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

Anonymous ACL submission

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. Towards bridging this gap, we propose an evaluation framework, CDQG (Curiosity-Driven Question Generation)[1]. 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.

1 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 surrounding environments, explore the mechanisms in processes, 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, improving comprehension (Acar et al., 2023). In education, questioning is closely linked to higher-level thinking skills like analysis, synthesis, and evaluation (Kurdi et al., 2020). The complexity & depth of questions asked often reflect a person's grasp and understanding of a topic (Kotov and Zhai, 2010).

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 diversity — scores with roots in the literature of psychology (Zhao et al., 2023). We use state-of-the-art LLMs to compute

[1] Upon acceptance of this paper, the complete details of our research along with the CDQG dataset will be made available at `url_here`.

these scores. The scores are validated by human judgment as well as 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 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 formalize CDQG, a novel framework to quantitatively evaluate an LLM's intrinsic capability to seek knowledge.
- We compile the CDQG dataset, which includes varied and challenging content to test the questioning capabilities of LLMs.
- We validate the CDQG evaluation protocol through both human evaluation and a novel automated noise-injection 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.

2 Related Works

2.1 Question Generation

Question generation is a crucial task in educational settings, underscored by various studies (Elkins et al., 2023; Kurdi et al., 2020). This domain has transitioned from early rule-based systems (Yao et al., 2022) to the utilization of advanced transformer-based models, and more recently, to large language models (LLMs). This progression towards employing deep learning techniques has not only improved the relevance and quality of questions but also facilitated more dynamic interactions within educational software (Abbasiantaeb et al., 2024) and conversational systems (Wang et al., 2024b). Differing from conversational frameworks like those studied by (Scialom and Staiano, 2020), our research presents a unique evaluation framework that assesses LLMs' capacity to generate curiosity-driven questions based on static scientific statements. This approach emphasizes intrinsic curiosity and a pursuit of knowledge, moving away from reliance on predefined question templates or task-specific objectives.

2.2 Evaluation of Generative Models

In evaluating text generation from LLMs, recent methodologies have shifted towards multifaceted approaches that resonate more with human judgment. GPTScore (Fu et al., 2023) and UniEval (Leiter et al., 2023) leverage LLMs' natural language understanding to tailor evaluations to specific criteria, with GPTScore focusing on fluency and UniEval employing a boolean question-answering format to assess multiple quality dimensions. CheckEval (Lee et al., 2024) utilizes a structured checklist for reliability, while X-Eval (Liu et al., 2024b) dynamically selects evaluation aspects to enhance adaptability. The zero-shot comparative methodology (Liusie et al., 2024) and the Unified Framework (Zhong et al., 2022) combine traditional and novel approaches for direct quality judgments. PlanBench (Valmeekam et al., 2023) investigates LLM reasoning capabilities through planning tasks, and TIGERSCORE (Jiang et al., 2023) emphasizes explainability. Complementing these are strategies that test LLMs' instruction-following skills (He et al., 2024) and a composite metric system that aggregates scores for holistic assessment (Verga et al., 2024). Unlike these methodologies, which focus on how LLMs answer questions or execute tasks, our work uniquely assesses their capacity to generate meaningful questions, introducing a new dimension to LLM evaluation.

2.3 LLMs for Reasoning

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 et al., 2024; Ghafarollahi and Buehler, 2024). Thoughtful questions can open doors to new ideas and solutions.

2.4 LLMs for Evaluation

Recent studies highlight LLMs' potential to achieve human-level assessment quality in various tasks (Gilardi et al., 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). Wang et al. (2023) finds 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., 2023). 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 emphasizes the importance of transparency in LLM assessments (Leiter et al., 2024). Similar to previous works, we add incremental noises to validate the robustness of LLM evaluation.

3 CDQG framework

As summarized by Figure 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.

3.1 CDQG task

The CDQG task starts with sampling a statement from the CDQG dataset (which we'll explain in detail in Section 3.3).

Then, CDQG prompts the model to conceptualize itself as a human who encounters the statement for the first time and devoid of prior knowledge. CDQG then prompts the model to generate the top five questions that would instinctively arise. This prompt is constructed to accommodate the distinct instructional formats of multiple models, and allows us to elicit the models' inquisitive capabilities in a novel and controlled environment. Figure 1 shows a prompt example, Appendix B shows more. The full prompt template is listed in Appendix A.

3.2 CDQG evaluation

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 (Zhao et al., 2023; Sun et al., 2023).

Coherence: Coherence assesses how logically the questions within each set connect. Coherence in the chatbot literature checks if the sequence facilitates a structured exploration (Wang et al., 2020). Following this definition, a set of questions with a high coherence score forms a coherent line of inquiry that would logically progress a beginner's understanding of the topic.

Diversity: Diversity describes the range of aspects covered by the questions to the statement. The questions with a high diversity score collectively offer a broad exploration of the topic, including but not limited to definitions, implications, applications, or theoretical underpinnings (Puranik et al., 2023; Sultan et al., 2020; Guo et al., 2024).

We chose these metrics because together they comprehensively capture essential dimensions of curiosity-driven inquiry: relevance assesses direct pertinence to the initial statement, coherence evalu-

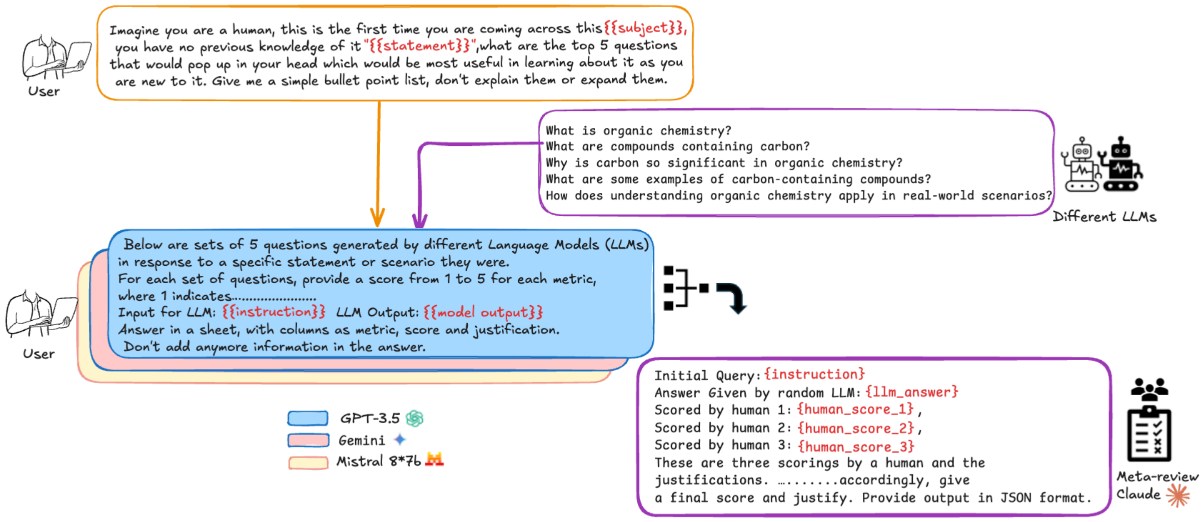

Figure 1: The CDQG framework. The top half shows the CDQG task, and the lower half shows the evaluation method of the generated questions.

ates logical depth and meaningful sequential exploration, and diversity ensures breadth by covering multiple aspects or perspectives. Combined, these metrics effectively encapsulate key attributes associated with curiosity.

Scoring procedure We use LLMs to score the generations on relevance, coherence, and diversity, following the recent LLM-as-a-judge trend (Li et al., 2024). We select three LLMs, 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 each of the three LLM judges to generate a score on a 5-point Likert scale and the corresponding justifications (the prompt template is included in Appendix A). Then, we use Claude[2] 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 metareviewer model to provide a fair score.

In case one of these models is used for question generation, our scoring procedure mitigates its potential biases: each metric is scored independently by 3 models to reduce reliance on any single model's perspective or biases. We also set up 2 validation studies to show the validity of this evaluation protocol: an automatic noise-injection experiment and a human validation experiment. The details of the 2 validation studies are described in Section 6.

[2]claude-3-5-sonnet-20241022

3.3 CDQG dataset

The CDQG dataset facilitates the CDQG evaluation framework. We leverage GPT-4's generative capabilities under human oversight, with domain experts, qualified PhD students familiar with the respective subject categorizing statements into basic, intermediate, and advanced levels based on educational standards, to assemble the dataset incrementally (Xu et al., 2023), selecting statements that span diverse topics and complexity levels. Table 1 shows the dataset's splits and their corresponding sizes. We consider the following desiderata when constructing the CDQG dataset.

Multiple subjects and Difficulty Levels 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. For each academic subject, we split the dataset into distinct difficulty levels to allow for stratified assessments of LLM knowledge-seeking behavior. Each level contains a balanced number of statements. To validate this categorization, we conducted a human evaluation process involving three independent annotators. The inter-annotator agreement, measured using pairwise Cohen's Kappa against the original labels, was found to be substantial, yielding an average Cohen's κ of 0.639. (Refer Appendix subsection C.1 for details)

Wrong statements A unique feature of our dataset is the inclusion of these intentionally er-

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.

roneous 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.

4 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. Standard hyperparameters recommended by model documentation were used without modification. 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’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 are accessed via Huggingface.

5 Results

Table 2, Figure 3 and Figure 2 illustrate our main results, with the rest in the Appendix D.

5.1 Performance by model

GPT-4: Dominates in most 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.

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: Despite its 2.7-billion model size, Phi-2 produces highly relevant and coherent questions at basic- to intermediate-level tasks. Phi-2 benefits significantly from high-quality, curated training data that emphasizes “textbook-quality” content (Mojan Javaheripi, 2023), enhancing its logical reasoning and commonsense understanding abilities. 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 even 70b consistently perform below other models in the evaluation, though occasionally achieve comparable scores. The Llama models have broad knowledge bases and excel in chat and dialogue tasks. However, their performance in CDQG tasks suggests that while they have strong general capabilities, they may benefit from further tuning to excel specifically in the academic question generation domain.

Gemini: Gemini’s performance is mixed, often appearing in the lower tier across several categories. It particularly struggles with Relevance and Coherence for Basic Physics questions and with Coherence in intermediate and advanced Maths.

