Dissecting Physics Reasoning in Small Language Models: A Multi-Dimensional Analysis from an Educational Perspective

Anonymous ACL submission

Abstract

Small Language Models (SLMs) offer computational efficiency and accessibility, making them promising for educational applications. However, their capacity for complex reasoning, particularly in domains such as physics, remains underexplored. This study investigates the high school physics reasoning capabilities of state-of-the-art SLMs (under 4 billion parameters), including instruct versions of Llama 3.2, Phi 4 Mini, Gemma 3, and Qwen series. We developed a comprehensive physics dataset from the OpenStax High School Physics textbook, annotated according to Bloom's Taxonomy, with LaTeX and plaintext mathematical notations. A novel cultural contextualization approach was applied to a subset, creating culturally adapted problems for Asian, African, and South American/Australian contexts while preserving core physics principles. Using an LLM-as-a-judge framework with Google's Gemini 2.5 Flash, we evaluated answer and reasoning chain correctness, along with calculation accuracy. The results reveal significant differences between the SLMs. Qwen 3 1.7B achieved high 'answer accuracy' ($\approx$ 85%), but 'fully correct reasoning' was substantially low ($\approx$38%). The format of the mathematical notation had a negligible impact on performance. SLMs exhibited varied performance across the physics topics and showed a decline in reasoning quality with increasing cognitive and knowledge complexity. In particular, the consistency of reasoning was largely maintained in diverse cultural contexts, especially by better performing models. These findings indicate that, while SLMs can often find correct answers, their underlying reasoning is frequently flawed, suggesting an over-reliance on pattern recognition. For SLMs to become reliable educational tools in physics, future development must prioritize enhancing genuine understanding and the generation of sound, verifiable reasoning chains over mere answer accuracy.

1 Introduction

Small Language Models (SLMs) are neural language models distinguished by their smaller parameter counts and greater computational efficiency relative to Large Language Models (LLMs). This compact design allows SLMs to operate effectively on common consumer hardware without requiring specialized high-performance infrastructure that is typically essential for LLMs. Notable examples of SLMs, which generally range from several hundred million to a few billion parameters, include the Phi series (Gunasekar et al., 2023; Li et al., 2023; Abdin et al., 2024b), Gemma (Team et al., 2024, 2025), Pythia (Biderman et al., 2023), Llama (Grattafiori et al., 2024), Qwen (Qwen et al., 2025; Yang et al., 2025), and TinyLlama (Zhang et al., 2024).

The capacity for reasoning in language models has traditionally been associated with substantial scale, often emerging in models with hundreds of billions of parameters (Team et al., 2023; Hurst et al., 2024). Initial research indeed suggested that complex, multi-step reasoning was primarily a feature of these LLMs. However, this scale-centric view is increasingly being contested by newer findings (Yang et al., 2025; Srivastava et al., 2025).

The physics domain is characterized by a wide array of sub-disciplines, each with distinct conceptual frameworks and problem-solving paradigms. Effective engagement with physics requires a broad spectrum of cognitive skills, from foundational recall of laws and definitions to the application of principles, analysis of complex systems, evaluation of evidence, and even creative problem formulation, aligning with the hierarchical levels of Bloom's Taxonomy (Krathwohl, 2002). This process is further complicated by the mathematical problem solving required for physics proficiency, adding another layer of complexity for language models (Xu et al., 2025). A common limitation in evaluating reasoning for such complex tasks is an over-reliance on the correctness of the final answer, often neglecting the reasoning steps or the chain of thought that led to the solution (Srivastava et al., 2025). For this study, we define physics reasoning as the ability to work effectively with physics knowledge and problems by recalling facts and basic principles, understanding concepts, applying physical laws through single-step or multi-step processes (both conceptual and mathematical), analyzing scenarios, explaining phenomena, solving problems, and reaching well-supported conclusions about physical systems.

SLMs enable efficient on-device processing without requiring internet connectivity, significantly improving data privacy through local computation (Sun et al., 2020; Abdin et al., 2024b). These attributes are particularly advantageous within educational frameworks where SLMs can promote more equitable access to AI-driven learning resources, a significant benefit in contexts with limited internet stability or financial constraints (Wei et al., 2025; Schick and Schütze, 2021). The confidentiality of student data, a critical factor in digital learning environments, is significantly improved through local processing, avoiding the privacy risks associated with the API-based LLM service (Das et al., 2025). Beyond these operational benefits, effective application of SLMs in education also requires appropriate pedagogical approaches, such as contextualizing learning materials to specific cultural or regional settings to enhance student engagement and comprehension (Cordova and Lepper, 1996).

However, the performance of SLM reasoning across multiple dimensions, such as evaluating the reasoning chain, navigating the various cognitive demands and types of knowledge of physics, and adapting culturally contextualized educational content for diverse regions, remains largely underexplored, highlighting a significant gap in current research.

To systematically evaluate these reasoning capabilities in SLMs, our approach included selecting state-of-the-art models and developing a specialized physics dataset from the OpenStax High School Physics textbook (Urone and Hinrichs, 2020). This dataset is annotated according to Bloom's Taxonomy to categorize questions across cognitive and knowledge dimensions, spanning multiple physics topics. To assess the impact of the representation format, we maintain parallel versions with both the LaTeX notation and plain text equivalents. Furthermore, we develop a novel cultural contextualization approach, systematically adapting a substantial subset of problems to incorporate authentic elements from underrepresented regions in Asia, Africa, and South America while preserving the underlying physics principles. The evaluation framework assesses the correctness of both the answers and the reasoning chains.

Based on the identified research gaps, our study addresses the following research questions: (1) How effectively can SLMs perform high school physics reasoning? (2) Does mathematical symbol representation alter the quality of physics reasoning? (3) Do SLMs exhibit consistent performance across different physics topics? (4) How does cognitive and knowledge complexity influence physics reasoning in SLMs? (5) Can SLMs maintain consistent physics reasoning chains across different cultural contexts?

2 Related Work

Recent advances in training methodologies, such as specialized fine-tuning of SLMs using reasoning-intensive datasets (Li et al., 2023; Gunasekar et al., 2023), knowledge distillation from larger models (Guo et al., 2025), and targeted post-training compression techniques (Egashira et al., 2024) are enhancing the ability of SLMs. However, the extent to which these improvements translate across diverse domains and reasoning types requires systematic investigation, particularly in complex fields like physics, where multi-step problem solving and conceptual understanding are essential (Polverini and Gregorcic, 2024; Kahaleh and Lopez, 2025).

Evaluating reasoning capabilities presents significant challenges. While rule-based parsing offers precision, it struggles with format variations. Human evaluation remains the gold standard but faces scalability constraints. LLM-as-a-Judge frameworks have emerged as effective alternatives, with studies showing a strong correlation between LLM judgments and human assessments (Thakur et al., 2024; Chiang and Lee, 2023; Gu et al., 2024; Chang et al., 2024). Though promising, systematic comparisons of evaluation methodologies specifically for assessing SLM reasoning capabilities remain limited.

Recent work by (Karim et al., 2025) demonstrates that LLM reasoning processes can be affected by contextual changes. While evaluation frameworks like WorldView-Bench (Mushtaq et al., 2025) exist for assessing cultural perspectives in larger models, the specific impacts of contextualization on SLM reasoning pathways and error patterns represent an underexplored research area requiring dedicated investigation.

3 Methodology

To answer the research questions, we require a suite of small language models, a dataset with physics questions, a contextualization framework to include culturally relevant components, inference with SLMs and evaluation of generated responses. In what follows, we describe each part of our methodology.

3.1 Model selection

For the purposes of the study, SLMs are defined as models with fewer than 4 billion (4B) parameters. The chosen models include instruct versions of Llama 3.2 (1B and 3B) (Grattafiori et al., 2024), Phi 4 Mini (3.84B) (Abdin et al., 2024a) and its reasoning-focused variant (3.84B) (Abdin et al., 2025), Gemma 3 (1B and 4B) (Team et al., 2025), Qwen 2.5 Instruct (1.5B) (Qwen et al., 2025), Qwen 2.5 DeepSeek Distil (1.5B) (Guo et al., 2025), and Qwen 3 Instruct (0.6B and 1.7B) (Yang et al., 2025). This selection represents the current state-of-the-art SLMs, with each model chosen for its specific architectural innovations, training methodologies, or performance characteristics.

The inclusion of multiple model sizes from the same families (Llama 3.2, Gemma 3, and Qwen 3) enables analysis of how parameter count affects physics reasoning capabilities within the same architecture choices of SLMs. Similarly, the comparison between Phi 4 Mini and its reasoning variant provides insight into how specialized training for reasoning tasks affects performance on physics problems.

