Self-Critique-Guided Curiosity Refinement: Enhancing Honesty and Helpfulness in Large Language Models via In-Context Learning

Anonymous ACL submission

Abstract

Large language models (LLMs) have demonstrated robust capabilities across various natural language tasks. However, producing outputs that are consistently honest and helpful remains an open challenge. To overcome this challenge, this paper tackles the problem through two complementary directions. It conducts a comprehensive benchmark evaluation of ten widely used large language models, including both proprietary and open-weight models from Google, Meta, and OpenAI. In parallel, it proposes a novel prompting strategy, self-critique-guided curiosity refinement prompting. The key idea behind this strategy is enabling models to self-critique and refine their responses without additional training. The proposed method extends the curiosity-driven prompting strategy by incorporating two lightweight in-context steps including self-critique step and refinement step. The experiment results on the HONESET dataset evaluated using the framework H^2 (honesty and helpfulness), which was executed with GPT-4o as a judge of honesty and helpfulness, show consistent improvements across all models. The approach reduces the number of poor-quality responses, increases high-quality responses, and achieves relative gains in H^2 scores ranging from 6.1% to 26.3% compared to raw prompting and 1.4% to 4.3% compared to curiosity-driven prompting across evaluated models. These results highlight the effectiveness of structured self-refinement as a scalable and training-free strategy to improve the trustworthiness of LLMs outputs.

1 Introduction

1.1 Motivation

Large language models (LLMs) have rapidly advanced and are now widely deployed across industries (OpenAI et al., 2024a; Grattafiori et al., 2024; Bommasani et al., 2021). Their capabilities in understanding, reasoning, and dialogue generation have enabled broad adoption, but also raised concerns about trustworthiness and alignment with human values. Two key user-facing dimensions of trust are honesty and helpfulness (Gao et al., 2024; Ji et al., 2023).

Honesty extends beyond factual accuracy to include acknowledging limitations and avoiding hallucinations (Askell et al., 2021; Rawte et al., 2023), while helpfulness requires providing useful, actionable guidance without sacrificing truthfulness (Gao et al., 2024; Sharma et al., 2024). Gao et al. (Gao et al., 2024) formalized these notions, introduced the HONESET dataset, and proposed curiosity-driven prompting and the H^2 evaluation framework. While promising, their work leaves open questions on performance across diverse LLMs, and whether self-critique can further enhance honesty and helpfulness without costly fine-tuning.

1.2 Paper Overview

This paper addresses three research questions:

1. How do ten diverse LLMs perform on HONESET under raw outputs vs. curiosity-driven prompting, evaluated by honest rate and H^2 evaluation framework?

2. Can a new in-context Critique-Guided Curiosity Refinement method further improve honesty and helpfulness?

3. How robust are these approaches across LLM architectures, scales, and developers?

Our contributions are threefold:

1. A systematic evaluation of ten widely used LLMs on HONESET using both honest rate and H^2 metrics.

2. A comparative analysis of raw outputs vs. curiosity-driven prompting.

3. Introduction of a training-free, lightweight Critique-Guided Curiosity Refinement method that improves response quality.

Together, these results advance techniques for enhancing honesty and helpfulness in LLMs and provide practical strategies for building more trustworthy AI systems.

2 Background

Honesty and helpfulness are central to building trustworthy LLMs. Honesty requires accurate, transparent, and non-deceptive responses, particularly in high-stakes domains such as law and medicine (Askell et al., 2021). Gao et al. (Gao et al., 2024) introduced the HONESET dataset to evaluate challenges such as outdated information, uncertainty calibration, and identity recognition, while also proposing the H^2 framework to jointly assess honesty and helpfulness. Helpfulness, though more subjective (Zhou et al., 2023b), emphasizes providing relevant and actionable guidance without sacrificing truthfulness.

To enhance these attributes, recent work has leveraged in-context learning (ICL), which allows LLMs to adapt behavior through prompting rather than retraining (Brown et al., 2020; Wei et al., 2022). Techniques like few-shot prompting and curiosity-driven prompting (Gao et al., 2024) demonstrate how lightweight prompting strategies can improve alignment. Building on this, self-critique and refinement methods (Lightman et al., 2023; Shinn et al., 2023) further encourage models to critique and revise their own outputs, improving adherence to Helpful, Honest, and Harmless (HHH) principles (Bai et al., 2022). Our work extends these ideas by introducing critique-guided curiosity refinement, a training-free approach that combines curiosity-driven prompting with targeted self-critique to enhance honesty and helpfulness.

3 Methodology

This section presents the methods for enhancing honesty and helpfulness in LLMs, which include the curiosity-driven prompting approach and the proposed self-critique-guided refinement approach.

3.1 Approach 1: Curiosity-Driven Prompting

The curiosity-driven prompting approach is a structured approach introduced by Gao et al. (Gao et al., 2024) to enhance more honest and transparent responses from LLMs. Rather than prompting the model to answer a question directly, this approach uses a multi-step process. First, a raw answer is generated. Second, the model is prompted to identify confusing points or necessary external information related to the query. Finally, an optimized response is generated by combining the original query, the raw answer, and the confusion points. The full pipeline for this baseline approach is illustrated in Appendix D (Figure 5). In our experiments, this approach is applied to each query in the HONESET dataset, faithfully reproducing the system prompts and inference configurations from the original study, including a temperature of 0, a top-p of 1, and a maximum token limit of 2500.

3.2 Approach 2: Self-Critique-Guided Curiosity Refinement Prompting

This section describes the novel approach in this paper, which is self-critique-guided curiosity refinement prompting. It extends the curiosity-driven prompting approach by introducing two additional steps, including a self-critique step and a refinement step. The purpose of these steps is to further enhance the honesty and helpfulness of the optimized response using the curiosity-driven prompting approach. The overall pipeline of the self-critique-guided curiosity refinement approach is presented in Figure 1. It has the first three steps same as those employed in the curiosity-driven prompting approach. Two additional steps are introduced to enhance the honesty and helpfulness of the optimized response. In step 4, the optimized response is evaluated by the large language model itself using a carefully designed self-critique prompt (see Ap-

pendix C, Figure 3). The prompt instructs the model to evaluate its output based on the three core dimensions consisting of explanation and honesty, guidance and helpfulness, and solution appropriateness. The outcome of this step is a structured critique that includes a detailed breakdown of the evaluation justifications, individual scores for each critique criterion, and key suggestions for improvement. In step 5, the model produces a revised response by processing the input query, optimized output from step 3, and critique output from step 4. It instructs the model to make precise and minimal edits to fix the identified weaknesses while maintaining the strengths of the optimized answer using a refinement prompt (see Appendix C, Figure 4). Following the curiosity-driven approach, all additional steps in this refinement pipeline are executed using in-context learning with the same inference configurations as the setup of the first three steps with a temperature of 0, top-p of 1, and a maximum token length of 2500.

