DeepQuestion: Systematic Generation of Real-World Challenges for Evaluating LLMs Performance

Ali Khoramfar, Ali Ramezani, Mohammad Mahdi Mohajeri
Mohammad Javad Dousti, Majid Nili Ahmadabadi, Heshaam Faili
Department of Electrical and Computer Engineering, College of Engineering,
University of Tehran, Tehran, Iran
{khoramfar, ali.ramezani.96, mehdimohajeri, mjdousti, mnili, hfaili}@ut.ac.ir

Abstract

While Large Language Models (LLMs) achieve near-human performance on standard benchmarks, their capabilities often fail to generalize to complex, real-world problems. To bridge this gap, we introduce DeepQuestion, a scalable, automated framework that systematically elevates the cognitive complexity of existing datasets through controlled task transformations grounded in explicit cognitive hierarchies. Based on Bloom's taxonomy, DeepQuestion generates (1) scenario-based problems to test the application of knowledge in noisy, realistic contexts, and (2) instruction-based prompts that require models to create new questions from a given solution path, assessing synthesis and evaluation skills. Our extensive evaluation across ten leading open-source and proprietary models, covering both general-purpose and reasoning LLMs, reveals a stark performance decline—with accuracy dropping by up to 70%—as tasks ascend the cognitive hierarchy across evaluation settings. These findings underscore that current benchmarks overestimate true reasoning abilities and highlight the critical need for cognitively diverse evaluations to guide future LLM development.

Keywords: Evaluation Methodologies, Cognitive Methods, Learning Science inspired Evaluation

1. Introduction

Recent advances in Large Language Models (LLMs) have driven remarkable improvements across a wide spectrum of benchmarks, from arithmetic reasoning in GSM8K to expert-level performance on MMLU and GPQA (Achiam et al., 2023; Team et al., 2023; Grattafiori et al., 2024; Team et al., 2025; Cobbe et al., 2021; Hendrycks et al., 2021b,a; Wang et al., 2024; Rein et al., 2024). Yet, as scores approach saturation, evidence increasingly suggests that such benchmarks do not reflect the demands of real-world reasoning. When confronted with tasks that include irrelevant details, ambiguous contexts, or creative synthesis, even state-of-the-art LLMs exhibit significant failures—revealing an overreliance on pattern recognition rather than genuine understanding (Arias-Duart et al., 2025; Myrzakhan et al., 2024).

Emerging studies have begun to expose this gap. Models that perform flawlessly on structured academic datasets often falter on problems drawn from authentic or modern settings, such as new mathematics contests, clinical scenarios, or symbolic reasoning variations (Shojaee et al., 2025; Mirzadeh et al., 2025; Petrov et al., 2025; Alaa et al., 2025). These shortcomings highlight that current benchmarks primarily target lower-order cognitive abilities—such as recall and comprehension—while neglecting the deeper reasoning, analysis, and creativity required for practical decision-making. Without cognitively diverse evaluation frameworks, progress risks being misinterpreted as intelligence rather than memorization.

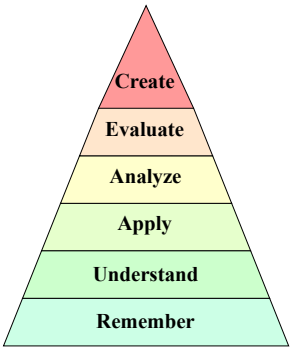


Figure 1: Bloom's taxonomy hierarchy.

To address this limitation, we introduce DeepQuestion, a systematic framework for generating cognitively enriched evaluation datasets grounded in Bloom's taxonomy of learning (Forehand, 2010; Leung, 2000), a widely recognized framework outlining six cognitive skill levels, from basic recall to higher-order reasoning (see Figure 1). DeepQuestion transforms existing question–answer pairs into (1) scenario-based problems that test the application of knowledge in realistic, constraint-rich situations, and (2) instruction-based prompts that assess creation and evaluation skills by requiring models to design new questions aligned with given solution paths. This taxonomy-guided approach adds interpretable layers of complexity that correspond directly to human cognitive processes.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 11451–11460
11–16 May 2026.

Through extensive experiments across ten leading LLMs—spanning general-purpose and reasoning-oriented architectures—we observe sharp performance declines, up to 70%, as the cognitive level of tasks increases. These results provide evidence that LLMs' strong benchmark performance does not generalize to deeper reasoning. Beyond evaluation, the DeepQuestion framework offers a replicable foundation for constructing cognitively meaningful benchmarks across diverse disciplines, from mathematics and physics to law and medicine. Our main contributions are:

- Proposed DeepQuestion, a framework for generating questions from existing datasets for better LLM evaluation based on Bloom's taxonomy.
- Introduced the DeepQuestion dataset, created using our framework.
- Conducted a comprehensive evaluation of LLMs with the DeepQuestion dataset, highlighting their knowledge limitations across Bloom's taxonomy.