3.2 Dataset creation

A dataset of physics questions was developed on the basis of the end-of-chapter exercises in the OpenStax High School Physics Textbook (Urone and Hinrichs, 2020). This textbook encompasses 23 chapters covering diverse physics domains, including Introduction, Mechanics, Electricity and Magnetism, Waves and Acoustics, Thermodynamics, Optics, and Modern Physics. The extraction process resulted in a set of physics questions, encompassing various question types such as conceptual items, critical thinking challenges, short answer questions, true/false statements, extended response tasks, and numerical problems. This diversity ensured a broad assessment across physics topics and cognitive demands. The challenges and specific processes involved in dataset generation are detailed in the Appendix A.1.

We converted mathematical equations from images to LaTeX format using OCR tools, followed by a rigorous data cleaning process to address inconsistencies in the source material. This iterative refinement produced a final curated dataset ($\mathcal{D}_{\text{openstax}}$) containing 1,306 questions. We also created a plain text version ($\mathcal{D}_{\text{plaintext}}$) by systematically converting all LaTeX expressions to standard text notation.

We annotated each physics problem according to the knowledge and cognitive dimensions of Bloom's Taxonomy to enable a systematic analysis of the cognitive skills these tasks require from SLMs. This annotation scheme was essential for systematically evaluating how SLMs handle increasingly complex reasoning tasks, particularly as they progress from lower-order thinking skills (Remember, Understand) to higher-order cognitive processes (Apply, Analyze, Evaluate, Create). Similarly, distinguishing between knowledge types (Factual, Conceptual, Procedural, and Metacognitive) allowed us to identify specific strengths and limitations in how these models process and manipulate different forms of physics knowledge. The details and composition of the dataset is given in Appendix A.2.

3.3 Contextualization of dataset

From the comprehensive dataset, we selected a subset of 393 questions to investigate how contextualization affects SLM performance. By incorporating diverse cultural and geographical elements into standard physics problems, we could evaluate whether these models exhibit consistent reasoning abilities across differently contextualized versions of identical physics concepts.

To ensure geographical and cultural diversity, we developed a cultural context database drawing from countries selected using the United Nations Geoscheme, focusing on underrepresented regions in Asia, Africa, South America, and Australia, combining South America and Australia into a single regional dataset. This was an intentional choice due to the smaller number of countries within these continents individually. For each region, we compiled authentic cultural elements including common names, festivals, landmarks, foods, transportation, sports, and traditions. Using Google's Gemini models with integrated search capabilities, we generated and verified this cultural information.

For each original question, we created five distinct contextualized variations that maintained the original physics principles while incorporating cultural elements. This approach produced three culturally adapted datasets ($\mathcal{D}_{\text{Asia}}$, $\mathcal{D}_{\text{Africa}}$, and $\mathcal{D}_{\text{SA\_AU}}$), each containing 1,965 questions. Details on the contextualization process, including cultural database creation, verification procedures, and the adaptation methodology, are provided in Appendix B. An example of a contextual question is given in Figure 3.

3.4 Model inference

We conducted inference using several models mentioned in Section 3.1 to address the research questions given in Section 1. For multiple choice questions, we supplied the question text and all options in the prompt, requiring the models to generate the selected option, explanation, and supporting reasoning. For open-ended questions, the models generated both answers and detailed reasoning without predefined options.

To perform a systematic comparison, six distinct evaluation sets were used as follows. (1) the entire original dataset with LaTeX notation ($\mathcal{D}_{\text{openstax}}$) consisting of 1,306 physics questions as our primary baseline; (2) a plain text version of the entire dataset ($\mathcal{D}_{\text{plaintext}}$); (3) the subset of original questions selected for contextualization ($\mathcal{D}_{\text{contextual}}$) comprising 393 questions; and three sets of culturally adapted versions including (4) questions adapted with Asian cultural elements ($\mathcal{D}_{\text{Asia}}$); (5) African cultural elements ($\mathcal{D}_{\text{Africa}}$); and (6) South American and Australian cultural elements ($\mathcal{D}_{\text{SA\_AU}}$). This structured approach enables us to systematically evaluate: (1) baseline reasoning capabilities (using dataset 1, $\mathcal{D}_{\text{openstax}}$); (2) mathematical notation effects (by comparing datasets 1 and 2); (3) performance across cognitive and knowledge dimensions (through the Bloom's Taxonomy annotations applied to dataset 1); (4) variations across physics topics (using the topic categorizations within dataset 1); and (5) the impact of cultural adaptation across different regions (by comparing dataset 3 with datasets 4, 5, and 6).

3.5 Model evaluation

In this study, the LLM-as-a-judge model evaluation approach with Google's Gemini 2.5 Flash was utilized. Multiple evaluation models were initially tested, and Gemini was ultimately selected on the basis of its superior balance of cost-effectiveness and evaluation accuracy.

The evaluation strategy implemented three specialized assessment prompts tailored to different question formats: (1) Multiple choice: for responses where answer selection and reasoning were clearly delineated, this prompt compared the model's selected option and reasoning directly with the ground truth; (2) Multiple choice unstructured response: for free-form responses to multiple-choice questions, this prompt first extracted the selected option and reasoning from the generated text before performing comparative assessment; and (3) Open ended: for questions without predefined answers, this prompt assessed whether the response adequately addressed the required physics concepts while allowing for valid alternative approaches.

All evaluation prompts assessed responses across three dimensions: answer correctness (binary classification of correct/incorrect), reasoning quality (categorized as fully correct, partially correct, or incorrect), and calculation accuracy (not required, correct, or incorrect when calculations were present). Reasoning quality distinguished between responses with complete understanding of physics (fully correct), those with partly correct reasoning (partially correct), and those containing fundamental misunderstandings (incorrect).

To provide a more nuanced evaluation of reasoning capabilities, we implemented a weighted reasoning accuracy measure. This metric assigned different weights to each level of reasoning quality: 2 points for fully correct reasoning, 1 point for partially correct reasoning, and 0 points for incorrect reasoning. The weighted reasoning accuracy was calculated as

$$\text{WRA} = \frac{\sum_{i=1}^{n} w_i}{2n} \times 100\%,$$

where w_i represents the reasoning score (0, 1, or 2) for the i-th question, and n is the total number of questions.

The reliability of this automated approach was verified by a manual review of randomly selected samples across different types of questions and physics topics for all SLMs. We examined approximately 185 randomly selected questions covering various question formats and topic areas to verify the quality and consistency of the automated evaluations. After benchmarking several LLMs for their evaluation capabilities, Gemini 2.5 Flash emerged as the best choice, demonstrating the most reliable evaluation performance while maintaining computational cost.

This methodology enabled a quantitative comparison of model performance in different notation formats, cognitive and knowledge dimensions, physics topics, and cultural contexts. Detailed evaluation prompts are provided in Figure 4, 5 and 6.

4 Results and Analysis

Our findings address all research questions, revealing patterns in SLMs' physics reasoning capabilities across notation formats, topics, knowledge, and cognitive demands, and cultural contexts. Table 1 summarizes the evaluation metrics across datasets.

4.1 How effectively can SLMs perform high school physics reasoning?

The data in Table 1 reveals a striking gap between answer and reasoning correctness for SLMs. While Qwen 3 1.7B achieves the highest answer accuracy at 84.68%, its fully correct reasoning accuracy is only 38.25%, demonstrating that even the best-performing model produces flawless reasoning chains in fewer than two fifths of cases. This substantial discrepancy between getting the right answer and showing entirely correct reasoning is consistent across all SLMs.

The Phi 4 Reasoning 3.8B significantly outperforms the standard Phi 4 3.8B in both answer accuracy (77.18% compared to 50.61%) and fully correct reasoning (30.47% compared to 11.22%), highlighting the impact of reasoning-focused training. Smaller models like Gemma 3 1B and Llama 3.2 1B demonstrate particularly low fully correct reasoning rates of 9.65% and 10.41% respectively, despite achieving answer accuracies above 40%.

The difference between weighted reasoning accuracy (which considers both fully and partially correct reasoning) and fully correct reasoning further reveals issues with the reasoning chains generated by the SLMs. For instance, Qwen 3 1.7B shows 84.99% weighted reasoning but only 38.25% fully correct reasoning, indicating that in many cases, models reach correct answers despite reasoning that contains errors or misconceptions.

Calculation accuracy shows considerable variation across models, with Qwen 3 1.7B (87.85%) and Qwen 2.5 Distil 1.5B (83.67%) demonstrating particularly strong mathematical capabilities. This suggests that some models can execute calculations correctly for the reasoning chain generated.

Our analysis revealed that in multiple choice questions, SLMs often select options closest to their derived answers, which partially explains the higher answer correctness relative to reasoning quality. Additionally, these models are typically optimized during training to produce correct answers rather than flawless reasoning chains, potentially leading them to leverage pattern recognition to select correct options even without complete physical understanding. These factors collectively contribute to the significant gap observed between the models' ability to select correct answers and their capacity to produce sound reasoning chains.

4.2 Does mathematical symbol representation alter the quality of physics reasoning?

The comparison between performance in $\mathcal{D}_{\text{openstax}}$ and $\mathcal{D}_{\text{plaintext}}$ (Table 1) reveals that the format of representation of mathematical symbols has a negligible effect on the quality of physics reasoning in all the SLM tested.