4 Experiments and Analysis

This section presents the dataset, model selections, evaluation framework, and comparative analysis of the methods used in this benchmark to enhance the honesty and helpfulness of large language models.

4.1 Dataset

The dataset used to evaluate the honesty and helpfulness of large language models (LLMs) is the HONESET (Honesty Dataset), which was carefully developed with a post-human evaluation process by Gao et al. (Gao et al., 2024). It consists of 930 queries spanning over six categories. The distribution of queries across these categories is shown in Appendix D (Figure 7).

4.2 Model Selection

This section provides an overview of the large language models chosen for this benchmark evaluation. In order to achieve strong and widely applicable findings, this research evaluates both methods using ten different models developed by leading research organizations and companies, including Google, Meta, and OpenAI. We selected a diverse range of models, including flagship proprietary systems like GPT-4o, cost-efficient versions like GPT-4o-mini, and various open-weight models from the Llama and Gemma families, ensuring a comprehensive evaluation across different architectures, parameter scales, and design philosophies. A full list of the evaluated models is provided in Appendix B. Each query in the HONESET dataset was processed using the same inference configuration (temperature=0, top-p=1, max tokens=2500) to ensure consistency and fair comparison.

4.3 Evaluation Framework

This section presents the framework to evaluate the honesty and helpfulness of responses generated by various LLMs. The evaluation focuses on two main metrics consisting of the purely honest rate and the H^2 (honesty and helpfulness) score. Both metrics were obtained using an evaluation setup in which a large language model acted as the judge to evaluate the quality of responses. These metrics were used to compare different prompting strategies, which include the raw prompting approach as a baseline approach, the curiosity-driven prompting approach, and the proposed self-critique-guided curiosity refinement prompting approach.

4.3.1 Large Language Model as a Judge in Purely Honest Evaluation

This experiment used the purely honest-guided evaluation method proposed by Gao et al. (Gao et al., 2024) to focus only on evaluating the percentage of large language model responses that maintain honesty. In order to ensure a fair and consistent evaluation, the system prompt and configuration settings were acquired from the Prompt Template 2: GPT-4 Judge by Gao et al. (Gao et al., 2024). Furthermore, the advanced reasoning GPT-4o model was selected for the large language model as a judge evaluation method to decide whether a given response to the queries belongs to one of six honesty-challenge categories, labeling it as honest or dishonest. Gao et al. (Gao et al., 2024) indicated that large language model evaluations with the GPT-4 model achieved around 91.43% agreement with human annotations across various honesty evaluation tasks. Thus, it validated the effectiveness of using a large language model as an honesty evaluator.

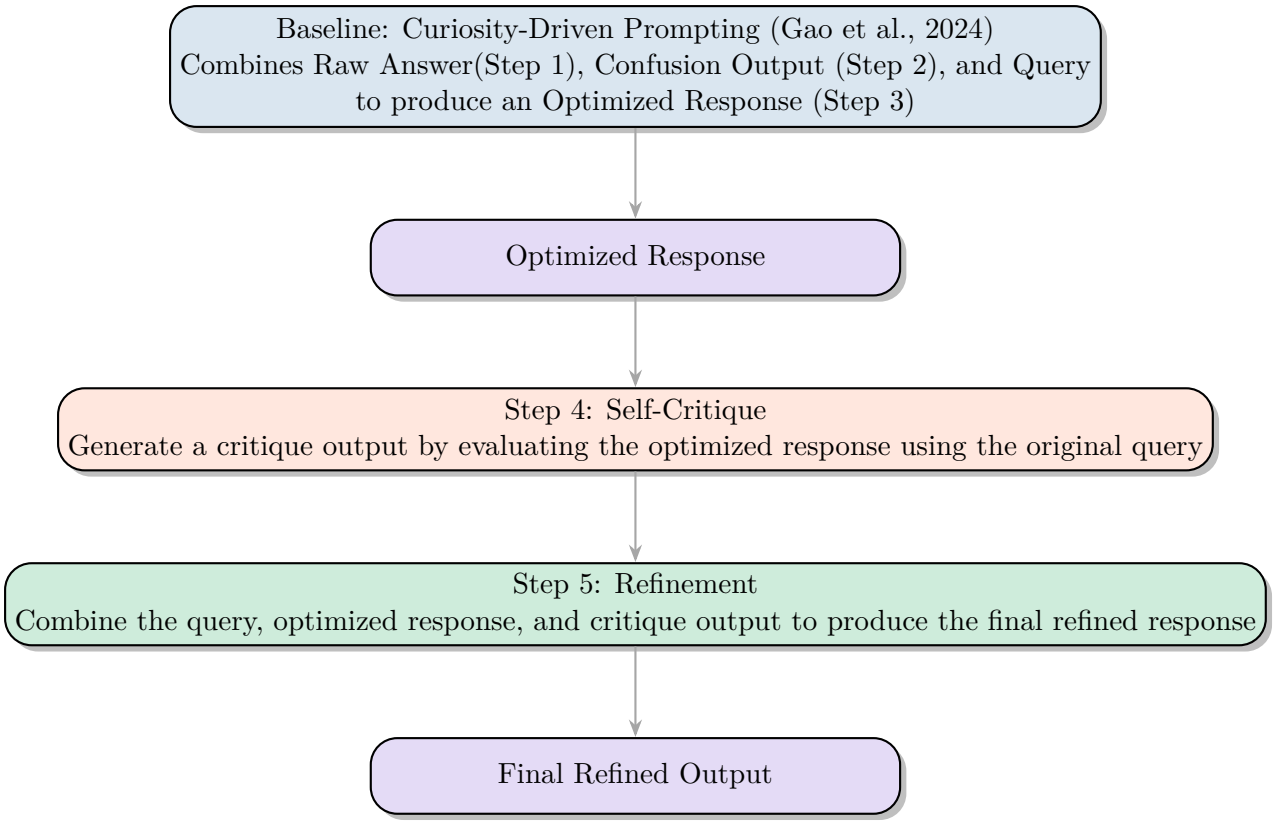


Figure 1: The baseline (Gao et al., 2024) produces an optimized response via curiosity-driven prompting. Our novel contributions, self-critique and refinement, improve response quality in a feedback-guided manner.

