SATQUEST: A VERIFIER FOR LOGICAL REASONING EVALUATION AND REINFORCEMENT FINE-TUNING OF LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) exhibit strong general reasoning, yet the community lacks *controllable, scalable, and verifiable* tools to analyze and improve these abilities. We present `SATQuest`, a verifier that generates diverse SAT-based reasoning tasks directly from Conjunctive Normal Form (CNF) instances and checks answers objectively with PySAT. `SATQuest` factorizes evaluation along three orthogonal dimensions—*instance*, *problem type*, and *question format*—enabling fine-grained, multi-dimensional analysis and reinforcement fine-tuning. Randomized CNF generation mitigates memorization and supports reproducible experiments. Using `SATQuest`, we benchmark a range of open- and closed-weight LLMs and uncover persistent gaps in logical reasoning, particularly on higher-complexity tasks and in transfer beyond familiar mathematical notation to machine or narrative formats. We further show that reinforcement fine-tuning with `SATQuest` rewards substantially boosts targeted performance and generalizes to larger instances, while cross-format robustness remains challenging. Collectively, `SATQuest` provides verifier-backed infrastructure for controlled, scalable, and reproducible *empirical research* on LLM logical reasoning and its training.

Anonymous GitHub

1 INTRODUCTION

Large Language Models (LLMs) have demonstrated remarkable proficiency in general reasoning tasks, including complex problem-solving and code generation, with models like `o3-mini`(OpenAI, 2025), `DeepSeek-R1`(DeepSeek-AI et al., 2025), and `QwQ-32B`(Team, 2024) excelling in programming, mathematics, and scientific question-answering(Brown et al., 2020; OpenAI et al., 2024; Wei et al., 2022). This advanced reasoning capability, a cornerstone for Artificial General Intelligence (AGI), is a critical indicator of models' deep understanding and generalization.

To further explore the reasoning capabilities of LLMs, we urgently need evaluation and training tools that are both controllable and scalable. Fine-grained and reliable performance analysis requires systematic variable control, which is fundamental to empirical scientific research. However, existing benchmarks and datasets have significant limitations in variable control, making it difficult to support multi-dimensional, systematic analysis and training experiments—hindering deeper understanding of reasoning mechanisms in LLMs.

Although benchmarks such as *ZebraLogic* (Lin et al., 2025) and *Knights and Knaves* (Xie et al., 2024) have introduced controllable difficulty dimensions, their question types and formats remain narrow, supporting only the *instance* dimension. *MATH* (Hendrycks et al., 2021b) and *LiveBench* (White et al., 2025) provide rich content across mathematics and programming; however, they lack structured relationships and rely heavily on human judgment for difficulty labeling. In comparison, general-purpose benchmarks like *GPQA* (Rein et al., 2024), *MMLU* (Hendrycks et al., 2021a), and *Big-Bench* (Srivastava et al., 2023) offer broad coverage, yet suffer from issues such as data leakage, lack of continuity, and insufficient support for multi-dimensional controlled analysis.

In this work, we aim to construct a multi-dimensionally controllable and scalable verifier to support the evaluation and training of LLM reasoning capabilities, enabling the tracking of progress and discovery of limitations in LLM reasoning. Our main contributions are as follows:

(1) **CNF-native, controllable verifier.** We present `SATQuest`, which generates SAT-based reasoning tasks directly from CNF and organizes them along three orthogonal axes—*instance*, *problem type*, and *question format*. Randomized CNF construction, binary-string targets, and objective PySAT checking enable leakage-resistant, verifiable evaluation and training; we also release 🤗 `ANON/SATQuest` for reproducibility.

(2) **Multi-dimensional evaluation.** Using `SATQuest`, we provide a comprehensive study of open/closed-weight LLMs across all three axes, revealing strong cross-format sensitivity beyond `Math`, accuracy degradation with solver-measured complexity, and a systematic shift toward hallucination (fabricated execution and invented shortcuts) under high complexity (App. I).

(3) **Verifier-driven Reinforcement Fine-tuning.** We implement reinforcement fine-tuning with `SATQuest` rewards, yielding targeted gains and longer reasoning chains. Improvements transfer to larger instances and exhibit *asymmetric* cross-task generalization, while cross-format robustness remains limited.

2 SATQUEST CHALLENGE DESIGN

Overview. `SATQuest` is a systematic verifier engineered to comprehensively evaluate and enhance the logical reasoning capabilities of LLMs. Its primary goal is to offer a framework for fine-grained analysis, providing deeper insights into the strengths and limitations of LLMs in logical deduction. `SATQuest` is not designed to train LLMs as general-purpose solvers or to make them surpass specialized symbolic solvers in speed or accuracy.

To achieve its objectives, `SATQuest` automatically generates a variety of logical reasoning tasks directly derived from CNF instances. These tasks are meticulously organized along three orthogonal dimensions: *instance*, *problem type*, and *question format*, each targeting distinct aspects of logical reasoning. This multi-dimensional structure creates a comprehensive and controllable challenge space suitable for nuanced evaluation and effective reinforcement fine-tuning. CNF instances are utilized as the foundational elements due to their formal clarity, their established role as a standard representation in propositional logic, and their inherent compatibility with established SAT solvers for objective answer verification, making them an ideal medium for systematically probing the logical capabilities of LLMs.

Data Generation. `SATQuest` supports evaluation using any CNF instance dataset stored in the standard `DIMACS` format. For reproducibility, we generated two CNF datasets using the procedure outlined in Algorithm 1:

- 🤗 `ANON/SATQuest`: This dataset is generated for evaluation purposes (Sec. 3), specifically to assess logical reasoning across varying instance scales and difficulties. It consists of randomly generated CNF instances with $n \in [3, 16]$ variables and a fixed clause-to-variable ratio resulting in $m = 4n$ clauses. For each (n, m) configuration, 10 CNF instance pairs (one satisfiable and one unsatisfiable) were generated, resulting in a total of 140 CNF pairs. The two CNF instances share the same number of literals and nearly identical CNF structures.
- 🤗 `ANON/SATQuest-RFT-3k`: This dataset is generated for reinforcement fine-tuning (Sec. 4). It consists of CNF instances with $n \in [3, 8]$ variables and clause counts m determined by varying the clause-to-variable ratio from 2.1 to 4.0 in increments of 0.1 (i.e., m ranges from $2.1n$ to $4.0n$). For each (n, m) configuration, 25 CNF instance pairs were generated, resulting in a total of $3,000$ CNF pairs.

Challenge Dimensions 1: *Instance*. We categorize instances by their *scale* (number of variables n, clauses m, and literals) and inherent *difficulty*. Computational *difficulty* is assessed using SAT solver statistics: *decisions*, indicating search breadth; *conflicts*, reflecting constraint-driven backtracking; and *propagations*, quantifying chained logical implications. These structural and solver-derived

metrics provide a multi-faceted characterization of an instance's combinatorial complexity, where higher values generally correspond to more challenging problems.

Challenge Dimensions 2: *Problem Type*. Let $\alpha : X \to \{0, 1\}$ be an assignment for $X = \{x_1, \ldots, x_n\}$, and let $F = \bigwedge_{i=1}^{m} C_i$ be a CNF formula. Each C_i is a clause of the form $C_i = \ell_{i,1} \vee \ell_{i,2} \vee \cdots \vee \ell_{i,k_i}$, where $\ell_{i,j} \in X \cup \{\neg x | x \in X\}$. We define five fundamental SAT-based problems:

- `SATDP` (SAT Decision Problem): Determine whether F is satisfiable:

$$\texttt{SATDP}(F) = \begin{cases} 1, & \exists\, \alpha \text{ s.t. } F(\alpha) = 1, \\ 0, & \text{otherwise.} \end{cases}$$

  Tests the fundamental ability to determine logical consistency.
- `SATSP` (SAT Search Problem): If F is satisfiable, find an assignment α:

$$\alpha \text{ s.t. } F(\alpha) = 1.$$

  Probes constructive reasoning by requiring the generation of a satisfying assignment.
- `MaxSAT` (Maximum Satisfiability): Find the assignment α^* that maximizes the number of satisfied clauses:

$$\alpha^* = \arg\max_{\alpha} \sum_{i=1}^{m} \mathbf{1}[C_i(\alpha) = 1].$$

  Evaluates optimization skills when maximizing clause satisfaction under conflicting constraints.
- `MCS` (Minimal Correction Subset): For an unsatisfiable F, find a minimal set $S \subseteq \{1, \ldots, m\}$ whose removal yields a satisfiable formula:

$$\texttt{SATDP}(\bigwedge_{i \in S} C_i) = 1 \bigwedge \forall\, S' \subset S,\ \texttt{SATDP}(\bigwedge_{i \in S'} C_i) = 0.$$

  Tests diagnostic reasoning through the identification of minimal corrections for unsatisfiability.
- `MUS` (Minimal Unsatisfiable Subset): For an unsatisfiable F, find a minimal unsatisfiable core $S \subseteq \{1, \ldots, m\}$:

$$\texttt{SATDP}(\bigwedge_{i \in S} C_i) = 0 \bigwedge \forall\, S' \subset S,\ \texttt{SATDP}(\bigwedge_{i \in S'} C_i) = 1.$$

  Probes diagnostic reasoning by localizing minimal sources of logical inconsistency.

