Disentangling Logic: The Role of Context in Large Language Model Reasoning Capabilities

Anonymous ACL submission

Abstract

This study intends to systematically disentangle pure logic reasoning and text understanding by investigating the contrast across abstract and contextualized logical problems from a comprehensive set of domains. We explore whether LLMs demonstrate genuine reasoning capabilities across various domains when the underlying logical structure remains constant. We focus on two main questions (1) Can abstract logical problems alone accurately benchmark LLMs' reasoning ability in real-world scenarios, disentangled from contextual support in practical settings? (2) Does fine-tuning LLMs on abstract logic problems generalize to contextualized logic problems and vice versa? To investigate these questions, we focus on standard propositional logic, specifically propositional deductive and abductive logic reasoning. We construct datasets for both reasoning types with four difficulty levels across 12 distinct domains based on the Wikipedia categorization in additional to those with purely abstract variables. Our experiments aim to provide insights into disentangling context in logical reasoning, the genuine reasoning capabilities of LLMs, and their generalization potential. Coda and data are available at `https://anonymous.4open.science/r/ContextHub-957E`.

1 Introduction

Large language models (LLMs) (Team et al., 2024; Arrieta et al., 2025; Guo et al., 2025) have demonstrated significant potential in reasoning capabilities across a variety of reasoning benchmarks (Cobbe et al., 2021; Hendrycks et al., 2021; Wei et al., 2022a; Liang et al., 2023; bench authors, 2023; Zhu et al., 2024; Fan et al., 2023a; Fu et al., 2024), broadening their potential applications in fields such as psychology, education, and social sciences (Gandhi et al., 2024; Li et al., 2024a; Fan et al., 2023b). The widespread use of LLMs necessitates rigorous evaluation of their reasoning abilities, particularly in context-rich scenarios that reflect real-world complexities.

While assessments on abstract logical problems (Sawada et al., 2023; Zhu et al., 2024; Fan et al., 2023a, 2024; Fu et al., 2024) showcase LLMs' theoretical reasoning capacities, they do not entirely capture their practical utility in real-life applications where context drastically affects outcomes. Conversely, focusing exclusively on context-specific tasks (Guha et al., 2024; Liévin et al., 2024; Han et al., 2022; Clark et al., 2020; Hendrycks et al., 2020) may conceal the fundamental mechanisms that empower LLMs to process and reason with information. Thus, exploring the disparity between contextualized and abstract reasoning (Tang et al., 2023; Saparov and He, 2022) is vital for advancing LLMs and ensuring their effectiveness across different domains.

To this end, we introduce ContextHub, a benchmark designed to systematically and meticulously disentangle and evaluate the core reasoning capabilities of LLMs from the influences of contextual information. By leveraging a dual-assessment framework, ContextHub compares LLMs' performance on identical logical constructs within both abstract and richly contextualized settings (a contextualized example can be seen in step 3 of Figure 2). This approach not only identifies the significant impacts of context on reasoning but also provides a scalable and flexible methodology adaptable across various domains.

ContextHub aims to address two main questions: (1) **Evaluation disentanglement**: how accurate and reliable are abstract logic problems versus contextualized problems in evaluating LLMs' reasoning abilities? We examine this by comparing LLM performance across both problem types to understand context's role in reasoning. (2) **Fine-tuning disentanglement**: how do abstract versus contextualized logic problems affect model generalization during fine-tuning? We analyze LLM performance

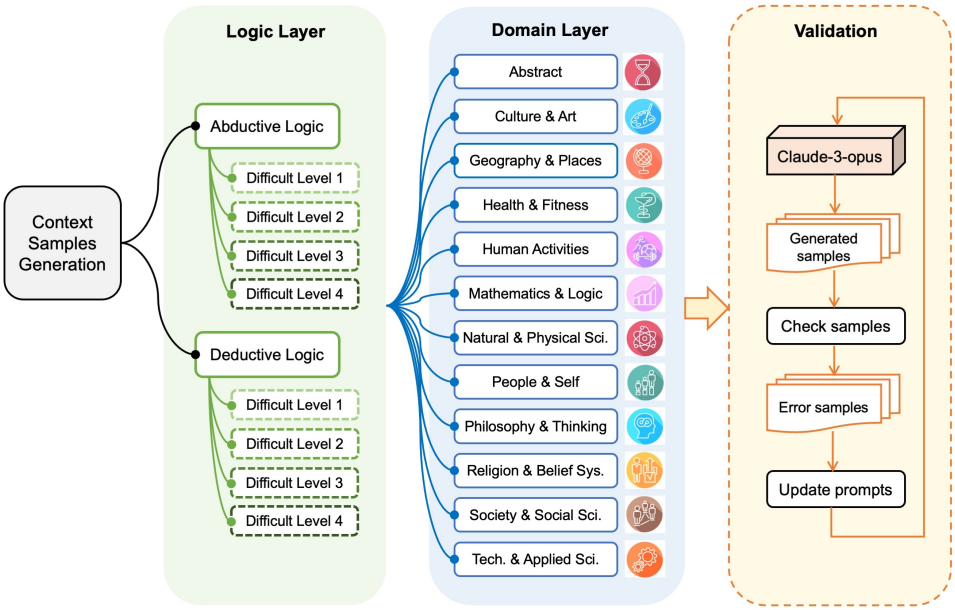

Figure 1: Benchmark Construction Procedure

on unseen problems of both types to determine which training data best improves reasoning while maintaining cross-domain consistency.

We employ a dynamic graph-based construction method (Zhu et al., 2024) to generate formal logic templates at four levels of difficulty. These templates are then instantiated in 13 distinct domains, comprising 12 contexts with specific settings and 1 using purely abstract variables for comparison. Every instantiated question undergoes a rigorous two-step quality control process to ensure reliability. We evaluate these datasets using a variety of LLMs, including GPT-4o, GPT-3.5-Turbo, Qwen series, Yi-series, LLaMA-2, and LLaMA-3.1 series. In the fine-tuning phase, we compare three settings: models fine-tuned on abstract data only, on mixed contextualized data from all domains, and on single domain data. These settings enable us to investigate how abstract and contextualized logical problems affect LLMs' generalization abilities. Our key findings are: (1) The relative performance of LLMs on abstract logic problems and corresponding contextualized logic problems is **dependent on model size or general model performance**. Stronger models tend to perform better on abstract logic, while smaller models typically rely on contextual cues. (2) The domain of contextualization has a **statistically significant** impact on model performance, suggesting the choice of contextualization domain can affect the accuracy and reliability of LLMs for logical reasoning tasks. (3) The **generalization power of abstract logic data is limited** compared with that of contextualized logic data. This indicates that LLMs fine-tuned on contextualized logic data may be better equipped to handle a wider range of real-world logical reasoning tasks.

2 Related Work

Evaluating the reasoning abilities of LLMs has garnered significant attention across various disciplines, from biomedical informatics (Liévin et al., 2024; Chen et al., 2024; Jin et al., 2024b) and humanities (Hua et al., 2023; Lin et al., 2024; Jin et al., 2024a) to social sciences (Ziems et al., 2024; Gandhi et al., 2024; Li et al., 2024a; Fan et al., 2023b). Research has predominantly concentrated on diverse logical reasoning tasks, including deduction, induction, and abduction, addressed through neural models (Pan et al., 2023; Li et al., 2024b; Dasgupta et al., 2022; Han et al., 2022; Del and Fishel, 2022). LogicBench (Parmar et al., 2024) focuses on natural language logical reasoning questions. FOLIO (Han et al., 2022), RuleTaker (Clark et al., 2020), and FLD (Morishita et al., 2023) build logic questions based on deduction rules focusing only on validity without semantics. Nevertheless, no existing benchmark offers a systematic, fine-grained investigation into how additional contextual detail or scenario-specific variations can affect the stability and reliability of LLMs' logical reasoning—particularly in terms of how prior knowledge might bolster performance or, conversely, how the absence of such knowledge may lead to greater degradation than performance in a purely abstract

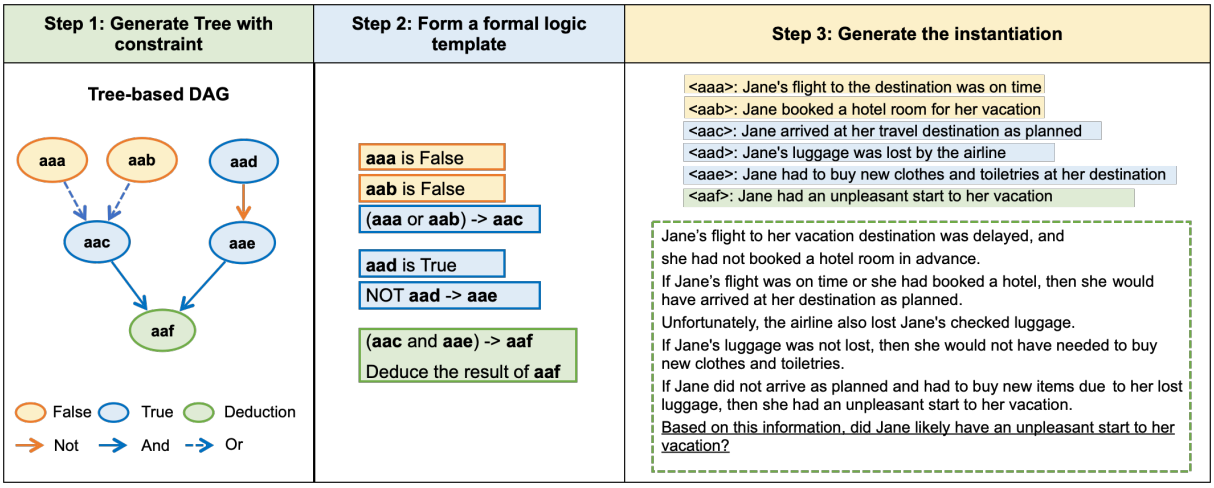

Figure 2: An illustration of abstract and contextualized logical problems.

context. While LLMs have demonstrated notable successes in certain reasoning tasks (Cobbe et al., 2021), their generalizable logical reasoning capabilities remain uncertain (Tang et al., 2023; Saparov and He, 2022).