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 better (Hoffmann et al., 2022), suggesting a nuanced approach to model selection based on specific task.

5.2 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

Subject	Level	Relevance (High / Low)	Coherence (High / Low)	Diversity (High / Low)
PHYSICS	Basic	**G4** / *Gem*	**Mis** / *Gem*	**G4** / *G3.5*
	Intermediate	**G4** / *Gem*	**Mis** / *Gem*	**Mis** / *L7*
	Advanced	**Mis** / *L13*	**Mis** / *L7*	**Mis** / *L7*
	Wrong	**G4** / *L7*	**G4** / *L7*	**G4** / *L7*
CHEMISTRY	Basic	**G4** / *L13*	**G4** / *Gem*	**Mis** / *L7*
	Intermediate	**G4** / *L13*	**Phi** / *L7*	**Mis** / *G3.5*
	Advanced	**Mis** / *L13*	**Phi** / *Gem*	**Mis** / *G3.5*
	Wrong	**Phi** / *L7*	**G4** / *L7*	**Mis** / *L7*
MATHS	Basic	**Mis** / *G3.5*	**Phi** / *G3.5*	**L13** / *G3.5*
	Intermediate	**G3.5** / *L13*	**G4** / *Gem*	**Mis** / *G3.5*
	Advanced	**G3.5** / *L7*	**Mis** / *Gem*	**Mis** / *L7*
	Wrong	**G4** / *L13*	**G4** / *L7*	**G4** / *G3.5*
GENERAL	All splits	**G4** / *L7*	**G4** / *L7*	**G4** / *L7*

Table 2: Best (bold) and worst (italics) models per metric. G3.5 = GPT-3.5, G4 = GPT-4, Gem = Gemini, L13 = Llama-13B, L70 = Llama-70B, L7 = Llama-7B, Mis = Mistral-8*7B, Phi = Phi-2. GPT-4 and Mistral capture most top slots, while Llama-7B is the predominant laggard, evidencing a clear quality gap across tasks.

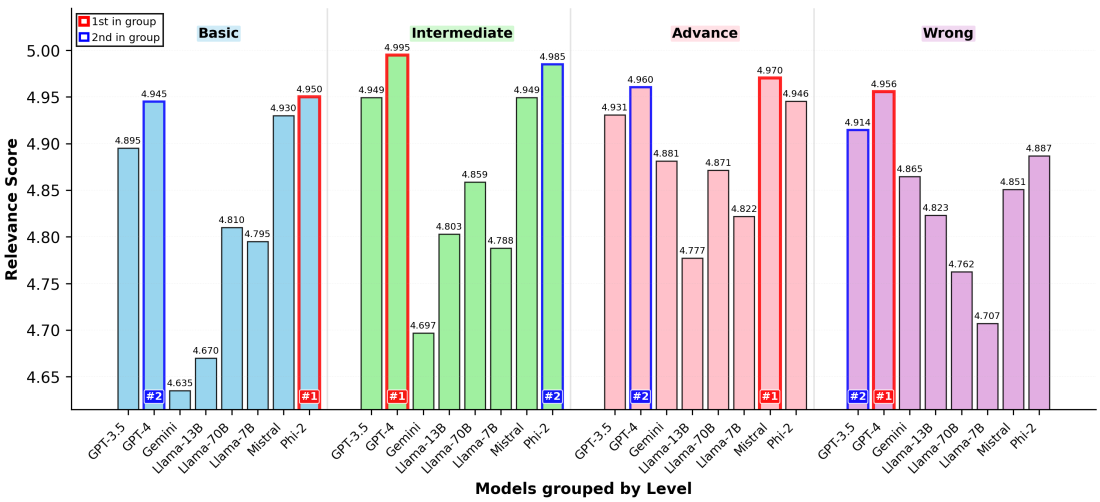

Figure 2: Relevance scores for all models on the Physics dataset, split by **Basic , Intermediate , Advanced , and Wrong** subsets. The red bar marks the best model in each subset, blue bar marks the runner-up. Across all four difficulty bands, either *Phi-2* or *GPT-4* ranks first or second, whereas the Llama series never breaks the top.

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 their frequency of challenging a statement's truth varies. 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.

5.3 Metric Correlations

We analyze model-level correlations between relevance, coherence, and diversity (Figure 4). GPT-4 and GPT-3.5 show strong inter-metric correlations, highlighting a balanced capability. Conversely, Llama models demonstrate weaker correlations, suggesting metric inconsistency and potential specialization trade-offs. The strong positive correlation between relevance and coherence ($r_{\text{Rel-Coh}} \approx 0.72$) is particularly noteworthy, suggesting that for top models, on-topic questions go hand-in-hand with logical structure. Additional scatter plots (Figures 11 and 12 in Appendix) illustrate this performance gap, showing that models like GPT-4 effectively balance focused questioning with breadth while others prioritize one dimension, proving model scale alone does not guarantee this well-rounded ability.

6 Ensuring the validity of CDQG

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

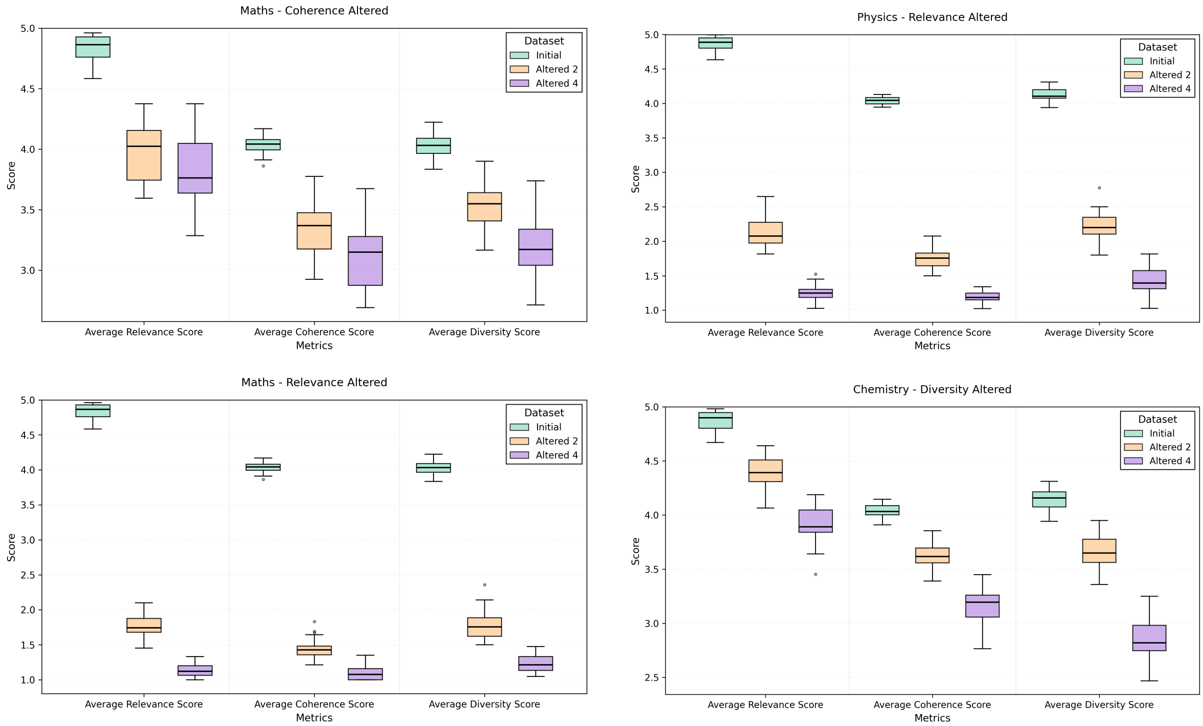

Figure 3: Boxplots showing average scores for relevance, coherence, and diversity across three disciplines **(Physics, Chemistry, Mathematics)**. Each dataset version **("Initial," "2-Altered," "4-Altered")** represents increasing levels of noise introduced into the generated questions.

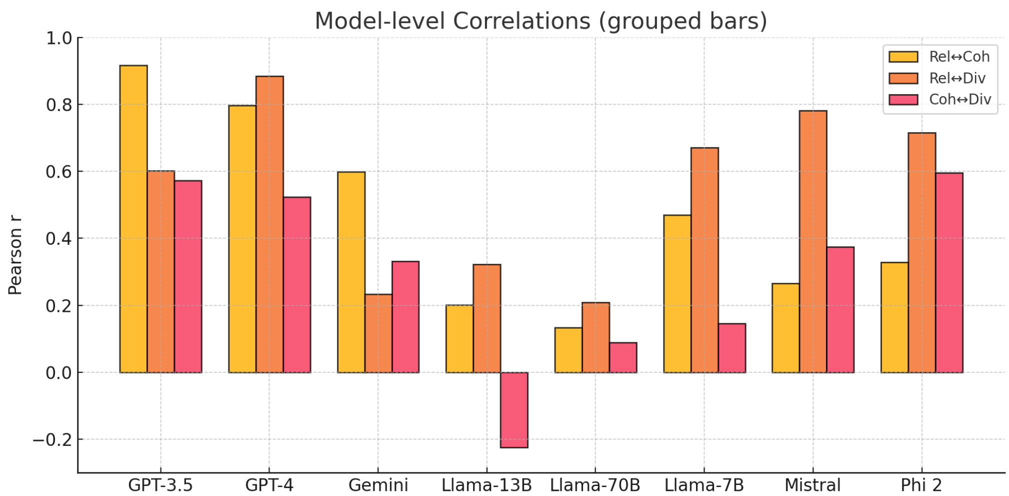

Figure 4: Model-level Pearson correlations among the three metrics—**Relevance ↔ Coherence, Relevance ↔ Diversity, and Coherence ↔ Diversity**—for all evaluated LLMs. *GPT-4* and *GPT-3.5* show the strongest, well-balanced couplings, while Llama variants display weaker links.

6.1 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 A 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 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 As shown in Figure 3 and Figure 9 (in Appendix), added noise consistently degrades scores across all metrics, though the magnitude varies. The relevance metric is most sensitive to noise, with scores dropping sharply from 4.8 to 1.0. In contrast, the coherence metric shows a smaller decrease, as individual question alterations do not always disrupt the logical flow. The reduction in the diversity score is also less pronounced, partly because manipulating this metric is uniquely challenging and requires deep subject matter understanding. Ultimately, these results validate that LLM judges can effectively differentiate between high-quality and noise-compromised content, which supports the robustness of our evaluation framework.