Each CNF instance pair consists of one satisfiable and one unsatisfiable formula. The satisfiable instance is used for `SATSP`, while the unsatisfiable instance is used for `MaxSAT`, `MCS`, and `MUS`. For `SATDP`, both instances are evaluated, forming two sub-tasks (`SATDP-sat` and `SATDP-unsat`). A response to `SATDP` is considered correct only if both sub-tasks are answered correctly, thereby discouraging random guessing. These problems progressively challenge LLM reasoning capabilities, from foundational deduction and solution construction (`SATDP`/`SATSP`), to constrained optimization (`MaxSAT`), and finally to minimal cause identification and correction (`MCS`/`MUS`).

Challenge Dimensions 3: *Question Format*. Recognizing that the presentation of a problem can significantly impact an LLM's reasoning process, this dimension introduces four logically equivalent representational formats for each CNF instance. This variation aims to test different reasoning skills and reduce reliance on superficial pattern matching.

- `Math` (mathematical notation): Uses $\wedge$, $\vee$, and $\neg$ to represent logic formulas. Balances between formality and readability.
  Example: $x_1 \vee \neg x_2 \vee x_3$
- `DIMACS` (machine format): A minimal, line-based format for representing Boolean formulas in CNF, the standard input for many SAT symbolic solvers.
  Example:

```
p cnf 3 1
1 -2 3 0
```

- `Story` (OR semantics, cookie day scenario): Wraps clauses as friendly narratives—"Alice is happy if..."—to test LLMs' ability to ground disjunctions in natural language.
  Example: "Alice will be happy if she gets crunchy choco (x_1), chewy vanilla ($\neg x_2$), or crunchy peanut (x_3)."
- `DualStory` (AND semantics, cookie day scenario): Presents the negated form—"Alice will be unhappy only if..." —turning OR into AND and requiring semantic tracking.
  Example: "Alice will be unhappy only if she is served crunchy choco (x_1), chewy vanilla ($\neg x_2$), and crunchy peanut (x_3)."

`Math` is common in training data and accessible to math-tuned LLMs, whereas `DIMACS` is a compact, noise-free, machine-readable format that tests a model's ability to interpret raw clause structures. We adopt it specifically to evaluate whether LLMs can interpret raw clause structures without relying on mathematical training. `Story` and `DualStory` introduce narrative elements that add informational noise and require the model to translate natural-language logical structures into formal logic before reasoning. The deterministic CNF→narrative conversion maps variables to cookie names (from a predefined pool), positive/negative literals to "crunchy"/"chewy" variants, and clauses to individual friend preferences; the complete mapping procedure with pseudocode is given in App. B.

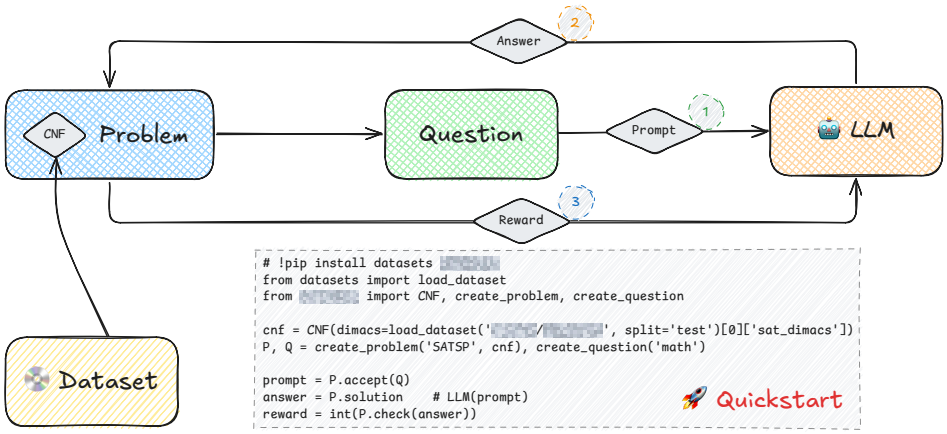

Figure 1: `SATQuest` Pipeline and Quickstart.

Answer Verification. Prior logical reasoning benchmarks (e.g., (Lin et al., 2025; Xie et al., 2024; Mondorf & Plank, 2024)) typically require LLMs to produce `JSON` outputs, choose from preset options, or use complex extraction—methods that hinder automation and scalability despite reduced manual overhead. Moreover, multiple-choice formats can significantly alter the true difficulty of problems, as the complexity of `SATSP` differs fundamentally from SAT verification. `SATQuest` uses a unified binary-string output format—1 bit for `SATDP`, n bits for `SATSP`/`MaxSAT`, and m bits for `MCS`/`MUS` —to ensure consistency across tasks and to control for variation in the output schema. These strings are extracted using regex and checked against CNF constraints via PySAT, allowing for multiple valid answers. A known limitation is that long binary outputs may challenge smaller models' ability to adhere to the expected format. To shed light on this, Fig. 7 presents format correctness statistics for mainstream models evaluated. For the complete prompt and detailed output format instructions, refer to App. C. The `SATQuest` Pipeline and Quickstart are illustrated in Fig. 1.

3 Evaluation

Overview. We conduct a comprehensive evaluation of the logical reasoning performance of various LLMs using the `SATQuest` benchmark. The analysis spans different *instance*, *problem types*, and *question formats*, enabling fine-grained, multi-dimensional insights. We begin with the overall benchmark results, followed by detailed analyses along each dimension.

Setup. We evaluate a diverse set of state-of-the-art open-weight and closed-weight LLMs, including vanilla models, reasoning models, and distilled variants. The evaluation uses the

🤗 ANON/SATQuest dataset, comprising 140 CNF instance pairs categorized by scale ($n \in [3, 16]$, clauses $m = 4n$). Each CNF pair yields tasks across five logical reasoning types (SATDP, SATSP, MaxSAT, MCS, MUS) and four question formats (Math, DIMACS, Story, DualStory), resulting in 20 evaluations per CNF pair. For detailed evaluation configurations and parameters, see App. F.

Overall Results. Fig. 2 shows model accuracies on SATQuest. o3-mini leads with 0.56 accuracy, followed by DeepSeek-R1 (0.42), QwQ-32B (0.40), and DeepSeek-R1-Distill-Qwen-32B (0.39), indicating that reasoning-enhanced models outperform vanilla LLMs. Large vanilla models like GPT-4.1 (0.38) and DeepSeek-V3-0324 (0.36) perform competitively despite lacking explicit reasoning training. In contrast, smaller vanilla models (e.g., Qwen2.5-7B-Instruct) achieve below 0.1 accuracy, revealing limited reasoning capabilities.

These results point to two notable trends. First, reasoning models consistently outperform vanilla counterparts, particularly on more complex tasks. Second, the modest overall accuracy across models reflects the challenging nature of SATQuest and its effectiveness in distinguishing reasoning capabilities. Moreover, the model ranking induced by SATQuest is broadly consistent with that of other recent reasoning benchmarks such as *GPQA* (Rein et al., 2024) and *ZebraLogic* (Lin et al., 2025), lending confidence that SATQuest reliably differentiates reasoning capability within the evaluated LLM population.

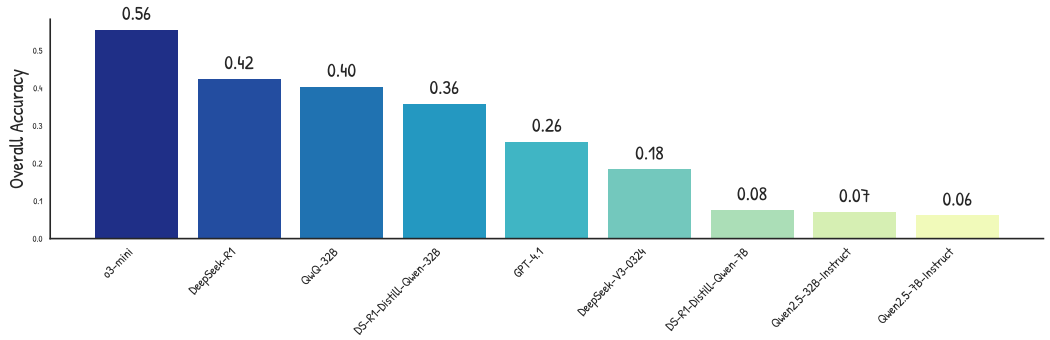

Figure 2: Overall accuracy of evaluated LLMs on the SATQuest benchmark, averaged across all *problem types* and *question formats*.

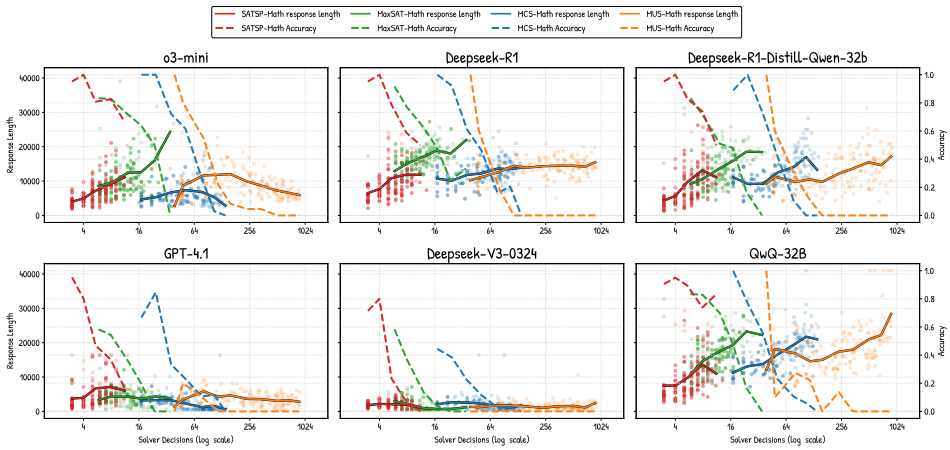

Figure 3: Accuracy (dashed lines, left axis) and response length (solid lines, right axis) vs. instance complexity (solver decisions) for SATSP-Math, MaxSAT-Math, MCS-Math and MUS-Math tasks.

