SUPERNOVA: Eliciting General Reasoning in LLMs with Reinforcement Learning on Natural Instructions

Anonymous ACL submission

Abstract

Reinforcement Learning with Verifiable Rewards (RLVR) has substantially improved reasoning in formal domains such as mathematics and code, but extending these gains beyond STEM remains challenging. Extending RLVR beyond STEM is fundamentally constrained by the lack of high-quality verifiable training data. In this work, we introduce SUPERNOVA, a framework for curating RLVR data from natural instruction datasets, which are a rich source of expert-annotated data but are underexplored for RLVR training. Through 100+ controlled RL experiments, we systematically study how to utilize these dataset for RLVR and how data curation decisions affect downstream reasoning performance . In particular, we investigate three data designs: (a) source task selection, (b) task mixing, and (c) synthetic interventions. Our analysis reveals that source task selection has a significant impact on downstream reasoning performance. Moreover, selecting tasks based on their performance for individual target tasks outperforms strategies based on overall average performance and synthetic interventions do not improve reasoning. Guided by these insights, we construct SUPERNOVA, a high-quality RLVR dataset of 25K instances curated from natural instruction datasets. We show that training Qwen3-0.6B on SUPERNOVA outperforms the base Qwen3-0.6B, yielding a relative gain of 64.4pp on BigBench Extra Hard (BBEH), a challenging benchmark comprising 23 complex reasoning tasks. Importantly, we find that gains from SUPERNOVA generalize to unseen benchmarks, larger model scales, and newer model families. Overall, our findings provide practical insights for curating human-annotated resources to extend RLVR to general reasoning. [1]

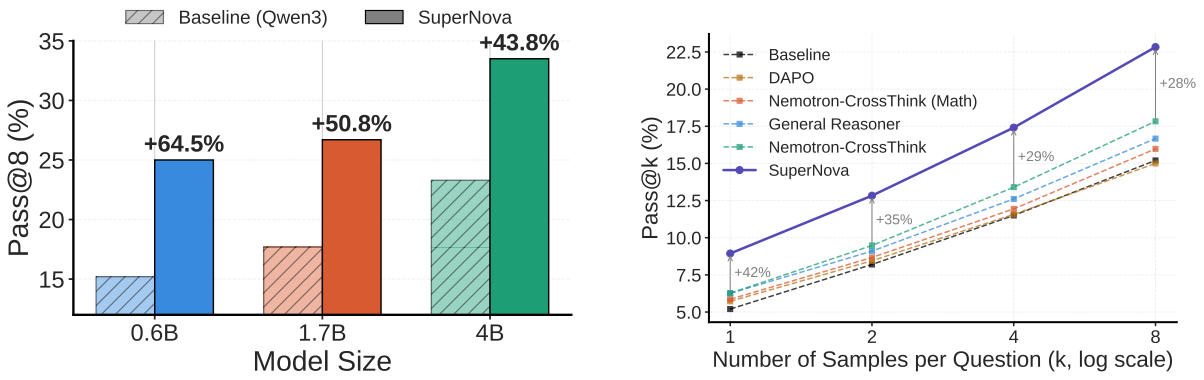
1 Introduction

Large language models (LLMs) have shown remarkable progress in reasoning capabilities for formal domains such as mathematics and code (Guo et al., 2025; Lambert et al., 2024; Guha et al., 2025; Ma et al., 2025; Zeng et al., 2025; Hu et al., 2025; Chen et al., 2025). A vast majority of these advances leverage reinforcement learning with verifiable rewards (RLVR) which relies on the ability to verify model outputs against a ground-truth final answer (Guo et al., 2025). The success of RLVR in STEM domains is closely tied to the abundance of high-quality verifiable data, such as MATH (Hendrycks et al., 2021), competitions (e.g., CodeForces, Art of Problem Solving) and forums (e.g., StackOverflow). Together, these resources have enabled the rapid adoption of RLVR for mathematical and code reasoning (Chen et al., 2025; Yu et al., 2025; Akter et al., 2026; Hu et al., 2025).

However, reasoning in real-world settings extends beyond STEM problem solving and requires a broader spectrum of capabilities, including temporal reasoning, causal inference, and logical deduction (Newell et al., 1972; Johnson-Laird, 2010; Griffiths, 2020). Yet recent work shows that training models on mathematical and coding data does not necessarily improve general reasoning capabilities (Bhaskar et al., 2025; Huan et al., 2025; Cheng et al., 2026). In particular, models trained on high-quality mathematical and coding reasoning data often achieve substantial gains on formal reasoning benchmarks while simultaneously degrading performance on more general reasoning tasks. For example, OpenReasoner-7B (Hu et al., 2025) and OpenThinker-7B (Guha et al., 2025) improve over their base models by more than 50% on challenging mathematical benchmarks such as AIME24 (Zhang and Math-AI, 2024), yet suffer performance drops of up to 8% on complex reasoning tasks in BBEH (Kazemi et al., 2025).

A key bottleneck in extending RLVR beyond STEM domains is the lack of high-quality verifiable training data that spans diverse reasoning skills. Prior work addresses this challenge through

[1] We will release the data and models upon acceptance.



(a) (b)

Figure 1: **SuperNova shows strong reasoning performance.** Training with our curated SuperNova data leads to consistent pass@k improvements on a challenging benchmark, BBEH-test. (a) SuperNova shows consistent performance gains across various model sizes from the Qwen3 family. (b) SuperNova shows superior performance to existing reasoning datasets across varying values of k under compute-matched comparisons.

domain-specific reward designs and extensive data filtering pipelines to curate reasoning data spanning diverse domains (Cheng et al., 2026; Ma et al., 2025; Akter et al., 2026). However, these approaches often depend on noisy web-scale corpora (Ma et al., 2025) or specialized datasets tailored to particular domains (Cheng et al., 2026), limiting their applicability to understudied reasoning tasks. Moreover, collecting human-annotated data for RLVR is expensive and labor-intensive.

In contrast, large amounts of high-quality human-annotated data already exist in instruction-following datasets. Resources such as SuperNI (Wang et al., 2022) and FLAN (Wei et al., 2021) contain thousands of expert-curated tasks spanning event understanding, question generation, and other reasoning-intensive settings (Appendix Table 8). However, these datasets cannot be directly used for RLVR because (a) many tasks are open-ended and lack reliable verification signals, (b) not all tasks elicit strong reasoning behavior, and (c) the principles underlying effective RLVR data curation remain underexplored beyond STEM domains.

In this work, we introduce SuperNova, a multi-stage pipeline for curating high-quality RLVR data from natural instruction datasets (Figure 2). We conduct 100+ compute-matched RL experiments to systematically study the principles underlying RLVR data curation. We explore three key design choices: (a) source task selection, (b) task mixing strategies, and (c) synthetic interventions to enhance task quality. In particular, we find that (a) task selection has a large impact on downstream reasoning performance, (b) our novel mixing strategy: micro-mixing yields further gains, and (c) synthetic interventions on tasks do not improve reasoning performance. Finally, based on these insights we construct SuperNova, a curated corpus of 25K verifiable training instances spanning diverse reasoning types for RLVR (Figure 3). Training Qwen3 models across scales (0.6B–4B) on SuperNova yields substantial gains on BBEH (Kazemi et al., 2025), achieving relative gains up to 64.4pp, while outperforming existing reasoning datasets such as Nemotron-Crossthink (Akter et al., 2026) by 42pp at pass@1 (Figure 1). Models trained on SuperNova also generalize strongly to challenging reasoning benchmarks including BBH (Suzgun et al., 2023), MMLU-Pro (Wang et al., 2024), and Zebralogic (Lin et al., 2025), while exhibiting cross-model transfer across both different model families. Moreover, these gains remain consistent at larger values of k, highlighting the importance of principled RLVR data curation for improving reasoning capabilities in LLMs.

In summary, our contributions are as follows:

- **Natural Instructions can elicit reasoning.** We demonstrate that natural instruction datasets are a rich source of human-annotated data which can be used for RL training to improve complex reasoning.

- **SuperNova Framework.** We introduce a framework to curate high-quality RLVR data beyond STEM. Through our controlled experiments, we study the impact of diverse data designs including (a) task selection, (b) task

mixing, and (c) synthetic interventions.

- **Generalization of SUPERNOVA.** We curate a high-quality RLVR dataset of 25K instances. Importantly, we show that gains from SUPERNOVA generalize to evaluation benchmarks that were unseen during data curation. Specifically, our data curated from a small Qwen3-0.6B model generalizes to larger model sizes and newer model families.

2 Preliminaries

Reinforcement Learning with Verifiable Rewards (RLVR). RLVR is widely adopted for training LLMs for reasoning in domains that rely on automatically verifiable ground truth such as mathematics and code. Given an input-target pair (q, t), RLVR samples G rollouts $o_{i=1}^{G}$ from a behavior policy $\pi_{\theta_{\text{old}}}$ and optimizes the GRPO (Shao et al., 2024) objective:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{\substack{(q,t)\sim\mathcal{D} \\ \{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)}} \left[\frac{1}{G}\sum_{i=1}^{G} \min\left(\rho_i(\theta)\,\hat{A}_i,\ \text{clip}(\rho_i(\theta), 1-\epsilon, 1+\epsilon)\,\hat{A}_i\right)\right] \tag{1}$$

where $\rho_i(\theta) = \frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}$ is the importance sampling ratio. The group-centered advantage $\hat{A}_i$ for each output is computed as $\hat{A}_i = r_i - \frac{1}{G}\sum_{j=1}^{G} r_j$ where $r_i = r(o_i, q)$, the computed reward. Following (Yu et al., 2025), we skip the KL penalty to improve training efficiency in our experiments.