6.2 Human evaluation

To rigorously validate our LLM-based evaluations, we constructed a human evaluation subset by randomly sampling approximately 1,000 data points from our complete dataset. This subset was partitioned into two distinct files of 500 data points each. We then engaged four independent graduate student annotators, assigning each file to a unique pair of raters. This two-by-two design ensured robust, independent judgments, with all annotators following a consistent set of guidelines detailed in Appendix C.1.

We measured the agreement between these human judgments and the Claude model's ratings using Cohen's Kappa (κ). The final scores represent the average agreement calculated across all four human evaluators against the model's corresponding ratings. The analysis revealed a substantial overall agreement for **Relevance** ($\kappa = 0.656$), and moderate agreement for **Coherence** ($\kappa = 0.608$) and **Diversity** ($\kappa = 0.550$). According to established benchmarks, these Kappa values indicate a reliable alignment, validating our LLM-based evaluation.

7 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).

8 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 validated by 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 potential applications to improve the performance and usability of knowledge systems.

9 Limitations

While this study introduces an innovative framework for evaluating the questioning capabilities of LLMs, it primarily utilizes well-defined metrics. Though robust, these metrics do not consider the pragmatic factors in human-like questioning, which lead to different human question types such as clarification questions, knowledge acquisition questions, curiosity-driven questions, etc. Future research could explore the integration of metrics that assess these human-centric qualities to better mimic real-world applications. Additionally, this study only considers one-round questioning, which might not fully reflect the complexities of human-in-the-loop questioning that usually involves multiple rounds.

References

Zahra Abbasiantaeb, Yifei Yuan, Evangelos Kanoulas, and Mohammad Aliannejadi. 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*, pages 8–17, Merida Mexico. ACM.

Selcuk Acar, Kelly Berthiaume, and Rebecca Johnson. 2023. What kind of questions do creative people ask? *Journal of Creativity*, 33(3):100062.

Nischal Ashok Kumar, Nigel Fernandez, Zichao Wang, and Andrew Lan. 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)*, pages 247–259, Toronto, Canada. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Sabina Elkins, Ekaterina Kochmar, Iulian Serban, and Jackie C. K. Cheung. 2023. How Useful Are Educational Questions Generated by Large Language Models? In *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky*, pages 536–542, Cham. Springer Nature Switzerland.

Guglielmo Faggioli, Laura Dietz, Charles L. A. Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, and Henning Wachsmuth. 2023. Perspectives on Large Language Models for Relevance Judgment. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 39–50, Taipei Taiwan. ACM.

Javier Ferrando, Oscar Obeso, Senthooran Rajamanoharan, and Neel Nanda. 2024. Do I Know This Entity? Knowledge Awareness and Hallucinations in Language Models. *arXiv preprint*.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. GPTScore: Evaluate as You Desire. *arXiv preprint*. ArXiv:2302.04166 [cs].

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Gemini Team, Rohan Anil, and et al. 2024. Gemini: A Family of Highly Capable Multimodal Models. *arXiv preprint*. ArXiv:2312.11805 [cs].

Alireza Ghafarollahi and Markus J. Buehler. 2024. Sciagents: Automating scientific discovery through multi-agent intelligent graph reasoning. *Preprint*, arXiv:2409.05556.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks.

Shash Guo, Lizi Liao, Cuiping Li, and Tat-Seng Chua. 2024. A survey on neural question generation: methods, applications, and prospects. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Wang, and Zhiting Hu. 2023. Reasoning with Language Model is Planning with World Model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8154–8173, Singapore. Association for Computational Linguistics.

Qianyu He, Jie Zeng, Wenhao Huang, Lina Chen, Jin Xiao, Qianxi He, Xunzhe Zhou, Jiaqing Liang, and Yanghua Xiao. 2024. Can Large Language Models Understand Real-World Complex Instructions? *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):18188–18196.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Fan Huang, Haewoon Kwak, Kunwoo Park, and Jisun An. 2024. ChatGPT rates natural language explanation quality like humans: But on which scales? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3111–3132, Torino, Italia. ELRA and ICCL.

Yizheng Huang and Jimmy Huang. 2024. A survey on retrieval-augmented text generation for large language models. *Preprint*, arXiv:2404.10981.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of Experts. *arXiv preprint*. ArXiv:2401.04088 [cs].

Dongfu Jiang, Yishan Li, Ge Zhang, Wenhao Huang, Bill Yuchen Lin, and Wenhu Chen. 2023. TIGERScore: Towards Building Explainable Metric for All Text Generation Tasks. *arXiv preprint*. ArXiv:2310.00752 [cs].

Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. 2022. Maieutic prompting: Logically consistent reasoning with recursive explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1266–1279, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Moussa Kamal Eddine, Guokan Shang, Antoine Tixier, and Michalis Vazirgiannis. 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)*, pages 1305–1318, Dublin, Ireland. Association for Computational Linguistics.

Nan Rosemary Ke, Danny P. Sawyer, Hubert Soyer, Martin Engelcke, David P Reichert, Drew A. Hudson, John Reid, Alexander Lerchner, Danilo Jimenez Rezende, Timothy P Lillicrap, Michael Mozer, and Jane X Wang. 2024. Can foundation models actively gather information in interactive environments to test hypotheses? *Preprint*, arXiv:2412.06438.

Tom Kocmi and Christian Federmann. 2023. Large Language Models Are State-of-the-Art Evaluators of Translation Quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Alexander Kotov and ChengXiang Zhai. 2010. Towards natural question guided search. In *Proceedings of the 19th international conference on World wide web*, pages 541–550, Raleigh North Carolina USA. ACM.

Satyapriya Krishna, Kalpesh Krishna, Anhad Mohananey, Steven Schwarcz, Adam Stambler, Shyam Upadhyay, and Manaal Faruqui. 2024. Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation. *Preprint*, arXiv:2409.12941.

Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. 2020. A Systematic Review of Automatic Question Generation for Educational Purposes. *International Journal of Artificial Intelligence in Education*, 30(1):121–204.

Yukyung Lee, Joonghoon Kim, Jaehee Kim, Hyowon Cho, and Pilsung Kang. 2024. CheckEval: Robust Evaluation Framework using Large Language Model via Checklist. *arXiv preprint*. ArXiv:2403.18771 [cs].

Christoph Leiter, Piyawat Lertvittayakumjorn, Marina Fomicheva, Wei Zhao, Yang Gao, and Steffen Eger. 2024. Towards Explainable Evaluation Metrics for Machine Translation. *Journal of Machine Learning Research*, 25(75):1–49.

Christoph Leiter, Juri Opitz, Daniel Deutsch, Yang Gao, Rotem Dror, and Steffen Eger. 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*, pages 117–138, Bali, Indonesia. Association for Computational Linguistics.

Belinda Z. Li, Alex Tamkin, Noah Goodman, and Jacob Andreas. 2023. Eliciting human preferences with language models. *Preprint*, arXiv:2310.11589.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2024. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *Preprint*, arXiv:2411.16594.

Jiwei Li, Alexander H. Miller, Sumit Chopra, Marc'Aurelio Ranzato, and Jason Weston. 2017. Learning through Dialogue Interactions by Asking Questions. *arXiv preprint*. ArXiv:1612.04936 [cs].

Kevin Liu, Stephen Casper, Dylan Hadfield-Menell, and Jacob Andreas. 2024a. Cognitive Dissonance: Why Do Language Model Outputs Disagree with Internal Representations of Truthfulness? In *EMNLP*.

Minqian Liu, Ying Shen, Zhiyang Xu, Yixin Cao, Eunah Cho, Vaibhav Kumar, Reza Ghanadan, and Lifu Huang. 2024b. X-Eval: Generalizable Multi-aspect Text Evaluation via Augmented Instruction Tuning with Auxiliary Evaluation Aspects. *arXiv preprint*. ArXiv:2311.08788 [cs].

Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. 2023. Calibrating LLM-Based Evaluator. *arXiv preprint*. ArXiv:2309.13308 [cs].

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

Evan Lucas, Kelly S. Steelman, Leo C. Ureel, and Charles Wallace. 2024. For those who don't know (how) to ask: Building a dataset of technology questions for digital newcomers. *arXiv preprint*. ArXiv:2403.18125 [cs].

Tula Masterman, Sandi Besen, Mason Sawtell, and Alex Chao. 2024. The Landscape of Emerging AI Agent Architectures for Reasoning, Planning, and Tool Calling: A Survey. *arXiv preprint*. ArXiv:2404.11584 [cs].

Sébastien Bubeck Mojan Javaheripi. 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, Josh Achiam, Steven Adler, and et al. 2024. GPT-4 Technical Report. *arXiv preprint*. ArXiv:2303.08774 [cs].

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 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)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Vinayak Puranik, Anirban Majumder, and Vineet Chaoji. 2023. PROTEGE: Prompt-based diverse question generation from web articles. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5449–5463, Singapore. Association for Computational Linguistics.

Thomas Scialom and Jacopo Staiano. 2020. Ask to learn: A study on curiosity-driven question generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2224–2235, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Preprint*, arXiv:2303.11366.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2024. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *Preprint*, arXiv:2409.04109.

Md Arafat Sultan, Shubham Chandel, Ramón Fernandez Astudillo, and Vittorio Castelli. 2020. On the importance of diversity in question generation for QA. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5651–5656, Online. Association for Computational Linguistics.

Bin Sun, Yitong Li, Fei Mi, Weichao Wang, Yiwei Li, and Kan Li. 2023. Towards diverse, relevant and coherent open-domain dialogue generation via hybrid latent variables. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press.

Mirac Suzgun, Tayfun Gur, Federico Bianchi, Daniel E. Ho, Thomas Icard, Dan Jurafsky, and James Zou. 2024. Belief in the Machine: Investigating Epistemological Blind Spots of Language Models. *arXiv preprint*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint*. ArXiv:2307.09288 [cs].

Karthik Valmeekam, Matthew Marquez, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. 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.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. 2024.

Replacing Judges with Juries: Evaluating LLM Generations with a Panel of Diverse Models. *arXiv preprint*. ArXiv:2404.18796 [cs].

