BloomXplain: A Framework and Benchmark Dataset for Pedagogically Sound LLM-Generated Explanations Based on Bloom's Taxonomy

Maria-Eleni Zoumpoulidi[1], **Eleni Batsi**[1], **Georgios Paraskevopoulos**[1]
Vassilis Katsouros[1], **Alexandros Potamianos**[2]
[1]Institute for Language and Speech Processing, Athena Research Center, Greece
[2]School of Electrical and Computer Engineering, National Technical University of Athens, Greece
{m.zoumpoulidi, eleni.mpatsi, g.paraskevopoulos, vsk}@athenarc.gr
potam@central.ntua.gr

Abstract

The ability of Large Language Models (LLMs) to generate accurate and pedagogically sound instructional explanations is necessary for their effective deployment in educational applications, such as AI tutors and teaching assistants. However, little research has systematically evaluated their performance across varying levels of cognitive complexity. Believing that such a direction serves the dual goal of not only producing more educationally sound and human-aligned outputs, but also fostering more robust reasoning and, thus, leading to more accurate results, we introduce BloomXplain, a framework designed to generate and assess LLM-generated instructional explanations across Bloom's Taxonomy levels. We first construct a STEM-focused benchmark dataset of question–answer pairs categorized by Bloom's cognitive levels, filling a key gap in NLP resources. Using this dataset and widely used benchmarks, we benchmark multiple LLMs with diverse prompting techniques, assessing correctness, alignment with Bloom's Taxonomy and pedagogical soundness. Our findings show that BloomXplain not only produces more pedagogically grounded outputs but also achieves accuracy on par with, and sometimes exceeding, existing approaches. This work sheds light on the strengths and limitations of current models and paves the way for more accurate and explainable results.

1 Introduction

"If you can't explain it simply, you don't understand it well enough": Attributed to Einstein, this quote underscores how clear explanations both aid teaching and signal true understanding. As LLMs are increasingly used in educational tools, assessing their explanatory capabilities is becoming vital.

Recent work has explored these capabilities. Chain-of-Thought (CoT) prompting enables stepwise reasoning [Wei et al., 2023], with SEA-CoT aligning it with context [Wei Jie et al., 2024]. Other approaches enhance explanation quality via KG-to-text generation [Axelsson and Skantze, 2023], assertion-based few-shot learning [Shahriar et al., 2024], logic scaffolding [Rahdari et al., 2024], and iterative refinement [Madaan et al., 2023]. In parallel, NLP research on cognitive alignment evaluates model behavior against frameworks like Bloom's Taxonomy [1] (fig. 2). Early work used traditional classifiers [Ullrich and Geierhos, 2023], while newer approaches leverage LLMs [Raz et al., 2023] to

[1]a multi-tiered model of classifying thinking according to six cognitive levels of complexity

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: LLM-eval@NeurIPS: Workshop on Evaluating the Evolving LLM Lifecycle – Benchmarks, Emergent Abilities, and Scaling.

classify questions in Bloom's taxonomy levels, generate Bloom-aligned questions [Scaria et al., 2024, Hwang et al., 2023], and analyze model performance in educational contexts [Maiti and Goel, 2024]. Benchmarks for Bloom-aligned generation [Chen et al., 2024] and taxonomy coverage of existing benchmarks [Huber and Niklaus, 2025] reveal imbalances, especially at higher levels.

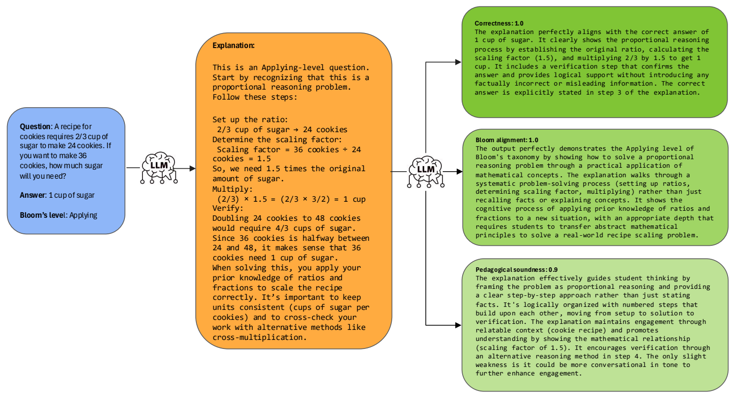


Figure 1: Overview of BloomXplain (BAQ).The input (question,answer, Bloom's level) is shown in the blue box, followed by the generated explanation (orange box) and the evaluation (green boxes). For further prompt details, the reader can refer to Appendix D.

Yet, most studies focus on question generation rather than explanation quality, exposing a gap: the lack of systematic methods for generating and evaluating cognitively aligned explanations. Our work addresses this by shifting focus from answer correctness to explanation quality, investigating whether LLMs can generate outputs aligned with human learning processes. We introduce **BloomXplain**, a framework for generating and evaluating explanations across Bloom's Taxonomy. We develop a STEM-focused benchmark dataset annotated with Bloom levels and test prompting strategies to elicit tutor-like responses from four LLMs. Outputs are evaluated by human reviewers and via an LLM-as-a-judge technique, focusing on accuracy, Bloom alignment and pedagogical soundness. An overview can be found in Fig. 1.

Our main contributions are: 1) a STEM QA benchmark dataset annotated with Bloom levels, offering a cognitively aligned benchmark (our code and data is released under the Apache 2.0 license[2] 2) Prompting strategies for eliciting Bloom-aligned explanations, highlighting LLMs' educational and reasoning potential 3) A multi-metric evaluation of LLMs across Accuracy, Bloom Alignment, and Pedagogical Effectiveness, revealing trade-offs between precision, cognitive depth, and pedagogical quality (for the broader impact see K).

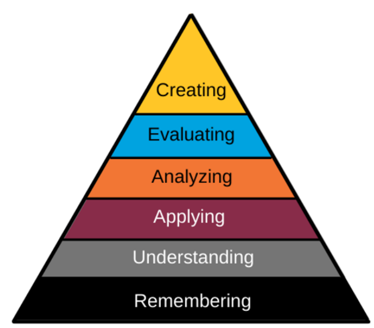


Figure 2: Bloom's Taxonomy (as revised by Anderson and Krathwohl [2001]).

2 BloomXplain

Our goal is to develop a framework that generates and evaluates explanations that are accurate, aligned with Bloom's Taxonomy, and pedagogically sound. An overview is shown in Fig. 1.