There are also many logic reasoning benchmarks such as LogicBench (Parmar et al., 2024) and FOLIO (Han et al., 2022) using contextualized language over logic templates. Benchmarks such as RuleTaker (Clark et al., 2020) and FLD (Morishita et al., 2023) build synthetic logic questions based on logic templates without coherence semantics. However, none of these benchmarks systematically examine the influence of contextual factors on reasoning performance or explore the stability of logical reasoning skills across diverse scenarios, including those that are purely abstract.

Valmeekam et al. (Valmeekam et al., 2022) argued that LLMs often struggle with common planning and reasoning tasks, which are typically straightforward for humans. Similarly, Wei et al. (Wei et al., 2022b) noted that while chain-of-thought (CoT) techniques stimulate human-like thought processes, they do not necessarily indicate genuine neural reasoning. Further illustrating these limitations, Tang et al. (Tang et al., 2023) reported that LLaMA-2 predominantly relied on template matching for reasoning tasks and lacked the ability to generalize beyond learned logic rules, a challenge exemplified in their studies using Symbolic Trees and ProofWriter. They questioned whether LLMs truly possess human-like inductive, deductive, and abductive reasoning capabilities. Adding to this discourse, Saparov and He (Saparov and He, 2022) introduced a synthetic question-answering dataset, PrOntoQA, to assess the logical reasoning abilities of LLMs. Their findings indicated that while LLMs can correctly perform individual deduction steps, they struggle with complex scenarios requiring the exploration of multiple valid deduction pathways.

3 Benchmark Construction

This section demonstrates the construction process of ContextHub on deductive logic and abductive logic. Deductive reasoning infers a logically certain conclusion from general statements, whereas abductive reasoning hypothesizes the most likely explanation based on observed data. As illustrated in Figure 1, constructing instantiated logical reasoning benchmarks involves three steps:

(1) Creating Formal Logical Reasoning Question Templates. We begin by developing formal logic templates for deductive and abductive reasoning across four levels of difficulty, utilizing the DyVal (Zhu et al., 2024) framework, which employs a tree structure to dynamically generate formal logic templates. These templates serve as the foundation for subsequent contextualization.

(2) Instantiation. Each logical template T is instantiated across 12 different domains drawn from Wikipedia categories, plus one domain with purely abstract variables. For each contextual domain, we randomly select a sub-category to ensure diversity, then instruct LLMs to contextualize the template T accordingly. An example instantiation can be found in Figure 2.

(3) Quality Control. To ensure the correctness of the instantiated logic problems, we implement a two-step quality control process. Initially, the samples are assessed by *Claude-3-Opus* which validates the samples against specified constraints.

Subsequently, human verification with 5 annotators is applied iteratively to refine the quality of generated questions.

3.1 Creating Formal Logical Reasoning Question Templates

We generate formal logic templates $\mathcal{X}$ based on the dynamic evaluation framework: DyVal (Zhu et al., 2024). For deductive and abductive logic, DyVal utilize tree structure to generate template samples on the fly with controllable difficulty. The tree structure naturally aligns with the inference process of a logic reasoning question. Take deductive logic as an example, the premises are given by the leaf nodes, where the intermediate nodes represent the intermediate inference steps, and the final result is shown by the root node.

Tree-based DyVal consists of three components: (1) **Constraint $\mathcal{C}$.** It aims to modulate the evaluation samples' complexity and validity. In our experiment, we define the complexity level of formal logic template by the depth of the generated tree. The validity constraints ensure the correctness of the generated formal logic template, for example, constrain the 'NOT' operation to have only one children node. (2) **Tree generation algorithm $\mathcal{G}$.** After defining the constraints, the generation algorithm $\mathcal{G}$ generates fixed complexity evaluation samples following the constraint $\mathcal{C}$, during the generation process, the final answer is also calculated automatically. (3) **Description function $\mathcal{F}$.** It translates each node in the graph into natural languages and finally forms all nodes into a formal logic template. For example, in deductive logic, a leaf node 'A' with truth value 'True' will be translated as "A is True.", a non-leaf node 'C' with 'OR' operation and its children 'A' and 'B' will be translated as "(A OR B) $\rightarrow$ C", where $\rightarrow$ means deductive operation.

3.2 Instantiation

The 13 instantiated domains include 12 domains based on Wikipedia's categorization in addition to a purely abstract instantiation. These Wikipedia domains spanning from Culture and the arts to Geography, Health, and Human activities. Each category is further subdivided into specific sub-categories $\mathcal{S}$ to ensure diverse contextual challenges. We exclude the "History and events" domain because its fact-based questions often bypass the need for logical reasoning. Further details about the domains and sub-categories can be found in Appendix D.

The instantiation process comprises two main transformations: (10 **Variable-based Transformation $\mathcal{T}_v$:** Each variable $\mathcal{V}$ in a template $\mathcal{X}$ is instantiated into a sentence specific to a chosen sub-category $\mathcal{S}$. For example, within the "Mathematics Education" sub-category, $\mathcal{V}$ might be instantiated as "All of Galois theory was developed by Galois alone." (2) **Template-based Transformation $\mathcal{T}_t$:** This step transforms the collection of instantiated sentences $\{s_{\mathcal{V}}\}$ into a coherent natural language query, preserving logical structure of original templates.

3.3 Quality Control

The quality verification of our instantiated benchmarks is managed using a hybrid model involving Claude-3-Opus, and a diverse panel of 5 human annotators. These verification steps are implemented to maintain a high standard of quality and relevance in our benchmarks, ensuring that they not only test logical reasoning but also engage with the domain knowledge in a meaningful way.

LLMs verification We deploy three primary checks to ensure the validity of our benchmarks: (1) *Common Sense Checking:* Ensures each problem requires logical inference rather than recognition of well-known facts. (2) *Sensibility Checking:* Ensures the scenarios and questions are coherent and clear, without internal contradictions or ambiguity. (3) *Tautology Checking:* Identifies any tautological statements, ensuring problems remain intellectually challenging and meaningful.

Human verification Five human annotators further assess the benchmarks to ensure quality and relevance: (1) *Template Adherence:* Verifies that the instantiated logic retains the structure and intent of the original templates, maintaining the fundamental logical framework. (2) *Fact Reckoning:* Checks that the problems necessitate logical deduction rather than simple fact recall, preserving the benchmark's complexity and educational value.

The human annotators[1] are required to answer these questions: (1) whether the contextualized question matches with the logic template (2) whether the contextualized question is against actual facts. Four rounds of annotation and re-

[1]The panel of annotators include four males and one female, each holding a Ph.D. degree in a diverse range of fields, including computer science, informatics, civil engineering, and medicine. The participants' ages ranged from 24 to 30, with an average age of 27.

view are conducted: in each round, every annotator manually reviewed a random sample of 220 questions (20 for each category) and identified questions that were not appropriately generated by the LLMs. Then all annotators participated in group review sessions to discuss potential issues with the generated questions and ways to improve the prompt design. After four rounds of verification and discussion, no further issues were identified in the questions generated by the LLMs across all categories. The following table 1 presents the accuracy of another randomly samples 220 questions for template adherence and fact reckoning:

data level	template adherence	fact reckoning
1	100%	95.45%
2	100%	94.54%
3	100%	96.36%
4	96.36%	93.18%

Table 1: Accuracy of Synthetically-generated Data

4 Experiment

Our experimental setup defines 4 levels of difficulty for logical reasoning tasks, with tree depths of $(2, 3, 4, 5)$ and a uniform width of 2 across nodes. The dataset comprises 10 deductive and 6 abductive formal logic templates at difficulty level 1, the maximum number of distinct templates of reasoning that can be generated at this level. We then include 40 deductive and 40 abductive reasoning formal logic templates at levels 2, 3, and 4, respectively. To ensure balanced data distribution, we assign equal counts of 'True', 'False', and 'N/A' truth values across all questions. The dataset spans 12 domains for contextualization, generating 5 unique instances per domain, resulting in $18,240$ total datapoints across various levels and types of logic.

The experiments are structured into two phases: benchmarking and fine-tuning. In the benchmarking phase, we assess model performance across various domains and compare between contextualized and abstract logic, to explore whether LLMs consistently grasp the underlying logical structures. The fine-tuning phase uses the generated data points to investigate how different instantiation types (abstract vs. contextualized) and data domains influence model generalization. This setup allows us to scrutinize the effects of model scaling and domain specificity on performance.

Benchmarking We benchmark the logical reasoning performance of several advanced models, including GPT series, Qwen-1.5 series (Bai et al., 2023), LLaMA-2 series (Touvron et al., 2023), LLaMA-3.1 series (Touvron et al., 2023), and Yi-1.5 series (Young et al., 2024). This evaluates whether models understand the underlying logic structures consistently, irrespective of contextual variations. We use an average weighted F_1 score for model evaluation, detailed in Appendix E.

Fine-tuning In the fine-tuning phase, we examine how different data types affect model generalization in logic reasoning. Models are fine-tuned using three settings: (1) solely with abstract logic data to test base logical reasoning capabilities. (2) with a sample of contextualized data across all domains to evaluate generalization across varied contexts. (3) with contextualized data from single domains to investigate the impacts of domain specificity and diversity. We utilize models ranging from Qwen1.5-0.5b to GPT-3.5-turbo and employ QLoRA for fine-tuning processes. Further details are available in Appendix F.

4.1 Results Analysis

In this subsection, we present a comprehensive analysis of the results obtained from our experiments. The benchmark results provide a general analysis of the performance trends, as well as a statistical analysis of the impact of the domain on model performance. Meanwhile, the fine-tuning results offer insights into the factors that influence model generalization for logic reasoning. By examining these two areas in detail, we aim to provide a thorough understanding of the behavior of large language models in the context of logic reasoning. Full benchmark results can be found in Appendix I.

4.1.1 Benchmarking

Overview of Model Performance Table 2 presents a selected benchmark result for the Qwen1.5 series, LLaMA-3.1 series, and GPT-4o models on deductive logic data across all 4 difficulty levels. The highest and lowest performance scores for each row are highlighted in bold and underlined, respectively. The full results are presented in Figure 8 in Appendix I, which demonstrate the varying performance by heatmap distributions.

