

BloomXplain: A framework and dataset for pedagogically sound LLM-generated explanations based on the Bloom’s Taxonomy

Anonymous ACL submission

Abstract

The ability of Large Language Models (LLMs) to generate accurate and pedagogically sound instructional explanations is a sine qua non 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 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 interpretable results.

1 Introduction

“If you can’t explain it simply, you don’t understand it well enough”: This quote, attributed to Albert Einstein, captures the dual significance of clear explanations: they are both practically useful for teaching and serve as indicators of true understanding. For this reason, research into the explanation and solution-planning capabilities of LLMs is of growing importance—particularly as these models are increasingly integrated into educational tools and AI-driven tutoring systems.

Several lines of research have begun to explore the explanation capabilities of LLMs. For example, Chain-of-Thought (CoT) prompting enables step-by-step reasoning (Wei et al., 2023), with SEA-CoT aligning reasoning with context (Wei Jie et al., 2024). Other works have focused on zero-shot KG-to-text generation for coherent explanations (Axelson and Skantze, 2023), Assertion-Enhanced Few-Shot Learning for clearer reasoning paths (Shahriar et al., 2024), Logic-Scaffolding for logical consistency (Rahdari et al., 2024), and Self-Refine for iterative self-feedback (Madaan et al., 2023).

At the same time, a growing body of NLP research has explored cognitive alignment—evaluating how well models conform to established frameworks such as Bloom’s Taxonomy. Early work in this area focused on question classification using traditional classifiers (Ullrich and Geierhos, 2023), while more recent approaches leverage appropriately trained LLMs (Raz et al., 2023). Other studies investigate question generation techniques tailored to Bloom levels (Scaria et al., 2024), (Hwang et al., 2023), revealing limitations in eliciting higher-order cognitive skills. Similar strategies have been applied to question answering tasks (Sahu et al., 2021) and to the analysis of LLM behavior in interactive educational environments (Maiti and Goel, 2024). In parallel, new benchmarks have emerged to assess how well LLMs generate cognitively-aligned questions (Chen et al., 2024) or to evaluate the Bloom’s taxonomy coverage of existing benchmarks (Huber and Niklaus, 2025), highlighting the generally imbalanced coverage and the absence of coverage at the higher levels.

While these works primarily focus on question generation and Bloom’s taxonomy coverage, they underscore a broader issue: existing frameworks often fail to adequately represent cognitive depth, particularly at higher levels. This observation exposed a critical gap—namely, the absence of sys-

tematic methods for generating and evaluating explanations and plans across varying levels of cognitive complexity. Our work addresses this gap by shifting the focus from mere answer generation to the explainability of outputs, exploring not only whether models can provide correct answers but also whether they can generate explanations that align with human learning processes.

To address the need for cognitively aligned explanation generation and evaluation, we present BloomXplain, a comprehensive framework for assessing LLMs’ ability to generate explanations, guidelines, and solution plans across the six levels of Bloom’s Taxonomy. Our approach begins with the development of a STEM-focused question-answering benchmark dataset, systematically annotated with Bloom levels to ensure clear alignment with cognitive complexity. Leveraging both this dataset and widely used benchmarks, we design and test various prompting strategies aimed at eliciting tutor-like responses from four LLMs. We assess their performance via both human reviewers and an LLM-as-a-judge approach. Our evaluation focuses on the models’ ability to generate accurate, pedagogically sound explanations tailored to the cognitive level of each query. An overview of our framework, which will be thoroughly presented in subsequent sections, can be found in Fig. 2. Our main contributions are:

- We introduce BloomXplain, a novel STEM-focused question-answering dataset annotated with Bloom’s Taxonomy levels, providing a fine-grained, cognitive-aligned benchmark for LLMs. Our code and dataset will be available to the research community under the Apache 2.0 license¹.
- We design a range of prompting techniques that elicit Bloom’s-taxonomy-aligned explanations and solution plans from LLMs, demonstrating their potential both as educational tools and as mechanisms for robust cognitive reasoning.
- We conduct a comprehensive evaluation of multiple state-of-the-art LLMs, benchmarking them across Accuracy, Bloom Alignment, and Pedagogical Effectiveness, revealing critical trade-offs between precision, cognitive complexity and pedagogical effectiveness.

¹https://osf.io/mg3c4/?view_only=6fe1767ade4c4852a312baf163fa43cb

More generally, BloomXplain highlights a new direction for LLMs, leveraging cognitively structured explanations and solution plans to enhance both educational soundness and reasoning depth. Moreover, our findings shed light on the capabilities and limitations of LLMs across different Bloom levels, revealing how well they handle varying cognitive complexities.

2 Related Work

2.1 Alignment with cognitive principles

Alignment with cognitive principles and levels is widely utilized in education and NLP research. One prominent direction involves categorizing questions according to Bloom’s taxonomy. Early approaches employed classifiers to determine question complexity levels (Ullrich and Geierhos, 2023), while more recent advancements, such as Raz et al. (2023), leverage appropriately trained LLMs.

Another significant avenue is question generation. Works such as Scaria et al. (2024) and Hwang et al. (2023) explore advanced prompting techniques to generate questions at various Bloom’s levels, highlighting LLMs’ limitations in generating questions requiring higher-order cognitive skills. Similar methodologies have been applied to question answering, as demonstrated in Sahu et al. (2021).