Task-Specific Instruction Datasets. Instruction-tuning datasets such as SuperNI (Wang et al., 2022), and Flan-Collection (Wei et al., 2021) are a collection of well-structured, distinct tasks spanning diverse reasoning abilities. These datasets are constructed from high-quality human supervision including task definitions, instructions and ground-truth annotations. We observe that these large instruction-tuning datasets often encode reasoning structures that are not explicitly annotated but can be inferred from the examples and task structure. Consider an instruction dataset $D = \{D_1, D_2, D_3 ... D_K\}$ comprising K tasks where each subset D_k is a well-defined task targeting a particular skill.

Problem Setup. In this work, we focus on the curation of high-quality training data to enable strong general reasoning capabilities via reinforcement learning. Given a pool of candidate datasets $D = \{D_1, D_2, \ldots, D_K\}$, a model M, and a training algorithm A, we seek a subset of tasks $S \subseteq D$ that maximizes downstream performance after training. Following the data curation formulations propose for SFT in math reasoning (Guha et al., 2025) and multimodal reasoning (Bansal et al., 2025), we define our objective as:

$$S^* = argmax_{S\subseteq D}\Phi(A(M, S), V) \tag{2}$$

where $A(M, S)$ denotes the model after applying A to M on the selected subset S, V is the validation set and ϕ measures downstream performance on V.

3 SUPERNOVA

We outline our SUPERNOVA framework (Figure 2), which consists of multiple stages: (a) task selection, which assesses the impact of source task choice (§ 3.1); (b) mixing, which identifies the best strategy to mix the diverse tasks (§ 3.2); and (c) data interventions, which aim to enhance the quality of our data (§ 3.3). Finally, we combine the best performing strategy from each step and curate SUPERNOVA, comprising of $25K$ verifiable training samples spanning 9 diverse reasoning types as shown in Figure 3. SUPERNOVA consists of 31 tasks selected from SuperNI with 25% of the tasks targeting temporal and causal reasoning.[2] We provide qualitative examples from SUPERNOVA in Appendix Table 4.

3.1 Task Selection

Extracting reasoning data from instructions. The quality of the input and the coverage of reasoning types is critical for determining the reasoning skills imparted to the LLM. For example, a LLM exposed to temporal graphs will excel in downstream temporal understanding tasks (Xiong et al., 2024). In this work, we leverage instruction-tuning data D to source diverse tasks D_k for general reasoning. Since, instructions are formatted for supervised-finetuning they are not directly usable for RLVR as they may incorporate hard to verify ground-truth. Thus, for every instruction p in D_k, we **reformat the instruction** to a verifiable question q. To further identify the most effective data from D, we sample 8 rollouts from model M for each q and compute the **per-question win-rate**.

[2] Most SuperNI tasks target various reasoning types, we consider the primary reasoning type identified by Claude-Opus-4.6 based on the task description. More details in App. § F.

Figure 2: **SUPERNOVA Framework:** We study the impact of various data design choices for RLVR data curation from natural instruction datasets. First, we study the impact of task selection on downstream reasoning performance. Then, we explore strategies to mix diverse tasks in source data. Finally, we examine whether synthetic data interventions can enhance data quality and improve downstream reasoning.

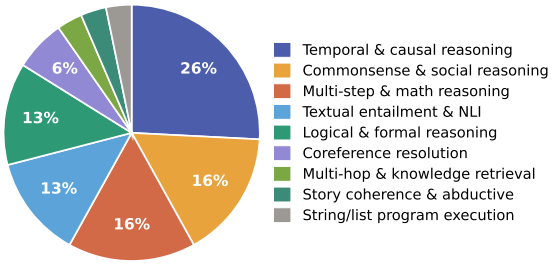


Figure 3: **SUPERNOVA Distribution.** SUPERNOVA comprises of 25K training samples that target 9 diverse reasoning types.

Finally, we remove all questions which are too easy for the model (win-rate=1) or too challenging (win-rate=0).

Task Ranking. For each task D_k, we define a task-utility score $u_k \in R$ that indicates how effective D_k is for RLVR training. Then, we rank the K tasks according to their utility scores, producing a ranking: $u_{D_1} > u_{D_2} > u_{D_3} > \cdots > u_{D_K})$. The task utility scores enable us to select high-quality tasks while downweighting poor and irrelevant tasks. We explore various approaches to compute task-utility: (a) we compute the semantic and lexical similarity between the task questions and the questions from our validation benchmark V; (b) we compute the difficulty of the task based on the average win-rate of the task under model M; and (c) we train model M on D_k and evaluate performance on V.

3.2 Mixing

After obtaining high-quality tasks, we determine how to combine them to construct an effective training mixture. Mixing strategy is a key design choice in data curation and prior work in LLM reasoning have shown to yield superior datasets by mixing subsets from various sources. Consider the K tasks from § 3.1 and number of tasks to be mixed $N \in \{1, 2, 4, 8, 16\}$, we want to determine the optimal value of N under two mixing strategies:

- **Macro Mixing:** Consider the ranking from § 3.1: $u_{D_1} > u_{D_2} > u_{D_3} .. > u_{D_K}$ where u_{D_k} is the macro average of model performance on V_{BBEH}. We select the top-ranked N tasks $u_{D_1} > u_{D_2} > u_{D_3} > \cdots > u_{D_N}$ for our mixture.

- **Micro Mixing:** Here, we leverage the sub-tasks of our V and produce a ranking for each sub-task V_i. Specifically, we define $u_k^{(i)}$ as the performance of model M trained on D_k and evaluated on sub-task V_i, yielding a per-sub-task ranking: $u_{D_1}^{(i)} > u_{D_2}^{(i)} > \cdots > u_{D_K}^{(i)}$ for each $V_i \in V$. We then select the top-ranked N tasks per sub-task and take the unique set of selected tasks for our mixture.

3.3 Data Interventions

Starting from the best mixture from §3.2, we assess whether we can enhance the data quality through targeted data interventions. RLVR datasets primarily focus on the questions since interventions on the target may hinder the verifiability of the answer. Thus, we apply a set of interventions to transform the difficulty of the questions while preserving the target answer. These interventions aim to increase the difficulty of the questions by introducing diverse reasoning types such very long-context dependency, information that prompts model to go against a strong prior or needle in haystack. Let $D_{\text{base}} = \{(q, t)\}$ be the base data with original question-target pairs. We apply an intervention I that transforms each question while preserving the target, producing $D' = \{(I(q), t)\}$. We provide the implementation details of applying these interventions in Appendix § E.

4 Experimental Setup

Training Data. We use SuperNI (Wang et al., 2022) as our data source. It consists of 1600 tasks

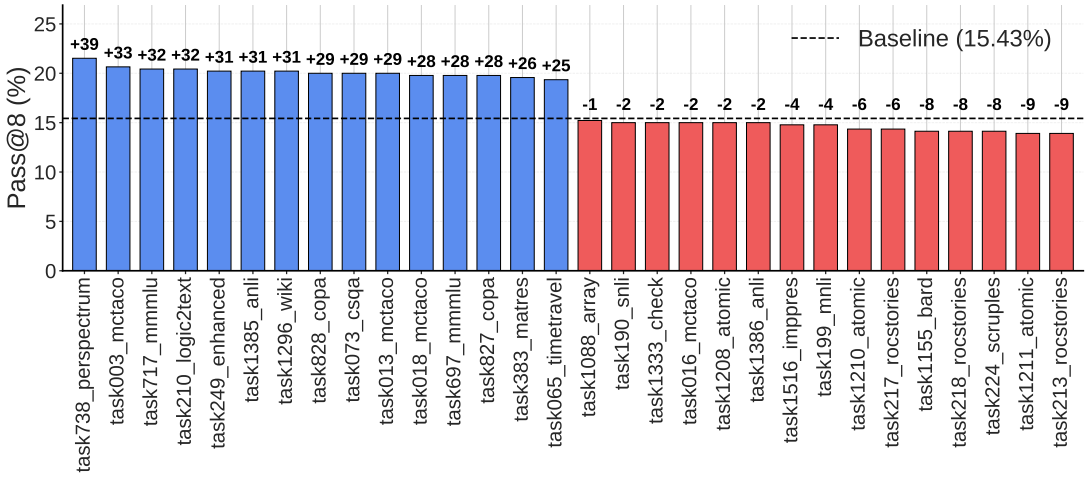


Figure 4: **Impact of Task Selection.** We train the baseline (Qwen3-0.6B) on each task individually under compute-matched settings. We report relative pass@8 gains on BBEH-mini for each task and highlight the tasks that improve and degrade the baseline.

spanning various tasks types such as question answering, question generation and commonsense reasoning. Each task consists of the task description and the instruction-response pair, annotated by experts. We sample a candidate pool of 83 tasks for our experiments.[3]

Training. We train models from Qwen3 (Yang et al., 2025) family (0.6B, 1.7B and 4B) with GRPO (Shao et al., 2024) for all our experiments (§ 2). We adopt binary rewards for training: 1 if the output is judged correct under rule-based verification and 0 otherwise. The rule-based verifier extracts the final answer, normalizes the output and applies string matching. For our data curation experiments, we use Qwen3-0.6B for faster training iterations. All our data curation experiments were run for 250 RL steps. Finally, we train the SuperNova models for 5000 RL steps. We present more details about the training setup in Appendix §C.

