretation: high-ROC rigidity produces late-stage uncertainty (high entropy), while Long-CoT, having explored more broadly, converges decisively (low entropy).

7 Related Work

Chain-of-Thought (CoT) prompting markedly elevates Large Language Models' reasoning by eliciting explicit intermediate steps that mirror human cognition (Wei et al. 2022; Chen et al. 2024a,b; Cheng et al. 2025b). Yet its standard, concise variant, Short-CoT, remains brittle (Qin et al. 2023; Chen et al. 2025b). To improve robustness, later work introduced more expressive reasoning topologies. Tree-of-Thought (Yao et al. 2023) and Graph-of-Thoughts (Besta et al. 2024), for example, let models explore multiple reasoning paths concurrently, enabling branching and backtracking to boost the likelihood of a correct answer (Hu et al. 2024).

A model's ability to report accurate confidence is crucial for its reliability, especially in high-stakes applications and for detecting hallucinations (Amodei et al. 2016; Hou et al. 2024). However, a significant body of work has shown that LLMs are often poorly calibrated and exhibit overconfidence, expressing high certainty in answers that are incorrect or incomplete (Jiang et al. 2020; Singh et al. 2023). Research has explored mitigating this issue through methods like probability calibration (Fisch, Jaakkola, and Barzilay 2022; Guo et al. 2017) or by developing prompting strategies that leverage self-consistency or relative ranking to better estimate verbal confidence (Li et al. 2024a; Yang et al. 2024; Shrivastava, Kumar, and Liang 2025). Overconfidence is pronounced in CoT reasoning: orderly steps can falsely signal correctness, leading models to commit before exploring the full solution space (Ling et al. 2023; Cheng et al. 2025a). Although recent work shows Long-CoT models are better calibrated than Short-CoT ones (Yoon et al. 2025), these studies, and calibration research in general, still judge confidence only by the accuracy of a single final answer.

Prior efforts that allow multiple solutions either restrict the solution space to narrow fact-checking tasks (Lin, Hilton, and Evans 2021; Li et al. 2024b) or apply standard calibration metrics (Lin, Hilton, and Evans 2022) without addressing the unique failure modes of exhaustive generation. To examine a key failure mode in multi-solution scenarios, we formally define and analyze Reasoning Overconfidence: a model's unwarranted certainty that its generated solution set is complete. We shift evaluation from confidence-accuracy to confidence-completeness correlations, offering a fresh lens for enhancing LLM reliability.

8 Conclusion

This paper identifies reasoning overconfidence in LLMs as a critical failure mode on multi-solution tasks, where standard Short-CoT yields incomplete yet highly confident solutions. To better understand this phenomenon, our analysis delves into its key influencing factors. In response to this problem, we demonstrate that the emerging Long-CoT effectively mitigates this issue by improving both solution diversity and confidence calibration. We attribute the success of Long-CoT to its ability to overcome our proposed cognitive-rigidity hypothesis: a state where Short-CoT locks the model into a narrow search space. These findings underscore the limitations of conventional CoT and call for more exploratory reasoning paradigms to build reliable AI.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (NSFC) via grants 62236004, 62476073, 92570120, and 62306342. This work was sponsored by the CCF-Zhipu Large Model Innovation Fund (NO.CCF-Zhipu202406).

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; and Mané, D. 2016. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.

Besta, M.; Blach, N.; Kubicek, A.; Gerstenberger, R.; Podstawski, M.; Gianinazzi, L.; Gajda, J.; Lehmann, T.; Niewiadomski, H.; Nyczyk, P.; et al. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 17682–17690.

Chen, Q.; Peng, D.; Liu, J.; Su, H.; Guan, J.; Qin, L.; and Che, W. 2025a. Aware First, Think Less: Dynamic Boundary Self-Awareness Drives Extreme Reasoning Efficiency in Large Language Models. *arXiv preprint arXiv:2508.11582*.