2.1 Bloom's Taxonomy-Aligned Dataset Creation

We present a novel STEM-focused QA benchmark dataset annotated with Bloom's Taxonomy levels. Each entry includes a question, a concise gold answer, and a Bloom level. To our knowledge, no existing NLP dataset offers structured Bloom annotations in STEM. Prior work also highlights limited coverage of higher-order skills by the LLMs and Bloom-level imbalance. To address this, we built the dataset from scratch. Initial QA–Bloom triples were generated using Claude-3.7-Sonnet[3], with

[2] https://github.com/marilena1123/BloomXplain/tree/main

[3] https://www.anthropic.com/news/claude-3-7-sonnet

Table 1: Main Results across models and methods

Model	Method	Correctness	Bloom Alignment	Pedagogical Soundness	Overall Score
deepseek-r1	BAQ	94.99	**92.75**	88.75	**92.00**
	AQ	93.75	87.00	**89.83**	90.00
	Baseline	**96.16**	-	76.16	85.99
llama3.1 70b	BAQ	91.16	**86.83**	**79.49**	**85.66**
	AQ	79.91	72.08	66.75	73.00
	Baseline	**96.66**	-	53.41	75.08
llama3.1 8b	BAQ	89.91	**82.33**	**70.83**	**80.99**
	AQ	93.41	78.41	63.66	78.41
	Baseline	**95.75**	-	49.50	72.66
gpt-4o-mini	BAQ	92.50	**89.08**	**77.83**	**86.50**
	AQ	89.91	80.08	72.08	80.58
	Baseline	**93.99**	-	48.58	71.33

prompts designed for diversity and minimal repetition (Appendix C- the few shot examples used are human-generated). All samples were then reviewed by two annotators for factual accuracy, diversity, correct Bloom classification and appropriate educational level categorization (QA pairs which did not meet the criteria were replaced). The dataset contains 360 QA pairs across mathematics, science, and technology with 30 questions per domain at each of four educational levels: elementary, junior high, high school, and undergraduate. While major subfields are covered, we did not enforce subfield balance to reflect natural curricular distributions. Focusing on educational content, where Bloom's Taxonomy is widely used and ,thus, sources abound, helped ensure cognitively aligned questions—especially at higher Bloom levels, which LLMs often struggle to produce. For the interested reader, information about our dataset's size and scope extensibility can be found at Appendix H and I.

2.2 Prompting Strategies

For our implementation, we employed the DSPy framework [Khattab et al., 2024]. The prompt templates and few-shot examples used for each strategy can be found in Appendix D. We propose two prompting strategies designed to elicit Bloom-aligned instructional outputs from LLMs:

Question + Answer + Bloom Level, Level-Specific Prompt (BAQ): The model received the question, gold answer, and Bloom level, and was prompted to generate tutor-style explanations using a prompt tailored to that level.

Question + Answer Only, Generic Prompt (AQ): Given the question and answer (but not the Bloom level), the model first inferred the level and then generated an aligned explanation using a generic Bloom Taxonomy prompt.

2.3 Evaluation

We evaluated explanations using both automated and human assessments across three criteria: **Correctness**, **Bloom Alignment**, and **Pedagogical Soundness**. For automated evaluation, we used the GEval [4] framework from Deepeval[5], shown to approximate human judgment [Liu et al., 2023b] (criteria in Appendix E). For human evaluation, a high school teacher with Physics background assessed 12 explanations-from GPT-4o-mini and DeepSeek-R1 under BAQ, AQ, and baseline prompts (216 instances), scoring each criterion on a 0–10 scale using a rubric (details in J) [6].

3 Experiments

3.1 Main Results

Table 1 summarizes our main results (baseline prompts in Appendix F). Across models, **BAQ** outperforms other methods in *pedagogical soundness* (avg. 79.23 vs. 73.08 for AQ, 56.91 for Baseline) and *Bloom alignment* (87.75 vs. 79.39 for AQ; not applicable for Baseline). It also maintains high *correctness* (92.14), close to AQ (89.25) and Baseline (95.64). These results show that BAQ's explicit Bloom-level guidance achieves the best balance of pedagogical depth and factual

[4] G-Eval is a trade-off between cost and sufficiently high quality

[5] `https://github.com/confident-ai/deepeval`

[6] We use a 0–10 scale because it's more intuitive for human evaluators, and our analysis is mapped to the same scale for consistency.

accuracy. While Baseline scores highest in correctness (95.64), its lack of structure leads to the lowest pedagogical score (56.91), revealing a trade-off between factual accuracy and instructional quality. AQ, which infers Bloom levels, underperforms BAQ in both Bloom alignment (79.39 vs. 87.75) and pedagogy (73.08 vs. 79.23), suggesting model-driven inference introduces alignment errors. BAQ's slight drop in correctness is offset by stronger pedagogical alignment, making it the most holistic method—especially for educational tasks requiring cognitive scaffolding.

Table 2: Human evaluation scores comparing correctness, Bloom alignment and pedagocical soundness for different methods (in a scale 0-10)

Model-Method	Correct.	Bloom Align.	Pedag. Sound.
deepseek-r1 BAQ	**8.79**	**9.08**	**8.67**
deepseek-r1 AQ	8.33	8.58	8.17
deepseek-r1 base	7.88	-	7.25
gpt-4o-mini BAQ	**8.88**	**8.71**	**8.25**
gpt-4o-mini AQ	8.08	7.54	7.33
gpt-4o-mini base	7.96	-	7.25

3.2 Human evaluation

Table 2 shows human scores. BAQ ranks highest in pedagogy and Bloom alignment across both evaluation types. For correctness, automated scores favor Baseline, while humans prefer BAQ and AQ—likely due to holistic judgments vs. factual precision. T-tests between BAQ and AQ show no significance for deepseek-r1 ($p = 0.35, 0.32, 0.26$) but significant differences for gpt-4o-mini ($p = 0.002, 0.032, 0.033$), reflecting narrower gaps in reasoning-optimized models.