For the 'answer accuracy' metric, most models show only slight variations between $\mathcal{D}_{\text{openstax}}$ and $\mathcal{D}_{\text{plaintext}}$. Qwen 3 1.7B achieves 84.68% with $\mathcal{D}_{\text{openstax}}$ and 86.13% with $\mathcal{D}_{\text{plaintext}}$, while Phi 4 3.8B shows 50.61% with $\mathcal{D}_{\text{openstax}}$ and 49.12% with $\mathcal{D}_{\text{plaintext}}$.

Similarly, for the 'fully correct reasoning accuracy' metric, there are only marginal differences between notation formats. Qwen 3 1.7B achieves 38.25% with $\mathcal{D}_{\text{openstax}}$ and 39.35% with plain text, while Llama 3.2 3B shows 24.85% with $\mathcal{D}_{\text{openstax}}$ and 23.83% with $\mathcal{D}_{\text{plaintext}}$.

The 'weighted reasoning accuracy' metric also demonstrates this pattern, with most models showing differences of roughly ±2% between the two datasets. For the 'calculation accuracy' metric, some models show slightly better performance with $\mathcal{D}_{\text{plaintext}}$ (Phi 4 3.8B improves from 48.33% to 53.12%), while others perform marginally better with $\mathcal{D}_{\text{openstax}}$ notation (Gemma 3 4B decreases from 64.72% to 62.54%), but the differences remain relatively small.

The comparable performance across notation formats indicates that modern SLMs effectively process both LaTeX and plain text mathematical representations in physics problems. This indicates that mathematical symbol representation has only a minor influence on the quality of physics reasoning in these models.

4.3 Do SLMs exhibit consistent performance across different physics topics?

Our analysis reveals notable variations in SLM performance across different physics topics, as illustrated in the heat map (Figure 7 in the Appendix). Most models demonstrate stronger performance on Introduction and Thermodynamics topics while struggling relatively more with Optics and Modern Physics.

Table 1: Performance metrics of SLMs across different datasets.

Model	$\mathcal{D}_{\text{openstax}}$	$\mathcal{D}_{\text{plaintext}}$	$\mathcal{D}_{\text{contextual}}$	$\mathcal{D}_{\text{Asia}}$	$\mathcal{D}_{\text{Africa}}$	$\mathcal{D}_{\text{SA\_AU}}$
Answer Accuracy (%)						
Qwen 3 0.6B	66.85	68.81	60.05	68.45	67.91	66.80
Gemma 3 1B	43.79	45.21	33.33	30.68	29.38	29.50
Llama 3.2 1B	45.71	46.59	33.84	34.66	34.98	34.36
Qwen 2.5 1.5B	66.76	64.52	55.73	56.71	58.96	56.66
Qwen 2.5 Distil 1.5B	69.61	69.50	73.79	67.81	67.31	67.66
Qwen 3 1.7B	**84.68**	**86.13**	**87.53**	**87.59**	**86.61**	**86.72**
Llama 3.2 3B	65.69	65.82	56.74	56.16	56.11	57.43
Phi 4 3.8B	50.61	49.12	46.56	47.12	45.93	44.86
Phi 4 Reasoning 3.8B	77.18	79.16	75.32	71.54	71.89	71.92
Gemma 3 4B	71.67	70.11	70.74	70.23	69.40	68.32
Fully Correct Reasoning Accuracy (%)						
Qwen 3 0.6B	25.46	26.36	25.19	25.35	24.72	24.71
Gemma 3 1B	9.65	10.42	5.98	5.20	4.53	5.09
Llama 3.2 1B	10.41	11.23	7.25	8.50	8.25	8.92
Qwen 2.5 1.5B	22.82	22.11	17.68	17.89	18.89	18.68
Qwen 2.5 Distil 1.5B	23.39	23.64	23.66	24.65	24.64	24.48
Qwen 3 1.7B	**38.25**	**39.35**	**39.44**	**38.93**	**38.72**	**38.95**
Llama 3.2 3B	24.85	23.83	18.83	20.53	19.96	20.58
Phi 4 3.8B	11.22	12.15	12.60	12.64	12.55	12.37
Phi 4 Reasoning 3.8B	30.47	30.92	28.88	27.22	26.83	26.84
Gemma 3 4B	27.99	27.20	25.45	24.95	23.73	23.90
Weighted Reasoning Accuracy (%)						
Qwen 3 0.6B	65.92	67.51	67.94	68.01	67.54	66.78
Gemma 3 1B	37.17	38.28	25.83	23.99	22.58	24.51
Llama 3.2 1B	38.51	40.96	31.30	33.35	32.15	33.50
Qwen 2.5 1.5B	61.95	60.99	53.05	54.59	56.64	55.63
Qwen 2.5 Distil 1.5B	67.31	67.70	68.70	68.31	68.71	68.47
Qwen 3 1.7B	**84.99**	**86.48**	**87.91**	**87.01**	**86.84**	**87.10**
Llama 3.2 3B	64.58	63.03	56.74	58.48	58.15	58.16
Phi 4 3.8B	45.94	46.82	46.95	47.28	46.92	46.81
Phi 4 Reasoning 3.8B	78.59	79.39	76.59	74.24	74.31	74.25
Gemma 3 4B	70.63	69.73	69.47	68.29	67.31	67.13
Calculation Accuracy (%)						
Qwen 3 0.6B	69.92	71.15	70.07	72.35	72.95	70.21
Gemma 3 1B	19.09	20.00	22.11	20.05	18.54	20.58
Llama 3.2 1B	23.25	24.71	22.26	24.50	23.84	23.55
Qwen 2.5 1.5B	52.88	52.42	47.64	51.48	52.49	51.48
Qwen 2.5 Distil 1.5B	83.67	82.85	83.89	78.25	78.65	78.15
Qwen 3 1.7B	**87.85**	**88.18**	**88.40**	**89.17**	**89.14**	**88.47**
Llama 3.2 3B	50.43	48.54	48.67	51.21	51.02	53.16
Phi 4 3.8B	48.33	53.12	45.79	48.32	49.32	49.08
Phi 4 Reasoning 3.8B	78.36	80.72	78.31	78.43	76.83	77.27
Gemma 3 4B	64.72	62.54	68.51	65.81	63.76	65.12

For fully correct reasoning, the variation between topics is substantial. Models generally produce better reasoning chains for foundational topics compared to those requiring advanced mathematical formalism or abstract conceptualization. This variation in reasoning performance is more pronounced than differences in answer accuracy, suggesting that SLMs may be leveraging pattern recognition rather than physics understanding for more challenging topics.

Performance inconsistency is most evident in smaller models. Gemma 3 1B and Llama 3.2 1B have dramatic reasoning performance drops for complex topics. In contrast, larger models and those with specialized training demonstrate some robustness. Qwen 3 1.7B and Phi 4 Reasoning 3.8B maintain more consistent reasoning capabilities across different physics domains, though they still show a decline in fully correct reasoning for more abstract topics.

These topic-dependent variations likely stem from differences in conceptual complexity, mathematical demands, and the representation of training data. These findings indicate that current SLMs have not yet achieved physics reasoning capabilities across different topics.

4.4 How do cognitive and knowledge complexity influence the physics reasoning in SLMs?

The complexity of cognitive and knowledge significantly affects the performance of SLM physics reasoning (Figures 8 and 9 in the appendix). Across all models, there is a clear performance gradient along both complexity dimensions, with capabilities declining as tasks become more cognitively demanding or require more sophisticated knowledge application.

In the cognitive dimension, SLMs show strong performance on lower-order thinking skills (Remember, Understand) but struggle progressively with higher-order cognitive processes (Apply, Analyze, Evaluate, Create). The fully correct reasoning metric reveals this pattern most prominently, with even the best-performing model (Qwen 3 1.7B) showing substantial degradation from Remember (88.46%) to Create (35.71%) tasks. This decline indicates that current SLMs have not yet developed robust capabilities for complex reasoning processes that require evaluation and creation.

The results of the knowledge dimension reveal that factual knowledge is handled most effectively across models, while procedural knowledge presents the greatest challenge. This pattern is evident in both reasoning and calculation metrics, suggesting that SLMs struggle most with knowledge requiring systematic application of procedures to solve problems or reach conclusions. Models with specialized training (Phi 4 Reasoning 3.8B) show relatively better performance on procedural knowledge, indicating that targeted training can partially address these limitations.

Our findings suggest that performance degradation is most pronounced at the higher levels of Bloom's Taxonomy, likely due to the compounding nature of errors that becomes especially problematic for achieving fully correct reasoning. These advanced tasks require coherent multi-stage reasoning and accurate execution of procedures at each step, creating a cascade effect where even minor inaccuracies in intermediate steps make it increasingly difficult to maintain a coherent reasoning chain throughout the entire process.

4.5 Can SLMs maintain consistent physics reasoning chains across different cultural contexts?

SLMs demonstrate varying degrees of consistency in physics reasoning when identical principles are presented within different cultural frameworks (Table 1). The data reveal important patterns in how contextual variations affect reasoning chains.

