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

NEURAL INFORMATION PROCESSING SYSTEMS

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 [2] National Technical University of Athens

In a nutshell

We introduce a framework and a STEM-benchmark dataset for Pedagogically sound LLM-generated explanations based on Bloom's Taxonomy.

Motivation

Why explanations?

- Useful for tutoring
- Better explanations ⇒ better reasoning

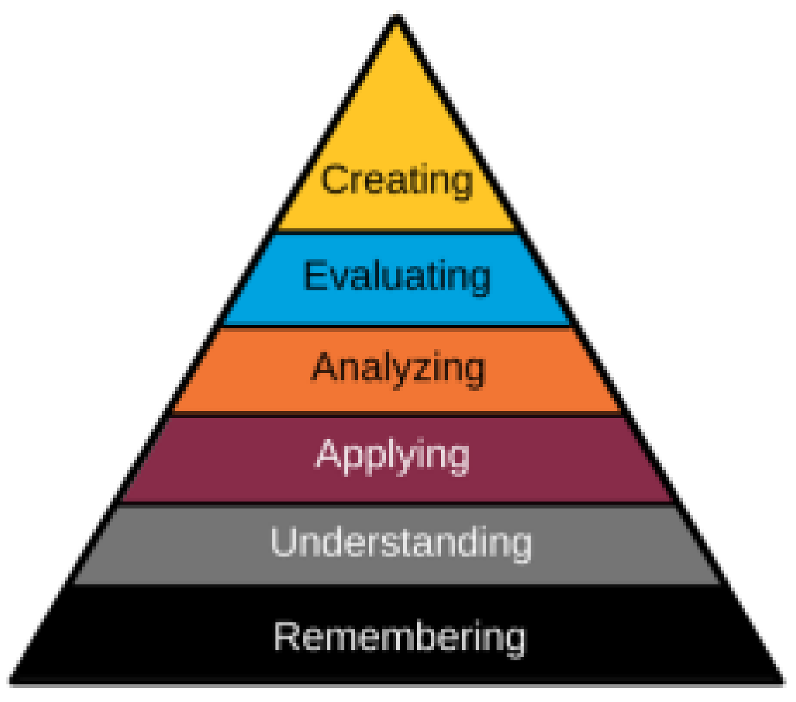
Why Bloom's Taxonomy?

- Structured framework
- Human-aligned, explainable results

Contributions

- A STEM QA benchmark dataset annotated with Bloom levels
- Prompting strategies for eliciting Bloom-aligned explanations
- A multi-metric evaluation of LLMs across Correctness, Bloom Alignment, and Pedagogical Soundness