Manya Wadhwa, Xinyu Zhao, Junyi Jessy Li, and Greg Durrett. 2024. Learning to refine with fine-grained natural language feedback. *Preprint*, arXiv:2407.02397.

Haonan Wang, James Zou, Michael Mozer, Anirudh Goyal, Alex Lamb, Linjun Zhang, Weijie J. Su, Zhun Deng, Michael Qizhe Xie, Hannah Brown, and Kenji Kawaguchi. 2024a. Can AI Be as Creative as Humans? *arXiv preprint*. ArXiv:2401.01623 [cs].

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is ChatGPT a Good NLG Evaluator? A Preliminary Study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11, Hybrid. Association for Computational Linguistics.

Shuliang Wang, Dapeng Li, Jing Geng, Longxing Yang, and Hongyong Leng. 2020. Learning to balance the coherence and diversity of response generation in generation-based chatbots. *International Journal of Advanced Robotic Systems*, 17:172988142095300.

Wenxuan Wang, Juluan Shi, Chaozheng Wang, Cheryl Lee, Youliang Yuan, Jen tse Huang, and Michael R. Lyu. 2024b. Learning to Ask: When LLMs Meet Unclear Instruction. *Preprint*, arXiv:2409.00557.

Yating Wu, Ritika Mangla, Alexandros G. Dimakis, Greg Durrett, and Junyi Jessy Li. 2024. Which questions should i answer? salience prediction of inquisitive questions. *Preprint*, arXiv:2404.10917.

Siheng Xiong, Ali Payani, Ramana Kompella, and Faramarz Fekri. 2024. Large Language Models Can Learn Temporal Reasoning. *arXiv preprint*. ArXiv:2401.06853 [cs].

Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Wang, and Lei Li. 2023. INSTRUCTSCORE: Towards Explainable Text Generation Evaluation with Automatic Feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5967–5994, Singapore. Association for Computational Linguistics.

Bingsheng Yao, Dakuo Wang, Tongshuang Wu, Zheng Zhang, Toby Li, Mo Yu, and Ying Xu. 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)*, pages 731–744, Dublin, Ireland. Association for Computational Linguistics.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Eric Zelikman, Georges Harik, Yijia Shao, Varuna Jayasiri, Nick Haber, and Noah D. Goodman. 2024. Quiet-STaR: Language Models Can Teach Themselves to Think Before Speaking. *arXiv preprint*. ArXiv:2403.09629 [cs].

Wei Zhao, Michael Strube, and Steffen Eger. 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*, pages 3865–3883, Dubrovnik, Croatia. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a Unified Multi-Dimensional Evaluator for Text Generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yuqi Zhu, Xiaohan Wang, Jing Chen, Shuofei Qiao, Yixin Ou, Yunzhi Yao, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2024. LLMs for Knowledge Graph Construction and Reasoning: Recent Capabilities and Future Opportunities. *arXiv preprint*.

A 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.

Collection of Physics Statements for Curiosity (Synthetic Data Collection)

Prompt: *"I want to create a collection of physics statements that naturally provoke curiosity and make people think. These statements should be intriguing enough that when someone reads them, they automatically want to ask 'why' or 'how' questions."*

Requirements:

- Make readers naturally curious
- Encourage critical thinking
- Lead to deeper questions
- Cover different levels of physics (basic, intermediate, advanced)
- Don't include the questions themselves - just the thought-provoking statements that would generate questions in someone's mind.

B Example of Question Generation and Evaluation

Initial Statement: "Organic chemistry focuses on compounds containing carbon."

Prompt Given to Models

Imagine you are a human, this is the first time you are coming across this Chemistry Statement, you have no previous knowledge of it "Organic chemistry focuses on compounds containing carbon.", what are the top 5 questions that would pop up in your head which would be most useful in learning about it as you are new to it. Give me a simple bullet point list, don't explain them or expand them.

Model Output (13B)

- What exactly is carbon, and why is it so important in organic chemistry?
- How do organic compounds differ from inorganic compounds, and what are some common types of organic compounds?
- What is the difference between saturated and unsaturated hydrocarbons, and how do they behave differently in chemical reactions?
- How do functional groups affect the properties and behavior of organic compounds?

Evaluation Results

Evaluator	Relevance	Coherence	Diversity	Justification
Mistral	5/5	4/5	4/5	Questions follow logical progression, moving from basic concepts to more specific topics. All questions highly relevant to initial statement.
Gemini	4/5	4/5	3/5	Questions directly related to statement, form coherent line of inquiry. Cover range of aspects but could explore more applications.
GPT-4	5/5	4/5	5/5	Questions directly pertain to concepts, logical progression but incomplete final question. Wide range of topics covered.

Table 3: Combined Evaluation Scores and Justifications

Metric	Score	Description
Relevance	4.5/5	Questions strongly aligned with core concept
Coherence	4/5	Good logical flow from basic to advanced concepts
Diversity	4/5	Balanced coverage of different aspects
Human Validation Scores		
Relevance	3/5	Human evaluators consistently rated the questions lower than AI evaluators, suggesting potential gaps between AI and human assessment criteria.
Coherence	3/5	
Diversity	3/5	

Table 4: Final Claude Evaluation (Meta-evaluator) Results and Human Evaluations

Instruction/Prompt for Chemistry Advance

Imagine you are a human, this is the first time you are coming across this Chemistry Statement, you have no previous knowledge of it *"Zeolites are microporous, aluminosilicate minerals used as commercial adsorbents and catalysts"*. What are the top 5 questions that would pop up in your head which would be most useful in learning about it as you are new to it? Give me a simple bullet point list, don't explain them or expand them.

Model Outputs

Model	Generated Questions
Llama 13B	• What exactly are zeolites, and where do they come from? • How do zeolites' microporous structure and composition affect their adsorption and catalytic properties? • What types of applications use zeolites as commercial adsorbents and catalysts? • Can zeolites be modified or tailored for specific uses? • Are there any potential drawbacks or limitations to using zeolites?
Llama 70B	• What are microporous materials? • What are aluminosilicate minerals? • What does it mean for a material to be used as a commercial adsorbent? • What are some examples of how zeolites are used as catalysts? • How do zeolites compare to other materials used in similar applications?
Llama 7B	• What are zeolites? • What are micropores? • What is an aluminosilicate? • How are zeolites used as commercial adsorbents? • What are catalysts and how do zeolites act as catalysts?
Mistral	• What are zeolites made of and how do they form their unique structure? • How do zeolites exhibit microporosity and what does it mean for their properties? • In what ways can zeolites be used as commercial adsorbents? • How do zeolites function as catalysts and what types of reactions do they catalyze? • What gives zeolites their ability to act as both adsorbents and catalysts?
Phi-2	• What are zeolites? • How do they work as adsorbents and catalysts? • What are their applications in industry? • What makes zeolites unique compared to other materials? • Are there any potential environmental concerns with using zeolites?
Gemini	• What are microporous minerals? • Why are zeolites aluminosilicate? • What does it mean for zeolites to be used as adsorbents? • How do zeolites function as catalysts? • What are the commercial applications of zeolites?

Continued on next page

Model	Generated Questions
GPT-3.5	• What are zeolites? • How are zeolites used as adsorbents? • How are zeolites used as catalysts? • What makes zeolites microporous? • What are the practical applications of zeolites?
GPT-4	• What are Zeolites and what is their fundamental structure? • What does microporous mean and why is it significant for Zeolites? • What is meant by aluminosilicate minerals? • How are Zeolites used as commercial adsorbents and why are they suitable for this? • How do Zeolites act as catalysts, and in what kind of reactions or processes are they commonly used?

Table 5: Questions Generated by Different Models

Instruction/Prompt, Chemistry Basic

> Imagine you are a human, this is the first time you are coming across this Chemistry Statement, you have no previous knowledge of it *"Matter is composed of atoms"*. What are the top 5 questions that would pop up in your head which would be most useful in learning about it as you are new to it? Give me a simple bullet point list, don't explain them or expand them.

Model Outputs

Model	Generated Questions
Llama 13B	• What exactly is an atom? • How do atoms combine to form different substances? • Is everything in the world made up of atoms? • How can we see atoms if they are too small to observe directly? • Can atoms be broken down into smaller parts, and if so, what are those parts called?
Llama 70B	• What are atoms? • How do atoms combine to form matter? • What are some examples of matter that are composed of atoms? • How small are atoms compared to everyday objects? • Can atoms be broken down further into smaller parts?

Continued on next page

Model	Generated Questions
Llama 7B	• What is an atom? • How do atoms make up matter? • Are there any other building blocks of matter besides atoms? • How can I see or observe atoms? • What are some examples of different types of atoms and how do they differ from each other?
Mistral	• What is an atom? • How are atoms arranged in matter? • Are all atoms the same or do they differ from each other? • Can matter be broken down into anything smaller than an atom? • How do atoms combine to form different substances?
Phi-2	• What is matter? • What are atoms and how do they make up matter? • How many types of atoms exist? • Can atoms combine to form molecules? • How does the arrangement of atoms affect the properties of matter?
Gemini	• What are atoms and how small are they? • How do atoms combine to form larger substances? • What determines the properties of different substances? • Can atoms be created or destroyed? • How does the concept of atoms advance our understanding of the world?
GPT-3.5	• What is matter? • What are atoms? • How are atoms related to matter? • How do atoms combine to form matter? • Can atoms be broken down further into smaller particles?
GPT-4	• What is matter? • What is an atom? • How are atoms structured? • How do atoms compose matter? • What are some examples of different types of atoms?

Table 6: Questions Generated by Different Models

Instruction/Prompt, Maths Intermediate

> Imagine you are a human, this is the first time you are coming across this Mathematics Statement, you have no previous knowledge of it *"$log_a(x * y) = log_a(x) + log_a(y)$"*. What are the top 5 questions that would pop up in your head which would be most useful in learning about it as you are new to it? Give me a simple bullet point list, don't explain them or expand them.