The remainder of this paper is organized as follows: Section 2 reviews the related works, Section 3 introduces the DeepQuestion framework, Section 4 details our experimental setup, and the final sections present our results and conclusions.

2. Related Works

This paper draws on ideas from two main areas: the levels of learning, derived from the learning sciences, and the limitations of LLMs in solving complex and real-world problems. In this section, we review the previous work related to these two domains.

2.1. Learning Levels

In the learning sciences, various definitions of learning have been proposed. Some of these frameworks conceptualize learning as occurring at different levels. Among the most influential are the SOLO taxonomy and Bloom's taxonomy, both of which define learning across hierarchical stages (İlhan et al., 2017).

In the SOLO taxonomy (Biggs and Collis, 2014), the lowest level of learning is Pre-structural, where no meaningful learning has occurred. The highest level, known as Extended Abstract, indicates that the learner is able to extend their understanding and apply their knowledge to new domains.

Bloom's taxonomy (Bloom, 1971; Anderson and Krathwohl, 2001; Forehand, 2010) offers a more widely adopted framework. It also defines learning in multiple levels and exists in several versions. In its most recent revision, the lowest level is *Remember*, in which the learner simply recalls information without deeper understanding. A higher level, apply (the third level), represents the learner's ability to use acquired knowledge to solve real-world problems. The highest level, *Create*, reflects the learner's capacity to generate new ideas or concepts based on previously acquired knowledge.

Employing these taxonomies to assess the performance of LLMs can provide a more human-aligned perspective on evaluating their knowledge and reasoning capabilities.

2.2. LLMs Evaluation in Complex and Real-world Problems

With the rapid advancement of LLMs, their ability to solve complex problems has become one of the key questions in the field. Moreover, as the potential use of LLMs in real-world contexts grows, evaluating their performance on real-world problems has become increasingly important.

MMLU (Hendrycks et al., 2021a) serves as a standard benchmark for evaluating LLMs across diverse academic and professional domains. However, as model performance on MMLU has approached saturation, MMLU-Pro (Wang et al., 2024) was introduced to assess models on more complex, reasoning-intensive tasks. Establishing such advanced benchmarks is essential to accurately measure progress and ensure meaningful evaluation of LLMs' capabilities in challenging real-world scenarios. This aligns with findings from The Illusion of Thinking (Shojaee et al., 2025), which highlight the limitations of current LLMs when confronted with more complex reasoning tasks.

Recent studies (Chauhan et al., 2025; Alaa et al., 2025) have examined the performance of LLMs in real-world scenarios and compared it with their performance on standard benchmarks. These studies indicate a notable reduction in LLM performance when applied to real-world contexts. Given the growing interest in deploying these models in practical applications, assessing their performance in real-world settings is crucial. Consequently, rigorous evaluation of LLMs under such conditions has become increasingly important.

3. Method

This study is guided by a key research question: Can existing benchmarks be systematically deepened to reflect real-world task complexity, rather than introducing artificial difficulty?

To address this, we leverage insights from learning science—particularly Bloom's taxonomy—which offers a principled framework for systematically controlling and measuring cognitive

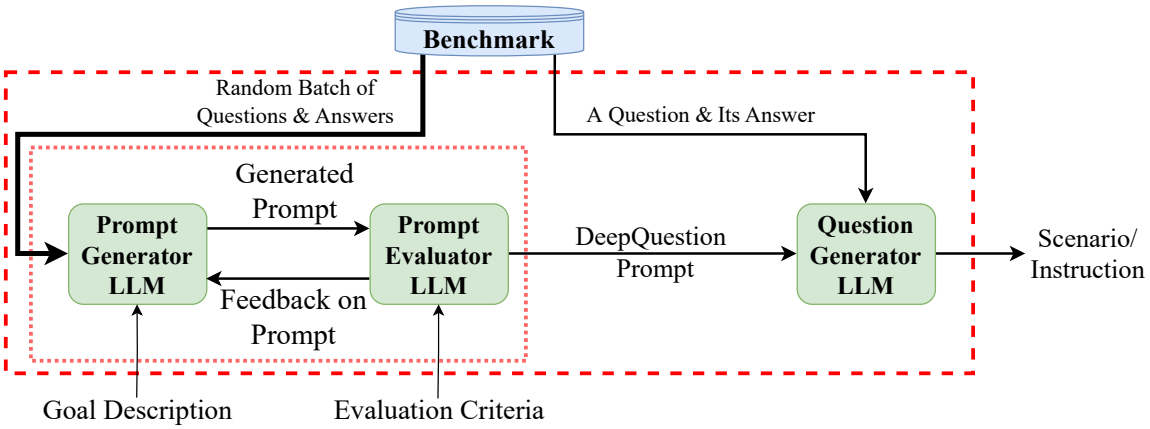


Figure 2: Overview of DEEPQUESTION framework. It begins with the selection of a random batch of questions and answers. Then, by conversation between the prompt generator and the prompt evaluator LLMs, the deep-question prompt is generated. The question generator LLM with the deep-question prompt converts each question and answer pair to the deep question and answer.