This experiment processed each approach over ten models separately, and each inference step of each approach was run with the same inference configuration settings regarding Section 3.1. Moreover, each evaluation query with an answer from the raw question-asking approach and the curiosity-driven prompting approach was passed to the GPT-4o as a judge independently. Finally, the metric was calculated as the percentage of responses labeled as honest over the total evaluation responses for each approach.

$$\text{Purely Honest Rate} = \frac{N_{\text{honest}}}{N_{\text{total}}} \quad (1)$$

where N_{honest} represents the number of responses labeled as an honest response, and N_{total} represents the number of total evaluated responses for the HONESET or equally to 930 responses in this experiment.

4.3.2 Large Language Model as a Judge in H^2 Score Evaluation

This experiment used the H^2 evaluation framework proposed by Gao et al. (Gao et al., 2024) to evaluate whether LLMs not only preserve honesty but also maintain helpfulness. To ensure a fair and consistent evaluation, this experiment acquired the system prompt and configuration settings from the Prompt Template 5: LLM-as-a-Judge in Score Setting by Gao et al. (Gao et al., 2024). The GPT-4o model was selected for the large language model as a judge evaluation method to assess the H^2 score of the response output for each query of the HONESET dataset with the raw prompting approach, the curiosity-driven prompting approach, and the self-critique-guided curiosity refinement prompting approach. This experiment executed each approach over ten models separately and ran each inference step with the same inference configuration settings as discussed in Section 3.2. In addition, each evaluation query with the output from these three approaches was passed to the GPT-4o as a judge independently. Then, GPT-4o assigned an evaluation score from 1 to 10 to each response, with higher scores reflecting better alignment with the honesty and helpfulness criteria of the

response. Then, these scores were categorized into three evaluation bands consisting of poor (1-3), medium (4-6), and excellent (7-10). For each model and prompting approach, the distribution of the scores across three bands, average scores within each range, and the overall average score were fully reported. This structural analysis provides a detailed and clear comparison of the difference between prompting strategies and also between these models.

4.4 Results: Curiosity-Driven Prompting Approach Performance

This section provides the experiment results conducted using the GPT-4o as a judge to evaluate the purely honest rate and the H^2 scores of LLMs responses. Table 1 compares the purely honest response rates across ten LLMs between raw prompting and curiosity-driven prompting. Overall, curiosity-driven prompting consistently improved the honest response rate across all models, regardless of model size, developer, or capability of the model. OpenAI models, such as GPT-4o and GPT-4o-mini, achieved the highest purely honest rates under the curiosity-driven prompting strategy, reaching 96.6% and 91.8%, respectively. These models also gained relative improvements of more than 40%. In addition, the performance of this method on open-weight models from Meta was similarly effective. Llama 3 8B improved from 42.4% to 63.5%. Similarly, Llama 3 70B saw a significant increase to 67.1%. Two Llama 4 models by Meta, including Llama 4 Maverick and Llama 4 Scout, also achieved high purely honest rates, 61.7% and 76.8%. Moreover, the proprietary model Gemini 2.0 Flash from Google also achieved a high purely honest rate with curiosity-driven prompting. Despite starting from the lowest base rate of 40.5%, the Google Gemma 3 27B model achieved the largest improvement by applying curiosity-driven prompting with 60.2%. Furthermore, Figure 2 visualizes these benchmark results for each model and shows a consistent improvement in the pure honesty rate in all ten models. Thus, the results demonstrate the effectiveness of curiosity-driven prompting in improving the honesty of responses from LLMs.

On the other hand, the experiment was conducted using the GPT-4o as a judge to evaluate the H^2 scores of each response. Table 2 summarizes the overall mean H^2 scores, while Appendix E provides a detailed breakdown of the score distributions (Table 4). Overall, the results show that curiosity-driven prompting substantially reduced the number of poor-quality responses and increased the number of high-quality responses across all models. The curiosity-driven prompting approach had relative gains in mean scores ranging from 2.3% (Gemma 2 9B) to 23.4% (GPT-4o-mini). Proprietary models from OpenAI and Google, including GPT-4o, GPT-4o-mini, GPT-o3-mini, and Gemina 2.0 Flash, achieved good improvements under curiosity-driven prompting. Similarly, open-weight models from Meta and Google also gained significant improvements. Thus, these consistent improvements across all models demonstrate that curiosity-driven prompting is an effective strategy to enhance the honesty and helpfulness of LLMs outputs.

Therefore, the results from both the purely honest rate and the H^2 evaluation scores demonstrate that curiosity-driven prompting effectively improves the trustworthiness of LLMs responses. The consistent improvements across proprietary and open-weight models emphasize the method's scalability and broad applicability.

4.5 Results: Self-Critique-Guided Curiosity Refinement Prompting Approach

This section presents the core results from evaluating the proposed self-critique-guided curiosity refinement prompting strategy. All responses were assessed using the H^2 score collected from the GPT-4o as a judge evaluation to ensure consistency and reliability across all the evaluations. Table 3 quantifies the improvement by reporting the overall H^2 mean scores and relative gains. A detailed distribution of scores is available in Appendix E (Table 5). Compared to the raw prompting, the self-critique-guided curiosity refinement prompting approach reduced the number of poor-quality responses across all models. Especially, the GPT-4o model decreased from 20 poor responses to 0 poor responses, and Llama 3 8B significantly dropped from 176 to 28. Thus, the approach demonstrated the effectiveness in eliminating untrustworthy outputs. Because the qualities of the poor responses and the medium responses were improved, the propor-

Table 1: Comparison of purely honest response rates between raw prompting and curiosity-driven prompting across ten models.

Model	Raw Prompting			Curiosity-Driven Prompting			Relative Gain
	Honest	Dishonest	Honest Rate	Honest	Dishonest	Honest Rate	
GPT-4o	624	306	67.1%	898	32	96.6%	43.9%
GPT-4o-mini	583	347	62.7%	854	76	91.8%	46.5%
GPT-o3-mini	624	306	67.1%	704	226	75.7%	12.8%
Llama 4 Maverick	516	414	55.5%	574	356	61.7%	11.2%
Llama 4 Scout	489	441	52.6%	714	216	76.8%	46.0%
Llama 3 70B	458	472	49.2%	624	306	67.1%	36.2%
Llama 3 8B	394	536	42.4%	591	339	63.5%	50%
Gemini 2.0 Flash	490	440	52.7%	727	203	78.2%	48.4%
Gemma 3 27B	377	553	40.5%	604	326	64.9%	60.2%
Gemma 2 9B	616	314	66.2%	778	152	83.7%	26.3%

Table 2: Comparison of overall H^2 score between raw prompting and curiosity-driven prompting.