3.3 Performance per model

Our results show consistent trade-offs across models and methods. For deepseek-r1, BAQ combines high correctness (94.99), top Bloom alignment (92.75), and strong pedagogy (88.75). Baseline scores highest in correctness (96.16) but lower in pedagogy (76.16), while AQ offers strong pedagogy (89.83) but weaker alignment (87.00). LLaMA3.1-70B shows sharper contrasts: BAQ balances correctness (91.16), alignment (86.83), and pedagogy (79.49); Baseline excels in correctness (96.66) but drops in pedagogy (53.41); AQ underperforms overall (79.91, 72.08, 66.75). LLaMA3.1-8B relies on BAQ scaffolding to stabilize pedagogy (70.83) and alignment (82.33) with acceptable correctness (89.91). AQ scores higher in correctness (93.41) but sacrifices pedagogy (63.66) and alignment (78.41), while Baseline again drops pedagogically (49.50) despite high correctness (95.75). GPT-4o-mini follows the same trend: BAQ offers balanced performance (92.50, 89.08, 77.83), AQ lags in pedagogy (72.08) and alignment (80.08), and Baseline favors correctness (93.99) over pedagogy (48.58). Across all models, BAQ consistently resolves the correctness–pedagogy trade-off, outperforming AQ in Bloom alignment and avoiding Baseline's instructional weaknesses.

For the interested reader, ablations and extended analysis as well as results for problems consisting of multiple Bloom's Taxonomy levels can be found at Appendices A and B.

4 Conclusion

BloomXplain is a framework for generating and evaluating LLM explanations aligned with Bloom's Taxonomy. Using a Bloom-annotated STEM QA benchmark, we test two prompting strategies: BAQ, where the level is given, and AQ, where the LLM infers it. Evaluated across four LLMs using both an LLM-as-a-judge method and human evaluation, BAQ outperforms AQ and generic prompts in pedagogical soundness with minimal loss in correctness, with Deepseek-r1 as the top-performing among models.

Limitations

We acknowledge that while our benchmark dataset provides valuable insights for analysis, its utility could be significantly enhanced by increasing its size. Achieving this expansion would require additional research efforts or extensive human annotation, as the generation of high-quality, Bloom-aligned data by LLMs remains constrained. Additionally, although our LLM-based evaluation framework delivers reliable assessments, it does not fully capture human characteristics, including inherent biases.

Acknowledgments

This research was funded by the European Union's Horizon Europe research and innovation programme under the AIXPERT project (Grant Agreement No. 101214389), which aims to develop an agentic, multi-layered, GenAI-powered framework for creating explainable, accountable, and transparent AI systems. We also acknowledge the EuroHPC Joint Undertaking for awarding this project access to the EuroHPC supercomputer LEONARDO, hosted by CINECA (Italy) and the LEONARDO consortium through an EuroHPC AI Data Intensive Applications Access call.

References

L. W. Anderson and D. R. Krathwohl. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Longman, 2001.

Agnes Axelsson and Gabriel Skantze. Using large language models for zero-shot natural language generation from knowledge graphs. In Albert Gatt, Claire Gardent, Liam Cripwell, Anya Belz, Claudia Borg, Aykut Erdem, and Erkut Erdem, editors, *Proceedings of the Workshop on Multimodal, Multilingual Natural Language Generation and Multilingual WebNLG Challenge (MM-NLG 2023)*, pages 39–54, Prague, Czech Republic, September 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.mmnlg-1.5/.

Yuyan Chen, Chenwei Wu, Songzhou Yan, Panjun Liu, and Yanghua Xiao. Dr.Academy: A benchmark for evaluating questioning capability in education for large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3138–3167, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.173. URL https://aclanthology.org/2024.acl-long.173/.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168, 2021. URL http://arxiv.org/abs/2110.14168.

Thomas Huber and Christina Niklaus. LLMs meet bloom's taxonomy: A cognitive view on large language model evaluations. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5211–5246, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.350/.

Kevin Hwang, Sai Challagundla, Maryam M. Alomair, Lujie Karen Chen, and Fow-Sen Choa. Towards ai-assisted multiple choice question generation and quality evaluation at scale: Aligning with bloom's taxonomy. In *Proceedings of the NeurIPS'23 Workshop on Generative AI for Education (GAIED)*, New Orleans, LA, USA, 2023.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. Dspy: Compiling declarative language model calls into self-improving pipelines. 2024.

Tengxiao Liu, Qipeng Guo, Yuqing Yang, Xiangkun Hu, Yue Zhang, Xipeng Qiu, and Zheng Zhang. Plan, verify and switch: Integrated reasoning with diverse X-of-thoughts. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2807–2822, Singapore, December 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.169. URL https://aclanthology.org/2023.emnlp-main.169/.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore, December 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback, 2023. URL https://arxiv.org/abs/2303.17651.

Pratyusha Maiti and Ashok K. Goel. How do students interact with an llm-powered virtual teaching assistant in different educational settings? In *Proceedings of the 17th International Conference on Educational Data Mining (EDM 2024), Workshop on Leveraging Large Language Models for Next-Generation Educational Technologies*. CEUR Workshop Proceedings, Volume 3840, 2024. URL https://arxiv.org/abs/2407.17429.

Behnam Rahdari, Hao Ding, Ziwei Fan, Yifei Ma, Zhuotong Chen, Anoop Deoras, and Branislav Kveton. Logic-scaffolding: Personalized aspect-instructed recommendation explanation generation using llms. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, WSDM '24, page 1078–1081, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703713. doi: 10.1145/3616855.3635689. URL https://doi.org/10.1145/3616855.3635689.

Tuval Raz, Simone Luchini, Roger Beaty, and Yoed Kenett. Automated scoring of open-ended question complexity: A large language model approach. 2023.

Nidhin Scaria, Dharani Chenna, and Dhanamjayulu Subramani. Automated educational question generation at different bloom's skill levels using large language models: Strategies and evaluation. In A.M. Olney, I.A. Chounta, Z. Liu, O.C. Santos, and I.I. Bittencourt, editors, *Artificial Intelligence in Education. AIED 2024*, volume 14830 of *Lecture Notes in Computer Science*. Springer, Cham, 2024. doi: 10.1007/978-3-031-64299-9_12. URL https://doi.org/10.1007/978-3-031-64299-9_12.

Tasmia Shahriar, Kelly Ramos, and Noboru Matsuda. Assertion enhanced few-shot learning: Instructive technique for large language models to generate educational explanations, 2024. URL https://arxiv.org/abs/2312.03122.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. Challenging big-bench tasks and whether chain-of-thought can solve them, 2022. URL https://arxiv.org/abs/2210.09261.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.824. URL https://aclanthology.org/2023.findings-acl.824/.

Sabine Ullrich and Michaela Geierhos. Using bloom's taxonomy to classify question complexity. In *Proceedings of the Research Institute CODE*, Germany, 2023. Universität der Bundeswehr München.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

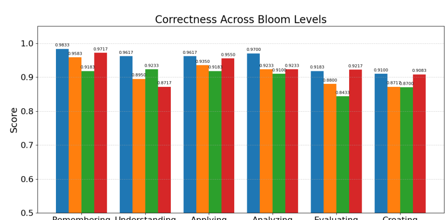
Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL `https://arxiv.org/abs/2201.11903`.

