BloomEval: A Bloom's Cognitive Taxonomy-Based Benchmark for Evaluating LRMs via Cognitive Hierarchy Trace

Zhiyi Duan[1], Lei Gao[1], Jiangshan Guan[1], Qi Wang[2], Rui Liu[1]*,
[1]Inner Mongolia University, Hohhot, China,
[2]Jilin University, Changchun, China,
duanzy@imu.edu.cn, 32409159@mail.imu.edu.cn, 32409140@mail.imu.edu.cn, qiwang@jlu.edu.cn, imucslr@imu.edu.cn

Abstract

Current benchmarks for Large Reasoning Models (LRMs) primarily rely on answer correctness, failing to assess the structural coherence and cognitive soundness of the reasoning process itself. To address this gap, we introduce Cognitive Hierarchy Trace (CHT), a novel evaluation framework grounded in Bloom's Cognitive Taxonomy (BCT). CHT provides a structured, step-wise mapping of a model's reasoning trajectory onto hierarchical cognitive levels, enabling the detection of structural anomalies such as hierarchy jumps, breaks, and overthinking. Based on CHT, we present BloomEval, the first large-scale benchmark designed for fine-grained cognitive capability assessment. It comprises 94,602 math problems, each annotated with Bloom's cognitive levels, CHT trajectories, a three-tier knowledge hierarchy, and problem difficulty. To ensure scalable yet reliable annotation, we develop an Expert-LLM collaborative pipeline with a three-stage reconciliation mechanism. Our comprehensive evaluation reveals a critical finding: models often arrive at correct answers through cognitively flawed or opaque reasoning paths. The CHT-based analysis uncovers prevalent structural inconsistencies that are invisible to outcome-only metrics, demonstrating that answer accuracy is an insufficient proxy for reasoning quality.

1 Introduction

Human cognition operates through two distinct systems of thought: the fast, intuitive, and automatic processes of System 1, and the slow, deliberate, and goal-directed processes of System 2 thinking (Li et al., 2024; Lombardi, 2023; Li et al., 2025). Recent advances in Chain-of-Thought (CoT) have empowered Large Reasoning Models (LRMs) like DeepSeek-R1 and OpenAI-o1, to achieve remarkable breakthroughs in tasks, such as mathematical reasoning (Huang et al., 2025; Xia et al., 2025; Gao et al., 2024), code generation (Dong et al., 2025; Nejjar et al., 2025; Ouyang et al., 2025), and commonsense reasoning (Lamsiyah et al., 2025).

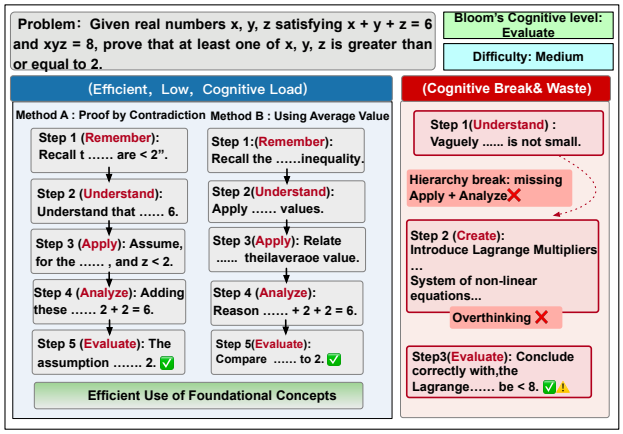

Figure 1: An example of cognitive bias under Bloom's Cognitive Taxonomy.

However, evaluation practices have not kept pace (Zhang et al., 2025). Mainstream benchmarks (e.g., GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021), HumanEval (Li and Murr, 2024)) remain largely answer-based, measuring outcome accuracy but offering little insight into the organization of the reasoning process. Studies show that LRMs can produce superficially plausible reasoning traces through pattern completion, even when the underlying cognitive structure is fragmented or poorly grounded. Consequently, final answer correctness alone is inadequate for assessing the coherence and quality of reasoning.

Recent efforts have begun to integrate cognitive frameworks into evaluation, yet most work assigns cognitive labels at the task level rather than examining step-wise reasoning trajectories. Crucially, effective human reasoning follows an ascending hierarchy of cognitive operations, progressing from lower-order processes to higher-order ones, regardless of the specific solution strategy. As shown in

*Corresponding author

Findings of the Association for Computational Linguistics: ACL 2026, pages 25228–25248

Figure 1, different reasoning paths for the same problem may yield the same answer, but only those with a coherent cognitive progression ensure interpretability and efficiency. However, there remains a lack of methods that systematically assess the cognitive capabilities demonstrated during an LRM's reasoning process.

To fill this gap, we introduce **Cognitive Hierarchy Trace** (CHT) grounded in Bloom's Cognitive Taxonomy (BCT), a structured representation that maps each inference step in a model's reasoning trajectory to a specific cognitive level. BCT defines six hierarchical levels of cognition and provides a theoretical basis for constructing rigorous, structured, and quantifiable evaluation dimensions. Thus, CHT allows for the systematic detection of cognitive-level errors within reasoning trajectories, including hierarchical jumps and breaks. Based on CHT, we present **BloomEval**, the first large-scale, fine-grained benchmark that systematically assesses the cognitive capability of LRMs through the lens of cognitive hierarchy. BloomEval is composed of 94,602 math problems spanning K-12 to higher education, each annotated with Bloom's cognitive levels, CHT, knowledge point hierarchies, and problem difficulty. Our contributions can be summarized as follows:

- We propose Cognitive Hierarchy Trace, a structured representation that maps each step of a model's reasoning to a level in Bloom's Taxonomy, enabling a fine-grained, hierarchical analysis of cognitive progression.
- We introduce BloomEval, a large-scale benchmark constructed via an expert-LLM collaborative framework. Our annotation pipeline employs a multi-stage workflow with a dedicated reconciliation mechanism to ensure high consistency and reliability.
- We define a set of diagnostic metricsincluding Hierarchy jump, Hierarchy break, and Overthinking, to quantify cognitive structural flaws in reasoning. Our analysis reveals that correct answers frequently conceal opaque and incoherent reasoning processes.

2 Related Work

2.1 Benchmarks for Evaluating LLMs

Early benchmark evaluations of LLMs primarily used the accuracy of the final answer as the core indicator of ability (Hendrycks et al., 2021; Cobbe et al., 2021). However, as model size increased and capabilities emerged, this outcome-oriented assumption gradually revealed its limitations. Existing studies show that models often rely on large-scale pattern matching to achieve high scores in tests, rather than true semantic understanding (Zhao et al., 2025; Wu et al., 2025; Liu et al., 2024). To address this gap, recent evaluation paradigms have gradually shifted from outcome-oriented to process-oriented, focusing on examining the reasoning paths and intermediate steps in generating answers (Xia et al., 2025; Song et al., 2025; Zeng et al., 2024). In particular, with the emergence of LRMs based on System 2 thinking, assessments have begun to encompass human-like cognitive dimensions such as critical thinking and self-reflection (Zhou et al., 2024; Guan et al., 2025; Hao et al., 2025; Figueras and Agerri, 2025). However, existing methods mostly characterize cognition using surface-level behavioral indicators, making it difficult to systematically distinguish between deep reasoning with coherent cognitive structures and superficial reasoning behaviors that are merely superficially reasonable.

2.2 Application of Bloom's Cognitive Taxonomy

BCT, originating from educational psychology, constructs a hierarchical system of human cognitive activities from low to high order, encompassing six dimensions: Remember, Understand, Apply, Analyze, Evaluate, and Create, emphasizing the systematic progression of cognitive abilities (Krathwohl, 2002). In recent years, Research has provided a structured perspective beyond single accuracy rates for understanding model capabilities by mapping standard benchmarks to BCT (Dou et al., 2025; Zoumpoulidi et al.; Phung et al., 2023; Fu et al., 2022). For example, (Huber and Niklaus, 2025) used this taxonomy to perform fine-grained annotation of programming tasks, empirically distinguishing the robust performance of current models on low-order memory tasks from their significant shortcomings on high-order evaluation and creative tasks. However, existing research often limits this framework to static task-level labels, focusing on identifying the cognitive needs of the task while neglecting the dynamic unfolding of cognition during the reasoning process.

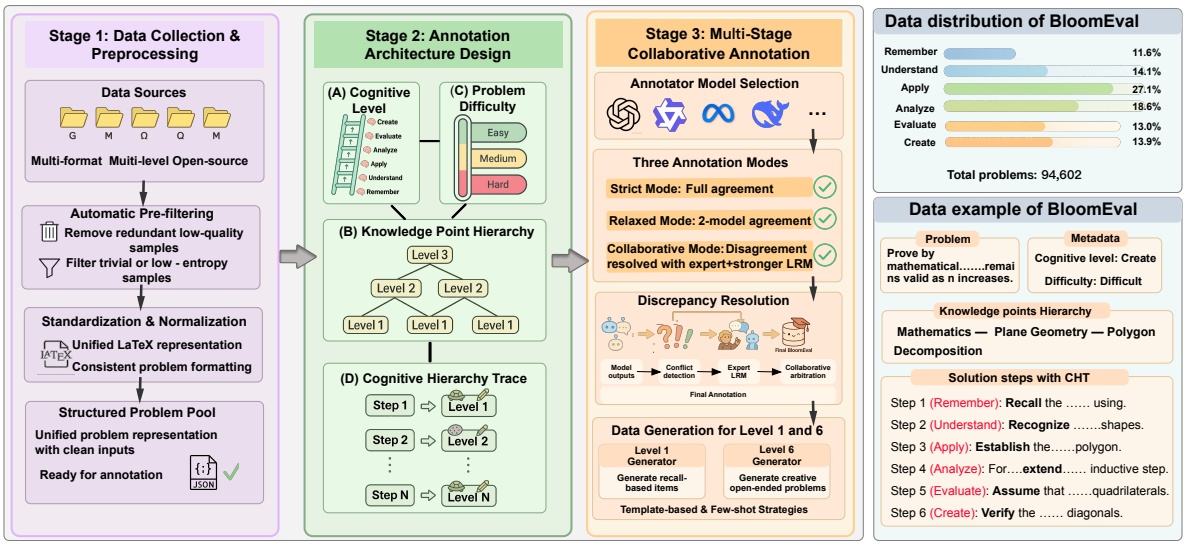

Figure 2: Construction workflow of the BloomEval Benchmark.