Larger models and those with specialized training maintain remarkably stable reasoning in all contexts. Qwen 3 1.7B shows almost identical fully correct reasoning accuracy in all cultural adaptations (38.93-38.95% for Asian, African, and South American/Australian contexts compared to 39.44% for the baseline contextual dataset). Similarly, Qwen 3 0.6B and Qwen 2.5 Distil 1.5B exhibit minimal variation in reasoning performance in different cultural contexts.

Smaller models show modest variations in reasoning performance in different cultural contexts. Gemma 3 1B shows differences of approximately 1.5 percentage points between the contextual baseline dataset (5. 98%) and the culturally adapted versions (4.53-5.20%). Llama 3.2 3B and Gemma 3 4B exhibit similar patterns with performance differences of approximately 2-4 percentage points in different contexts.

Interestingly, the weighted reasoning metric reveals that, while complete reasoning chains may vary across contexts, models often maintain partially correct reasoning chains. This suggests that cultural adaptations primarily affect specific reasoning steps rather than fundamental physics understanding. Calculation accuracy remains particularly stable across contexts for most models, suggesting that mathematical operations maintain consistency even when the same physics problems are presented with different cultural elements.

These findings highlight a critical challenge for educational applications of SLMs: while they can often produce correct answers to physics problems, generating completely sound reasoning chains remains difficult. This distinction is particularly important in educational contexts, where the quality of the explanation may be as valuable as the correctness of the answers.

5 Discussion

The evaluation of physics reasoning capabilities of SLMs reveals several significant insights. A critical observation is the substantial discrepancy between answer accuracy and reasoning quality across all models. Qwen 3 1.7B, despite not being the largest model tested, consistently outperformed larger counterparts like Gemma 3 4B and Phi 4 3.8B, achieving the highest answer accuracy across all datasets. Qwen 3 0.6B also demonstrated notable capabilities despite its lower parameter count. However, even the best-performing model reached only 40% fully correct reasoning, highlighting a disturbing pattern throughout the model spectrum. These findings suggest that SLMs often rely on pattern recognition rather than genuine physical understanding, particularly in multiple choice scenarios. Specialized training significantly improves reasoning capabilities, as demonstrated by Phi 4 Reasoning 3.8B outperforming the standard Phi 4 3.8B, indicating that targeted optimization can improve reasoning without increasing the model size.

The mathematical symbol representation has minimal impact on the quality of reasoning in all the SLMs evaluated. This notation-agnostic performance indicates robust capabilities for handling diverse mathematical representations, which is valuable for educational applications where content appears in various formats. For educational applications, this highlights that improvements in SLM-assisted learning should prioritize enhancing fundamental reasoning capabilities rather than adapting to specific mathematical formats.

Performance across Bloom's Taxonomy shows a clear decline from foundational to advanced reasoning tasks, with fully correct reasoning suffering most dramatically as cognitive demands increase. This suggests SLMs struggle with the compounding nature of errors in multi-step reasoning chains, where small inaccuracies cascade through the problem-solving process. Although factual knowledge is handled adequately, conceptual and procedural knowledge that is essential for physics presents challenges across all models tested. In physics education, particularly, where procedural problem-solving and development of increasingly complex cognitive skills are central to building expertise, SLMs with flawed reasoning chains may reinforce superficial understanding and impede students' progression toward advanced analytical and creative problem-solving abilities, especially when these models demonstrate correct answers despite faulty reasoning processes.

SLMs exhibit notable performance variations across physics domains, excelling in foundational topics while struggling with advanced topics, likely due to differing conceptual complexity and mathematical demands. Despite topic-dependent performance, models maintain stable reasoning across cultural contexts when physics principles remain unchanged, with calculation accuracy particularly consistent across contextual variations. This contextual robustness suggests promising applications for supporting physics education in diverse settings, underrepresented geographies and underserved communities.

6 Conclusion

Our study provided a systematic and multi-dimensional evaluation of the physics reasoning capabilities of contemporary SLMs. Our investigation, spanning diverse SLMs, physics topics, knowledge, and cognitive demands as per Bloom's Taxonomy, mathematical notation formats, and cultural contextualizations, reveals critical insights into the current strengths and limitations of these SLMs. The findings consistently demonstrate a notable disparity between SLMs' ability to produce correct final answers and their capacity for generating entirely sound reasoning chains.

The implications of these findings are significant for both the development of SLMs and their application in educational settings. For SLM advancement, efforts should prioritize enhancing genuine physics understanding, multi-step reasoning abilities, and the generation of coherent and correct reasoning chains, rather than solely optimizing for final answer correctness. Specialized training appears to be a promising avenue for such improvements. This could involve developing sophisticated verifiers capable of scrutinizing step-by-step deductions and exploring methods to improve the grounding of SLM outputs in fundamental physical laws and validated knowledge. Investigating hybrid architectures that synergize SLMs with other reasoning paradigms, such as symbolic systems or knowledge graphs, also presents a promising direction. For physics education, while SLMs offer potential benefits in terms of accessibility, privacy, and contextual adaptability, educators and developers must exercise caution. The tendency of SLMs to provide correct answers despite flawed reasoning could inadvertently reinforce superficial learning if not carefully managed. The quality of explanatory reasoning is paramount in educational tools.

With continued focus on reasoning quality rather than mere answer correctness, SLMs have the potential to evolve from pattern matching systems to genuine reasoning assistants. By addressing the fundamental limitations identified in this study, future developments could transform these efficient models into valuable educational tools that support physics education across different dimensions and resource settings, potentially expanding access to quality physics instruction where specialized teaching resources remain limited.

Limitations

The LLM-as-a-judge evaluation method, while scalable for rapidly evolving models, has inherent limitations such as potential evaluator bias and reduced nuance compared to resource-intensive human expert review. However, human evaluation is not scalable with new models arriving fast and considering the range of reasoning tasks. Furthermore, the cultural contextualization, though regionally diverse, was not globally exhaustive. The country-level focus might have missed finer local nuances and the full depth of cultural integration.

Ethics Statement

We obtained necessary permissions from OpenStax for the use of their high school physics textbook content in our evaluation of SLMs. The dataset creation and model evaluation processes were designed to respect intellectual property rights while facilitating research on physics reasoning capabilities.

References

Marah Abdin, Sahaj Agarwal, Ahmed Awadallah, Vidhisha Balachandran, Harkirat Behl, Lingjiao Chen, Gustavo de Rosa, Suriya Gunasekar, Mojan Javaheripi, Neel Joshi, and 1 others. 2025. Phi-4-reasoning technical report. *arXiv preprint arXiv:2504.21318*.

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024a. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, and 1 others. 2024b. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, and 1 others. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, and 1 others. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.

Cheng-Han Chiang and Hung-Yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631.

Diana I Cordova and Mark R Lepper. 1996. Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice. *Journal of educational psychology*, 88(4):715.

Badhan Chandra Das, M Hadi Amini, and Yanzhao Wu. 2025. Security and privacy challenges of large language models: A survey. *ACM Computing Surveys*, 57(6):1–39.

Kazuki Egashira, Mark Vero, Robin Staab, Jingxuan He, and Martin Vechev. 2024. Exploiting llm quantization. In *Advances in Neural Information Processing Systems*, volume 37, pages 41709–41732. Curran Associates, Inc.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, and 1 others. 2023. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*.

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

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Rabih Kahaleh and Victor Lopez. 2025. Evaluating large language models in high school physics education: addressing misconceptions and fostering conceptual understanding. *Physics Education*, 60(2):025013.

Aabid Karim, Abdul Karim, Bhoomika Lohana, Matt Keon, Jaswinder Singh, and Abdul Sattar. 2025. Lost in cultural translation: Do llms struggle with math across cultural contexts? *arXiv preprint arXiv:2503.18018*.

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

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*.

Abdullah Mushtaq, Imran Taj, Rafay Naeem, Ibrahim Ghaznavi, and Junaid Qadir. 2025. Worldview-bench: A benchmark for evaluating global cultural perspectives in large language models. *arXiv preprint arXiv:2505.09595*.

Giulia Polverini and Bor Gregorcic. 2024. How understanding large language models can inform the use of chatgpt in physics education. *European Journal of Physics*, 45(2):025701.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Timo Schick and Hinrich Schütze. 2021. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352.

Gaurav Srivastava, Shuxiang Cao, and Xuan Wang. 2025. Towards reasoning ability of small language models. *arXiv preprint arXiv:2502.11569*.

Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. 2020. Mobilebert: a compact task-agnostic bert for resource-limited devices. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2158–2170.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 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, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, and 1 others. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. 2024. Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges. *arXiv preprint arXiv:2406.12624*.

Paul Peter Urone and Roger Hinrichs. 2020. *Physics*. OpenStax, Houston, Texas.

Yumou Wei, Paulo Carvalho, and John Stamper. 2025. Small but significant: On the promise of small language models for accessible aied. *arXiv preprint arXiv:2505.08588*.