Analysis by *Instance*. We examine how model performance scales with instance complexity, measured using the number of decisions made by established SAT solvers (Glucose 4.1 (Audemard & Simon, 2018) for SATDP/SATSP, RC2 (Ignatiev et al., 2019) for MaxSAT, LBX (Mencía et al., 2015)

for MCS, and MUSX (Marques-Silva, 2010) for MUS). Fig. 3 visualizes model accuracy and response length against this complexity metric. A consistent trend across tasks is that as instance complexity increases (more solver decisions), model accuracy tends to decline, while response length generally increases. Notably, we observe a concerning hallucination phenomenon when models encounter highly complex instances: they often fabricate solver calls or invent simplified reasoning paths rather than engaging with the full logical complexity of the problem. This hallucination is particularly evident for o3-mini on MCS-Math and MUS-Math tasks, where response length actually decreases at high complexity, indicating the model abandons complete reasoning in favor of hallucinated shortcuts. Overall, all models exhibit reduced accuracy on larger and more difficult instances. Top-performing models like o3-mini and DeepSeek-R1 show more gradual degradation, indicating better scalability, whereas less capable models experience a sharp performance drop, often failing completely on moderately complex instances and resorting to increasingly severe hallucinations when faced with logical complexity beyond their reasoning capacity.

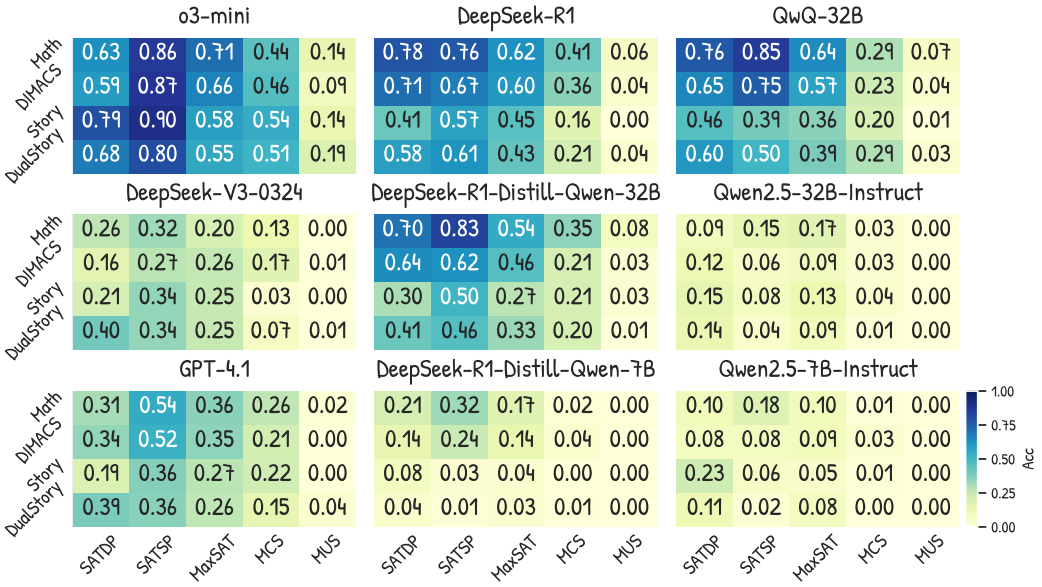

Figure 4: Accuracy heatmaps from a single run for each model, showing performance by *problem type* (columns) and *question format* (rows).

Analysis by *Problem Type*. We analyze performance variations across the five distinct SAT-based problem types: SATDP, SATSP, MaxSAT, MCS, and MUS. The heatmaps in Fig. 4 reveal performance differences across these tasks. Consistently across models, performance tends to be highest on SATDP (especially SATDP-sat) and SATSP, and lowest on MCS and MUS, forming a clear difficulty hierarchy. Models generally handle the basic decision (SATDP) and search (SATSP) tasks better than tasks requiring optimization or diagnosis, though performance on SATSP is often lower than SATDP. Performance on the optimization task MaxSAT typically sits between the basic tasks and the diagnostic tasks. The diagnostic tasks, MCS and MUS, which require identifying minimal subsets, prove particularly challenging, with accuracy dropping significantly for almost all models. While top models like o3-mini maintain some capability even on harder tasks, the gap between task types is pronounced across the board. This performance stratification aligns with the solver complexity shown in Fig. 3 (measured by solver decisions), suggesting LLMs struggle progressively more with tasks demanding global optimization, minimality constraints, and diagnostic reasoning over combinatorial spaces. Overall, the results highlight LLM limitations in tackling the full spectrum of logical reasoning challenges represented by these diverse SAT-based tasks.

Analysis by *Question Format*. The way a logical problem is presented can significantly affect an LLM's ability to solve it. We analyze this impact by evaluating performance across four distinct question formats: Math, DIMACS, Story, and DualStory. Fig. 4 illustrates how accuracy varies across these formats. All models perform best in the Math format, generally achieving their highest

accuracy, followed by DIMACS, with Story and DualStory formats yielding the lowest accuracy. o3-mini demonstrates relatively stable performance across the four formats, indicating strong reasoning robustness regardless of presentation style. However, other open-weight reasoning models like DeepSeek-R1 and QwQ-32B, while performing well in the Math format, exhibit a significant drop in accuracy in other formats, suggesting higher sensitivity to the presentation style.

It is noteworthy that the Story and DualStory problems introduce narrative elements, adding informational noise and requiring the model to translate the natural language logical structure into formal logic before reasoning. The increased difficulty and subsequent lower accuracy are thus expected. However, the DIMACS format is structurally similar to Math, contains no redundant information, and has higher information density. Despite this, open-weight reasoning models still show a marked decrease in accuracy compared to Math (e.g., DeepSeek-R1 and QwQ-32B accuracy dropped by 9% and 10% respectively on SATSP-DIMACS compared to SATSP-Math).

Through case studies presented in App. E, we observe that DeepSeek-R1 and QwQ-32B often attempt to reason directly within the DIMACS format rather than translating it into formal mathematical notation. This approach involves working with the raw DIMACS clauses, which requires tracking multiple variable assignments simultaneously across numerous constraints. The models frequently make errors when attempting to verify clause satisfaction or when determining the implications of specific variable assignments, particularly misinterpreting the disjunctive nature of clauses or conflating the semantic meaning of positive and negative literals. This direct approach appears to lead to a higher error rate during the reasoning process, as the models struggle to maintain consistency across the complex network of logical constraints represented in the DIMACS format.

Conversely, vanilla models like GPT-4.1 and DeepSeek-V3-0324, although performing worse overall, show relatively balanced performance across different formats, indicating lower format sensitivity. These models sometimes employ structured thinking approaches by first translating DIMACS inputs into the Math format before proceeding with reasoning, or by introducing meaningful symbolic notation during their reasoning process, which appears to enhance reasoning stability. The shorter reasoning chains produced by vanilla models may also contribute to their format robustness, as briefer deductions have fewer opportunities for errors to accumulate. This contrasts with DeepSeek-R1 and QwQ-32B, whose struggles outside the Math format seem to stem from relying more on potentially error-prone direct reasoning or trial-and-error within unfamiliar formats, rather than employing systematic format translation or structured analysis.

Achieving AGI likely requires LLMs to reason effectively across diverse formats, enabling the integration of knowledge from different domains and fostering more powerful, generalized reasoning capabilities. SATQuest thus serves as a valuable benchmark for assessing LLMs' adaptability and robustness in logical reasoning across various presentation styles.

4 Reinforcement Fine-Tuning

Overview. As demonstrated by the evaluation in Sec. 3, current LLMs exhibit significant limitations in logical reasoning, with notable deficiencies in generalization across *instance*, *problem type*, and *question format*. This section explores avenues for enhancing LLM logical reasoning capabilities through Reinforcement Fine-Tuning (RFT), directly utilizing reward signals from the SATQuest verifier. Our investigation particularly focuses on two aspects: First, we assess whether SATQuest-driven RFT can stimulate LLMs to construct longer reasoning chains, thereby fostering deeper logical deduction. Second, we delve into the relationship between RFT and the generalization deficiencies identified in Sec. 3, aiming to elucidate the specific effects and potential bottlenecks of RFT in improving cross-task and cross-format generalization.

Setup. We select the Qwen2.5-7B-Instruct model as the baseline, utilizing the Group Relative Policy Optimization (GRPO) algorithm (Shao & et al., 2024) during fine-tuning. This baseline model is nearly a blank slate, a small vanilla model that only shows marginal performance on SATSP-Math. The training leverages the 🤗 ANON/SATQuest-RFT-3k dataset, comprising $3,000$ CNF instance pairs with $n \in [3, 8]$ variables and clause-to-variable ratios ranging from 2.1 to 4.0. We train three distinct models, each focusing on a specific *problem type* and *question format* combination: SATSP-Math, SATSP-Story, MaxSAT-Math.