At a granular level, GPT-4o models frequently appear to excel, particularly in higher difficulty levels. In contrast, smaller models like Qwen-0.5,

Diff. Level	LLM	Domain									
		Abs.	Culture	Geo.	Math	Sci.	People	Phi.	Religion	Tech.	Health
Level 1	Qwen1.5-0.5	34.15	38.27	38.81	32.26	**43.77**	32.83	36.64	34.92	36.54	30.30
	Qwen1.5-7	**72.55**	54.42	52.92	58.33	58.87	69.56	56.36	69.39	67.89	67.70
	LLaMA-3.1-8	78.31	87.65	85.93	85.11	82.76	**89.96**	76.27	84.17	83.48	88.51
	LLaMA-3.1-70	**100.00**	74.03	77.59	73.93	77.73	70.54	74.84	76.71	77.73	71.71
	Qwen1.5-72	76.88	**85.58**	82.29	68.69	80.39	76.02	82.31	74.56	74.78	74.00
	Qwen1.5-110	**90.94**	78.85	84.92	73.75	85.50	83.80	74.00	81.68	74.84	78.00
	LLaMA-3.1-405	**100.00**	84.91	80.56	71.87	80.34	85.57	74.69	86.47	72.14	81.26
	GPT-4o	**100.00**	88.04	81.45	67.37	81.15	84.59	79.53	83.90	89.50	86.48
Level 2	Qwen1.5-0.5	29.57	31.23	**34.45**	34.38	31.70	33.70	33.00	32.89	31.03	33.09
	Qwen1.5-7	59.53	**61.45**	52.50	45.22	48.73	44.44	53.09	48.84	47.07	61.07
	LLaMA-3.1-8	51.84	67.98	67.11	69.52	**77.12**	71.91	68.14	66.66	66.41	71.69
	LLaMA-3.1-70	**92.04**	69.22	68.80	68.58	76.72	76.75	71.76	68.76	62.87	75.89
	Qwen1.5-72	**81.16**	72.22	68.53	63.03	74.70	74.40	69.92	64.12	64.04	71.95
	Qwen1.5-110	65.53	65.96	68.57	63.43	69.04	**72.53**	66.09	65.21	60.93	67.80
	LLaMA-3.1-405	**97.95**	76.41	72.80	66.36	79.05	81.89	73.05	72.39	68.23	79.14
	GPT-4o	**98.50**	69.23	68.12	62.65	72.76	74.51	68.60	66.96	63.30	68.14
Level 3	Qwen1.5-0.5	25.39	34.04	33.96	33.57	31.90	33.11	32.51	**34.79**	33.96	32.98
	Qwen1.5-7	43.86	49.76	49.91	42.46	49.88	44.80	47.37	48.69	42.42	**50.38**
	LLaMA-3.1-8	51.48	55.04	54.26	50.43	60.70	54.71	59.13	**61.07**	49.01	56.99
	LLaMA-3.1-70	**74.71**	56.42	59.20	49.34	59.31	64.67	59.94	64.46	50.62	54.11
	Qwen1.5-72	**71.17**	63.10	58.52	53.90	55.55	60.47	58.00	54.98	54.28	56.48
	Qwen1.5-110	**60.37**	55.28	53.51	47.80	46.26	58.80	53.60	59.58	47.47	50.46
	LLaMA-3.1-405	**91.01**	64.13	58.63	51.14	65.51	63.03	59.47	63.32	51.31	63.15
	GPT-4o	**91.65**	52.30	50.90	46.96	56.56	61.96	51.75	59.58	45.83	56.26
Level 4	Qwen1.5-0.5	30.38	33.10	**35.46**	33.10	34.44	34.09	33.67	34.39	32.86	33.50
	Qwen1.5-7	51.49	56.02	55.48	51.48	57.00	**60.34**	54.64	50.48	51.68	49.19
	LLaMA-3.1-8	45.07	51.88	**55.18**	49.57	48.84	50.51	43.97	49.16	45.53	46.45
	LLaMA-3.1-70	**55.06**	50.92	48.05	42.27	45.72	52.01	45.41	37.05	41.93	45.61
	Qwen1.5-72	**57.41**	51.39	50.73	42.24	50.92	52.70	44.39	45.54	49.16	49.86
	Qwen1.5-110	49.57	**51.42**	47.35	48.13	46.65	49.72	45.85	47.08	48.58	49.45
	LLaMA-3.1-405	**76.94**	54.28	46.42	49.66	52.12	54.18	47.46	44.59	46.04	50.04
	GPT-4o	**83.51**	46.93	47.19	45.56	48.03	52.87	43.94	39.80	44.74	41.62

Table 2: Selected Experiment Results on Benchmarking

Qwen-7, LLaMA-3.1-8 often lag, struggling notably with abstract reasoning tasks. This disparity underscores the influence of model size. When aggregating results across all models and logic levels, certain domains consistently present more challenges. Specifically, the domains of Math and Philosophy appear to be the most demanding, likely due to their intrinsic requirement for deep logical structuring and abstract reasoning. Conversely, the domain labeled People generally shows the best performance, which is indeed less abstract, more intuitive, and features more contextual cues. The observed difference in performance across the different domains has been tested with statistical significance, the details of which are provided in the following section. *This indicates the significant impact that contexts can have on LLMs' logical reasoning performance.*

Influence of Model Size A pivotal observation from our data is the interaction between model size and sample type, presented in Figure 3. Larger models demonstrate a marked proficiency in abstract logical reasoning samples compared to their performance with their corresponding instantiated samples. This trend holds regardless of the difficulty level, suggesting that as models scale, their ability to decipher and apply abstract logic patterns improves significantly. While smaller models, like Qwen-0.5, LLaMA-2-7, LLaMA-3.1-8, demonstrate either better performance on instantiated samples than abstract samples or less difference between these two types. This discovery differs from previous observations (Tang et al., 2023; Saparov and He, 2022) which in general state that LLMs are better at instantiated data.

Inter-domain Disparities Further analysis of specific model performance within different do-

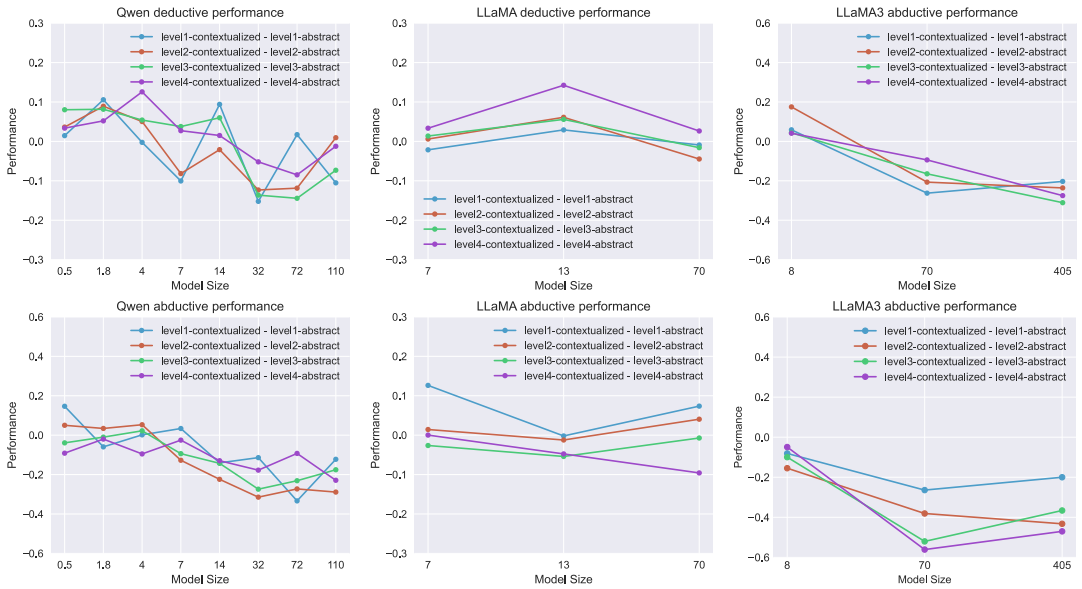

Figure 3: Abstract performance vs. Contextualized performance

mains reveals notable patterns. For abstract reasoning tasks, performance is highly variable: smaller models like Qwen-0.5 and Qwen-1.8 perform significantly worse, while larger configurations often excel. In the domain of Math, both Yi and Qwen series models exhibit consistently lower performance, reinforcing the notion of this domain's complexity and implying the extent to which logic reasoning performance can be influenced by the context. Interestingly, we also observe a general trend from observation where models that generally perform well show more pronounced disparities across domains, suggesting that higher capabilities amplify domain-specific challenges or advantages.

Statistical Analysis on Domain-specific Performance Difference The statistical results are presented in Figures 9 and 10 in Appendix I.1. Each row in either figure consists of four distinct subfigures. The two sub-figures on the left side illustrate the performance of the respective model for abductive reasoning, while the two on the right side demonstrate the deductive performance. In each pair of two sub-figures, the barplot shows the weighted F1-score for each category across difficulty levels calculated using equation (4), while the heatmap displays the results of the chi-square test (Bolboacă et al., 2011) with each cell corresponding to the p-value of the test regarding any pairwise categories. The application of the chi-square test in this regard aims to determine whether there is a significant association between two distributions. As shown in each heatmap, the darker blue ($p - value = 0.05$ at different thresholds) implies a significant difference between the distributions of the two categories, while the lightest blue ($p - value > 0.05$) suggests no significance.

Based on the bar plots and heatmaps, there are several observations to highlight in terms of models' performance. *First*, most of these models perform better in deductive reasoning tasks than abductive reasoning tasks. This observed pattern is consistent across most of the models and categories under investigation in this study. *Second*, the models' performance varies significantly across different categories. For instance, based on the results of Qwen-32, the weighted F1-score for the abstract category is much higher than that of other categories. However, it is also noted that the math category consistently displays a comparatively lower weighted F1-score across these categories. *Third*, the abstract category is more likely to display significant differences when compared to other models.

4.1.2 Generalization after Finetuning

This segment of our research is dedicated to examining the generalization capabilities of models fine-tuned on different types of logic data. We address several key research questions: (1) How do models trained on abstract data compare in generalization to those trained on contextualized data? (2) What is the effect of model scaling on generalization performance across different data types? (3) How well can models trained on data from a single domain generalize across multiple domains, and what role does the diversity of training domains play in this context?