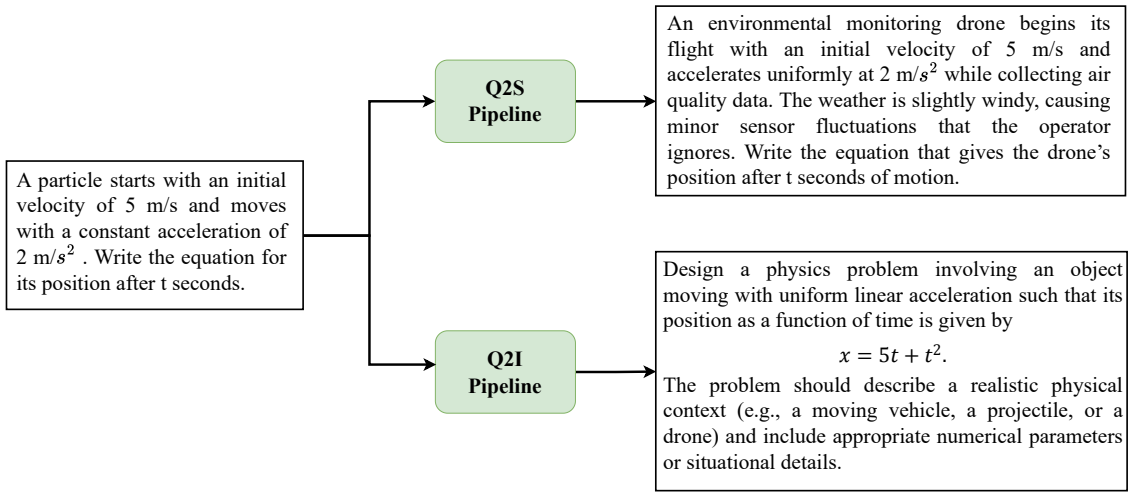


Figure 3: Examples of question transformations produced by the Q2S and Q2I pipelines

complexity. Unlike ad-hoc increases in difficulty, Bloom's hierarchy allows us to add real, interpretable complexity to benchmark items, specifying exactly how and by how much each transformation increases the cognitive demands.

Prior analyses reveal that most benchmarks remain at lower Bloom levels—*Remember* and *Understand* (Alaa et al., 2025; Chauhan et al., 2025)—and rarely stress the higher-order reasoning required in practical tasks. To bridge this gap, we focus on the *apply* and *create* levels, which represent critical forms of real-world reasoning often underrepresented in evaluation. Specifically, we propose two aligned methods: (1) question-to-scenario (Q2S), which increases complexity by embedding items in realistic, constraint-rich contexts to elicit *apply*-level reasoning; and (2) question-to-instruction (Q2I), which reframes items as multi-step, constructive procedures with explicit acceptance criteria, targeting the *create* level of synthesis and innovation. These transformations go beyond surface-level difficulty and require test-takers to transfer, adapt, and synthesize knowledge in meaningful, task-driven ways.

By systematically deepening benchmarks along Bloom's hierarchy using Q2S and Q2I, our DEEPQUESTION framework enables precise diagnosis of model weaknesses—revealing, for example, whether a model fails at context adaptation, constraint handling, or constructive synthesis. This fine-grained understanding directly informs concrete intervention strategies for model improvement, such as targeted data augmentation, objective design, or architectural changes. Implementation details and our automated prompt-generation pipeline are described in the following section.

3.1. Question-to-Scenario Generation (Q2S)

To deepen questions, Q2S targets the third level of Bloom's taxonomy, namely *Apply*, which emphasizes using knowledge in practical contexts. Each original question is transformed into a scenario-based version embedding the core problem within a realistic narrative, often including extraneous details and some distractions to simulate real-world complexity. For example, as shown in Figure 3, a basic physics question asking for the position-time equation of an object, is reframed within a scenario involving an environmental drone, where irrelevant elements like weather acted as distractors. The core question data and required reasoning remain unchanged. Q2S employs an automated prompt-based pipeline powered by an LLM (see Subsection 3.3 for more details on the pipeline). The output is a benchmark of scenario-driven questions that better reflect authentic problem-solving contexts.

These scenario-based questions resemble real-world problems. To solve them, LLMs must extract relevant information from the scenario and apply their knowledge accordingly. This process reflects a higher cognitive level than the *Remember* and *Understand* levels.