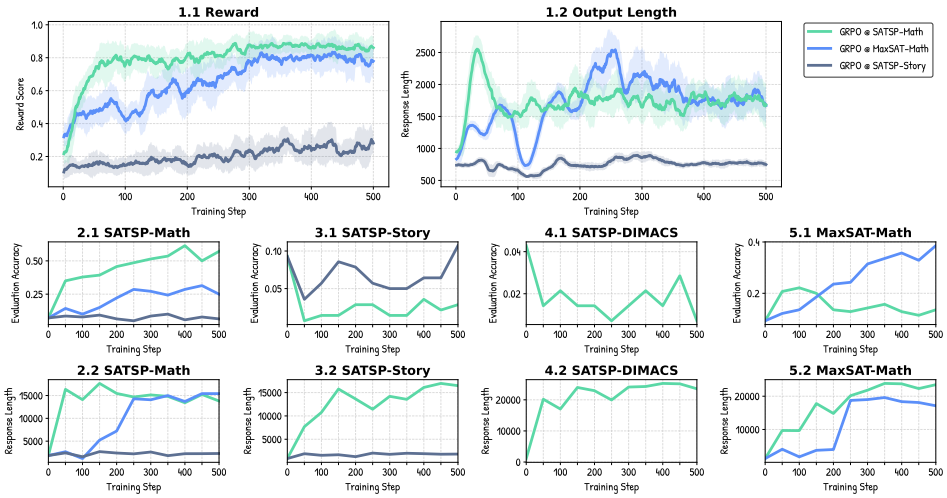

Figure 5: GRPO fine-tuning using `SATQuest` rewards. Training dynamics (Top: reward, response length) and evaluation performance (Bottom: accuracy, response length on target and generalization tasks) vs. training steps for models fine-tuned on `SATSP-Math`, `MaxSAT-Math`, and `SATSP-Story`. Solid lines show the mean; shaded regions denote ± 1 standard deviation.

Our implementation is based on the TRL library (von Werra et al., 2020), adopting the GRPO objective as described in (Shao & et al., 2024). The prompt template follows the structure proposed in (DeepSeek-AI et al., 2025). The reward function is designed using the `SATQuest` verifier, assigning a reward of 1.0 for correct answers and 0.0 for incorrect ones. Additionally, two format correctness rewards, detailed in App. G.1, are incorporated with weights of 0.05 each, complementing the primary reward weight of 1.0. For training we set max_prompt_length to 2048 and $max_completion_length$ to 8192, for evaluation we set $max_prompt_length + max_completion_length$ to 32768. The training and evaluation parameters are detailed in App. G.2.

Training Dynamics. Training curves (Fig. 5, top row) reveal that models trained on `Math`-based tasks (`SATSP-Math` and `MaxSAT-Math`) achieve higher rewards and generate longer responses than those trained on `SATSP-Story`. The `Math` format appears to better facilitate extended reasoning chains that receive positive reinforcement from the verifier. Our `SATQuest` verifier effectively stimulates extended reasoning development within few training steps, especially with the mathematical format. Response length curves show distinct patterns—rapid initial growth as models learn longer reasoning chains, temporary decline when adapting to format constraints at the training response limit (8192 tokens), followed by stabilization.

Training on `SATSP-Story` proved less effective, largely due to the baseline model's weak narrative reasoning abilities. While *Logic-RL* (Xie et al., 2025) has successfully stimulated narrative reasoning, our tasks involve substantially higher complexity and scale. Convergence efficiency correlates with task complexity and initial model capabilities, though these factors require further investigation to fully separate.

Generalization Across *Instance*. We observe positive generalization concerning problem complexity. Models fine-tuned on `SATSP-Math` and `MaxSAT-Math` demonstrated improved accuracy when evaluated on the corresponding tasks within the evaluation set (Fig. 5, subplots 2.1, 5.1). Crucially, these evaluation instances involved larger scales ($n > 8$) than those used during training ($n \in [3, 8]$), indicating that the learned reasoning skills generalize to more complex instances within the same problem and format.

Generalization Across *Problem Types*. Our results reveal an interesting asymmetry in cross-problem generalization. Fine-tuning on the more complex `MaxSAT-Math` task led to performance improvements not only on `MaxSAT-Math` itself but also conferred benefits to the simpler `SATSP-Math` task. However, the model trained solely on `SATSP-Math` did not show a corresponding improvement on `MaxSAT-Math` (compare improvements patterns in Fig. 5, bottom row). This

Generalization Type	Positive Transfer?	Observation
Across Instances	✓	Improved accuracy on larger unseen CNF instances.
Across Problem Types	✓ (Asymmetric)	MaxSAT → SATSP transfer stronger.
Across Question Formats	✗	Math → DIMACS/Story weak transfer.

Table 1: Summary of generalization findings across SATQuest dimensions.

suggests that the reasoning capabilities required for MaxSAT may encompass those needed for SATSP. Strategically, this implies that training on more complex and diverse logical problems could be more effective for fostering robust reasoning skills that generalize to simpler, related problems.

Generalization Across *Question Formats*. Cross-format generalization remains notably difficult. The model fine-tuned on SATSP-Math shows minimal improvement when evaluated on other formats such as SATSP-Story (Fig. 5, subplot 3.1) and even on the structurally similar SATSP-DIMACS task (subplot 4.1). This suggests that reasoning capabilities acquired in the Math format do not readily transfer to logically equivalent tasks presented in alternative formats, whether narrative or machine-readable. Further analysis of failure cases reveals that, after SATSP-Math fine-tuning, the model tends to generate verbose but flawed reasoning when confronted with DIMACS inputs—mirroring the issues described in Sec. 3 and App. E. The model appears to have overfitted to a specific Math-style reasoning pattern, at the expense of its initial structured thinking ability, and fails to effectively translate DIMACS representations into a suitable reasoning form. These findings indicate that small-scale RFT may not be sufficient to overcome format generalization barriers, and that the performance discrepancies across formats observed in Sec. 3 may partially stem from the limitations of format-specific fine-tuning itself.

5 Related Work

Researchers have developed numerous benchmarks to evaluate LLMs' capabilities. Popular evaluations like (Rein et al., 2024; Hendrycks et al., 2021b;a; Srivastava et al., 2023) comprehensively assess LLMs but suffer from data leakage and lack of continuity, often being solved by advanced models within 18 months of introduction. While newer benchmarks (Suzgun et al., 2023; Kazemi et al., 2025; Team et al., 2025; Gema et al., 2025; Wang et al., 2024; Glazer et al., 2024) offer improvements, core issues remain. White et al. (2025) introduced dynamic question banks and automatic scoring but still relies on manual difficulty annotation without multi-dimensional analysis controls. These evaluations primarily assess capabilities rather than providing insights into internal mechanisms. Some studies explore more nuanced approaches: Lin et al. (2025); Xie et al. (2024) use formalized templates with controllable difficulty dimensions for finer-grained analysis. Xie et al. (2025) conducts RL training on *K&K* to investigate how RL enhances reasoning capabilities. He et al. (2024) provides evaluation through multilingual coverage and multi-turn design. Huang et al. (2025); Yu et al. (2025) test mathematical reasoning robustness through minimal perturbations. Research by Hazra et al. (2025) investigates LLM reasoning capabilities through 3-SAT phase transitions. Our work, SATQuest, uses randomly generated CNF instances to prevent data leakage and ensure continuity. We provide five interrelated SAT-based *problem types* and four *question formats* with different information densities. These three orthogonal dimensions—*instance*, *problem type*, and *question format*—enable flexible experimental control for future LLM reasoning research.

6 Conclusion

We introduce SATQuest, a CNF-native verifier that systematically structures SAT-based reasoning evaluation and training along three orthogonal axes: *instance*, *problem type*, and *question format*. Our results reveal that while targeted reinforcement fine-tuning with verifiable rewards enables measurable gains—especially for harder or larger instances—persistent weaknesses remain in diagnostic reasoning and robustness to presentation changes. The key patterns of generalization and their asymmetries are summarized in Table 1.

Takeaway. SATQuest provides a controllable and verifiable feedback loop for probing LLM logical reasoning. It reliably exposes non-obvious failure modes—particularly format sensitivity and difficulty with minimal diagnoses—while enabling principled improvements without manual heuristics. Future progress will require richer reward signals and explicit curriculum bridging across formats and domains, to close the gap between instance-specific improvements and generalizable, format-robust reasoning, and we release datasets and code to catalyze progress.

7 LIMITATIONS

Scope and intent. SATQuest is designed as a *controllable, verifiable, and factorized* probe of LLM logical reasoning along three axes (*instance* / *problem type* / *question format*), prioritizing clean manipulation of variables over breadth. MCS/MUS are general diagnostic notions beyond propositional logic, but in this work they are *instantiated* as clause-level subsets over CNF; richer logics (e.g., first-order, modal), multi-tool interaction, and open-ended natural-language tasks are out of scope. The evaluation set fixes $m{=}4n$ with $n{\leq}16$, the RFT training set uses $n{\leq}8$, and narratives are synthetic and English-only. Hence our conclusions target *relative* patterns across instance/problem/format rather than absolute real-world capability.

Technical. By design, our setup prioritizes *verifiable, controlled analysis*: models must return binary strings that we extract and check against the CNF with PySAT, which removes grading ambiguity but constrains the response space and can be brittle for long clause-mask outputs. All evaluations are conducted in a *strict zero-shot* regime using standardized prompts (no tools, no few-shot exemplars), to isolate reasoning from prompt engineering; the prompt templates used for evaluation are provided in the paper and appendix. On the training side, we only instantiate reinforcement fine-tuning with the GRPO algorithm, optimizing a *binary* (0/1) verifier reward; alternative policy-optimization methods, decoding strategies, and shaped rewards are not explored here. Task difficulty is proxied by classical solver statistics—decisions, conflicts, and propagations—which we compute with established solvers per task; the complete per-instance statistics and the solver identifiers are provided with the dataset metadata.