Alignment with cognitive principles has also been explored in interactive educational settings. For example, Maiti and Goel (2024) examines how an LLM-powered teaching assistant engages with students, analyzing question types and complexity using Bloom’s Revised Taxonomy.

In the context of LLM benchmarking, Chen et al. (2024) introduce a benchmark to assess LLMs’ ability to generate educational questions, utilizing Anderson and Krathwohl’s revised taxonomy (Anderson and Krathwohl, 2001). Additionally, Huber and Niklaus (2025) systematically evaluate widely used LLM benchmarks to determine their coverage of Bloom’s taxonomy levels. The findings reveal significant gaps and imbalances in how current benchmarking methods cover cognitive skills, with higher-order skills being predominantly underrepresented.

2.2 Explainability

Traditional explainability approaches include post-hoc explanation methods (e.g., IG (Bhat and Ray-

chowdhury, 2023), LIME (Ribeiro et al., 2016), SHAP (Liu and Barnard, 2021)), built-in interpretability via attention (Tull et al., 2024), human-in-the-loop explanations (Eiband et al., 2018), (Martens et al., 2025) and prompting frameworks for explainable reasoning. Focusing on the latter, Chain-of-Thought (CoT) prompting encourages models to generate step-by-step reasoning (Wei et al., 2023), making their thought processes transparent. SEA-CoT extends this by aligning reasoning paths with context (Wei Jie et al., 2024). Axelsson and Skantze (2023) explored zero-shot KG-to-text generation, transforming structured knowledge triples into coherent text without fine-tuning, enhancing the interpretability of reasoning. Assertion-Enhanced Few-Shot Learning (Shahriar et al., 2024) utilizes domain-specific assertions to produce clearer and more faithful reasoning paths. Logic-Scaffolding combines aspect-based personalization (Rahdari et al., 2024) with intermediate steps, ensuring logically grounded outputs. Finally, Self-Refine (Madaan et al., 2023) introduces iterative self-feedback, where the model critiques and improves its reasoning through self-assessment and revision, leading to more interpretable and human-like responses.

2.3 LLMs and Educational Explanations

Many research approaches use LLMs to generate personalized explanations and study plans, making learning more understandable and engaging (Laak and Aru, 2025), (Ng and Fung, 2024). Other works (Abu-Rasheed et al., 2023) highlight that LLMs can create graph-based explanations to organize syllabi systematically, helping students follow each step and understand the purpose behind each topic, thereby increasing their interest.

3 Preliminaries

3.1 Bloom’s Taxonomy

Bloom’s Taxonomy is a multi-tiered model of classifying thinking according to six cognitive levels of complexity. In the original version of the Taxonomy, the lowest three levels are: remembering, understanding, and applying. The highest three levels are: analyzing, synthesizing, and evaluating. The taxonomy is hierarchical, as shown in Fig. 1, where each level is subsumed by the higher levels. In 2001, the taxonomy was revised. The new structure of the taxonomy is: remembering, understanding, applying, analyzing, evaluating, and

creating. Our work is based on the revised taxonomy of Anderson and Krathwohl (2001). The steps

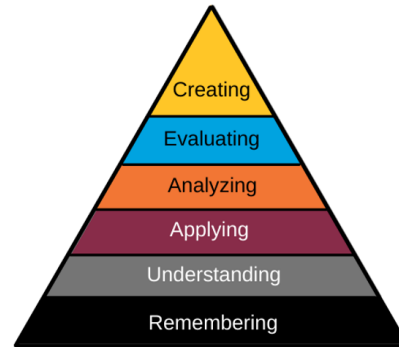


Figure 1: Bloom’s Taxonomy (as revised by Anderson and Krathwohl (2001)).

used in the Taxonomy are defined as follows (Forehand et al., 2005):

Remembering: Retrieving, recognizing, and recalling relevant knowledge from long-term memory.

Understanding: Constructing meaning from oral, written, and graphic messages through interpreting, exemplifying, classifying, summarizing, inferring, comparing, and explaining.

Applying: Carrying out or using a procedure through executing, or implementing.

Analyzing: Breaking material into constituent parts, determining how the parts relate to one another and to an overall structure or purpose through differentiating, organizing, and attributing.

Evaluating: Making judgments based on criteria and standards through checking and critiquing.

Creating: Putting elements together to form a coherent or functional whole; reorganizing elements into a new pattern or structure through generating, planning, or producing.

4 BloomXplain

Our goal is to develop a framework that generates and evaluates explanations and solution plans that are accurate, aligned with Bloom’s Taxonomy, and pedagogically sound. An overview is shown in Fig. 2.