Chen, Q.; Qin, L.; Liu, J.; Peng, D.; Guan, J.; Wang, P.; Hu, M.; Zhou, Y.; Gao, T.; and Che, W. 2025b. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint arXiv:2503.09567*.

Chen, Q.; Qin, L.; Wang, J.; Zhou, J.; and Che, W. 2024a. Unlocking the capabilities of thought: A reasoning boundary framework to quantify and optimize chain-of-thought. *Advances in Neural Information Processing Systems*, 37: 54872–54904.

Chen, Q.; Qin, L.; Zhang, J.; Chen, Z.; Xu, X.; and Che, W. 2024b. M ^{3} CoT: A Novel Benchmark for Multi-Domain Multi-step Multi-modal Chain-of-Thought. *arXiv preprint arXiv:2405.16473*.

Cheng, J.; Su, T.; Yuan, J.; He, G.; Liu, J.; Tao, X.; Xie, J.; and Li, H. 2025a. Chain-of-Thought Prompting Obscures Hallucination Cues in Large Language Models: An Empirical Evaluation. *arXiv preprint arXiv:2506.17088*.

Cheng, Z.; Chen, Q.; Xu, X.; Wang, J.; Wang, W.; Fei, H.; Wang, Y.; Wang, A. J.; Chen, Z.; Che, W.; et al. 2025b. Visual thoughts: A unified perspective of understanding multimodal chain-of-thought. *arXiv preprint arXiv:2505.15510*.

Cui, G.; Zhang, Y.; Chen, J.; Yuan, L.; Wang, Z.; Zuo, Y.; Li, H.; Fan, Y.; Chen, H.; Chen, W.; et al. 2025. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*.

Fisch, A.; Jaakkola, T.; and Barzilay, R. 2022. Calibrated selective classification. *arXiv preprint arXiv:2208.12084*.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. In *International conference on machine learning*, 1321–1330. PMLR.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Hou, B.; Zhang, Y.; Andreas, J.; and Chang, S. 2024. A probabilistic framework for llm hallucination detection via belief tree propagation. *arXiv preprint arXiv:2406.06950*.

Hu, M.; Mu, Y.; Yu, X. C.; Ding, M.; Wu, S.; Shao, W.; Chen, Q.; Wang, B.; Qiao, Y.; and Luo, P. 2024. Tree-Planner: Efficient Close-loop Task Planning with Large Language Models. In *The Twelfth International Conference on Learning Representations*.

Jaech, A.; Kalai, A.; Lerer, A.; Richardson, A.; El-Kishky, A.; Low, A.; Helyar, A.; Madry, A.; Beutel, A.; Carney, A.; et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Jiang, Z.; Xu, F. F.; Araki, J.; and Neubig, G. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8: 423–438.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213.

Kumar, A.; Morabito, R.; Umbet, S.; Kabbara, J.; and Emami, A. 2024. Confidence under the hood: An investigation into the confidence-probability alignment in large language models. *arXiv preprint arXiv:2405.16282*.

Li, L.; Chen, Z.; Chen, G.; Zhang, Y.; Su, Y.; Xing, E.; and Zhang, K. 2024a. Confidence matters: Revisiting intrinsic self-correction capabilities of large language models. *arXiv preprint arXiv:2402.12563*.

Li, M.; Wang, W.; Feng, F.; Zhu, F.; Wang, Q.; and Chua, T.-S. 2024b. Think twice before trusting: Self-detection for large language models through comprehensive answer reflection. *arXiv preprint arXiv:2403.09972*.

Li, Z.-Z.; Zhang, D.; Zhang, M.-L.; Zhang, J.; Liu, Z.; Yao, Y.; Xu, H.; Zheng, J.; Wang, P.-J.; Chen, X.; et al. 2025. From system 1 to system 2: A survey of reasoning large language models. *arXiv preprint arXiv:2502.17419*.