3.2. Question-to-Instruction Generation (Q2I)

Encouraged by the initial result analysis at the *Apply* level of Bloom's taxonomy with Q2S, we extend our approach to the higher *Evaluate* and *Create* levels. At these levels, learners are expected to assess information critically and produce novel outputs. We hypothesize that individuals or models with deep conceptual understanding are not only able to solve problems but also to design meaningful questions. Designing a question requires the evaluation of existing knowledge in order to shape a coherent concept. It also involves the ability to establish a logical and structured process that leads to the formulation of a meaningful question. This process demands both critical reflection and creative synthesis of ideas. Consequently, the capacity to design questions can be used to assess competencies at both the evaluate and create levels of cognition. This informs our second experimental setup, which reverses the direction of the task: instead of solving a question, the model is instructed to design one.

For each original dataset question, Q2I constructs an instruction designed to prompt an LLM to generate a new question that preserves the same topic and solution path as the original. This instruction is formulated in a way that ensures a different LLM, when given the instruction, would also produce a question aligned with the original in both topic and reasoning process. For example, if the original question in Figure 3 (position-time equation of an object) yields the solution $x = 5t + t^2$, the instruction would ask the model to design a physics problem whose solution would be exactly that equation. Designing questions based on instruction implicitly tests three levels of the model's reasoning: conceptual understanding of the domain and relevant equations, the ability to define appropriate variables and their interrelations, and the selection of values that yield the target solution.

Q2I employs an automated prompt-based pipeline similar to Q2S; however, the output of Q2I is the set of instructions for question generation. The next subsection details the pipeline.

3.3. Prompt Generation Pipeline

Since transforming each benchmark into its Q2S and Q2I variants requires carefully crafted prompts tailored to the specific domain and style of that benchmark, a key component of the DEEPQUESTION framework is a prompt generation pipeline which automates the creation of these task-specific prompts and is shown in Figure 2. This pipeline eliminates the dependency on human expertise for prompt design and enables scalable adaptation to different benchmarks.

The pipeline operates through an iterative dialogue between two LLMs: a prompt generator and a prompt evaluator. Initially, a randomly selected batch of questions and their answers are sampled from the source benchmark, along with a high-level *goal description*, are provided to the prompt generator LLM. The goal is to produce a prompt that can transform similar input questions into scenario-based questions in Q2S or instruction prompts in Q2I.

The generated prompt is then passed to the evaluator LLM, which assesses it based on predefined *evaluation criteria*. The evaluator assigns a numerical score ranging from 0 to 10 and provides qualitative feedback describing strengths and weaknesses of the prompt.

If the score surpasses a threshold (e.g., 8), the prompt is accepted for use in the corresponding generation task. Otherwise, the prompt generator revises the prompt using the evaluator's feedback, and the process repeats iteratively until a satisfactory prompt is produced. Once a prompt is accepted, it is then reviewed by a human domain expert, who either approves or rejects the generated prompts. In our experiments on the GSM8K and physics question datasets, the produced prompts were approved by the experts. See the generated prompt for Q2S for physics questions in the following.

This automatic prompt optimization is described in Algorithm 1. By leveraging this pipeline, the

DEEPQUESTION framework achieves robust prompt design tailored to the specific domain and style of each benchmark, enabling effective question transformation without manual intervention.

Algorithm 1: DEEPQUESTION Framework Algorithm

Input: QA dataset $B = \{(q_i, a_i)\}$
Output: Deeper Question Benchmark: DEEPQUESTION

$score \leftarrow 0$;
$feedback \leftarrow$ `None`;
`goal_description` $\leftarrow$ Description of the intended purpose of generated prompts;
`evaluation_criteria` $\leftarrow$ Instructions for how the LLM should assess generated prompt quality;
Sample a random subset $B_s \subset B$;
while $score < 8$ **do**
  $prompt \leftarrow$ `LLM`$(goal\_description, B_s, feedback)$;
  $(score, feedback) \leftarrow$ `LLM`$(evaluation\_criteria, prompt)$;
`DeepQuestion` $\leftarrow$ empty list;
foreach $(q, a) \in B$ **do**
  $q' \leftarrow$ `LLM`$(prompt, q, a)$;
  `DeepQuestion.append`(q', a);

Deep-Question Prompt Q2S for Physics Questions

My goal is to design questions based on Bloom's *Apply* level. I will give you a question, and your task is to create a scenario in the form of a narrative that includes redundant material not related to solving the question. Ultimately, the explicit question I provide must be solved to find the answer, but the student must infer this question from the narrative.
The narrative should include a real-world story and irrelevant numbers that are ultimately simplified to the original problem.
* You should never give any hints about which information is necessary and which is irrelevant.
* The story should not reference the formulas or concepts required to solve the question. It should put the solver in a situation where they must apply their knowledge independently. * All formulas and numbers should be presented in LaTeX format. * All your answers must be in Persian.