Model Outputs

Model	**Generated Questions**
Llama 13B	• What does the symbol "log" represent? • What is the meaning of "$\log_a(x)$" and "$\log_a(y)$" in this equation? • How do I calculate the logarithm of a product like x * y? • Is there a way to simplify this equation or make it more intuitive?
Llama 70B	• What does the "$\log_a$" symbol represent? • What is the purpose of the equation? • How do the variables x and y relate to each other in the equation? • Is there a specific context or application where this equation is commonly used? • Can the equation be simplified or rewritten in a different form?
Llama 7B	• What does the symbol "log" mean? • How do I interpret the variables "x" and "y"? • Why does the equation say "= $\log_a(x) + \log_a(y)$" instead of just "= x + y"? • Is there a way to simplify or break down the equation into smaller parts? • Can I use any real-world examples or applications to help me understand the equation better?
Mistral	• What does the "log" function represent and how is it defined? • What does the subscript "a" mean in the notation "$\log_a(x)$"? • How is the equation "$\log_a(x * y) = \log_a(x) + \log_a(y)$" derived or proven? • What are the conditions or restrictions on the variables x and y for the equation to hold true? • Can the equation be generalized or extended to more than two variables?
Phi-2	• What does the symbol 'log' stand for? • How do I interpret the equation $\log_a(x * y)$? • Can you give an example of how to use this equation? • Are there any special rules or properties associated with logarithms? • How can I simplify expressions involving logarithms?

Continued on next page

Model	Generated Questions
Gemini	• What is the exponent symbol and what does it represent? • How does exponentiation affect the number being raised? • What is the base number and how does it relate to the result? • Why is the answer 8 instead of a different number? • How can I use this statement to solve similar problems in the future?
GPT-3.5	• What does "log" mean? • What does the subscript "a" represent in "$\log_a$"? • What is the purpose of "$\log_a(x * y)$"? • How does the equation relate to multiplication? • How can I apply this equation to solve mathematical problems?
GPT-4	• What does the $\log_a$ notation signify? • What is 'a' in the function? • What does x and y represent? • How does multiplication in the log argument relate to addition? • What are some of the practical applications of this formula?

Table 7: Questions Generated by Different Models

Instruction/Prompt, Maths Intermediate 2-Altered (Diversity)

Imagine you are a human, this is the first time you are coming across this Mathematics Statement, you have no previous knowledge of it *"The diagonal of a cube is $\sqrt{3}$ times the length of a side"*. What are the top 5 questions that would pop up in your head which would be most useful in learning about it as you are new to it? Give me a simple bullet point list, don't explain them or expand them.

Model Outputs

Model	Generated Questions
Mistral	• What is a cube and how is it defined mathematically? • What is a cube and what is its volume? • How is the diagonal of a cube related to the length of a side? • What is the formula for finding the diagonal of a cube? • How can I prove or disprove this statement?

Continued on next page

Model	Generated Questions
Llama 70B	• What is a cube? • What is a cube made up of? • What does it mean to say that the diagonal of a cube is $\sqrt{3}$ times the length of a side? • What is the formula for finding the diagonal of a cube? • How can I visualize this equation?
Phi-2	• What does ’diagonal’ mean? • What other meaning does ’diagonal’ have in geometry? • What is the formula for finding the diagonal of a cube? • Can you give another formula related to the diagonal? • Are there any other properties or formulas related to cubes that I should know about?
GPT-4	• What is a cube in terms of geometry? • What does a cube look like? • What does ’diagonal of a cube’ mean? • How is the length of the side of a cube defined? • What does the term ’$\sqrt{3}$’ mean and how is it calculated?
Llama 7B	• What is a cube? • How do we represent a cube geometrically? • How do I calculate the length of a side of a cube? • What units are used to measure the side of a cube? • How can I use this equation to solve problems involving cubes and their dimensions?
Llama 13B	• What is a cube? • Can you give a definition of a cube? • How do I find the length of a side of a cube? • Can you explain how to measure a side of a cube? • How do I use this equation to solve real-world problems?
Gemini	• What is a cube? • What is a cube from a mathematical perspective? • What does ’times the length of a side’ mean? • What does ’times the length of a side’ mean in geometry? • Can I apply this formula to cubes of different sizes?

Continued on next page

Model	Generated Questions
GPT-3.5	• What is a cube? • What is a diagonal? • What is a cube made of? • How do you calculate the length of a side? • What are some properties of cubes?

Table 8: Questions Generated by Different Models

Instruction/Prompt, Physics Advance 4-Altered (Relevance)

> Imagine you are a human, this is the first time you are coming across this Physics Statement, you have no previous knowledge of it *"Black hole thermodynamics links the laws of thermodynamics to black holes, suggesting that black holes have entropy proportional to their event horizon area and temperature related to their surface gravity"*. What are the top 5 questions that would pop up in your head which would be most useful in learning about it as you are new to it? Give me a simple bullet point list, don't explain them or expand them.

Model Outputs

Model	Generated Questions
Llama 7B	• What is black hole at the first place? • What is the color of black holes? • Can we travel in time using black holes? • Are there any black holes in our solar system?
Phi-2	• What is a neutron star? • How does hydrodynamics relate to water bodies? • What is weather change and how is it related to global warming? • What is photosynthesis? • How does geological activity affect the Earth's surface?
Llama 70B	• When was the black hole discovered? • Can we travel to black holes? • What was the first film about space? • Is the black hole considered to be a form of dark matter?
GPT-4	• What is the scientific study of black bears? • What is the meaning of entropy in the context of digital communication? • Can you explain the process of measuring the height of Mount Everest? • What are some laws of thermodynamics implicate in a home microwave?

Continued on next page

Model	Generated Questions
Llama 13B	• Can you explain the history of astronomy? • What's the contribution of Newton to physics? • How is the solar system formed? • What is the relation between string theory and quantum physics?
GPT-3.5	• What is the eating habit of a black panther? • How do we measure the event horizon area in a desert? • What is the temperature of a boiling egg? • How is gravity related in water buoyancy?
Mistral	• What are the different types of fruit? • Can a black hole have a favorite color? • How far is it to the nearest galaxy? • What's the capital of France?
Gemini	• What is astrobiology? • How does a plant photosynthesize? • Who was Albert Einstein and what are his contributions? • What are the characteristics of a white dwarf star?

Table 9: Questions Generated by Different Models

C Human Annotations and Validations

C.1 Human Evaluation of Difficulty Levels

To ensure an unbiased and robust validation of our difficulty levels, we implemented a rigorous human annotation protocol. The methodology was designed to mitigate potential biases and produce reliable agreement scores.

First, all statements from the three academic subjects (Physics, Chemistry, and Maths) were combined into a single dataset. This dataset was then randomly shuffled and partitioned into three distinct, non-overlapping files. Each file contained a balanced and mixed-subject distribution of statements.

We then engaged three university students as independent annotators, assigning one file to each. The annotators were provided with a detailed set of guidelines (as outlined in Appendix X) to score the difficulty of each statement. This blind, non-overlapping distribution ensured that each annotator evaluated a unique set of statements without being influenced by the judgments of others.

Finally, to measure the consistency of our original difficulty labels, we calculated the pairwise Cohen's Kappa (κ) between each annotator's ratings and the ground-truth labels. The individual agreement scores were as follows:

- Annotator 1 vs. Reference: $\kappa = 0.695$
- Annotator 2 vs. Reference: $\kappa = 0.712$
- Annotator 3 vs. Reference: $\kappa = 0.511$

The average of these scores provides a final, substantial inter-annotator agreement of $\kappa = 0.639$.

Annotation Guidelines: Classifying Statement Difficulty

1. Your Task & Objective

Your goal is to classify a list of scientific statements into one of three difficulty levels: **Basic**, **Intermediate**, or **Advanced**.
The objective is consistency. We are not grading the statements for correctness or style, but rather for their conceptual depth. To ensure everyone rates consistently, please follow this rubric closely.

2. The 3-Level Difficulty Rubric

This rubric is aligned with Bloom's Taxonomy to provide a standard educational framework for our classifications.

Label to Write	Bloom Cognitive Band	What the Statement is Doing	Fast Clues & Verbs
Basic	Remember & Understand	- States what something is (definition, property) - Describes everyday cause & effect in plain words	is, are, has, called, forms, shows, causes, occurs, appears
Intermediate	Apply & Analyze	- Gives an explicit formula or equation - Cites a topic-specific law or method - Describes a single-step mechanism or lab technique	calculate, relates, determines, depends on, increases, separates
Advanced	Evaluate & Create	- Mentions a field-level theorem, principle, or theory - Discusses a frontier/cutting-edge research topic	prove, generalize, model, optimize, predict broadly, govern

†Bloom levels are used only as anchors to ground the task; you are not grading students, just classifying statements.

3. One-Minute Decision Checklist

Use this quick checklist to make a fast and consistent decision.

Advanced Test:

Does the statement name a major theorem, principle, or theory or an obvious frontier research term?

If YES → Advanced.

Formula / Named-Law Test:

If not Advanced, does it contain a math symbol or cite a topic-specific quantitative law or lab tool?

If YES → Intermediate.

Default to Basic:

If the answer to both of the above is NO, classify it as **Basic**.

Golden Rule: If you are torn between two levels, choose the lower one unless an Advanced keyword is clearly present.

Human Annotation Guidelines for the Generated Questions

Overview

Your task is to evaluate the quality of answers generated by a Large Language Model (LLM). For each given **Instruction** and its corresponding **LLM Answer**, you will provide a score based on three key criteria: Relevance, Coherence, and Diversity. Your careful evaluation is crucial for helping us understand the model's performance.

Your Core Task

For each item, please provide a numeric score from **1 (worst) to 5 (best)** for the three metrics defined below.

Detailed Scoring Rubric

1. Relevance

How well does the response satisfy the user's instruction?

- **5 (Completely Relevant):** The response directly and fully addresses all parts of the instruction.
- **4 (Mostly Relevant):** The response addresses the main point but misses a minor part or includes slightly irrelevant information.
- **3 (Somewhat Relevant):** The response is on the general topic but fails to address a key part of the instruction.
- **2 (Mostly Irrelevant):** The response is related to keywords in the instruction but completely misses the point.
- **1 (Completely Irrelevant):** The response is off-topic and has no connection to the instruction.

2. Coherence

How logical, well-structured, and easy to understand is the answer?

