What Would You Ask When You First Saw $a^2 + b^2 = c^2$? Evaluating LLM on Curiosity-Driven Question Generation

Shashidhar Reddy Javaji[1], **Zining Zhu**[1]

[1]Stevens Institute of Technology, Hoboken, NJ, USA
{sjavaji, zzhu41}@stevens.edu

Abstract

Large language models (LLMs) are increasingly widely used as critical components of knowledge retrieval systems and agentic systems. These systems can benefit from knowledge-seeking capabilities of LLMs, in other words, curiosity. However, this capability has not been evaluated quantitatively. Toward bridging this gap, we propose an evaluation framework, CDQG (Curiosity-Driven Question Generation). The CDQG task prompts LLMs to generate questions about a statement introducing scientific knowledge, simulating a curious person when facing the statement for the first time. The CDQG dataset contains 1,988 statements including physics, chemistry, and mathematics with distinct levels of difficulty, general knowledge statements, and intentionally erroneous statements. We score the qualities of the questions generated by LLMs along multiple dimensions. These scores are validated by rigorous controlled ablation studies and human evaluations. While large models like GPT-4 and Mistral 8x7b can generate highly coherent and relevant questions, the smaller Phi-2 model is equally or more effective. This indicates that size does not solely determine a model's knowledge acquisition potential. CDQG quantifies a critical model capability, and opens up research opportunities for developing future knowledge retrieval systems driven by LLMs.

Introduction

Nowadays, large language models (LLMs) trained on internet-scale datasets are capable of storing and processing massive amounts of knowledge. LLMs are used as critical components of knowledge retrieval and processing systems, and the performance of these systems is related to the LLMs' capability to seek knowledge (Krishna et al. 2024; Huang and Huang 2024; Gao et al. 2024).

However, to the best of our knowledge, this capability has not been evaluated quantitatively. Previous works in the literature assessed the capability to store knowledge (Liu et al. 2024a; Petroni et al. 2019), to be aware of the knowledge (Suzgun et al. 2024; Ferrando et al. 2024) and the capability to use knowledge (Zhu et al. 2024). We take an alternate perspective, assessing the capability of LLMs to *seek* knowledge.

Our setup is inspired by how humans seek knowledge: asking questions out of curiosity. Questioning is a key cognitive skill that underpins learning and knowledge acquisition. By asking questions, humans seek to understand the world around them, explore how things work, and challenge existing beliefs. This act of inquiry not only helps humans learn new information but also sharpens their thinking, promotes critical analysis, and drives innovation. Effective questioning fuels intellectual growth by sparking curiosity, encouraging deeper exploration of subjects, and improving comprehension (Acar, Berthiaume, and Johnson 2023). In education, questioning is closely linked to higher-level thinking skills like analysis, synthesis, and evaluation (Kurdi et al. 2020). The complexity and depth of questions asked often reflect a person's grasp and understanding of a topic (Kotov and Zhai 2010).

Questions also play a crucial role in reasoning (Zelikman et al. 2024; Hao et al. 2023) since asking insightful questions requires logical thinking, clarifying assumptions, identifying knowledge gaps, and exploring alternative viewpoints (Lucas et al. 2024). OpenAI's o1 model uses its own "chain of thought" to engage in structured reasoning (OpenAI 2024). Thoughtful questions are essential for thorough and logical reasoning (Ashok Kumar et al. 2023). Questioning is equally important for fact-checking. Good questions guide the verification process by identifying gaps, biases, and inconsistencies in the information (Li et al. 2017). Questions like "Does this agree with other sources?" or "Is this consistent with historical data?" lead to careful checking of facts and encourage cross-referencing across multiple sources. Effective fact-checking requires context and nuance, and good questions can help reveal false or misleading information. Besides reasoning and fact-checking, questioning plays a major role in many other areas (Masterman et al. 2024), like encouraging creativity (Wang et al. 2024a), stimulating discussion, and driving innovation (Si, Yang, and Hashimoto 2024; Ghafarollahi and Buehler 2024). Thoughtful questions can open doors to new ideas and solutions.

Inspired by human questioning, we propose a framework, CDQG, that evaluates the LLMs' potential for discovering new knowledge. This framework is centered around a curiosity-driven question generation (CDQG) task, where a model is prompted to imagine itself as a human encountering a new statement for the first time, eliciting the most immediate questions that would arise. The questions are then scored along three metrics — relevance, coherence, and di-

versity — scores with roots in the literature of psychology (Zhao, Strube, and Eger 2023). We use state-of-the-art LLMs to compute these scores. The scores are validated by human judgment as well as rigorous ablation studies. Recent work by (Ke et al. 2024) explores how foundation models can independently gather information, highlighting parallel advancements in our field as we examine LLMs' curiosity-driven questioning.

We collect the CDQG dataset. The CDQG dataset contains 1,101 statements in physics, chemistry, and math, spanning across distinct levels of difficulty. Additionally, the CDQG dataset includes a section of 300 general knowledge statements and a special section of erroneous statements. CDQG challenges the models' critical inquiry skills and facilitates rigorous and generalizable evaluation.

Using the CDQG framework, we evaluate pretrained language models of varying sizes, ranging from smaller ones like Phi-2 (Mojan Javaheripi 2023) to larger models like GPT-4 (OpenAI et al. 2024) and Gemini (Gemini Team, Anil, and et al. 2024). While the larger models score high in coherence and relevance, the smaller Phi-2 model scores comparably well (or even better), indicating that the size might not be the only factor for the knowledge acquisition potential.

Our contributions can be summarized as follows:

- We introduce the CDQG framework, a novel approach for evaluating the ability of LLMs to generate questions given new information.
- We establish and validate a set of evaluation metrics, to systematically measure the depth and comprehensiveness of the questions generated by the LLMs.
- We compile the CDQG dataset, which includes varied and challenging content to test the questioning capabilities of LLMs.
- We conduct extensive testing with state-of-the-art LLMs to demonstrate the effectiveness of our framework through an ablation study.
- We highlight the practical applications of our findings in educational technology and AI-driven content creation.

To our knowledge, we are the first to introduce an evaluation framework assessing LLMs' questioning abilities based on knowledge statements. Our research encourages questioning-based evaluations to deepen the understanding of LLMs as critical components of knowledge-processing systems.

Related Works

Question Generation

Question generation has long been recognized as a critical task in education, with numerous studies underscoring its significance (Elkins et al. 2023; Kurdi et al. 2020). The evolution of this field has seen a progression from early rule-based question generation systems (Yao et al. 2022) to more sophisticated methods employing transformer-based models. Most recently, the application of LLMs represents the latest advancement in this area. The transition from rule-based systems to transformers, and ultimately to LLMs, highlights a shift towards utilizing deep learning techniques that better mimic human-like questioning abilities. This evolution enhances the relevance and quality of the generated questions, and also opens new possibilities for dynamic interactions within educational software (Abbasiantaeb et al. 2024) and conversation systems (Wang et al. 2024b).

Evaluation of Generative Models

In evaluating text generation from LLMs, recent methodologies have expanded beyond traditional metrics to include multifaceted approaches that align more closely with human judgment. GPTScore (Fu et al. 2023) and UniEval (Leiter et al. 2023) utilize the natural language understanding capabilities of LLMs to tailor evaluations to specific criteria, with GPTScore focusing on customized fluency and UniEval using a Boolean question-answering format for multiple quality dimensions. Similarly, CheckEval (Lee et al. 2024) employs a structured checklist to enhance reliability, while X-Eval (Liu et al. 2024b) dynamically selects evaluation aspects, enhancing adaptability and depth. Further enriching these approaches are frameworks like the zero-shot comparative methodology (Liusie, Manakul, and Gales 2024), which performs direct quality judgments, and the Unified Framework (Zhong et al. 2022), which combines traditional and specialized models for the assessment. PlanBench (Valmeekam et al. 2023) explores LLMs' reasoning through various planning tasks, while TIGERSCORE (Jiang et al. 2023) emphasizes explainability in evaluations. These are complemented by strategies that assess LLMs' ability to follow complex instructions (He et al. 2024) and a composite metric system that aggregates individual scores for a holistic view (Verga et al. 2024), enhancing the development and refinement of LLMs across different applications. However, these methodologies primarily center on how LLMs answer questions and perform predefined tasks, with little exploration into how effectively these models can generate meaningful questions themselves. Different from prior works, we focus explicitly on the questioning abilities of LLMs, introducing a new assessment dimension.

Prompt Engineering

Recent advancements in LLM evaluation have focused on optimizing prompting techniques to align more closely with human judgment. Studies show that LLM evaluations are more reproducible than human evaluations and effectively use a five-point Likert scale (Chiang and Lee 2023). The G-EVAL framework improves evaluation accuracy by leveraging GPT-4 with chain-of-thought prompting (Liu et al. 2023a). Emphasizing prompt engineering, research demonstrates that well-crafted instructions and score aggregation significantly enhance LLM performance (Baswani, Mukherjee, and Shrivastava 2023). Additionally, smaller LLMs, guided by effective prompts, can match larger models' evaluation performance, highlighting the importance of prompt design (Kotonya et al. 2023). Moreover, reference-free evaluation methods show that LLMs can assess text quality through comparative judgments and rationales (Chen et al. 2023). These advancements informed our prompt engineering strategies.