Lin, S.; Hilton, J.; and Evans, O. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Lin, S.; Hilton, J.; and Evans, O. 2022. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*.

Ling, Z.; Fang, Y.; Li, X.; Huang, Z.; Lee, M.; Memisevic, R.; and Su, H. 2023. Deductive verification of chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 36: 36407–36433.

Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Qin, L.; Chen, Q.; Feng, X.; Wu, Y.; Zhang, Y.; Li, Y.; Li, M.; Che, W.; and Yu, P. S. 2024. Large language models meet nlp: A survey. *arXiv preprint arXiv:2405.12819*.

Qin, L.; Chen, Q.; Wei, F.; Huang, S.; and Che, W. 2023. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. *arXiv preprint arXiv:2310.14799*.

Shrivastava, V.; Kumar, A.; and Liang, P. 2025. Language Models Prefer What They Know: Relative Confidence Estimation via Confidence Preferences. *arXiv preprint arXiv:2502.01126*.

Singh, A. K.; Devkota, S.; Lamichhane, B.; Dhakal, U.; and Dhakal, C. 2023. The confidence-competence gap in large language models: A cognitive study. *arXiv preprint arXiv:2309.16145*.

Wang, P. J.; and Kreminski, M. 2024. Guiding and diversifying LLM-based story generation via answer set programming. *arXiv preprint arXiv:2406.00554*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Wu, Y.; Sun, Z.; Li, S.; Welleck, S.; and Yang, Y. 2024. Inference scaling laws: An empirical analysis of compute-optimal inference for problem-solving with language models. *arXiv preprint arXiv:2408.00724*.

Xiong, M.; Hu, Z.; Lu, X.; Li, Y.; Fu, J.; He, J.; and Hooi, B. 2023. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*.

Xu, C.; Wen, B.; Han, B.; Wolfe, R.; Wang, L. L.; and Howe, B. 2025. Do Language Models Mirror Human Confidence? Exploring Psychological Insights to Address Overconfidence in LLMs. *arXiv preprint arXiv:2506.00582*.

Xu, W.; Jojic, N.; Rao, S.; Brockett, C.; and Dolan, B. 2024. Echoes in ai: Quantifying lack of plot diversity in llm outputs. *arXiv preprint arXiv:2501.00273*.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yang, R.; Rajagopal, D.; Hayati, S. A.; Hu, B.; and Kang, D. 2024. Confidence calibration and rationalization for llms via multi-agent deliberation. *arXiv preprint arXiv:2404.09127*.

Yao, S.; Yu, D.; Zhao, J.; Shafran, I.; Griffiths, T.; Cao, Y.; and Narasimhan, K. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36: 11809–11822.

Yoon, D.; Kim, S.; Yang, S.; Kim, S.; Kim, S.; Kim, Y.; Choi, E.; Kim, Y.; and Seo, M. 2025. Reasoning models better express their confidence. *arXiv preprint arXiv:2505.14489*.

Zeng, Z.; Cheng, Q.; Yin, Z.; Wang, B.; Li, S.; Zhou, Y.; Guo, Q.; Huang, X.; and Qiu, X. 2024. Scaling of search and learning: A roadmap to reproduce o1 from reinforcement learning perspective. *arXiv preprint arXiv:2412.14135*.

Zhang, Y.; Diddee, H.; Holm, S.; Liu, H.; Liu, X.; Samuel, V.; Wang, B.; and Ippolito, D. 2025. NoveltyBench: Evaluating Language Models for Humanlike Diversity. *arXiv preprint arXiv:2504.05228*.

Zhuang, Z.; Chen, Q.; Ma, L.; Li, M.; Han, Y.; Qian, Y.; Bai, H.; Feng, Z.; Zhang, W.; and Liu, T. 2023. Through the lens of core competency: Survey on evaluation of large language models. *arXiv preprint arXiv:2308.07902*.