Narrative format robustness. The CNF$\rightarrow$Story/DualStory conversion uses a fixed narrative template with entity names sampled from small predefined pools; we have not conducted systematic prompt-paraphrasing experiments to quantify how much the specific wording affects performance. The consistent ordering Math $\geq$ DIMACS $>$ Story $\approx$ DualStory across multiple models and tasks suggests that the main trend is not driven solely by a particular choice of phrasing, but we cannot rule out that alternative narrative framings would shift absolute accuracy. More diverse narrative templates and controlled paraphrasing ablations are an important direction for future work.

Future work. Future directions include expanding the diversity and realism of question formats—incorporating both LLM-generated and programmatically varied representations such as tabular, code-based, or diagrammatic/logical schemas—to better capture the heterogeneity of human reasoning contexts. We also plan to curate *industrial-scale CNF corpora* drawn from real verification, planning, and scheduling benchmarks with standardized metadata for open research use. On the learning side, designing *dense and shaped rewards* (e.g., clause coverage, unit-propagation progress, MCS/MUS cardinality margins, or partial-assignment consistency) may provide more informative optimization signals than binary correctness alone. Curriculum learning across scale (n, m), problem-type mixtures, and structural regimes can help bridge small synthetic tasks and large real-world instances. Finally, SATQuest is intended to serve as a practical foundation for *empirical research* on LLM logical reasoning—offering a controllable, verifier-backed testbed for standardized and reproducible studies of reward shaping, curricula, and cross-format transfer.

REFERENCES

Gilles Audemard and Laurent Simon. On the glucose sat solver. *International Journal on Artificial Intelligence Tools*, 27(01):1840001, 2018. doi: 10.1142/S0218213018400018. URL https://doi.org/10.1142/S0218213018400018.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, Claire Barale, Robert McHardy, Joshua Harris, Jean Kaddour, Emile Van Krieken, and Pasquale Minervini. Are we done with MMLU? In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5069–5096, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. URL https://aclanthology.org/2025.naacl-long.262/.

Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, Olli Järviniemi, Matthew Barnett, Robert Sandler, Matej Vrzala, Jaime Sevilla, Qiuyu Ren, Elizabeth Pratt, Lionel Levine, Grant Barkley, Natalie Stewart, Bogdan Grechuk, Tetiana Grechuk, Shreepranav Varma

Enugandla, and Mark Wildon. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai, 2024. URL https://arxiv.org/abs/2411.04872.

Rishi Hazra, Gabriele Venturato, Pedro Zuidberg Dos Martires, and Luc De Raedt. Have large language models learned to reason? a characterization via 3-sat phase transition, 2025. URL https://arxiv.org/abs/2504.03930.

Yun He, Di Jin, Chaoqi Wang, Chloe Bi, Karishma Mandyam, Hejia Zhang, Chen Zhu, Ning Li, Tengyu Xu, Hongjiang Lv, Shruti Bhosale, Chenguang Zhu, Karthik Abinav Sankararaman, Eryk Helenowski, Melanie Kambadur, Aditya Tayade, Hao Ma, Han Fang, and Sinong Wang. Multi-if: Benchmarking llms on multi-turn and multilingual instructions following, 2024. URL https://arxiv.org/abs/2410.15553.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021b. URL https://openreview.net/forum?id=7Bywt2mQsCe.

Shulin Huang, Linyi Yang, Yan Song, Shuang Chen, Leyang Cui, Ziyu Wan, Qingcheng Zeng, Ying Wen, Kun Shao, Weinan Zhang, Jun Wang, and Yue Zhang. Thinkbench: Dynamic out-of-distribution evaluation for robust llm reasoning, 2025. URL https://arxiv.org/abs/2502.16268.

Alexey Ignatiev, Antonio Morgado, and Joao Marques-Silva. Rc2: an efficient maxsat solver. *Journal on Satisfiability, Boolean Modelling and Computation*, 11(1):53–64, 2019. doi: 10.3233/SAT190116. URL https://journals.sagepub.com/doi/abs/10.3233/SAT190116.

Mehran Kazemi, Bahare Fatemi, Hritik Bansal, John Palowitch, Chrysovalantis Anastasiou, Sanket Vaibhav Mehta, Lalit K. Jain, Virginia Aglietti, Disha Jindal, Peter Chen, Nishanth Dikkala, Gladys Tyen, Xin Liu, Uri Shalit, Silvia Chiappa, Kate Olszewska, Yi Tay, Vinh Q. Tran, Quoc V. Le, and Orhan Firat. Big-bench extra hard, 2025. URL https://arxiv.org/abs/2502.19187.

Bill Yuchen Lin, Ronan Le Bras, Kyle Richardson, Ashish Sabharwal, Radha Poovendran, Peter Clark, and Yejin Choi. Zebralogic: On the scaling limits of llms for logical reasoning, 2025. URL https://arxiv.org/abs/2502.01100.

Joao Marques-Silva. Minimal unsatisfiability: Models, algorithms and applications (invited paper). In *Proceedings of the 2010 40th IEEE International Symposium on Multiple-Valued Logic*, ISMVL '10, pp. 9–14, USA, 2010. IEEE Computer Society. ISBN 9780769540245. doi: 10.1109/ISMVL.2010.11. URL https://doi.org/10.1109/ISMVL.2010.11.

Carlos Mencía, Alessandro Previti, and Joao Marques-Silva. Literal-based mcs extraction. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, pp. 1973–1979. AAAI Press, 2015. ISBN 9781577357384.

Philipp Mondorf and Barbara Plank. Liar, liar, logical mire: A benchmark for suppositional reasoning in large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7114–7137, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.404. URL https://aclanthology.org/2024.emnlp-main.404/.

OpenAI. OpenAI o3-mini System Card. https://cdn.openai.com/o3-mini-system-card-feb10.pdf, January 2025. Accessed: 2025-04-22.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

Daniel Selsam, Matthew Lamm, Benedikt Bünz, Percy Liang, Leonardo de Moura, and David L. Dill. Learning a SAT solver from single-bit supervision. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJMC_iA5tm.

Zhihong Shao and et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Johan Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew Kyle Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, Cesar Ferri, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Christopher Waites, Christian Voigt, Christopher D Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, C. Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodolà, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germàn Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Xinyue Wang, Gonzalo Jaimovitch-Lopez, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Francis Anthony Shevlin, Hinrich Schuetze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B Simon, James Koppel, James Zheng, James Zou, Jan Kocon, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh Dhole, Kevin Gimpel, Kevin Omondi, Kory Wallace Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros-Colón, Luke Metz, Lütfi Kerem Senel, Maarten Bosma, Maarten Sap, Maartje Ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramirez-Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael Andrew Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan Andrew Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter W Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan Le Bras, Rosanne

Liu, Rowan Jacobs, Rui Zhang, Russ Salakhutdinov, Ryan Andrew Chi, Seungjae Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel Stern Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima Shammie Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven Piantadosi, Stuart Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsunori Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Venkatesh Ramasesh, vinay uday prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Sophie Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=uyTL5Bvosj. Featured Certification.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13003–13051, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.824. URL https://aclanthology.org/2023.findings-acl.824/.

P Team, Xinrun Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, King Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, Chujie Zheng, Kaixin Deng, Shawn Gavin, Shian Jia, Sichao Jiang, Yiyan Liao, Rui Li, Qinrui Li, Sirun Li, Yizhi Li, Yunwen Li, David Ma, Yuansheng Ni, Haoran Que, Qiyao Wang, Zhoufutu Wen, Siwei Wu, Tyshawn Hsing, Ming Xu, Zhenzhu Yang, Zekun Moore Wang, Junting Zhou, Yuelin Bai, Xingyuan Bu, Chenglin Cai, Liang Chen, Yifan Chen, Chengtuo Cheng, Tianhao Cheng, Keyi Ding, Siming Huang, Yun Huang, Yaoru Li, Yizhe Li, Zhaoqun Li, Tianhao Liang, Chengdong Lin, Hongquan Lin, Yinghao Ma, Tianyang Pang, Zhongyuan Peng, Zifan Peng, Qige Qi, Shi Qiu, Xingwei Qu, Shanghaoran Quan, Yizhou Tan, Zili Wang, Chenqing Wang, Hao Wang, Yiya Wang, Yubo Wang, Jiajun Xu, Kexin Yang, Ruibin Yuan, Yuanhao Yue, Tianyang Zhan, Chun Zhang, Jinyang Zhang, Xiyue Zhang, Xingjian Zhang, Yue Zhang, Yongchi Zhao, Xiangyu Zheng, Chenghua Zhong, Yang Gao, Zhoujun Li, Dayiheng Liu, Qian Liu, Tianyu Liu, Shiwen Ni, Junran Peng, Yujia Qin, Wenbo Su, Guoyin Wang, Shi Wang, Jian Yang, Min Yang, Meng Cao, Xiang Yue, Zhaoxiang Zhang, Wangchunshu Zhou, Jiaheng Liu, Qunshu Lin, Wenhao Huang, and Ge Zhang. Supergpqa: Scaling llm evaluation across 285 graduate disciplines, 2025. URL https://arxiv.org/abs/2502.14739.

Qwen Team. Qwq: Reflect deeply on the boundaries of the unknown, November 2024. URL https://qwenlm.github.io/blog/qwq-32b-preview/.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. Trl: Transformer reinforcement learning. https://github.com/huggingface/trl, 2020.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=y10DM6R2r3.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=_VjQlMeSB_J.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh-Agrawal, Sandeep Singh Sandha, Siddartha Venkat Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. Livebench: A challenging, contamination-limited LLM benchmark. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=sKYHBTAxVa.