LLMs for Evaluation

Recent studies highlight LLMs' potential to achieve human-level assessment quality in various tasks (Gilardi, Alizadeh,

and Kubli 2023; Huang et al. 2024). The GEMBA framework, for instance, showcases the effectiveness of LLMs in reference-free machine translation evaluation (Kocmi and Federmann 2023), while FrugalScore offers a streamlined approach by combining LLM-based metrics with lightweight models for efficient assessment (Kamal Eddine et al. 2022). Literature such as "Is ChatGPT a Good NLG Evaluator?" underscores ChatGPT's strong alignment with human judgments across NLG tasks (Wang et al. 2023). AUTOCALIBRATE enhances LLM-human alignment by iteratively refining evaluation criteria with human feedback (Liu et al. 2023b). Additionally, LLMs have proven effective in delivering relevance judgments with natural language explanations (Faggioli et al. 2023). Evaluations in machine translation and chatbot conversations show LLMs closely align with human ratings (Zheng et al. 2023). Instruction tuning has been shown to improve the correlation between LLM evaluations and human judgments (Xiong et al. 2024), while the development of explainable metrics emphasize the importance of transparency in LLM assessments (Leiter et al. 2024). While numerous methods have been introduced, many still have limitations and lack full robustness. In this paper, we propose a new framework that includes incremental noise addition to demonstrate the robustness of LLM evaluation without relying on human evaluation references.

CDQG framework

As summarized by Fig 1, this section describes the CDQG framework. CDQG specifically prompts models to ask questions elicited from intrinsic curiosity. CDQG then systematically evaluates these models across three critical performance metrics.

CDQG task

The CDQG task starts with a statement sampled from the CDQG dataset (which we'll explain in detail in Section). A prompt is constructed to accommodate the distinct instructional formats of multiple models. The prompting approach is centered around personification, where we ask each model to conceptualize itself as a human. This hypothetical human, encountering a statement for the first time and devoid of prior knowledge, is prompted to generate the top five questions that would instinctively arise. This allows us to elicit the models' inquisitive capabilities in a novel and controlled environment. The full prompt template is listed in appendix .

CDQG evaluation

The questions generated by the models are scored via a multi-dimensional evaluation procedure. The following three scores are computed:

Relevance: Relevance assesses how directly each question pertains to the specific details, elements, or concepts presented in the statement or scenario. The relevance criterion checks if questions aim to clarify, expand upon, or directly explore the content of the statement, focusing on the immediate context rather than the topics not directly introduced by the statement.

Subject	Split				Total
	Basic	Intermediate	Advanced	Wrong	
Physics	100	101	100	225	526
Chemistry	161	161	161	181	664
Math	108	108	101	181	498
General					300
Total	369	370	362	587	1,988

Table 1: Splits and sizes of the CDQG dataset.

Coherence: Coherence assesses how logically the questions within each set connect to one another and whether they form a coherent line of inquiry that would logically progress a beginner's understanding of the topic. The coherence criterion checks if the sequence of questions or their thematic connection facilitates a structured exploration of the statement.

Diversity: Finally, diversity determines the range of aspects covered by the questions in relation to the statement, ensuring that each question brings a new dimension or perspective to understanding the statement. The diversity criterion checks if while maintaining direct relevance, the questions collectively offer a broad exploration of the topic, including but not limited to definitions, implications, applications, or theoretical underpinnings.

We use LLMs to score the generations on the aforementioned three dimensions, following the recent LLM-as-a-judge trend (Li et al. 2024). We select three large language models, GPT-3.5 Turbo, Mistral 8x7b, and Gemini, based on their accessibility, state-of-the-art performance characteristics, and diverse architectural approaches. For each specified metric, we prompt the LLM judge to generate a score in 5-point Likert scale and the corresponding justifications.

Then, we use Gemini as a "metareviewer" that summarizes the three evaluations (score with justification) into one final score, with a brief sentence as metareview. While the metareview sentence is not directly used to compare the models, it helps the Gemini models to provide a fair summary score.

To further ensure the validity of this evaluation protocol, we set up two validation experiments: an automatic noise-injection experiment and a human validation experiment. The details of the two validation studies are described in Section .

CQQG dataset

The CDQG dataset facilitates the CDQG evaluation framework. We leverage GPT-4's generative capabilities under human oversight to assemble the dataset incrementally (Xu et al. 2023), selecting statements that span diverse topics and complexity levels. Table 1 shows a breakdown of the dataset's splits and their corresponding sizes. We consider the following desiderata when constructing the CDQG dataset.

Multiple subjects We include three subjects: chemistry, physics, and mathematics, to encompass a range of academic scenarios that an LLM may be useful. We additionally include general statements reflecting everyday life scenarios to broaden the coverage of the dataset.

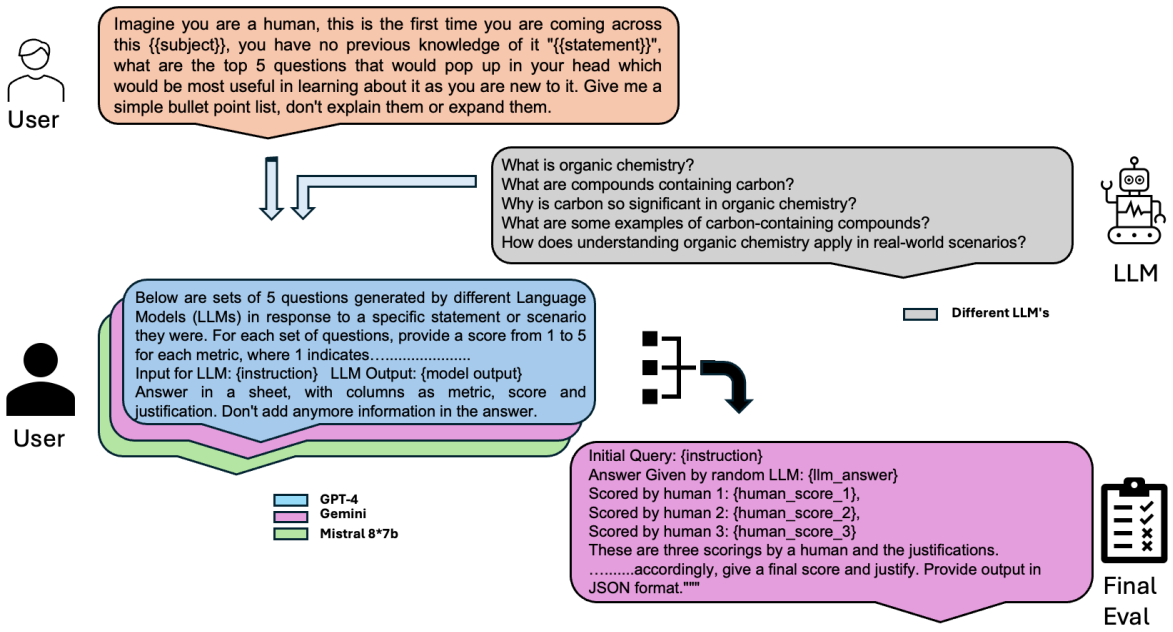

Figure 1: The CDQG framework. The top half shows the CDQG task, and the lower half shows the evaluation method of the generated questions.

Distinct difficulty levels For each of the academic subjects, we split the dataset into distinct difficulty levels, allowing stratified assessments of the LLMs' knowledge-seeking behavior regarding the statements with distinct levels of difficulty. Each level contains approximately the same number of statements to ensure a balanced distribution.

Wrong statements A unique feature of our dataset is the inclusion of these intentionally erroneous statements such as "The sum of 5 and 6 is 55", which probe the models' critical questioning abilities. These wrong statements span all three scientific domains, created by subtly modifying accurate statements. This subset tests whether models can identify and question statement veracity and logical consistency, particularly when treating the information as novel. We hypothesize that if a model operates as though it possesses prior knowledge, it will naturally question statement legitimacy. This dataset component serves as a critical test for evaluating models' depth of inquiry and their ability to critically engage with new information.

Models

We examine models ranging from a wide array of sizes: Llama 7b, Llama 13b, Llama 70b (Touvron et al. 2023), Mistral 8x7b (Jiang et al. 2024), Microsoft Phi-2 2.7b, Gemini, GPT 3.5 Turbo (Brown et al. 2020), and GPT-4. Our selection is based on practical considerations such as open-source availability and ease of access through APIs. Mistral's architecture, designed for handling complex queries, and Phi-2.7b's specialization in Q&A, make them well-suited for CDQG. By choosing models with varying architectures and parameter sizes, we ensure a broad comparison of model capabilities while maintaining accessibility and relevance to the task. The Gemini, GPT-3.5 turbo, and GPT-4 models are accessed using available APIs, and the other models (Llama-2 7b, Llama-2 13b, Llama-2 70b, Mistral 8x7b, Microsoft Phi-2) are accessed using the open-source weights, downloaded from Hugging face.

Results

Fig 2, Fig 2 and Fig 3 illustrate our main results, with the rest in the Appendix.

Performance by model

GPT-4: Dominates in almost all metrics and subjects, especially in advanced tasks. This superior performance can be attributed to its extensive training on a diverse dataset, which equips it with a broad knowledge base and sophisticated reasoning capabilities.

Mistral 8x7b: Frequently matches or exceeds GPT-4, showing exceptional strength in Chemistry and Maths. Its use of a sparse mixture-of-experts architecture allows it to efficiently manage specific query types, demonstrating the benefits of mixture-of-experts architecture.

Phi-2: Phi-2's performance is particularly noteworthy. Despite its smaller scale of 2.7 billion parameters, Phi-2 consistently produces relevant and coherent questions at basic to intermediate task levels. This model benefits significantly from high-quality, curated training data that emphasizes "textbook-quality" content (Mojan Javaheripi 2023), enhancing its capability in logical reasoning and common-sense understanding.