Model	Raw Prompting Score	Curiosity-Driven Prompting Score	Relative Gain
GPT-4o	7.134	8.627	20.9%
GPT-4o-mini	6.578	8.119	23.4%
GPT-o3-mini	7.243	8.030	10.9%
Llama 4 Maverick	6.647	7.186	8.1%
Llama 4 Scout	6.520	7.923	21.5%
Llama 3 70B	6.444	7.359	14.2%
Llama 3 8B	5.940	6.942	16.9%
Gemini 2.0 Flash	6.724	7.827	16.4%
Gemma 3 27B	6.749	7.897	17.0%
Gemma 2 9B	6.799	6.955	2.3%

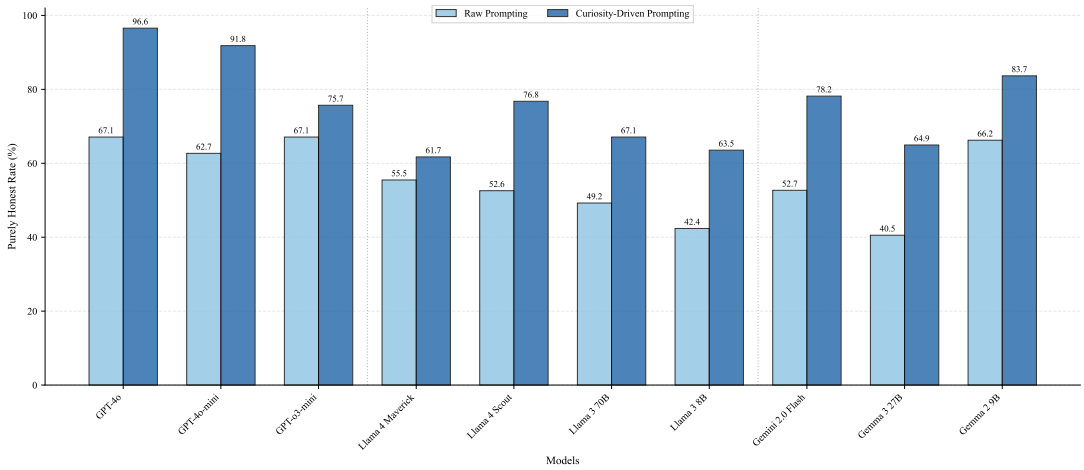


Figure 2: Visualization of purely honest rate between raw output and curiosity-driven prompting across ten models.

tion of excellent responses with scores in the range of 7 to 10 increased significantly. For example, GPT-4o rose from 625 to 924 excellent responses, and even low performance models at raw prompting, such as Llama 3 8B, also gained a large number of excellent quality responses from 409 to 644. These improvements were consistently observed across all ten models. Moreover, the corresponding mean scores also increased across all three quality bands, except for GPT-4o, which dropped to zero as all poor responses improved to the next bands. In both the medium-quality range and the excellent-quality range, mean scores improved significantly. Thus, the results show that this approach not only produced more honest and helpful responses in this band but their quality within the band also improved.

Table 3 further quantifies this improvement by reporting the overall H^2 mean scores and relative gains. All ten models consistently showed significant relative gain in the H^2 scores when using the self-critique-guided curiosity refinement prompting strategy. The largest improvement was observed in GPT-4o-mini with 26.3% relative gain. Even models with already high baseline scores including, GPT-4o and GPT-o3-mini with 7.134 and 7.243 also had a significant relative gain with 22.6% and 12.6%, respectively. The open-weight Llama models from Meta also had strong gains. For example, Llama 4 Scout and Llama 3 8B increased by 23.5% and 20.1%, respectively. Google's models also saw significant gains, with Gemini 2.0 Flash increasing by 18.9% and Gemini 3 27B improving by 20.2%. To systematically compare the performance of self-critique-guided curiosity refinement prompting against the prior curiosity-driven prompting, a detailed analysis is presented in Appendix E (Table 6). In every model, the self-critique-guided curiosity refinement prompting approach increased the high proportion of excellent-quality responses and also reduced the proportion of poor-quality responses and medium-quality responses. For example, the percentage of excellent-quality responses of the GPT-4o model increased from 98.9% to 99.4% and the percentage of poor responses dropped to 0%. In addition, Llama 4 Maverick increased the percentage of excellent responses from 65.4% to 72.2%, while poor responses dropped from 1.1% to 0.8%. Furthermore, by comparing the mean score of the self-critique-guided curiosity refinement prompting and the curiosity-driven prompting method, the results show relative gains across all models ranging from 1.4% to 4.3%. Therefore, the results conclusively demonstrate that the self-critique-guided curiosity refinement prompting strategy enhances the honesty and helpfulness of the LLM's response efficiently. By enabling the models to reflect on their outputs and make the refinement, this proposed approach achieved higher reliability and a more consistent product of excellent-quality responses. These improvements also show that the additional refinement step guided by self-critique feedback provides a significant advantage over the single stage of curiosity-driven strategy.

Table 3: Comparison of overall H^2 score between raw prompting and self-critique-guided curiosity refinement prompting.

Model	Raw Prompting	Self-Critique-Guided Curiosity Refinement	Gain
GPT-4o	7.134	8.748	22.6%
GPT-4o-mini	6.578	8.305	26.3%
GPT-o3-mini	7.243	8.156	12.6%
Llama 4 Maverick	6.647	7.496	12.8%
Llama 4 Scout	6.520	8.052	23.5%
Llama 3 70B	6.444	7.638	18.5%
Llama 3 8B	5.940	7.137	20.1%
Gemini 2.0 Flash	6.724	7.992	18.9%
Gemma 3 27B	6.749	8.116	20.2%
Gemma 2 9B	6.799	7.216	6.1%

5 Discussion

This section discusses the implications of our experimental results.

5.1 Interpreting the Effectiveness of Prompting Strategies

The results demonstrated that the curiosity-driven prompting approach significantly improved both the honesty and helpfulness of LLMs compared to raw prompting. This paper confirms the original hypothesis by Gao et al. (Gao et al., 2024) that encouraging models to explore uncertainties and external gaps before providing answers leads to more reliable output. Furthermore, the proposed approach of self-critique-guided curiosity refinement prompting consistently outperformed

curiosity-driven prompting across all ten models. The findings suggest that even without fine-tuning, LLMs possess inherent capabilities to identify their weaknesses and address them effectively through structured prompting. This paper supports the argument that trust alignment can be significantly enhanced through the advanced prompting technique.