Chulin Xie, Yangsibo Huang, Chiyuan Zhang, Da Yu, Xinyun Chen, Bill Yuchen Lin, Bo Li, Badih Ghazi, and Ravi Kumar. On memorization of large language models in logical reasoning. In *The 4th Workshop on Mathematical Reasoning and AI at NeurIPS'24*, 2024. URL https://openreview.net/forum?id=mxX8WdPCx9.

Tian Xie, Zitian Gao, Qingnan Ren, Haoming Luo, Yuqian Hong, Bryan Dai, Joey Zhou, Kai Qiu, Zhirong Wu, and Chong Luo. Logic-rl: Unleashing llm reasoning with rule-based reinforcement learning, 2025. URL https://arxiv.org/abs/2502.14768.

Tong Yu, Yongcheng Jing, Xikun Zhang, Wentao Jiang, Wenjie Wu, Yingjie Wang, Wenbin Hu, Bo Du, and Dacheng Tao. Benchmarking reasoning robustness in large language models, 2025. URL https://arxiv.org/abs/2503.04550.

A GENERATE CNF PAIR

The CNF generation algorithm is adapted from Selsam et al. (2019)[1]. The procedure takes four parameters: the number of variables n, the number of clauses m, the probability p_{k_2} of generating a unit clause, and the geometric distribution parameter p_{geo} that controls clause length. It first builds an unsatisfiable CNF formula by repeatedly sampling random clauses until the conjunction is unsatisfiable (verified by a SAT solver). It then derives a satisfiable partner by iteratively flipping literal polarities until the formula becomes satisfiable. This paired generation ensures that each satisfiable instance has a closely related unsatisfiable counterpart, enabling controlled evaluation across both cases.

Algorithm 1 Generate CNF Pair (Satisfiable and Unsatisfiable)

```
procedure GENCNFPAIR(n, m, p_{k_2}, p_{geo})
    unsat_clauses ← ∅
    while solve(unsat_clauses) do                          ▷ Generate unsatisfiable CNF
        unsat_clauses ← ∅
        while |unsat_clauses| < m do
            k ← 1 if rand() < p_{k_2} else 2 + Geometric(p_{geo})
            k ← min(k, n)
            clause ← RandomClause(n, k)                    ▷ Random literals with polarity
            unsat_clauses ← unsat_clauses ∪ {clause}
        end while
    end while
    sat_clauses ← unsat_clauses
    while ¬solve(sat_clauses) do                           ▷ Convert to satisfiable CNF
        sat_clauses ← RandomFlipClause(sat_clauses)        ▷ Flip literal polarities
    end while
    return unsat_clauses, sat_clauses
end procedure
```

[1] https://github.com/dselsam/neurosat/blob/master/python/gen_sr_dimacs.py

B CNF TO NARRATIVE CONVERSION

This appendix details the deterministic procedure that converts a CNF instance into the `Story` and `DualStory` question formats. The conversion preserves the logical structure of the original CNF while wrapping it in a cookie-day narrative. A CNF object has two fields: `num_vars` (n) and `clauses` (a list of clauses, each a list of integer literals in DIMACS convention, e.g. `[1, -2, 3]` for $x_1 \vee \neg x_2 \vee x_3$).

Variable and entity mapping. Given n variables and m clauses, the procedure deterministically assigns: a chef name, n cookie names (drawn from a predefined pool), and m friend names. Each variable x_i maps to a cookie; positive literal $+i$ maps to the "crunchy" variant, and negative literal $-i$ maps to the "chewy" variant.

Story format (OR semantics). Each clause becomes a friend's preference list connected by disjunction: "*Friend is happy if she gets crunchy A, chewy B, or crunchy C.*" The full formula is satisfiable iff every friend can be made happy.

DualStory format (AND semantics). Each clause is negated and presented as a dislike combination connected by conjunction: "*Friend is unhappy only if she is served chewy A and crunchy B and chewy C.*" This requires the model to track De Morgan's transformation.

Pseudocode. Algorithm 2 gives the `Story` conversion; the `DualStory` variant differs only in the friend description template (replacing "happy if ... or" with "unhappy only if ... and") and swapping literal polarities in the phrase mapping. The full object-oriented implementation covering all four formats and all five problem types is available in `satquest/question.py`.

Algorithm 2 CNF → `Story` Prompt (simplified)

```
procedure CnfToStory(cnf, problem_type)
    n, m ← cnf.num_vars, |cnf.clauses|
    chef, cookies[1..n], friends[1..m] ← AssignNames(n, m)
    for i ← 1 to n do                                   ▷ Literal → phrase
        phrase[+i] ← "crunchy cookies[i]"
        phrase[−i] ← "chewy cookies[i]"
    end for
    lines ← ["It's cookie day! Chef chef bakes n cookies ..."]
    for k ← 1 to m do                                   ▷ Clause → friend preference
        opts ← [phrase[ℓ] for ℓ ∈ cnf.clauses[k]]
        Append "k. friends[k] wants: opts[1], ..." to lines
    end for
    Append task instruction for problem_type to lines
    return lines joined by newline
end procedure
```

C PROMPT DETAILS

In this appendix, we present the complete prompt templates and representative example outputs used to instruct LLMs to produce binary string responses for SAT-based reasoning tasks. These templates specify the precise wording, formatting requirements, and task descriptions used during both evaluation and reinforcement fine-tuning phases. Our prompt construction strategy adapts and extends established templates from OpenAI's `simple-eval`[2] and `DeepSeek-R1` (DeepSeek-AI et al., 2025). The following subsections detail how we systematically assemble task prompts and incorporate system-level configurations for both evaluation (Sec. 3) and fine-tuning (Sec. 4) experiments. Fig. 6 illustrates the general structure of the task prompt used throughout our experiments.

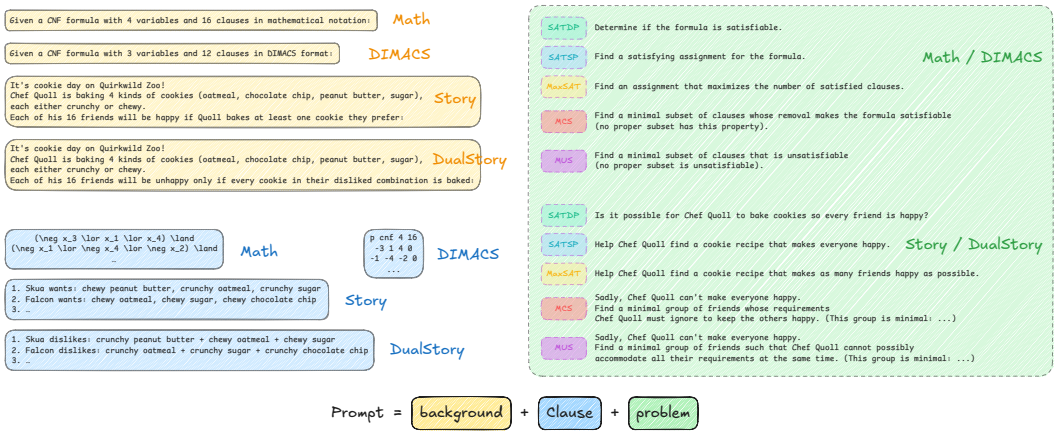

Figure 6: Task Prompt Construction in `SATQuest`

C.1 EVALUATION PROMPT TEMPLATE

In Sec. 3, we employ the following prompt template that includes Chain-of-Thought (CoT) guidance and specific output format instructions to ensure consistent responses across all models.

For vanilla model, add system prompt: `"You are a helpful assistant."`.

```
Solve the following problem step by step. The last line of your response should be of the form Answer:
    ↪ $ANSWER (without quotes) where $ANSWER is the answer to the problem.

Given a CNF formula with 4 variables and 16 clauses in DIMACS format:

p cnf 4 16
-3 1 4 0
-1 -4 -2 0
-1 4 2 0
2 -1 -4 0
2 -3 4 0
-3 -4 1 0
-4 1 -3 0
1 2 -4 0
-3 -2 1 0
4 -3 1 0
-1 -3 2 0
2 -3 4 0
-1 -2 3 0
2 3 4 0
2 3 1 0
1 3 -4 0

Find a satisfying assignment for the formula.
Output a binary string of length 4 ('1' for true, '0' for false).

Remember to put your answer on its own line after "Answer:", and you do not need to use a \boxed command.
```

[2] https://github.com/openai/simple-evals/blob/0f2cf3/math_eval.py

C.2 REINFORCEMENT FINE-TUNING PROMPT TEMPLATE

For Sec. 4, we adapt the prompt to encourage explicit CoT reasoning by combining system instructions with a structured user message format.

```
<|im_start|>system
You are a helpful AI Assistant that provides well-reasoned and detailed responses. You first think about
      ↪ the reasoning process as an internal monologue and then provide the user with the answer. Respond
      ↪ in the following format: <think>\n...\n</think>\n<answer>\n...\n</answer>
<|im_end|>
<|im_start|>user
Given a CNF formula with 4 variables and 16 clauses in DIMACS format:

p cnf 4 16
-3 1 4 0
-1 -4 -2 0
-1 4 2 0
2 -1 -4 0
2 -3 4 0
-3 -4 1 0
-4 1 -3 0
1 2 -4 0
-3 -2 1 0
4 -3 1 0
-1 -3 2 0
2 -3 4 0
-1 -2 3 0
2 3 4 0
2 3 1 0
1 3 -4 0

Find a satisfying assignment for the formula.
Output a binary string of length 4 ('1' for true, '0' for false).

Show your work in <think> </think> tags. And return the final answer in <answer> </answer> tags, for
      ↪ example <answer> 0101 </answer>.
<|im_end|>
<|im_start|>assistant
```

D PROMPT EXAMPLES

For more prompt examples, please refer to the `https://sat-quest-anon.vercel.app/examples/#prompt-gallery` website.