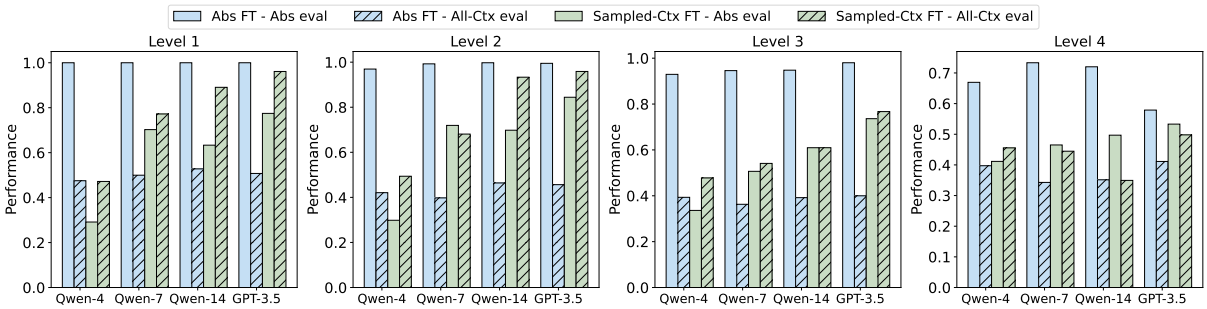

Figure 4: Comparative Generalization Performance on Abstract and Contextualized Data.

Abstract vs. Contextualized Data We conducted experiments to compare the generalization abilities of models trained on purely abstract data against those trained on contextualized data. We utilized four models for this purpose: Qwen-4, Qwen-7, Qwen-14, and GPT-3.5-turbo, each fine-tuned on both types of data. The abstract dataset comprised 1280 data points generated from 256 formal logic templates. For a balanced comparison, we selected a random sample of 1280 data points from a much larger pool of contextualized data, ensuring each template was equally represented.

Our results in Figure 4 show that models fine-tuned on abstract data, while performing well on similar abstract test cases, exhibit a marked decline in performance when applied to contextualized data, particularly as the difficulty level increases. On the other hand, models fine-tuned on contextualized data (sampled-ctx) demonstrate robust generalization capabilities, significantly outperforming those trained on abstract data across both similar and dissimilar tasks. This is especially evident in GPT-3.5, achieving near-perfect scores on the lower difficulty levels of contextualized data.

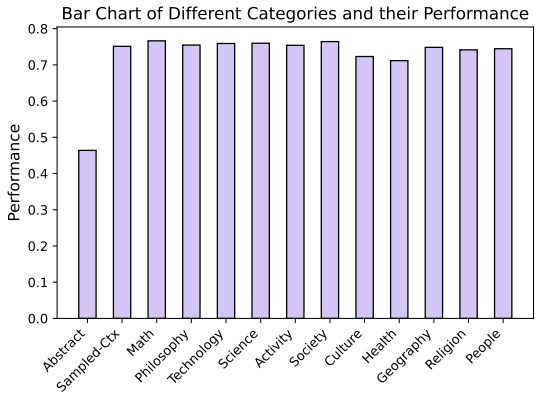

Figure 5: Model Performance by Finetuning on Different Domains.

Model Scale Effect on Generalization Our study also explores how the size of models influences their ability to generalize from training data. In Figure 4, we observe that while larger models exhibit only slight improvements when fine-tuned on abstract data, suggesting a saturation point in the complexity that abstract reasoning can model, the same models show significantly better performance improvements when fine-tuned on sampled-ctx data. This suggests that the richness of contextualized data may better support the models in learning to generalize across various logic tasks.

Single-domain vs. Multi-domain Generalization One potential explanation for the superior generalization capacity of sampled-ctx data is the diversity of domains from which it is sampled, as opposed to the relative homogeneity of abstract data. Thus, further investigations focused on whether training on data from a single domain could match or exceed the generalization capabilities achieved by training across multiple domains. We fine-tuned GPT-3.5 on individual domain datasets as well as on a mixed-domain dataset (sampled-ctx). Results in Figure 5 indicate that models trained on single-domain datasets often perform on par with or better than those trained on multi-domain data, challenging the assumption that greater domain diversity improves generalization.

5 Conclusion

This paper provides a comprehensive investigation into the logic reasoning abilities of LLMs through the ContextHub benchmark. Our approach effectively separates logical reasoning from textual comprehension, enabling a focused analysis of models' reasoning capabilities. The study reveals that a model's performance on reasoning tasks is substantially influenced by the context and domain-specific variables involved. Also, finetuning on instantiated data enhances the models' ability to generalize across various logic reasoning tasks, irrespective of the domain complexity or the diversity of domains.

6 Limitations

Our study has several limitations. First, the synthetic nature of the datasets, particularly abstract logic data, may not fully capture the complexity and variability of real-world reasoning tasks. This limitation could affect the external validity of our findings, as models trained on such data might not perform equivalently on natural, less structured tasks. Second, the study's focus on propositional logic might not translate directly to other forms of reasoning used in practical applications, such as probabilistic or causal reasoning. Future research should aim to address these limitations by incorporating more diverse and complex reasoning formats, extending beyond propositional logic to include other reasoning types that are prevalent in real-world scenarios.

6.1 Template-Based Data Generation and Limited Complexity of Logical Rules

A final area of concern relates to the *templatized* nature of our dataset and the relative simplicity of the inference rules involved, which some argue do not fully capture real-life reasoning:

Modeling Real-World Reasoning. While our dataset leverages structured templates, we note that it primarily seeks to evaluate whether the inherent reasoning abilities of LLMs generalize beyond strictly abstract logic. We design contexts that, although templated, reflect more naturalistic scenarios than purely formal problems. This design allows us to isolate the effects of contextual information on reasoning performance—an aspect often overlooked by benchmarks that focus either on entirely decontextualized logic or on unconstrained real-world data without rigorous comparisons. We acknowledge that fully simulating the variety and complexity of real-world language remains challenging; however, our approach provides an intermediate step that systematically tests reasoning in contexts richer than purely symbolic logic.

Potential for "Hacking" Templatized Data. Because the data is templated, one might worry that targeted fine-tuning could exploit superficial patterns. Yet, our experiments indicate that such fine-tuning does not close the performance gap between reasoning in abstract contexts and reasoning under richer, diverse circumstances, suggesting that the dataset's complexity is non-trivial. Furthermore, our generation pipeline can flexibly produce new templates or extended contexts, preserving diversity and countering the risk of overfitting. Thus, while we acknowledge that carefully crafted templates can, in principle, be exploited, our design aims to mitigate this vulnerability through continual expansion and variation of the dataset.

Scope of Logical Rules. Our benchmark focuses on relatively fundamental inference rules (*e.g.*, Modus Ponens) rather than the more intricate constructions found in advanced logic benchmarks (e.g., LogicBench (Parmar et al., 2024)). We select simpler rules to illuminate a primary objective: examining how contextual information affects model performance when abstract reasoning appears achievable in controlled settings. High performance on purely abstract tasks highlights an unexpected deficit once the same rules are embedded in richer contexts. We acknowledge that extending the benchmark to incorporate more complex rules remains a worthwhile direction, but our current approach demonstrates that even foundational inference (like Modus Ponens) can deteriorate significantly with the introduction of contextual complexity, underscoring the ongoing challenges for LLMs in real-world-like reasoning tasks.

References

Wikipedia categories.

Aitor Arrieta, Miriam Ugarte, Pablo Valle, José Antonio Parejo, and Sergio Segura. 2025. o3-mini vs deepseek-r1: Which one is safer? *arXiv preprint arXiv:2501.18438*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

BIG bench authors. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Sorana D Bolboacă, Lorentz Jäntschi, Adriana F Sestraş, Radu E Sestraş, and Doru C Pamfil. 2011. Pearson-

fisher chi-square statistic revisited. *Information*, 2(3):528–545.

Shan Chen, Yingya Li, Sheng Lu, Hoang Van, Hugo JWL Aerts, Guergana K Savova, and Danielle S Bitterman. 2024. Evaluating the chatgpt family of models for biomedical reasoning and classification. *Journal of the American Medical Informatics Association*, 31(4):940–948.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2020. Transformers as soft reasoners over language. *arXiv preprint arXiv:2002.05867*.

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

Ishita Dasgupta, Andrew K Lampinen, Stephanie CY Chan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. 2022. Language models show human-like content effects on reasoning. *arXiv preprint arXiv:2207.07051*.

Maksym Del and Mark Fishel. 2022. True detective: a deep abductive reasoning benchmark undoable for gpt-3 and challenging for gpt-4. *arXiv preprint arXiv:2212.10114*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Lizhou Fan, Wenyue Hua, Lingyao Li, Haoyang Ling, and Yongfeng Zhang. 2023a. Nphardeval: Dynamic benchmark on reasoning ability of large language models via complexity classes. *arXiv preprint arXiv:2312.14890*.

Lizhou Fan, Wenyue Hua, Xiang Li, Kaijie Zhu, Mingyu Jin, Lingyao Li, Haoyang Ling, Jinkui Chi, Jindong Wang, Xin Ma, et al. 2024. Nphardeval4v: A dynamic reasoning benchmark of multimodal large language models. *arXiv preprint arXiv:2403.01777*.

Lizhou Fan, Sara Lafia, Lingyao Li, Fangyuan Yang, and Libby Hemphill. 2023b. Datachat: Prototyping a conversational agent for dataset search and visualization. *arXiv preprint arXiv:2305.18358*.

Deqing Fu, Ghazal Khalighinejad, Ollie Liu, Bhuwan Dhingra, Dani Yogatama, Robin Jia, and Willie Neiswanger. 2024. Isobench: Benchmarking multimodal foundation models on isomorphic representations. *arXiv preprint arXiv:2404.01266*.

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah Goodman. 2024. Understanding social reasoning in language models with language models. *Advances in Neural Information Processing Systems*, 36.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, et al. 2024. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36.

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*.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Wenfei Zhou, James Coady, David Peng, Yujie Qiao, Luke Benson, et al. 2022. Folio: Natural language reasoning with first-order logic. *arXiv preprint arXiv:2209.00840*.

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

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

Wenyue Hua, Lizhou Fan, Lingyao Li, Kai Mei, Jianchao Ji, Yingqiang Ge, Libby Hemphill, and Yongfeng Zhang. 2023. War and peace (waragent): Large language model-based multi-agent simulation of world wars. *arXiv preprint arXiv:2311.17227*.