Evaluation. We evaluate our models on various benchmarks that target diverse reasoning capabilities. For our data curation experiments, we choose BBEH-mini as our validation benchmark. BBEH-mini is a small subset of BBEH (Kazemi et al., 2025) that comprises of 23 challenging tasks that target diverse reasoning capabilities. We use the remaining BBEH samples, which are not included in BBEH-mini as the unseen test set, BBEH-test. Thus, downstream reasoning performance is aggregated over all 23 tasks. After curating SuperNova, we evaluate our models on 4 additional *unseen* benchmarks including BBH (Suzgun et al., 2023), Zebralogic (Lin et al., 2025), MMLU-Pro (Wang et al., 2024) and MATH500 (Lightman et al., 2023). To ensure consistency, we use an identical prompt across all evaluations that encourages the model to think before answering, provided in Appendix §C.4.

Evaluation Metric. We adopt pass@k as our evaluation metric, which is well-suited for evaluating RL-trained models (Chen et al., 2021; Yue et al., 2025). As shown in Appendix § B, we find that pass@8 provides 2.5 times greater discriminability than pass@1 (σ: 0.76 $\rightarrow$ 1.92). We therefore utilize pass@8 for our data curation experiments.

Baselines. To further compare the quality of SuperNova with other reasoning datasets, we train Qwen3-0.6B under a compute-matched setup using three baseline datasets including Nemotron-CrossThink (Akter et al., 2026), which targets diverse domains and question-answer formats; General-Reasoner (Ma et al., 2025), which curates reasoning data across diverse STEM-focused domains; and DAPO (Yu et al., 2025), a high-quality math reasoning dataset sourced from competition websites. We provide additional details in Appendix §D. We evaluate several models as baselines for our experiments. **(1) Qwen3**: included to measure the gains obtained from training on SuperNova. **(2) OpenThinker3-7B** (Guha et al., 2025): a strong math reasoning model supervised fine-tuned on large math corpus. **(3) OpenReasoner-**

[3]The candidate pool allows us to implement SuperNova on tractable compute. Ideally, SuperNova can be applied to all tasks from SuperNI. Additional details in Appendix §C.

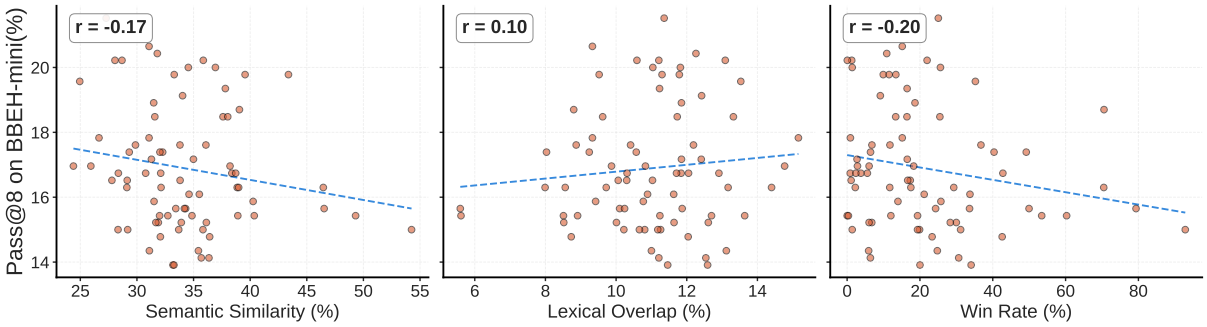


Figure 5: **(Left)** Correlation between semantic similarity and task performance. **(Middle)** Correlation between lexical similarity and task performance. **(Right)** Correlation between win rate and downstream reasoning performance.

Nemotron-7B (Ahmad et al., 2025): a strong reasoning model. **(4) Olmo3-7B-Think** (Olmo et al., 2025): a state-of-art reasoning model that has strong performance across various domains.

5 Experiments

Impact of Task Selection. We train `Qwen3-0.6B` on each task and report model performance on BBEH-mini in Figure 4. Our experiments show that task selection has a substantial impact on downstream reasoning performance (Figure 4). Specifically, we observe a 7.6 percentage point (pp) gap between the lowest-performing task (`task213-rocstories`, 13.9%) and the highest-performing task (`task738-perspectrum`, 21.5%). Notably, several tasks degrade performance relative to the baseline, underscoring that not all tasks are beneficial for improving reasoning under RLVR, and task selection based on task utility is critical for training strong reasoning models. Furthermore, we find that tasks involving multi-hop reasoning yield the largest gains over the baseline model (Appendix § F).

Task Utility Ranking. We find that semantic similarity and lexical similarity between the tasks and validation benchmark are poor predictors of task utility for RLVR. As shown in Fig. 5, both measures exhibit weak correlation with model performance on BBEH-mini. These approaches are attractive because they are cheap, fast to implement, and model-agnostic. However, our findings suggest that surface similarity is insufficient for task selection for RLVR. We also investigate whether task difficulty, measured by the average win-rate of the base model, predicts downstream reasoning performance in Fig. 5. Similar to surface similarity, we observe only a weak correlation between task difficulty and model performance on BBEH-mini, indicating that base-model difficulty is also a poor predictor of task utility for RLVR. Overall, our findings underscore that effective task selection is grounded in compute-matched RL training, and simpler proxies are insufficient for predicting downstream performance.

	Top 1	Top 2	Top 4	Top 8	Top 16
Micro Mixing					
pass@1	7.5	**8.9**	7.5	7.6	7.5
pass@8	18.3	**22.8**	18.7	18.0	20.2
Macro Mixing					
pass@1	7.6	8.2	6.6	6.4	7.5
pass@8	21.5	21.7	17.4	17.0	18.3

Table 1: **Impact of mixing.** We mix questions from tasks following two strategies: micro mixing and macro mixing. We find that micro mixing with top 2 tasks achieves the best performance **(bold)**.

Impact of Task Mixing. We present the results of two mixing strategies: Macro Mixing and Micro Mixing in Table 1. Across both strategies, we find that mixing top 4, 8 or 16 tasks did not yield better results than top 1 or 2 tasks at both pass@1 and pass@8. Thus, mixing strategies are highly dependent on the training data distribution and mixing more data sources does not always yield high reasoning performance. Furthermore, we find that micro-mixing with top-2 achieves the highest pass@8 scores of 22.8% our experiments. These results indicate that selecting top-ranked tasks per sub-task (Micro Mixing) yields better performance than selecting tasks based on overall ranking at compute-matched settings.

Impact of Data Interventions. We apply several data intervention strategies to the best-performing dataset from §5 (Micro-Top2) and report results

Model	MMLU-Pro	BBH	Zebralogic	MATH500	Average
Qwen3-0.6B	55.3	52.4	34.4	**71.9**	53.5
RL w/ General-Reasoner (Ma et al., 2025)	54.4	64.3	45.4	66.3	57.6
RL w/ Nemotron-Crossthink (Akter et al., 2026)	55.7	69.9	49.1	70.0	61.2
RL w/ SuperNova	**56.2**	**81.5**	**49.4**	71.4	**64.6**

Table 2: **SuperNova beats reasoning datasets on reasoning benchmarks**. We report pass@8 across four benchmarks that were unseen during data curation. We find that LLMs trained on SuperNova show improved pass@8 over the models trained on baseline datasets under compute-matched settings.

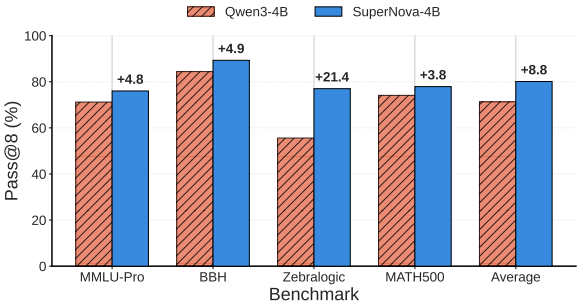


Figure 6: **SuperNova generalizes to OOD Benchmarks.** We report pass@8 across four reasoning benchmarks for Qwen3-4B and SuperNova-4B.

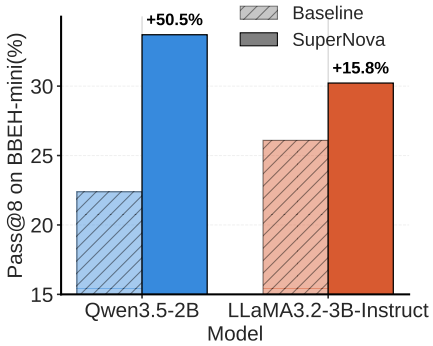


Figure 7: We train Qwen3.5-2B and LLaMA3.2-3B-Instruct with SuperNova and show relative gains over their respective baseline models on 23 complex reasoning tasks.

