CogSTEM: A Bloom's Taxonomy-Grounded Benchmark for Diagnosing High-Order Capabilities in Large Language Models

Anonymous ACL submission

Abstract

The rapid evolution of Large Language Models (LLMs) has sparked urgent demand for their integration as intelligent teaching assistants in STEM education. However, existing benchmarks often exhibit severe *distributional biases*, focusing disproportionately on factual recall or narrow procedural reasoning while neglecting the assessment of cognitive abilities essential for educational contexts. To address this, we introduce **CogSTEM**, a bilingual benchmark strictly aligned with the Revised Bloom's Taxonomy to achieve multi-dimensional equilibrium. Constructed through a rigorous human-in-the-loop annotation process, CogSTEM comprises 4,491 high-quality samples that evaluate models across Disciplinary, Knowledge, and Cognitive dimensions. Our extensive evaluation reveals a critical cognitive disparity: while models excel in foundational "Remembering" tasks, they struggle significantly with high-order "Analyzing" problems, with even SOTA models facing substantial challenges. Furthermore, we demonstrate CogSTEM's practical utility via fine-tuning; experimental results show that Qwen series models achieve significant gains—specifically a **7.90%** surge in high-order evaluation capabilities—without compromising general proficiency. CogSTEM serves as a rigorous diagnostic framework for assessing and enhancing LLMs.

1 Introduction

With the rapid development of LLMs, frontier models such as GPT-5, Gemini 3, and DeepSeek R1 (Guo et al., 2025) have demonstrated remarkable capabilities in complex reasoning and multimodal processing. This progress has sparked an urgent demand for leveraging them as intelligent teaching assistants in STEM (Science, Technology, Engineering, and Mathematics) education to facilitate student learning.

The core objective of STEM education extends beyond the mere acquisition of factual knowledge; it necessitates the cultivation of logical reasoning, problem-solving skills, and the ability to transfer acquired knowledge to novel contexts. Therefore, evaluating the proficiency of LLMs within the STEM domain serves not only as a metric of model intelligence but also as a decisive factor in their suitability as qualified educational agents.

To measure these capabilities, the research community has established several milestone benchmarks. Comprehensive evaluations such as MMLU (Hendrycks et al., 2020) and C-Eval (Huang et al., 2023) establish benchmarks for assessing broad, largely static knowledge. In parallel, with respect to reasoning depth, datasets such as GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021), and AIME focus on multi-step mathematical reasoning, effectively measuring the ability to solve complex multi-step reasoning problems through structured derivations. These benchmarks typically frame models as specialized mathematical solvers or coding assistants, rather than general reasoning agents.

However, a critical gap remains between these extremes. As illustrated in Table 1, a qualified educational agent must navigate a complex cognitive spectrum that current benchmarks fail to systematically cover. Existing benchmarks tend to emphasize different regions of this spectrum, challenging models' capabilities in either broad academic knowledge or complex procedural reasoning within specific domains. However, they generally lack a fine-grained framework for evaluating the intermediate yet crucial cognitive abilities required in everyday human problem-solving and knowledge application (Webb, 1997), ranging from diagnosing a circuit failure (Analyze), selecting appropriate materials for a task (Evaluate), or applying abstract concepts to physical reality (Apply). Without a pedagogical reference system, it is

Table 1: **Illustrative examples from the CogSTEM benchmark.** The dataset is structured across two primary axes: the Revised Bloom's Taxonomy (Panel A) and Knowledge Dimensions (Panel B). Unlike previous benchmarks that focus solely on factual recall or pure calculation, CogSTEM evaluates the model's ability to navigate the full pedagogical hierarchy—from basic understanding to complex evaluation in realistic scenarios.

Dimension	Problem Description
Panel A: Revised Bloom's Cognitive Levels	
Remember	The relationship between the magnitude of aircraft drag and the aircraft's flight speed is ( ).
Understand	The bell sound from the clock tower can be heard from a great distance because:
Apply	Which host IP address belongs to the same network segment as `10.110.12.29` (subnet mask `255.255.255.224`)?
Analyze	After Xiao Ming turned on the light switch, the bulb did not light up. Using a voltage tester, both terminals lit up the neon tube. This is because ( ).
Evaluate	Xiaoming wants to help his mother choose a wok for stir-frying. What material should the wok be made of?
Create	To transform a city into a highly self-healing and adaptive living system, which solution would you choose?
Panel B: Knowledge Dimensions	
Factual	According to current national regulations, what is the minimum age for operating tractors and combine harvesters?
Conceptual	Among gasoline engine, steam engine, diesel engine, and electric motor, which ones are internal combustion engines?
Procedural	The decimal number 257 is converted into binary as:

difficult to determine whether a model truly "understands" science or is merely retrieving data and executing algorithms.