Mingyu Jin, Beichen Wang, Zhaoqian Xue, Suiyuan Zhu, Wenyue Hua, Hua Tang, Kai Mei, Mengnan Du, and Yongfeng Zhang. 2024a. What if llms have different world views: Simulating alien civilizations with llm-based agents. *arXiv preprint arXiv:2402.13184*.

Mingyu Jin, Qinkai Yu, Dong Shu, Chong Zhang, Lizhou Fan, Wenyue Hua, S Zhu, Y Meng, Z Wang, M Du, et al. 2024b. Health-llm: Personalized retrieval-augmented disease prediction system. *arXiv preprint arXiv:2402.00746*.

Lingyao Li, Lizhou Fan, Shubham Atreja, and Libby Hemphill. 2024a. "hot" chatgpt: The promise of chatgpt in detecting and discriminating hateful, offensive, and toxic comments on social media. *ACM Transactions on the Web*, 18(2):1–36.

Yanda Li, Dixuan Wang, Jiaqing Liang, Guochao Jiang, Qianyu He, Yanghua Xiao, and Deqing Yang. 2024b. Reason from fallacy: Enhancing large language models' logical reasoning through logical fallacy understanding. *arXiv preprint arXiv:2404.04293*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan,

Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*.

Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. 2024. Can large language models reason about medical questions? *Patterns*, 5(3).

Shuhang Lin, Wenyue Hua, Lingyao Li, Che-Jui Chang, Lizhou Fan, Jianchao Ji, Hang Hua, Mingyu Jin, Jiebo Luo, and Yongfeng Zhang. 2024. Battleagent: Multi-modal dynamic emulation on historical battles to complement historical analysis. *arXiv preprint arXiv:2404.15532*.

Terufumi Morishita, Gaku Morio, Atsuki Yamaguchi, and Yasuhiro Sogawa. 2023. Learning deductive reasoning from synthetic corpus based on formal logic. In *International Conference on Machine Learning*, pages 25254–25274. PMLR.

Liangming Pan, Alon Albalak, Xinyi Wang, and William Yang Wang. 2023. Logic-lm: Empowering large language models with symbolic solvers for faithful logical reasoning. *arXiv preprint arXiv:2305.12295*.

Mihir Parmar, Nisarg Patel, Neeraj Varshney, Mutsumi Nakamura, Man Luo, Santosh Mashetty, Arindam Mitra, and Chitta Baral. 2024. Logicbench: Towards systematic evaluation of logical reasoning ability of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13679–13707.

Abulhair Saparov and He He. 2022. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. *arXiv preprint arXiv:2210.01240*.

Tomohiro Sawada, Daniel Paleka, Alexander Havrilla, Pranav Tadepalli, Paula Vidas, Alexander Kranias, John J Nay, Kshitij Gupta, and Aran Komatsuzaki. 2023. Arb: Advanced reasoning benchmark for large language models. *arXiv preprint arXiv:2307.13692*.

Xiaojuan Tang, Zilong Zheng, Jiaqi Li, Fanxu Meng, Song-Chun Zhu, Yitao Liang, and Muhan Zhang. 2023. On the paradox of generalizable logical reasoning in large language models.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Karthik Valmeekam, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. 2022. Large language models still can't plan (a benchmark for llms on planning and reasoning about change). *arXiv preprint arXiv:2206.10498*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022a. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Kaijie Zhu, Jiaao Chen, Jindong Wang, Neil Zhenqiang Gong, Diyi Yang, and Xing Xie. 2024. Dyval: Graph-informed dynamic evaluation of large language models. *International Conference on Learning Representations*.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

A Appendix

B Licenses for Existing Assets

All the methods we used for comparison are publicly available for academic usage.

C Environments and Resources

To reproduce the computational environment used in this study, an environment file, environment.yml, is provided in our code repository. This YAML file lists all the dependencies and their specific versions used in the study. Users can create an identical Conda environment using the command conda env create -f environment.yml. The computational experiments were conducted on machines equipped

with NVIDIA Tesla A100 GPUs (80GB of GPU memory each).

D Formal Logic Template Instantiation

D.1 Contextual Instantiation

Domains of contextualization We instantiate the above formal logic templates in the below contextual domains following the categorization of Wikipedia (Wik):

```
Culture and the arts, Geography and
    places, Health and fitness, Human
    activities, Mathematics and logic,
    Natural and physical science, People
     and self, Philosophy and thinking,
    Religion and belief systems, Society
     and social sciences, Technology and
     applied sciences.
```

Listing 1: Categories of Wikipedia

The domain of "History and events" is removed because instantiated sentences are often about known fact and the question can be answer without going through the reasoning process, thus nullify the reasoning problem.

Each instantiation of a domain is created based on a randomly selected sub-categories in the domain from above based on sub-categories established in Wikipedia to encourage diversity and specification. For example, "Culture and the arts" has the following sub-categories:

```
Classics, Critical theory, Cultural
    anthropology, Clothing, Folklore,
    Food and drink culture, Language,
    Literature, Museology, Mythology,
    Philosophy, Popular culture, Science
     and culture, Traditions, Arts and
    crafts, Celebrity, Censorship in the
     arts, Festivals, Humor, Literature,
     Museums, Parties, Poetry, Circuses,
     Dance, Film, Music, Opera,
    Storytelling, Theatre, Architecture,
     Comics, Crafts, Design, Drawing,
    Film Animation, New media art,
    Painting, Photography, Sculpture,
    Board games, Card games, Dolls,
    Puppetry, Puzzles, Role-playing
    games, Video games, Air sports,
    American football, Association
    football, Auto racing, Baseball,
    Basketball, Boating, Boxing,
    Canoeing, Cricket, Cycling, Exercise
    , Fishing, Golf, Gymnastics, Hobbies
    , Horse racing, Ice hockey, Lacrosse
    , Olympic Games, Rugby league, Rugby
     union, Sailing, Skiing, Swimming,
    Tennis, Track and field, Walking
    trails, Water sports, Whitewater
    sports
```

Listing 2: Sub-categories of Culture and the arts in Wikipedia

Contextualization process After obtaining the formal logic templates, for each domain, we first randomly selected one sub-category c, then we ask LLMs (in our experiment, Claude-3-Opus) to instantiate each variable in the original logic templates with the relevant context in the selected sub-category. This contextualizatin process is divided into 2 steps:

1. Variable-based Transformation: $\mathcal{T}_v$. For each variable $\mathcal{V}$ contained in the logic template $\mathcal{X}$, a instantiated sentence $s_{\mathcal{V}}$ is generated by $\mathcal{T}_v(c, \mathcal{V} \in \mathcal{X})$. For example, a leaf node variable $\mathcal{V}$ can be instantiated as "Alice studied hard for the following math test" in the sub-category of "Mathematics Education" in the category of "Mathematics and Logic".

2. Template-based Transformation: $\mathcal{T}_t$. After generating $\{s_{\mathcal{V}}\}$ for all $\{\mathcal{V}\}$ in the template $\mathcal{X}$, a coherent natural language description will be generated by $\mathcal{T}_t(\{s_{\mathcal{V}}\}, \mathcal{X})$ by forming a instantiated version of the original formal logic template.

E Evaluation Metrics for Benchmarking

To assess the reasoning capabilities of LLMs, we employ the average F1 score. The calculation of the average F1 score involves determining the average of the F1 scores for data points (d) that possess the same truth values. For datapoints with identical ground truth (gt) truth value $\mathcal{T}$, the F1 score is computed by first ascertaining the true positive ($T_p^{\mathcal{T}}$), false positive($F_p^{\mathcal{T}}$), and false negative($F_n^{\mathcal{T}}$):

$$T_p^{\mathcal{T}} = \{d \in D | f(d) = gt(d), gt(d) = \mathcal{T}\} \quad (1)$$

$$F_p^{\mathcal{T}} = \{d \in D | f(d) \neq gt(d), f(d) = \mathcal{T}\} \quad (2)$$

$$F_n^{\mathcal{T}} = \{d \in D | f(d) \neq gt(d), gt(d) = \mathcal{T}\} \quad (3)$$

$F_1^{\mathcal{T}}$ for for datapoints with the truth value $\mathcal{T}$ is then computed by:

$$F_1^{\mathcal{T}} = \frac{2T_p^{\mathcal{T}}}{2T_p^{\mathcal{T}} + F_p^{\mathcal{T}} + F_n^{\mathcal{T}}} \quad (4)$$

The average F1 score for the entire dataset is calculated by determining the average of the F1$^{\mathcal{T}}$ scores for all possible truth values $\mathcal{T}$.

F Hyperparameters for Finetuning

We leverage QLora (Dettmers et al., 2024) for finetuning on open-source models. Other relevant hyperparameters are: epochs = 1, warmup proportion = 0.01, learning rage = 3e-4, weight decay = 0.01, lora rank = 64, lora dropout = 0.05, lora alpha = 16, batch size = 4, accumulate gradient steps = 8.

F.1 Abstract Instantiation

Other than the 11 contextual domains from Wikipedia, we also create an "abstract" domain where we simply substitute by heuristic rules the propositional variables in the formal logic template with arbitrary character sequences of varying lengths, ranging from 3 to 5. The purpose of creating this domain is to augment the number of datapoints expressed in an abstract form, thereby enabling a fair comparison with other contextualized domains in terms of sample size. Furthermore, by employing multiple instantiations, we can mitigate the impact of any potential outliers and obtain a more reliable and generalizable estimate of the performance of abstract data as we have only 256 formal logic templates in total.

Below is an example of instantiation in Table 3 of a formal logic template of difficulty level 1, where propositional variables are represented by strings such as "aaa", "aab", and "aac".

```
(aaa or aab) - aac.  Given aac is False,
    what is the value of aab?
```

We provide an abstract instance and a contextualized instance on the domain of "Geography and Places", where we provide the instantiations of each proposition in the template and the final combined logic reasoning task based on propositional instantiations. More examples can be found in Appendix J.

G Length Correlation

It is possible that some may question whether the observed differences in model performance are due to the varying input lengths, rather than the effect of different instantiations. To address this potential concern, we have conducted a series of experiments to investigate the correlation between input length and model performance: we employ four models of varying sizes (Qwen-0.5, Qwen-7, Qwen-32, Qwen-110) and conduct length-based performance ablation. We then analyze the performance of each model based on the length of the input text. Specifically, for each graph presented below, the x-axis represents the text length, while the y-axis represents the corresponding model performance. The y-axis value for each x-axis value x is the model performance on the part of the data whose corresponding input text length is smaller than x.