4.1 Bloom’s Taxonomy-Aligned Dataset Creation

We introduce a novel STEM-focused QA dataset annotated with Bloom’s Taxonomy levels. Each entry consists of a question, a concise gold answer, and an associated Bloom level indicating

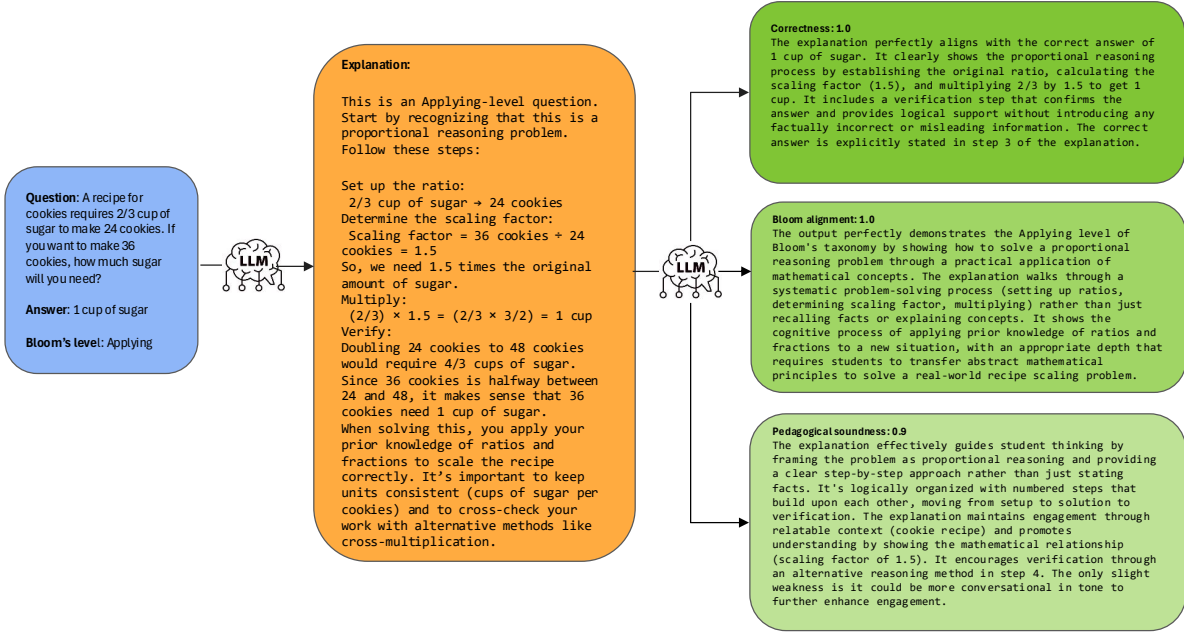


Figure 2: Overview of BloomXplain (BAQ). The input (question, answer, Bloom's level) is shown in the blue box, followed by the generated explanation (orange box) and the evaluation (green boxes). For further prompt details, the reader can refer to Appendix B.

cognitive complexity. To the best of our knowledge, no existing NLP dataset offers this form of cognitive-level annotation in a structured STEM context. Additionally, as noted in prior work, commonly used benchmarks show significant imbalance in Bloom's Taxonomy levels' coverage and gaps in the coverage of higher-order cognitive skills. Therefore, we constructed the dataset from scratch rather than augmenting an existing resource. For efficiency, we first generated initial Question–Answer–Bloom Level triples using Claude-3.7-Sonnet² (the prompts can be found in Appendix A). We explicitly prompted the model to produce diverse questions and avoid repetitions. A subset (30%) of these entries was then manually reviewed and validated by 2 human annotators to ensure factual accuracy, diversity, quality and correct classification according to Bloom's framework.

The dataset comprises 360 Question–Answer pairs, specifically covering the fields of mathematics, science, and technology. While we ensured inclusion of questions from major subfields within each domain, we did not enforce balanced representation across subfields, reflecting the natural uneven distribution of topics in educational curricula. Our priority was to cover a broad range of educational

explanations rather than artificially equalize sub-domain frequencies. The questions are distributed across four educational levels: elementary school, junior high school, high school, and undergraduate. This controlled selection ensures consistent coverage across both subject domains and cognitive development stages. Our focus on educational content was a deliberate choice, as Bloom's Taxonomy is widely used in educational contexts, making it more feasible to generate Bloom-aligned questions by leveraging educational data. This approach allowed us to overcome the well-known challenge of LLMs struggling to reliably generate Bloom-aligned questions, especially for higher cognitive levels. A detailed breakdown of the dataset by field and educational level is presented in Table 1.

Field	Elem.	Jr. High	High	Undergrad
Mathematics	30	30	30	30
Science	30	30	30	30
Technology	30	30	30	30
Total	90	90	90	90

Table 1: Distribution of dataset samples across fields and educational levels. Each field–level combination contains exactly 30 samples.

²<https://www.anthropic.com/news/claude-3-7-sonnet>

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

Table 2: Main Results across models and methods

4.2 Prompting Strategies

We propose three 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.

Question + Bloom Level Only, Level-Specific Prompt (Planning): The model was given the question and Bloom level (but not the answer) and was prompted to generate a solution plan aligned with that level without revealing the answer. For our implementation, we employed the DSPy framework (Khatab et al., 2024). The prompt templates and few-shot examples used for each strategy can be found in Appendix B.