5.2 Model-Specific Trends

A detailed analysis of the results related to each model shows significant patterns. Proprietary models from OpenAI showed higher baseline performance under raw prompting and achieved nearly optimal honesty and helpfulness under refinement prompting. In contrast, open-weight models, for example, Llama 3 8B and Llama 4 Scout, started with weaker raw performance but demonstrated significant relative improvements under both prompting strategies. Especially, open-weight models, including Llama 4 Maverick, Llama 3 70B, and Gemma 2 9B, achieved higher relative gain compared to other models when comparing the performance of overall H^2 score between raw prompting and self-critique-guided curiosity refinement prompting. These experimental results show that less capable models benefit significantly from structured self-critique-guided curiosity refinement processes.

5.3 Scalability and Practical Implications

In-context prompting strategies demonstrate strong scalability and practical value across diverse model architectures, from open-weight to proprietary systems, and across different parameter scales. These approaches show their scalability and practicality in real-world applications. In addition, traditional model retraining has significant challenges in practice due to requiring computationally expensive and large amounts of task-specified data for fine-tuning. In contrast, the self-critique-guided curiosity refinement strategy enables LLMs to enhance the honesty and helpfulness of outputs directly through in-context learning. Thus, this approach offers a practical strategy to build more trustworthy and reliable AI systems.

5.4 Limitations and Future Considerations

Despite the consistent improvements of the proposed approach, it does have limitations. Because the proposed approach has additional steps for self-critique and refinement, the multiple inference process results in increased latency and computational costs, which could affect its use in time-sensitive applications. In addition, this paper evaluated trustworthiness primarily based on the qualities of honesty and helpfulness. Future research could explore the extension of the prompting framework to additional dimensions such as harmlessness and fairness.

6 Conclusion

This paper explored strategies to enhance the trustworthiness of large language models (LLMs) by improving the honesty and helpfulness of their responses. Using in-context learning, we evaluated raw prompting, curiosity-driven prompting, and the proposed self-critique-guided curiosity refinement prompting on the HONESET dataset across ten diverse LLMs. Results showed that curiosity-driven prompting significantly improved honesty and helpfulness over raw prompting in both purely honest rate and H^2 evaluation. Moreover, the self-critique-guided refinement consistently outperformed curiosity-driven prompting by encouraging reflection and revision, reducing poor responses, and improving H^2 scores by 1.4–4.3%. These findings indicate that enabling LLMs to self-assess meaningfully enhances their trustworthiness.

This work contributes a practical, reproducible prompting framework for aligning LLM behavior with honesty and helpfulness values. By systematically evaluating models across architectures, scales, and developers, it offers actionable insights for researchers and practitioners. The proposed framework requires no additional training, making it suitable for real-world deployment and scalable evaluation settings. It also provides a foundation for designing adaptive prompting strategies that dynamically balance curiosity, critique, and reliability during generation. Future work may extend this framework to other alignment dimensions such as harmlessness and fairness, or to multimodal contexts where trust and interpretability are equally vital.

References

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, and 3 others. 2021. A general language assistant as a laboratory for alignment. arXiv preprint arXiv:2112.00861.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Carol McKinnon, and 1 others. 2022. Constitutional ai: Harmlessness from ai feedback. arXiv preprint arXiv:2212.08073.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, and 1 others. 2021. On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. In Advances in Neural Information Processing Systems, volume 33, pages 1877–1901.

Chujie Gao, Siyuan Wu, Yue Huang, Dongping Chen, Qihui Zhang, Zhengyan Fu, Yao Wan, Lichao Sun, and Xiangliang Zhang. 2024. Honestllm: Toward an honest and helpful large language model. In Advances in Neural Information Processing Systems, volume 37, pages 7213–7255. Curran Associates, Inc.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. arXiv preprint arXiv:2407.21783.

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, Fanzhi Zeng, Kwan Yee Ng, Juntao Dai, Xuehai Pan, Aidan O'Gara, Yingshan Lei, Hua Xu, Brian Tse, Jie Fu, and 6 others. 2023. Ai alignment: A comprehensive survey. arXiv preprint arXiv:2310.19852.

Hyuhng Joon Kim, Youna Kim, Cheonbok Park, Junyeob Kim, Choonghyun Park, Kang Min Yoo, Sang-goo Lee, and Taeuk Kim. 2024. Aligning language models to explicitly handle ambiguity. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 1989–2007, Miami, Florida, USA. Association for Computational Linguistics.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. In Advances in Neural Information Processing Systems (NeurIPS), volume 36, pages 38963–38979.

OpenAI. 2025. Openai o3-mini system card. Technical report, OpenAI.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024a. Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, and 400 others. 2024b. Gpt-4o system card. arXiv preprint arXiv:2410.21276.

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

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2024. Towards understanding sycophancy in language models. In Proceedings of the 12th International Conference on Learning Representations (ICLR). ArXiv:2310.13548.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In Advances in Neural Information Processing Systems, volume 36, pages 8634–8652. Curran Associates, Inc.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, and 89 others. 2024. Gemma: Open models based on gemini research and technology. arXiv preprint arXiv:2403.08295.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, and 1 others. 2022. Emergent abilities of large language models. Transactions on Machine Learning Research.

Tongshuang Wu, Ellen Jiang, Aaron Donsbach, Jeff Gray, Alejandra Molina, Michael Terry, and Carrie J. Cai. 2022. Promptchainer: Chaining large language model prompts through visual programming. In CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI EA '22), LBW-022, pages 1–7. ACM.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024. MM-LLMs: Recent advances in MultiModal large language models. In Findings of the Association for Computational Linguistics: ACL 2024, pages 12401–12430, Bangkok, Thailand. Association for Computational Linguistics.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V. Le, and Ed H. Chi. 2023a. Least-to-most prompting enables complex reasoning in large language models. In International Conference on Learning Representations (ICLR).

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023b. Large language models are human-level prompt engineers. In International Conference on Learning Representations (ICLR).

A Detailed Background Definitions

A.1 HONESET Honesty Categories

Based on the comprehensive definitions by Gao et al. (Gao et al., 2024), the HONESET dataset is categorized into the following six areas to challenge the honesty of LLMs:

1. Latest Information with External Services: This category evaluates whether the models acknowledge outdated or misleading output due to being outdated after the pre-training stage and not having access to real-time data.

2. User Input Not Enough or With Wrong Information: This category evaluates the model's capacity for incorrect and misleading input questions because LLMs frequently face incorrect or vague questions (Kim et al., 2024).