Xin Xu, Tong Xiao, Zitong Chao, Zhenya Huang, Can Yang, and Yang Wang. 2025. Can LLMs solve longer math word problems better? In *The Thirteenth International Conference on Learning Representations*.

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

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*.

A Dataset Development and Annotation

A.1 Dataset extraction and preprocessing

A challenge encountered during the data extraction phase was the representation of mathematical equations, which were predominantly embedded as images rather than text within the exercise documents. To address this challenge, we used multiple tools, including specialized optical character recognition (OCR) models (Google Vision API, Mathpix), to convert equation images into LaTeX format. This approach significantly reduced the need for manual transcription of mathematical notation and preserved the integrity of the mathematical content.

Following the extraction of questions, answer choices, correct answers, and the reasoning for why the answer is correct from the textbook and teacher resources from OpenStax (which served as our ground truth for evaluating the model responses), a comprehensive data cleaning and refinement process was implemented. This systematic review, which combined automated checks and manual verification, found several inconsistencies in the source material. These included references to non-existent textbook figures in the extracted context, explanations with logical flaws or incomplete reasoning, and questions that lacked sufficient standalone information as they referenced specific textbook sections. Each question was meticulously evaluated to ensure clarity, correctness, and completeness. Entries with irreparable flaws were removed, and others were revised to meet the quality standards required for the present study.

A.2 Bloom's taxonomy annotations

We annotated each physics problem according to the revised Bloom's Taxonomy (Krathwohl, 2002), which consists of two dimensions:

Cognitive Process Dimension:

- *Remember:* Retrieving relevant knowledge from long-term memory, such as recalling facts, basic concepts, or definitions
- *Understand:* Constructing meaning from instructional materials, including interpreting, summarizing, and explaining ideas
- *Apply:* Using procedures or learned methods in a given situation to solve problems or carry out tasks
- *Analyze:* Breaking down information into components, identifying relationships or patterns, and understanding structure and function
- *Evaluate:* Judging or determining the value of material or methods based on criteria or standards
- *Create:* Generating new ideas, products, or structures by combining elements into a coherent or functional whole

Knowledge Dimension:

- *Factual:* The basic elements or facts students must know to solve problems, including terminology and specific details
- *Conceptual:* The interrelationships between elements, such as theories, principles, and models, that enable function within a domain
- *Procedural:* Knowing how to perform tasks, techniques, and methods, and when to apply them
- *Metacognitive:* Knowledge about one's own cognition and how to regulate it, including self-awareness of learning strategies

The distribution of questions across the Cognitive Process and Knowledge Dimensions revealed important characteristics of high school physics education and resulting limitations in our dataset:

Table 2: Distribution of questions across Cognitive Process Dimension

Cognitive Level	**Percentage**	**Count**
Remember	18.3%	239
Understand	26.1%	341
Apply	29.5%	385
Analyze	16.7%	218
Evaluate	7.4%	97
Create	2.0%	26

As evidenced by this distribution, the dataset contained questions spanning different levels of cognitive skills, but with notable limitations. Lower-order cognitive processes (Remember, Understand, Apply) accounted for approximately 74% of the questions, while higher-order processes (Analyze, Evaluate, Create) comprised only 26%. This

Table 3: Distribution of questions across Knowledge Dimension

Knowledge Type	Percentage	Count
Factual	16.2%	212
Conceptual	42.5%	555
Procedural	41.3%	539
Metacognitive	0%	0

distribution reflects the typical emphasis in high school physics education.

Notably, no questions were classified as addressing metacognitive knowledge. This absence reflects the traditional focus of high school physics curricula on factual, conceptual, and procedural knowledge rather than on developing students' awareness of their own cognitive processes. Additionally, metacognitive questions are challenging to assess in standardized formats and are often addressed through reflective exercises or learning journals rather than end-of-chapter problems.

A.3 Topic composition in the dataset

The distribution of physics topics in our dataset reflects the comprehensive coverage of a typical high school physics curriculum. Table 4 shows the percentage and count breakdown of questions across different physics domains.

Table 4: Distribution of questions across physics topics

Physics Topic	Percentage	Count
Introduction	4.7%	61
Mechanics	32.5%	424
Electricity & Magnetism	21.9%	286
Thermodynamics	8.6%	112
Waves & Acoustics	10.2%	133
Optics	12.3%	161
Modern Physics	9.8%	129

This topic distribution ensured comprehensive coverage of the physics curriculum, allowing us to evaluate SLM performance across the full spectrum of physics concepts typically encountered in high school education. The contextualization process maintained this topic distribution in each culturally adapted dataset, ensuring that comparative analyses across different cultural contexts were not confounded by variations in topic coverage.

The Introduction category includes foundational concepts such as scientific notation, measurement, and dimensional analysis. Mechanics covers motion, forces, energy, and momentum. Electricity & Magnetism encompasses electric charge, current, circuits, and magnetic fields. Thermodynamics includes heat, temperature, and the laws of thermodynamics. Waves & Acoustics covers mechanical waves, sound, and basic wave phenomena. Optics includes light, mirrors, lenses, and optical instruments. Modern Physics covers topics such as quantum mechanics, atomic physics, and nuclear physics.

B Cultural Contextualization Methodology

Our cultural contextualization approach required developing comprehensive regional databases to ensure authentic representation. Countries were selected systematically based on the United Nations Geoscheme[1], with particular emphasis on underrepresented regions. We organized cultural information into three distinct regional datasets:

1. **Asian Context:** Included information from countries such as India, China, Indonesia, Philippines, and many more (51 countries in total)

2. **African Context:** Incorporated elements from Nigeria, Kenya, South Africa, Ethiopia, and many more (58 countries in total)

3. **South American and Australian Context:** Featured Brazil, Argentina, Colombia, Peru, Australia, and many more (41 countries in total)

For each country, we compiled structured information on common names and honorifics, cultural festivals and celebrations, geographical landmarks and natural features, local foods and culinary traditions, region-specific modes of transportation, popular sports and recreational activities, cultural traditions, rituals, and practices, and local industries and occupations.

We selected Google's Gemini 2.5 Pro model for generating cultural information because of Google's international presence and search index across virtually all countries provided the model with exposure to authentic cultural elements from the different countries across the regions. The

[1] https://unstats.un.org/unsd/methodology/m49/

model's integrated 'tool use' capabilities with Google Search enabled real-time retrieval and verification of cultural information. This search-augmented generation approach, combined with the model's built-in reasoning capabilities, allowed for systematic fact-checking and refinement of cultural details during the generation process itself. The generated cultural data was subsequently verified through additional targeted internet searches, review of materials from relevant cultural heritage websites and reputable online encyclopedias, and cross-referencing with publicly available demographic and cultural information. The verification process helped ensure cultural accuracy and representativeness while avoiding stereotypical portrayals. This carefully vetted cultural information was provided as contextual input to guide the model during the question generation process.

With these three regional cultural context databases, we proceeded to the question adaptation phase using Google's Gemini 2.5 Flash model, chosen for its ability to efficiently generate a large volume of adaptations while consistently producing high-quality, well-formed contextualized questions. For each of the 393 selected original physics questions, the model was instructed to analyze the original question, deconstruct its underlying physics principles, and then integrate elements from our cultural context database. A critical directive in this process was to maintain physics fidelity. The model was instructed to ensure that all contextualized variations retained the original core physics concepts, mathematical relationships, and formulae (preserved in $\LaTeX$ notation), and the overall difficulty level. To foster diversity, we generated five distinct contextualized variations for each original question. This approach produced three culturally adapted datasets, each containing $393 \times 5 = 1,965$ questions. The system maintained a history of previously generated questions for each country within the region in each generation instance, which helped prevent repetition and ensure authenticity, addressing the tendency of language models to produce a similar output when creating multiple items. For multiple choice questions, options and correct answers were generally preserved, unless contextual adaptation required modification for coherence. For open-ended questions, the underlying reasoning remained consistent with the physics tested in the original problem.

A custom implementation managed the entire workflow from question selection to final output processing. The system maintained comprehensive records of all generation attempts, producing a parallel dataset of original and culturally contextualized physics problems for comparative analysis of the physics reasoning abilities of SLMs across different cultural contexts. The prompts used for multiple choice and open ended question generations are given in Figure 1 and 2.

C Model Inference and Evaluation details

C.1 Parameters for cultural context generation

For the cultural context database creation and contextualized question generation phases, we used slightly different parameters. Based on Google's recommendations for their models, we set temperature to 0.2 and top_p to 0.95 when using Gemini 2.5 Pro (for cultural information generation) and Gemini 2.5 Flash (for question adaptation). This slightly higher temperature value provided an appropriate balance between creativity and consistency, allowing for diverse cultural elements and problem formulations while maintaining coherence and factual accuracy.

C.2 Parameters for SLM inference

For our primary experiments evaluating reasoning capabilities in SLMs, we used consistent inference parameters across all models to ensure fair comparison. All SLMs were run with a temperature of 0.1 and top_p of 0.95. These low temperature settings were selected to minimize randomness and promote deterministic outputs, which is particularly important for assessing reasoning capabilities.