4. Experiment Setup

We constructed the DEEPQUESTION benchmark by applying our framework to 60 randomly selected GSM8K questions (in English) and 60 physics questions from the Iranian University Entrance Exam (in Persian). While our method aims to minimize reliance on expert intervention, these datasets were chosen due to the familiarity of the authors to math and physics subjects. The generated deep questions were manually reviewed for quality and accuracy, while we evaluated the framework by comparing them to the original questions.

For evaluation, we utilized Gemini-2, GPT-4.1, Llama-3.1 (Grattafiori et al., 2024), Deepseek-V3 (Liu et al., 2024), Deepseek-R1 (Guo et al., 2025), O4-mini, Gemma3 (Team et al., 2025), Phi4 (Abdin et al., 2024), and Qwen3 (Yang et al., 2025). To ensure reproducibility, all models were run with a temperature setting of zero. Furthermore, within the framework itself, we employed Gemini 2.5 Pro in prompt generator LLM, prompt evaluator LLM, and question generator LLM.

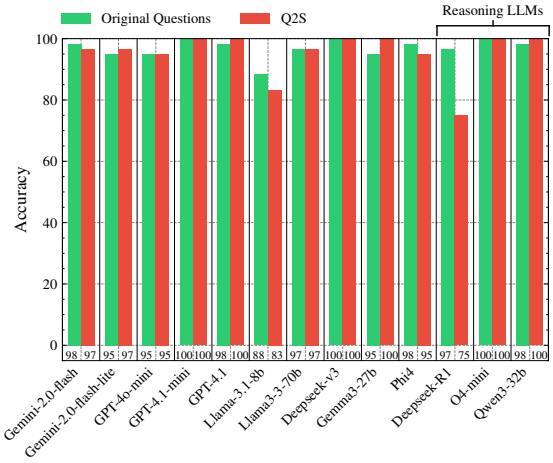
In the Q2S setting, the model generates a solution to a given scenario-question, and correctness is easily verified due to well-defined answers in math and physics problems in our setup.

In the Q2I setting, the model generates questions from instructions. Evaluation consisted of two steps: (1) Answerability Check: verifying whether a strong model (O4-mini) and the model itself can correctly solve the generated questions, indicating the validity of the questions. (2) LLM-as-Judge: since answerability alone does not capture question quality, a powerful language model (O4-mini) was employed for direct qualitative evaluation based on expert-defined criteria: Reasoning Demand — how much genuine thinking and decision-making the question requires; Numerical Quality — whether the numerical values are realistic and meaningful; Physical Realism — whether the scenario is plausible and internally consistent; Clarity and Brevity — how concise and understandable the question is; and Solution Spoiling — whether the problem avoids revealing its own answer or solution steps.

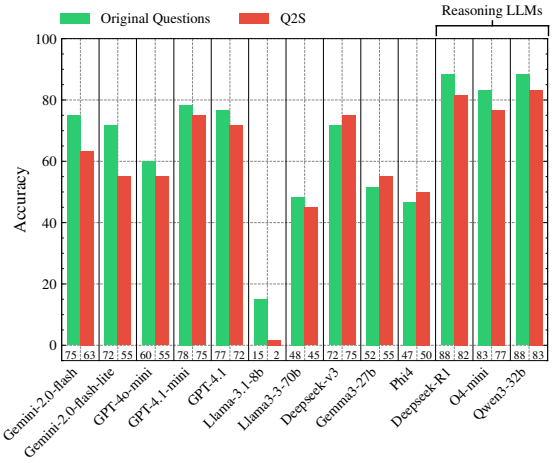
In future work, for open-ended questions that do not have a single correct answer, the same LLM-as-Judge approach could be extended. By defining domain-specific evaluation criteria, LLMs can provide consistent, qualitative assessments that capture nuanced aspects of question quality beyond simple answerability.

5. Results

Based on the distinctions between general-purpose models and reasoning-focused models, we evaluated LLMs from both categories and report our results. Our evaluation methodology is similar to that of Illusion of Thinking, which also compared these two families of models.