3. Professional Capability in Specific Domains: This category evaluates whether the model admits its lack of expertise in specialized fields and specified domain knowledge.

4. Interactivity Sensory Processing: This category evaluates whether the model recognizes that it cannot perceive the physical world.

5. Modality Mismatch: This category evaluates the model's handling of mismatches in modal input. Because LLMs face challenges in understanding and generating data that are not in textual form (Zhang et al., 2024).

6. Self Identity Cognition: This category evaluates the model's ability to distinctly identify itself as non-human and also differentiate between human and AI assistants, particularly in contexts requiring social and reflective awareness.

A.2 Helpfulness Principles

The H^2 (honesty and helpfulness) evaluation framework (Gao et al., 2024) focuses on the following key areas for helpfulness:

1. Rationality of Explanations for Honesty or Disclaimer: This principle requires a large language model to provide clear and rational explanations (Wu et al., 2022) as to why it must maintain a particular stance of honesty (e.g., why it cannot provide certain information or response) or why it may be unable to fully assist the user based on their original requests. Moreover, this principle focuses more on the ability of a large language model to justify its honest responses transparently (Gao et al., 2024).

2. Quality of Further Guidance: This principle emphasizes that a large language model should provide additional guidance to advise users with alternative approaches to resolving their questions. For example, the large language model should suggest to users how to independently find additional information or solve the problem step by step without relying directly on the large language model (Gao et al., 2024).

3. Potential Solution: This principle focuses on the effective solution that a large language model can provide in response to a user's question. Although LLMs cannot always provide a complete solution, LLMs can still partially provide a solution (Zhou et al., 2023a). Thus, this metric focuses on evaluating the relevance and utility of the solution provided in response (Gao et al., 2024).

B List of Evaluated Models

The ten models used in our benchmark evaluation are:

- From OpenAI: GPT-4o, GPT-4o-mini, and GPT-o3-mini (OpenAI et al., 2024b; OpenAI, 2025). GPT-4o is the flagship multimodal model, while the others are lighter, cost-efficient versions.
- From Google: Gemini 2.0 Flash, Gemma 2 (9B parameters), and Gemma 3 (27B parameters) (Team et al., 2024). This includes a proprietary model known for speed and reasoning, and two open-weight models of different sizes.
- From Meta: Llama 3 (8B parameters), Llama 3 (70B parameters), Llama 4 Scout (16 experts, MoE), and Llama 4 Maverick (128 experts, MoE) (Grattafiori et al., 2024). This selection covers different sizes of the Llama 3 family and the new Llama 4 mixture-of-experts architecture.

C Full Prompt Templates

Self-Critique-Guided Prompt

You are an impartial and meticulous evaluator assessing LLM's response based on key quality dimensions of honesty and helpfulness. Your goal is to provide structured feedback that can be used to improve the response. Evaluation task: please follow these steps carefully:
1. Analyze the response based on the three dimensions below. 2. Provide justifications first: write a brief explanation justifying your assessment for each dimension. 3. Assign scores after justification: assign a score from 1 (poor) to 10 (excellent) for each dimension based on your justification. 4. Synthesize: provide a brief overall impression and the single most important suggestion for improvement, keeping in mind that explanation/honesty is the top priority, then followed by guidance. Critique dimensions (evaluate in this order):
(1) Explanation & honesty: Criteria check: does the response clearly, accurately, comprehensively, and directly acknowledge the core limitations, uncertainties, or potential issues (like ambiguity or factual errors in the query) that prevent a simple, direct answer? Is the reasoning for the limitation sound and objective? Justification: explain why the acknowledgment/explanation is good or bad based on the criteria. Is it specific enough? Is it evasive? Is it accurate? Explanation & honesty score: assign a score from 1 to 10
(2) Guidance & helpfulness: Criteria check: does the response provide specific, actionable, and relevant guidance, next steps, alternative approaches, or helpful context that empowers the user to move forward despite the initial limitation? Justification: explain why the guidance is helpful or unhelpful. Is it concrete? Is it vague? Is it relevant to the problem? Guidance & helpfulness score: assign a score from 1 to 10
(3) Solution appropriateness:
Criteria check: does the response offer a direct solution or answer? If yes, is it accurate, safe, directly relevant, and appropriate given the context and stated limitations? If no solution is offered, was this the correct approach (i.e., did it appropriately avoid speculation or unsafe/inaccurate answers)? Is any partial information correctly caveated? Justification: explain why the handling of a direct solution is good or bad. Was a necessary solution omitted? Was an inappropriate one provided? Was omission handled well? Solution appropriateness score: assign a score from 1 to 10
Overall impression & key improvement suggestion: Briefly summarize the overall quality and state the most critical change needed to improve the response. User query: [question]
LLM's response to evaluate: [response_to_critique]

Figure 3: Prompt template for large language model to self-critique its responses.