Based on the two images Figure 6 and Figure 7, we cannot see any consistent correlation between model performance and input length after tokenization using model corresponding tokenizer.

H Error Analysis

We carefully selected 970 error samples of 5544 on GPT-4o's result on abductive and deductive level 1 and 2 datasets. We identified 4 different reasoning errors and have updated the manuscripts to elaborate them.

- Fail to reason on counter-factual questions (24.94%): In the example: If the Earth's oceans are warming due to climate change, or humans are emitting large amounts of greenhouse gases, then global temperatures will rise. Given that the statement "global temperatures are not rising" is false, the correct logical answer should be False—meaning that humans emitting large amounts of greenhouse gases is not true. However, the model might assume this outcome due to its pre-existing knowledge, leading to an incorrect conclusion of True.

- Laziness and shallow thinking (37.11%): The model occasionally exhibits a lazy approach, particularly when dealing with complex premises or pre-conditions. Instead of thoroughly analyzing the situation, it tends to give a "N/A" or a simplified response without fully utilizing all the available information.

- Inconsistent adherence to logical expressions (12.47%): At times, the model fails to follow basic logical rules in simple expressions. For example, in an "aaa OR aab" scenario, where either being true should lead to "aac" being true as well, the model sometimes incorrectly concludes otherwise. This suggests a lapse in the model's ability to consistently apply fundamental logical reasoning.

- Weak understanding of contrapositive logic (25.48%): The model consistently struggles with understanding and applying contrapositive logic (e.g., understanding that "If P, then Q" logically implies "If not Q, then not P").

Level 1 - Abstract	Level 1 - Geography and Places
aaa: vxkgr aab: caunc aac: ybyz reasoning task: (vxkgr or caunc) $\rightarrow$ ybyz. Given ybyz is False, what is the value of caunc?	aaa: The terrain has experienced significant uplift. aab: Powerful erosional forces have shaped the land. aac: The area features tall, steep mountains. reasoning task: If an area of land has experienced significant uplift or been shaped by powerful erosional forces, then the terrain will feature tall, steep mountains. Given that the area does not have tall, steep mountains, can it be determined if powerful erosional forces have shaped the land?

Table 3: Example of abstract and instantiated logic reasoning task based on the original formal logic template.

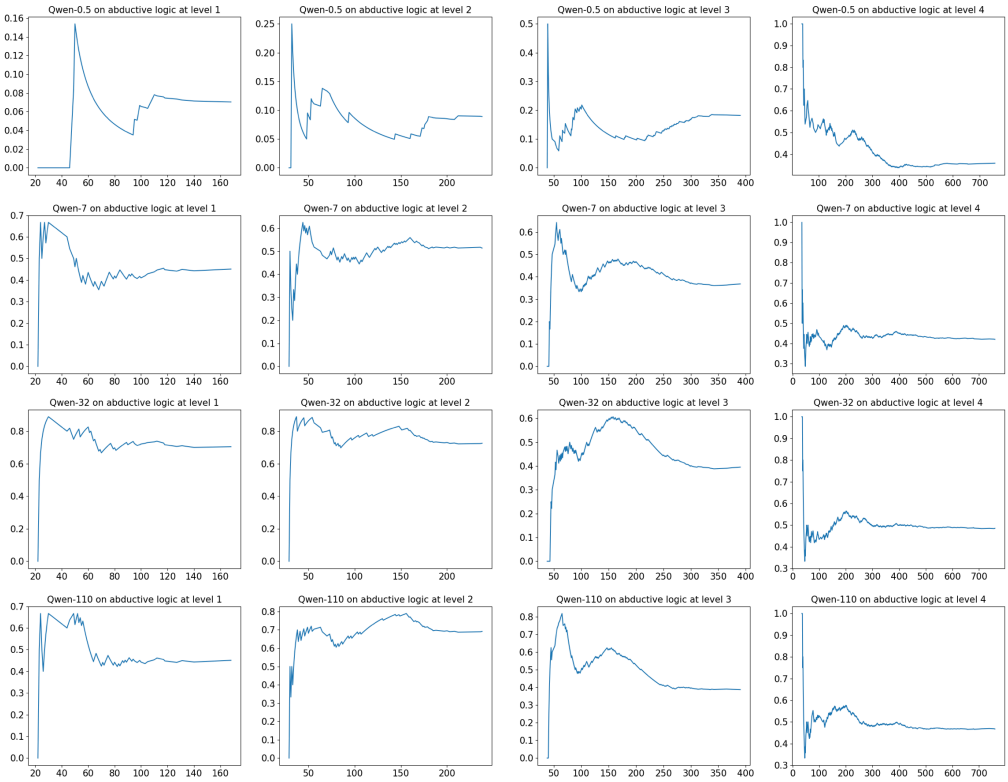

Figure 6: Length-based performance collection on abductive logic. The four rows correspond to four models, and four columns correspond to four difficulty levels.

I Main Benchmark Result

I.1 Statistical Analysis Result

Figure 7: Length-based performance collection on deductive logic. The four rows correspond to four models, and four columns correspond to four difficulty levels.