	Relevance		Coherence		Diversity	
Dataset	Highest	Lowest	Highest	Lowest	Highest	Lowest
Physics						
- Basic	GPT-4	Llama2-7b	Phi-2	Llama2-70b	GPT-4	Llama2-7b
- Intermediate	GPT-4	Gemini	GPT-4	Llama2-7b	Phi-2	Gemini
- Advanced	GPT-4	Llama2-70b	GPT-4	Llama2-7b	Gemini	Llama2-7b
- Wrong	GPT-3.5	Mistral 8x7b	GPT-3.5	Llama2-70b	GPT-3.5	GPT-4
Chemistry						
- Basic	GPT-4	Llama2-7b	GPT-4	Llama2-7b	Mistral 8x7b	Llama2-7b
- Intermediate	GPT-4	Llama2-7b	Mistral 8x7b	Llama2-7b	Mistral 8x7b	Llama2-70b
- Advanced	Mistral 8x7b	Llama2-7b	GPT-4	Llama2-7b	Mistral 8x7b	Llama2-70b
- Wrong	Gemini	Llama2-7b	Gemini	Phi-2	Gemini	Phi-2
Maths						
- Basic	Phi-2	Llama2-7b	Phi-2	Llama2-7b	GPT-4	GPT-3.5
- Intermediate	Mistral 8x7b	Llama2-7b	GPT-4	Llama2-7b	GPT-4	Gemini
- Advanced	GPT-4	Llama2-7b	GPT-4	Llama2-7b	Mistral 8x7b	Llama2-7b
- Wrong	Mistral 8x7b	Llama2-13b	Phi-2	GPT-3.5	GPT-4	Gemini
General	GPT-4	Llama2-7b	GPT-4	Llama2-7b	GPT-4	Llama2-7b

Table 2: Models with the highest and the lowest scores across datasets and expertise levels.

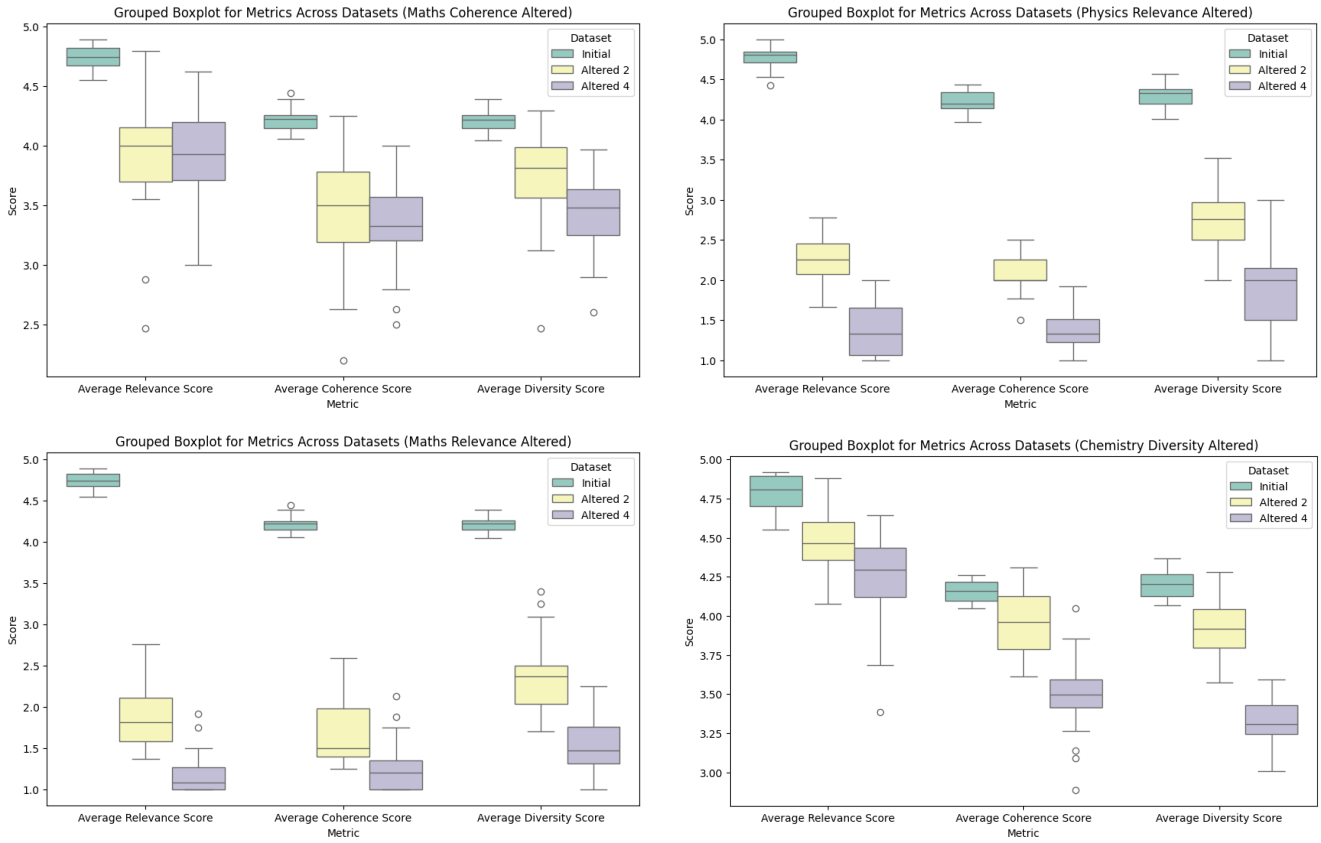

Figure 2: Boxplots showing average relevance, coherence, and diversity scores across altered datasets in Mathematics, Physics, and Chemistry. Each includes "Initial", "2-Altered", and "4-Altered" dataset versions.

Additionally, Phi-2's architecture leverages a scaled knowledge transfer (Mojan Javaheripi 2023) from its predecessor, Phi-1.5, which improves its performance on benchmark tests. These factors make Phi-2 an exceptional model within the specified tasks, demonstrating that well-planned training and design can yield high performance, challenging the prevailing notion that larger models are inherently superior.

Llama2 Models: These models consistently score lower, indicating possible limitations in their training or architectural adaptability. This underperformance highlights that

large models do not universally guarantee superior performance. This could be attributed to their primary optimization for other types of tasks, such as chat and dialogue scenarios, rather than the specific demands of generating novel and contextually deep questions in academic subjects. The models are trained on a new mix of publicly available online data, ensuring a broad knowledge base. However, their performance variability in CDQG tasks suggests that while they have strong general capabilities, they may require further tuning to excel specifically in the academic question generation domain.

Gemini: While generally showing lower overall performance, it excels in diversity, perhaps due to its multi-modal training. This suggests that it can generate more varied and creative outputs, which purely text-based models may not achieve.

Insights and implications While larger models like GPT-4 generally offered robust overall performance, smaller or specialized models like Phi-2 and Mistral 8x7b performed exceptionally well. This challenges the conventional notion that bigger is inherently better (Hoffmann et al. 2022), suggesting a nuanced approach to model selection based on specific task requirements.

Questioning the wrong statements

We expect to see the models doubt the credibility of the statements that are intentionally erroneous. While models generally follow the instructions by asking questions, their responses include questioning the credibility of dubious statements with probing questions like "*Are there any exceptions to this rule?*" While all the models do this, but how often they challenge a statement's truth varies. The models like Mistral, LLama 70b, and GPT-4 frequently ask this question in about 250 out of 600 cases the most. In contrast, GPT 3.5 and Llama 7b ask it less often, only about 100 to 150 times the least.

Ensuring the validity of CDQG

We validate the CDQG evaluation through an ablation study that incrementally add noise, as well as a human validation.

Noise-addition ablation

Setup For each entry in the output dataset containing five generated questions, we create two derivative entries by deliberately introducing disturbances. The first variant modifies two questions (**2 Altered**), while the second alters four questions (**4 Altered**). We execute this noise addition using GPT-4 (See Appendix for the prompt template) and verify that exactly 2 or 4 questions are modified in each respective variant, ensuring the noise addition diminishes question quality. This process yields six new datasets corresponding to each evaluation metric, divided between the two and four modified question scenarios. When we reintroduce these altered datasets to our evaluation process, we expect to observe a decline in scores across all metrics proportional to the added noise. This anticipated degradation aims to demonstrate an inverse correlation between LLM-generated content integrity

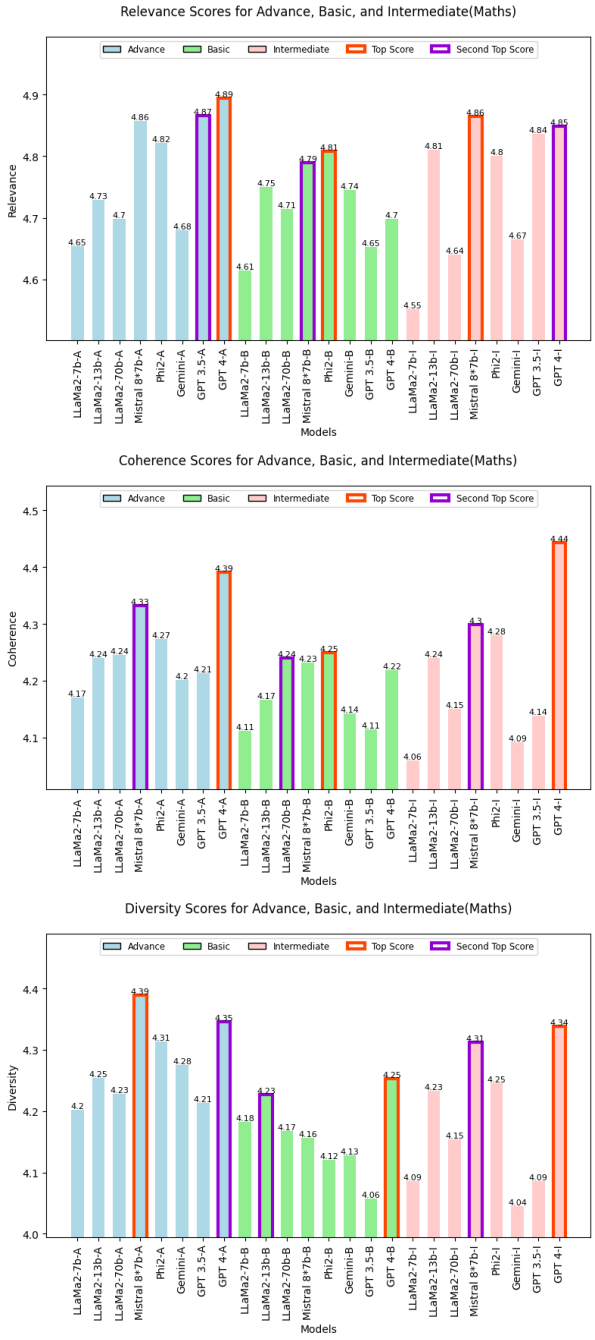

Figure 3: **Metric scores on Maths:** The set of bar charts provides a multidimensional analysis of various models, evaluated by three key performance metrics — Relevance (top), Coherence (middle), and Diversity (bottom). Each chart contrasts the scores across Advanced, Basic, and Intermediate expertise levels for maths, with distinct colors signifying the respective categories. Highlighted bars denote the top and second-highest scoring models within each metric, offering a visual synopsis of comparisons.