SATSP-Math Prompt Example

```
Given a CNF formula with 3 variables and 4 clauses in mathematical notation:

(x_1 \lor \neg x_2 \lor x_3) \land (\neg x_1 \lor x_2 \lor x_3) \land (x_1 \lor x_2) \land (\neg x_3)

Find a satisfying assignment for the formula.
Output a binary string of length 3 ('1' for true, '0' for false).
```

SATSP-DIMACS Prompt Example

```
Given a CNF formula with 3 variables and 4 clauses in DIMACS format:

p cnf 3 4
1 -2 3 0
-1 2 3 0
1 2 0
-3 0

Find a satisfying assignment for the formula.
Output a binary string of length 3 ('1' for true, '0' for false).
```

SATSP-Story Prompt Example

```
It's cookie day on Quirkwild Zoo!
Chef Orion is baking 3 kinds of cookies (chocolate comet, maple moon, ginger nebula), each either crunchy
      ↪ or chewy.
Each of his 4 friends will be happy if Orion bakes at least one cookie they prefer:


1. Aquila wants: crunchy chocolate comet, chewy maple moon, crunchy ginger nebula
2. Borealis wants: chewy chocolate comet, crunchy maple moon, crunchy ginger nebula
3. Cygnus wants: crunchy chocolate comet, crunchy maple moon
4. Draco wants: chewy ginger nebula

Find a satisfying assignment for the formula.
Output a binary string of length 3 ('1' for true, '0' for false).
```

SATSP-DualStory Prompt Example

```
It's cookie day on Quirkwild Zoo!
Chef Orion is baking 3 kinds of cookies (chocolate comet, maple moon, ginger nebula), each either crunchy
      ↪ or chewy.
Each of his 4 friends will be unhappy only if every cookie in their disliked combination is baked:


1. Aquila dislikes: chewy chocolate comet + crunchy maple moon + chewy ginger nebula
2. Borealis dislikes: crunchy chocolate comet + chewy maple moon + chewy ginger nebula
3. Cygnus dislikes: chewy chocolate comet + chewy maple moon
4. Draco dislikes: crunchy ginger nebula

Find a satisfying assignment for the formula.
Output a binary string of length 3 ('1' for true, '0' for false).
```

E CASE STUDIES

Our analysis reveals that both `DeepSeek-R1` and `QwQ-32B` predominantly reason directly with the `DIMACS` format rather than translating it into formal mathematical notation. Throughout the four case studies, segments highlighted in green demonstrate instances where the models correctly comprehend the CNF structure, especially regarding the disjunctive (`OR`) semantics within clauses. Conversely, red-highlighted sections expose critical reasoning flaws: models frequently misinterpret the disjunctive nature of clauses or conflate the semantic meaning of positive and negative literals (representing true versus false values). These observations suggest that direct reasoning within the `DIMACS` representation, without proper symbolic abstraction, frequently results in semantic drift and format-specific reasoning errors.

Case1: DeepSeek-R1 @ SATSP-DIMACS

```
Okay, so I need to find a satisfying assignment for this CNF formula with 9 variables and 36 clauses.

...

So, for each clause, at least one of the literals must be true in the assignment.

...

For each variable, check if all occurrences are positive, all negative, or mixed.

Let's start with variable 1:

Looking through the clauses where variable 1 appears.

Clause 5: -1 2 -6 5 8. So -1 here.

Clause 6:8 1 9 -6. Positive 1.

...

Clause1:  -2 -9 6 -> with 6=0, this becomes -2 AND -9 -> so both must be true.  So variables2 and 9 must be 0.

...

All clauses seem to be satisfied. So the assignment is valid.
```

Full prompt and response: [Link hidden for anonymity]

Case2: DeepSeek-R1 @ SATSP-DIMACS

```
Okay, so I need to find a satisfying assignment for this CNF formula with 10 variables and 40 clauses.
    ↪ Hmm, let's think. The variables are numbered 1 through 10, and each line in the DIMACS format
    ↪ represents a clause.
The clauses are ORs of literals, and the entire formula is the AND of all these clauses.
So the assignment has to make at least one literal in each clause true.

...

Clause 5: -1 -> negative occurrence.

Clause 6: 1 -> positive.

...

Ok, moving on.

Clause28:-5 -1 -> both must be false.  So variable5=0 and variable1=0.

Wait, ...

All other clauses seem to have been handled earlier. So this assignment should work. Thus, the binary
    ↪ string is 0001010100.
```

Full prompt and response: [Link hidden for anonymity]

Case3: QwQ-32B @ SATSP-DIMACS

```
Okay, so I have this CNF formula with 9 variables and 36 clauses, and I need to find a satisfying
    ↪ assignment.

...

Let me start by writing down some of the clauses:

1.  5 3 -4 9 0 -> variables 5,3 are true, 4 is false, 9 is true.
2.  -1 -2 7 -6 0 -> variables 1 and 2 are false, 7 and 6 are false?
Wait, no, the clause is satisfied if at least one literal is true.
So for this clause, either -1 (1 is false), -2 (2 false), 7 (true), or -6 (6 false). So ...

...

So this should be a valid assignment.
```

Full prompt and response: [Link hidden for anonymity]

Case4: QwQ-32B @ SATSP-DIMACS

```
Okay, so I need to find a satisfying assignment for this CNF formula with 10 variables and 40 clauses.
    ↪ Hmm, that's a bit of a challenge.
Let me think about how to approach this.
Since it's a CNF formula, each clause has to be satisfied by at least one literal being true.

...

Looking at the clauses:

The first few clauses are:

-9 4 -10 0 -> clause 1

6 -2 -8 -1 0 -> clause 2

....

Clause4:  -1 -2 -4 -> since variables1 and2 are 0, this clause requires -4 (so variable4 must be 0).

...

Yes, that works.

Thus, the satisfying assignment is 1000101101.
```

Full prompt and response: [Link hidden for anonymity]

F EVALUATION CONFIGS

In all experiments in Sec. 3, we use APIs for evaluation. The specific model APIs and IDs are listed in Table 2, and the parameters for vanilla models are shown in Table 3. For reasoning models, we use their default configurations.

Table 2: Model API / ID of our Evaluated Models.

Model Name	API / ID
o3-mini	Azure API[3]: o3-mini-2025-01-31
GPT-4.1	Azure API: gpt-4.1-2025-04-14
DeepSeek-R1	VolcEngine API[4]: deepseek-r1
DeepSeek-V3-0324	VolcEngine API: deepseek-v3-0324
DeepSeek-R1-Distill-Qwen-7B	VolcEngine API: deepseek-r1-distill-qwen-7b
DeepSeek-R1-Distill-Qwen-32B	VolcEngine API: deepseek-r1-distill-qwen-32b
QwQ-32B	Alibaba Cloud API: qwq-32b-plus[5]
Qwen2.5-7B-Instruct	Alibaba Cloud API: qwen2.5-7b-instruct
Qwen2.5-32B-Instruct	Alibaba Cloud API: qwen2.5-32b-instruct

Table 3: Evaluation Parameters for Vanilla Models

Parameter	Value
temperature	0.6
top_p	1.0
max_tokens	16384

[3] *Azure OpenAI API*: https://azure.microsoft.com/en-us/products/ai-services/openai-service

[4] *ByteDance VolcEngine AI platform API*: https://www.volcengine.com/

[5] *Alibaba Cloud API*: https://www.alibabacloud.com/en

G Reinforcement Fine-Tuning Configs

G.1 Format Reward Functions

The format reward functions are adapted from the `huggingface/Open-R1` library[6].

```
def tag_count_reward(completions, **kwargs) -> list[float]:
    def count_tags(text: str) -> float:
        count = 0.0
        if text.count("<think>") == 1:
            count += 0.25
        if text.count("</think>") == 1:
            count += 0.25
        if text.count("<answer>") == 1:
            count += 0.25
        if text.count("</answer>") == 1:
            count += 0.25
        return count

    contents = [completion[0]["content"] for completion in completions]
    return [count_tags(c) for c in contents]

def format_reward(completions, **kwargs):
    _PATTERN = re.compile(r"<think>.*?</think>\s?<answer>.*?</answer>", flags=re.DOTALL)
    completion_contents = [completion[0]["content"] for completion in completions]
    rewards = []
    for c in completion_contents:
        text = str(c)
        total_len = len(text)
        if total_len == 0:
            rewards.append(0.0)
            continue

        m = _PATTERN.search(text)
        match_len = len(m.group()) if m else 0
        rewards.append(match_len / total_len)

    return rewards
```

G.2 Training Parameters

The training parameters used for GRPO of the `Qwen2.5-7B-Instruct` model are summarized in Table 4.

Table 4: Training Parameters for GRPO

Parameter	**Value**
`learning_rate`	0.000002
`batch_size`	`num_generations` $\times 8 = 128$
`max_grad_norm`	0.3
`num_iterations`	1
`beta`	0.01
`max_steps`	500
`max_prompt_length`	2048
`max_completion_length`	8192
`mask_truncated_completions`	`True`
`num_generations`	16
`temperature`	1.0
`scale_rewards`	`True`

G.3 Experiments Compute Resources

All experiments were conducted on a single server node equipped with 8 `NVIDIA A100 80GB` GPUs, 2 `Intel Xeon Platinum 8350C` CPU, and `1600GB` memory. We allocated 4 GPUs for training and 4 GPUs for *VLLM* inference. The training time for `GRPO@SATSP-Math`,

[6] `https://github.com/huggingface/open-r1/blob/main/src/open_r1/rewards.py`

GRPO@MaxSAT-Math, and GRPO@SATSP-Story was approximately 30 hours, 26 hours, and 9 hours, respectively.