Yeo Wei Jie, Ranjan Satapathy, Rick Goh, and Erik Cambria. How interpretable are reasoning explanations from prompting large language models? In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2148–2164, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.138. URL `https://aclanthology.org/2024.findings-naacl.138/`.

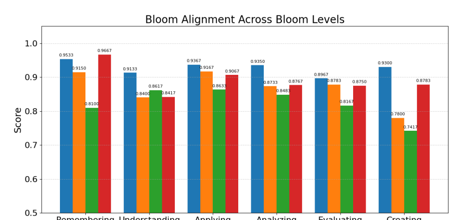
A Ablations and extended analysis

A.1 Performance per Bloom's level

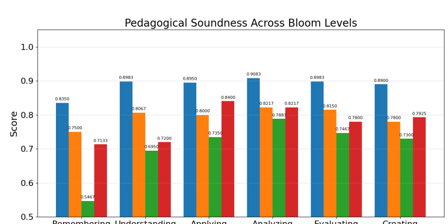
In Fig. 3 we show the performance of BAQ across different levels of the Bloom taxonomy for four LLMs.
Correctness:Deepseek-r1 consistently outperforms all other models across Bloom's taxonomy levels, followed closely by gpt-4o-mini and Llama 3.1 70b in most cases. Interestingly, Llama 3.1 8b ranks second in the "Understanding" level—potentially due to smaller models avoiding unnecessary complexity in comprehension tasks. Overall, performance declines as the taxonomy level increases (e.g., in "Evaluating"), indicating that LLMs generally struggle with abstract reasoning, regardless of their scale.
Bloom alignment: Deepseek-r1 excels in generating outputs that align with Bloom's taxonomy, while gpt-4o-mini and Llama 3.1 70b also perform well, exhibiting comparable results. Llama 3.1 8b consistently lags behind, except in the "Understanding" level, where it slightly outperforms other non-reasoning-optimized models.
Pedagogical soundness: Deepseek-r1 again demonstrates the highest performance, with gpt-4o-mini and Llama 3.1 70b closely competing for second place. In contrast, Llama 3.1 8b exhibits the lowest performance, suggesting that smaller or less sophisticated models struggle with pedagogical soundness, likely due to insufficient training in educational contexts.



(a) Correctness



(b) Bloom Alignment



(c) Pedagogical Soundness

(d) Legend

Figure 3: Performance of BAQ across Bloom's levels with four LLMs. We measure (a) Correctness, (b) Bloom alignment, (c) Pedagogical soundness.

A.2 Comparison with CoT on widely used benchmarks

We hypothesize that our Bloom-aligned prompting approach not only produces high-quality explanations but also enhances robust reasoning. To validate this, we compared our best-performing method (BAQ) against the widely used Chain-of-Thought (CoT) [Wei et al., 2022] approach on commonly used benchmarks in terms of accuracy. The selected benchmarks are BBH and GSM (Suzgun et al. [2023] and Cobbe et al. [2021]). The benchmark selection is based on the mapping of benchmarks to Bloom's taxonomy levels as defined in Huber and Niklaus [2025]. Given that this mapping covers only the first four levels of the taxonomy, we selected one benchmark (task) for each level and randomly selected 100 evaluation samples from each benchmark[7]. The results of this comparison are presented in Table 3. For BAQ, we used the same prompts as we used in our datasets (of course, we did not provide the gold answers to the LLM) with slight rephrasing in understanding, to make the prompt aligned to the task. The Understanding prompt and few-shot examples used for BAQ can be found in Appendix G. For CoT, we used the prompts and 3-shot examples from Suzgun et al. [2022] and Liu et al. [2023a]. We chose Deepseek-r1 and gpt-4o-mini to explore the differences between reasoning and non-reasoning-optimized models. Our experiments demonstrate that BAQ achieves competitive or superior performance compared to Chain-of-Thought (CoT) across Bloom's taxonomy levels, validating its efficacy in fostering robust reasoning. For deepseek-r1, BAQ outperforms CoT on all tasks. For gpt-4o-mini, BAQ excels in remembering, applying, and analyzing, but lags slightly in understanding (-6%). The results underscore the value of aligning prompts with Bloom's taxonomy—focusing not on what to think but how to think—enhancing performance, particularly for non-reasoning-optimized models like gpt-4o-mini. The Understanding-level task—referential ambiguity detection—revealed a key divergence in model reasoning: Deepseek-r1 predominantly anchors decisions to grammatical rules (e.g., pronoun-noun agreement), while gpt-4o-mini prioritizes pragmatic likelihood (e.g., real-world plausibility). This explains why gpt-4o-mini's CoT outperforms Deepseek-r1's in this task. Crucially, our BAQ method bridges this gap for Deepseek-r1 (+18% accuracy) by nudging it toward contextually probable interpretations, though it offers diminishing returns for gpt-4o-mini (+6%), whose default pragmatism already aligns with the task's demands.

Table 3: Comparison of our best method (BAQ) with CoT in terms of accuracy across widely used benchmarks. We selected 100 evaluation samples randomly for each benchmark and counted the correct answers

Model	Benchmark	CoT	BAQ (ours)
Deepseek-r1	BBH object counting (Remembering)	96	**100**
	BBH disambiguation qa (Understanding)	60	**78**
	GSM (Applying)	**99**	**99**
	BBH snarks (Analyzing)	90	**93**
gpt-4o-mini	BBH object counting (Remembering)	88	**95**
	BBH disambiguation qa (Understanding)	**74**	68
	GSM (Applying)	94	**98**
	BBH snarks (Analyzing)	78	**79**

B Multi-Bloom level problems

For completeness, we compared our best-performing method (BAQ) and the baseline on questions that align with multiple Bloom's taxonomy levels. Results are shown in Table 4. Deepseek-r1, a reasoning-optimized model, and GPT-4o-mini, a non-reasoning-optimized model, were evaluated. Our results show that Deepseek-r1 maintains similar (and slightly higher) performance on multi-level Bloom problems as it does on single-level problems. Specifically, BAQ and the baseline method demonstrate comparable correctness, while BAQ significantly outperforms the baseline in pedagogical soundness.