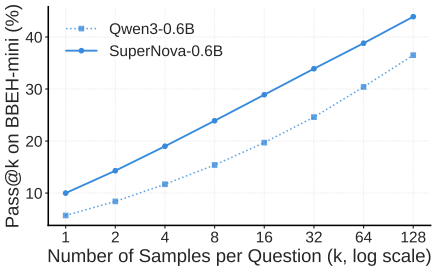


Figure 8: We show the performance comparison between the baseline model and SuperNova-0.6B by scaling values of k up to 128 on 23 complex reasoning tasks.

on BBEH-mini in Appendix Table 6. Surprisingly, none of the interventions improve over the original data. While Going Against Prior achieves the highest performance among the interventions (22.6%), is comparable to but does not exceed Micro Top2. Thus, such synthetically generated interventions on the question are not effective in improving downstream reasoning performance with RLVR.

6 Training Reasoners with SuperNova

SuperNova demonstrates strong reasoning performance. We find that training Qwen3-4B on SuperNova outperforms the base Qwen3-4B model with a relative gain of 43.8pp (Figure 1(a)). Moreover, SuperNova also demonstrates consistent gains over the base model across 0.6B and 1.7B models. We show a reasoning trace produced by SuperNova-4B on a challenging Boardgame-QA example in Appendix Figure 11. The model performs multi-step reasoning by chaining together multiple rules and resolving preference conflicts. This example demonstrates that training on SuperNova enables the model to perform structured reasoning over challenging tasks.

SuperNova beats SOTA reasoning datasets. We compare SuperNova against three state-of-the-art reasoning datasets that target diverse reasoning skills. To ensure a fair comparison of data quality, we perform a compute-matched analysis of all datasets (details in Appendix § D). Results on BBEH-test are shown in Figure 1(b). We find that SuperNova achieves relative gains of 42pp on pass@1 and 28pp on pass@8 over the strongest baseline, Nemotron-Crossthink. In contrast, both math reasoning datasets, DAPO and Nemotron-Crossthink (Math) show little to no improvement over the baseline. Additionally, we report pass@8 performance of all datasets on OOD benchmarks in Table 2. Models trained on SuperNova outperform those trained on all baseline datasets, achieving 81.5% on BBH and a 3.4pp improvement on average over Nemotron-Crossthink.

Model	**Avg.**	Brd.	Mov.	Dis.	Bool.	Geo.	T.Ar.	WoL	Word	Shuf.	Zebra	NYCC	M.Ar.	Hyp.
OLMo-3-7B-Think	15.1	6.1	**78.4**	48.5	0	0	16.2	2.4	24.4	0	0	20.8	0	0
OpenReasoning-7B	9.7	4.1	19.6	60.6	0	0	0	2.4	9.8	0	0	29.2	0	0
OpenThinker3-7B	8.8	12.2	9.8	54.5	0	0	8.1	0	17.1	0	0	12.5	0	0
Qwen3-8B	26.5	59.2	51	57.6	5.4	2.9	**67.6**	31	36.6	2.3	10	20.8	0	0
Qwen3-4B	25.8	65.3	64.7	51.5	5.4	0	48.6	26.2	36.6	0	10	25	2.2	0
SuperNova-4B	**46.8**	**71.4**	65.7	**65.2**	**56.8**	**54.4**	51.4	**47.6**	**42.7**	**41.9**	**33**	**30.2**	**25.6**	**22.2**

Table 3: **SUPERNOVA models exhibit strong performance across challenging reasoning tasks.** We report the pass@8results (%) on 13 tasks from BBEH-test. Best per column is **bolded** and second-best in underlined. Avg. is computed across the 13 tasks shown. (Brd.=Boardgame QA, Mov.=Movie Recommendation, Dis.=Disambiguation QA, Bool.=Boolean Expressions, Geo.=Geometric Shapes, T.Ar.=Time Arithmetic, WoL=Web of Lies, Word=Word Sorting, Shuf.=Shuffled Objects, Zebra=Zebra Puzzles, M.Ar.=Multistep Arithmetic, Hyp.=Hyperbaton.)

SUPERNOVA generalizes to Out-of-Distribution (OOD) Benchmarks. We evaluate SUPERNOVA-4B on challenging reasoning benchmarks that are unseen during data curation, as shown in Figure 6. Notably, SUPERNOVA achieves substantial gains on Zebralogic, where SUPERNOVA-4B outperforms Qwen3-4B by 21pp. SUPERNOVA-4B also shows gains of 4.8pp on MMLU-pro and 4.9pp on BBH over the base model. Finally, SUPERNOVA-4B shows gains of 3.8pp on MATH500 despite not being trained on math, indicating that training on SUPERNOVA transfers performance improvements to math reasoning benchmarks.

SUPERNOVA demonstrates reasoning gains over diverse tasks. To further understand the reasoning performance gains from SUPERNOVA, we report the pass@8 on 13 challenging tasks from BBEH-test on which SUPERNOVA-4B demonstrates substantial gains in Table 3. We find that SUPERNOVA-4B consistently outperforms Qwen3-4B across all tasks with an absolute gain of 19pp. SUPERNOVA-4B achieves substantial gains on tasks such as Hyperbaton, Geometric objects and Shuffled objects where Qwen3-4B achieves zero performance. This highlights that SUPERNOVA elicits reasoning behaviors beyond the base model's capabilities on challenging reasoning tasks. Finally, we observe that SUPERNOVA-4B outperforms Qwen3-8B across 12 tasks by 20.3pp despite having half the model capacity. This further highlights that SUPERNOVA is a high-quality dataset that improves parameter efficiency, enabling smaller models to outperform substantially larger ones.

SUPERNOVA gains are consistent at larger values of k. We analyze whether the performance gains from SUPERNOVA persist at higher values of k. As shown in Figure 8, SUPERNOVA-0.6B maintains consistent gains over Qwen3-0.6B across all values of k up to 128. This suggests that training on SUPERNOVA expands the model's exploration space even at large sample sizes, enabling more diverse reasoning behaviors than the baseline.

SUPERNOVA shows cross-model generalization. We study how training on SUPERNOVA generalizes across model families (Figure 7). In particular, LLaMA3.2-3B-Instruct (Grattafiori et al., 2024) trained on SUPERNOVA achieves gains of 15.8pp over its baseline. We further observe similar improvements on Qwen3.5-2B (Team, 2026), suggesting that data curation insights derived from earlier-generation models (e.g., Qwen3) transfer to newer-generation models. Overall, the benefits of SUPERNOVA generalize across both model families and generations.

7 Conclusion

In this work, we introduce SUPERNOVA, a framework for curating RLVR data from large-scale instruction-tuning datasets to improve the reasoning capabilities of LLMs beyond formal STEM domains. Through controlled RL experiments, we show that effective RLVR data curation depends critically on source task selection and fine-grained task mixing, while synthetic interventions designed to increase reasoning complexity do not reliably improve downstream performance. We further demonstrate that models trained on SUPERNOVA generalize across challenging reasoning benchmarks and transfer across model families and newer model generations. Our findings suggest that existing human-annotated instruction datasets contain rich underutilized signals for RLVR, but unlocking their potential requires principled empirical curation such as SUPERNOVA framework. Overall, we hope SUPERNOVA provides both a practical resource and a foundation for future work on data curation for RLVR in diverse domains.

Limitations

While SUPERNOVA provides key insights and exhibits substantial improvements on academic reasoning benchmarks, several important directions remain for future work. We apply SUPERNOVA to a sample of tasks from SuperNI to demonstrate our key findings and hypothesize that our findings will remain consistent scaling training tasks. Additionally, our work focuses on compute-constrained settings, and future work may explore how these data curation principles scale with substantially larger RL budgets. It is possible that continued RL training might yield even pronounced results (Liu et al., 2025). We use RL-training and empirical validation as measures of task utility in our curation pipeline which are expensive at scale, future work may explore how to apply gradient-based strategies for task selection (Xia et al., 2024) or reward oriented strategies (Wu et al., 2024) in RLVR. Our evaluation benchmarks give a comprehensive assessment of the impact of principled data curation on reasoning capabilties, however, they do not fully capture real-world setting and decision making capabilities. Future work may apply our insights to other complex reasoning tasks such as multi-agent systems, app development, healthcare, finance.

Ethical Concerns

SUPERNOVA is sourced from SuperNI tasks which were annotated by humans. Thus, SUPERNOVA inherits the biases from SuperNI and RL training on such data may amplify biases present in the base model or underlying source dataset. In this work, we primarily focus on enhancing the reasoning capabilities of LLMs and do not explicitly address fairness, bias mitigation or value alignment. Future work should systematically evaluate how RL on reaasoning data from natural instruction datasets affects model behavior across sensitive axes such as gender and race.

Disclosure of LLM use in both research and reviewing. We use ChatGPT and Claude as in our experiments and have provided the relevant prompts. Claude was used in formatting latex tables and code generation for the figures. Finally, we used ChatGPT and Claude to assist with grammar and proof-reading in our paper writing.