H ADDITIONAL FIGURES

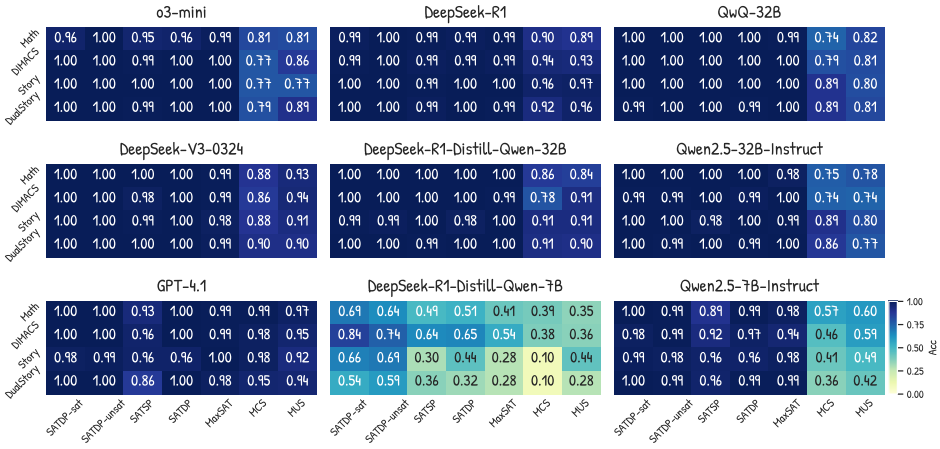

Figure 7: Format accuracy heatmaps showing LLM performance breakdown by *problem type* (columns) and *question format* (rows).

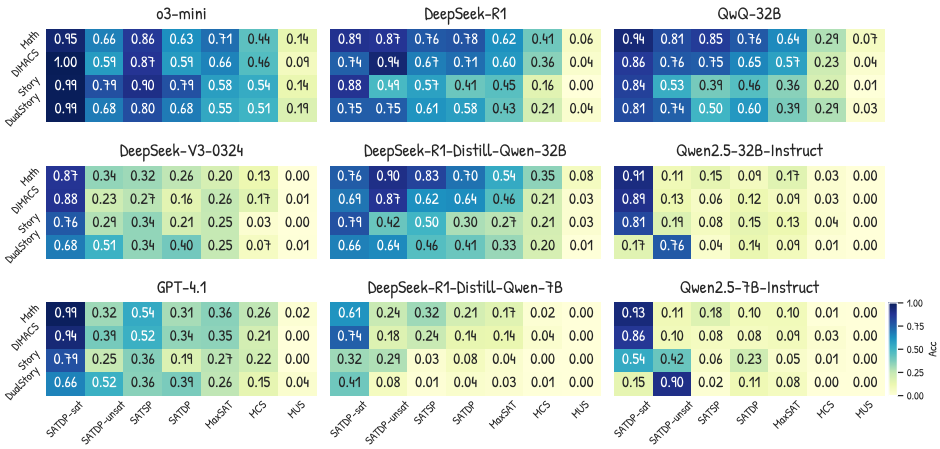

Figure 8: All accuracy heatmaps (including `SATDP-sat` and `SATDP-unsat`) showing LLM performance breakdown by *problem type* (columns) and *question format* (rows).

I HALLUCINATION AUDIT

To substantiate the claim that high-complexity instances trigger hallucination-dominated failures across models, we conducted a dedicated error-type audit following the protocol below.

Protocol. From the 🤗 ANON/SATQuest evaluation set (140 CNF pairs), we selected the 20 pairs whose pair ID is divisible by 7. For each pair we instantiated four Math-format tasks (SATSP, MaxSAT, MCS, MUS), yielding 80 instances. We collected responses from three reasoning models: o3-mini, QwQ-32B, and DeepSeek-R1. Error types were labelled by Claude-Sonnet-4.5 in an LLM-as-a-Judge protocol, with 10% of cases manually audited by a human expert to verify label quality.

Error taxonomy. We adopt a multi-label taxonomy (a single incorrect response may exhibit multiple types):

- **Type A** (Logic Error): a mostly honest, step-by-step reasoning attempt with a concrete calculation or tracking mistake.
- **Type B** (Loop): repetitive reasoning without real progress.
- **Type C** (Fabricated Execution): the model fabricates intermediate solver states or verification steps (e.g., asserting that a clause is satisfied when it is not).
- **Type D** (Invented Shortcut): the model invents non-existent logical rules or "shortcuts" to bypass the combinatorial reasoning required.

Results by task. Table 5 summarizes the failure-mode distribution. Across all three models, complex diagnostic/optimization tasks (MaxSAT, MCS, MUS) are dominated by Type C hallucination (fabricated execution), often supplemented by Type D shortcuts. Simple logic mistakes without fabrication (Type A) account for only a small fraction of errors on these tasks.

Table 5: Failure-mode distribution by task (% of total samples). Categories are non-exclusive (multi-label); rows may sum to $>100\%$.

Model	Task	A	B	C	D	C∪D
o3-mini	SATSP	5.0	0.0	10.0	0.0	10.0
	MaxSAT	45.0	0.0	45.0	0.0	45.0
	MCS	5.0	0.0	75.0	10.0	75.0
	MUS	0.0	0.0	50.0	0.0	50.0
QwQ-32B	SATSP	35.0	0.0	60.0	0.0	60.0
	MaxSAT	15.0	0.0	25.0	0.0	25.0
	MCS	0.0	0.0	95.0	5.0	100.0
	MUS	0.0	0.0	100.0	0.0	100.0
DeepSeek-R1	SATSP	20.0	0.0	65.0	0.0	65.0
	MaxSAT	10.0	0.0	75.0	0.0	75.0
	MCS	5.0	0.0	95.0	40.0	95.0
	MUS	35.0	0.0	70.0	30.0	80.0

Hallucination vs. complexity. Table 6 buckets instances by solver decisions and reports the hallucination rate (Type C or D) within each bucket. The pattern is consistent across architectures: as solver-measured complexity increases, the fraction of failures attributed to hallucination rises sharply.

These results provide quantitative evidence that, under low complexity, models sometimes fail due to ordinary logic errors (Type A), whereas under medium and especially high complexity, failures shift to hallucinated solver states or invented reasoning shortcuts (Types C/D). This shift appears consistently for all three models, even though their overall accuracy levels differ. All sampling scripts, raw model outputs, and judge logs are released with the SATQuest repository.

Table 6: Hallucination rate (Type C∪D, % of total samples) vs. instance complexity (solver decisions).

Complexity (decisions)	o3-mini	QwQ-32B	DeepSeek-R1
Low (3–10)	10.3	51.7	65.5
Medium (10–76)	60.0	64.0	84.0
High (76–903)	69.2	100.0	88.5

J ADDITIONAL EXPERIMENTS

Scaling to Larger Models. We applied the same RFT protocol to `Qwen2.5-14B-Instruct`, following the GRPO setup used for the 7B model. As shown in Table 7, the 14B model exhibits consistent improvements on all in-domain tasks, confirming `SATQuest`'s scalability. However, the cross-format generalization gap persists, suggesting that the limitation stems from current RFT methods rather than model size.

Model	Evaluation on `SATSP-Math`	Evaluation on `SATSP-Story`	Evaluation on `MaxSAT-Math`
`Qwen2.5-14B-Instruct`	12.86 ± 1.91	4.38 ± 1.77	12.94 ± 1.35
+ GRPO@`SATSP-Math`	70.80 ± 3.14	0.63 ± 0.80	17.05 ± 1.59
+ GRPO@`SATSP-Story`	11.96 ± 2.83	11.25 ± 1.31	14.20 ± 3.18
+ GRPO@`MaxSAT-Math`	41.52 ± 2.30	1.25 ± 0.63	46.33 ± 2.24
+ GRPO@`SATSP-Math`/`Story`	64.38 ± 3.71	10.54 ± 2.08	21.96 ± 2.88

Table 7: Additional RFT results on `Qwen2.5-14B-Instruct` (mean $\pm$ std over 8 runs).

Generalization Beyond SATQuest. To examine whether `SATQuest` training benefits broader reasoning tasks, we evaluated fine-tuned models on two external benchmarks: *AIME2024* (pass@4) and *DeepMind Mathematics*. As summarized in Table 8, `SATQuest`-trained models show consistent improvements across both datasets, indicating enhanced mathematical reasoning rather than overfitting to `SATQuest` formats.

Model	*AIME2024* (pass@4)	*DeepMind Mathematics*
`Qwen2.5-14B-Instruct`	24.17 ± 5.69	75.70 ± 0.74
+ GRPO@`SATSP-Math`	25.83 ± 3.19	79.15 ± 0.95
+ GRPO@`SATSP-Story`	20.84 ± 4.19	79.15 ± 0.55
+ GRPO@`MaxSAT-Math`	24.16 ± 1.67	77.72 ± 0.49
+ GRPO@`SATSP-Math`/`Story`	25.00 ± 4.30	79.30 ± 0.55

Table 8: Out-of-domain generalization on external reasoning benchmarks.

Summary. These extended results confirm that: (1) `SATQuest`-derived RFT scales to larger models; (2) improvements extend to independent reasoning benchmarks; and (3) the remaining cross-format gap reflects an intrinsic challenge in current RFT methods rather than dataset bias or model capacity.