3 BloomEval Benchmark

As shown in Figure 2, our proposed BloomEval comprises three components: Data Collection and Preprocessing, Annotation Architecture Design, Multi-stage Collaborative Annotation.

3.1 Data Collection and Preprocessing

To ensure broad sample coverage and diversity of problem types, we curate a collection from mainstream open-source mathematical datasets, including Omni-MATH (Gao et al., 2024), GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021), Orca-math (Mitra et al., 2024), and Mooc_Radar (Yu et al., 2023). These datasets span mathematics problems from K-12 to higher education and comprise three primary question formats: multiple-choice, fill-in-the-blank, and short-answer. All raw problems are standardized into a unified LaTeX format. During preprocessing, we address two main challenges: inconsistent structural representations across sources and the normalization of mathematical notations (e.g., $\sqrt{\cdot}$, $\sum$, π). This process ensures both structural consistency and semantic clarity, forming a reliable foundation for subsequent annotation. Further details are provided in Appendix A.

3.2 Annotation Architecture Design

To enable interpretable and fine-grained evaluation of cognitive processes and reasoning quality, we design a multi-dimensional annotation framework based on BCT. From the perspective of problem comprehension, it comprises four dimensions: Bloom's cognitive level, CHT, Knowledge point hierarchy, and Problem difficulty.

Bloom's Cognitive Level

We adopt BCT to establish a hierarchical cognitive scale, ranging from Remember (Level 1) to Create (Level 6). This quantification allows a precise assessment of the gap between the cognitive depth demonstrated in LRM's reasoning and the level required by the ground-truth solution.

Cognitive Hierarchy Trace

We model the reasoning process as a structured cognitive trajectory, defined by an ordered sequence $\mathcal{C} = \{(s_i, l_i)\}_{i=1}^{N}$, where s_i is an atomic reasoning step and l_i denotes its corresponding cognitive level. By decoding the cognitive hierarchy embedded in the reasoning flow, we map each step to a specific cognitive tier. To operationalize this mapping, Drawing from established taxonomies in educational sciences (Stanny, 2016), we employ a curated set of *Cognitive Process Verbs* (see Figure 3) as explicit indicators that project unstructured natural language onto the structured cognitive space. These verbs serve as semantic anchors, quantifying the cognitive depth of each step according to BCT. To resolve ambiguities arising from verb polysemy and implicit reasoning in natural language, we further introduce three heuristic disambiguation rules:

- **Maximum Pooling Principle**: a step involving multiple cognitive operations is assigned the highest cognitive level present among

them.

- **Taxonomy-Aligned Operation Mapping**: the classification of implicit mathematical actions is based on cognitive intent rather than surface linguistic form.
- **Structural Contribution Criterion**: only operations that actively alter the problem state or introduce inferential dependencies receive higher-level assignments, whereas mere restatements or paraphrases default to the lowest applicable level.

Detailed annotation guidelines are provided in Appendix B.

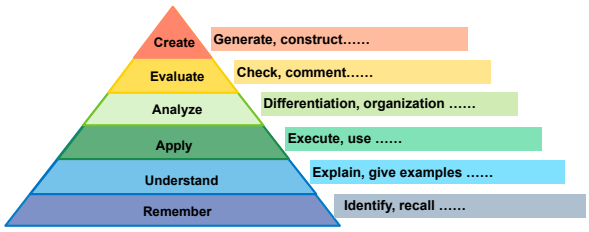

Figure 3: An example of cognitive process verbs under BCT.

Knowledge Point Hierarchy

Knowledge points are organized into three hierarchical levels: Level 1 covers foundational concepts, Level 2 intermediate constructs, and Level 3 fine-grained, domain-specific knowledge. This structure supports the evaluation of both breadth of knowledge coverage and the correct invocation of necessary knowledge points during reasoning.

Problem Difficulty

To enable a more objective assessment of problem difficulty within the benchmark, we adopt a three-level categorization, including easy, medium, and Hard, aligned with existing difficulty-rating schemes.

3.3 Multi-stage Collaborative Annotation

Given the high cost and complexity of large-scale manual annotation, we propose a multi-stage collaborative workflow that combines subject expertise with LRM reasoning to ensure consistency and scalability.

3.3.1 Collaborative Annotation

To enhance the stability and reliability of automatic annotation, we first construct a gold-standard test set $\mathcal{D}_{gold}$, comprising $|\mathcal{D}_{gold}| = 1{,}000$ samples. The set follows a stratified and balanced distribution across problem difficulty and cognitive levels to ensure unbiased evaluation. Using a unified few-shot prompting strategy with carefully designed examples and challenging stress-test samples, we systematically assess annotation robustness across candidate models. Under this controlled setup, we compare standard LLMs and LRMs on the gold-standard test set. Using annotation accuracy against expert labels and per-token inference cost as evaluation metrics, two main findings emerge:

- Despite their distinct architectures and reasoning mechanisms, the accuracy gap between the two model types remains within 1.2% on tasks dominated by lower-order cognitive operations, indicating that LLMs are sufficient for annotating such tasks.
- Due to the generation of longer reasoning chains, LRMs incur substantially higher computational costs, with their per-token inference overhead exceeding that of LLMs by more than 20 tokens.

Based on these findings, we select the three top-performing standard LLMs (i.e., GPT-3.5-turbo, LLaMA-3-70B, and Qwen-2.5-72B) on the gold test set as our primary automatic annotators. This selection significantly improves annotation efficiency without compromising quality. We also evaluate the agreement between the three selected LLMs and human expert annotations. The results indicate high consistency, with an average Cohen's κ of 0.81 ($p < 0.001$), demonstrating that automatic annotations align closely with expert judgment. Next, we employ few-shot prompting with carefully designed exemplars and adversarial stress tests, dynamically adjusting temperature, example order, and prompt framing to account for model variations. For samples requiring higher-order cognitive reasoning, we introduce a supplementary verification step in which stronger reasoning-oriented models generate and validate additional reasoning chains, further ensuring annotation reliability.

3.3.2 Discrepancy Reconciliation Mechanism

To ensure annotation reliability while maintaining broad coverage, we adopt a three-stage reconciliation mechanism:

- **Strict-mode:** Samples with unanimous agreement among all annotator models are accepted

directly as high-confidence annotations.

- **Relaxed-mode:** Samples with majority agreement are resolved via a consensus rule, retaining them as medium-confidence instances to balance coverage and reliability.

- **Collaborative-mode:** Samples with complete disagreement undergo a two-step refinement: a more capable LRM first re-evaluates the problem with specialized prompts, after which a human expert adjudicates the final label.

Quality Verification of Annotations To validate the reliability of the reconciliation mechanism, we perform expert verification on samples from each agreement regime. For high-confidence samples in Strict-mode, random auditing yields a correction rate below 1.5%, confirming the reliability of acceptance. For medium-confidence samples resolved by majority vote, stratified review shows a correction rate of 7.2%, indicating acceptable noise for large-scale annotation. In contrast, low-confidence samples with complete disagreement require substantial human intervention, with experts revising 33.4% of labels after re-annotation, underscoring the need for the collaborative resolution step. This process yields 74,602 high-quality annotated samples with multi-dimensional cognitive labels (see Figure 2). Detailed collaborative annotation are provided in Appendix C.

3.3.3 Data Generation

To address the scarcity of Level 1 (Remember) and Level 6 (Create) samples, each constituting less than 10% of the original dataset, we implement a targeted synthetic-data generation pipeline to achieve a balanced distribution across cognitive levels. This augmentation addresses potential evaluation bias in both low-level recall and high-level creative reasoning tasks. We employ OpenAI-o1 for generation, adopting distinct protocols aligned with BCT:

- **Level 1 generation:** Single-step recall tasks are generated by abstracting templates from existing problems and using few-shot prompting to produce varied questions that test basic facts, definitions, and formulas.

- **Level 6 generation:** Open-ended templates requiring multi-step reasoning and reverse engineering are designed to elicit solutions that traverse multiple cognitive levels and explore divergent paths.

The resulting augmentation yields a balanced distribution across all cognitive levels, with strong reliability confirmed by expert verification (Cohen's $\kappa = 0.87$). Full task definitions, templates, and prompting details are provided in Appendix C. In total, we generate 20,000 new problems across Level 1 and 6, expanding the final dataset to 94,602 problems with improved distributional balance.

4 Experiment

4.1 Experimental Setup

Models. We select LRMs and LLMs that have demonstrated strong performance in reasoning benchmarks over the past two years. See Appendix D for the list of Model Selection.

Settings. To ensure a controlled and reproducible experimental setup, we construct a balanced test set of 1,000 problems through stratified random sampling, ensuring even coverage across all six cognitive levels defined by BCT. All LLMs and LRMs are accessed via their official APIs. We strictly adhere to the prescribed chat templates and default system prompts for each model, and employ the standard sampling configurations (e.g., temperature parameters) as provided in the official implementations.