References

Wasi Uddin Ahmad, Sean Narenthiran, Somshubra Majumdar, Aleksander Ficek, Siddhartha Jain, Jocelyn Huang, Vahid Noroozi, and Boris Ginsburg. 2025. OpenCodeReasoning: Advancing Data Distillation for Competitive Coding. *arXiv preprint arXiv:2504.01943*.

Syeda Nahida Akter, Shrimai Prabhumoye, Matvei Novikov, Seungju Han, Ying Lin, Evelina Bakhturina, Eric Nyberg, Yejin Choi, Mostofa Patwary, Mohammad Shoeybi, and 1 others. 2026. Nemotron-crossthink: Scaling self-learning beyond math reasoning. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 984–1002.

Hritik Bansal, Devandra Singh Sachan, Kai-Wei Chang, Aditya Grover, Gargi Ghosh, Wen-tau Yih, and Ramakanth Pasunuru. 2025. Honeybee: Data recipes for vision-language reasoners. *arXiv preprint arXiv:2510.12225*.

Adithya Bhaskar, Xi Ye, and Danqi Chen. 2025. Language models that think, chat better. *Preprint*, arXiv:2509.20357.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, and 1 others. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Yang Chen, Zhuolin Yang, Zihan Liu, Chankyu Lee, Peng Xu, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2025. Acereason-nemotron: Advancing math and code reasoning through reinforcement learning. *arXiv preprint arXiv:2505.16400*.

Jorge Zhoujun Cheng, Shibo Hao, Tianyang Liu, Fan Zhou, Yutao Xie, Feng Yao, Yuexin Bian, Nilabjo Dey, Yonghao Zhuang, Yuheng Zha, and 1 others. 2026. Revisiting reinforcement learning for llm reasoning from a cross-domain perspective. *Advances in Neural Information Processing Systems*, 38.

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

Thomas L Griffiths. 2020. Understanding human intelligence through human limitations. *Trends in Cognitive Sciences*, 24(11):873–883.

Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, and 1 others. 2025. Openthoughts: Data recipes for reasoning models. *arXiv preprint arXiv:2506.04178*.

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

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. *Preprint*, arXiv:2103.03874.

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. 2025. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. *arXiv preprint arXiv:2503.24290*.

Maggie Huan, Yuetai Li, Tuney Zheng, Xiaoyu Xu, Seungone Kim, Minxin Du, Radha Poovendran, Graham Neubig, and Xiang Yue. 2025. Does math reasoning improve general llm capabilities? understanding transferability of llm reasoning. *Preprint*, arXiv:2507.00432.

Hugging Face. 2025. Open r1: A fully open reproduction of deepseek-r1.

Philip N Johnson-Laird. 2010. Mental models and human reasoning. *Proceedings of the National Academy of Sciences*, 107(43):18243–18250.

Mehran Kazemi, Bahare Fatemi, Hritik Bansal, John Palowitch, Chrysovalantis Anastasiou, Sanket Vaibhav Mehta, Lalit K Jain, Virginia Aglietti, Disha Jindal, Yuanzhu Peter Chen, and 1 others. 2025. Big-bench extra hard. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26473–26501.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, and 1 others. 2024. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. In *The twelfth international conference on learning representations*.

Bill Yuchen Lin, Ronan Le Bras, Kyle Richardson, Ashish Sabharwal, Radha Poovendran, Peter Clark, and Yejin Choi. 2025. Zebralogic: On the scaling limits of llms for logical reasoning. *arXiv preprint arXiv:2502.01100*.

Junteng Liu, Yuanxiang Fan, Zhuo Jiang, Han Ding, Yongyi Hu, Chi Zhang, Yiqi Shi, Shitong Weng, Aili Chen, Shiqi Chen, Yunan Huang, Mozhi Zhang, Pengyu Zhao, Junjie Yan, and Junxian He. 2025. Synlogic: Synthesizing verifiable reasoning data at scale for learning logical reasoning and beyond. *Preprint*, arXiv:2505.19641.

Yang Liu, Jiaqi Li, and Zilong Zheng. 2026. Rulereasoner: Reinforced rule-based reasoning via domain-aware dynamic sampling. *Preprint*, arXiv:2506.08672.

Ximing Lu, David Acuna, Jaehun Jung, Jian Hu, Di Zhang, Shizhe Diao, Yunheng Zou, Shaokun Zhang, Brandon Cui, Mingjie Liu, Hyunwoo Kim, Prithviraj Ammanabrolu, Jan Kautz, Yi Dong, and Yejin Choi. 2026. Golden goose: A simple trick to synthesize unlimited rlvr tasks from unverifiable internet text. *Preprint*, arXiv:2601.22975.

Xueguang Ma, Qian Liu, Dongfu Jiang, Ge Zhang, Zejun Ma, and Wenhu Chen. 2025. General-reasoner: Advancing llm reasoning across all domains. *arXiv preprint arXiv:2505.14652*.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori B Hashimoto. 2025. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20286–20332.

Allen Newell, Herbert Alexander Simon, and 1 others. 1972. *Human problem solving*, volume 104. Prentice-hall Englewood Cliffs, NJ.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, and 1 others. 2025. Olmo 3. *arXiv preprint arXiv:2512.13961*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and 1 others. 2023. Challenging big-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051.

Qwen Team. 2026. Qwen3. 5: Towards native multimodal agents. *URL: https://qwen. ai/blog*.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, and 1 others. 2022. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 5085–5109.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and 1 others. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Yang Wu, Huayi Zhang, Yizheng Jiao, Lin Ma, Xiaozhong Liu, Jinhong Yu, Dongyu Zhang, Dezhi Yu, and Wei Xu. 2024. Rose: A reward-oriented data selection framework for llm task-specific instruction tuning. *arXiv preprint arXiv:2412.00631*.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. Less: Selecting influential data for targeted instruction tuning. *arXiv preprint arXiv:2402.04333*.

Siheng Xiong, Ali Payani, Ramana Kompella, and Faramarz Fekri. 2024. Large language models can learn temporal reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10452–10470.

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

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. 2025. Limo: Less is more for reasoning. *arXiv preprint arXiv:2502.03387*.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, and 1 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. 2025. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*.

Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. 2025. Simplerl-zoo: Investigating and taming zero reinforcement learning for open base models in the wild. *arXiv preprint arXiv:2503.18892*.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2025. Instruction tuning for large language models: A survey. *Preprint*, arXiv:2308.10792.

Yifan Zhang and Team Math-AI. 2024. American invitational mathematics examination (aime) 2024. https://huggingface.co/datasets/math-ai/aime24.

Han Zhao, Haotian Wang, Yiping Peng, Sitong Zhao, Xiaoyu Tian, Shuaiting Chen, Yunjie Ji, and Xiangang Li. 2025. 1.4 million open-source distilled reasoning dataset to empower large language model training. *arXiv preprint arXiv:2503.19633*.

A Related Work

A.1 General-Purpose Reasoning in LLMs

Several works have explored expanding the general reasoning capabilities of LLMs. (Ma et al., 2025) constructs a large-scale dataset spanning multiple domains such as history, finance, and physics from web-scraped sources. (Akter et al., 2026) goes beyond mathematics by curating synthetically derived questions from CommonCrawl and open-source QA datasets. (Lu et al., 2026) leverages transformed pretraining data with structured templates and distractors to generate verifiable reasoning data in domains such as cybersecurity. However, these approaches largely rely on internet-sourced data, which can be noisy and of low quality. Other work has focused on rule-based tasks (Liu et al., 2026) and logic puzzles (Liu et al., 2025). (Cheng et al., 2026) employs data mixing across diverse domains such as math, logic and tabular and studies data curation for RLVR, however, they rely on domain-specific rewards and complex data filtering to combine data from existing domain specific datasets. While effective for specialized reasoning, these approaches rely on highly specialized domain-specific datasets that are challenging to scale and require complex reward design and filtering. In contrast, SuperNova leverages instruction-tuning datasets, which are human-annotated and can be made usable for RLVR with simple reformatting and scale easily across diverse reasoning types.

A.2 Data Curation for Reasoning

High-quality reasoning data is critical for training strong LLM reasoners. Prior work has focused on large-scale datasets for supervised fine-tuning (SFT) (Hugging Face, 2025; Zhao et al., 2025) and RLVR (Chen et al., 2025; Hu et al., 2025), typically by scraping competition websites or distilling knowledge from larger models. On the other hand, (Muennighoff et al., 2025; Ye et al., 2025) demonstrate that carefully curated, high-quality reasoning datasets can yield strong gains even with relatively small datasets. (Guha et al., 2025) systematically studies data design principles for SFT reasoning data at scale through controlled experiments, in a manner similar to SuperNova. However, these efforts primarily focus on reasoning in formal domains using SFT. SFT aims to improve instruction-following by training on demonstrations (Zhang et al., 2025; Wang et al., 2022), while RLVR optimizes a sparse, outcome-based reward (Guo et al., 2025). Moreover, SFT typically requires complete reasoning traces and solutions for training, whereas RLVR only requires the final answer. As a result, SFT-oriented data curation strategies do not directly transfer to RLVR. SuperNova addresses this gap by providing key insights to drive data curation for RLVR.