C.3 Parameters for evaluation

For our LLM-as-a-judge framework, we utilized Google's Gemini 2.5 Flash model with conservative sampling parameters (temperature = 0.1, top_p = 0.95). These settings were selected to minimize stochasticity in the evaluation process, ensuring consistent and reliable assessment of both answer correctness and reasoning quality across all model outputs.

Contextual_Multiple_Choice

You are tasked with creating culturally contextualized physics questions.
Input:
Question:

{question}

Options:

{options}

Ground Truth:

- Correct Option: {correct_option}
- Correct Option Answer: {correct_option_answer}
- Reasoning: {ground_truth_reasoning}

Context:

{context}

Question History:

{question_history}

Step-by-step Instructions:
Step 1: Carefully read and understand the physics question
- Analyze the physical scenario described
- Identify the core physics concepts and principles involved
- Note any formulas or equations used

Step 2: Examine the provided answer options
- Understand what each option represents
- Note the format and units of measurement

Step 3: Identify the correct answer and understand why it's correct
- Review the correct option letter/number
- Study the reasoning explanation thoroughly
- Understand the solution method and calculations involved

Step 4: Review the question history for this country
- Analyze previously generated questions in the question history
- Note which cultural elements, scenarios, and contexts have already been used
- Identify patterns to avoid repeating

Step 5: Analyze the cultural context provided in the context JSON
- Identify the country
- Review the available cultural elements:
  * names
  * festivals
  * locations
  * foods
  * transportation
  * sports
  * other_elements (clothing, music_and_dance, art_and_crafts, traditions_and_customs, etc.)
- Prioritize cultural elements that have NOT been used in the question history

Step 6: Create cultural variations of the question
- While keeping the core physics problem identical:
  a) Replace Western/generic names with culturally specific names from the context
  b) Change the setting to culturally relevant locations from the context
  c) Incorporate cultural elements like festivals, foods, transportation, sports, etc.
  d) Use traditional objects, instruments, or clothing when applicable
  e) Maintain the same level of difficulty and mathematical relationships
  f) Preserve all mathematical formulas using LaTeX notation
  g) Ensure the new questions are distinct from those in the question history

Step 7: Generate 5 distinct cultural variations
- Ensure each variation uses different combinations of cultural elements like names
- Avoid repetition of the same cultural details across questions
- Each variation should focus on different aspects of the culture (e.g., one on festivals, one on sports, etc.)
- Thoroughly avoid Western cultural elements and previously used scenarios
- Each variation should feel authentic to the specified country

Step 8: Format the output as a JSON array with 5 objects
- Include the country name in each object
- Keep the original options, correct answer, reasoning, and answer text unchanged

Output format:

```
[
   {
      "Country":  "country name from context",
      "ContextualQuestion":  "Culturally adapted question text 1",
      "ContextOptions":  "same as original options if context doesn't affect options", //modify if changes are
required be made to options taking context into account like ["<first>", "<second>", ..]
      "ContextCorrectOption":  "same as original",
      "ContextReasoning":  "same as original if context doesn't affect reasoning", // modify if changes are
required be made to reasoning taking context into account
      "ContextCorrectOptionAnswer":  "same as original if context doesn't affect answer"
   },
   // 4 more similar JSON objects with different cultural elements
]
```

IMPORTANT: Return ONLY this JSON object with no additional text.

Figure 1: Prompt template used for creating 5 contextual multiple choice physics questions for each question in the dataset.

Contextual_Open_Ended

You are tasked with creating culturally contextualized physics questions.

Input:

Question:

{question}

Ground Truth:

- Reasoning: {ground_truth_reasoning}

Context:

{context}

Question History:

{question_history}

Step-by-step Instructions:
Step 1: Carefully read and understand the reference physics question
- Analyze the physical scenario described
- Identify the core physics concepts and principles involved
- Note any formulas or equations used

Step 2: Identify the expected approach to answering
- Study the reasoning explanation thoroughly
- Understand the solution method and explanations involved

Step 3: Review the question history for this country
- Analyze previously generated questions in the question history
- Note which cultural elements, scenarios, and contexts have already been used
- Identify patterns to avoid repeating

Step 4: Analyze the cultural context provided in the context JSON
- Identify the country
- Review the available cultural elements:
  * names
  * festivals
  * locations
  * foods
  * transportation
  * sports
  * other_elements (clothing, music_and_dance, art_and_crafts, traditions_and_customs, etc.)
- Prioritize cultural elements that have NOT been used in the question history

Step 5: Create cultural variations of the question
- While keeping the core physics problem identical:
  a) Replace Western/generic names with culturally specific names from the context
  b) Change the setting to culturally relevant locations from the context
  c) Incorporate cultural elements like festivals, foods, transportation, sports, etc.
  d) Use traditional objects, instruments, or clothing when applicable
  e) Maintain the same level of difficulty and mathematical relationships
  f) Preserve all mathematical formulas using LaTeX notation
  g) Ensure the new questions are distinct from those in the question history

Step 6: Generate 5 distinct cultural variations
- Ensure each variation uses different combinations of cultural elements like names
- Avoid repetition of the same cultural details across questions
- Each variation should focus on different aspects of the culture (e.g., one on festivals, one on sports, etc.)
- Thoroughly avoid Western cultural elements and previously used scenarios
- Each variation should feel authentic to the specified country

Step 7: Format the output as a JSON array with 5 objects
- Include the country name in each object
- Keep the original reasoning unchanged

Output format:

```
[
   {
       "Country":  "country name from context",
       "ContextualQuestion":  "Culturally adapted question text 1",
       "ContextReasoning":  "same as original if context doesn't affect reasoning", // modify if
changes are required be made to reasoning taking context into account
   },
   // 4 more similar JSON objects with different cultural elements
]
```

IMPORTANT: Return ONLY this JSON object with no additional text.

Figure 2: Prompt template used for creating 5 contextual open ended physics questions for each question in the dataset.

Contextual_Open_Ended

Original Question:

How many 10Ω resistors must be connected in series to make an equivalent resistance of 80Ω ?

Options: A. 80, B. 8, C. 20, D. 40

Asian Context (India):

Aarav is building a circuit for his science project in Delhi. He has a box full of 10Ω resistors. **How many of these** 10Ω resistors must **he** connect in series to **achieve a total** equivalent resistance of 80Ω **for a specific part of his circuit**?

Options: A. 80, B. 8, C. 20, D. 40

African Context (Nigeria):

Adetokunbo is setting up a sound system for a community gathering in Lagos. He needs a specific part of the circuit to have an equivalent resistance of 80Ω. **If he is connecting identical resistors of** 10Ω **each in series**, how many 10Ω resistors must be connected in series to make an equivalent resistance of 80Ω?

Options: A. 80, B. 8, C. 20, D. 40

South American Context (Brazil):

João is working on the electrical system for a small 'barraca' (stall) at a Festa Junina celebration. He needs a section of the circuit to have an equivalent resistance of 80Ω. **If he only has** 10Ω **resistors available**, how many **of these** 10Ω resistors must be connected in series to make an equivalent resistance of 80Ω?

Options: A. 80, B. 8, C. 20, D. 40

Correct Answer: B. 8

Figure 3: Example of a contextual question across different regions.

Multiple_Choice

You are an expert physics teacher evaluating AI-generated responses to physics problems.

Analyze the following with great care:

Question: {question}
Options: {options}

Ground Truth:

Correct Option: {correct_option}
Correct Option Answer: {correct_option_answer}
Reasoning: {ground_truth_reasoning}

Model Response:

Selected Option: {model_answer}
Model Reasoning: {model_reasoning}
Model Explanation: {model_explanation}

Evaluation Instructions:
Step 1: Check if the model's selected option ({model_answer}) matches the correct option ({correct_option}).
Step 2: Carefully trace through the model's reasoning step-by-step.
Step 3: Compare the model's explanation with the ground truth reasoning.
Step 4: For ALL responses, identify which parts of the reasoning are correct and which are incorrect.

Step 5: Categorize the response's answer into one of these numeric categories:
(1) Correct answer: Model's selected option matches the correct option
(0) Wrong answer: Model's selected option does not match the correct option

Step 6: Categorize the reasoning into one of these numeric categories:
(2) Fully correct reasoning: All physics principles, concepts, and logic are correct
(1) Partially correct reasoning: Some correct physics principles but with errors or misconceptions
(0) Incorrect reasoning: Fundamental flaws in the physics concepts, formulas, or approach

Step 7: For calculations, use these categories:
(2) No calculations required for this problem
(1) Calculations required and performed correctly
(0) Calculations required but performed incorrectly or with errors

Verification Guidelines:
For partially correct reasoning, identify both the correct reasoning elements and the specific errors or misconceptions
For incorrect reasoning, identify the fundamental flaws in the physics understanding
For calculation errors, specify exactly what went wrong in the mathematical steps
For numerical problems: approximations within $\approx$ 1-2% of calculated values are reasonable

IMPORTANT: Your response MUST be in the following JSON format:

```
{
"answer": <1 for correct or 0 for wrong>,
"reasoning": <2 for fully correct, 1 for partially correct, or 0 for incorrect>,
"calculations": <2 for no calculations needed, 1 for correct calculations, 0 for incorrect
calculations>,
"explanation": "<brief explanation highlighting verification of correct and incorrect elements
in reasoning>"
}
```

Return ONLY this JSON object with no additional text.