4.3 Evaluation

4.3.1 LLM-as-a-judge

We employed the Deepeval GEval framework³ for evaluating the generated explanations, as it has been shown to closely approximate human judgments (Liu et al., 2023b). The evaluation was conducted across three dimensions, each reflecting a distinct aspect of explanation quality. First, we assessed *factual accuracy and logical consistency* (**Correctness**), ensuring that the explanation was correct, coherent, and aligned with the provided answer, without introducing misleading or contradictory information. Second, we evaluated *alignment with Bloom’s Taxonomy* (**Bloom Alignment**),

judging whether the explanation reflected the cognitive processes associated with the specified Bloom level (e.g., recall for *Remembering*, analysis for *Analyzing*). This criterion focused purely on cognitive alignment, independent of factual correctness. Third, we assessed the *instructional quality* (**Pedagogical soundness**) of the explanation—its clarity, structure, engagement, guidance effectiveness and pedagogical effectiveness—without considering correctness or taxonomy alignment. Each dimension was evaluated independently, allowing for a holistic assessment of the generated content’s educational value. The criteria used for GEval evaluation can be found in Appendix C.

4.3.2 Human evaluation

Under the human evaluation setting, an educator (physician), with teaching experience at the school level—independently assessed 12 samples of explanations (generated when prompting deepseek-r1 and gpt-4o-mini using BAQ, AQ and baseline) using the same three criteria applied in the LLM-based evaluation: Correctness, Bloom Alignment, and Pedagogical Soundness. The evaluator assigned a score from 0 to 10 for each criterion, guided by a detailed rubric to ensure consistent interpretation.

5 Experiments

5.1 Experimental setting

We benchmark deepseek-r1 (DeepSeek-AI et al., 2025), Llama3.1 8b, Llama3.1 70b (Grattafiori et al., 2024) and gpt-4o-mini⁴.

5.2 Main Results

In Table 2, we present our main results. The prompt and few shot examples for the baseline method can be found in Appendix D. Across all

³<https://github.com/confident-ai/deepeval>

⁴<https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>

models, BAQ outperforms the other methods in pedagogical soundness-achieving an average score of 79.23 while AQ and Baseline achieve 73.08 and 56.91 respectively- and Bloom alignment, where the scores are 87.75, 79.39 for BAQ and AQ (this metric was not applicable for the baseline). BAQ maintains a high correctness score of 92.14 while AQ and baseline achieve 89.25 and 95.64. The results demonstrate that BAQ’s explicit integration of Bloom-level guidance achieves the strongest balance between pedagogical rigor and factual accuracy. While the Baseline method prioritizes correctness (95.64), its lack of pedagogical structuring leads to the lowest pedagogical soundness (56.91), highlighting a critical trade-off between factual robustness and instructional utility. AQ, which infers Bloom levels, underperforms BAQ in both Bloom alignment (79.39 vs. 87.75) and pedagogical soundness (73.08 vs. 79.23), suggesting that model-driven inference introduces errors that propagate to explanation quality. BAQ’s slightly lower correctness (92.14) compared to the Baseline is offset by its superior pedagogical alignment, positioning it as a holistic solution for educational applications where scaffolding and cognitive targeting are paramount. (for the interested reader, results for problems consisting of multiple Bloom’s Taxonomy levels are shown in Appendix F).

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

Table 3: Human evaluation scores comparing correctness, Bloom alignment and pedagogical soundness for different methods (in a scale 0-10)

5.3 Human evaluation

Human evaluation scores can be found in Table 3. The results show that, based on both human and automated evaluations, BAQ is the top performer in terms of pedagogical soundness and Bloom alignment. Regarding correctness, automated and human evaluations disagree: the former identifies baseline as the top performers, while the latter favors BAQ and AQ. This discrepancy can be attributed to the fact that automated evaluation favors straightforward, factually correct responses, while human evaluators tend to adopt a more holistic approach, even when instructed otherwise.

To ensure our human evaluation results are statistically significant, we performed a t-test between BAQ and AQ (the top-performing methods based on human evaluation). The test showed no statistical significance for deepseek-r1 (p-values: 0.35, 0.32, 0.26 for correctness, Bloom alignment, and pedagogical soundness, respectively), but significant differences for gpt-4o-mini (p-values: 0.002, 0.032, and 0.033). This outcome aligns with expectations, as differences are typically narrower in reasoning-optimized models.

5.4 Performance per model

Our experiments reveal systematic trade-offs across models and methods. For deepseek-r1, BAQ achieves near-perfect correctness (94.99), the highest Bloom alignment (92.75), and strong pedagogy (88.75)—while the Baseline prioritizes correctness (96.16) at the cost of pedagogy (76.16), and AQ lags in Bloom alignment (87.00) despite competitive pedagogy (89.83). Llama3.1 70B exhibits stark contrasts: BAQ balances correctness (91.16) with robust Bloom alignment (86.83) and pedagogy (79.49), whereas the Baseline collapses pedagogically (53.41) despite extreme correctness (96.66), and AQ struggles across metrics (correctness: 79.91, Bloom alignment: 72.08, pedagogy: 66.75). Smaller models like Llama3.1 8B rely on BAQ’s scaffolding to stabilize pedagogy (70.83) and Bloom alignment (82.33) despite a correctness dip (89.91), while AQ’s higher correctness (93.41) sacrifices pedagogy (63.66) and alignment (78.41), and the Baseline’s pedagogy plummets (49.50) despite high accuracy (95.75). GPT-4o-mini mirrors this pattern: BAQ balances correctness (92.50) and pedagogy (77.83) with strong Bloom alignment (89.08), while the Baseline’s pedagogy collapses (48.58) despite high correctness (93.99), and AQ trails in pedagogy (72.08) and alignment (80.08). Critically, BAQ’s explicit scaffolding resolves the correctness-pedagogy trade-off universally, outperforming AQ in Bloom alignment and avoiding the Baseline’s pedagogical failures.