Background: Bloom's Taxonomy



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

Methodology

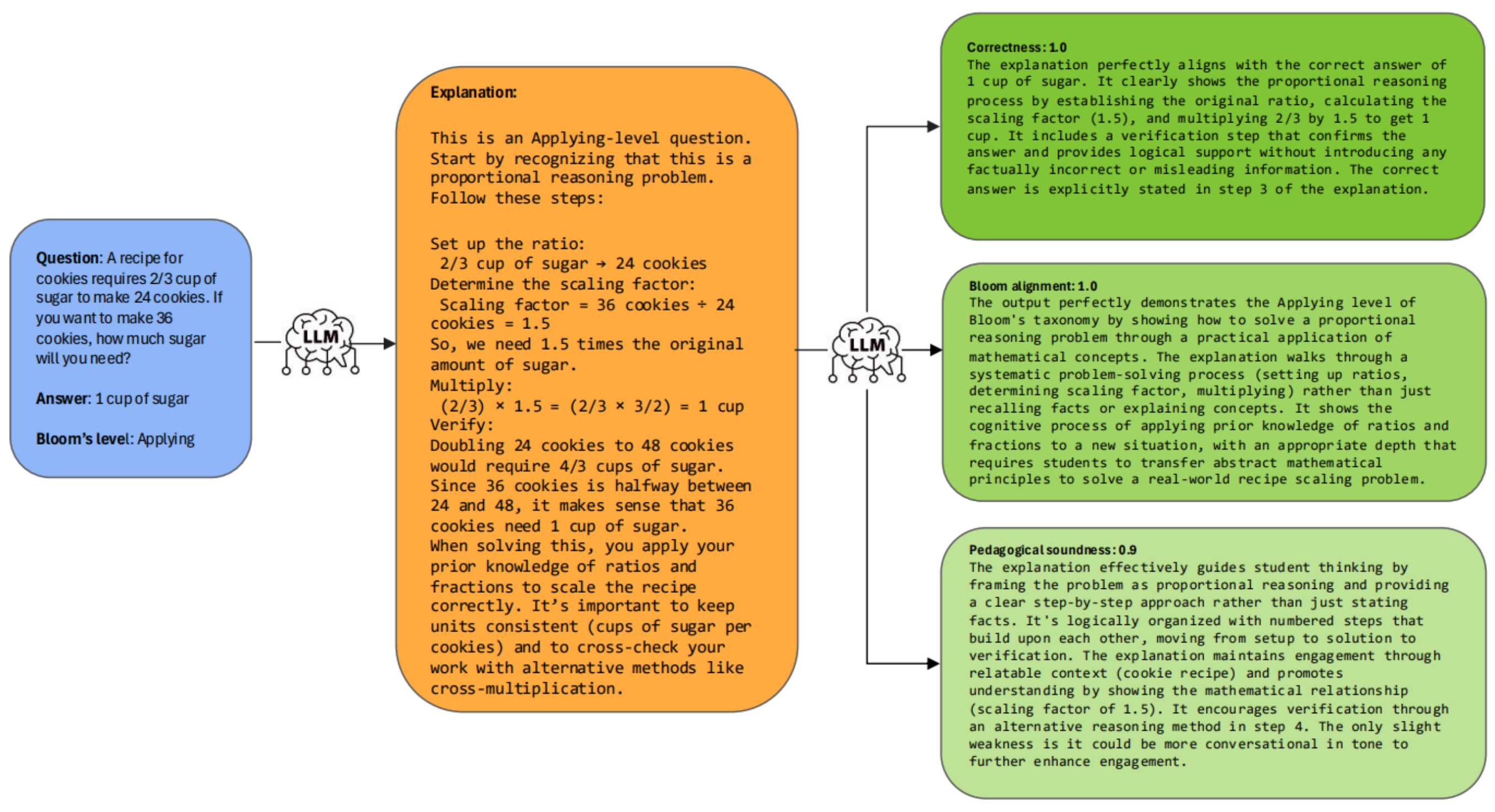
Dataset: 360 Bloom-aligned STEM QA pairs, spanning elementary → undergraduate, LLM-generated (Claude 3.7 sonnet) and human-validated

Prompting strategies:

Prompting strategy	Input	Output
BAQ	Question, Answer, Bloom's level	Bloom-aligned explanation
AQ	Question, Answer	Inferred Bloom's level, Bloom-aligned explanation
Baseline	Question, Answer	Explanation

Evaluation: LLM-as-a-Judge (Claude 3.7 sonnet) and human evaluation across three criteria: Correctness, Bloom Alignment, and Pedagogical Soundness

Framework (BAQ)



Question: A recipe for cookies requires 2/3 cup of sugar to make 24 cookies. If you want to make 36 cookies, how much sugar will you need?

Answer: 1 cup of sugar

Bloom’s level: Applying

Explanation:

```
This is an Applying-level question.
Start by recognizing that this is a
proportional reasoning problem.
Follow these steps:

Set up the ratio:
 2/3 cup of sugar → 24 cookies
Determine the scaling factor:
 Scaling factor = 36 cookies ÷ 24
cookies = 1.5
So, we need 1.5 times the original
amount of sugar.
Multiply:
 (2/3) × 1.5 = (2/3 × 3/2) = 1 cup
Verify:
Doubling 24 cookies to 48 cookies
would require 4/3 cups of sugar.
Since 36 cookies is halfway between
24 and 48, it makes sense that 36
cookies need 1 cup of sugar.
When solving this, you apply your
prior knowledge of ratios and
fractions to scale the recipe
correctly. It’s important to keep
units consistent (cups of sugar per
cookies) and to cross-check your
work with alternative methods like
cross-multiplication.
```