- **5 (Completely Coherent):** The response is well-structured, logical, and flows smoothly. It is grammatically correct and easy to read.
- **4 (Mostly Coherent):** The response is generally understandable but may have minor issues, like an awkward transition or a confusing sentence.
- **3 (Somewhat Coherent):** The response contains some logical connections, but there are significant gaps or contradictions that make it difficult to follow.
- **2 (Mostly Incoherent):** The response is a jumble of related ideas with no clear logical structure.
- **1 (Completely Incoherent):** The response is nonsensical and impossible to understand.

3. Diversity

How varied is the vocabulary and sentence structure? Does it avoid significant repetition?

- **5 (Excellent Diversity):** The response uses a wide range of vocabulary and varied sentence structures.
- **4 (Good Diversity):** The response shows some variety but may have minor instances of repetition.
- **3 (Moderate Diversity):** The response is functional but relies on a limited vocabulary and simple sentence structures.
- **2 (Low Diversity):** The response is highly repetitive, using the same words and sentence patterns frequently.
- **1 (No Diversity):** The response repeats the exact same phrase or structure to an extreme degree.

IMPORTANT: How to Record Your Scores

Please enter only the numeric score (1-5) for each metric in the corresponding column of your spreadsheet.

D Plots

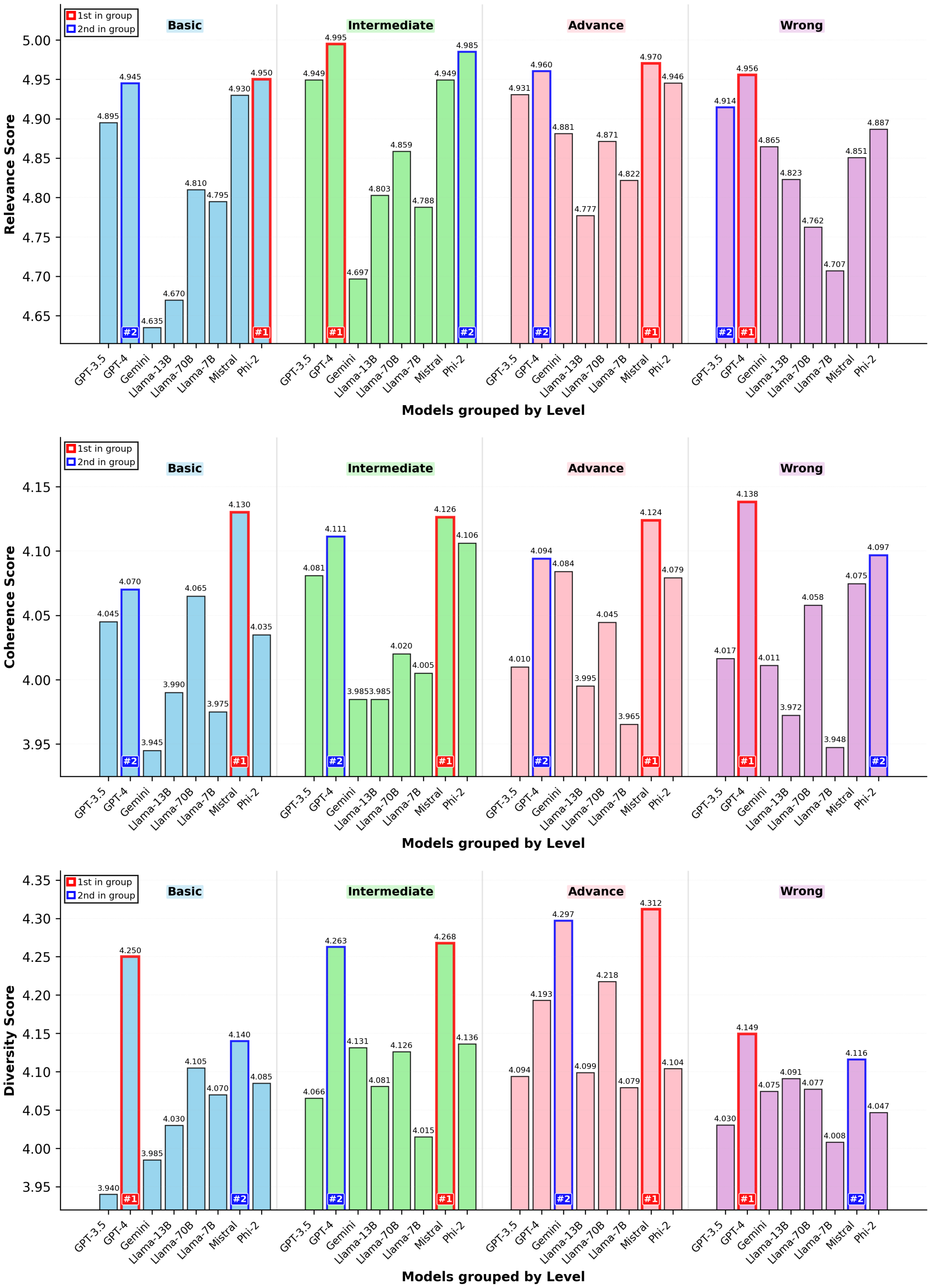

Figure 5: **Metric scores on Physics:** The set of bar charts provides a multidimensional analysis of various models, evaluated by three key performance metrics — Relevance, Coherence , and Diversity. Each chart contrasts the scores across Advanced , Basic , and Intermediate expertise levels for maths, 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.

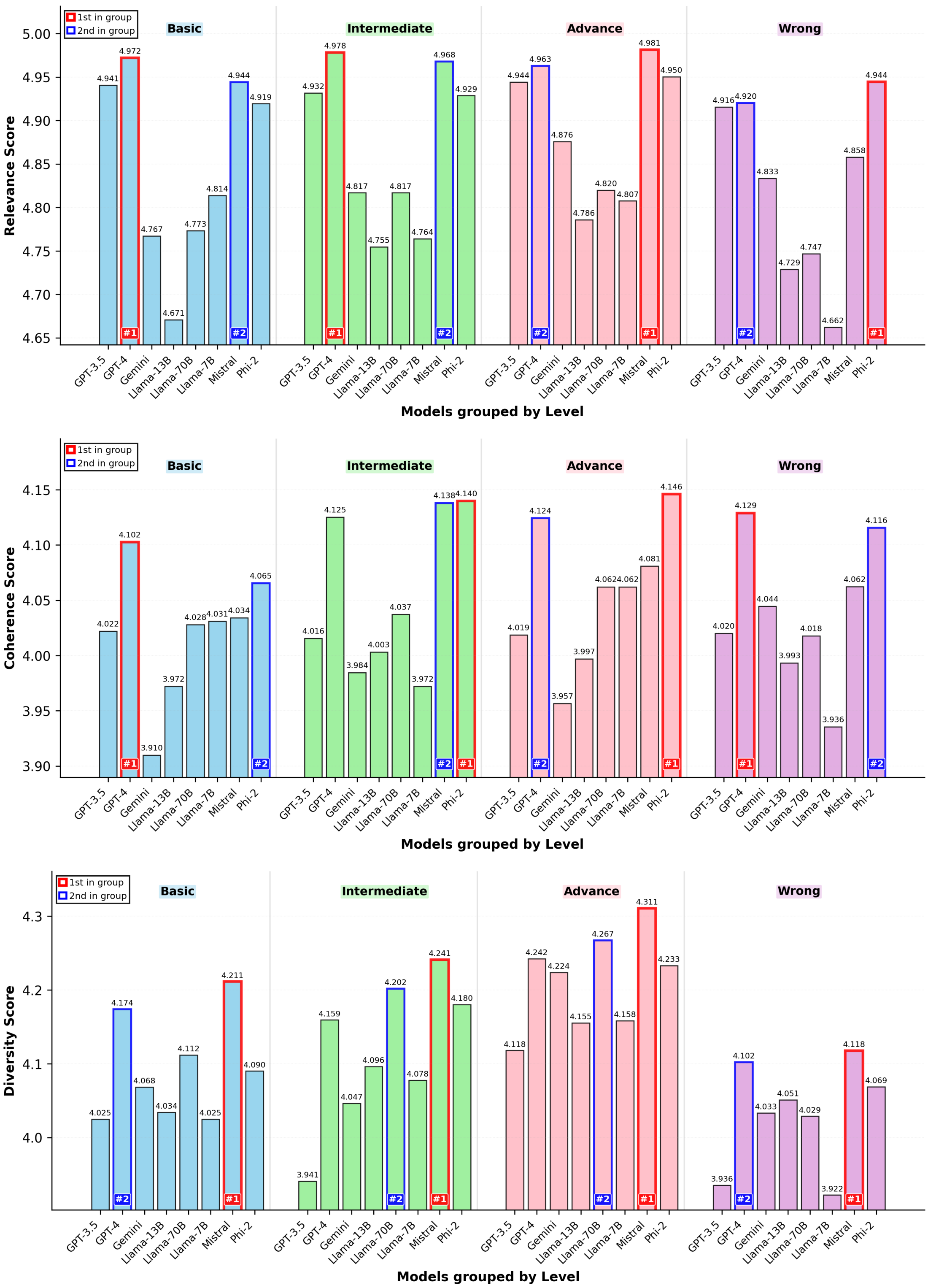

Figure 6: **Metric scores on Chemistry:** The set of bar charts provides a multidimensional analysis of various models, evaluated by three key performance metrics — Relevance, Coherence , and Diversity. Each chart contrasts the scores across Advanced , Basic , and Intermediate expertise levels for Chemistry, 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.