and noise level. This approach validates our hypothesis that LLMs can effectively differentiate between high-quality (signal) and compromised (noise) data inputs. By showing that introduced inaccuracies result in predictable evaluation score decreases, we employ a logical framework similar to mathematical proof by contradiction. This method demonstrates LLMs' effectiveness in judging relevance, coherence, and diversity.

Results The added noise significantly impacts all metrics, showing consistent decline as noise increases from no alterations to **2 Altered** and **4 Altered**, though the magnitude varies across metrics. The relevance metric exhibits the most pronounced trend, with scores dropping from approximately 4.8 to 2.2 to 1.0. This demonstrates that question precision and topic relevance are most sensitive to noise-induced disruptions. The coherence metric shows a less significant decrease, as alterations to individual questions do not always disrupt the overall logical flow and order. Diversity presents unique challenges, as effectively reducing this metric requires deep subject matter understanding and awareness of topic interconnections. As shown in Figure 2 and Section 8, while our modifications decrease the diversity score, the reduction is less pronounced than the relevance and coherence, reflecting the complexity of perturbing question diversity while maintaining topical consistency.

Our analysis confirms that noise introduction leads to degradation in LLM performance across all metrics. This validates our hypothesis that LLMs effectively differentiate between high-quality and noise-compromised content, and supports the robustness of our evaluation framework.

Human evaluation

Setup The human study starts by selecting a subset of the data. To maintain a manageable workload, we select questions from the first 10 statements of three models (out of eight) for two subject areas (out of four). We include all three variations for each subject area to ensure the generalizability of validation results. The selected 1,320 statements for human evaluation represent approximately 19.6% of our subdataset's 6,708 statements. A PhD student manually rated the questions on relevance, coherence, and diversity.

Results To analyze the agreement between human and LLM evaluations, we employ Cohen's kappa with linear weights (Doewes, Kurdhi, and Saxena 2023). This approach accounts for the ordinal nature of the rating scale and appropriately weights the proximity of agreement on scores, reducing penalties for minor discrepancies between evaluators. The resulting agreement scores demonstrate strong correlations: 0.736 for relevance, 0.698 for coherence, and 0.697 for diversity, indicating robust alignment between LLM and human evaluations.

Discussion

Questioning for better LM agents The ability to raise curiosity-driven questions is crucial for agentic systems that involve knowledge. Current technologies like tree-of-thought (Yao et al. 2024), maieutic prompting (Jung et al. 2022) and Reflexion (Shinn et al. 2023) incorporate functions resembling self-questioning. With improved questioning capabilities, future LM-based agents can better recognize low-quality information and reason about it, eventually being more robust against misinformation. A particularly useful use case for LM agents involves the external memory. Questioning equips the LM agents to inspect and potentially fix the errors within the memory.

Questioning for scientific discovery Curiosity-driven questioning has always been a critical step in scientific discovery. Human scientists raise questions along many steps of the endeavor of discovery. Questions like "Why can't an alternative method work here?" and "Why can't an alternative theory explain the data?" are the initial steps toward novel scientific discoveries.

Questioning in human-machine collaborations Language models have shown capabilities to elicit human preference (Li et al. 2023). As LMs appear more widely used in chatbots and other human-machine interaction systems, questioning becomes an increasingly important function that improves personalization. Questions can allow the models to clarify the human users' unspoken thoughts and intentions, improving the overall quality of communication (Wadhwa et al. 2024; Wu et al. 2024).

Conclusion

We propose CDQG and start the exploration for assessing an important capability of LLMs: the potential to seek knowledge driven by curiosity. The CDQG framework includes a task that elicits curiosity-driven questions, a dataset covering statements with varying levels of difficulty and supporting stratified studies, and an LLM-based evaluation setting which is validated by both noise-addition ablation and human evaluations. We find that across various subject domains, LLMs exhibit a strong capability to formulate relevant and coherent questions, underscoring their potential to engage in meaningful inquiry. The automated questioning setting has broad potential applications to improve the performance and usability of knowledge systems.

Limitations

While this study introduces an innovative framework for evaluating the questioning capabilities of LLMs, it primarily utilizes three metrics: relevance, coherence, and diversity. Though robust, these metrics may not capture the full depth of human-like questioning, such as emotional intelligence, knowledge acquisition, factual reasoning, etc. Future research could explore the integration of metrics that assess these human-centric qualities to better mimic real-world applications. Additionally, the evaluations are performed within a controlled academic setting, which might not fully reflect the complexities of natural environments where LLMs typically operate. Extending the evaluation to more dynamic settings or incorporating unstructured, real-world conversation data could enhance the applicability of the findings. Moreover, while our noise addition ablation study strengthens the assessment of model robustness, exploring more varied distur-

bances could provide a richer understanding of how LLMs perform under realistic and unpredictable conditions.

References

Abbasiantaeb, Z.; Yuan, Y.; Kanoulas, E.; and Aliannejadi, M. 2024. Let the LLMs Talk: Simulating Human-to-Human Conversational QA via Zero-Shot LLM-to-LLM Interactions. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, 8–17. Merida Mexico: ACM. ISBN 9798400703713.

Acar, S.; Berthiaume, K.; and Johnson, R. 2023. What kind of questions do creative people ask? *Journal of Creativity*, 33(3): 100062.

Ashok Kumar, N.; Fernandez, N.; Wang, Z.; and Lan, A. 2023. Improving Reading Comprehension Question Generation with Data Augmentation and Overgenerate-and-rank. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, 247–259. Toronto, Canada: Association for Computational Linguistics.

Baswani, P.; Mukherjee, A.; and Shrivastava, M. 2023. LTRC_IIITH's 2023 Submission for Prompting Large Language Models as Explainable Metrics Task. In *Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems*, 156–163. Bali, Indonesia: Association for Computational Linguistics.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, 1877–1901. Curran Associates, Inc.

Chen, Y.; Wang, R.; Jiang, H.; Shi, S.; and Xu, R. 2023. Exploring the Use of Large Language Models for Reference-Free Text Quality Evaluation: An Empirical Study. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, 361–374. Nusa Dua, Bali: Association for Computational Linguistics.

Chiang, C.-H.; and Lee, H.-y. 2023. Can Large Language Models Be an Alternative to Human Evaluations? ArXiv:2305.01937 [cs].

Doewes, A.; Kurdhi, N.; and Saxena, A. 2023. Evaluating Quadratic Weighted Kappa as the Standard Performance Metric for Automated Essay Scoring. In Feng, M.; Käser, T.; and Talukdar, P., eds., *Proceedings of the 16th International Conference on Educational Data Mining*, 103–113. International Educational Data Mining Society (IEDMS). 16th International Conference on Educational Data Mining, EDM 2023, EDM 2023 ; Conference date: 11-07-2023 Through 14-07-2023.

Elkins, S.; Kochmar, E.; Serban, I.; and Cheung, J. C. K. 2023. How Useful Are Educational Questions Generated by Large Language Models? In Wang, N.; Rebolledo-Mendez, G.; Dimitrova, V.; Matsuda, N.; and Santos, O. C., eds., *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky*, 536–542. Cham: Springer Nature Switzerland. ISBN 9783031363368.

Faggioli, G.; Dietz, L.; Clarke, C. L. A.; Demartini, G.; Hagen, M.; Hauff, C.; Kando, N.; Kanoulas, E.; Potthast, M.; Stein, B.; and Wachsmuth, H. 2023. Perspectives on Large Language Models for Relevance Judgment. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval*, 39–50. Taipei Taiwan: ACM. ISBN 9798400700736.

Ferrando, J.; Obeso, O.; Rajamanoharan, S.; and Nanda, N. 2024. Do I Know This Entity? Knowledge Awareness and Hallucinations in Language Models.

Fu, J.; Ng, S.-K.; Jiang, Z.; and Liu, P. 2023. GPTScore: Evaluate as You Desire. ArXiv:2302.04166 [cs].

Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; Wang, M.; and Wang, H. 2024. Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv:2312.10997.

Gemini Team; Anil, R.; and et al. 2024. Gemini: A Family of Highly Capable Multimodal Models. ArXiv:2312.11805 [cs].

Ghafarollahi, A.; and Buehler, M. J. 2024. SciAgents: Automating scientific discovery through multi-agent intelligent graph reasoning. arXiv:2409.05556.

Gilardi, F.; Alizadeh, M.; and Kubli, M. 2023. ChatGPT outperforms crowd workers for text-annotation tasks.