To address this, we propose adopting a rigorous educational syllabus as the evaluation standard. We introduce **CogSTEM**, a benchmark grounded in the **Revised Bloom's Taxonomy** (Anderson and Krathwohl, 2001), which provides a theoretically sound framework for distinguishing different levels of human cognition. Unlike competitors that act as unstructured problem collections, CogSTEM enforces a *multi-dimensional equilibrium*. As shown in the examples in Table 1, CogSTEM distinguishes itself in three key aspects: (1) **Cognitive Systematization**: It aligns with Bloom's Taxonomy to cover the full journey from *Remembering* to *Creating*, avoiding the cognitive skewness seen in previous works; (2) **Knowledge Completeness**: It balances factual, conceptual, and procedural knowledge types, ensuring models are tested on understanding mechanisms rather than just rote memorization; and (3) **Disciplinary Coverage**: It provides substantial coverage across all four STEM pillars, correcting the marginalization of Technology and Engineering found in pure science or math datasets.

The CogSTEM dataset was constructed using a rigorous Human-in-the-loop annotation process. High-quality original problems were selected from standardized STEM examinations and authoritative textbooks ranging from high school to university levels. We assembled an annotation team of subject-matter experts who, beyond verifying content correctness, performed multi-round double-blind annotation and consistency arbitration. Empirical evaluations and fine-tuning experiments were conducted to demonstrate the dataset's reliability in diagnosing cognitive gaps and enhancing model capabilities.

The primary contributions of this paper are summarized as follows:

1. We introduce CogSTEM, a large-scale STEM dataset comprising 4,491 samples derived from authentic educational scenarios. By leveraging the Revised Bloom's Taxonomy, it serves as the first benchmark to systematically evaluate LLMs through a pedagogical lens rather than just task performance.

2. Fine-tuning Qwen series models on CogSTEM yields significant gains without compromising general capabilities: Qwen2.5-7B shows a **2.66%** overall improvement, while Qwen3-32B achieves a **7.90%** surge in the higher-order *Evaluate* dimension.

3. We evaluate diverse representative LLMs, revealing a critical cognitive disparity: models

excel at "Remember" tasks but struggle with high-order "Analyze" problems. Even SOTA Gemini 3 scores only 80.90%, indicating that substantial challenges remain in establishing true educational intelligence.

2 Related Work

2.1 General Knowledge Benchmarks

Foundational evaluations, represented by MMLU (Hendrycks et al., 2020) and C-Eval (Huang et al., 2023), established the paradigm for assessing Knowledge Breadth across diverse subjects. Subsequently, AGIEval (Zhong et al., 2024) introduced human-centric standardized tests, while recent benchmarks like MMLU-Pro (Wang et al., 2024) and GPQA (Rein et al., 2024) have elevated evaluation standards by targeting expert-level difficulty.

However, a critical limitation persists: these benchmarks typically treat difficulty as a singular scalar, often conflating foundational fact retrieval with high-order reasoning. They lack a rigorous pedagogical grounding—specifically the Revised Bloom's Taxonomy (Anderson and Krathwohl, 2001)—to disentangle cognitive levels. More importantly, existing benchmarks overlook the *distributional reality* of educational applications. Consequently, prior works suffer from *cognitive skewness*, failing to provide a valid diagnosis of a model's utility in practical pedagogical contexts. Our work aims to align with the authentic distribution of cognitive demands found in real-world educational settings.

2.2 STEM and Complex Reasoning Evaluation

Evaluations in the STEM domain have predominantly bifurcated into two extremes. On one hand, datasets like GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021) effectively quantify CoT capabilities but are largely confined to symbolic operations within closed systems. On the other hand, scientific QA datasets like SciQ (Welbl et al., 2017) focus heavily on low-level factual retrieval.

This dichotomy reveals a significant gap in measuring Procedural Knowledge Transfer—the capacity to retrieve abstract principles and apply them to solve novel problems in open scientific contexts. Existing benchmarks struggle to capture the transition from rote memorization to flexible application. To address this, we propose CogSTEM, which serves not merely as a test suite but as a diagnostic framework strictly aligned with the dual dimensions of knowledge and cognition. By integrating Bloom's taxonomy, CogSTEM bridges this gap, enabling a precise assessment of how LLMs handle the rigorous demands of high-order scientific reasoning beyond simple arithmetic derivation.