Tasks. We design three complementary evaluation tasks to systematically assess the cognitive capabilities of LRMs: Cognitive Level Perception, Hierarchical Knowledge Alignment, Cognitive Coherence in Reasoning Trajectories.

4.2 Evaluation of Cognitive Level Perception

This task assesses whether LRMs can accurately identify and distinguish the specific Bloom's cognitive levels required by a given problem. To systematically evaluate the impact of cognitive hierarchical reasoning on cognitive level prediction, we assess models under two distinct settings: **(1) Zero-shot:** where the model predicts Bloom's cognitive level relying solely on the problem statement; and **(2) Few-shot:** the model can refer to the structured reasoning steps provided by CHT as context when making predictions, but these steps do not contain any explicit cognitive hierarchy labels.

Table 1 shows the average accuracy of each model across the six Bloom's cognitive levels. The

Type	Model	Remember	Understand	Apply	Analyze	Evaluate	Create	Average
				Zero-shot Accuracy				
	Grok-3	**0.575**	**0.522**	**0.483**	**0.448**	**0.422**	**0.403**	**0.476**
	DeepSeek-R1	0.573	0.497	0.446	0.423	0.398	0.384	0.453
	Claude 3.7	0.552	0.484	0.426	0.417	0.391	0.364	0.439
	OpenAI o1-mini	0.452	0.398	0.381	0.363	0.344	0.324	0.377
LRMs	s1-32B	0.429	0.358	0.318	0.232	0.155	0.284	0.296
	Kimi k1.5	0.396	0.340	0.366	0.339	0.141	0.300	0.314
	Gemini-2.5-Pro	0.563	0.478	0.432	0.384	0.353	0.316	0.421
	DeepSeek v3	0.518	0.450	0.421	0.405	0.385	0.347	0.421
	OpenAI o3	0.424	0.347	0.321	0.298	0.283	0.245	0.320
	GPT-4o	0.548	0.477	0.421	0.384	0.353	0.320	0.417
	Gemini-1.5	0.522	0.459	0.424	0.397	0.379	0.354	0.423
LLMs	Claude 3.5	0.392	0.351	0.437	0.354	0.297	0.390	0.370
	Qwen3-Max	0.504	0.451	0.424	0.397	0.375	0.356	0.418
	OpenAI o4-mini	0.474	0.427	0.398	0.378	0.362	0.332	0.395
				Few-shot Accuracy				
	Grok-3	**0.698**	**0.652**	**0.601**	**0.553**	**0.503**	0.472	**0.580**
	DeepSeek-R1	0.666	0.648	0.498	**0.553**	0.479	**0.476**	0.553
	Claude 3.7	0.631	0.619	0.577	0.527	0.469	0.451	0.546
	OpenAI o1-mini	0.599	0.548	0.497	0.448	0.428	0.398	0.486
LRMs	s1-32B	0.525	0.538	0.521	0.453	0.419	0.376	0.472
	Kimi k1.5	0.554	0.547	0.447	0.403	0.378	0.347	0.446
	Gemini-2.5-Pro	0.697	0.648	0.596	0.518	0.448	0.449	0.559
	DeepSeek v3	0.649	0.599	0.549	0.497	0.467	0.446	0.535
	OpenAI o3	0.547	0.496	0.452	0.397	0.353	0.296	0.424
	GPT-4o	0.668	0.548	0.596	0.517	0.437	0.448	0.536
	Gemini-1.5	0.647	0.596	0.548	0.498	0.428	0.439	0.526
LLMs	Claude 3.5	0.662	0.616	0.572	0.475	0.428	0.399	0.525
	Qwen3-Max	0.598	0.552	0.501	0.449	0.417	0.396	0.486
	OpenAI o4-mini	0.618	0.578	0.528	0.477	0.461	0.432	0.516

Table 1: Bloom's cognitive level prediction accuracy. Bold values denote the best performance; underlined values indicate the second-best.

results reveal a clear performance gap that all models consistently outperform on lower-order cognitive tasks compared to higher-order ones. For instance, DeepSeek-R1 achieves an accuracy of 0.573 at the Remember level but drops to 0.384 at the Create level. After introducing few-shot learning, all models improve across all levels, with the most substantial gains observed at higher-order cognitive tiers. This indicates that few-shot prompting with detailed solution steps not only reinforces the structural coherence of the reasoning process but also extends the effective cognitive trajectory, thereby enhancing model capability on complex tasks.

4.3 Evaluation of Hierarchical Knowledge Alignment

This task assesses the model's ability to dynamically invoke knowledge units that match the required cognitive level during reasoning. We measure this alignment via knowledge-point matching accuracy, which quantifies the semantic correspondence between the generated reasoning trajectory and the reference knowledge hierarchy using a predefined synonym dictionary and text similarity metrics.

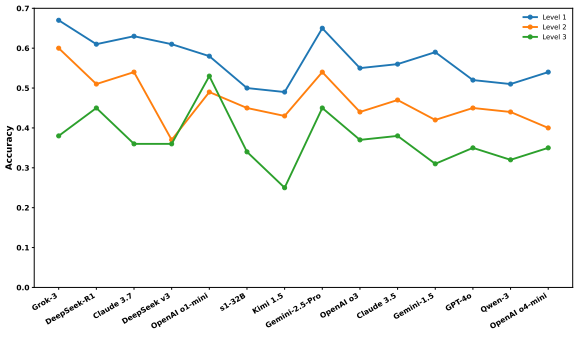

Figure 4: Knowledge point matching accuracy.

As shown in Figure 4, models achieve reliable accuracy on Level 1 knowledge points, ranging from 0.530 to 0.675. Performance declines at Level 2, with accuracy falling between 0.425 and 0.575, attributable to increased structural complexity. The most pronounced drop occurs at Level 3, where accuracy decreases to 0.300 for s1-32B and 0.265 for Kimi k1.5. This downward trend, however, does not necessarily reflect cognitive failure. Given the fine-grained nature of Level 3 concepts, models often internalize such detailed knowledge within

Type	Model	Answer-Correct			Answer-Incorrect		
		Hierarchy Break	Hierarchy Jump	Overthinking	Hierarchy Break	Hierarchy Jump	Overthinking
LRMs	Grok-3	0.088	0.185	0.091	0.337	0.142	**0.219**
	DeepSeek-R1	0.088	0.162	0.072	**0.252**	0.125	0.228
	Claude 3.7	0.077	0.145	**0.061**	0.368	0.138	0.289
	DeepSeek v3	**0.067**	0.112	0.109	0.358	0.110	0.281
	OpenAI o1-mini	0.083	0.245	0.092	0.262	0.095	0.268
	s1-32B	0.079	**0.052**	0.121	0.436	0.098	0.289
	Kimi k1.5	0.158	0.068	0.078	0.417	0.105	0.352
	Gemini-2.5-Pro	0.078	0.135	0.067	0.362	0.122	0.313
	OpenAI o3	0.093	0.215	0.069	0.382	0.118	0.231
LLMs	GPT-4o	0.095	0.095	0.068	0.345	0.085	0.282
	Gemini-1.5	0.115	0.088	0.063	0.360	**0.072**	0.307
	Claude 3.5	0.085	0.092	0.074	0.355	0.089	0.286
	Qwen3-Max	0.095	0.105	0.115	0.340	0.092	0.315
	OpenAI o4-mini	0.075	0.082	0.073	0.340	0.075	0.247

Table 2: Structure analysis of CHT, comparing the rates of hierarchy break, hierarchy jump, and overthinking across correct and incorrect answer conditions.

the problem-solving flow rather than expressing it explicitly.

4.4 Evaluation of Cognitive Coherence in Reasoning Trajectories

Moving beyond final-answer accuracy, we investigate whether the model-generated CHT maintains an explicit and traceable cognitive trajectory. We acknowledge that, in certain contexts, applying higher-order methods to relatively elementary problems may indeed reflect greater mathematical maturity or enhanced generalization capability. However, within standardized mathematical and logical problem-solving frameworks, progressive hierarchical structures remain the normative reasoning pathway. Following Madaus et al.(Madaus et al., 1973), who demonstrated through causal modeling that mastery of lower-order cognitive objectives constitutes a necessary prerequisite for higher-order achievement, and Liu et al.(Liu et al., 2025), who emphasize aligning LLM reasoning with coherent human cognitive patterns, CHT evaluates not the ingenuity of individual cognitive strategies but rather the structural consistency of the reasoning process, whether it adheres to a logically ordered progression. Thus, we define three specific metrics to quantify the coherence and validity of the cognitive process:

- **Hierarchy break** describes a reasoning trace that does not reach the cognitive level required by the task. For instance, if a task is labeled as Level 6 (Create) but the model's reasoning only exhibits operations up to Level 3 (Apply), the trace is recorded as a hierarchy break.

- **Hierarchy jump** occurs when a reasoning trace omits essential intermediate cognitive operations. For example, a model transitioning directly from Level 2 (Understand) to Level 6 (Create), while bypassing the required stages of Apply, Analyze, and Evaluate, is recorded as exhibiting a hierarchy jump.

- **Overthinking** occurs when a reasoning trace employs knowledge or cognitive operations that exceed the task's required cognitive level. For example, introducing advanced concepts such as series expansions to solve a problem that only necessitates basic inequalities is categorized as overthinking.

As shown in Table 2, cognitive-structural anomalies occur more frequently in samples with incorrect answers. Hierarchy break is the most pronounced anomaly in such cases, with an incidence of 0.417 for Kimi k1.5, indicating a strong correlation between structural incompleteness and erroneous outcomes. Notably, even when answers are correct, models frequently exhibit structural deviations; for instance, Grok-3 shows a Hierarchy jump rate of 0.185. This pattern demonstrates that answer correctness does not guarantee a normatively structured reasoning process. While the final result may be accurate, the transparency and evaluability of the underlying reasoning are often compromised. Furthermore, Overthinking is more prevalent in incorrect samples, as illustrated by a rate of 0.352 for Kimi k1.5, suggesting that models tend to invoke irrelevant or advanced knowledge under cognitive uncertainty, thereby further obscuring the reasoning structure.