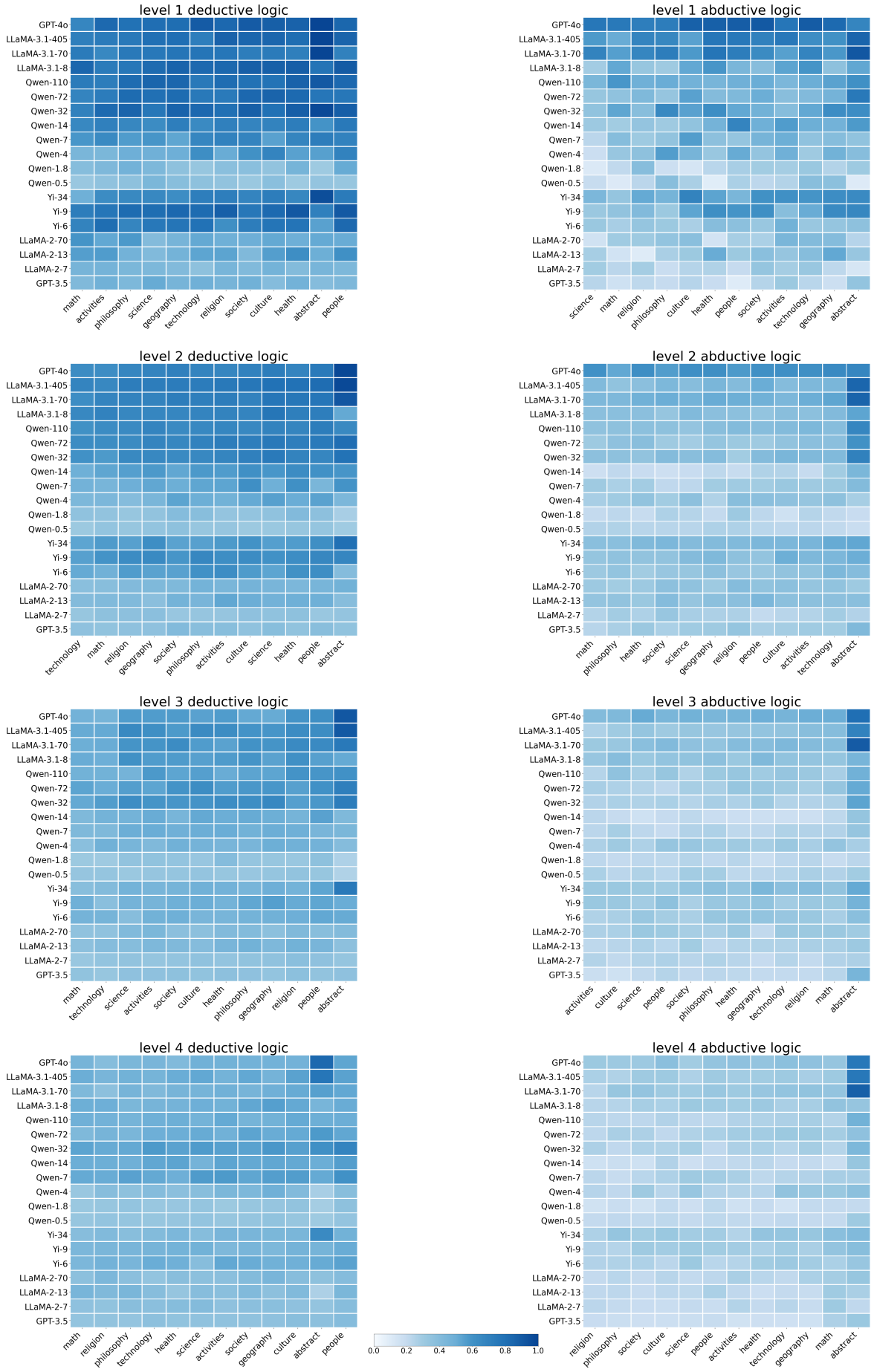

Figure 8: Main Benchmark Performance

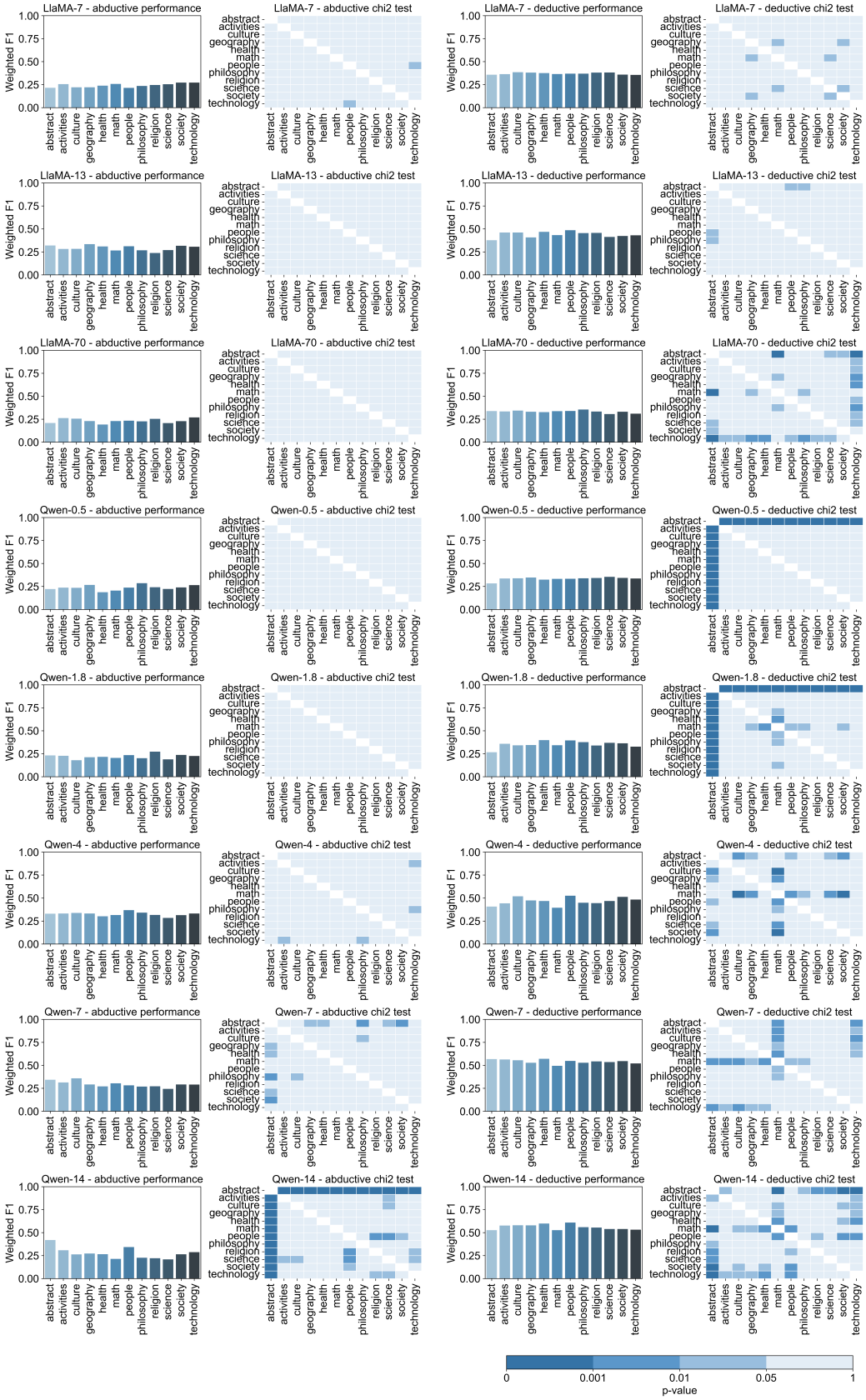

Figure 9: Results of weighted F1-score and Chi-square test

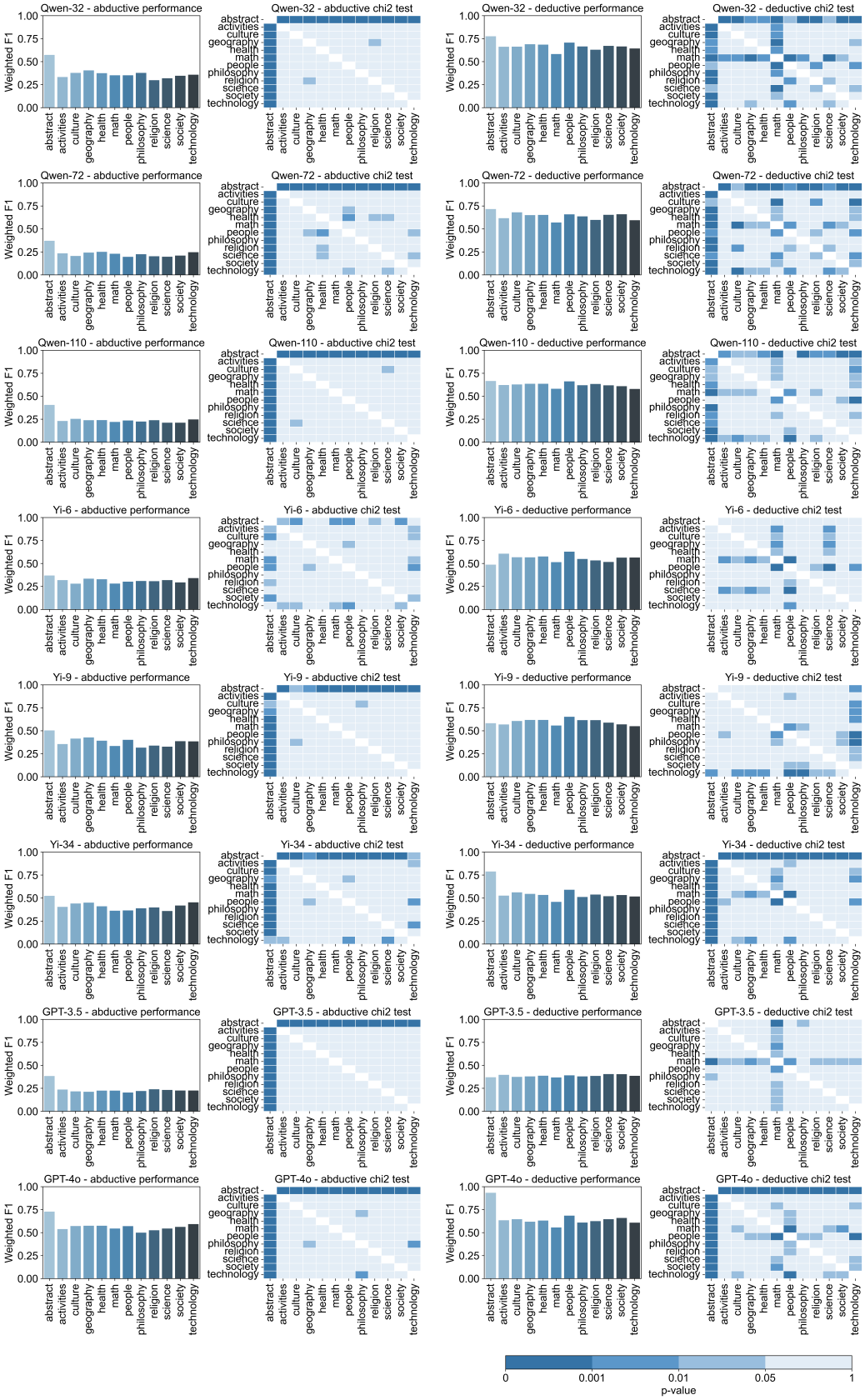

Figure 10: Results of weighted F1-score and Chi-square test (Cont.)

J Data Examples

The following table presents several examples showing abductive and deductive reasoning with their respective difficulty levels and domains. The left column shows examples of abstract instantiations, while the right column shows contextually instantiated examples in specific domains.

Table 4: Examples of abductive and deductive reasoning.

Abstract Example	Specific Domain Contextualized Example
	Abductive Reasoning
Level 1 - Abstract	Level 1 - Geography and Places
aaa: vxkgr aab: caunc aac: ybyz reasoning task: (vxkgr or caunc) - ybyz. Given ybyz is False, what is the value of caunc?	aaa: The terrain has experienced significant uplift. aab: Powerful erosional forces have shaped the land. aac: The area features tall, steep mountains. reasoning task: If an area of land has experienced significant uplift or been shaped by powerful erosional forces, then the terrain will feature tall, steep mountains. Given that the area does not have tall, steep mountains, can it be determined if powerful erosional forces have shaped the land?
Level 2 - Abstract	Level 2 - Mathematics and Logic
aaa: ttjmx aab: kottz aac: wqeq aad: mnze aae: zkx aaf: pofk reasoning task: (wqeq or mnze) - zkx. (NOT ttjmx) - kottz. (kottz or zkx) - pofk. Given pofk is False, what is the value of ttjmx?	aaa: The prior probability is a uniform distribution aab: The prior probability expresses existing beliefs about the parameters. aac: A prior probability distribution is specified. aad: New data is collected. aae: The posterior probability is calculated using Bayes' theorem. aaf: The posterior probability provides an improved estimate of the parameters. reasoning task: In Bayesian statistics, if a prior probability distribution is specified or new data is collected, then the posterior probability can be calculated using Bayes' theorem to update the probability based on the new evidence. If the prior probability is not a uniform distribution, then it expresses existing beliefs or knowledge about the values of the parameters. If the prior probability expresses existing beliefs or the posterior probability is calculated, then there is sufficient information to update the probability distribution. Given that the statement "The posterior probability provides an improved estimate of the parameters" is false, can it be determined whether the prior probability is a uniform distribution or not?
Level 3 - Abstract	Level 3 - Technology and Applied Sciences
aaa: dmacf aab: my aac: qnvj aad: lxnf aae: jf aaf: ors aag: kuyl aah: jal aai: rqo aaj: vrmxo aak: mcwe aan: pdzyf aao: guwls aap: xjgwm	aaa: Regular vulnerability scans are performed. aab: Penetration testing is conducted quarterly. aac: Security weaknesses are proactively identified. aad: Operating systems are up to date with patches. aae: Antivirus software is installed on all computers. aaf: Endpoint devices are protected. aag: The overall attack surface is minimized. aah: The firewall is properly configured. aai: Intrusion detection systems are active. aaj: The network perimeter is secure. aak: Employees have completed security training. aan: Security policies are strictly enforced. aao: Employees follow secure computing practices. aap: Internal systems and data are well-defended.