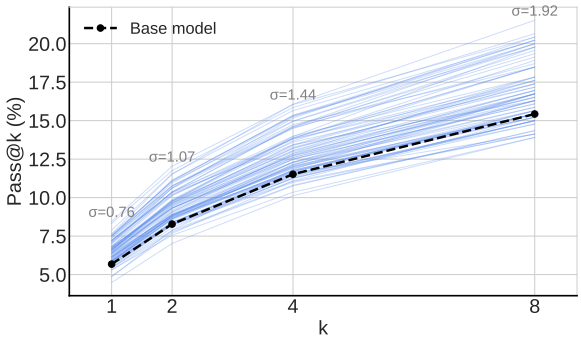


Figure 9: Pass@k accuracy of task-specific models across various values of k.

B Pass@k Analysis

Following (Chen et al., 2021) and (Yue et al., 2025), we analyze the pass@k curves of our task-specific models. Across 80+ RL curves, we observe that the spread and distinguishability of model performance increases at k=8, with maximum overlap at k=1. We show the pass@k curves in Figure 9.

C Detailed Experimental Setup

C.1 Task Selection

To ensure we can conduct a controlled study with our limited compute and keep our search space tractable, we prompt an LLM (Claude-opus-4.6) to first classify each task in SuperNI as suitable for reasoning or not, then we randomly sample 83 tasks from all the reasoning suitable tasks to prepare our candidate pool. We would like to note that such capable LLMs are useful for initial task sampling but they do not allow on-policy training and thus exhibit poor correlation with downstream reasoning performance. Ideally, SuperNova can be applied to all tasks from SuperNI and similar instruction-tuning datasets to create large scale RLVR data. We use a minimal binary classification prompt and do not consider the validation benchmark while preparing this candidate pool. This was done to ensure that the task ranking is done purely on task utility scores from §3.1. The prompt follows:

Task A: `task738_perspectrum_classification` — claim/perspective stance detection (classification, support/undermine).	
Shared raw task definition: "In this task you will be given a claim and a perspective. You should determine whether that perspective supports or undermines the claim. If the perspective could possibly convince someone with different view, it is supporting, otherwise it is undermining."	
Raw	claim: Children should not be allowed to inherit vast wealth as this damages them and society. perspective: Inherited wealth demotivates the recipients so that they put less effort into training, education and social skills. **Output:** `support`
Reformat *(MCQ, 2-way)*	Claim: Children should not be allowed to inherit vast wealth as this damages them and society. Perspective: Inherited wealth demotivates the recipients so that they put less effort into training, education and social skills. Question: Does the perspective support or undermine the claim? Choose one option: (A) Supports (B) Undermines **Answer:** `A`
Raw	claim: Domestic intelligence agencies have a legitimate role to play in democracy. perspective: The government does not have the right to spy on its citizens. **Output:** `undermine`
Reformat *(open-ended)*	Claim: Domestic intelligence agencies have a legitimate role to play in democracy. Perspective: The government does not have the right to spy on its citizens. Rule: If the perspective could possibly convince someone with a different view, it is supporting; otherwise it is undermining. Question: Does the perspective support or undermine the claim? **Answer with a single word: support or undermine.** **Answer:** `undermine`
Raw	claim: Marriage is an outdated institution. perspective: Those who are observant religiously think that marriage is important. **Output:** `undermine`
Reformat *(MCQ, 10-way)*	Determine whether the perspective supports or undermines the claim. Claim: Marriage is an outdated institution. Perspective: Those who are observant religiously think that marriage is important. Choose the best option: (A) Supports (B) Undermines (C) Both supports and undermines (D) Neither (E) Neutral or irrelevant (F) Depends on context (G) Ambiguous (H) Contradicts (I) Unrelated (J) Cannot determine **Answer:** `B`

Table 4: **Examples from SuperNova dataset.** We show examples from our top performing task task738. In the reformatting step of our pipeline, we employs LLMs to (i) inline the shared task description in each sample, (ii) produce both open-ended and multiple-choice variants, and (iii) vary MCQ option counts from 2-way to 10-way. This reformatting across diverse tasks, normalizes the answer format and allows for rule-base verification.

```
This is an instruction-following task used to
train LLMs. Consider the given task description
and examples. Now assess the suitability of the
task for RL training reasoning models. Think
step by step and only respond with yes/no.
Task ID: {task_id}
Task Description: {description}
Example Input: {input}
Example Output: {output}
```

For reformatting the instruction tuning tasks to verifiable questions, we prompt GPT-5-mini with G. To estimate the quality of the reformatting, we manually inspect 100 samples from 8 tasks and find that GPT-5-mini follows the prompt accurately on 98.4% of the samples while preserving the ground-truth and original task structure.

C.2 Training

All our experiments were done on 4xH100 gpus. We use the GRPO implementation from TRL[4] for our training. All our data curation experiments utilize Qwen3-0.6B with 500 prompts, learning rate of 1e-6, 8 generations per prompt, batch size of 8, decoding temperature of 0.7 and maximum generation length 4096. We run our training for 250 steps (1 epoch). For our large scale experiments, we run 5000 steps (1 epoch) across 10,000 prompts and use a learning rate of 1e-6 for 0.6B models and 4e-6 for 1.7B and 4B models.

[4]https://github.com/huggingface/trl

C.3 Reward Design

We adopt binary rewards for all our experiments: 1 if output is judged correct by the reule-based verification and 0 otherwise. The verifier extracts the final answer from the output and matches it against the reference under a normalization-and-fuzzy-match rule: stripping answer-prefix sentinels and LaTeX wrappers (e.g., $\boxed{\{}$}, "The answer is:"), lowercasing, and admitting numeric equality, multiple-choice (A)-A equivalence, and list-bracket equivalence.

C.4 Evaluation

We use the following benchmarks for our evaluations:

- BBEH (Kazemi et al., 2025): Comprises of 23 challenging tasks that require diverse reasoning skills such as sarcasm detection, humour detection, constraint satisfaction, boolean algebra, obect ordering, temporal reasoning, commonsense reasoning and logical deduction. Strong reasoning models such as Gemini and O3 achieve 50% on this benchmark. BBEH was designed to resist saturation and evaluate models on diverse reasoning tasks in a robust manner.
- BBH (Suzgun et al., 2023): Comprises of 23 tasks, similar to BBEH but easier in difficulty. Most modern LLMs have saturated and chain of thought prompting usually achieves great improvements on this benchmark.
- Zebralogic (Lin et al., 2025): Comprises of hard and challenging constraint satisfaction logical grid puzzles.
- MMLU-pro (Wang et al., 2024): tests knowledge intensive reasoning across diverse domains.
- MATH500 (Lightman et al., 2023): A samller subset of the MATH benchmarch that is used to evaluate mathematical reasoning performance of LLMs.

For our evaluations, we use the following prompt across all benchmarks:

```
Think step by step, and when you are ready to
provide the final answer, use the prefix "The
answer is:" followed by the answer directly,
with no formatting and no markup. For instance:
"The answer is: 42", or "The answer is: yes",
or "The answer is: (a)" For multi-choice
questions, provide the letter, e.g. "The answer
is: (a)
```

All evaluations were conducted on 1xH100 with a batch size of 8. We use decoding temperature=0.7 with maximum generation length of 4096 across all our experiments.

D Implementation Details of Training Baseline Datasets

For fair comparison across dataset quality, we train Qwen3-0.6B on the fixed budget of 250 RL steps across 500 prompts and the same learning rate for all datasets. Since Nemotron-Crossthink, Dapo and General-Reasoner are large-datasets, we report their performance as average pass@8 across three runs trained on three random samples of 500 prompts.

E Data Interventions

Following (Kazemi et al., 2025), we design the given 7 interventions to improve data quality (Table 5) and prompt GPT-5-mini with given prompt. Since, we want to preserve the ground-truth answer, we apply these interventions only to the problem statement. Finally, to ensure the that the final answer is preserved, filter the augmented data with based on win-rate computed again with the augmented problem statements. In our experiments, we combine the original data and the intervened data in a ratio of 1:1.

F Task Performance Analysis.

We prompt an LLM (Claude-Opus-4.6) with the task descriptions from each task and generate coarse category labels. We find that Multi-hop Reasoning and Coreference resolution emerge as the strongest categories, while narrative and surface-formatting tasks (e.g., Story Coherence, Date/Temporal format) consistently underperform (Figure 10). However, these aggregate trends obscure variations at the task-level. Despite Textual Entailment & NLI ranking in the middle at the category-level, `task738_perspectrum` emerges as the top-ranked task by large margin. This highlights that coarse category labels are insufficient for task selection and effective data curation for RL should be driven by fine-grained task utility analysis.

G Micro Mixing

We provide the top tasks ranked per sub-task in Table 7. For Micro-Top1, 16 unique tasks are selected while 31 unique tasks are included in Micro-Top2. Additionally, we show the distribution of reasoning skills as categorized in § F in SUPERNOVA which is scaled from Micro-Top2 and comprises 31 unique tasks.