Figure 4: Evaluation prompt template used for assessing model responses to multiple choice physics questions.

Multiple_Choice_Unstructured_Response

You are an expert physics teacher evaluating AI-generated responses to physics problems.

Analyze the following with great care:

Question: {question}
Options: {options}

Ground Truth:

- Correct Option: {correct_option}
- Correct Option Answer: {correct_option_answer}
- Reasoning: {ground_truth_reasoning}

Model Response:

{model_response}

Evaluation Instructions:
Step 1: Extract the model's selected option from the Model Response text. Look for patterns like "ANSWER: [letter]", "The answer is [letter]", or clear indication of option selection.
Step 2: Extract the model's reasoning from the Model Response text. Look for sections marked "REASONING:" or explanatory paragraphs.
Step 3: Verify if the extracted answer matches the correct option ({correct_option}).
Step 4: Carefully trace through the extracted reasoning step-by-step.
Step 5: Identify which parts of the reasoning are correct and which are incorrect.

Step 6: Categorize the response's answer into one of these numeric categories:
(1) Correct answer: Model's selected option matches the correct option
(0) Wrong answer: Model's selected option does not match the correct option

Step 7: Categorize the reasoning into one of these numeric categories:
(2) Fully correct reasoning: All physics principles, concepts, and logic are correct
(1) Partially correct reasoning: Some correct physics principles but with errors or misconceptions
(0) Incorrect reasoning: Fundamental flaws in the physics concepts, formulas, or approach

Step 8: For calculations, use these categories:
(2) No calculations required for this problem
(1) Calculations required and performed correctly
(0) Calculations required but performed incorrectly or with errors

Verification Guidelines:
Document what was extracted from the Model Response
For partially correct reasoning, identify both the correct reasoning elements and the specific errors or misconceptions.
For incorrect reasoning, identify the fundamental flaws in the physics understanding.
If no clear answer or reasoning can be extracted, categorize as 0
For calculation errors, specify exactly what went wrong in the mathematical steps.
For numerical problems: approximations within 1–2% of calculated values are reasonable.

IMPORTANT: Your response MUST be in the following JSON format:

```
{
"answer": <1 for correct or 0 for wrong>,
"reasoning": <2 for fully correct, 1 for partially correct, or 0 for incorrect>,
"calculations": <2 for no calculations needed, 1 for correct calculations, 0 for incorrect
calculations>,
"explanation": "<brief explanation highlighting verification of correct and incorrect elements
in reasoning>"
}
```

Return ONLY this JSON object with no additional text.

Figure 5: Evaluation prompt template used for assessing unstructured responses to multiple choice physics questions.

Open_Ended

You are an expert physics teacher evaluating AI-generated responses to open-ended physics problems. This question does not have a single correct answer, so evaluate whether the response adequately addresses the question.

Question: {question}

Expected Approach/ Reasoning: {ground_truth_reasoning}

Topic Context:

- Physics Topic: {topic}
- Knowledge Type: {knowledge_dimension}
- Cognitive Level: {cognitive_dimension}

Model Response:

{model_response}

Evaluation Instructions:
Step 1: Determine if the model's response actually addresses the question asked.
Step 2: Carefully trace through the model's approach and reasoning.
Step 3: For performance tasks, check if all parts (Part A, B, C, etc.) are addressed.
Step 4: For numerical problems, verify calculations. For theoretical problems, verify concepts.
Step 5: Compare the model's approach with the expected reasoning, but allow for valid alternative approaches.

Step 6: Categorize the response's adequacy:
(1) Adequate answer: Response appropriately addresses the question with valid physics
(0) Inadequate answer: Response fails to address the question or contains major errors

Step 7: Categorize the reasoning quality:
(2) Fully correct reasoning: Excellent physics understanding, complete and accurate
(1) Partially correct reasoning: Good physics understanding with minor gaps or errors
(0) Incorrect reasoning: Poor physics understanding or significant errors

Step 8: For calculations, use these categories:
(2) No calculations required for this problem
(1) Calculations required and performed correctly
(0) Calculations required but performed incorrectly or with errors

Verification Guidelines:
Accept valid alternative approaches that differ from the expected reasoning
For experimental design, evaluate practicality and physics validity
For multi-part questions, assess completeness of coverage
Focus on physics accuracy rather than exact match to expected answer
For numerical problems: approximations within 1-2% of calculated values are reasonable

IMPORTANT: Your response MUST be in the following JSON format:

```
{
"answer":  <1 for adequate or 0 for inadequate>,
"reasoning":  <2 for fully correct, 1 for partially correct, or 0 for incorrect>,
"calculations":  <2 for no calculations needed, 1 for correct calculations, 0 for incorrect
calculations>,
"explanation":  "<analysis of how well the response addresses the question and physics
accuracy>"
}
```

Return ONLY this JSON object with no additional text.

Figure 6: Evaluation prompt template used for assessing model responses to open ended physics questions.

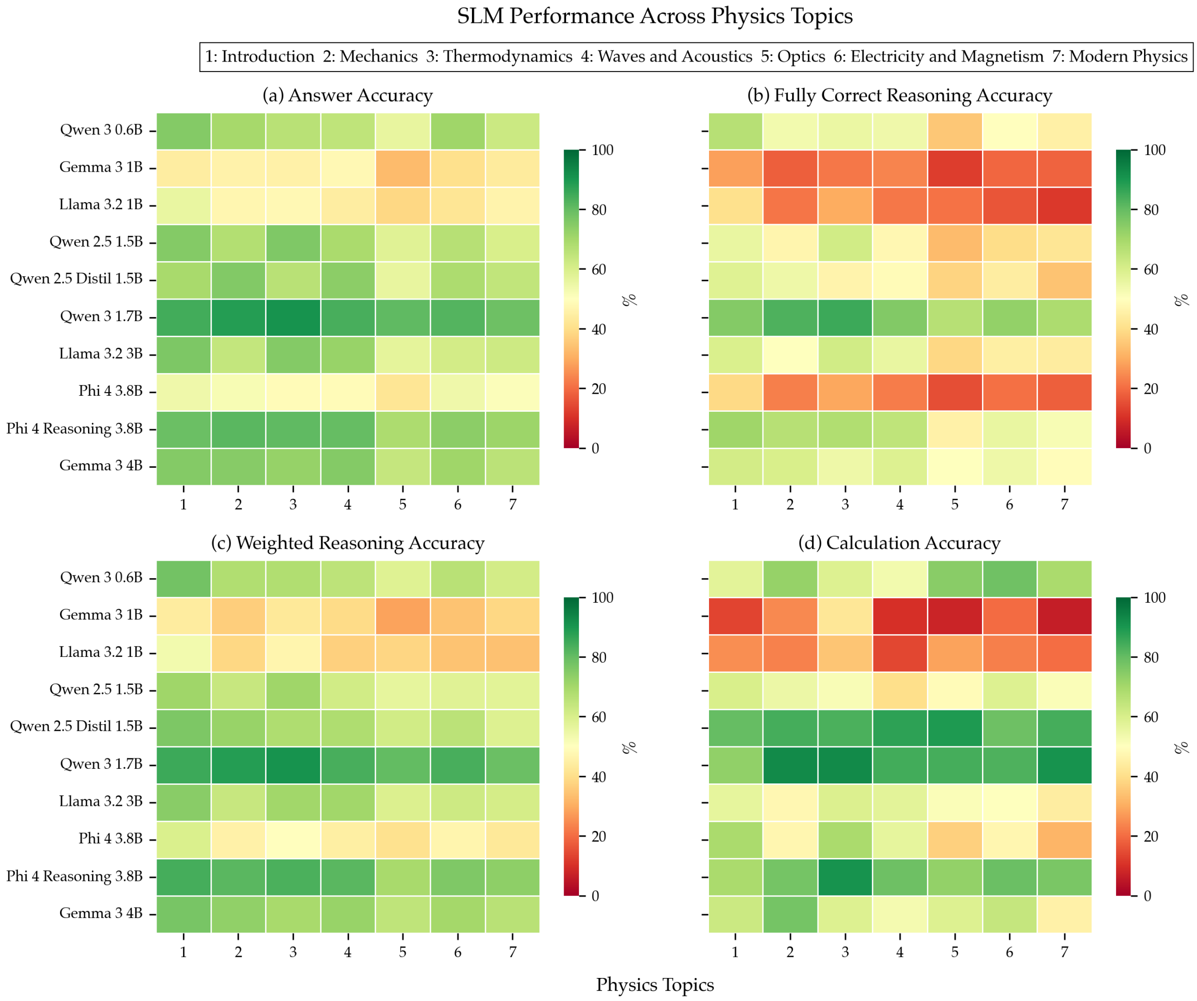

Figure 7: SLM Performance Across Physics Topics. The heatmaps show: (a) Answer Accuracy, (b) Fully Correct Reasoning Accuracy, (c) Weighted Reasoning Accuracy, and (d) Calculation Accuracy across physics topics. Topics: 1: Introduction, 2: Mechanics, 3: Thermodynamics, 4: Waves and Acoustics, 5: Optics, 6: Electricity and Magnetism, 7: Modern Physics.