In contrast, GPT-4o-mini shows a sharper decline in correctness when using BAQ on multi-level problems, performing worse than both single-level BAQ and the multi-level baseline. However, its Bloom alignment remains similar to that of the single-level case. Notably, BAQ achieves higher pedagogical soundness than the baseline, with both models showing better pedagogical performance on multi-level problems than on single-level ones.

[7] This choice was made to limit API costs

Table 4: Results for multi-Bloom level problems

Model	Method	Correctness	Bloom Alignment	Pedagogical Soundness	Overall
Deepseek-r1	BAQ	0.94	0.9667	0.9067	0.9367
	Baseline	0.94	-	0.7900	0.8633
gpt-4o-mini	BAQ	0.8600	0.8833	0.8400	0.8600
	Baseline	0.9167	-	0.5067	0.7133

These findings suggest that reasoning-optimized models like Deepseek-r1 remain unaffected by the complexity of multi-Bloom-level problems, maintaining high performance. Conversely, non-reasoning-optimized models like GPT-4o-mini tend to produce more verbose explanations, leading to improved pedagogical soundness but reduced correctness.

C Dataset creation prompts

The prompt templates and few-shot examples used for data generation are provided in the Tables 5 and 6.

Table 5: Prompt templates used for QA pairs' creation

Level	Prompt Template
Remember	Create a problem at the Remembering level of Bloom's Taxonomy for the topic **topic** at a **difficulty level** level. The problem should require students to recall specific facts, definitions, or basic concepts.
Understand	Create a problem at the Understanding level of Bloom's Taxonomy for the topic **topic** at a **difficulty level** level. The problem should require students to explain, summarize, or paraphrase key concepts.
Apply	Create a problem at the Applying level of Bloom's Taxonomy for the topic **topic** at a **difficulty level** level. The problem should require students to apply a concept to a real-world scenario or novel situation.
Analyze	Create a problem at the Analyzing level of Bloom's Taxonomy for the topic **topic** at a **difficulty level** level. The problem should require students to identify relationships, patterns, or underlying structures.
Evaluate	Create a problem at the Evaluating level of Bloom's Taxonomy for the topic **topic** at a **difficulty level** level. The problem should require students to evaluate an argument, solution, or theory and justify their reasoning with evidence.
Create	Create a problem at the Creating level of Bloom's Taxonomy for the topic **topic** at a **difficulty level** level. The problem should prompt students to synthesize knowledge and generate a new idea, design, or alternative solution.

D Explanation creation prompts

The prompts and few-shot examples for explanation (BAQ and AQ) generation can be found in Tables 7, 8 and 9.

E GEval Evaluation Criteria

The criteria used for the evaluation (both human and automated) are shown in Table 10.

Table 6: Few-shot examples for dataset creation

Question	Answer
What is the formula for the area of a triangle?	$(1/2) \times$ base $\times$ height
What is the atomic number of carbon?	6
Explain the role of the CPU in a computer system.	The CPU (Central Processing Unit) is the brain of the computer; it processes instructions and manages tasks to ensure the system functions properly.
Explain Newton's First Law of Motion in simple terms.	An object will stay at rest or keep moving at the same speed and direction unless something forces it to change.
How does natural selection contribute to evolution?	Natural selection favors individuals with traits that help them survive and reproduce, gradually leading to evolutionary changes in a population.
How would you reduce your carbon footprint in daily life?	By using public transport, reducing energy consumption, recycling, and eating more plant-based foods.
Analyze why the concept of derivatives represents the rate of change in a function.	Derivatives measure how a function's output value changes as its input changes, indicating the function's rate of change at any point.
Compare the processes of mitosis and meiosis. What do their differences reveal about their roles in the body?	Mitosis produces identical cells for growth and repair, while meiosis creates genetically diverse gametes for reproduction, showing their distinct roles in bodily maintenance and genetic diversity.
Assess whether the solution to the equation $2x + 5 = 15$ is correct if $x = 5$.	Yes, because $2(5) + 5 = 15$, so $x = 5$ is a valid solution.
Evaluate whether using a solid-state drive (SSD) instead of a hard disk drive (HDD) significantly improves overall computer performance.	Yes, SSDs significantly improve performance because they have faster data access speeds, reduce boot time, and make programs load more quickly.
Design an experiment to test how different types of liquids affect plant growth.	Plant identical seeds in separate pots and water each with a different liquid (e.g., water, soda, juice). Keep all other conditions the same and measure growth over time.
Design a real-world word problem that requires using the Pythagorean theorem to solve.	A ladder is leaning against a wall. The foot of the ladder is 3 meters from the wall and it reaches 4 meters up. How long is the ladder?

F Baseline Prompt and few-shot examples

The baseline prompt and few shot examples are shown in Table 11.

G Prompts and few-shot examples for widely used benchmarks

Few-shots examples of BAQ from widely used benchmarks are shown in Tables 12, 13, 14 and 17. The "understanding" prompt is shown below:
Understanding Prompt: This is an Understanding-level question. Guide the student toward the most probable interpretation of the pronoun based on the context of the sentence. If there are contextual clues suggesting that one option is more likely than the others, prefer that option over choosing

Table 7: Prompts for explanations' creation (Question + Answer + Bloom Level, Level-Specific Prompt-BAQ)

Level	Guidance Prompt
Remembering	This is a Remembering-level question. Give a brief explanation that helps the student recall key facts or terms to answer the question. Include a mnemonic or simple strategy to remember.
Understanding	This is an Understanding-level question. Explain how the student can paraphrase or explain the concept in their own words, and walk through the reasoning that leads to the answer.
Applying	This is an Applying-level question. Describe how the student should use prior knowledge in a new situation. Walk through the steps of applying a concept to solve the problem.
Analyzing	This is an Analyzing-level question. Guide the student to break down the concept into parts, find patterns, or relationships. Walk through how to think critically to reach the answer.
Evaluating	This is an Evaluating-level question. Help the student assess or justify a position or solution. Walk through the logic used to check or critique the answer.
Creating	This is a Creating-level question. Guide the student through combining ideas to produce something new. Walk through the steps of designing or generating a solution.

Table 8: Prompts for explanations' creation (Question + Answer, Generic Prompt-AQ)

	Generic prompt
Tutor Guidance	You are a helpful tutor. Use the following guidance to craft your explanation depending on the level of the question: • **Remembering:** Help the student recall key facts or terms. Use mnemonics or memory strategies. • **Understanding:** Guide the student to paraphrase the concept and walk through the reasoning that leads to the answer. • **Applying:** Show how to use prior knowledge in a new situation. Walk through the application. • **Analyzing:** Break down the concept and compare parts or relationships. Walk through how to think critically to reach the answer. • **Evaluating:** Help the student justify or critique a position using logic or evidence. • **Creating:** Guide the student in combining ideas to produce something new. Walk through the steps of the design process.