3 Dataset Construction Pipeline

To construct an evaluation benchmark that strictly adheres to the pedagogical principles of STEM education while maintaining robust discriminative power, we designed a rigorous, expert-driven pipeline featuring iterative Human-AI collaboration. As illustrated in Figure 1, the development of CogSTEM unfolds through four distinct stages: (1) broad-spectrum data acquisition, (2) multi-round model-assisted cleaning, (3) fine-grained cognitive annotation, and (4) diversity-enhancing data augmentation.

3.1 Real-World Oriented Data Collection

The data acquisition for CogSTEM extends beyond conventional K-12 assessments to encompass higher education and professional domains, with a strategic focus on interdisciplinary fields such as Medical Physics and Bioengineering. Unlike existing benchmarks such as MMLU (Hendrycks et al., 2020), which rely heavily on open-source web crawling, our primary data sources are directly derived from **non-public instructional resources**. These materials are curated by a team of 10 subject-matter experts who are currently active in STEM pedagogy.

The corpus comprises curriculum-standard textbooks and actual classroom exercises that assess a model's grasp of sophisticated theoretical foundations, such as the physical properties of ultrasound and the spectral analysis of physiological signals. Rather than focusing on abstract symbolic computation, these items emphasize the application of theoretical knowledge within authentic engineering constraints. For instance, questions are designed to compel models to apply physical principles—such as the attenuation characteristics of fiber optics—to solve practical safety and design scenarios.

To ensure a comprehensive diagnostic capability, we implemented a **hierarchical sampling strategy** strictly aligned with the Revised Bloom's Taxonomy. While many existing datasets suffer from a "cognitive collapse" into factual recall, we explic-

Dataset	Domains	Language	Explanation	Core Capability
MMLU (Hendrycks et al., 2020) C-Eval (Huang et al., 2023)	General	EN ZH	No No	Declarative knowledge & Facts
GSM8K (Cobbe et al., 2021) MATH (Hendrycks et al., 2021)	Math	EN EN	Yes Yes	Symbolic & Closed reasoning
SciQ (Welbl et al., 2017)	Science	EN	No	Fact retrieval
GPQA (Rein et al., 2024) MMLU-Pro (Wang et al., 2024)	Expert	EN EN	No No	Expert knowledge & Derivation
CogSTEM (Ours)	**STEM**	**ZH & EN**	**Yes**	**Procedural transfer & Application**

Table 2: Systematic comparison between CogSTEM and existing benchmarks across key dimensions.

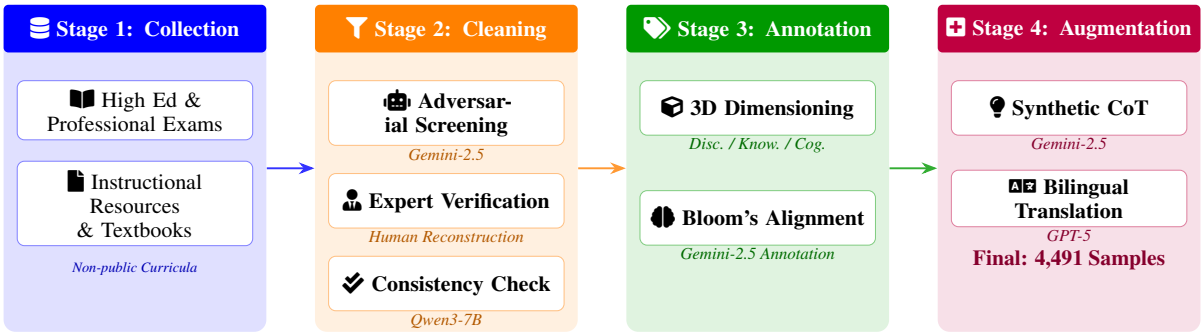

Figure 1: The construction pipeline of CogSTEM, featuring iterative Human-AI collaboration.

itly prioritized the inclusion of items necessitating *Understanding*, *Applying*, and *Analyzing*. We maintained a controlled distribution where foundational *Remembering* tasks form the base (49%) to verify knowledge breadth, while higher-order cognitive levels (Apply, Analyze, Evaluate, and Create) constitute a significant portion to evaluate reasoning depth. This structured distribution enables CogSTEM to systematically pinpoint whether a model's failure stems from a lack of factual knowledge or an inability to perform complex cognitive processing. Detailed information regarding the annotation team, 3D framework, and quality control procedures is provided in Appendix B.