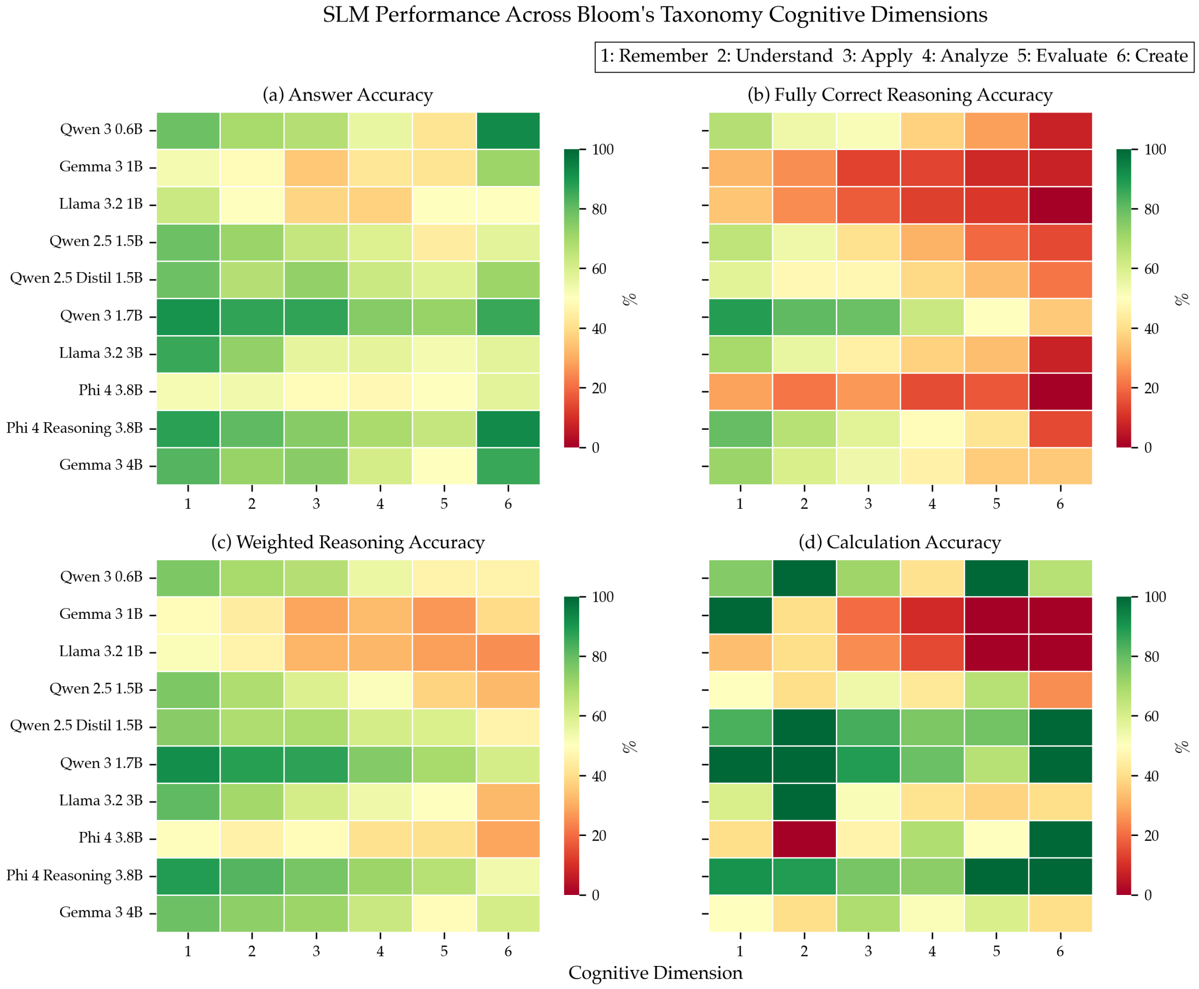

Figure 8: SLM Performance Across Bloom's Taxonomy Cognitive Dimensions. The heatmaps show: (a) Answer Accuracy, (b) Fully Correct Reasoning Accuracy, (c) Weighted Reasoning Accuracy, and (d) Calculation Accuracy across cognitive dimensions. Cognitive dimensions: 1: Remember, 2: Understand, 3: Apply, 4: Analyze, 5: Evaluate, 6: Create.

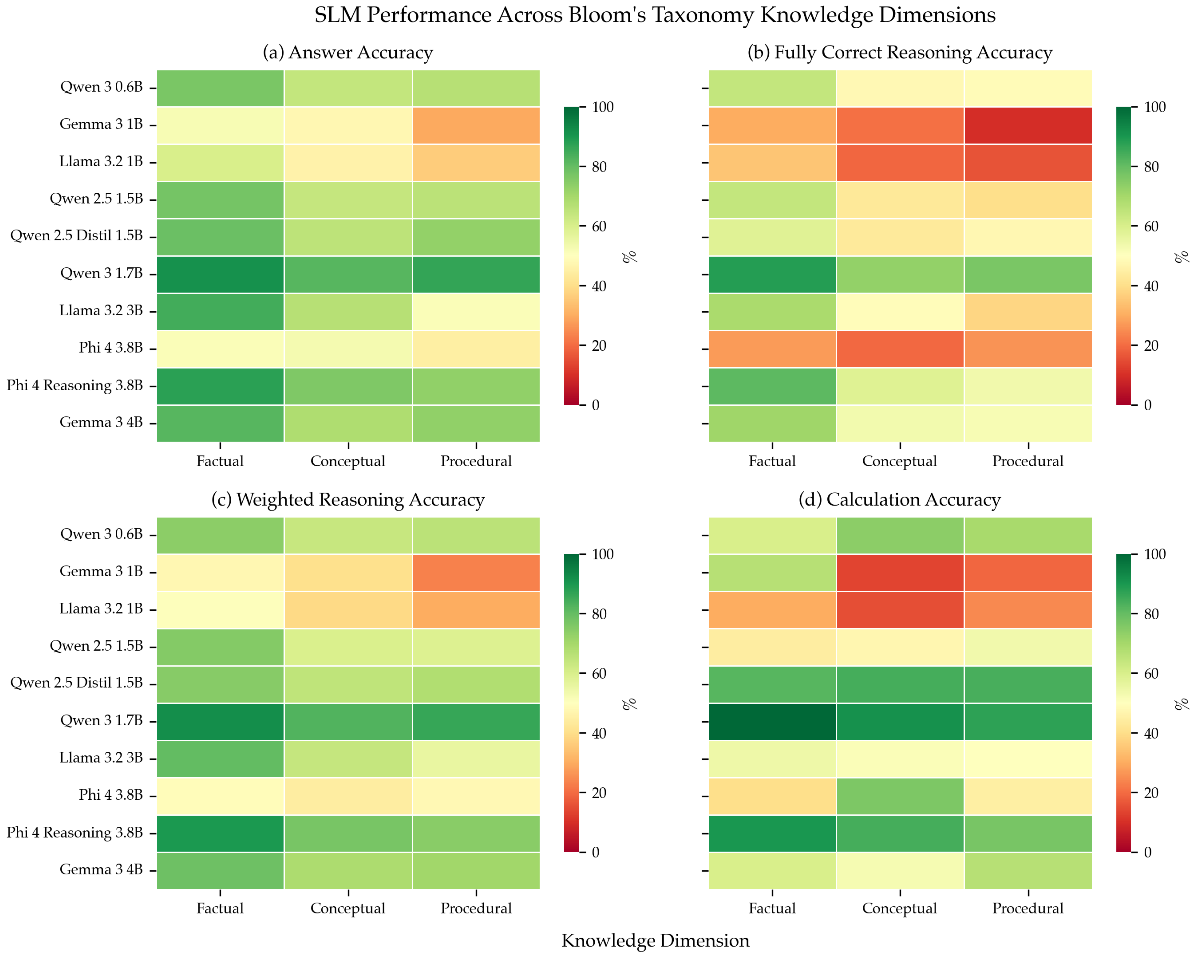

Figure 9: SLM Performance Across Bloom's Taxonomy Knowledge Dimensions. The heatmaps show: (a) Answer Accuracy, (b) Fully Correct Reasoning Accuracy, (c) Weighted Reasoning Accuracy, and (d) Calculation Accuracy across knowledge dimensions (Factual, Conceptual, Procedural).