6 Ablations and extended analysis

6.1 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 taxon-

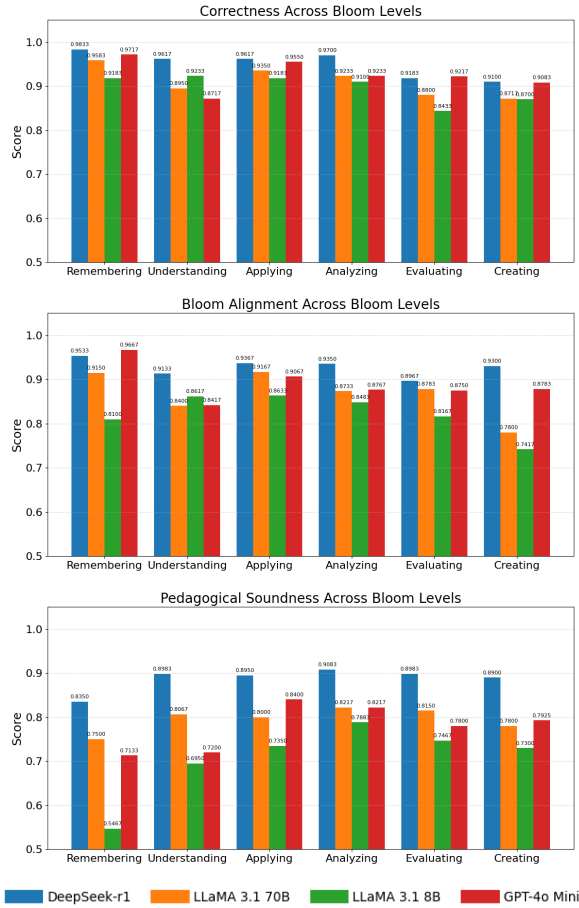


Figure 3: Performance of BAQ across Bloom’s levels with four LLMs. We measure a) Correctness (top), b) Bloom alignment (middle), c) Pedagogical soundness (bottom)

omy 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.

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

Table 4: 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

6.2 Comparison with CoT

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 benchmark⁵. The results of this comparison are presented in Table 4. 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 E. 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

⁵This choice was made to limit API costs

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.

Model	Correctness	Bloom Align.	Pedag. Sound.
Deepseek-r1	89.17	82.75	87.91
Llama 3.1 70b	76.67	69.91	75.75
Llama 8b	62.33	59.50	55.99
GPT-4o-mini	69.50	58.75	58.00

Table 5: Planning results

6.3 Planning

Since planning is a valuable strategy in both LLM reasoning and educational applications, we compared our best-performing method with the baseline in planning generation. The model was provided with the question and Bloom level (but not the answer) and was tasked with generating a solution plan aligned with that level without disclosing the answer. The results are presented in Table 5. Our findings indicate that, in the absence of an answer and within a more abstract task, the results exhibit greater variability. Regarding correctness, Deepseek-r1 significantly outperformed the other models (89.17), underscoring the superiority of reasoning-optimized models in abstract tasks. It was followed by Llama 3.1 70b (76.67), gpt-4o-mini (69.50), and Llama 3.1 8b (62.33). These marked differences suggest that larger and more sophisticated LLMs have a distinct advantage in such scenarios. A similar trend is observed in Bloom alignment and pedagogical soundness. Notably, for Bloom alignment, the two smallest models achieve comparable scores (Llama 3.1 8b: 59.50 vs. gpt-4o-mini: 58.75), while the reverse is true for ped-

agogical soundness (Llama 3.1 8b: 55.99 vs. gpt-4o-mini: 58.00).

7 Conclusions

We introduce BloomXplain, a novel framework designed to generate and evaluate explanations aligned with Bloom’s Taxonomy. Our approach begins with the creation of a STEM benchmark dataset consisting of question-answer (QA) pairs annotated with their corresponding Bloom’s Taxonomy levels. Additionally, we design two prompting strategies for LLM explanation generation: BAQ, where the Bloom level is provided and the LLM is prompted to generate a level-specific explanation, and AQ, where the LLM receives a general prompt covering all Bloom levels and is prompted to infer the Bloom’s level and then generate a Bloom-aligned explanation. We conduct extensive benchmarking of these two methods across four widely-used LLMs, evaluating their performance using an LLM-as-a-judge approach, the validity of which is supported with human validations. Our evaluation metrics focus on three key aspects: explanation correctness, alignment with Bloom’s taxonomy, and pedagogical soundness. Our findings indicate that the BAQ method consistently achieves an optimal balance between correctness and pedagogical soundness, outperforming a baseline approach that uses a generic explanation prompt. In contrast, the AQ method, which relies on automatic Bloom-level inference, demonstrates lower performance. Among the models tested, the reasoning-optimized Deepseek-r1 consistently outperforms others. Moreover, we conduct an extensive analysis of performance across levels of Bloom’s taxonomy and LLMs’ aptitude in Bloom-level-aligned planning, offering insights into their respective strengths and weaknesses. Furthermore, we benchmark our best-performing method (BAQ) against the widely-used Chain-of-Thought (CoT) prompting approach on previously Bloom-annotated tasks from popular datasets. Our results show that BAQ not only matches but occasionally surpasses CoT in terms of accuracy. Finally, we explore the capabilities of our method in generating plans.