Hao, S.; Gu, Y.; Ma, H.; Hong, J.; Wang, Z.; Wang, D.; and Hu, Z. 2023. Reasoning with Language Model is Planning with World Model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 8154–8173. Singapore: Association for Computational Linguistics.

He, Q.; Zeng, J.; Huang, W.; Chen, L.; Xiao, J.; He, Q.; Zhou, X.; Liang, J.; and Xiao, Y. 2024. Can Large Language Models Understand Real-World Complex Instructions? *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16): 18188–18196.

Hoffmann, J.; Borgeaud, S.; Mensch, A.; Buchatskaya, E.; Cai, T.; Rutherford, E.; Casas, D. d. L.; Hendricks, L. A.; Welbl, J.; Clark, A.; et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Huang, F.; Kwak, H.; Park, K.; and An, J. 2024. ChatGPT Rates Natural Language Explanation Quality like Humans: But on Which Scales? In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 3111–3132. Torino, Italia: ELRA and ICCL.

Huang, Y.; and Huang, J. 2024. A Survey on Retrieval-Augmented Text Generation for Large Language Models. arXiv:2404.10981.

Jiang, A. Q.; Sablayrolles, A.; Roux, A.; Mensch, A.; Savary, B.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Hanna, E. B.; Bressand, F.; Lengyel, G.; Bour, G.; Lample, G.; Lavaud, L. R.; Saulnier, L.; Lachaux, M.-A.; Stock, P.; Subramanian, S.; Yang, S.; Antoniak, S.; Scao, T. L.; Gervet, T.; Lavril, T.;

Wang, T.; Lacroix, T.; and Sayed, W. E. 2024. Mixtral of Experts. ArXiv:2401.04088 [cs].

Jiang, D.; Li, Y.; Zhang, G.; Huang, W.; Lin, B. Y.; and Chen, W. 2023. TIGERScore: Towards Building Explainable Metric for All Text Generation Tasks. ArXiv:2310.00752 [cs].

Jung, J.; Qin, L.; Welleck, S.; Brahman, F.; Bhagavatula, C.; Le Bras, R.; and Choi, Y. 2022. Maieutic Prompting: Logically Consistent Reasoning with Recursive Explanations. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 1266–1279. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Kamal Eddine, M.; Shang, G.; Tixier, A.; and Vazirgiannis, M. 2022. FrugalScore: Learning Cheaper, Lighter and Faster Evaluation Metrics for Automatic Text Generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1305–1318. Dublin, Ireland: Association for Computational Linguistics.

Ke, N. R.; Sawyer, D. P.; Soyer, H.; Engelcke, M.; Reichert, D. P.; Hudson, D. A.; Reid, J.; Lerchner, A.; Rezende, D. J.; Lillicrap, T. P.; Mozer, M.; and Wang, J. X. 2024. Can foundation models actively gather information in interactive environments to test hypotheses? arXiv:2412.06438.

Kocmi, T.; and Federmann, C. 2023. Large Language Models Are State-of-the-Art Evaluators of Translation Quality. In Nurminen, M.; Brenner, J.; Koponen, M.; Latomaa, S.; Mikhailov, M.; Schierl, F.; Ranasinghe, T.; Vanmassenhove, E.; Vidal, S. A.; Aranberri, N.; Nunziatini, M.; Escartín, C. P.; Forcada, M.; Popovic, M.; Scarton, C.; and Moniz, H., eds., *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, 193–203. Tampere, Finland: European Association for Machine Translation.

Kotonya, N.; Krishnasamy, S.; Tetreault, J.; and Jaimes, A. 2023. Little Giants: Exploring the Potential of Small LLMs as Evaluation Metrics in Summarization in the Eval4NLP 2023 Shared Task. In *Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems*, 202–218. Bali, Indonesia: Association for Computational Linguistics.

Kotov, A.; and Zhai, C. 2010. Towards natural question guided search. In *Proceedings of the 19th international conference on World wide web*, 541–550. Raleigh North Carolina USA: ACM. ISBN 9781605587998.

Krishna, S.; Krishna, K.; Mohananey, A.; Schwarcz, S.; Stambler, A.; Upadhyay, S.; and Faruqui, M. 2024. Fact, Fetch, and Reason: A Unified Evaluation of Retrieval-Augmented Generation. arXiv:2409.12941.

Kurdi, G.; Leo, J.; Parsia, B.; Sattler, U.; and Al-Emari, S. 2020. A Systematic Review of Automatic Question Generation for Educational Purposes. *International Journal of Artificial Intelligence in Education*, 30(1): 121–204.

Lee, Y.; Kim, J.; Kim, J.; Cho, H.; and Kang, P. 2024. CheckEval: Robust Evaluation Framework using Large Language Model via Checklist. ArXiv:2403.18771 [cs].

Leiter, C.; Lertvittayakumjorn, P.; Fomicheva, M.; Zhao, W.; Gao, Y.; and Eger, S. 2024. Towards Explainable Evaluation Metrics for Machine Translation. *Journal of Machine Learning Research*, 25(75): 1–49.

Leiter, C.; Opitz, J.; Deutsch, D.; Gao, Y.; Dror, R.; and Eger, S. 2023. The Eval4NLP 2023 Shared Task on Prompting Large Language Models as Explainable Metrics. In *Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems*, 117–138. Bali, Indonesia: Association for Computational Linguistics.

Li, B. Z.; Tamkin, A.; Goodman, N.; and Andreas, J. 2023. Eliciting Human Preferences with Language Models. arXiv:2310.11589.

Li, D.; Jiang, B.; Huang, L.; Beigi, A.; Zhao, C.; Tan, Z.; Bhattacharjee, A.; Jiang, Y.; Chen, C.; Wu, T.; Shu, K.; Cheng, L.; and Liu, H. 2024. From Generation to Judgment: Opportunities and Challenges of LLM-as-a-judge. arXiv:2411.16594.

Li, J.; Miller, A. H.; Chopra, S.; Ranzato, M.; and Weston, J. 2017. Learning through Dialogue Interactions by Asking Questions. ArXiv:1612.04936 [cs].

Liu, K.; Casper, S.; Hadfield-Menell, D.; and Andreas, J. 2024a. Cognitive Dissonance: Why Do Language Model Outputs Disagree with Internal Representations of Truthfulness? In *EMNLP*.

Liu, M.; Shen, Y.; Xu, Z.; Cao, Y.; Cho, E.; Kumar, V.; Ghanadan, R.; and Huang, L. 2024b. X-Eval: Generalizable Multi-aspect Text Evaluation via Augmented Instruction Tuning with Auxiliary Evaluation Aspects. ArXiv:2311.08788 [cs].

Liu, Y.; Iter, D.; Xu, Y.; Wang, S.; Xu, R.; and Zhu, C. 2023a. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2511–2522. Singapore: Association for Computational Linguistics.

Liu, Y.; Yang, T.; Huang, S.; Zhang, Z.; Huang, H.; Wei, F.; Deng, W.; Sun, F.; and Zhang, Q. 2023b. Calibrating LLM-Based Evaluator. ArXiv:2309.13308 [cs].

Liusie, A.; Manakul, P.; and Gales, M. 2024. LLM Comparative Assessment: Zero-shot NLG Evaluation through Pairwise Comparisons using Large Language Models. In Graham, Y.; and Purver, M., eds., *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 139–151. St. Julian's, Malta: Association for Computational Linguistics.

Lucas, E.; Steelman, K. S.; Ureel, L. C.; and Wallace, C. 2024. For those who don't know (how) to ask: Building a dataset of technology questions for digital newcomers. ArXiv:2403.18125 [cs].

Masterman, T.; Besen, S.; Sawtell, M.; and Chao, A. 2024. The Landscape of Emerging AI Agent Architectures for Reasoning, Planning, and Tool Calling: A Survey. ArXiv:2404.11584 [cs].

Mojan Javaheripi, S. B. 2023. Phi-2: The surprising power of small language models.

OpenAI. 2024. Learning to Reason with Large Language Models. https://openai.com/index/learning-to-reason-with-llms/. Accessed: 2024-09-13.

OpenAI; Achiam, J.; Adler, S.; and et al. 2024. GPT-4 Technical Report. ArXiv:2303.08774 [cs].

Petroni, F.; Rocktäschel, T.; Riedel, S.; Lewis, P.; Bakhtin, A.; Wu, Y.; and Miller, A. 2019. Language Models as Knowledge Bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2463–2473. Hong Kong, China: Association for Computational Linguistics.

Shinn, N.; Cassano, F.; Berman, E.; Gopinath, A.; Narasimhan, K.; and Yao, S. 2023. Reflexion: Language Agents with Verbal Reinforcement Learning. arXiv:2303.11366.

Si, C.; Yang, D.; and Hashimoto, T. 2024. Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers. arXiv:2409.04109.

Suzgun, M.; Gur, T.; Bianchi, F.; Ho, D. E.; Icard, T.; Jurafsky, D.; and Zou, J. 2024. Belief in the Machine: Investigating Epistemological Blind Spots of Language Models.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; Bikel, D.; Blecher, L.; Ferrer, C. C.; Chen, M.; Cucurull, G.; Esiobu, D.; Fernandes, J.; Fu, J.; Fu, W.; Fuller, B.; Gao, C.; Goswami, V.; Goyal, N.; Hartshorn, A.; Hosseini, S.; Hou, R.; Inan, H.; Kardas, M.; Kerkez, V.; Khabsa, M.; Kloumann, I.; Korenev, A.; Koura, P. S.; Lachaux, M.-A.; Lavril, T.; Lee, J.; Liskovich, D.; Lu, Y.; Mao, Y.; Martinet, X.; Mihaylov, T.; Mishra, P.; Molybog, I.; Nie, Y.; Poulton, A.; Reizenstein, J.; Rungta, R.; Saladi, K.; Schelten, A.; Silva, R.; Smith, E. M.; Subramanian, R.; Tan, X. E.; Tang, B.; Taylor, R.; Williams, A.; Kuan, J. X.; Xu, P.; Yan, Z.; Zarov, I.; Zhang, Y.; Fan, A.; Kambadur, M.; Narang, S.; Rodriguez, A.; Stojnic, R.; Edunov, S.; and Scialom, T. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. ArXiv:2307.09288 [cs].