3.2 Iterative Human-Model Collaborative Cleaning

To eliminate ambiguities, formatting errors, or ground-truth fallacies, we designed a Human-in-the-loop and Model-in-the-loop fusion cleaning framework. This employs a unique "dual-model, dual-human" funnel strategy:

- **Step 1: Adversarial Screening by SOTA Model.** We utilize Gemini-2.5 (Team et al., 2024a) as a "gold-standard validator" in a zero-shot setting. If a SOTA-level model fails to answer correctly, the item is flagged for potential ambiguity or error.
- **Step 2: Expert Disambiguation and Correction.** Experts review flagged samples to distinguish between model reasoning failure and inherent problem defects. For the latter, experts perform text reconstruction or condition supplementation.
- **Step 3: Consistency Check by Lightweight Model.** To prevent overfitting to specific large-model logic, we introduce Qwen3-7B (Team et al., 2024b) for secondary verification. This ensures that the problem is sufficiently clear for any model with adequate knowledge to comprehend.
- **Step 4: Final Quality Assurance.** A final expert review is conducted on items that still fail the consistency check to ensure the elimination of subjective puzzles.

3.3 Multi-Dimensional Fine-Grained Annotation

We established a "Disciplinary-Knowledge-Cognitive" 3D annotation framework based on the **Revised Bloom's Taxonomy** (Anderson and Krathwohl, 2001).

QUESTION 1

Xiaogang's optical network terminal (ONU) at home cannot go online. He wants to unplug the fiber optic cable to check if there is an optical signal. Is this action feasible?

Options:
A. Feasible, to see if there is an optical signal
B. Feasible, but it might damage the terminal
C. **Not allowed, because the laser may injure the eyes**
D. Uncertain, because the signal might not be visible

Explanation: The passive optical network (PON) uses laser light (e.g., 1490nm) which is invisible but carries high energy density.

Answer: C

Category: Technology-targeted **Knowledge:** Conceptual
Bloom: Evaluate **Field:** Fiber Optic

QUESTION 2

The water crisis forces reliance on desalination, but the toxic "brine" byproduct damages marine ecosystems. How can a chemical engineering closed-loop system be designed to transform this brine?

Options:
A. Inject the brine into volcanoes or deep underground
B. **Deploy a refining system (electrodialysis) to create a "mineral extraction facility"**
C. Construct massive evaporation ponds for industrial salt
D. Develop tech to reduce brine concentration

Explanation: Electrodialysis and chemical precipitation are key techniques. By extracting minerals, waste is transformed into commodities.

Answer: B

Category: Technology **Knowledge:** Conceptual
Bloom: Create **Field:** Chem-Engineering

Figure 2: Side-by-side comparison of two questions from the CogSTEM dataset. Both follow the dataset's structured format including problem statement, multiple-choice options, explanation, correct answer, and metadata (category, knowledge type, Bloom's taxonomy level, and field).

- **Subject Dimension:** Covers Science, Technology, Engineering, and Mathematics.
- **Knowledge Dimension:** Categorizes items into *Factual*, *Conceptual*, and *Procedural*. We place significant emphasis on **Procedural knowledge**—the techniques, methods, and criteria for applying skills.
- **Cognitive Dimension:** Defines the depth of thinking from *Remember* and *Understand* to higher-order *Apply* and *Analyze*.

We utilized Gemini-2.5 as an "expert annotator" with structured prompts. As shown in Table 3, the Knowledge Dimension achieves a near-perfect 1:1:1 ratio between Factual (1,490), Conceptual (1,510), and Procedural (1,491) items. In the Cognitive Dimension, while *Remember* (49%) forms the base, *Apply* and *Analyze* account for 30%, providing a robust sample for testing flexible application.

3.4 Data Augmentation

To extend the benchmark's utility for reasoning analysis and multilingual settings, we augmented the dataset with the following components:

Synthetic Chain-of-Thought: We utilized Gemini-2.5, validated by human experts, to generate detailed reasoning paths. These paths decompose problems into *Knowledge Localization*, *Logical Deduction*, and *Distractor Analysis*, supporting rigorous CoT evaluation (Wei et al., 2022).

High-Fidelity Parallel Translation: Using GPT-5, we created a high-quality parallel Chinese-English version of the dataset. This addition serves to evaluate Cross-Lingual Consistency, ensuring that model performance reflects genuine subject mastery rather than language-specific overfitting.

4 Experiments

4.1 Experimental Setup

To conduct our experiments, we selected a diverse range of representative models. Our evaluation set includes the open-source Qwen2.5-7B and the Qwen3 series (spanning 8B, 32B, and 235B parameters) (Team et al., 2024b; Yang et al., 2025), alongside the SOTA proprietary models GPT-5.1 and Gemini 3 (Team et al., 2024a).