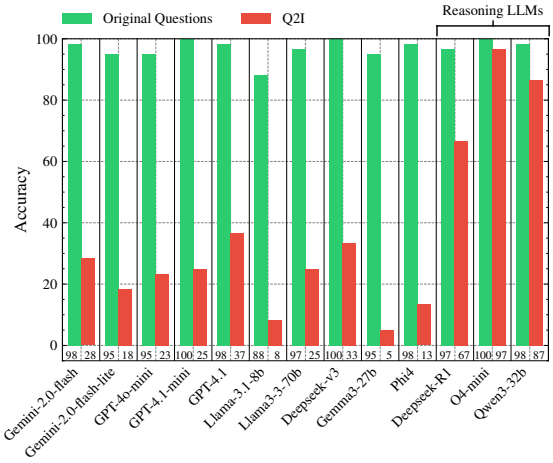


(a) Original vs. Q2S in GSM8K

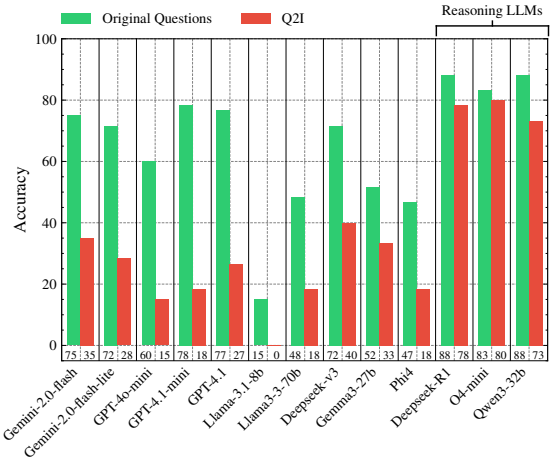


(b) Original vs. Q2S in Physics

Figure 4: Evaluation of different LLMs in original and scenario-based questions

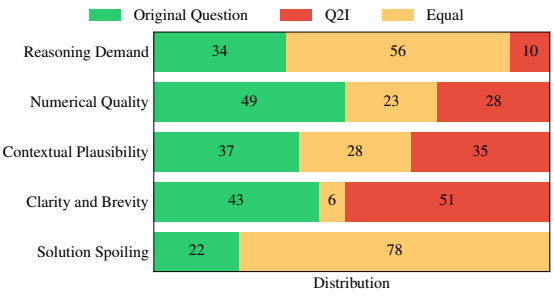


(a) Original vs. Q2I in GSM8K

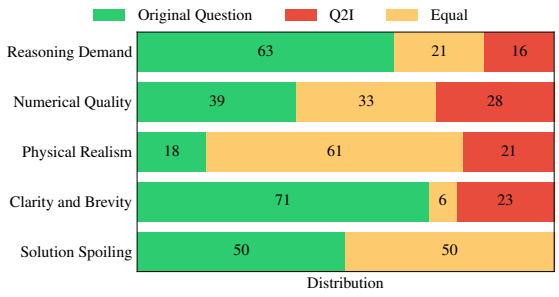


(b) Original vs. Q2I in Physics

Figure 5: Evaluation of different LLMs in original and instruction-based questions



(a) GSM8K

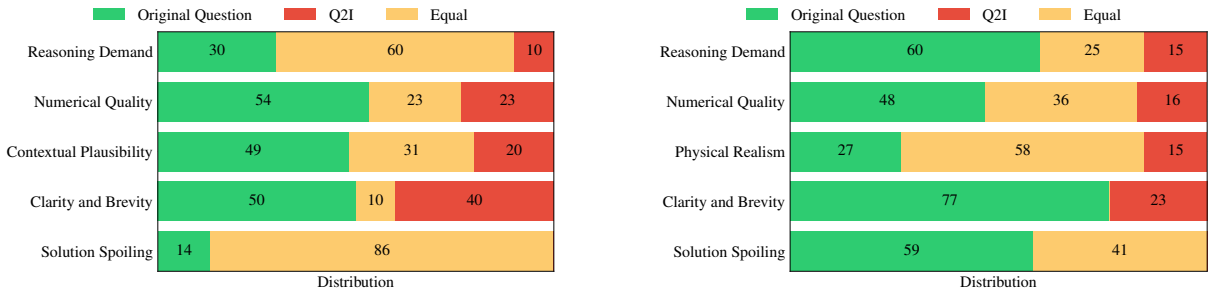


(b) Physics

Figure 6: Win rate of original against Q2I questions for O4-mini

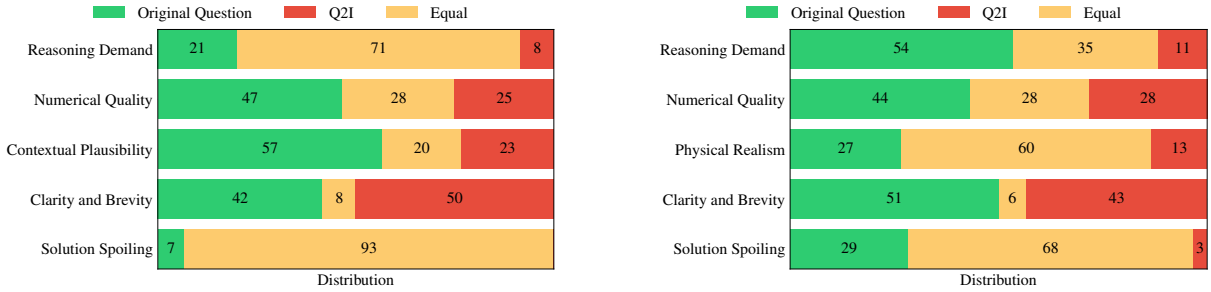
5.1. General-Purpose LLMs

First, we evaluated general-purpose LLMs on question generation using Q2S. As shown in Figure 4, general-purpose models exhibit a moderate decrease in performance on Q2S questions. Some models, such as Gemini-2-Flash and LLaMA-3.1-8B, demonstrate a noticeable reduction in accuracy, whereas others, including Gemma-3-27B and DeepSeek-v3, achieve comparable or even improved performance on Q2S questions. These findings indicate that general-purpose LLMs can solve scenario-based questions with only a slight