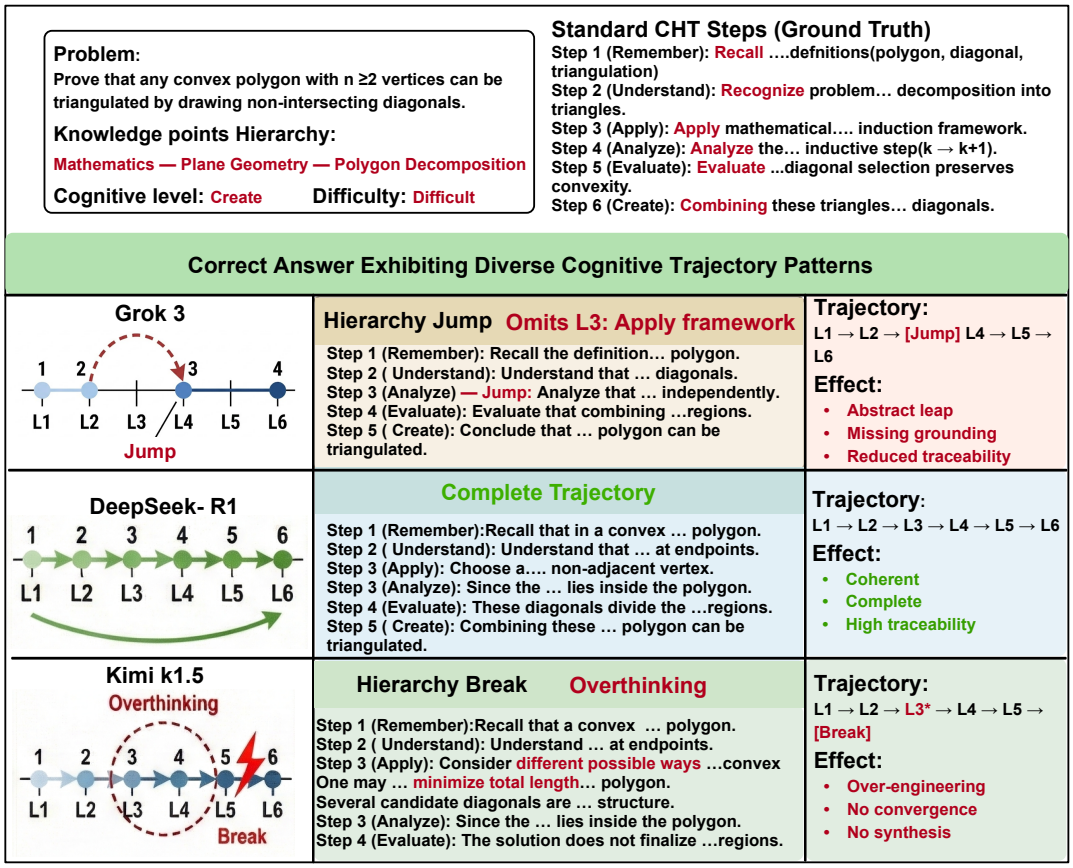

Figure 5: Comparative analysis of cognitive trajectories across LRMs based on the CHT.

4.5 Case Study

Figure 5 visualizes the cognitive trajectories produced by different models on a representative geometric task. DeepSeek-R1 demonstrates a complete and sequential cognitive hierarchy, progressing coherently from lower-order operations through procedural grounding to higher-order synthesis. In contrast, Grok-3 arrives at a correct answer but lacks structural integrity, exhibiting a Hierarchy jump by omitting the Apply stage and transitioning directly from Understand to Analyze, which reduces the transparency of its reasoning. A more pronounced deviation is observed in Kimi k1.5, after beginning with a correct initial trajectory, it diverges into overthinking by introducing extraneous optimization objectives such as minimizing diagonal length. This results in a Hierarchy break, where the reasoning fails to converge into a coherent synthesis, thereby compromising traceability despite ultimately yielding a correct answer.

4.6 Ablation Study

To validate the contribution of each component in the CHT framework, we perform an ablation study using Grok 3. Since low-order cognitive tasks often fail to reveal subtle differences in model performance, we construct an evaluation subset comprising 400 high-order cognitive problems. The reasoning paths generated on this subset are further assessed via expert manual evaluation to ensure label consistency. Detailed ablation studies are provided in Appendix E.

As shown in Table 3, the ablation study reveals three key findings. First, removing hierarchical constraints substantially degrades the integrity of the model's cognitive structure, with Hierarchy break rate rising from 0.284 to 0.425 and Hierarchy jump rate from 0.156 to 0.323. This confirms that explicit hierarchical scaffolding is essential for organizing reasoning into steps that align with human cognitive progression. Second, disabling cognitive process verbs causes the model's alignment with human reasoning to drop sharply; Cohen's κ falls from 0.810 to 0.422. This indicates that such verbs provide necessary granularity and coherence for distinguishing between cognitive operations. Finally, the full model achieves the lowest overthinking rate (0.112). Removing either hierarchical con-

Model / Method	Hierarchy Break ↓	Hierarchy Jump ↓	Overthinking ↓	Consistency (κ) ↑	Avg. Acc.↑
Grok 3 + CHT	**0.284**	**0.156**	**0.112**	**0.810**	**0.712**
- w/o Hierarchy Constraints	0.425	0.323	0.215	0.653	0.685
- w/o Cognitive Process Verbs	0.455	0.353	0.248	0.422	0.614
Standard CoT	0.490	0.453	0.285	0.380	0.552

Table 3: Ablation study for the key components of CHT. **Consistency** (κ) denotes Cohen's Kappa coefficient, which quantifies the structural agreement between the generated cognitive trajectory and the ground truth hierarchy. **Avg. Acc.** denotes the average accuracy across the high-order cognitive problems.

straints or process verbs increases this rate to 0.215 and 0.248, respectively. These results demonstrate that both components act as effective regularizers, curbing the injection of irrelevant knowledge and constraining the reasoning path to what is functionally required for the solution.

5 Conclusion

In this work, we introduce CHT to assess the cognitive validity of LRMs based on Bloom's Cognitive Taxonomy. We construct **BloomEval**, a large-scale benchmark with fine-grained cognitive-level annotations developed through an expert-LLM collaborative pipeline. Comprehensive experiments reveal a critical disconnect that models often produce correct answers via cognitively misaligned or structurally unsound reasoning traces, with frequent failures such as hierarchical jumps and breaks, particularly in complex tasks. These findings underscore the insufficiency of outcome-based accuracy and advocate for reasoning evaluation that accounts for underlying cognitive coherence.

Acknowledgments

This research of Zhiyi Duan was funded by the National Natural Science Foundation of China (No. 62567005), and Natural Science Foundation of Inner Mongolia Autonomous Region of China (No. 2025MS06004). This research of Rui Liu was funded by the General Program (No.62476146) of the National Natural Science Foundation of China, the Young Elite Scientists Sponsorship Program by CAST (2024QNRC001), the Outstanding Youth Project of Inner Mongolia Natural Science Foundation (2025JQ011), the Key R&D and Achievement Transformation Program of Inner Mongolia Autonomous Region (2025YFHH0014), the Central Government Fund for Promoting Local Scientific and Technological Development (2025ZY0143).

Limitations

Although the mathematical domain provides a controlled and verifiable foundation for establishing and validating the CHT framework, the proposed methodology is not inherently confined to mathematics. Illustrative examples demonstrating the applicability of CHT to non-mathematical contexts Appendix F. These preliminary extensions indicate the framework's potential for broader deployment. Furthermore, while the synthetic data generation pipeline employed in this work effectively alleviates the scarcity of high-order cognitive samples, constructing authentic, high-quality datasets to support tasks at advanced cognitive levels continues to pose critical challenges, advancing cognitive evaluation from basic capability verification toward systematic assessment of higher-order reasoning, remains a research issue that warrants sustained attention.

References

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*.

Yihong Dong, Jiazheng Ding, Xue Jiang, Ge Li, Zhuo Li, and Zhi Jin. 2025. Codescore: Evaluating code generation by learning code execution. *ACM Transactions on Software Engineering and Methodology*, 34(3):1–22.

Shaoyu Dou, Yutian Shen, Mofan Chen, Zixuan Wang, Jiajie Xu, Qi Guo, Kailai Shao, Chao Chen, Haixiang Hu, Haibo Shi, and 1 others. 2025. Fineval-kr: A financial domain evaluation framework for large language models' knowledge and reasoning. *arXiv preprint arXiv:2506.21591*.

Banca Calvo Figueras and Rodrigo Agerri. 2025. Benchmarking critical questions generation: A challenging reasoning task for large language models. *arXiv preprint arXiv:2505.11341*.

Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2022. Complexity-based prompting for multi-step reasoning. *arXiv preprint arXiv:2210.00720*.

Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, Lei Li, Chenghao Ma, Liang Chen, Runxin Xu, and 1 others. 2024. Omni-math: A universal olympiad level mathematic benchmark for large language models. *arXiv preprint arXiv:2410.07985*.

Xinyu Guan, Li Lyna Zhang, Yifei Liu, Ning Shang, Youran Sun, Yi Zhu, Fan Yang, and Mao Yang. 2025. rstar-math: Small llms can master math reasoning with self-evolved deep thinking. *arXiv preprint arXiv:2501.04519*.