Valmeekam, K.; Marquez, M.; Olmo, A.; Sreedharan, S.; and Kambhampati, S. 2023. PlanBench: An Extensible Benchmark for Evaluating Large Language Models on Planning and Reasoning about Change. *Advances in Neural Information Processing Systems*, 36: 38975–38987.

Verga, P.; Hofstatter, S.; Althammer, S.; Su, Y.; Piktus, A.; Arkhangorodsky, A.; Xu, M.; White, N.; and Lewis, P. 2024. Replacing Judges with Juries: Evaluating LLM Generations with a Panel of Diverse Models. ArXiv:2404.18796 [cs].

Wadhwa, M.; Zhao, X.; Li, J. J.; and Durrett, G. 2024. Learning to Refine with Fine-Grained Natural Language Feedback. arXiv:2407.02397.

Wang, H.; Zou, J.; Mozer, M.; Goyal, A.; Lamb, A.; Zhang, L.; Su, W. J.; Deng, Z.; Xie, M. Q.; Brown, H.; and Kawaguchi, K. 2024a. Can AI Be as Creative as Humans? ArXiv:2401.01623 [cs].

Wang, J.; Liang, Y.; Meng, F.; Sun, Z.; Shi, H.; Li, Z.; Xu, J.; Qu, J.; and Zhou, J. 2023. Is ChatGPT a Good NLG Evaluator? A Preliminary Study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, 1–11. Hybrid: Association for Computational Linguistics.

Wang, W.; Shi, J.; Wang, C.; Lee, C.; Yuan, Y.; tse Huang, J.; and Lyu, M. R. 2024b. Learning to Ask: When LLMs Meet Unclear Instruction. arXiv:2409.00557.

Wu, Y.; Mangla, R.; Dimakis, A. G.; Durrett, G.; and Li, J. J. 2024. Which questions should I answer? Salience Prediction of Inquisitive Questions. arXiv:2404.10917.

Xiong, S.; Payani, A.; Kompella, R.; and Fekri, F. 2024. Large Language Models Can Learn Temporal Reasoning. ArXiv:2401.06853 [cs].

Xu, W.; Wang, D.; Pan, L.; Song, Z.; Freitag, M.; Wang, W.; and Li, L. 2023. INSTRUCTSCORE: Towards Explainable Text Generation Evaluation with Automatic Feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 5967–5994. Singapore: Association for Computational Linguistics.

Yao, B.; Wang, D.; Wu, T.; Zhang, Z.; Li, T.; Yu, M.; and Xu, Y. 2022. It is AI's Turn to Ask Humans a Question: Question-Answer Pair Generation for Children's Story Books. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 731–744. Dublin, Ireland: Association for Computational Linguistics.

Yao, S.; Yu, D.; Zhao, J.; Shafran, I.; Griffiths, T.; Cao, Y.; and Narasimhan, K. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Zelikman, E.; Harik, G.; Shao, Y.; Jayasiri, V.; Haber, N.; and Goodman, N. D. 2024. Quiet-STaR: Language Models Can Teach Themselves to Think Before Speaking. ArXiv:2403.09629 [cs].

Zhao, W.; Strube, M.; and Eger, S. 2023. DiscoScore: Evaluating Text Generation with BERT and Discourse Coherence. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 3865–3883. Dubrovnik, Croatia: Association for Computational Linguistics.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; Zhang, H.; Gonzalez, J. E.; and Stoica, I. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems*, 36: 46595–46623.

Zhong, M.; Liu, Y.; Yin, D.; Mao, Y.; Jiao, Y.; Liu, P.; Zhu, C.; Ji, H.; and Han, J. 2022. Towards a Unified Multi-Dimensional Evaluator for Text Generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2023–2038. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Zhu, Y.; Wang, X.; Chen, J.; Qiao, S.; Ou, Y.; Yao, Y.; Deng, S.; Chen, H.; and Zhang, N. 2024. LLMs for Knowledge Graph Construction and Reasoning: Recent Capabilities and Future Opportunities.

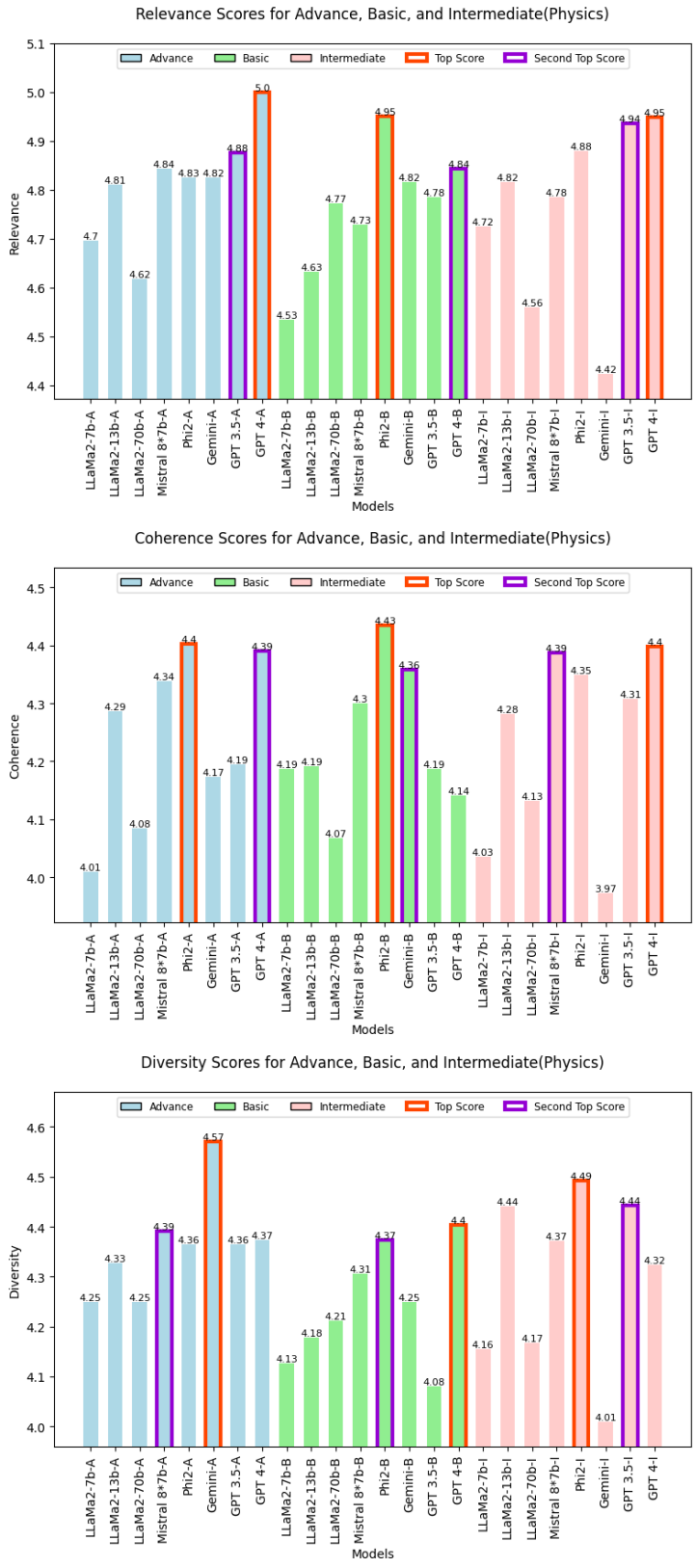

Figure 4: **Metric scores on Physics:** The set of bar charts provides a multidimensional analysis of various models, evaluated by three key performance metrics — Relevance, Coherence , and Diversity. Each chart contrasts the scores across Advanced , Basic , and Intermediate expertise levels for maths, with distinct colors signifying the respective categories. Highlighted bars denote the top and second-highest scoring models within each metric, offering a visual synopsis of comparisons.