To explore the potential of targeted optimization for higher-order cognitive abilities, we partitioned the CogSTEM dataset to ensure uniformity across multiple classification dimensions. This yielded 1,000 high-quality training examples (*CogSTEM-Train*) distinct from the evaluation set, while the remaining 3,491 examples were reserved for testing (*CogSTEM-Test*). Building upon *CogSTEM-Train*, we applied Group Relative Policy Optimization (GRPO) (Shao et al., 2024) to Qwen2.5-7B, Qwen3-8B, and Qwen3-32B to develop the STEM-specialized versions: Qwen2.5-7B-S, Qwen3-8B-S, and Qwen3-32B-S. We conducted a comparative evaluation of these models before and after training to determine if domain-specific alignment enhances procedural knowledge transfer across different parameter scales without sacrificing general reasoning capabilities.

Dimension	Category	Description	Train	Test	Total
Subject	Science	Natural sciences (Physics, Chemistry, Biology, etc.)	479	1,676	2,155
	Technology	Computing, programming, and tool usage	143	492	635
	Engineering	Structural design, electronics, and mechanical systems	307	1,064	1,371
	Mathematics	Abstract quantities, structures, and change	71	259	330
Knowledge[†]	Factual	Terminology, specific elements, and details	332	1,158	1,490
	Conceptual	Classifications, principles, and theoretical models	337	1,173	1,510
	Procedural	Subject-specific skills, algorithms, and methodologies	331	1,160	1,491
Cognitive[‡]	Remember	Recognizing or retrieving information	491	1,723	2,214
	Understand	Explaining meanings, exemplifying, or classifying	172	602	774
	Apply	Executing procedures in given situations	240	830	1,070
	Analyze	Breaking down material and determining relationships	71	236	307
	Evaluate	Judging based on criteria and standards	21	76	97
	Create	Assembling elements into a new functional whole	5	24	29

[†,‡] Categorization criteria are aligned with the Revised Bloom's Taxonomy (Anderson and Krathwohl, 2001).

Table 3: Detailed statistics of CogSTEM dataset. The table shows the distribution of samples across Subject, Knowledge, and Cognitive dimensions, partitioned into training and testing sets.

Implementation Details. We fine-tuned the models using the GRPO algorithm with a group sampling size of $G = 5$ to estimate the baseline. The training process spanned 3 epochs with a global batch size of 128 and a learning rate of 1×10^{-6}. To accommodate the extensive reasoning chains required for STEM problems, we set the maximum prompt and response lengths to 2,048 and 4,096 tokens, respectively. Stability was maintained using a KL penalty coefficient of $\beta = 0.001$ and a PPO clip ratio of 0.2. All experiments were conducted on a server node equipped with 8 NVIDIA H200 GPUs.

4.2 Experimental Results and Analysis

We evaluate the performance of various Large Language Models (LLMs) on CogSTEM across three key dimensions: Knowledge, Cognitive, and Disciplinary. Table 4 and Table 5 present the detailed results on the English and Chinese versions, respectively, while Table 6 illustrates the impact of STEM-tuning on general capabilities.

4.2.1 Overall Performance Comparison on Base Models

As shown in Table 4 and Table 5, **Gemini 3** consistently achieves the highest total scores across both languages (79.78 in English and 80.90 in Chinese), establishing a strong state-of-the-art baseline.

Scaling Law Anomalies and Model Efficiency A notable observation is the performance of **Qwen3-32B**. Despite having significantly fewer parameters, it consistently outperforms the much larger **Qwen3-235B** and the proprietary **GPT-5.1** in both language settings. For instance, on the English dataset, Qwen3-32B achieves a score of 77.84, surpassing Qwen3-235B (72.02) by over 5 points. This suggests that for STEM-specific reasoning, model architecture and training data quality may play a more critical role than sheer parameter scale.

Language Capability Disparity We observe a distinct performance gap between languages for certain models. While **GPT-5.1** remains competitive in English (76.93), its performance drops precipitously to 64.41 on the Chinese benchmark, lagging behind even the 7B-parameter model in the fine-tuned setting. in contrast, the **Qwen se-**

Table 4: Performance evaluation on the **English version** of CogSTEM. Accuracy across three dimensions: Knowledge (Fact: Factual, Conc: Conceptual, Proc: Procedural), Cognitive (Rem: Remember, Und: Understand, App: Apply, Ana: Analyze, Eva: Evaluate, Cre: Create), and Disciplinary (Sci: Science, Tech: Technology, Eng: Engineering, Mat: Mathematics). The best and second-best results are highlighted.