Dimension	Description
Many-hop reasoning	Add information that increases the number of reasoning steps needed to reach the answer.
Going against strong prior	Add context that creates a misleading prior belief which conflicts with the correct answer, tempting the model to answer incorrectly based on surface-level associations.
Learning on the fly	Introduce a new rule, definition, or convention within the problem that must be understood and applied to solve it.
Long-context	Pad the problem with additional (but non-answer-changing) context to increase overall length.
Finding errors in reasoning traces	Include a flawed reasoning chain within the problem that the model must recognize as incorrect.
Inductive reasoning	Provide a set of examples that establish a pattern, requiring the model to induce and apply the pattern.
Constraint satisfaction	Add extra constraints that the model must track, even though they do not affect the final answer.
Compositional understanding	Fuse an independent sub-problem into the main problem, requiring the model to separate and solve them independently.
Knowledge-intensive reasoning	Add domain-specific terminology or context that requires specialized knowledge to parse, even though it does not change the answer.

Table 5: Following (Kazemi et al., 2025), we design these interventions to improve the data quality. We provide the interventions and their definitions here.

Intervention	pass@8
Micro-Top2	**22.8**
Going Against Prior	22.6
Long-Context	21.3
Inductive Reasoning	20.4
Finding Errors	20.0
Many-hop Reasoning	20.0
Knowledge-intensive Reasoning	19.8
Compositional Understanding	19.6
Learning on the Fly	18.3

Table 6: **Impact of interventions.** We compare the performance of models trained on datasets transformed using synthetic interventions. We find that the base dataset is superior to all interventions.

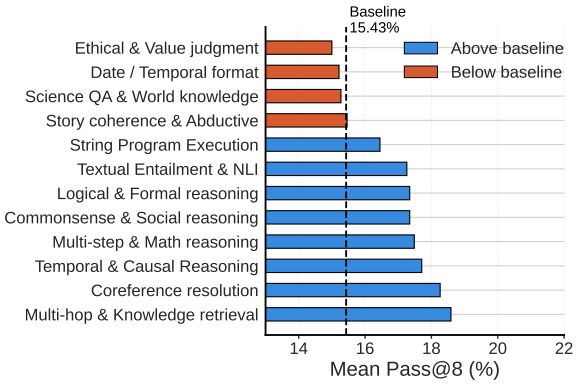


Figure 10: (a)We categorize the source tasks based on target reasoning skill and task type. We report Mean Pass@8 across each task category and highlight the categories which degrade baseline (Qwen3-0.6b) model on BBEH-mini.

Task. Given facts about animals and 30 inference rules (with explicit preferences when rules conflict), decide whether *"the bee invests in the company owned by the butterfly"* is `proved`, `disproved`, or `unknown`.

Model output (verbatim).

> *The fish has a cappuccino (something to drink), so by Rule 12 (preferred over Rule 30), it invests in the dragonfly's company. This triggers Rule 18, making the monkey unite with the coyote. The snake acquires a photo of the akita, so Rule 13 applies, leading to shouting at the lizard (Rule 16), which means it doesn't reveal secrets to the vampire. This activates Rule 20 (preferred over Rule 25), so the snake invests in the leopard's company. Rule 7 then makes the snake take over the crab, triggering Rule 5 (crab pays to llama). Rule 15 ensures the walrus doesn't call the bee. For the bee's investment in the butterfly, Rule 19 requires: (1) walrus not calling bee (satisfied) and (2) liger not paying to bee. Rule 23 connects this to the goat borrowing a weapon from the liger. Rule 17 is triggered if any animal stops the starling's victory. Rule 14 (preferred over Rule 3) applies to the snake, so it stops the starling, leading to the goat borrowing from the liger. Thus, liger doesn't pay to bee, satisfying Rule 19. Rule 27 (preferred over Rule 19) is irrelevant without bison info. Therefore, the bee invests in the butterfly.*

Step-by-step trace. Each row is one inferential move; preference overrides are noted in the second column.

#	Rule (over)	Inferential move	Conclusion
1	R12 (R30)	fish has a drink ⇒ invests in dragonfly	fish → dragonfly
2	R18	some animal invests in dragonfly	monkey unites with coyote
3	R13	snake has akita photo ⇒ shouts at lizard	snake shouts at lizard
4	R16	shouts at lizard ⇒ no secret to vampire	snake vampire
5	R20 (R25)	no vampire secret ⇒ invests in leopard	snake → leopard
6	R7	invests in leopard ⇒ takes over crab	snake takes crab
7	R5	takes over crab ⇒ crab pays llama	crab → llama
8	R15	someone pays llama ⇒ walrus $\not\to$ bee	**Cond. 1 of R19**
9	R14 (R3)	snake $\notin$ beetle, $\notin$ woodpecker ⇒ stops starling	snake stops starling
10	R17	stops starling ⇒ goat borrows from liger	goat borrows from liger
11	R23	goat borrows from liger ⇒ liger $\not\to$ bee	**Cond. 2 of R19**
12	R19	Cond. 1 *and* Cond. 2 hold	**bee → butterfly**
13	R27 (R19)	no animal invests in bison ⇒ override does not fire	R19's conclusion stands

Figure 11: Verbatim output from SuperNova-4B on a Boardgame-QA instance with 30 rules and preference orderings, together with the step-by-step inference trace it implies. The model produces the correct answer (proved) by chaining 13 rule applications and correctly resolves four rule-preference conflicts.

Prompt for Data Intervention

```
You are an expert data augmentation assistant. Your task is to take an existing (problem, answer) pair from an NLP dataset and inject a distractor into the problem.
The distractor must make the problem harder for an AI model to solve, but it must NOT change the ground-truth answer.

## DISTRACTOR TYPE
You MUST use the following distractor type:
**{distractor_name}**: {distractor_description}

## RULES
1. **Answer preservation (CRITICAL)**: The ground-truth answer MUST remain exactly the same after distractor injection. Do not alter the core reasoning chain.
2. **Naturalness**: The distractor must read naturally within the problem. It should not feel artificially inserted or out of place.
3. **Plausibility**: The distractor should be plausible and contextually relevant enough that a model might be misled by it.
4. **Minimal invasion**: Modify only what is necessary. Do not rewrite the entire problem. Inject the distractor into or around the existing text.
5. **Difficulty calibration**: The distractor should make the problem meaningfully harder, not trivially so. Aim for a difficulty increase that would cause a mid-tier model to fail while a strong model would still succeed.
```

```

## OUTPUT FORMAT
You MUST respond with a valid JSON object and nothing else. No markdown, no explanation outside the JSON.
Use the following schema:
{{
  "original_problem": "",
  "original_solution": "",
  "augmented_problem": "",
  "augmented_solution": "",

  "distractor_metadata": {{
    "distractor_types_used": [
      {{
        "name": "",
        "description": ""
      }}
    ],
    "injected_text_summary": "",
    "why_answer_unchanged": "",
    "estimated_difficulty_increase": ""
  }}
}}

## IMPORTANT GUIDELINES
- Think step by step before generating the output.
- First, understand what the problem is asking and why the given answer is correct.
- Second, identify which parts of the problem can be augmented without breaking the answer.
- Third, apply the specified distractor type as naturally as possible.
- Fourth, draft the distractor text.
- Fifth, verify that the answer is still correct with the distractor in place.
- Only then produce the final JSON output.

## FINAL CHECKLIST (verify before outputting)
- [ ] Is the output valid JSON?
- [ ] Is the answer in augmented_problem identical to the original answer?
- [ ] Does the distractor read naturally in context?
- [ ] Is the specified distractor type used with a clear description?
- [ ] Is the why_answer_unchanged field filled with a logical explanation?
- [ ] Would the augmented problem genuinely be harder for a model to solve?

Now, process the following input and return the JSON output:

Problem: Statement: {problem}
Solution: {solution}
"""
```

Prompt for Reformatting Instruction-Tuning Dataset

```
Role: You are an expert Dataset Engineer specializing in Reinforcement Learning from Human Feedback (RLHF) and Verifiable Rewards.

Objective: Transform a raw task description, input, and output into a structured Problem and Solution pair. This pair must be suitable for RL training where the reward is calculated via exact-match verification.

Constraints:
The Problem: Must incorporate all necessary context from the input and output without giving away the output.
The Solution: Must contain only the final answer. No explanations, no "The answer is...", and no punctuation unless it is part of the value.
Verifiability: The solution must be uniquely extractable via simple string matching or regex.

Formatting Logic:
Open-ended: Use this if the answer is a unique value (e.g., a number, a specific name, or a constant).
MCQ (Multiple Choice): Use this if the task is subjective, has multiple valid answers, or involves Yes/No.
MCQ Format: Provide options labeled (A) through (J). If the task is multi-correct, the solution should be a comma-separated list of letters (e.g., "A, C").
Output Format: Return a valid JSON object with the keys "formatting logic", "problem" and "solution".

Task Description: {def_task}
Input: {example_input}
Output: {example_output}
```

Finally, for win-rate filtering we generate 8 samples from Qwen3-0.6B at temperature=0.7 (generation length: 4096) , our baseline model and compute the per-question win-rate across these 8 samples. We filter all questions with a win-rate of 0 (too hard) and a win-rate of 1(too easy).

Table 7: Top-5 training tasks per BBEH task.