'Ambiguous'. Choose the best option from the list. Only select 'Ambiguous' if there is truly no way to reasonably infer the referent. Walk through the reasoning that leads to the answer.

H Dataset extensibility

Although our dataset is a benchmark, we do recognize that a larger dataset would enable further generalisation and fine-tuning applications. To demonstrate the potential of our dataset to be extended when resources allow, we generated an additional 120 mathematics samples. The results for these additional samples (evaluated with DeepSeek-R1 and GPT-4o-mini due to cost constraints) can be found in Table 15.

I Dataset scope

The main focus of our analysis is the STEM educational domain. Nonetheless, it is important to demonstrate the potential of our dataset to extended to other domains. Thus, we generated 60 samples corresponding to the humanities, which can be found in Table 16.

J Human Evaluation Instructions

Goal

We aim to automatically create high-quality explanations aligned with Bloom's Taxonomy. Your task is to evaluate explanations generated under different conditions. Settings (e.g., model or method) will not be revealed to avoid bias.

Background: Bloom's Taxonomy

- **Remembering:** Retrieving, recognizing, recalling knowledge.
- **Understanding:** Constructing meaning via interpreting, exemplifying, summarizing, comparing, explaining.
- **Applying:** Executing or implementing a procedure.
- **Analyzing:** Breaking down material, identifying relations and structure.
- **Evaluating:** Making judgments based on criteria and standards.
- **Creating:** Producing or reorganizing elements into a new whole.

Evaluation Criteria

For each explanation you will be given: the question, a concise gold answer and the corresponding level of Bloom's Taxonomy. Each explanation is rated on three criteria, using a scale of 0 (very poor) to 10 (excellent).

1. **Correctness (0–10)** Evaluate whether the explanation is factually accurate and logically consistent with the correct answer. The explanation must not contain any incorrect or misleading information. It should support or justify the correct answer, either directly or indirectly. Elaboration is acceptable as long as it aligns with the correct answer and does not introduce confusion or contradictions. It is acceptable if the correct answer is clearly implied, even if it is not explicitly stated; do not penalize for lack of explicit restatement.
    - 10: Perfectly correct; no errors.
    - 8–9: Mostly correct; minor issues only.
    - 5–7: Partial correctness; noticeable gaps.
    - 3–4: Major errors, some fragments correct.
    - 0–2: Mostly incorrect; severe misconceptions.

    Tips: Focus only on correctness (not style). Check calculations, reasoning, and units.
2. **Bloom Alignment (0–10)** Assess whether the explanation demonstrates the thinking style or cognitive demand associated with the specified Bloom's level (e.g., factual recall for Remembering, conceptual explanation for Understanding, real- world application for Applying). Do not evaluate for factual correctness or instructional quality
    - 10: Fully demonstrates intended level.
    - 8–9: Strong alignment; slight drift.
    - 5–7: Partial alignment; mixed levels.
    - 3–4: Minimal alignment.
    - 0–2: No alignment.

Tips: Use Bloom's descriptions; correctness does not affect this score.

3. **Pedagogical Soundness (0–10)** Evaluate how well the explanation functions as a teaching tool. Consider its clarity, organization, engagement. Place emphasis on how effectively it guides the student's thought process towards arriving at the answer or verifying it rather than just providing details. Do not evaluate for factual accuracy or alignment with the specified Bloom level. Some examples of explanations which effectively work as teaching tools are: -Question: 12*6? -Answer: 72 -Explanation: To answer this question, you need to multiply 12 with 6. In order to make this easier, you can do 10*6 and then add 6+6. The result is 72. -Evaluation: This explanation is clear, organized and easy to understand by someone who is in elementary school. It effectively guides the learner towards the answer by explaining the process and giving tricks to simplify it.

    - 10: Exceptionally clear, well-structured.
    - 8–9: Strong pedagogy; minor mismatches.
    - 5–7: Adequate but too simple/complex.
    - 3–4: Poorly adapted or disorganized.
    - 0–2: Confusing, unhelpful, inappropriate.

    Tips: Check for guidance, examples, and analogies. Do not penalize factual errors or Bloom alignment here.

General Instructions

- Evaluate each explanation independently.
- Apply the same rubric and scale consistently.
- Do not infer generation settings.
- Use the full 0–10 scale when justified.
- Follow rubric definitions strictly.

Procedural Rules

- Independence: Do not revisit earlier scores.
- Time: Spend ∼1–2 minutes per evaluation.
- Calibration: Practice with 3–5 examples before official scoring.

K Broader impacts

BloomXplain advances the educational use of LLMs by systematically evaluating and improving their ability to produce explanations aligned with Bloom's Taxonomy, potentially enabling more pedagogically sound AI tutors and scalable, personalized learning. By guiding LLMs to generate explanations matched to different cognitive levels, it can help learners build deeper understanding, expand access to quality guidance in underserved regions, and support teachers by augmenting rather than replacing their role. However, there is a risk of over-reliance on AI-generated explanations, which may sometimes reflect biases or inaccuracies that could affect learning outcomes, if not carefully reviewed and supplemented by human oversight.