Xiaoshuai Hao, Lei Zhou, Zhijian Huang, Zhiwen Hou, Yingbo Tang, Lingfeng Zhang, Guang Li, Zheng Lu, Shuhuai Ren, Xianhui Meng, and 1 others. 2025. Mimo-embodied: X-embodied foundation model technical report. *arXiv preprint arXiv:2511.16518*.

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*.

Zhongzhen Huang, Gui Geng, Shengyi Hua, Zhen Huang, Haoyang Zou, Shaoting Zhang, Pengfei Liu, and Xiaofan Zhang. 2025. O1 replication journey–part 3: Inference-time scaling for medical reasoning. *arXiv preprint arXiv:2501.06458*.

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.

David R Krathwohl. 2002. A revision of bloom's taxonomy: An overview. *Theory into practice*, 41(4):212–218.

Salima Lamsiyah, Kamyar Zeinalipour, Matthias Brust, Marco Maggini, Pascal Bouvry, Christoph Schommer, and 1 others. 2025. Arabicsense: A benchmark for evaluating commonsense reasoning in arabic with large language models. In *Proceedings of the 4th Workshop on Arabic Corpus Linguistics (WACL-4)*, pages 1–11.

Daniel Li and Lincoln Murr. 2024. Humaneval on latest gpt models–2024. *arXiv preprint arXiv:2402.14852*.

Deyi Li, Jialun Yin, Tianlei Zhang, Wei Han, and Hong Bao. 2024. The four most basic elements in machine cognition. *Data Intelligence*, 6(2):297–319.

Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, and 1 others. 2025. From system 1 to system 2: A survey of reasoning large language models. *arXiv preprint arXiv:2502.17419*.

Hongwei Liu, Zilong Zheng, Yuxuan Qiao, Haodong Duan, Zhiwei Fei, Fengzhe Zhou, Wenwei Zhang, Songyang Zhang, Dahua Lin, and Kai Chen. 2024. Mathbench: Evaluating the theory and application proficiency of llms with a hierarchical mathematics benchmark. *arXiv preprint arXiv:2405.12209*.

Jiayu Liu, Zhenya Huang, Wei Dai, Cheng Cheng, Jinze Wu, Jing Sha, Song Li, Qi Liu, Shijin Wang, and Enhong Chen. 2025. Cogmath: Assessing llms' authentic mathematical ability from a human cognitive perspective. *arXiv preprint arXiv:2506.04481*.

Doug Lombardi. 2023. On the horizon: The promise and power of higher order, critical, and critical analytical thinking. *Educational Psychology Review*, 35(2):38.

George F Madaus, Elinor M Woods, and Ronald L Nuttall. 1973. A causal model analysis of bloom's taxonomy. *American Educational Research Journal*, 10(4):253–262.

Arindam Mitra, Hamed Khanpour, Corby Rosset, and Ahmed Awadallah. 2024. Orca-math: Unlocking the potential of slms in grade school math. *arXiv preprint arXiv:2402.14830*.

Mohamed Nejjar, Luca Zacharias, Fabian Stiehle, and Ingo Weber. 2025. Llms for science: Usage for code generation and data analysis. *Journal of Software: Evolution and Process*, 37(1):e2723.

Shuyin Ouyang, Jie M Zhang, Mark Harman, and Meng Wang. 2025. An empirical study of the non-determinism of chatgpt in code generation. *ACM Transactions on Software Engineering and Methodology*, 34(2):1–28.

Tung Phung, Victor-Alexandru Pădurean, José Cambronero, Sumit Gulwani, Tobias Kohn, Rupak Majumdar, Adish Singla, and Gustavo Soares. 2023. Generative ai for programming education: benchmarking chatgpt, gpt-4, and human tutors. In *Proceedings of the 2023 ACM Conference on International Computing Education Research-Volume 2*, pages 41–42.

Mingyang Song, Zhaochen Su, Xiaoye Qu, Jiawei Zhou, and Yu Cheng. 2025. Prmbench: A fine-grained and challenging benchmark for process-level reward models. *arXiv preprint arXiv:2501.03124*.

Claudia J Stanny. 2016. Reevaluating bloom's taxonomy: What measurable verbs can and cannot say about student learning. *Education Sciences*, 6(4):37.

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.

Xiaobao Wu, Liangming Pan, Yuxi Xie, Ruiwen Zhou, Shuai Zhao, Yubo Ma, Mingzhe Du, Rui Mao, Luu Anh Tuan, and William Yang Wang. 2025. Antileakbench: Preventing data contamination by automatically constructing benchmarks with updated real-world knowledge. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18403–18419.

Shijie Xia, Xuefeng Li, Yixin Liu, Tongshuang Wu, and Pengfei Liu. 2025. Evaluating mathematical reasoning beyond accuracy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 27723–27730.

Jifan Yu, Mengying Lu, Qingyang Zhong, Zijun Yao, Shangqing Tu, Zhengshan Liao, Xiaoya Li, Manli Li, Lei Hou, Hai-Tao Zheng, and 1 others. 2023. Moocradar: A fine-grained and multi-aspect knowledge repository for improving cognitive student modeling in moocs. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2924–2934.

Zhongshen Zeng, Yinhong Liu, Yingjia Wan, Jingyao Li, Pengguang Chen, Jianbo Dai, Yuxuan Yao, Rongwu Xu, Zehan Qi, Wanru Zhao, and 1 others. 2024. Mr-ben: A meta-reasoning benchmark for evaluating system-2 thinking in llms. *Advances in Neural Information Processing Systems*, 37:119466–119546.

Xilin Zhang, Zhixin Mao, Ziwen Chen, and Shen Gao. 2025. Effective tool augmented multi-agent framework for data analysis. *Data Intelligence*, 6(4):923–945.

Qihao Zhao, Yangyu Huang, Tengchao Lv, Lei Cui, Qinzheng Sun, Shaoguang Mao, Xin Zhang, Ying Xin, Qiufeng Yin, Scarlett Li, and 1 others. 2025. Mmlu-cf: A contamination-free multi-task language understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13371–13391.

Pei Zhou, Jay Pujara, Xiang Ren, Xinyun Chen, Heng-Tze Cheng, Quoc V Le, Ed Chi, Denny Zhou, Swaroop Mishra, and Huaixiu Steven Zheng. 2024. Self-discover: Large language models self-compose reasoning structures. *Advances in Neural Information Processing Systems*, 37:126032–126058.

Maria-Eleni Zoumpoulidi, Eleni Batsi, Georgios Paraskevopoulos, Vassilis Katsouros, and Alexandros Potamianos. Bloomxplain: A framework and benchmark dataset for pedagogically sound llm-generated explanations based on bloom's taxonomy. In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*.

A Data Collection and Preprocessing Details

A.1 Details of Data Sources

Table 4 summarizes five core datasets used to construct the **BloomEval** benchmark, spanning K-12 to higher education. Omni-MATH includes symbolic and numerical problems from high school to Olympiad level. GSM8K focuses on grade-school arithmetic requiring multi-step reasoning. MATH covers competition-style problems from high school and college. Orca-MATH provides 200K synthetic elementary-level problems. Mooc_radar compiles college-level problems from MOOC platforms with learner interaction data.

Dataset	Problems	Venue
Omni-MATH	4,428	ICLR 2025
GSM8K	8,500	NeurIPS 2021
MATH	12,500	NeurIPS 2021
Orca-MATH	200,000	arXiv 2024
Mooc_radar	8,430	SIGIR 2023

Table 4: Overview of data sources.

A.2 Preprocessing Pipeline

To standardize the diverse formats from different sources, we implemented a three-stage preprocessing pipeline:

- **Format Unification & LaTeX Normalization:** All mathematical expressions were converted to standard LaTeX format. We applied regex-based normalization to unify variable representations (e.g., converting all Greek letters to their command forms, such as $\backslash alpha$) and standardize operator symbols.
- **Deduplication:** To eliminate duplication and high similarity among the five source datasets and avoid data leakage, we calculated the Jaccard similarity between sample pairs based on 13-gram text fragments and marked all sample pairs with a similarity greater than 0.8 as duplicates and removed them.
- **Low-Entropy Filtering:** We filtered out trivial samples and questions with insufficient context by calculating the semantic entropy of the problem text. Samples with a token count < 10 or lacking explicit mathematical predicates were discarded.

After applying the rigorous preprocessing and filtering pipeline described, the final **BloomEval** benchmark consists of **74,602** high-quality samples.

B Details of Cognitive Hierarchy Trace Construction

We aim to accurately project unstructured natural language reasoning into a structured cognitive space via CHT. The solution consists of two core parts:

B.1 Cognitive-Driven Step Formulation

CHT adopts a construction mechanism that strictly adheres to cognitive hierarchical constraints. This process unfolds through the following hierarchical trajectory:

- **Cognitive Decomposition**: The system initially identifies the core cognitive objective of the problem (e.g., the Apply level). Based on this benchmark, it maps the solution process onto a hierarchically progressive cognitive reasoning sequence. For instance, when the objective is defined as "Apply", the system constructs a progressive reasoning path: Remember (Knowledge Retrieval) $\rightarrow$ Understand (Condition Analysis) $\rightarrow$ Apply (Solution Execution).
- **Lexicon-Guided Instantiation**: Each cognitive stage in the planned path is instantiated as a concrete natural language atomic step, denoted as s_i. To operationalize this mapping, we employ a curated lexicon of *Cognitive Process Verbs* grounded in established educational taxonomies (Stanny, 2016) (refer to Table 5). The system strictly selects action verbs corresponding to the current stage's hierarchy to construct the syntactic structure.