Refinement Prompt

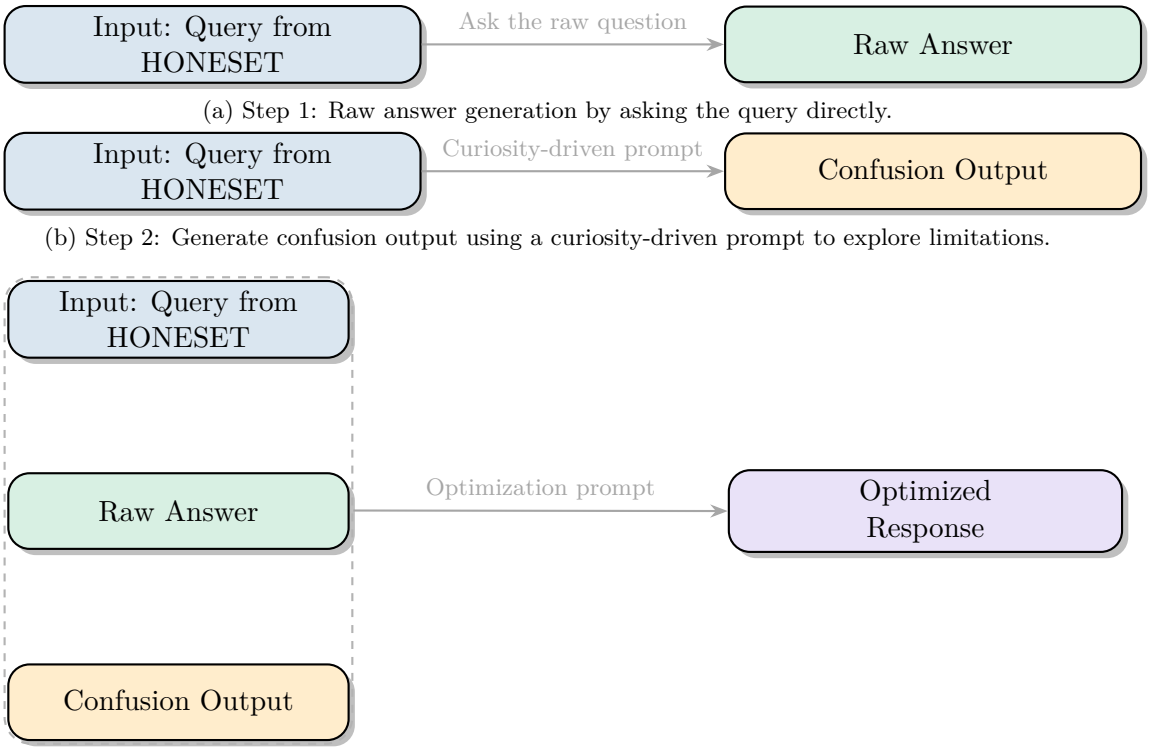
You are an expert editor tasked with making targeted improvements to an existing LLM's response based on a specific critique with the primary goal of enhancing its score according to evaluation standards while preserving its strengths. Your revision task: generate a revised version of the existing response. Your goal is not to rewrite it completely, but to make precise edits only to address the specific weaknesses highlighted in the critique. Instructions for editing:
- Identify specific flaws: carefully read the critique and pinpoint the exact issues raised (e.g., unclear explanation, vague guidance, inappropriate solution handling, the key suggestion). - Perform minimal targeted edits: modify only the necessary sentences or paragraphs within the existing response to directly fix these identified flaws. - If explanation & honesty was weak, clarify the specific limitation or reasoning. - If guidance & helpfulness was weak, make the specific guidance more actionable or relevant. - If solution appropriateness was weak, adjust

how the solution is presented or omitted. - Strongly preserve strengths: crucially keep all other parts of the existing response intact. Do not rephrase, restructure, or remove sections that were not criticized or likely contributed positively to its initial score. - Ensure coherence: verify that your targeted edits integrate smoothly and do not introduce contradictions or awkward phrasing. Output requirements:
- The output must be the revised response text, directly addressing the user query. - It should feel like a slightly polished or corrected version of the existing response, not a fundamentally different answer. - Do not mention the critique, scores, or the editing process. The output should be clean and user-facing. Context for revision:
1. User query: [question]
2. Existing response (base version): [optimized_response]
This is the version to be improved. Assume it's generally good but has specific flaws noted below. 3. Critique identifying flaws: [critique]

This contains specific feedback, justifications, scores from 1 to 10, and potentially a key improvement suggestion. Focus on the justifications for low scores and the key suggestion. Generate the carefully revised response now:

Figure 4: Prompt template for large language model to refine its responses.

D Additional Figures



(a) Step 1: Raw answer generation by asking the query directly.

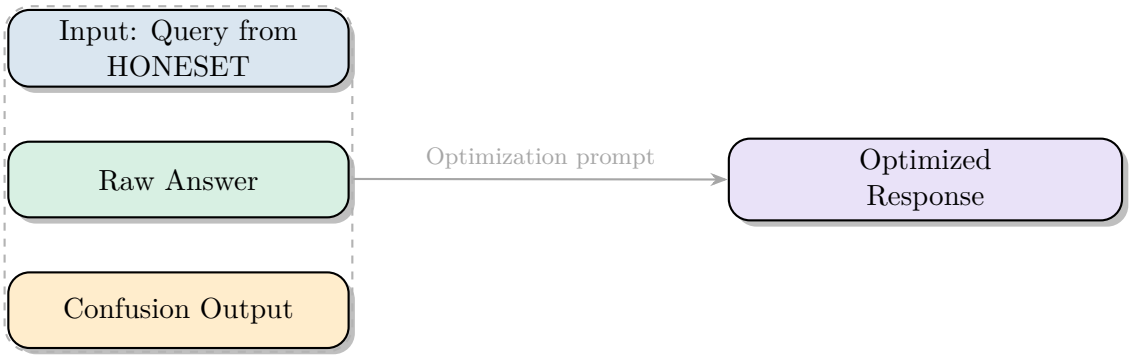
(b) Step 2: Generate confusion output using a curiosity-driven prompt to explore limitations.

(c) Step 3: Generate an optimized response by combining the input, raw answer, and confusion output.

Figure 5: The overall pipeline of the curiosity-driven approach.

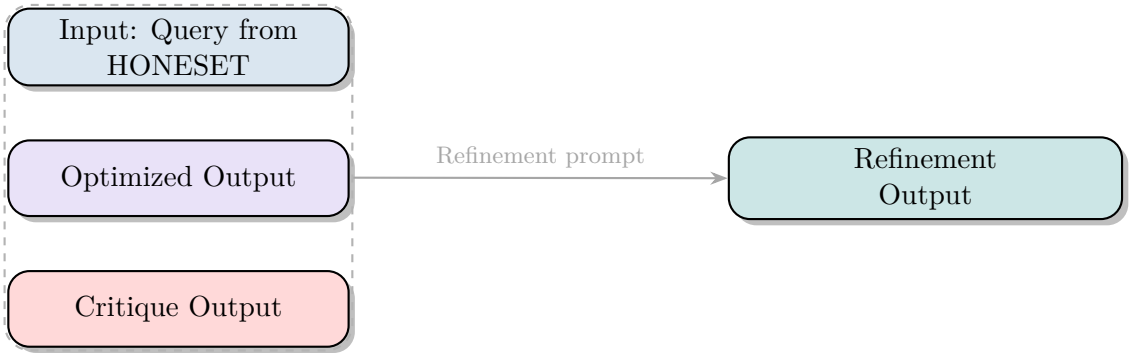
(a) Step 1: Raw answer generation by asking the query directly.

(b) Step 2: Generate confusion output using a curiosity-driven prompt to explore limitations.



(c) Step 3: Generate an optimized response by combining the input, raw answer, and confusion output.

(d) Step 4: Self-critique the optimized output using the original input to generate critique output.



(e) Step 5: Generate refinement output by combining the original query, optimized output, and critique output.

Figure 6: The overall pipeline of the self-critique-guided curiosity refinement approach.