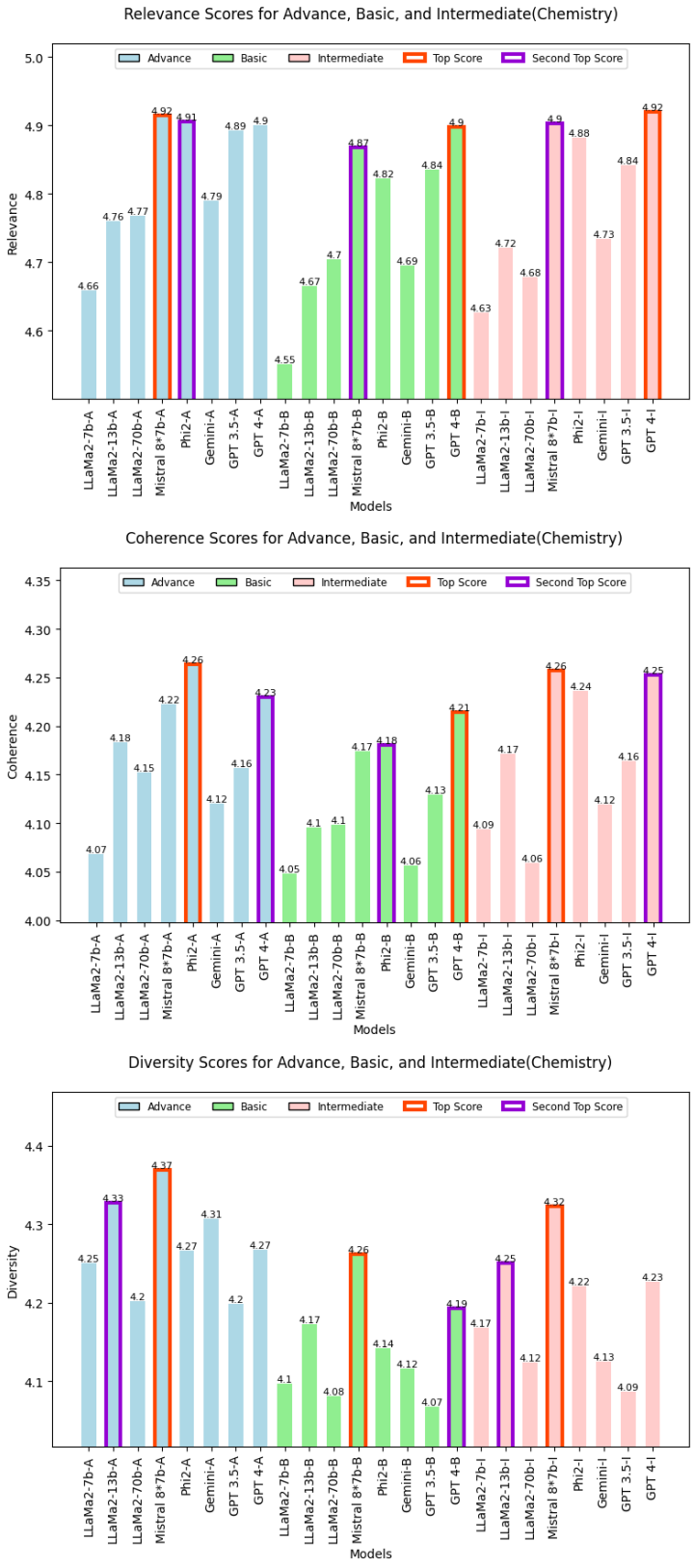

Figure 5: **Metric scores on Chemistry:** The set of bar charts provides a multidimensional analysis of various models, evaluated by three key performance metrics — Relevance, Coherence , and Diversity. Each chart contrasts the scores across Advanced , Basic , and Intermediate expertise levels for Chemistry, with distinct colors signifying the respective categories. Highlighted bars denote the top and second-highest scoring models within each metric, offering a visual synopsis of comparisons.

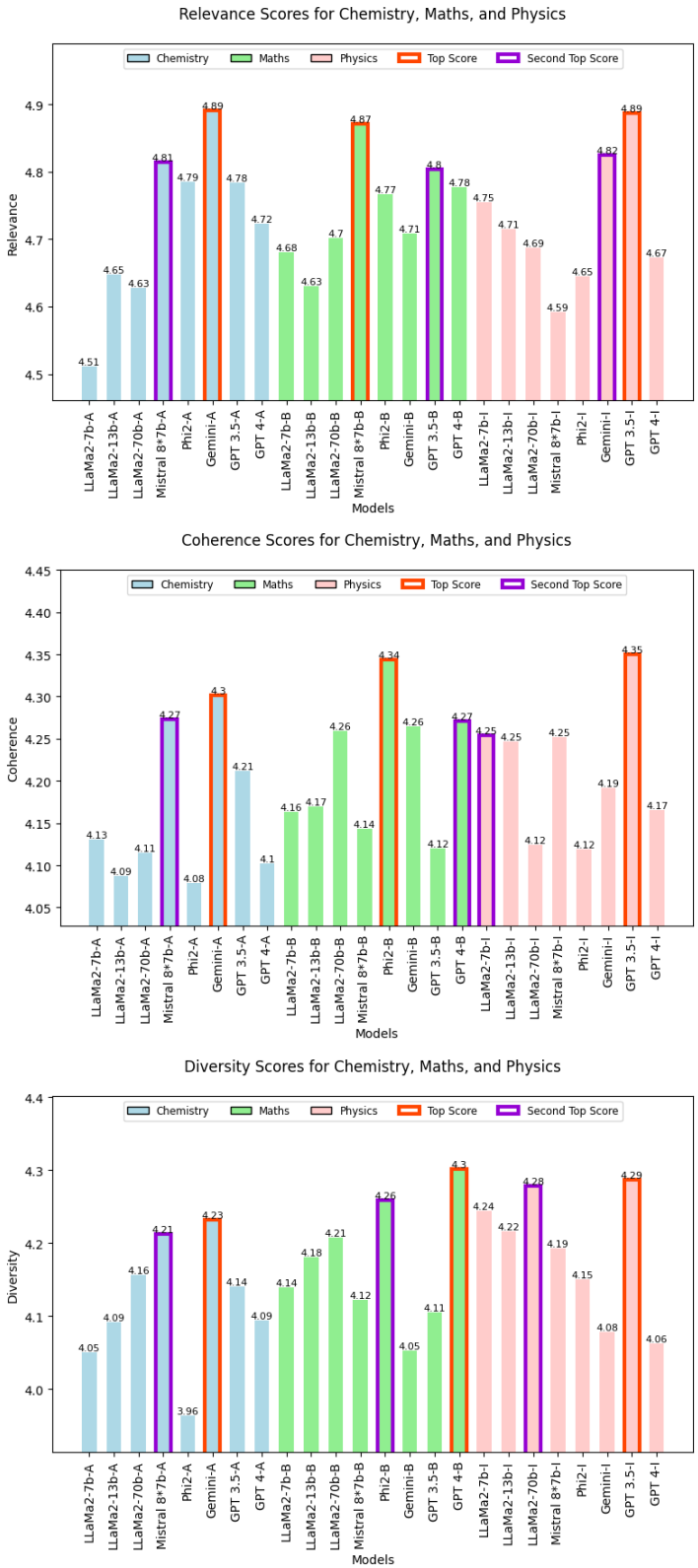

Figure 6: **Metric scores on Wrong Statements:** The set of bar charts provides a multidimensional analysis of various models, evaluated by three key performance metrics — Relevance, Coherence , and Diversity. Each chart contrasts the scores across Chemistry , Maths , and Physics subjects , with distinct colors signifying the respective categories. Highlighted bars denote the top and second-highest scoring models within each metric, offering a visual synopsis of comparisons.

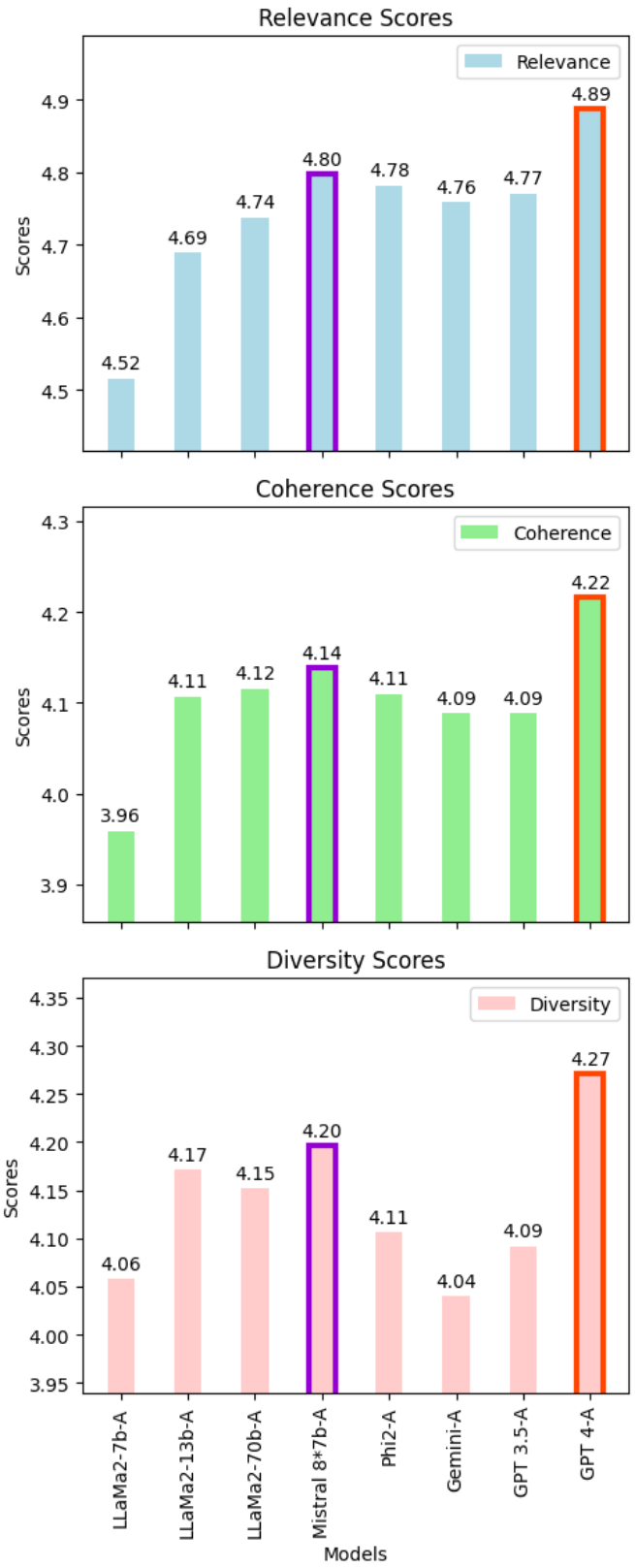

Figure 7: **Metric scores on General Statements:** The set of bar charts provides a multidimensional analysis of various models, evaluated by three key performance metrics — Relevance, Coherence , and Diversity. Each chart contrasts the scores across Relevance , Coherence , and Diversity . Highlighted bars denote the top and second-highest scoring models within each metric, offering a visual synopsis of comparisons.