Table 9: Few-shot examples used in BAQ and AQ

Level	Question + Answer	Tutor Explanation
Remembering	**Q:** What is the chemical symbol for sodium? **A:** Na	This is a Remembering-level question. To answer it, you need to recall the symbol for sodium from the periodic table. A helpful strategy is to remember that some elements use Latin roots. Sodium's Latin name is 'Natrium', which is why its symbol is 'Na'. A quick mnemonic: 'Na – Not Apparent from English!' to help remember it doesn't match the English spelling.
Understanding	**Q:** Explain why the heart has four chambers. **A:** The heart has four chambers to separate oxygen-rich and oxygen-poor blood, ensuring efficient circulation throughout the body.	This is an Understanding-level question. To answer it, try to explain in your own words what each chamber does. The right side of the heart handles blood returning from the body and sends it to the lungs to pick up oxygen. The left side receives the oxygenated blood and pumps it to the body. Walking through the process helps you understand how the heart keeps the oxygen-rich and oxygen-poor blood separate to work efficiently.
Applying	**Q:** Use the Pythagorean theorem to find the length of the hypotenuse in a right triangle with legs of 3 and 4 units. **A:** The hypotenuse is 5 units. Using $a^2 + b^2 = c^2$: $3^2 + 4^2 = 9 + 16 = 25$, so $\sqrt{25} = 5$.	This is an Applying-level question. Start by recognizing that this is a right triangle, so the Pythagorean theorem applies: $a^2+b^2 = c^2$. Use the values: $a = 3$ and $b = 4$. Compute $3^2 = 9$ and $4^2 = 16$, then add them: $9 + 16 = 25$. Now take the square root: $\sqrt{25} = 5$. This shows how to use prior knowledge in a new situation.
Analyzing	**Q:** Compare mitosis and meiosis in terms of genetic variation and number of resulting cells. **A:** Mitosis creates two identical cells; meiosis creates four genetically unique cells. Mitosis is for growth, meiosis for reproduction.	This is an Analyzing-level question. Start by breaking down each process: mitosis results in two identical cells, and is used for growth and repair. Meiosis results in four cells with half the DNA, all genetically different, and is used for reproduction. Notice the patterns: mitosis keeps the DNA the same; meiosis increases variation. You're analyzing by comparing structure, outcome, and function to understand deeper relationships.
Evaluating	**Q:** Which programming language would you recommend for beginners and why? **A:** Python, because its simple syntax makes it easy to learn, and it's widely used in various fields from web development to AI.	This is an Evaluating-level question. To answer, you must assess programming languages based on clarity, ease of learning, and real-world use. Python stands out due to its readable syntax and broad application. For example, 'print(Ḧello)̈' in Python is much simpler than in Java or C++. You're making a justified recommendation by comparing choices and applying logical criteria like simplicity and flexibility.
Creating	**Q:** Design an experiment to test the effect of light on plant growth. **A:** Place identical plants under different light conditions (sunlight, LED, darkness) and measure growth over 2 weeks to compare results.	This is a Creating-level question. Start by defining your goal — testing how light affects growth. Then combine your knowledge of experimental design: choose a variable (light type), control others (same plant species, water, soil), and decide on a measurable outcome (growth in cm). By generating this plan, you're combining ideas to build a new solution from scratch — the hallmark of creative thinking.

Table 10: Evaluation criteria provided to GEval for explanation evaluation (in pedagogical soundness, we used examples to guide the LLM because due to the abstract nature of the task, the scores were unreasonably high- due to formatting constraints, we only provide one of the examples provided to the model). The above criteria correspond to BAQ. For AQ, we added a disclaimer in the "correctness" criterion, prompting the model to determine correctness only in terms of alignment with the gold answer and not prediction of Bloom's level, as this is evaluated in the "Bloom alignment" section.

Criterion	Description
Correctness	Evaluate whether the explanation is factually accurate and logically consistent with the correct answer. The explanation must not contain any incorrect or misleading information. It should support or justify the correct answer, either directly or indirectly. Elaboration is acceptable as long as it aligns with the correct answer and does not introduce confusion or contradictions. It is acceptable if the correct answer is clearly implied, even if it is not explicitly stated; do not penalize for lack of explicit restatement.
Alignment with Bloom's Taxonomy	Assess whether the explanation demonstrates the thinking style or cognitive demand associated with the specified Bloom's level (e.g., factual recall for Remembering, conceptual explanation for Understanding, real-world application for Applying). Do not evaluate for factual correctness or instructional quality.
Pedagogical soundness	Evaluate how well the explanation functions as a teaching tool. Consider its clarity, organization, engagement. Place emphasis on how effectively it guides the student's thought process towards arriving at the answer or verifying it rather than just providing details. Do not evaluate for factual accuracy or alignment with the specified Bloom level. Some examples of explanations which effectively work as teaching tools are: -Question: 12*6? -Answer: 72 -Explanation: To answer this question, you need to multiply 12 with 6. In order to make this easier, you can do 10*6 and then add 6+6. The result is 72. -Evaluation: This explanation is clear, organized and easy to understand by someone who is in elementary school. It effectively guides the learner towards the answer by explaining the process and giving tricks to simplify it.

Table 11: Baseline Prompt and Few-Shot Examples (not all examples are listed due to formatting constraints)

Prompt	Create an explanation for why this is the correct answer.
Q: What is the chemical symbol for sodium?	A: Na. Explanation: The chemical symbol for sodium is 'Na', derived from its Latin name 'Natrium'.
Q: Why does the heart have four chambers?	A: To separate oxygen-rich and oxygen-poor blood. Explanation: The heart has two sides with two chambers each. One side sends blood to the lungs to get oxygen, and the other side pumps oxygen-rich blood to the rest of the body. This setup keeps the two types of blood from mixing.

Table 12: BAQ Few-Shot Examples for BBH Object Counting QA

Question	**Explanation**
I have a blackberry, a clarinet, a nectarine, a plum, a strawberry, a banana, a flute, an orange, and a violin. How many fruits do I have?	This is a Remembering-level question. First, you need to remember the definition of fruits:To answer this, you need to recall which items are fruits. Fruits are edible plant parts that often contain seeds and are commonly sweet. In this list, blackberry, nectarine, plum, strawberry, banana, and orange are all fruits. Clarinet, flute, and violin are musical instruments, not edible items. Now, you need to recall basing counting: 1+1+1+1+1+1=6. The final answer is 6.
I have an orange, a raspberry, two peaches, a blackberry, an apple, a grape, a nectarine, and three plums. How many fruits do I have?	This is a Remembering-level question. First, you need to remember the definition of fruits: fruits are edible plant parts that often contain seeds and are commonly sweet. All of the items listed—orange, raspberry, peaches, blackberry, apple, grape, nectarine, and plums—are fruits. Next, recall how to count quantities correctly: orange (1), raspberry (1), peaches (2), blackberry (1), apple (1), grape (1), nectarine (1), plums (3): 1 + 1 + 2 + 1 + 1 + 1 + 1 + 3 = 11. The final answer is 11.
I have a lettuce head, a head of broccoli, an onion, a stalk of celery, two carrots, a garlic, and a yam. How many vegetables do I have?	This is a Remembering-level question. First, you need to remember the definition of vegetables: vegetables are edible parts of plants such as roots, stems, and leaves, usually used in savory dishes. In this list, lettuce, broccoli, onion, celery, carrots, garlic, and yam are all vegetables. Next, recall basic counting: lettuce (1), broccoli (1), onion (1), celery (1), carrots (2), garlic (1), yam (1). Now, add the amounts: 1 + 1 + 1 + 1 + 2 + 1 + 1 = 8. The final answer is 8.