Correctness: 1.0
The explanation perfectly aligns with the correct answer of 1 cup of sugar. It clearly shows the proportional reasoning process by establishing the original ratio, calculating the scaling factor (1.5), and multiplying 2/3 by 1.5 to get 1 cup. It includes a verification step that confirms the answer and provides logical support without introducing any factually incorrect or misleading information. The correct answer is explicitly stated in step 3 of the explanation.

Bloom alignment: 1.0
The output perfectly demonstrates the Applying level of Bloom's taxonomy by showing how to solve a proportional reasoning problem through a practical application of mathematical concepts. The explanation walks through a systematic problem-solving process (setting up ratios, determining scaling factor, multiplying) rather than just recalling facts or explaining concepts. It shows the cognitive process of applying prior knowledge of ratios and fractions to a new situation, with an appropriate depth that requires students to transfer abstract mathematical principles to solve a real-world recipe scaling problem.

Pedagogical soundness: 0.9
The explanation effectively guides student thinking by framing the problem as proportional reasoning and providing a clear step-by-step approach rather than just stating facts. It's logically organized with numbered steps that build upon each other, moving from setup to solution to verification. The explanation maintains engagement through relatable context (cookie recipe) and promotes understanding by showing the mathematical relationship (scaling factor of 1.5). It encourages verification through an alternative reasoning method in step 4. The only slight weakness is it could be more conversational in tone to further enhance engagement.

Main Results

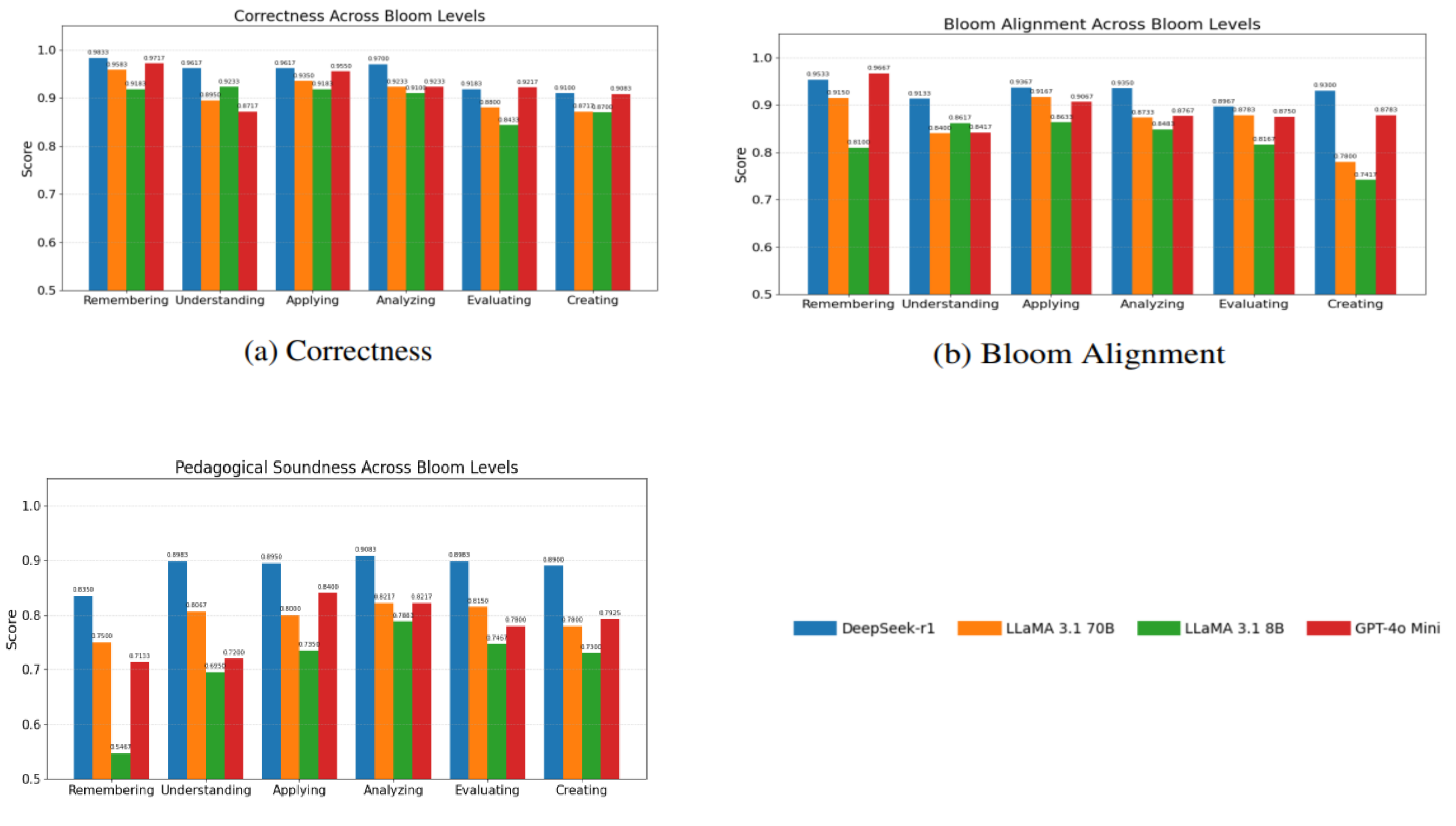
Model	Method	Correctness	Bloom Alignment	Pedagogical Soundness	Overall Score
deepseek-r1	BAQ	94.99	**92.75**	88.75	**92.00**
	AQ	93.75	87.00	**89.83**	90.00
	Baseline	**96.16**	-	76.16	85.99
llama3.1 70b	BAQ	91.16	**86.83**	**79.49**	**85.66**
	AQ	79.91	72.08	66.75	73.00
	Baseline	**96.66**	-	53.41	75.08
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