(a) GSM8K (b) Physics

Figure 7: Win rate of original against Q2I questions for Deepseek-R1



(a) GSM8K (b) Physics

Figure 8: Win rate of original against Q2I questions for Qwen3-32b

decrease in accuracy, suggesting their capability to apply knowledge at the applying level of Bloom's taxonomy.

To further assess LLMs' abilities at higher levels, such as the creation level, we were motivated to design the Q2I experiment. In this setting, models exhibited a substantial drop in accuracy when tackling tasks that require deeper understanding and more complex reasoning. No model surpassed 38% accuracy on instruction-based question generation for GSM8K or physics (see Figure 5), despite over 95% on original questions. These results support the use of Bloom's taxonomy as an evaluation framework and reveal the limitations of general-purpose LLMs on higher-order cognitive tasks.

5.2. Reasoning LLMs

Based on the observed results from general-purpose models and their performance decline on Q2I questions, we evaluated reasoning models on both Q2I and Q2S questions. Reasoning models exhibited a performance trend similar to that of general-purpose models on Q2S questions. However, for Q2I questions, the performance reduction in reasoning models was smaller. Despite this, accuracy in the Q2I setting still decreased by 8–30% (Figure 5), highlighting limitations of these models at higher cognitive levels. For example, Deepseek-R1's performance declined by 30% on GSM8K and 10% on physics datasets, whereas Qwen-3-32B experienced an 11% drop on both datasets. Although reasoning models generally outperform general-purpose models, significant gaps remain in achieving deep and comprehensive understanding. These findings indicate that neither general-purpose nor reasoning models achieve strong performance at the creation level of Bloom's taxonomy, suggesting that current LLMs are still distant from high-level cognitive reasoning.

5.3. Question Quality Analysis

While reasoning models can generate questions that follow instructions and maintain topical consistency, it is crucial to assess their quality beyond surface-level compliance. To this end, we use a preference-based evaluation with O4-mini as a judge, directly comparing generated and original questions across five criteria: reasoning demand, numerical quality, physical realism, clarity and brevity, and solution spoiling. This approach addresses concerns about the model's ability to assess quality by focusing on direct comparison.

As illustrated in Figures 6 to 8, the physics questions generated by the models continue to underperform compared to the original questions. Specifically, in the dimensions of Reasoning Demand, Clarity, and Brevity, the original questions outperform the generated ones in more than 50% of cases across all three models. Furthermore, in terms of Solution Spoiling, Q2I exhibits negligible improve-

Table 1: Model performance across translation tasks.

(a) Physics – Translate to English

Models	Llama3-3-70b	Llama3-3-70b-Trans	GPT-4.1	GPT-4.1-Trans	O4-mini	O4-mini-Trans
Original Questions	48.33%	61.67%	76.67%	78.33%	83.33%	78.33%
Q2S	45.00%	50.00%	71.67%	75.00%	76.67%	80.00%
Q2I	18.33%	23.33%	26.67%	33.33%	80.00%	73.33%

(b) GSM8K – Translate to Persian

Models	Llama3-3-70b	Llama3-3-70b-Trans	GPT-4.1	GPT-4.1-Trans	O4-mini	O4-mini-Trans
Original Questions	96.67%	88.33%	98.33%	96.67%	100.00%	100.00%
Q2S	96.67%	75.00%	100.00%	95.00%	100.00%	98.33%
Q2I	25.00%	11.67%	36.67%	31.67%	96.67%	95.00%

ment, achieving wins in only approximately 3% of cases. These findings underscore the persistent challenges faced by reasoning models in producing high-quality questions. In contrast, generated GSM8K questions are more comparable to the originals, likely reflecting the relative simplicity of this dataset. Overall, the results demonstrate the effectiveness of our framework in enhancing question depth without introducing artificial difficulty. Nevertheless, a substantial quality gap remains, emphasizing the ongoing difficulty of generating high-quality questions even with advanced reasoning models.

5.4. Disentangling Linguistic Effects from Cognitive Complexity

To rule out linguistic artifacts, we conducted a cross-lingual experiment by translating the Persian Physics dataset into English (see Table 1a) and the English GSM8K dataset into Persian (see Table 1b). We evaluated three representative models, including Llama3-70B, GPT-4.1, and O4-mini, to test whether performance declines in DeepQuestion stem from increased cognitive demand or from language-related confounds.