Model	Total	Knowledge Dimension			Cognitive Dimension						Disciplinary Dimension			
	Score	Fact.	Conc.	Proc.	Rem.	Und.	App.	Ana.	Eva.	Cre.	Sci.	Tech.	Eng.	Mat.
Qwen2.5-7B	63.02	65.63	56.18	67.36	66.76	60.80	59.16	53.39	65.79	70.83	64.10	69.51	60.06	55.98
Qwen3-8B	74.05	76.25	69.57	76.40	76.10	71.76	74.58	63.56	71.05	79.17	76.03	77.64	69.36	73.75
Qwen3-32B	77.84	81.78	72.21	79.59	79.52	75.58	78.54	68.64	76.32	83.33	80.56	81.71	72.37	75.29
Qwen3-235B	72.02	78.58	65.22	72.35	76.16	67.61	70.72	60.59	64.47	66.67	73.76	77.03	69.08	63.32
GPT-5.1	76.93	83.33	70.25	77.31	80.56	70.76	77.32	67.37	69.74	75.00	78.88	81.50	72.18	75.19
Gemini 3	79.78	85.66	73.40	80.36	82.95	74.92	81.81	66.10	68.42	75.00	81.10	84.55	75.38	80.31
Average	**73.94**	**78.54**	**66.14**	**75.56**	**77.01**	**70.24**	**73.69**	**63.28**	**69.30**	**75.00**	**75.74**	**78.66**	**69.74**	**70.64**

Table 5: Performance comparison on the **Chinese version** of CogSTEM, analyzing the impact of domain-specific fine-tuning. Models marked with **(S)** are STEM-tuned on the Chinese training set. Comparison deltas (Δ) indicate improvements relative to the base versions.

Model	Total	Knowledge Dimension			Cognitive Dimension						Disciplinary Dimension			
	Score	Fact.	Conc.	Proc.	Rem.	Und.	App.	Ana.	Eva.	Cre.	Sci.	Tech.	Eng.	Mat.
Qwen2.5-7B	61.86	63.64	55.07	66.93	64.79	61.13	60.96	45.34	60.53	66.67	62.13	68.70	59.68	55.98
Qwen3-8B	74.11	76.80	70.93	74.59	75.81	72.76	73.14	67.87	50.00	79.17	76.80	77.64	67.86	75.68
Qwen3-32B	78.72	82.82	73.40	80.02	81.15	75.91	79.04	69.07	71.05	83.33	81.87	81.71	72.18	79.54
Qwen3-235B	72.71	79.10	63.17	75.97	76.91	66.94	72.29	60.59	67.11	66.67	74.18	77.44	70.11	64.86
GPT-5.1	64.41	73.66	56.78	65.89	71.40	60.47	60.24	51.27	67.11	70.83	67.02	74.19	62.50	50.79
Gemini 3	80.90	86.36	74.68	87.74	84.28	73.92	82.77	69.49	76.32	75.00	82.77	83.94	76.69	80.31
Qwen2.5-7B-S	64.52 (+2.66)	65.98 (+2.34)	59.34 (+4.27)	68.30 (+1.37)	67.17 (+2.38)	64.62 (+3.49)	63.25 (+2.29)	49.15 (+3.81)	65.79 (+5.26)	62.50 (-4.17)	64.88 (+2.75)	71.95 (+3.25)	62.03 (+2.35)	58.30 (+2.32)
Qwen3-8B-S	74.60 (+0.49)	78.50 (+1.70)	71.18 (+0.25)	74.16 (-0.43)	76.91 (+1.10)	43.75 (-29.0)	73.13 (-0.01)	66.10 (-1.77)	71.05 (+21.1)	75.00 (-4.17)	77.28 (+0.48)	78.65 (+1.01)	68.42 (+0.56)	74.90 (-0.78)
Qwen3-32B-S	79.05 (+0.33)	83.59 (+0.77)	74.34 (+0.94)	79.29 (-0.73)	81.25 (+0.10)	77.08 (+1.17)	78.65 (-0.39)	69.49 (+0.42)	78.95 (+7.90)	79.17 (-4.16)	81.86 (-0.01)	81.30 (-0.41)	74.25 (+2.07)	76.36 (-3.18)
Average	**72.32**	**76.72**	**66.54**	**74.77**	**75.52**	**66.29**	**71.50**	**60.93**	**67.55**	**73.15**	**74.31**	**77.28**	**68.19**	**68.52**

ries demonstrates robust bilingual alignment, with Qwen3-32B showing slightly better performance in Chinese (78.72) than in English (77.84), indicating superior cross-lingual transferability in STEM contexts.