B.2 Context Disambiguation Rules

To address potential verb polysemy (e.g., "find" can indicate simple retrieval or complex derivation depending on the context), we apply the following heuristic rules:

- **Rule 1: Maximum Pooling Principle**. For steps involving mixed operations, we assign the cognitive level of the most demanding operation. For instance, a step combining recall (Remember) and calculation (Apply) is mapped to $l_i = Apply$.

Level	Definition	Cognitive process verbs	Examples
Remember	Retrieve mathematical facts, definitions, or formulas from memory.	Recall, Identify, State, Define, List, Retrieve, Cite, Label, Name, Quote, Reproduce	• Recall the quadratic formula. • State the Pythagorean theorem. • List the factors of 24.
Understand	Construct meaning from problem statements or symbols; translation.	Explain, Recognize, Interpret, Translate, Paraphrase, Classify, Describe, Illustrate, Represent	• Translate the word problem into $2x + 5 = 15$. • Interpret $f'(x)$ as the slope. • Classify the triangle as isosceles.
Apply	Use procedures or algorithms to solve specific problems (Calculation).	Execute, Solve, Calculate, Compute, Use, Apply, Demonstrate, Implement, Find, Determine, Operate	• Calculate the determinant value. • Solve the linear system for x. • Apply L'Hôpital's rule.
Analyze	Break down problems; derive logical relationships or structures.	Derive, Deduce, Differentiate, Decompose, Infer, Correlate, Distinguish, Organize, Break down, Factor	• Derive the volume-radius relation. • Deduce that the function is increasing. • Decompose into partial fractions.
Evaluate	Make judgments based on criteria; verification and checking.	Verify, Check, Assess, Critique, Compare, Validate, Justify, Test, Confirm, Judge, Argue	• Verify $x = 3$ satisfies the equation. • Check boundary conditions. • Justify the step using Mean Value Theorem.
Create	Combine elements to form a new structure or strategy.	Construct, Formulate, Generate, Design, Hypothesize, Synthesize, Devise, Invent, Generalize, Propose	• Construct a proof with auxiliary lines. • Formulate the general term a_n. • Generalize the result to n-dimensions.

Table 5: Cognitive process verbs for mapping reasoning steps to Bloom's cognitive levels.

- **Rule 2: Taxonomy-Aligned Operation Mapping**. We classify implicit mathematical steps based on their cognitive mechanism:
  - *Remember*: Automatized retrieval of atomic facts (e.g., $3 \times 3 = 9$, identities).
  - *Understand*: Verification of properties or concept instantiation (e.g., substituting to check a root).
  - *Apply*: Algorithmic derivation requiring multi-step procedures (e.g., solving integrals).
  - *Analyze*: Logical inference driven by connectors (e.g., "therefore", "implies") without computation.
- **Rule 3: Structural Contribution**. Levels are assigned based on the step's role in the *Cognitive Hierarchy Trace*. We prioritize steps that actively modify the problem state or introduce dependencies, while mere restatements or reports are defaulted to the lowest applicable level.

To scale this method for the BloomEval, we embed these rules into the prompt design of our Expert-LLM collaborative framework (See Figure 12 and Figure 13).

C Multi-Stage Collaborative Annotation

C.1 Description of the Gold Standard Subset

Table 6 shows the Bloom's cognitive level distribution in the gold standard subset, ensuring full task coverage.

Bloom's Cognitive Level	Proportion	Total Count
1. Remember	10%	100
2. Understand	15%	150
3. Apply	20%	200
4. Analyze	20%	200
5. Evaluate	20%	200
6. Create	15%	150

Table 6: Distribution of gold standard subset.

C.2 Discrepancy Reconciliation Mechanism

When the three annotator models assign mutually inconsistent labels to the same problem, this stage resolves disagreements through a structured, stepwise procedure that combines a LRM with expert verification, ensuring both annotation validity and cognitive interpretability. The discrepancy resolution process consists of the following steps:

- **LRM Re-annotation** The disputed sample is first re-evaluated by a reasoning-oriented OpenAI-o1 under explicit CHT constraints, with the goal of producing a structurally co-

herent cognitive trace that serves as a reference for subsequent adjudication. The details of the prompt are shown in Figure14.

Expert Adjudication Criteria We enlisted three doctoral-level mathematics experts to conduct independent and rigorous assessments using a Likert scale, grounded in educational measurement standards. This evaluation framework encompasses four core dimensions:

- **Hierarchy Completeness:** All prerequisite cognitive levels required by the task are present in the reasoning trace.
- **Hierarchy Validity:** The reasoning trace contains hierarchy jumps or hierarchy breaks.
- **Cognitive Sufficiency:** Higher-order cognitive operations are included only when they are necessary for solving the problem.
- **Knowledge Alignment:** The invoked knowledge points align with both the annotated cognitive level and the conceptual scope of the problem.

Each dimension is scored from 1 to 5, with higher scores indicating superior performance. Regarding the final cognitive label determination, a majority vote mechanism is employed, where the final label is adopted if at least two experts reach a consensus (e.g., two experts select "Analyze" while one selects "Evaluate").

Problem: Let a, b, c > 0 with a + b + c = 1. Prove that a/(b + c) + b/(c + a) + c/(a + b) ≥ 3/2.

Initial Model Annotations — Fully Disputed

- LLaMA-3-70B: Apply
  Rationale: Treats the task as direct application of a known inequality.
- Qwen-2.5-72B: Analyze
  Rationale: Focuses on structural transformation of the expression before applying inequalities.
- GPT-3.5-turbo: Evaluate
  Rationale: Emphasizes verification of the lower bound and equality case.

Strong LRM Re-annotation

Step 1 (Remember): Recall the .... Titu's lemma.
Step 2 (Understand): Recognize ...are positive.
Step 3 (Apply): Apply … rational bound.
Step 4 (Analyze): Simplify t… on pairwise products.
Step 5 (Evaluate): Compare …and verify the equality case a = b = c = 1/3.

Expert Decision

The expert determines that the decisive cognitive operation is not routine execution of a known inequality, but the structural decomposition of the expression and the selection of an appropriate inequality framework. Verification of the bound is a necessary but secondary step. Therefore, the expert assigns the task to the Analyze level rather than Apply or Evaluate.

Final Annotation

- Cognitive Level: Analyze
- Difficulty: Medium
- Knowledge Point Hierarchy: Mathematics → Inequalities → Cauchy–Schwarz
- Cognitive Goal Analysis: Analyze the structure of a symmetric inequality and match it to an appropriate inequality tool to derive the required bound.

Figure 6: An example of discrepancy resolution.

Figure 6 illustrates a fully disputed case in which the three annotator models assign different Bloom's cognitive levels (Apply, Analyze, and Evaluate) to the same mathematical inequality problem. Although all model judgments appear locally plausible, the disagreement indicates ambiguity in identifying the dominant cognitive operation required by the task.

C.3 Data Generation for Remember Levels

To mitigate data sparsity at the Remember level, we generate single-step recall problems using a prompt template that restricts reasoning to factual retrieval only. The prompt enforces a one-step CHT-aligned solution and is applied uniformly across all generated samples (see Figure 15) that guides models to generate Level 1 (Remember) problems and output structured annotations.

Problem:
What is the formula for the surface area of a regular icosahedron with edge length a.

Knowledge points Hierarchy: Mathematics — Geometry — Surface Area

Bloom's Cognitive level: Remember **Difficulty**: Difficult

Solution steps:
Step 1 (Remember): Recall that the surface area of a regular icosahedron is given by $5\sqrt{3}$ times the square of the edge length.

Figure 7: An example of Level 1 (Remember) problem.

As illustrated in Figure 7, the Level 1 (Remember) problem targets students' ability to recall factual knowledge. The example for the surface area formula of a regular icosahedron, requiring no reasoning or transformation beyond direct memory retrieval. Demonstrating how factual recall is concretely instantiated in our benchmark.

C.4 Data Generation for Create Levels

For Create-level augmentation, we employ a structured prompt template that explicitly requires multi-step reasoning progressing from lower-order cognitive operations to final construction. This template is fixed across generation rounds to ensure consistency in cognitive hierarchy representation (see Figure 16).

As illustrated in Figure 8, the Level 6 (Create) problem asks students to design a recursive algorithm for counting binary strings of length n that avoid consecutive 1s, and to prove the resulting recurrence relation using induction. The problem integrates multiple stages of reasoning from recalling binary string structure, understanding constraints, and analyzing small cases, to deriving and

Type	Model	Key Features	Release Year
LRMs	Grok-3	Conversational reasoning, factual alignment	2024
	DeepSeek-R1	Symbolic math, bilingual training	2025
	Claude 3.7	Long-context, structured reasoning	2025
	DeepSeek v3	Hierarchical tasks, compact tuning	2025
	OpenAI o1-mini	Fast inference, balanced performance	2024
	s1-32B	Large model, factual extraction	2023
	Kimi k1.5	Long CoT reasoning, hierarchy modeling	2024
	Gemini-2.5-Pro	Structured QA, multimodal alignment	2025
	OpenAI o3	Generalist reasoning, robustness	2024
LLMs	Claude 3.5	Logical consistency, CoT prompts	2023
	Gemini-1.5	Balanced summarization, instruction tuning	2024
	GPT-4o	Multimodal, tool use, code generation	2024
	Qwen3-Max	Chinese-English CoT, inference efficiency	2025
	OpenAI o4-mini	Lightweight, task-oriented reasoning	2024

Table 7: Overview of baseline models and key features.

proving a recurrence relation. Each step in the solution explicitly aligns with a distinct Bloom's cognitive level, making this a canonical example of higher-order combinatorial problem solving. After

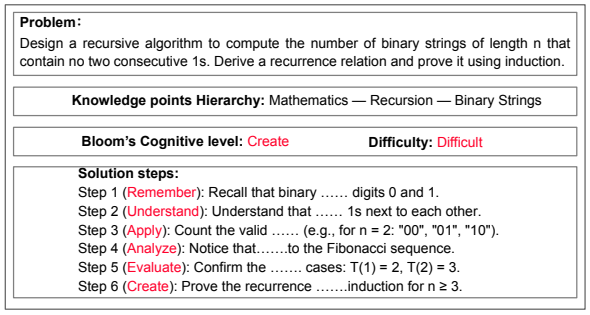

Figure 8: An example of a Level 6 (Create) problem.

generating 20,000 Level 1 and Level 6 samples, the dataset expanded to 94,602 problems. Partial data has been provided in the attachment. After the paper is published, all data and code will be open sourced.