Across both datasets and translation directions, the performance hierarchy (Original > Q2S > Q2I) remained consistent. For instance, in the translated Physics dataset, O4-mini scored 83.33% (Original), 80.00% (Q2S), and 73.33% (Q2I); GPT-4.1 followed the same pattern (78.33% → 75.00% → 33.33%). Similarly, in Persian-translated GSM8K, Llama3-70B dropped from 88.33% to 75.00% and then to 11.67%. This stability confirms that our transformations genuinely elevate cognitive complexity independent of language.

These results confirm that DeepQuestion isolates cognitive complexity as defined by Bloom's Taxonomy. Observed declines arise from reasoning demands rather than language artifacts, validating our framework's robustness across languages and model families.

6. Conclusion

This work presented DeepQuestion, a systematic and scalable framework for generating cognitively diverse benchmarks that extend existing datasets through the lens of Bloom's taxonomy. By transforming conventional question–answer pairs into scenario-based and instruction-driven tasks, DeepQuestion provides a structured means to probe LLMs across multiple levels of cognition, from application to creation.

Our findings across state-of-the-art LLMs reveal a persistent decline in performance as task complexity and cognitive depth increase—reaching up to 70% accuracy loss in higher-order reasoning tasks. These results highlight that current LLMs, despite their success on standardized benchmarks, still exhibit shallow generalization and limited conceptual transfer when confronted with real-world or creative problem-solving scenarios.

Beyond performance assessment, DeepQuestion offers a replicable pathway for future benchmark development. The framework can be readily applied to other domains such as law, medicine, or engineering to examine domain-specific reasoning. Moreover, it opens new directions for exploring automated benchmark construction, self-improving evaluation pipelines, and curriculum-style training data aligned with cognitive learning theories.

In summary, DeepQuestion bridges educational psychology and AI evaluation, revealing fundamental gaps in LLM reasoning and offering tools to measure and close them. As language models continue to evolve, cognitively grounded and contextually rich evaluation frameworks like DeepQuestion will be essential for steering their progress toward genuine understanding and human-aligned intelligence.

7. Bibliographical References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ahmed Alaa, Thomas Hartvigsen, Niloufar Golchini, Shiladitya Dutta, Frances Dean, Inioluwa Deborah Raji, and Travis Zack. 2025. Medical large language model benchmarks should prioritize construct validity. *arXiv preprint arXiv:2503.10694*.

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.

Anna Arias-Duart, Pablo Agustin Martin-Torres, Daniel Hinjos, Pablo Bernabeu-Perez, Lucia Urcelay Ganzabal, Marta Gonzalez Mallo, Ashwin Kumar Gururajan, Enrique Lopez-Cuena, Sergio Alvarez-Napagao, and Dario Garcia-Gasulla. 2025. Automatic evaluation of healthcare llms beyond question-answering. *arXiv preprint arXiv:2502.06666*.

John B Biggs and Kevin F Collis. 2014. *Evaluating the quality of learning: The SOLO taxonomy (Structure of the Observed Learning Outcome)*. Academic press.

Benjamin S Bloom. 1971. *Taxonomy of educational objectives: The classification of educational goals: By a committee of college and university examiners*. David McKay.

Archana Chauhan, Farah Khaliq, and Kirtana Raghurama Nayak. 2025. Assessing quality of scenario-based multiple-choice questions in physiology: Faculty-generated vs. chatgpt-generated questions among phase i medical students. *International Journal of Artificial Intelligence in Education*.

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

Mary Forehand. 2010. Bloom's taxonomy. *Emerging perspectives on learning, teaching, and technology*, 41(4).

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 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. 2021a. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the math dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Mustafa İlhan, Melehat Gezer, et al. 2017. A comparison of the reliability of the solo-and revised bloom's taxonomy-based classifications in the analysis of the cognitive levels of assessment questions. *Pegem Egitim ve Ogretim Dergisi= Pegem Journal of Education and Instruction*, 7(4):637.

CF Leung. 2000. Assessment for learning: Using solo taxonomy to measure design performance of design & technology students. *International Journal of Technology and Design Education*, 10(2).

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Seyed Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. 2025. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. In *The Thirteenth International Conference on Learning Representations*.

Aidar Myrzakhan, Sondos Mahmoud Bsharat, and Zhiqiang Shen. 2024. Open-llm-leaderboard: From multi-choice to open-style questions for llms evaluation, benchmark, and arena. *arXiv preprint arXiv:2406.07545*.

Ivo Petrov, Jasper Dekoninck, Lyuben Baltadzhiev, Maria Drencheva, Kristian Minchev, Mislav Balunović, Nikola Jovanović, and Martin Vechev. 2025. Proof or bluff? evaluating llms on 2025 USA math olympiad. *arXiv preprint arXiv:2503.21934*.

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*.

Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. 2025. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. *arXiv preprint arXiv:2506.06941*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. 2024. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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