4.2.2 Multi-dimensional Weakness Analysis

Analyzing the average scores across dimensions reveals universal bottlenecks in current LLMs' STEM capabilities:

- **Cognitive Bottleneck (Analyze):** Across both languages, the *Analyze* dimension proves to be the most challenging cognitive task, with the lowest average scores (63.28 in English and 60.93 in Chinese). This indicates that while models excel at *Remembering* (Avg 76) and *Applying* (Avg 72), they struggle to decompose complex systems into constituent parts and understand their organizational structures.

- **Knowledge Gap (Conceptual vs. Factual):** In the Knowledge dimension, models score significantly lower on *Conceptual* knowledge (Avg $\approx$ 66) compared to *Factual* knowledge (Avg $\approx$ 77-78). This disparity suggests that current models rely more on rote memorization of facts rather than grasping the interrelationships between basic elements within a larger structure.

- **Disciplinary Variances:** Models generally perform better in *Technology* and *Science* but

face greater difficulties in *Engineering* and *Mathematics*. Specifically, Mathematics remains a hurdle, with average scores hovering around 70, reflecting the inherent difficulty of rigorous logical derivation compared to descriptive knowledge retrieval.

4.2.3 STEM Fine-tuning Analysis

To address the aforementioned weaknesses, we applied our domain-specific fine-tuning strategy. The results in the bottom section of Table 5 and Table 6 demonstrate the efficacy of our approach:

Significant Gains in Smaller Models The fine-tuning process yields the most substantial relative improvements in smaller models. **Qwen2.5-7B-S** achieves a total score increase of **+2.66** points (from 61.86 to 64.52), significantly narrowing the gap with larger base models. This implies that targeted STEM data can effectively activate the latent reasoning capabilities of compact models.

Enhancement in Higher-Order Thinking For larger models, STEM-tuning specifically optimizes higher-order cognitive processes. **Qwen3-8B-S** and **Qwen3-32B-S** show massive improvements in the *Evaluate* dimension, with gains of **+21.1** and **+7.90**, respectively. This indicates that our training strategy successfully encourages the models to make judgments based on criteria and standards, addressing a key cognitive deficiency. However, we note a slight trade-off in the *Create* and *Mathematics* dimensions for the 32B model, suggesting a potential shift in the model's focus towards critical analysis over generative tasks.

Robustness on General Benchmarks Crucially, Table 6 confirms that this domain specialization does not come at the cost of catastrophic forgetting. **Qwen3-8B-STEM** not only maintains its general capability but achieves slightly higher scores on MMLU-Pro (+0.64) and MATH (+0.68) compared to its base version. This validates that CogSTEM serves as a high-quality alignment target that enhances domain expertise while preserving, and potentially reinforcing, the model's general reasoning fundamental.

Model	C-Eval	MMLU-Pro	MATH	Avg.
Qwen2.5-7B	78.36	55.77	72.66	68.91
Qwen3-8B	86.34	73.11	93.10	84.18
Qwen2.5-7B-STEM	78.04	55.24	72.74	68.67
Qwen3-8B-STEM	**86.34**	**73.75**	**93.78**	**84.60**

Table 6: Comparison of general capabilities before and after STEM reinforcement training. Results show that targeted STEM tuning maintains or slightly enhances general reasoning performance.

5 Conclusion

In this work, we presented CogSTEM, a novel and theoretically grounded benchmark designed to mitigate the severe *distributional biases* prevalent in existing LLM evaluations. By strictly integrating the Revised Bloom's Taxonomy, CogSTEM provides a fine-grained diagnostic framework that effectively decouples foundational memory from high-order scientific reasoning across disciplinary, knowledge, and cognitive dimensions. Our rigorous human-in-the-loop pipeline ensures a high-quality dataset of 4,491 problems that achieves a multi-dimensional equilibrium in assessing STEM capabilities.

Our experiments on state-of-the-art models revealed critical insights into AI cognition: (1) a persistent "Remembering-Analyzing Gap", where even top-tier models excel in factual recall but falter significantly in decomposing complex problems; (2) the efficacy of targeted fine-tuning, as demonstrated by the Qwen series achieving substantial gains—particularly a 7.90% surge in the *Evaluate* dimension—without compromising general capabilities. These findings suggest that current model bottlenecks lie less in knowledge accumulation and more in the depth of cognitive processing. Ultimately, we hope CogSTEM will serve as a catalyst for developing more reliable AI systems capable of authentic scientific reasoning.

6 Limitations

We acknowledge that the current version of CogSTEM primarily focuses on text-based symbolic reasoning within an objective framework. While this design ensures scalable and reliable evaluation, it limits the assessment of multimodal scientific problem-solving (e.g., diagram interpretation), which remains an essential aspect of real-world STEM practice. Future work will extend this framework to multimodal contexts to achieve a more holistic evaluation of intelligent teaching assistants.