D Experimental Details

D.1 Model Selection

We select LRMs and LLMs that have demonstrated strong performance in reasoning benchmarks over the past two years. (Table 7).

D.2 Bloom's Cognitive Level Prediction

This task evaluates whether LRMs can predict the Bloom's cognitive level required to solve a problem, using few-shot reasoning with cognitive level-aligned solution steps. The following example (Figure 9) illustrates how the introduction of a reasoning chain enables the model to perform more accurate Bloom's cognitive level analysis.

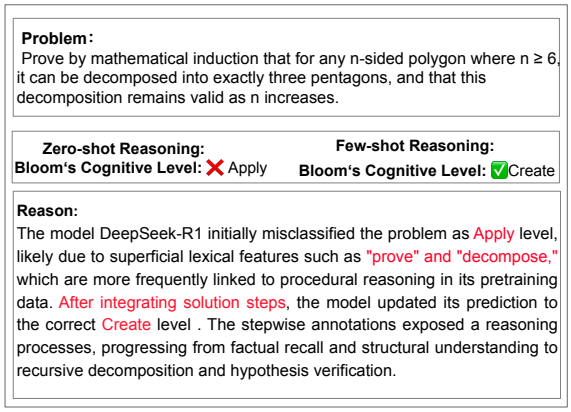

Figure 9: An example of Bloom's cognitive level prediction.

E Detailed Ablation Study

To strictly validate the marginal contribution of each component within the CHT) framework, we conducted a controlled ablation study. This appendix details the experimental setup, the definition of ablation variants, and a qualitative analysis of the failure modes observed when specific constraints were removed.

E.1 Experimental Setup

Dataset Construction We constructed an evaluation subset comprising 400 high-order cognitive problems. To ensure representativeness, the samples were stratified across difficulty levels, cognitive levels and covered diverse mathematical subdomains (Algebra, Geometry, Calculus).

Base Model We utilized **Grok-3** as the backbone reasoner for all ablation experiments due to its superior performance in the main evaluation (achieving the highest overall accuracy among tested models).

Human Evaluation Protocol Three doctoral-level mathematics experts manually evaluated the generated reasoning trajectories. They assessed two key dimensions:

- **Structural Integrity:** Whether the reasoning follows a logical progression without unexplained jumps.
- **Cognitive Alignment:** Whether the labeled cognitive level matches the actual semantic action of the step.

For each ablation setting, experts independently annotated the reasoning trajectories to identify specific structural anomalies: hierarchy break, hierarchy jump, and overthinking. Regarding the final determination of modes, we adopted a majority vote mechanism, a structural error was confirmed only if at least two experts reached a consensus.

E.2 Definition of Ablation Variants

We compared the full CHT framework against two degraded variants to isolate the impact of **Hierarchical Constraints** and **Cognitive Process Verbs**.

- **Full CHT (Ours):** The standard setting using the prompt defined in Appendix B, which includes strict atomic step segmentation, the full Cognitive Process Verbs dictionary, and Disambiguation Rules (1–3).
- **w/o Hierarchy Constraints:** In this setting, we removed the explicit instructions regarding Atomic Step Segmentation and Disambiguation Rules (e.g., Rule 2 regarding implicit actions). The model was asked to "think step-by-step" but was not forced to adhere to strict logical precedence rules.
- **w/o Cognitive Process Verbs:** In this setting, we removed the **Cognitive Process Verbs Dictionary** (semantic anchors) from the prompt. The model was given only the abstract definitions of Bloom's levels without the specific verb mapping table.
- **Standard CoT:** This setting represents the vanilla Chain-of-Thought prompting strategy (Wei et al., 2022). The model is simply instructed to "solve the problem step-by-step" without any constraints related to BCT, cognitive process verbs, or structured formatting. This serves as a control group to evaluate the intrinsic cognitive coherence of the model's native reasoning process compared to the structured CHT generation.

F Cross-Domain Generalization of CHT

Bloom's Cognitive Taxonomy offers a stable and generalizable method for defining cognitive levels, serving as a reliable reference across diverse domains. Our core mechanism, the Cognitive Hierarchy Trace (CHT), does not rely on mathematical symbols or specific formats tied to a single subject. Therefore, the framework can be easily extended to other areas, such as logical reasoning (e.g., the CLUTRR dataset) or legal case analysis (e.g., LexGLUE). Below are two brief examples to demonstrate how the framework can be transferred to non-mathematical tasks.

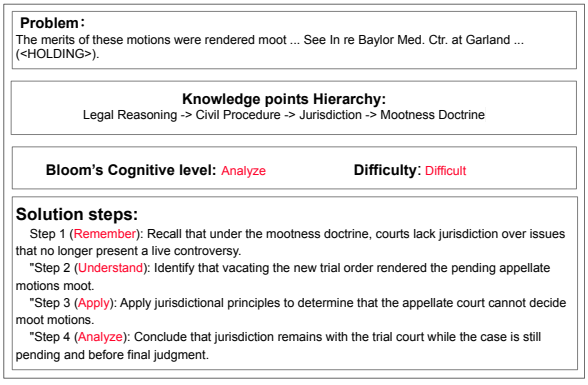

Figure 10: An example of legal case analysis (LexGLUE).

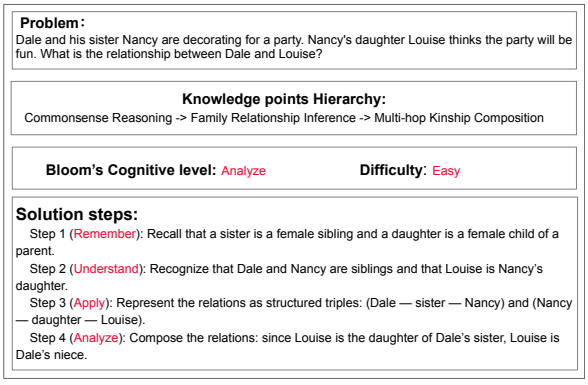

Figure 11: An example of Logical Reasoning (CLUTRR Dataset).

You are a teacher, you aim to evaluate each problem in terms of Bloom's cognitive level, difficulty, knowledge point hierarchy.

 # OBJECTIVE #

You are required to evaluate each problem from four perspectives:

1 Bloom's cognitive level
2 Problem difficulty
3 Knowledge point hierarchy

Your detailed tasks:

a. Notational Equivalence: Carefully examine LaTeX expressions to ensure equivalent notation and semantic accuracy.
b. Cognitive Level Analysis: Determine the Bloom level (1–6) using cues like "memory", "application", "comparison", "design", etc., and justify the level.
c. Step-by-Step Reasoning Aligned to Bloom's cognitive levels: Decompose the problem-solving process from lower to higher cognitive levels (Remember → Understand → Apply →Analyze → Evaluate → Create), and avoid steps that exceed the final Bloom's cognitive level.
d. Difficulty Classification: Score difficulty as 1 (Easy), 2 (Medium), or 3 (Difficult), specifically within the identified Bloom's cognitive level.
e. Knowledge Point Hierarchy: Summarize the layered knowledge structure (e.g., Algebra → Equations → Quadratics).

 # STYLE #

Structured report format.

 # TONE #

Scientific, professional.

 # AUDIENCE #

This prompt is for educational researchers and students. The goal is to support students in understanding how problems reflect different Bloom's cognitive levels, help instructors assess learning objectives, and guide curriculum alignment.

 # EXAMPLES DESIGN #

#Example 1#

#Example 2#

#Example 3#

……..

 # OUTPUT TEMPLATE #

[Problem]

[Insert problem in LaTeX]

[Knowledge Point Hierarchy]

[Summarize the domain → subdomain → concept, e.g., Algebra → Functions → Linear Recurrence]

[Bloom's Cognitive level]

[1–6, based on Bloom's Taxonomy]

[Difficulty]

[1: Easy / 2: Medium / 3: Difficult]

[Final Answer]

[Final answer only]

 # END #

Figure 12: Prompt template for annotation based on Bloom's Cognitive Taxonomy.

You are an expert Cognitive Scientist and Mathematician specializing in Bloom's Cognitive Taxonomy. you aim to analyze the cognitive depth of mathematical reasoning steps with high precision.

 # OBJECTIVE #

Your task is to map a specific Atomic Reasoning Step from a mathematical solution to its corresponding Bloom's Cognitive Level. You must strictly adhere to the provided verb dictionary and disambiguation rules.

#COGNITIVE PROCESS VERB DICTIONARY (SEMANTIC ANCHORS)#

Remember: Recall, Identify, State, Define, List, Retrieve, Cite.
Understand: Explain, Interpret, Translate, Recognize, Paraphrase.
Apply: Calculate, Solve, Compute, Use, Execute, Determine (numeric).
Analyze: Derive, Deduce, Decompose, Infer, Organize, Differentiate.
Evaluate: Verify, Check, Assess, Compare, Justify, Validate.
Create: Construct, Hypothesize, Design, Formulate, Synthesize.