Table 13: BAQ Few-Shot Examples for BBH Disambiguation QA

Question	**Explanation**
The chief told the counselor that they took the day off. Who took the day off? A) The chief, B) the counselor, C) Ambiguous	This is an Understanding-level question. To answer it, you need to figure out who most likely took the day off. The sentence says "The chief told the counselor that they took th day off." Now think: if the counselor had taken the day off, they would already know that—there would be no reason to be told. But if the chief took the day off, it makes sense that the counselor is being informed. So, based on how people typically communicate, it's more likely that the chief is the one who took the day off. The final answer is A.
The manager sent a message to the secretary, but he didn't reply yet. Who didn't reply? A) The secretary, B) The manager, C) Ambiguous	This is an Understanding-level question. You're being asked to decide who most likely didn't reply. The sentence says "The manager sent a message to the secretary, but he didn't reply yet." Normally, when someone sends a message, the other person is expected to respond. So it's more probable that the secretary didn't reply. The sentence structure makes "he" slightly ambiguous, but based on typical communication flow, the most likely answer is A: the secretary didn't reply yet. The final answer is A.
Bailey will plan to meet the director at his office. At whose office? A) Bailey's, B) The director's, C) Ambiguous	This is an Understanding-level question. You need to decide whose office they're most likely meeting at. The sentence says "Bailey will plan to meet the director at his office." The pronoun "his" could refer to either Bailey or the director. There's no strong clue in the sentence to tell us which one for sure. Since both interpretations are reasonable and no extra context is given, we can't decide with confidence. That makes the sentence ambiguous. The final answer is C.

Table 14: BAQ Few-Shot Examples for GSM

Question	**Explanation**
Olivia has \$23. She bought five bagels for \$3 each. How much money does she have left?	To solve this, apply your understanding of multiplication and subtraction in real-life money scenarios. First, use multiplication to find out how much Olivia spent: 5 bagels × \$3 per bagel = \$15. Next, apply subtraction to find out how much money she has left: \$23 - \$15 = \$8. The final answer is 8.
Michael had 58 golf balls. On Tuesday, he lost 23 golf balls. On Wednesday, he lost 2 more. How many golf balls did he have at the end of Wednesday?	To solve this, apply subtraction in a multi-step context. Start with what you know: Michael had 58 golf balls. Use subtraction to find out how many he had after Tuesday: 58 - 23 = 35. Then subtract again to include Wednesday's loss: 35 - 2 = 33. The final answer is 33.
There were nine computers in the server room. Five more computers were installed each day, from Monday to Thursday. How many computers are now in the server room?	To solve this, apply multiplication and addition in a repeated event scenario. Recognize that 5 computers were added for 4 days (Monday through Thursday), so multiply: 5 × 4 = 20.Then add the new computers to the original: 9 + 20 = 29. The final answer is 29.

Table 15: Results for extension-mathematics

Model	Method	Correctness	Bloom Alignment	Pedagogical Soundness
deepseek-r1	BAQ	97.2	96.5	90.3
	AQ	92.5	91.8	85.6
	Baseline	97.2	-	76.8
gpt-4o-mini	BAQ	96.7	92.0	82.0
	AQ	91.8	89.5	66.5
	Baseline	97.7	-	64.8

Table 16: Results for extension- humanities

Model	Method	Correctness	Bloom Alignment	Pedagogical Soundness
deepseek-r1	BAQ	91	94	90
	AQ	87.5	75	87
	Baseline	96	-	83
gpt-4o-mini	BAQ	90	85	77
	AQ	87.5	80	67
	Baseline	94	-	63

Table 17: BAQ Few-Shot Examples for BBH Snarks

Question	Explanation
Which statement is sarcastic? Options: (A) Yes, because having interests and actively researching them is a huge waste (B) Yes, because having interests and actively researching them is a huge deal	This is an Analyzing-level question. To determine sarcasm, we need to look for a statement where the surface meaning contrasts with the intended meaning—often implying mockery o irony. Let's analyze Option (A): It says that having interests and researching them is "a huge waste." At face value, this sounds negative, but in reality, pursuing interests is seen as productive and meaningful. This contradiction suggests the speaker is not being sincere—they likely mean the opposite. That's a sign of sarcasm. Now let's look at Option (B): It calls those efforts "a huge deal," which agrees with the general view that pursuing interests is important. There's no contradiction here, and the tone is sincere. The final answer is A.
Which statement is sarcastic? Options: (A) No one is going to disagree with you on this. Avoiding ad hominem attacks really help your case (B) No one is going to disagree with you on this. Ad hominem attacks really help your case	This is an Analyzing-level question. We're looking for sarcasm—a mismatch between what's said and what's meant. Option (A) says avoiding ad hominem attacks helps your case. That makes logical sense—attacking the argument, not the person, is a better debate strategy. There's no irony or contradiction here. Option (B), on the other hand, praises ad hominem attacks—saying they "really help your case." But we know that such attacks usually weaken an argument by shifting focus from logic to personal insults. This mismatch between what's said and what is commonly understood creates a sarcastic tone. The final answer is B.
Which statement is sarcastic? Options: (A) Consistency in the league's punishments? What do you think this is supposed to be, politics? (B) Consistency in the league's punishments? What do you think this is supposed to be, moral?	This is an Analyzing-level question.To find sarcasm, we need to identify a mismatch between expectation and reality that's presented with irony. Option (A) compares consistency to politics, which is often viewed as inconsistent or hypocritical. So, the speaker might be ironically pointing out the lack of consistency by pretending to suggest it's too much to expect—this is sarcasm. Option (B) compares consistency to morality, which is a more straightforward comparison. It doesn't involve an ironic twist; it just asks whether the league should base its decisions on moral grounds Because Option (A) uses an ironic tone to criticize inconsistency, the sarcastic statement is Option (A).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Abstract, 1

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: The paper does not include theoretical results

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: 1 (code and data), 2 (framework description)

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: 1

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: 3, 1 (code)

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: 3

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We rely on the API calls to the language models (Information can be found in Section 3.1).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Although due to limited space no analysis is provided, we ensured that our work is aligned with the NeurIPS Code of Ethics

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Appendix K

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: No significant risk is identified in the data produced.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: 1, Appendix A2

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: A README file is provided with each asset.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: Appendix J

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Although human subjects are involved, there is no need for IRB approval (based on the country's and institution's procedure and given the nature of the paper).

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The LLM is used solely for editing purposes.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.