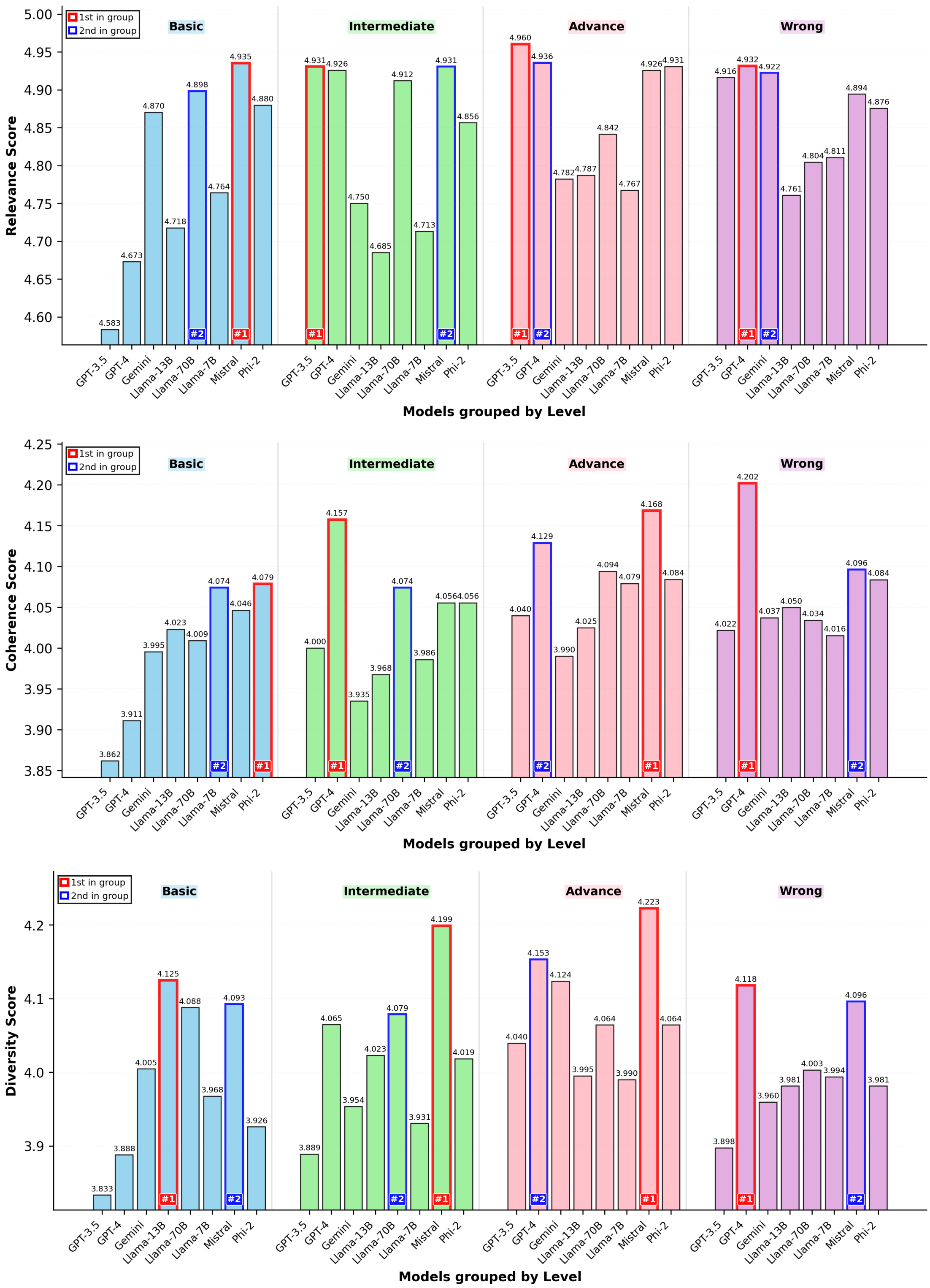

Figure 7: **Metric scores on Maths:** The set of bar charts provides a multidimensional analysis of various models, evaluated by three key performance metrics — Relevance, Coherence , and Diversity. Each chart contrasts the scores across Advanced, Basic, and Intermediate expertise levels for Chemistry, 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.

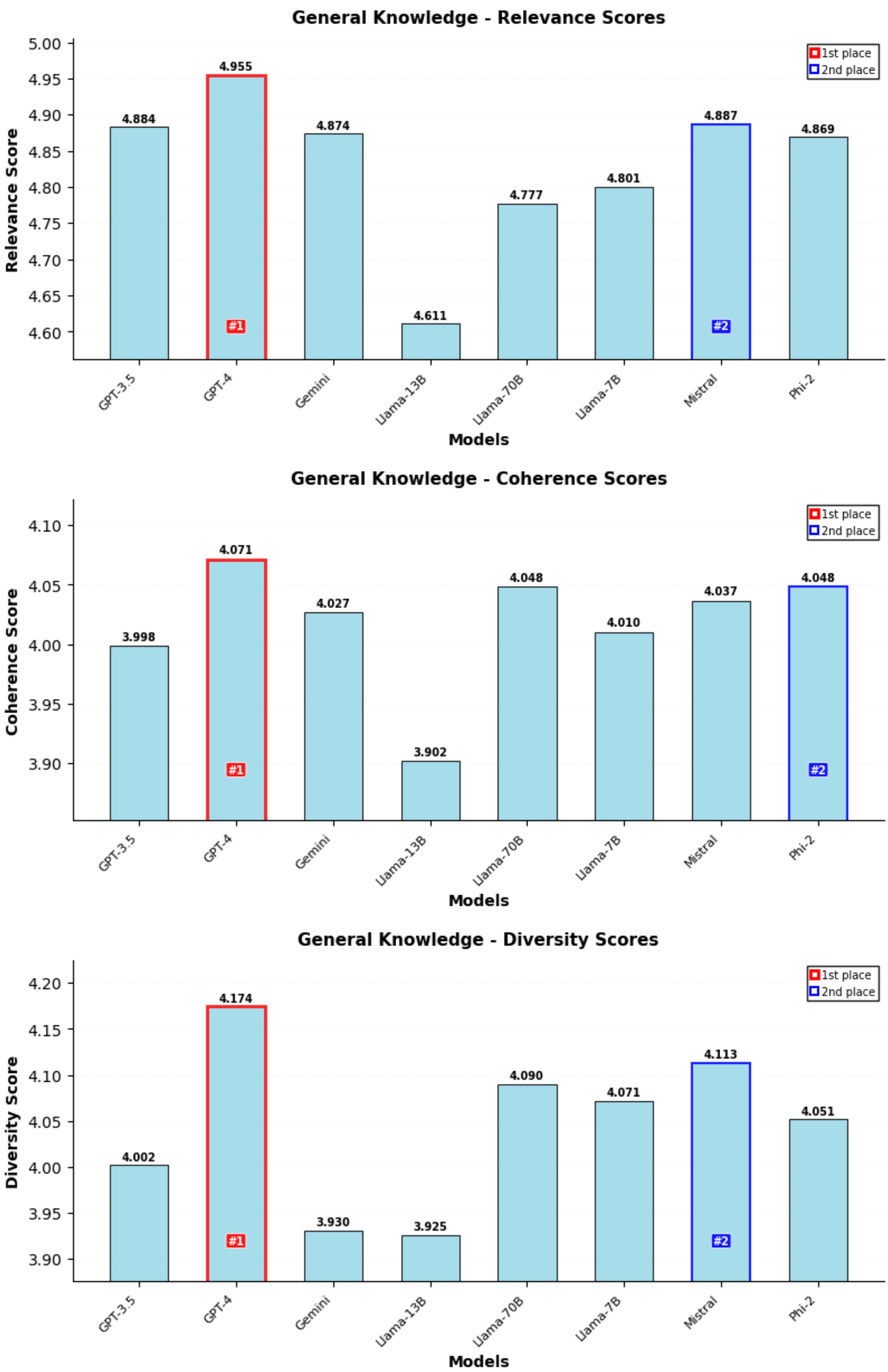

Figure 8: **Metric scores on General Statements:** The set of bar charts provides a multidimensional analysis of various models, evaluated by three key performance metrics — Relevance, Coherence, and Diversity. Each chart contrasts the scores across Relevance, Coherence, and Diversity. Highlighted bars denote the top and second-highest scoring models within each metric, offering a visual synopsis of comparisons.

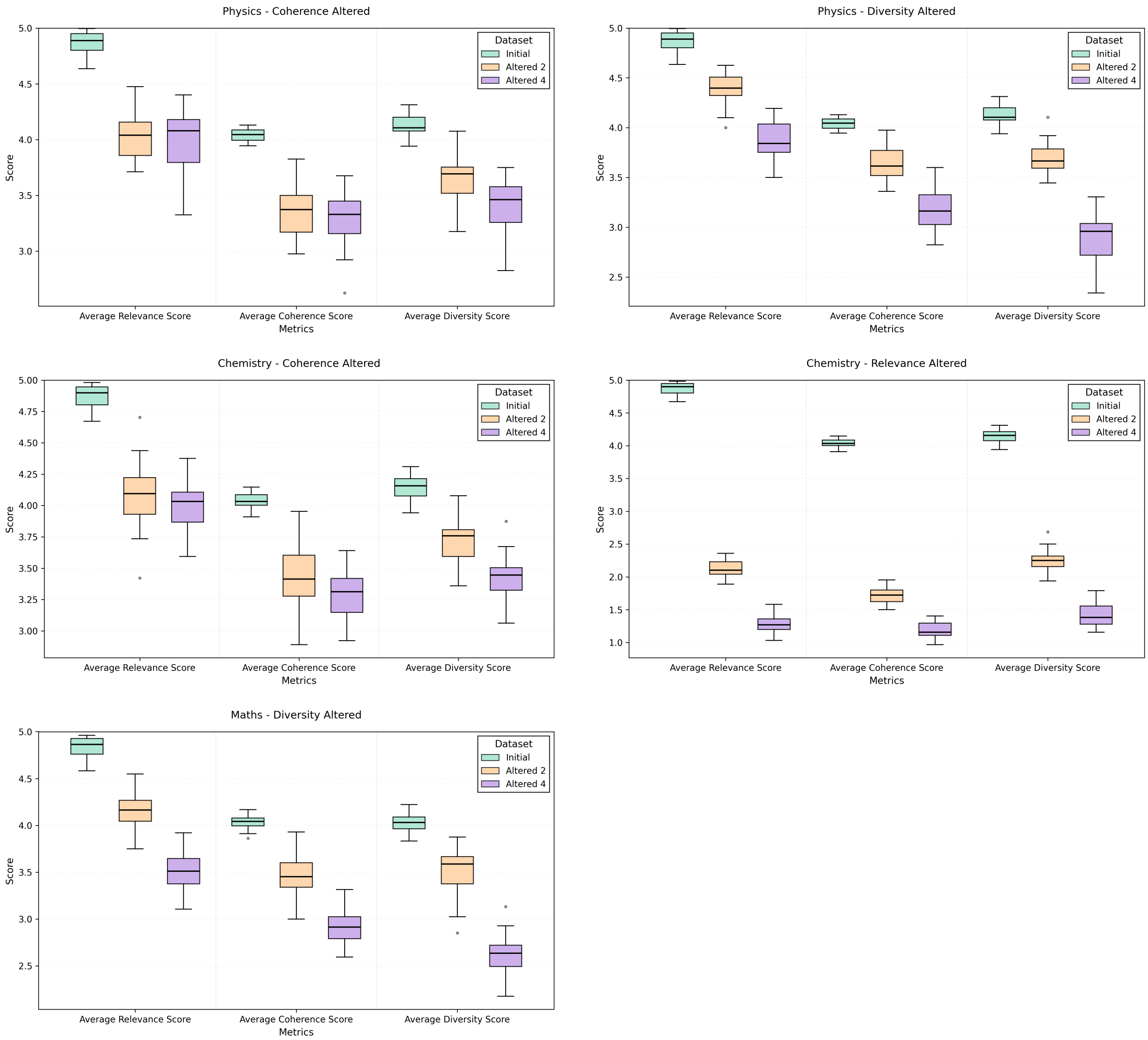

Figure 9: Boxplots showing average relevance, coherence, and diversity scores across altered datasets in Physics (first row), Chemistry (second row), and Maths (third row). Each includes **"Initial", "2-Altered", and "4-Altered"** dataset versions.