 # DISAMBIGUATION RULES (STRICT CONSTRAINTS)#

Rule 1: Maximum Pooling Principle
If a step involves multiple operations, assign the highest cognitive level found (e.g., Recall + Calculate -> Apply).
Rule 2: Taxonomy-Aligned Operation Mapping
Classify mathematical operations based on cognitive intent, not just surface form:
L1 Remember (Fact Retrieval): Instantaneous retrieval of atomic facts or constants (e.g., "3*3=9", "pi=3.14") without algorithmic processing.
L2 Understand (Verification): Operations performed solely to verify a concept or check a property (e.g., substituting x=2 to check if it's a root).
L3 Apply (Derivation): Non-trivial execution of algorithms to derive new values (e.g., solving quadratic equations, complex arithmetic like "23*45").
L4 Analyze (Inference): Steps driven by logical markers ("Therefore", "Implies") representing synthesis rather than computation.
Rule 3: Structural Contribution
Assign higher levels only to steps that actively modify the problem state or derive new information. Mere restatements or summaries must be assigned to the lowest applicable level.

 # STYLE #

Structured report format.

 # TONE #

Scientific, professional.

 # AUDIENCE #

This prompt is for educational researchers and students. The goal is to support students in understanding how problems reflect different Bloom's cognitive levels, help instructors assess learning objectives, and guide curriculum alignment.

 # EXAMPLES DESIGN #

#Example 1#

#Example 2#

#Example 3#

……..

 # OUTPUT TEMPLATE #

[Problem]

[Insert problem in LaTeX]

[Knowledge Point Hierarchy]

[Summarize the domain → subdomain → concept, e.g., Algebra → Functions → Linear Recurrence]

[Bloom's Cognitive level]

[1–6, based on Bloom's Taxonomy]

[Difficulty]

[1: Easy / 2: Medium / 3: Difficult]

[Solution steps]

[“Step 1 (Remember):……”

“Step 2 (Understand):……”

“Step 3 (Apply):……”

……]

[Final Answer]

[Final answer only]

 # END #

Figure 13: Prompt template for annotation on CHT.

You are a Cognitive Scientist acting as an arbitrator for a cognitive evaluation benchmark.

 # OBJECTIVE #

You are provided with a mathematics problem where junior annotator models have disagreed on its Bloom's Cognitive Level. Your goal is to resolve this dispute by generating a rigorous “Solution steps” to reveal the underlying cognitive structure.

Your detailed tasks:

a. Solution steps: Solve the problem step-by-step. For EACH step, strictly assign one of the 6 Bloom's Cognitive levels (Remember, Understand, Apply, Analyze, Evaluate, Create).

b. Verb-Driven Constraints: You MUST start the content of each step with a specific cognitive verb (e.g., "Recall", "Calculate", "Derive", "Verify") that accurately reflects the mental action.

c. Atomic Granularity: Ensure each step represents a single, atomic cognitive operation. Do not merge multiple levels (e.g., specific calculation vs. abstract derivation) into one step.

d. Peak Load Determination: Identify the "Dominant Cognitive Level" of the entire problem. This is defined as the highest cognitive level required by the most critical/difficult step in the solution path (Maximum Pooling Principle).

 # STYLE #

Structured, logical, and rigorous.

 # TONE #

Objective, analytical, and authoritative.

 # AUDIENCE #

Human domain experts who need a clear reference trace to make a final decision on the cognitive label.

 # COGNITIVE PROCESS VERB LEXICON (Reference) #

- Remember: Recall, Identify, State, Define
- Understand: Explain, Recognize, Interpret, Translate
- Apply: Calculate, Solve, Apply, Determine
- Analyze: Derive, Deduce, Decompose, Distinguish
- Evaluate: Verify, Check, Justify, Assess
- Create: Construct, Design, Generalize, Formulate

 # EXAMPLES DESIGN #

#Example 1#

#Example 2#

#Example 3#

……..

 # OUTPUT TEMPLATE #

[Problem]

[Insert problem in LaTeX]

[Cognitive Hierarchy Trace]

Step 1 (Level): [Verb] [Content]

Step 2 (Level): [Verb] [Content]

...

Step N (Level): [Verb] [Content]

[Adjudication Analysis]

[Briefly explain which step carries the highest cognitive load and why]

[Final Proposed Label]

[One of: Remember / Understand / Apply / Analyze / Evaluate / Create]

[Difficulty]

[1: Easy / 2: Medium / 3: Difficult]

END

Figure 14: Prompt template for re-annotation.

You are a teacher tasked with generating problems targeting Bloom's cognitive level of Remember . You aim to generate problems that evaluate a student's ability to recall basic facts, formulas, or definitions, across diverse topics and difficulty levels.

 # OBJECTIVE #
The objective is to generate JSON-formatted problems that strictly conform to Bloom's cognitive level of Remember and encompass varying degrees of factual recall difficulty:
a. The problem must require only recall (e.g., definition, formula, arithmetic fact) without involving comprehension or inference.
b. The output format must match the exact field names and formatting as defined in the instruction, including: problem, Bloom's cognitive level, problem difficulty, cognitive level-aligned solution steps, and knowledge points hierarchy and final answer.

 # STYLE #
Responses must be valid JSON objects, structured consistently.

 # TONE #
Precise, rigorous, and rich.

 # AUDIENCE #
AI model developers and educators evaluating capabilities in LRMs.

 # OUTPUT TEMPLATE #
Each problem must return a valid JSON object using the following structure:

```
{
  [Problem]: [A single factual recall question. Use LaTeX with double backslashes.],
  [Knowledge Point Hierarchy]: [Mathematics domain, e.g., Mathematics -> Geometry -> Area Formulas],
  [Bloom's cognitive level]: 1,
  [Difficulty]: [1 | 2 | 3],
  [Cognitive Goal Analysis]: [Remember: Recall [fact/definition/formula].],
  [Solution steps]: [Step 1 (Remember): [Describe the specific memory recall process].],
  [Final Answer]: [Correct recalled value or expression].
}
```

 # INSTRUCTIONS #
- Only generate problems aligned with Bloom's remember level (Level 1).
- Do NOT include multi-step reasoning, explanation, or any operations beyond factual recall.
- Ensure variety across topics (e.g., arithmetic, algebra, geometry, number theory).
- Format LaTeX with escaped slashes (e.g., \\times, \\frac).
- Output only valid JSONs separated by `----` with no extra explanations.
- Avoid omission or renaming of keys.

 # EXAMPLES DESIGN #
#Example 1#
#Example 2#
#Example 3#
……..
 # END #

Figure 15: Prompt template for generating Level 1 (Remember) problems.

You are a teacher tasked with generating problems targeting Bloom's cognitive level of Create. These problems are intended to evaluate a student's ability to synthesize knowledge, build new structures, and perform generative reasoning across multiple domains.

 # OBJECTIVE #

I need you to generate JSON-formatted problems that strictly align with Bloom's cognitive level of Create and embed the full cognitive hierarchy (from remember to create construction). Each problem must:

a. Require reverse reasoning, multi-step synthesis, or creative conjecture-making.

b. Progress through Bloom's cognitive levels (Remember → Understand → Apply → Analyze → Evaluate → Create).

c. Lead to the construction, proof, or invention of a new idea, formula, or structure.

d. Be original and span diverse topics like algebra, number theory, geometry, combinatorics, or functions.

e. The output format must match the exact field names and formatting as defined in the instruction, including:
problem, Bloom's cognitive level, problem difficulty, cognitive level-aligned solution steps, and the knowledge points hierarchy and final answer.

 # STYLE #

Responses must be valid JSON objects, structured consistently.

 # TONE #

Precise, rigorous, and rich.

 # AUDIENCE #

AI model developers and educators evaluating higher-order thinking capabilities in LRMs.

 # OUTPUT TEMPLATE #

Each problem must return a valid JSON object with the following structure:

```
{
  [Problem]: [An original, non-trivial creative problem involving conjecture, reverse reasoning, or generative proof. Use LaTeX with double backslashes.],
  [Knowledge Point Hierarchy]: [Mathematics -> Subdomain -> Topic],
  [Bloom's Cognitive level]: 6,
  [Difficulty]: [1 | 2 | 3],
  [Cognitive Goal Analysis]: Remember: Recall [basic fact]. Understand: Interpret [structure or constraint].Apply: Apply [technique]. Analyze: Identify [pattern or relationship]. Evaluate: Judge [hypothesis or structure]. Create: Construct or prove [conjecture, formula, or design.],
  [Solution steps]: [
    "Step 1 (Remember): [Recall a relevant formula or definition.]",
    "Step 2 (Understand): [Interpret the given setting or parameters.]",
    "Step 3 (Apply): [Apply known techniques to specific or small cases.]",
    "Step 4 (Analyze): [Identify structure or pattern emerging.]",
    "Step 5 (Evaluate): [Test the validity, generality, or limitations of the observed structure.]",
    "Step 6 (Create): [Construct a new expression, prove a general result, or build a creative formulation.]"
  ],
  "Final Answer": "[A clearly stated formula, structure, proof idea, or construction.]"
}
```

 # INSTRUCTIONS #

- Only generate problems aligned with Bloom's Create level.
- Must involve multi-layered reasoning with increasing cognitive depth.
- The final task must require the learner to create, not just solve.
- Do not reuse textbook questions; all content should be original.
- Format LaTeX with escaped slashes (\\times, \\frac, \\sum, etc.).
- Output only valid JSON. Separate multiple problems using `----`.

 # EXAMPLES DESIGN #

#Example 1#

#Example 2#

#Example 3#

……..

 # END #

Figure 16: Prompt template for generating Level 6 (Create) problems.