- BAQ outperforms other methods in pedagogical soundness and Bloom alignment while maintaining high correctness.
- AQ, which infers Bloom levels, underperforms BAQ in both Bloom alignment and pedagogical soundness
- While Baseline scores highest in correctness, its lack of structure leads to the lowest pedagogical score

BAQ’s explicit Bloom-level guidance achieves the best balance of pedagogical depth and factual accuracy.

Reasoning-optimized models achieve strong performance overall, while other models exhibit a much sharper pedagogy–correctness trade-off.

BAQ's Performance per Bloom’s level



(a) Correctness

(b) Bloom Alignment

(c) Pedagogical Soundness

- **Correctness:** Deepseek-r1 consistently leads across all Bloom levels, with GPT-4o-mini and LLaMA-3.1-70B close behind in most cases; performance drops for all models at higher cognitive levels (e.g., Evaluating).
- **Bloom Alignment:** Deepseek-r1 also achieves the strongest alignment, with GPT-4o-mini and LLaMA-3.1-70B performing similarly; LLaMA-3.1-8B generally lags except in Understanding tasks.
- **Pedagogical Soundness:** Deepseek-r1 again ranks highest, followed by GPT-4o-mini and LLaMA-3.1-70B; LLaMA-3.1-8B shows the weakest pedagogy, indicating smaller models struggle to provide instructional explanations.

Comparison with CoT on widely used benchmarks (100 samples/task)

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**

BAQ achieves competitive or superior performance compared to Chain-of-Thought (CoT) across Bloom’s taxonomy levels, validating its efficacy in fostering robust reasoning

Funded by the European Union

AIXPERT

ATHENA Research & Innovation Information Technologies

Workshop on Evaluating the Evolving LLM Lifecycle – Benchmarks, Emergent Abilities, and Scali