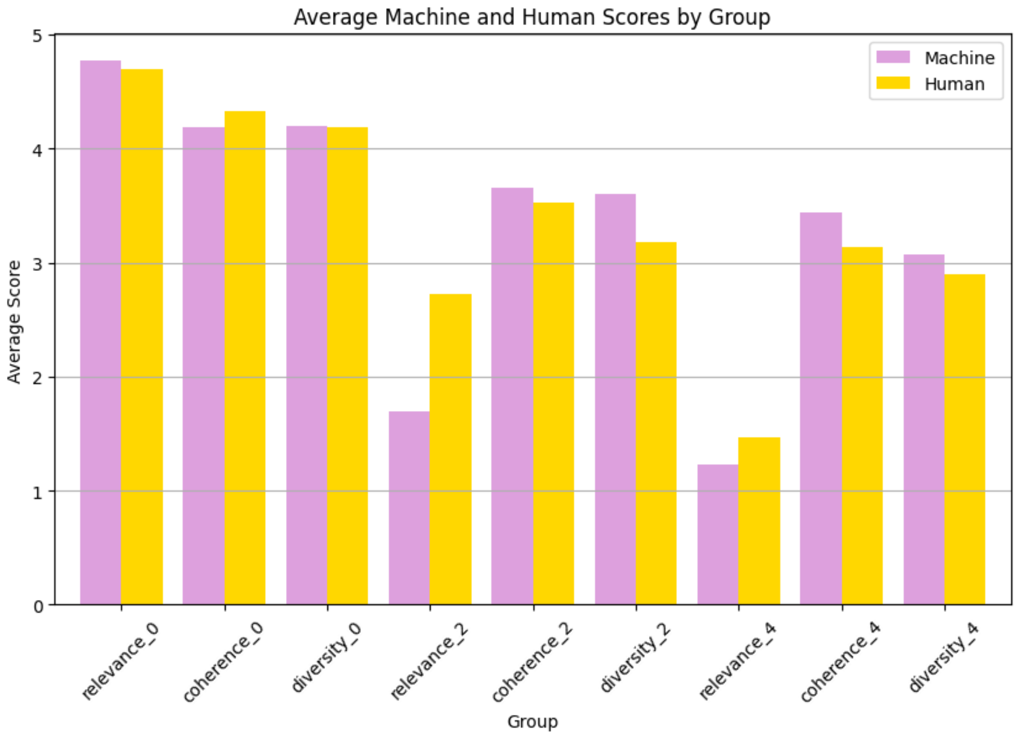

Figure 10: The bar chart compares average scores for machine and human evaluations across different evaluation metrics with alterations. “0-Altered”, “2-Altered” and “4-Altered”.

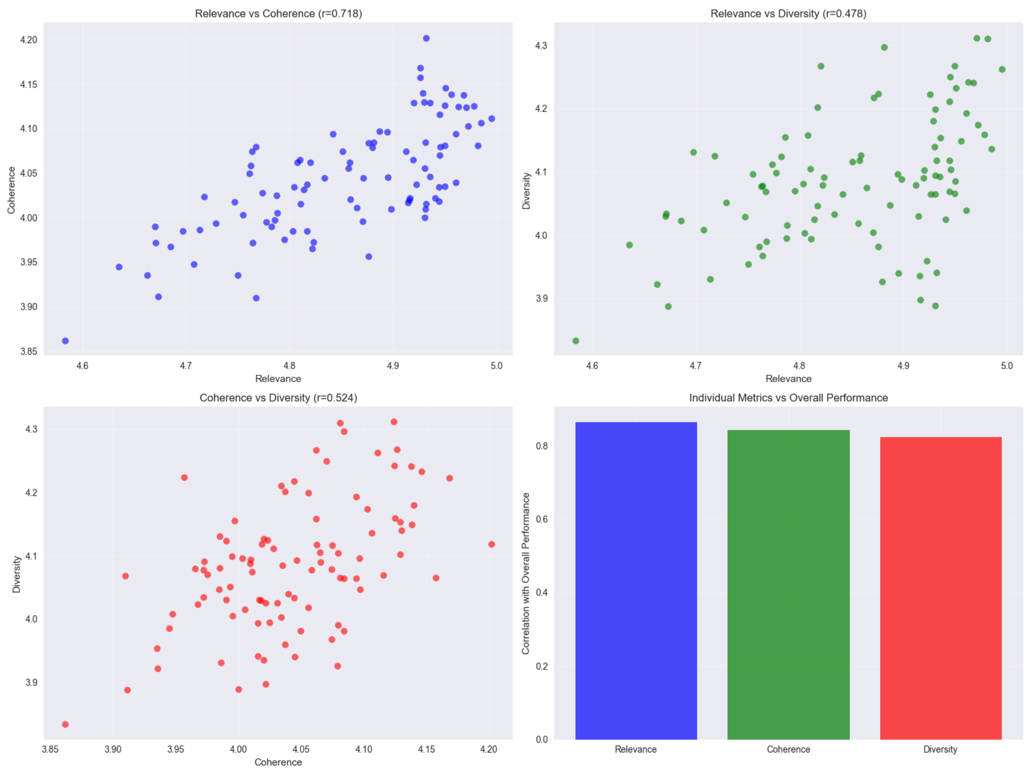

Figure 11: Pairwise correlations among the three curiosity metrics **(Relevance–Coherence, Relevance–Diversity, Coherence–Diversity)** and each metric’s correlation with the overall score. Relevance is the strongest single predictor of overall quality, yet the moderate cross-metric r values confirm that Coherence and Diversity **contribute complementary information**.

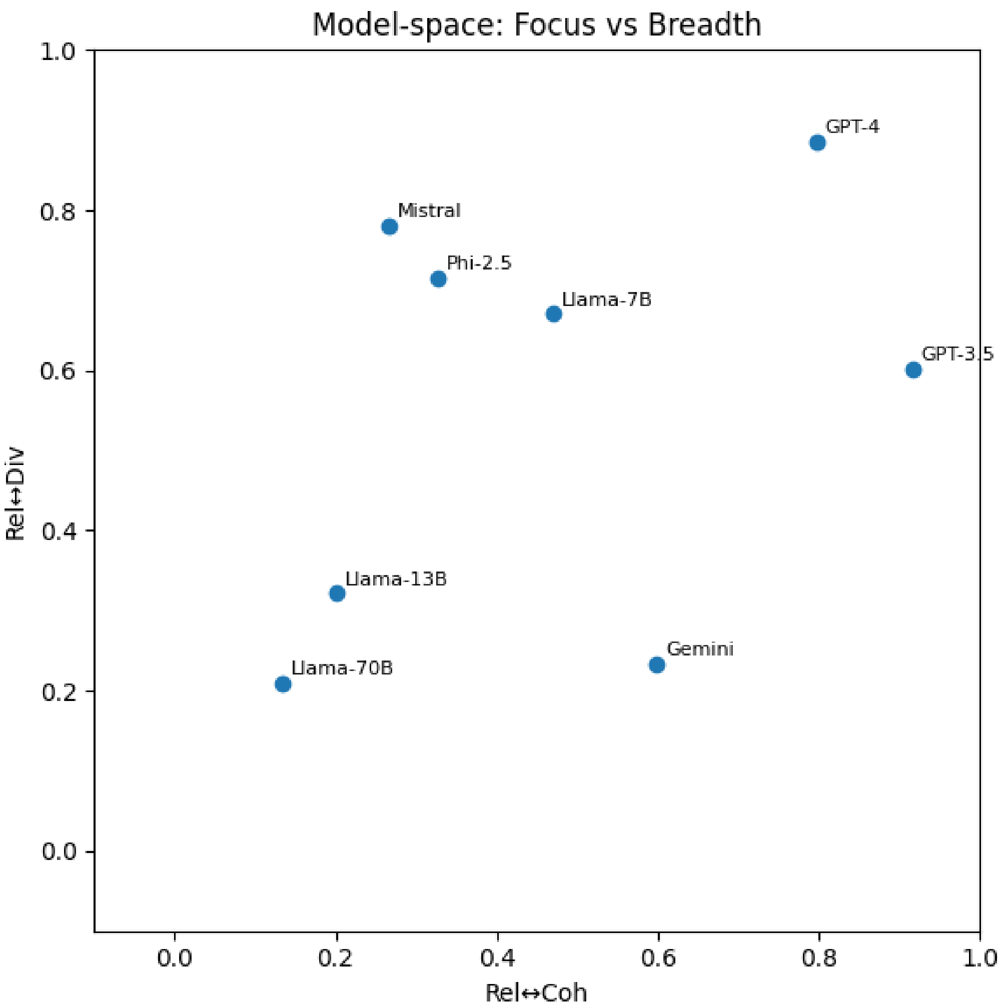

Figure 12: **Model-space map:** each point shows a model's correlation between Relevance–Coherence (focus, x-axis) and Relevance–Diversity (breadth, y-axis). GPT-4 occupies the upper-right corner—balancing focus and breadth—while Llama variants cluster low, indicating weaker alignment independent of parameter count.

E 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.