Our comparison between BAQ and CoT yielded actionable insights, highlighting that further exploration of Bloom-aligned benchmarking and prompt design is a promising direction for future research.

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.

Ethics Statement

The authors declare no known conflict of interests. We acknowledge that the education domain is a sensitive area for the deployment of AI systems. In this work, we focus on generating pedagogically sound explanations through prompting methods, with the goal of supporting—rather than replacing—human educators. Our work conforms to the [ACL Ethics Policy](#).

References

- Hasan Abu-Rasheed, Mareike Dornhöfer, Christian Weber, Gábor Kismihók, Ulrike Buchmann, and Madjid Fathi. 2023. [Building contextual knowledge graphs for personalized learning recommendations using text mining and semantic graph completion](#). In *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*, page 36–40. IEEE.
- L. W. Anderson and D. R. Krathwohl. 2001. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom’s Taxonomy of Educational Objectives*. Longman.
- Agnes Axelsson and Gabriel Skantze. 2023. [Using large language models for zero-shot natural language generation from knowledge graphs](#). In *Proceedings of the Workshop on Multimodal, Multilingual Natural Language Generation and Multilingual WebNLG Challenge (MM-NLG 2023)*, pages 39–54, Prague, Czech Republic. Association for Computational Linguistics.
- Ashwin Bhat and Arijit Raychowdhury. 2023. [Non-uniform interpolation in integrated gradients for low-latency explainable-ai](#). *Preprint*, arXiv:2302.11107.
- Yuyan Chen, Chenwei Wu, Songzhou Yan, Panjun Liu, and Yanghua Xiao. 2024. [Dr.Academy: A benchmark for evaluating questioning capability in education for large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*,

pages 3138–3167, Bangkok, Thailand. Association for Computational Linguistics.

- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. [Training verifiers to solve math word problems](#). *CoRR*, abs/2110.14168.
- DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *Preprint*, arXiv:2501.12948.
- Malin Eiband, Hanna Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug, and Heinrich Hussmann. 2018. [Bringing transparency design into practice](#). pages 211–223.
- Mary Forehand and 1 others. 2005. Bloom’s taxonomy: Original and revised. *Emerging Perspectives on Learning, Teaching, and Technology*, 8:41–44.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Thomas Huber and Christina Niklaus. 2025. [LLMs meet bloom’s taxonomy: A cognitive view on large language model evaluations](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5211–5246, Abu Dhabi, UAE. Association for Computational Linguistics.
- Kevin Hwang, Sai Challagundla, Maryam M. Aloomair, Lujie Karen Chen, and Fow-Sen Choa. 2023. 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.
- 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. 2024. Dspy: Compiling declarative language model calls into self-improving pipelines.
- Kristjan-Julius Laak and Jaan Aru. 2025. [Ai and personalized learning: bridging the gap with modern educational goals](#). *Preprint*, arXiv:2404.02798.
- Tengxiao Liu, Qipeng Guo, Yuqing Yang, Xiangkun Hu, Yue Zhang, Xipeng Qiu, and Zheng Zhang. 2023a. [Plan, verify and switch: Integrated reasoning with diverse X-of-thoughts](#). In *Proceedings of the 2023*