Continued

Table 4 – continued from previous page

Abstract Example	Specific Domain Contextualized Example
aaq: vv reasoning task: (wqeq or mnze) - zkx. (NOT ttjmx) - kottz. (kottz or zkx) - pofk. Given pofk is False, what is the value of ttjmx?	aaq: The organization has strong cybersecurity posture. reasoning task: If the firewall is properly configured or intrusion detection systems are active, then the network perimeter is secure. When employees have completed security training and security policies are strictly enforced, it implies that employees follow secure computing practices. If the network perimeter is secure or employees follow secure practices, then internal systems and data are well-defended. Having up-to-date operating systems with the latest patches or antivirus software installed on all computers means the endpoint devices are protected. Performing regular vulnerability scans or conducting quarterly penetration testing allows security weaknesses to be proactively identified. If security weaknesses are proactively identified or endpoint devices are protected, then the overall attack surface is minimized. When the attack surface is minimized and internal systems and data are well-defended, it indicates the organization has a strong cybersecurity posture. Given that the organization does not have a strong cybersecurity posture, can it be determined if operating systems are up to date with patches?
Level 4 - Abstract	Level 4 - Culture and Arts
aaa: cg	aaa: Sophie cannot practice her beam routine.
aab: ysjeo	aab: Sophie needs to prepare new skills.
aac: uby	aac: Sophie requires dedicated practice time.
aad: vwwf	aad: The springboard is broken.
aae: lj	aae: The vault is not stable.
aaf: qd	aaf: Performing vault runs is risky.
aag: miz	aag: Sophie is not able to practice effectively.
aah: tfxbc	aah: Sophie's coach is at practice.
aai: aaw	aai: Sophie does not have supervision.
aaj: oftr	aaj: Sophie is allowed to train.
aak: fzsq	aak: Sophie is not making progress in her gymnastics.
aan: yxt	aan: The balance beam is set up properly.
aao: ln	aao: Sophie cannot practice her beam routine.
aap: qa	aap: The uneven bars are not at the correct height.
aaq: py	aaq: Sophie cannot work on her bar skills.
aar: qe	aar: Sophie faces a major hindrance to her practice.
aas: ng	aas: The floor mat has tears and needs to be replaced.
aat: bhjb	aat: The floor area is not large enough for a full floor routine.
aau: djay	aau: It is unsafe for Sophie to practice floor exercises.
aav: pvize	aav: With an upcoming competition, Sophie needs to prepare new skills.
aaw: tk	aaw: Sophie does not have enough energy to train effectively.
aax: vod	aax: Sophie's gymnastics career is at risk.
aay: dngja	aay: Sophie's gymnastics performance will likely be impacted negatively.
aaz: ozyue	aaz: Sophie may need to consider withdrawing from competitions.

Continued

Table 4 – continued from previous page

Abstract Example	Specific Domain Contextualized Example
reasoning task: (NOT yxt) -> ln. (NOT qa) -> py. (ln or py) -> qe. (ng or bhjb) -> djay. (vwwf or lj) -> qd. (NOT tfxbc) -> aww. (NOT aww) -> oftr. (NOT pvize) -> tk. (cg or ysjeo) -> uby. (uby or qd) -> miz. (miz or oftr) -> fzsq. (djay or tk) -> vod. (qe or vod) -> dngja. (fzsq or dngja) -> ozyue. Given ozyue is False, what is the value of yxt?	reasoning task: The balance beam not being set up properly means Sophie cannot practice her beam routine. Similarly, if the uneven bars are not at the correct height, Sophie cannot work on her bar skills. If Sophie is unable to train on at least one apparatus, she faces a major hindrance to her practice. Torn floor mats needing replacement or insufficient floor space makes it unsafe for Sophie to practice floor exercises. A broken springboard or unstable vault makes performing vault runs risky. If it is unsafe to practice floor or vault exercises, Sophie cannot train safely or productively. Sophie's coach not being at practice means she does not have supervision. Having supervision allows Sophie to train. If Sophie did not fuel properly before practice, she will not have enough energy to train effectively. With an upcoming competition, Sophie needs to prepare new skills, requiring dedicated practice time. If Sophie's training is compromised by risky apparatus or lack of practice time, she will not be able to practice effectively. If Sophie's training session is unproductive or she faces major hindrances, then she is not making progress in her gymnastics. Lack of progress or likely negative performance impacts put Sophie's gymnastics career at risk. Given that Sophie is not considering withdrawing from competitions, what can be determined about the balance beam being set up properly?
	Deductive Reasoning
Level 1 - Abstract	Level 1 - Natural and Physical Sciences
aaa: pusvu aab: hs aac: ivl reasoning task: pusvu is True. hs is False. (pusvu or hs) - ivl. Deduce the result of ivl.	aaa: A cold front is approaching the region aab: A warm air mass is stagnant over the area aac: Atmospheric instability is likely to develop reasoning task: A cold front is approaching the region, but there is no warm air mass stagnant over the area. If a cold front approaches or a warm air mass is stagnant, then atmospheric instability is likely to develop. Can we say that atmospheric instability will likely develop in this scenario?
Level 2 - Abstract	Level 2 - Society and Social Sciences
aaa: jd aab: bfk aac: wng aad: vko aae: cva aaf: qymwa aag: cr	aaa: John Lee was born in the United States aab: John Lee's parents immigrated from South Korea aac: John Lee has Korean ancestry aad: The Lee family speaks Korean fluently aae: The Lee family identifies as Korean-American aaf: The Lee family has a connection to Korean culture aag: John Lee is considered Korean-American

Continued

Table 4 – continued from previous page

Abstract Example	Specific Domain Contextualized Example
reasoning task: cva is True. vko is False. (vko or cva) - qymwa. jd is True. bfk is True. (jd or bfk) - wng. (wng and qymwa) - cr. Deduce the result of cr.	reasoning task: The Lee family identifies as Korean-American, but they do not speak Korean fluently. If the Lee family speaks Korean fluently or identifies as Korean-American, then they have a connection to Korean culture. John Lee was born in the United States, and his parents immigrated from South Korea. If John Lee was born in the U.S. or his parents immigrated from South Korea, then he has Korean ancestry. If John Lee has Korean ancestry and his family has a connection to Korean culture, then he is considered Korean-American. Can we conclude that John Lee is considered Korean-American based on the given information?
Level 3 - Abstract	Level 3 - Culture and Arts
aaa: rfx	aaa: The opera house was empty
aab: gurl	aab: The soprano sang the aria beautifully
aac: imnsi	aac: Some people attended the opera
aad: wjgx	aad: The sets malfunctioned
aae: tg	aae: The costumes were delivered late
aaf: kopg	aaf: There were technical difficulties
aag: khh	aag: The show faced some challenges
aah: ozro	aah: The orchestra played flawlessly
aai: pg	aai: The tenor forgot his lines
aaj: bill	aaj: The performance went smoothly
aak: mek	aak: There was a major disruption
aan: jp	aan: The opening night was eventful
reasoning task: gurl is True. pg is False. ozro is True. (ozro or pg) - bill. rfx is False. (rfx or gurl) - imnsi. tg is False. wjgx is False. (wjgx or tg) - kopg. (imnsi or kopg) - khh. (NOT bill) - mek. (khh or mek) - jp. Deduce the result of jp.	reasoning task: The soprano sang her aria beautifully and the orchestra played flawlessly, but the tenor forgot his lines. If the orchestra played well or the tenor forgot his lines, then the performance did not go entirely smoothly. The opera house was not empty since the soprano's beautiful aria meant some people attended. The costumes were not delivered late and the sets did not malfunction, so there were no technical difficulties. If some people attended or there were technical difficulties, the show would have faced some challenges. Since the performance did not go smoothly, it implies there was a major disruption. If the show faced challenges or had a major disruption, the opening night of this opera was quite eventful. Given this, was the opening night of the opera eventful?
Level 4 - Abstract	Level 4 - Health and Fitness
aaa: msta	aaa: Sue did push-ups yesterday
aab: fo	aab: Sue did not do pull-ups yesterday
aac: jfnrh	aac: Sue did some upper body exercises yesterday
aad: ssb	aad: Sue did squats yesterday
aae: ac	aae: Sue did not do squats yesterday
aaf: dzda	aaf: Sue only trained upper body yesterday
aag: hujcf	aag: Sue did burpees yesterday
aah: pil	aah: Sue did not do burpees yesterday
aai: dyue	aai: Sue trained her core muscles yesterday
aaj: sgniu	aaj: Sue did planks yesterday
aak: stbf	aak: Sue had an effective core workout yesterday
aan: pswg	aan: Sue did an intense workout yesterday
aao: fkyxi	aao: Sue had a focused or intense workout yesterday

Continued

Table 4 – continued from previous page

Abstract Example	Specific Domain Contextualized Example
aap: outm aaq: ybjj aar: eek aas: wmejd aat: rdbk aau: rqmc aav: bw aaw: xvd aax: pg aay: qbli aaz: qvb reasoning task: fo is False. msta is True. (msta or fo) -> jfnrh. dyue is True. xvd is False. (NOT xvd) -> pg. ssb is True. (NOT ssb) -> ac. (jfnrh and ac) -> dzda. sgniu is True. (dyue or sgniu) -> stbf. outm is True. rdbk is False. ybjj is True. (outm or ybjj) -> eek. wmejd is True. (wmejd or rdbk) -> rqmc. (eek and rqmc) -> bw. (NOT pg) -> qbli. (bw or qbli) -> qvb. hujcf is True. (NOT hujcf) -> pil. (pil and stbf) -> pswg. (dzda or pswg) -> fkyxi. (fkyxi or qvb) -> abc. Deduce the result of abc.	aap: Sue did lunges yesterday aaq: Sue did step-ups yesterday aar: Sue trained her leg muscles yesterday aas: Sue did wall sits yesterday aat: Sue did not do calf raises yesterday aau: Sue did some quad and hamstring exercises yesterday aav: Sue had an effective lower body workout yesterda aaw: Sue did not do a full body workout yesterday aax: Sue did a partial body workout yesterday aay: Sue did a full body workout yesterday aaz: Sue had a comprehensive workout yesterday reasoning task: Sue did push-ups but not pull-ups yesterday. If she did push-ups or pull-ups, then she did some upper body exercises. Sue trained her core by doing planks. Since she did not do a full body workout, it means she did a partial body workout. Sue did squats yesterday, so it is not true that she did not do squats. If Sue did some upper body exercises and did not do squats, then she only trained upper body. If Sue did planks or trained her core muscles, then she had an effective core workout. Sue did lunges and step-ups, but not calf raises. If she did lunges or step-ups, then she trained her leg muscles. If Sue did wall sits or calf raises, then she did some quad and hamstring exercises. If Sue trained her leg muscles and did some quad/hamstring exercises, then she had an effective lower body workout. If Sue did not do a partial body workout, then she did a full body workout.If Sue had an effective lower body workout or did a full body workout, then she had a comprehensive workout. Sue did burpees yesterday, so it is not true that she did not do burpees. If Sue did not do burpees and had an effective core workout, then she did an intense workout. If Sue only trained upper body or did an intense workout, then she had a focused or intense workout. If Sue had a focused/intense workout or a comprehensive workout, then she had a productive bodyweight training session. Did Sue have a productive bodyweight training session yesterday?