References

Lorin W Anderson and David R Krathwohl. 2001. *A taxonomy for learning, teaching, and assessing: A*

revision of Bloom's taxonomy of educational objectives: complete edition. Addison Wesley Longman, Inc.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, and 1 others. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36:62991–63010.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2024. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024a. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Qwen Team and 1 others. 2024b. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2(3).

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and 1 others. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290.

Norman L Webb. 1997. Criteria for alignment of expectations and assessments in mathematics and science education. research monograph no. 6.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Johannes Welbl and 1 others. 2017. Crowdsourcing a magdeburg-sized dataset for science questions. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 34–40.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2024. Agieval: A human-centric benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2299–2314.

A Detailed Dataset Statistics and Comparisons

This appendix provides a granular comparison of CogSTEM with mainstream benchmarks. Labels and dataset names are abbreviated for readability.

As shown in Table 7, CogSTEM distinguishes itself through two key features:

- **Cognitive Diversity:** Unlike SciQ (skewed $>$ 90% towards *Remember*), CogSTEM offers a balanced distribution with significant *Apply* and *Analyze* components to better evaluate reasoning.
- **Knowledge Equilibrium:** CogSTEM maintains a near 1:1:1 ratio across *Conceptual*, *Factual*, and *Procedural* types, providing a more comprehensive assessment than single-mode benchmarks like GSM.

A.1: Cognitive Dimension (Revised Bloom's Taxonomy)

Lvl.	CogS.	C-Ev.	GPQA	GSM	MMLU	Math	SciQ
Rem.	**2,214**	5,980	102	0	2,316	32	12,608
Und.	**774**	3,730	52	0	1,837	14	1,049
App.	**1,070**	3,061	670	8,667	5,905	6,878	19
Ana.	**307**	1,054	404	124	1,613	3,193	1
Eva.	**97**	110	6	0	322	0	1
Cre.	**29**	11	18	0	37	2,378	0
Total	**4,491**	**13,946**	**1,252**	**8,791**	**12,030**	**12,495**	**13,678**

A.2: Knowledge Dimension

Type	CogS.	C-Ev.	GPQA	GSM	MMLU	Math	SciQ
Conc.	**1,510**	6,971	365	0	5,013	1,358	3,094
Fact.	**1,490**	3,615	11	0	1,488	19	10,510
Proc.	**1,491**	3,362	876	8,792	5,245	0	74

A.3: Disciplinary Dimension

Disc.	CogS.	C-Ev.	GPQA	GSM	MMLU	Math	SciQ
Sci.	**2,155**	7,054	1,221	11	6,854	15	13,503
Math	**330**	3,434	12	6	3,679	12,484	40
Tech.	**635**	1,813	14	4	781	0	79
Eng.	**1,371**	1,443	5	8,771	500	0	57

Note: Abbreviations align with Bloom's Taxonomy (Rem: Remember, Und: Understand, App: Apply, Ana: Analyze, Eva: Evaluate, Cre: Create) and Knowledge types (Conc: Conceptual, Fact: Factual, Proc: Procedural).

Table 7: Detailed statistics of CogSTEM dataset across dimensions.

B Data Annotation and Cost Details

B.1 Annotation Team Composition

The construction of the CogSTEM dataset followed a rigorous human-in-the-loop (HITL) paradigm. We assembled a team of 10 subject-matter experts, including university and high school educators specializing in Science, Technology, Engineering, and Mathematics (STEM). These experts were responsible for verifying the scientific accuracy of the problems and performing fine-grained cognitive labeling.

B.2 Annotation Framework and Quality Control

To ensure high annotation quality and consistency, we implemented the following procedures:

- **3D Annotation Framework**: Annotators categorized each sample across three dimensions: Disciplinary (Subject), Knowledge type (Factual, Conceptual, Procedural), and Cognitive level (aligned with the Revised Bloom's Taxonomy).
- **Multi-round Double-blind Annotation**: Each item underwent multiple rounds of double-blind labeling. In cases of disagreement, a senior expert performed consistency arbitration to eliminate subjective bias.
- **Human-AI Fusion Cleaning**: We utilized a "dual-model, dual-human" funnel strategy. This involved adversarial screening by SOTA models followed by expert disambiguation and text reconstruction to ensure logical rigor.

B.3 Workload and Budgetary Expenditure

The final version of the CogSTEM dataset comprises 4,491 high-quality bilingual samples.

- **Unit Compensation**: To incentivize experts for the high-intensity task of deep cognitive analysis (especially for higher-order levels like "Analyze" and "Evaluate"), the compensation was set at 20 RMB per item.