728		Conference on Empirical Methods in Natural Language Processing, pages 2807–2822, Singapore. Association for Computational Linguistics.	785
729			786
730			787
731	Tommy Liu and Amanda S Barnard. 2021.	Fast	788
732		derivation of shapley based feature importances	789
733		through feature extraction methods for nanoinformatics. <i>Machine Learning: Science and Technology</i> ,	790
734		2(3):035034.	
735			
736	Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang,		791
737	Ruochen Xu, and Chenguang Zhu. 2023b.	G-eval:	792
738		NLG evaluation using gpt-4 with better human alignment.	793
739		In <i>Proceedings of the 2023 Conference on</i>	794
740		<i>Empirical Methods in Natural Language Processing</i> ,	795
741		pages 2511–2522, Singapore. Association for Com-	796
742		putational Linguistics.	797
743			798
744	Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler		
745	Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon,		
746	Nouha Dziri, Shrimai Prabhunoye, Yiming Yang,		
747	Shashank Gupta, Bodhisattwa Prasad Majumder,		
748	Katherine Hermann, Sean Welleck, Amir Yazdan-		
749	bakhsh, and Peter Clark. 2023.	Self-refine: It-	
750		erative refinement with self-feedback. <i>Preprint</i> ,	
		arXiv:2303.17651.	
751	Pratyusha Maiti and Ashok K. Goel. 2024.	How do	799
752		students interact with an llm-powered virtual teach-	800
753		ing assistant in different educational settings? In	801
754		<i>Proceedings of the 17th International Conference on</i>	802
755		<i>Educational Data Mining (EDM 2024), Workshop</i>	803
756		<i>on Leveraging Large Language Models for Next-</i>	
757		<i>Generation Educational Technologies</i> . CEUR Work-	
758		shop Proceedings, Volume 3840.	
759	David Martens, James Hinns, Camille Dams, Mark Ver-		
760	gouwen, and Theodoros Evgeniou. 2025.	Tell me a	
761		story! narrative-driven xai with large language mod-	
762		els. <i>Decision Support Systems</i> , 191:114402.	
763	Chee Ng and Yuen Fung. 2024.	Educational person-	804
764		alized learning path planning with large language	805
765		models. <i>Preprint</i> , arXiv:2407.11773.	806
766	Behnam Rahdari, Hao Ding, Ziwei Fan, Yifei Ma,		807
767	Zhuotong Chen, Anoop Deoras, and Branislav Kve-		808
768	ton. 2024.	Logic-scaffolding: Personalized aspect-	809
769		instructed recommendation explanation generation	810
770		using llms. In <i>Proceedings of the 17th ACM Interna-</i>	811
771		<i>tional Conference on Web Search and Data Mining</i> ,	
772		WSDM '24, page 1078–1081, New York, NY, USA.	
773		Association for Computing Machinery.	
774	Tuval Raz, Simone Luchini, Roger Beaty, and Yoed		
775	Kenett. 2023.	Automated scoring of open-ended	
776		question complexity: A large language model ap-	
777		proach.	
778	Marco Tulio Ribeiro, Sameer Singh, and Carlos		
779	Guestin. 2016.	"why should i trust you?": Explain-	
780		ing the predictions of any classifier. In <i>Proceedings</i>	
781		<i>of the 22nd ACM SIGKDD International Conference</i>	
782		<i>on Knowledge Discovery and Data Mining</i> , KDD '16,	
783		page 1135–1144, New York, NY, USA. Association	
784		for Computing Machinery.	
	Prithish Sahu, Michael Cogswell, Ajay Divakaran, and		
	Sara Rutherford-Quach. 2021.	Comprehension based	
		question answering using bloom's taxonomy. In	
		<i>Proceedings of the 6th Workshop on Representation</i>	
		<i>Learning for NLP (RepL4NLP-2021)</i> , pages 20–28,	
		Online. Association for Computational Linguistics.	
	Nidhin Scaria, Dharani Chenna, and Dhanamjayulu Sub-		
	ramani. 2024.	Automated educational question gen-	
		eration 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, <i>Artificial Intelligence in Educa-</i>	
		<i>tion. AIED 2024</i> , volume 14830 of <i>Lecture Notes in</i>	
		<i>Computer Science</i> . Springer, Cham.	
	Tasmia Shahriar, Kelly Ramos, and Noboru Mat-		
	suda. 2024.	Assertion enhanced few-shot learn-	
		ing: Instructive technique for large language mod-	
		els to generate educational explanations. <i>Preprint</i> ,	
		arXiv:2312.03122.	
	Mirac Suzgun, Nathan Scales, Nathanael Schärli, Se-		
	bastian Gehrman, Yi Tay, Hyung Won Chung,		
	Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny		
	Zhou, and Jason Wei. 2023.	Challenging BIG-bench	
		tasks and whether chain-of-thought can solve them.	
		In <i>Findings of the Association for Computational Lin-</i>	
		<i>guistics: ACL 2023</i> , pages 13003–13051, Toronto,	
		Canada. Association for Computational Linguistics.	
	Mirac Suzgun, Nathan Scales, Nathanael Schärli, Se-		
	bastian Gehrman, Yi Tay, Hyung Won Chung,		
	Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi,		
	Denny Zhou, and Jason Wei. 2022.	Challenging	
		big-bench tasks and whether chain-of-thought can	
		solve them. <i>Preprint</i> , arXiv:2210.09261.	
	Sean Tull, Robin Lorenz, Stephen Clark, Ilyas Khan,		
	and Bob Coecke. 2024.	Towards compositional inter-	
		pretability for xai. <i>Preprint</i> , arXiv:2406.17583.	
	Sabine Ullrich and Michaela Geierhos. 2023.	Using	
		bloom's taxonomy to classify question complexity.	
		In <i>Proceedings of the Research Institute CODE</i> , Ger-	
		many. Universität der Bundeswehr München.	
	Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten		
	Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and		
	Denny Zhou. 2023.	Chain-of-thought prompting elic-	
		its reasoning in large language models. <i>Preprint</i> ,	
		arXiv:2201.11903.	
	Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten		
	Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le,		
	and Denny Zhou. 2022.	Chain-of-thought prompt-	
		ing elicits reasoning in large language models. In	
		<i>Proceedings of the 36th International Conference on</i>	
		<i>Neural Information Processing Systems, NIPS '22</i> ,	
		Red Hook, NY, USA. Curran Associates Inc.	
	Yeo Wei Jie, Ranjan Satapathy, Rick Goh, and Erik		
	Cambria. 2024.	How interpretable are reasoning ex-	
		planations from prompting large language models?	
		In <i>Findings of the Association for Computational Lin-</i>	
		<i>guistics: NAACL 2024</i> , pages 2148–2164, Mexico	

City, Mexico. Association for Computational Linguistics.