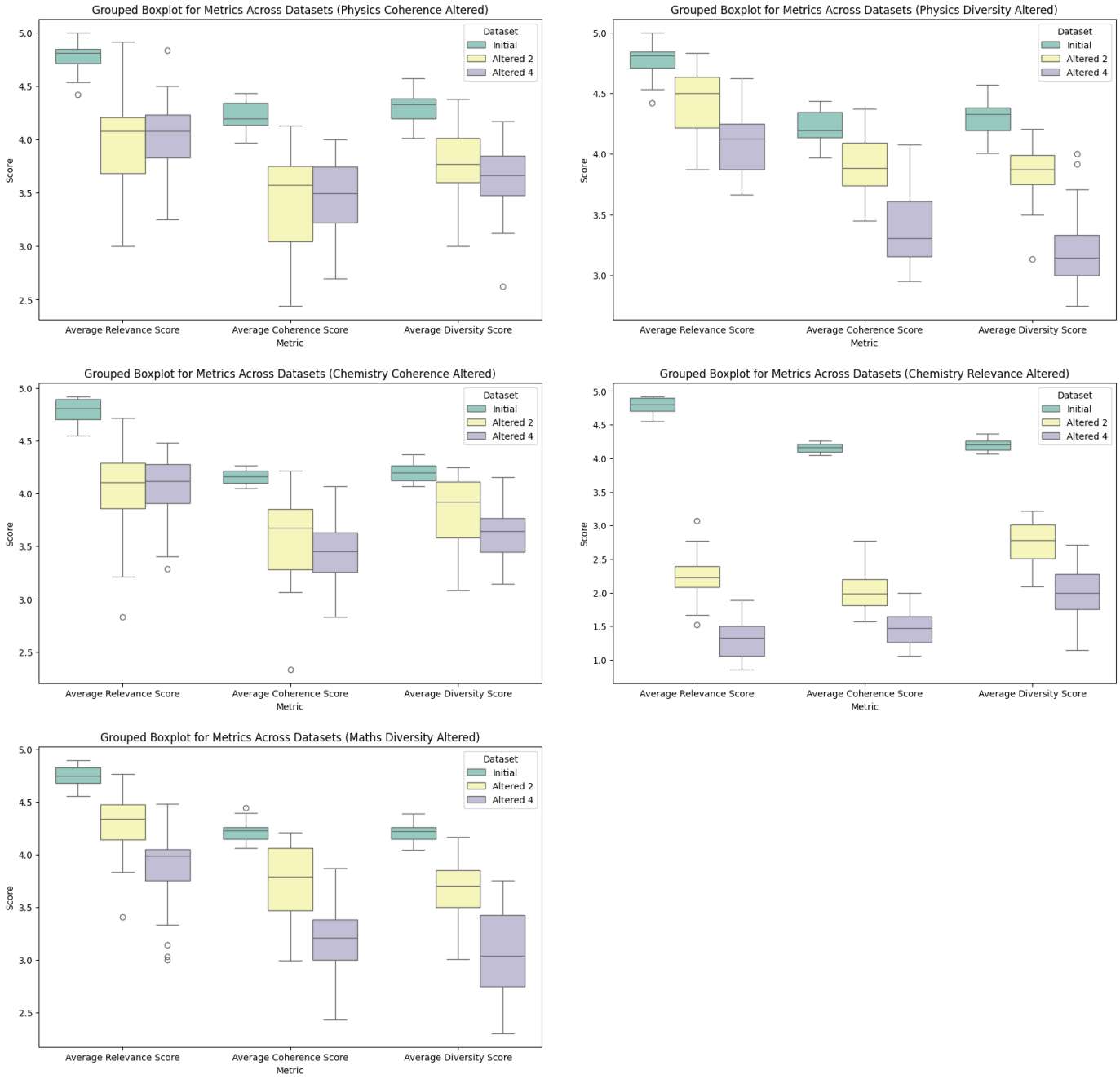

Figure 8: Boxplots showing average relevance, coherence, and diversity scores across altered datasets in Physics (first row), Chemistry (second row), and Maths (third row). Each includes 'Initial', 'Altered 2', and 'Altered 4' dataset versions.

List of prompt templates

Prompt 1: Curiosity-Driven Question Generation

Imagine you are a human encountering this *{subject}* for the first time: "*{scenario}*". List the top 5 questions that would come to your mind, useful for learning about it as you are new to it. Provide your questions in a simple bullet point list.

Prompt 3: Combining Scoring and Justification using Gemini

Initial Query: {instruction}
Answer Given by LLM: {llm_answer}
Scores by humans: Human 1: {human_score_1}, Human 2: {human_score_2}, Human 3: {human_score_3}.

These are three scorings by a human and the justifications. Now, consider all the scorings and their justifications and give final scores for relevance, coherence, and diversity. Don't just take the average of scores or support one scorer; instead, read the justifications and, accordingly, give a final score and justify. Provide output in JSON format.

Prompt 2: Evaluation Task

Below are sets of 5 questions generated by different Language Models (LLMs) in response to a specific statement or scenario they were presented with for the first time. Your task is to evaluate these questions based on the following three metrics: Coherence, Relevance, and Diversity. Each set of questions is aimed at uncovering and understanding the elements and concepts within the given statement.

Criteria for each metric:

- **Relevance:** Assess how directly each question pertains to the specific details, elements, or concepts presented in the statement or scenario. Questions should aim to clarify, expand upon, or directly explore the content of the statement, focusing on the immediate context rather than peripheral or advanced topics not directly introduced by the statement.
- **Coherence:** Evaluate how logically the questions within each set connect to one another and whether they form a coherent line of inquiry that would logically progress a beginner's understanding of the topic. Consider if the sequence of questions or their thematic connection facilitates a structured exploration of the statement.
- **Diversity:** Determine the range of aspects covered by the questions in relation to the statement, ensuring that each question brings a new dimension or perspective to understanding the statement. While maintaining direct relevance, the questions should collectively offer a broad exploration of the topic, including but not limited to definitions, implications, applications, or theoretical underpinnings.

For each set of questions, provide a score from 1 to 5 for each metric, where 1 indicates that the questions poorly meet the criteria and 5 indicates excellent adherence to the criteria. Additionally, provide brief justifications for your scores, highlighting strengths and areas for improvement in relation to the three metrics.

Your evaluation will help determine which LLM produced the most effective set of questions for fostering an understanding of the given statement or scenario, balancing direct relevance to the statement, logical coherence in inquiry, and diversity in exploration.

Input for LLM: {*instruction*}
LLM Output: {*model_output*}

Prompt 4: Alteration Prompt
Initial Query to random LLM: {instruction} and the Output given by that LLM: {model_output},
Given a set of questions related to a specific statement provided by an LLM, modify exactly 4 questions for each metric to intentionally introduce noise. The objective is to decrease the values of three specified metrics: relevance, coherence, and diversity, in relation to the original statement.

For Relevance: Alter 4 random questions to make them less directly connected to the main topic of the statement. The goal is to subtly shift focus without completely diverging into unrelated topics.

For Coherence: Revise the sequence or content of 4 random questions to break the logical flow of inquiry. Adjustments should make the progression less structured and more challenging to follow, thus impacting the coherence of the set.

For Diversity: Change or add 4 random questions to concentrate more narrowly on similar aspects or repeat themes. This reduces the range of explored topics, affecting the overall diversity of the question set.

After making these modifications, specify the number of questions you altered for each metric and provide the altered list of questions. Your output should demonstrate the impact of introduced noise on the measurement of each metric.

Required Output Format

Your response should be structured in JSON format, comprising three sections corresponding to the metrics: Relevance, Coherence, and Diversity. Each section must detail the number of questions modified ('changed') and include the revised list of questions after changes ('questions'). Avoid including explanations or content beyond this structured format.

Model Configuration Details

Gemini Settings: The Gemini model was configured with a low temperature setting of 0.1 to ensure predictable and consistent outputs. The top_p and top_k parameters were both set to 1, constraining the model to the most likely outcomes. The maximum output tokens were limited to 400 to balance detail with computational efficiency. Safety settings were established to minimize the risk of generating harmful content, with no blocks applied across categories such as harassment, hate speech, sexually explicit content, and dangerous content.

Mistral Model Setup: The Mistral model utilized a tokenizer and model settings specifically tailored for instruction-based tasks. This setup included using the AutoTokenizer and AutoModelForCausalLM from a pretrained snapshot, equipped with BitsAndBytesConfig for efficient quantization. The configuration ensured operations were optimized for 4-bit quantization and the compute dtype set to float16, enhancing the model's performance while reducing memory usage. The text-generation pipeline was adjusted with a temperature of 0.1 and a repetition penalty of 1.1 to generate more coherent and less repetitive text, with a limit of 128 new tokens per generation instance.

Llama Model Configurations: For the Llama models, including, Llama 7b, Llama 13b and Llama 70b, configurations were similarly tailored to enhance performance and efficiency. Both models used quantization settings conducive to low-memory consumption while maintaining computational precision. These settings were crucial for managing the large parameter size inherent to these models. Each model's generation pipeline was configured to produce full-text outputs with controlled temperature settings and repetition penalties to ensure relevance and diversity in the generated text.

Phi2 Model Configuration: The Phi2 model from Microsoft was set up with advanced quantization techniques to support efficient processing. The model and tokenizer were loaded from a specific snapshot with settings that enabled high-performance text generation. The generation settings included a controlled temperature for predictability, a sampling strategy to introduce variety, and a repetition penalty to avoid redundant content, making it well-suited for generating diverse and engaging text.

Compute Resources: For models accessed via API, computations were performed using CPU resources. In contrast, models retrieved from HuggingFace were run on a single NVIDIA GPU setup equipped with 48GB of RAM. Notably, all models utilized in this study were quantized versions, optimizing computational efficiency and resource usage.