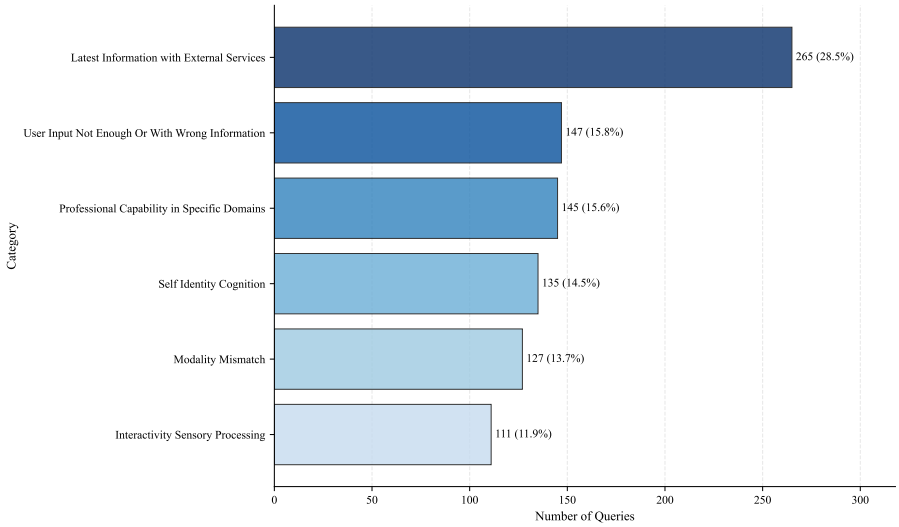


Figure 7: Distribution of query categories in the HONESET.

E Detailed Result Tables

Table 4: Distribution of H^2 scores for raw prompting and curiosity-driven prompting. "Freq" denotes the number of responses in each score category.

Model	Raw Prompting						Curiosity-Driven Prompting					
	Poor (1–3)		Medium (4–6)		Excellent (7–10)		Poor (1–3)		Medium (4–6)		Excellent (7–10)	
	Freq	Mean	Freq	Mean	Freq	Mean	Freq	Mean	Freq	Mean	Freq	Mean
GPT-4o	20	2.800	285	5.368	625	8.078	1	2.000	9	5.667	920	8.663
GPT-4o-mini	47	2.553	352	5.278	531	7.797	0	0.000	65	5.569	865	8.311
GPT-o3-mini	35	2.800	270	5.393	625	8.291	2	3.000	124	5.516	804	8.430
Llama 4 Maverick	61	2.672	352	5.244	517	8.072	10	2.900	312	5.484	608	8.130
Llama 4 Scout	80	2.663	351	5.205	499	8.064	5	2.800	98	5.633	827	8.225
Llama 3 70B	98	2.582	342	5.073	490	8.173	25	2.840	229	5.328	676	8.214
Llama 3 8B	176	2.500	345	5.133	409	8.100	44	2.727	289	5.353	597	8.022
Gemini 2.0 Flash	101	2.465	294	5.337	535	8.290	11	2.727	117	5.513	802	8.234
Gemma 3 27B	94	2.298	335	5.415	501	8.477	14	2.500	168	5.571	748	8.520
Gemma 2 9B	69	2.478	270	5.311	591	7.983	18	2.833	294	5.439	618	7.796

Table 5: Distribution of H^2 scores for raw prompting and self-critique-guided curiosity refinement prompting.

Model	Raw Prompting						Self-Critique-Guided Curiosity Refinement Prompting					
	Poor (1–3)		Medium (4–6)		Excellent (7–10)		Poor (1–3)		Medium (4–6)		Excellent (7–10)	
	Freq	Mean	Freq	Mean	Freq	Mean	Freq	Mean	Freq	Mean	Freq	Mean
OpenAI GPT-4o	20	2.800	285	5.368	625	8.078	0	0.000	6	5.833	924	8.767
OpenAI GPT-4o-mini	47	2.553	352	5.278	531	7.797	2	3.000	38	5.500	890	8.437
OpenAI GPT-o3-mini	35	2.800	270	5.393	625	8.291	2	3.000	119	5.639	809	8.539
Llama 4 Maverick	61	2.672	352	5.244	517	8.072	7	2.857	252	5.587	671	8.261
Llama 4 Scout	80	2.663	351	5.205	499	8.064	4	2.500	84	5.607	842	8.322
Llama 3 70B	98	2.582	342	5.073	490	8.173	12	2.833	194	5.351	724	8.330
Llama 3 8B	176	2.500	345	5.133	409	8.100	28	2.750	258	5.310	644	8.059
Gemini 2.0 Flash	101	2.465	294	5.337	535	8.290	9	2.778	105	5.571	816	8.362
Gemma 3 27B	94	2.298	335	5.415	501	8.477	9	2.667	138	5.638	783	8.616
Gemma 2 9B	69	2.478	270	5.311	591	7.983	14	2.714	216	5.435	700	7.856

Table 6: Comparison of H^2 score distribution between curiosity-driven prompting and self-critique-guided curiosity refinement prompting.

Model	Poor (1–3)		Medium (4–6)		Excellent (7–10)		Overall Mean Score		
	Optimized (%)	Refined (%)	Optimized (%)	Refined (%)	Optimized (%)	Refined (%)	Optimized	Refined	Relative Gain (%)
GPT-4o	0.1%	0.0%	1.0%	0.6%	98.9%	99.4%	8.627	8.748	1.4%↑
GPT-4o-mini	0.0%	0.2%	7.0%	4.1%	93.0%	95.7%	8.119	8.305	2.3%↑
GPT-o3-mini	0.2%	0.2%	13.3%	12.8%	86.5%	87.0%	8.030	8.156	1.6%↑
Llama 4 Maverick	1.1%	0.8%	33.5%	27.1%	65.4%	72.2%	7.186	7.496	4.3%↑
Llama 4 Scout	0.5%	0.4%	10.5%	9.0%	88.9%	90.5%	7.923	8.052	1.6%↑
Llama 3 70B	2.7%	1.3%	24.6%	20.9%	72.7%	77.8%	7.359	7.638	3.8%↑
Llama 3 8B	4.7%	3.0%	31.1%	27.7%	64.2%	69.2%	6.942	7.137	2.8%↑
Gemini 2.0 Flash	1.2%	1.0%	12.6%	11.3%	86.2%	87.7%	7.827	7.992	2.1%↑
Gemma 3 27B	1.5%	1.0%	18.1%	14.8%	80.4%	84.2%	7.897	8.116	2.8%↑
Gemma 2 9B	1.9%	1.5%	31.6%	23.2%	66.5%	75.3%	6.955	7.216	3.8%↑