A Dataset creation prompts

The prompt templates and few-shot examples used for data generation are provided in the Tables 6 and 7.

B Explanation/planning creation prompts

The prompts and few-shot examples for explanation (BAQ and AQ) and planning generation can be found in Tables 8, 9, 10, 11 and 12.

C GEval Evaluation Criteria

The criteria used for the evaluation (both human and automated) are shown in Table 13.

D Baseline Prompt and few-shot examples

The baseline prompt and few shot examples are shown in Table 14.

E Prompts and few-shot examples for widely used benchmarks

Few-shots examples of BAQ from widely used benchmarks are shown in Tables 15, 16, 17 and 18.

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 ‘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.

F 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 19. 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.

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.

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.

Table 6: Prompt templates used for QA pairs’ creation

Question	Answer
What is the formula for the area of a triangle?	$(1/2) \times \text{base} \times \text{height}$
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?

Table 7: Few-shot examples for dataset creation

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 + Bloom Level, Level-Specific Prompt-BAQ)

Generic prompt	
Tutor Guidance	<p>You are a helpful tutor. Use the following guidance to craft your explanation depending on the level of the question:</p> <ul style="list-style-type: none"> • 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.

Table 9: Prompts for explanations' creation (Question + Answer, Generic Prompt-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, <code>print('Hello')</code> 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: Few-shot examples used in BAQ and AQ

Prompt Type	Prompt Text
Instructional Planning	You are a tutor and I am a student. Write a step-by-step instructional plan to help me work toward the answer independently. Do NOT include the answer or try to guess it. Focus only on guiding my thinking.
Remembering	This is a Remembering-level question. Guide me in recalling facts, definitions, or basic information, and encourage using simple strategies to remember or mnemonics.
Understanding	This is an Understanding-level question. Guide me to put concepts into my own words, explain, summarize, or give examples to demonstrate comprehension.
Applying	This is an Applying-level question. Guide me on how to use prior knowledge in a new situation.
Analyzing	This is an Analyzing-level question. Guide me to break down the concept into parts, find patterns, or relationships.
Evaluating	This is an Evaluating-level question. Guide me to assess or justify a position or solution.
Creating	This is a Creating-level question. Guide me through combining ideas to produce something new.

Table 11: Instructional Planning Prompt

Bloom Level	Instructional Plan
Remembering	Q: What is the chemical symbol for sodium? Start by thinking about where you've seen sodium on the periodic table. Try to recall if it has a symbol that matches its English name or if it uses a Latin root. Think back to any mnemonics or tricks you've used to memorize common element symbols — especially ones that seem different from the name.
Understanding	Q: Explain why the heart has four chambers. Begin by identifying what each of the four chambers does. Think about how blood flows through the heart and lungs, and how oxygen is picked up and delivered. Try to put into your own words how this structure supports the function of the circulatory system, especially in separating oxygenated and deoxygenated blood.
Applying	Q: Use the Pythagorean theorem to find the length of the hypotenuse in a right triangle with legs of 3 and 4 units. Recognize that this is a right triangle, so the Pythagorean theorem applies. Identify the two legs of the triangle and assign them to a and b. Plug these into the formula $a^2 + b^2 = c^2$, calculate the sum, and then take the square root to find the hypotenuse. Show each step of the process clearly.
Analyzing	Q: Compare mitosis and meiosis in terms of genetic variation and number of resulting cells. Start by breaking down the purpose and process of mitosis and meiosis separately. Think about how many cells each one produces and what happens to the genetic material. Then compare the results — look at patterns, such as whether the cells are identical or unique, and how many are produced. Focus on key differences and what they imply about each process.
Evaluating	Q: Which programming language would you recommend for beginners and why? Begin by considering what makes a programming language beginner-friendly — factors like readability, simplicity, available learning resources, and how widely it's used. Think about a few popular options and weigh the pros and cons of each. Use reasoning to support your recommendation, rather than just stating a preference.
Creating	Q: Design an experiment to test the effect of light on plant growth. Start by defining the purpose of the experiment — what exactly are you trying to find out? Then decide what variables you'll test (e.g., type of light) and what you'll keep constant (e.g., plant type, soil, water). Think about how you'll measure plant growth and how long you'll run the experiment. Put together a step-by-step plan that someone else could follow to carry it out.

Table 12: Few-shot Planning Examples

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 13: 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. For Planning, we applied the same criteria as BAQ, replacing the term “explanation” with “planning” and including a disclaimer stating that the final answer should not be revealed. The same criteria were used as guidelines for human evaluators.

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

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 basic counting: $1+1+1+1+1=6$. 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 15: BAQ Few-Shot Examples for BBH Object Counting 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 16: BAQ Few-Shot Examples for BBH Disambiguation QA

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 \times \$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 \times 4 = 20. Then add the new computers to the original: 9 + 20 = 29. The final answer is 29.

Table 17: BAQ Few-Shot Examples for GSM

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 or 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).

Table 18: BAQ Few-Shot Examples for BBH Snarks

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

Table 19: Results for multi-Bloom level problems