BBEH Task	Rank	Task ID	Training Task
movie recommendation	1	task827	copa_commonsense_reasoning
	2	task069	abductivenli_classification
	3	task212	logic2text_classification
	4	task1297	qasc_question_answering
	5	task1209	atomic_classification_objectuse
word sorting	1	task828	copa_commonsense_cause_effect
	2	task1548	wiqa_binary_classification
	3	task1385	anli_r1_entailment
	4	task835	mathdataset_answer_generation
	5	task383	matres_classification
object counting	1	task1210	atomic_classification_madeupof
	2	task1211	atomic_classification_hassubevent
	3	task1155	bard_analogical_reasoning_trash_or_treasure
	4	task827	copa_commonsense_reasoning
	5	task004	mctaco_answer_generation_event_duration
geometric shapes	1	task249	enhanced_wsc_pronoun_disambiguation
	2	task1209	atomic_classification_objectuse
	3	task1385	anli_r1_entailment
	4	task697	mmmlu_answer_generation_formal_logic
	5	task717	mmmlu_answer_generation_logical_fallacies
nycc	1	task1297	qasc_question_answering
	2	task073	commonsenseqa_answer_generation

continued on next page

Table 7 – *continued from previous page*

BBEH Task	Rank	Task ID	Training Task
	3	task212	logic2text_classification
	4	task213	rocstories_correct_ending_classification
	5	task828	copa_commonsense_cause_effect
boardgame qa	1	task004	mctaco_answer_generation_event_duration
	2	task116	com2sense_commonsense_reasoning
	3	task062	bigbench_repeat_copy_logic
	4	task1726	mathqa_correct_answer_generation
	5	task1387	anli_r3_entailment
buggy tables	1	task007	mctaco_answer_generation_transient_stationary
	2	task1390	wscfixed_coreference
	3	task600	find_the_longest_common_substring_in_two_strings
	4	task391	causal_relationship
	5	task004	mctaco_answer_generation_event_duration
linguini	1	task004	mctaco_answer_generation_event_duration
	2	task1209	atomic_classification_objectuse
	3	task640	esnli_classification
	4	task085	unnatural_addsub_arithmetic
	5	task738	perspectrum_classification
boolean expressions	1	task850	synthetic_longest_palindrome
	2	task600	find_the_longest_common_substring_in_two_strings
	3	task1390	wscfixed_coreference
	4	task018	mctaco_temporal_reasoning_presence
	5	task210	logic2text_structured_text_generation
multistep arithmetic	1	task004	mctaco_answer_generation_event_duration
	2	task1210	atomic_classification_madeupof
	3	task1211	atomic_classification_hassubevent
	4	task007	mctaco_answer_generation_transient_stationary
	5	task1390	wscfixed_coreference
time arithmetic	1	task212	logic2text_classification
	2	task835	mathdataset_answer_generation
	3	task383	matres_classification
	4	task1209	atomic_classification_objectuse
	5	task1153	bard_analogical_reasoning_affordance
object properties	1	task828	copa_commonsense_cause_effect
	2	task004	mctaco_answer_generation_event_duration
	3	task1210	atomic_classification_madeupof
	4	task1211	atomic_classification_hassubevent
	5	task007	mctaco_answer_generation_transient_stationary
hyperbaton	1	task249	enhanced_wsc_pronoun_disambiguation
	2	task213	rocstories_correct_ending_classification
	3	task393	plausible_result_generation
	4	task600	find_the_longest_common_substring_in_two_strings
	5	task827	copa_commonsense_reasoning
sarc triples	1	task210	logic2text_structured_text_generation
	2	task640	esnli_classification
	3	task970	sherliic_causal_relationship
	4	task850	synthetic_longest_palindrome
	5	task717	mmmlu_answer_generation_logical_fallacies
zebra puzzles	1	task828	copa_commonsense_cause_effect
	2	task863	asdiv_multiop_question_answering
	3	task210	logic2text_structured_text_generation
	4	task1726	mathqa_correct_answer_generation
	5	task738	perspectrum_classification
spatial reasoning	1	task738	perspectrum_classification
	2	task087	new_operator_addsub_arithmetic
	3	task019	mctaco_temporal_reasoning_category
	4	task080	piqa_answer_generation
	5	task697	mmmlu_answer_generation_formal_logic
shuffled objects	1	task249	enhanced_wsc_pronoun_disambiguation

continued on next page

Table 7 – *continued from previous page*

BBEH Task	Rank	Task ID	Training Task
	2	task018	mctaco_temporal_reasoning_presence
	3	task827	copa_commonsense_reasoning
	4	task1297	qasc_question_answering
	5	task717	mmmlu_answer_generation_logical_fallacies
temporal sequence	1	task004	mctaco_answer_generation_event_duration
	2	task1210	atomic_classification_madeupof
	3	task1211	atomic_classification_hassubevent
	4	task007	mctaco_answer_generation_transient_stationary
	5	task1390	wscfixed_coreference
sportqa	1	task270	csrg_counterfactual_context_generation
	2	task210	logic2text_structured_text_generation
	3	task062	bigbench_repeat_copy_logic
	4	task600	find_the_longest_common_substring_in_two_strings
	5	task391	causal_relationship
web of lies	1	task1152	bard_analogical_reasoning_causation
	2	task828	copa_commonsense_cause_effect
	3	task1211	atomic_classification_hassubevent
	4	task080	piqa_answer_generation
	5	task640	esnli_classification
causal understanding	1	task383	matres_classification
	2	task004	mctaco_answer_generation_event_duration
	3	task1390	wscfixed_coreference
	4	task828	copa_commonsense_cause_effect
	5	task291	semeval_2020_task4_commonsense_validation
disambiguation qa	1	task697	mmmlu_answer_generation_formal_logic
	2	task1296	wiki_hop_question_answering
	3	task717	mmmlu_answer_generation_logical_fallacies
	4	task018	mctaco_temporal_reasoning_presence
	5	task065	timetravel_consistent_sentence_classification
dyck languages	1	task1390	wscfixed_coreference
	2	task1386	anli_r2_entailment
	3	task004	mctaco_answer_generation_event_duration
	4	task1210	atomic_classification_madeupof
	5	task1152	bard_analogical_reasoning_causation

Table 8: Task descriptions.

Task Name	Summary
task738 perspectrum classification	Decide whether the given perspective supports or undermines the given claim.
task003 mctaco question generation event duration	Writing questions that involve commonsense understanding of "event duration".
task717 mmmlu answer generation logical fallacies	Answering multiple choice questions on logical fallacies.
task249 enhanced wsc pronoun disambiguation	Given a sentence and a pronoun, decide which one of the choices the pronoun is referring to.
task1385 anli r1 entailment	Given a premise and hypothesis, determine if the hypothesis entails, contradicts, or is neutral to the premise.
task1296 wiki hop question answering	Given a subject, a relation, and a context, find the object with that relation to the subject.
task828 copa commonsense cause effect	Given a pair of sentences, judge whether the second sentence is the cause or effect of the first one.

Continued on next page

Task Name	Summary
task073 commonsenseqa answer generation	Answer questions based on commonsense knowledge.
task018 mctaco temporal reasoning presence	Checking the presence of temporal reasoning in a question.
task697 mmmlu answer generation formal logic	Answering multiple choice questions on formal logic.
task827 copa commonsense reasoning	Given a premise and two alternatives, select the alternative that more plausibly has a causal relation with the premise.
task383 matres classification	Given a context and a verb, answer if the given verb can be anchored in time or not.
task065 timetravel consistent sentence classification	Choosing the option that makes a given short story consistent.
task640 esnli classification	Given a premise and hypothesis, determine if the hypothesis entails, contradicts, or is neutral to the premise.
task1387 anli r3 entailment	Given a premise and hypothesis, determine if the hypothesis entails, contradicts, or is neutral to the premise.
task863 asdiv multiop question answering	Given a mathematical question involving multiple operations, find the most suitable numerical answer.
task1209 atomic classification objectuse	Given a tuple, determine whether the Head is used for the Tail or not.
task212 logic2text classification	Given a command, classify the command in one of seven logic types.
task750 aqua multiple choice answering	Given a mathematical question, find the most suitable numerical answer.
task010 mctaco answer generation event ordering	Answering questions that involve commonsense understanding of event ordering.
task1297 qasc question answering	Given two facts and a multiple-choice question, answer the question.
task007 mctaco answer generation transient stationary	Answering questions that involve commonsense understanding of transient vs. stationary events.
task1390 wscfixed coreference	Given a context, a pronoun, and a noun, determine if the pronoun in the context refers to the noun or not.
task600 find the longest common substring in two strings	Given two strings return the longest common substring in those two strings.
task080 piqa answer generation	Generate a solution to a goal regarding physical knowledge about the world.
task1726 mathqa correct answer generation	Generate correct answers for math questions.
task835 mathdataset answer generation	Find the numerical answer for a math word problem.
task580 socialiqa answer generation	Given a context, a question and three options, provide the correct answer based on the context.
task1393 superglue copa text completion	Given a premise sentence, two possible options and a question word, choose the best option.
task1727 wiqa what is the effect	Find the effect of an event on another event, based on an introduced process.
task170 hotpotqa answer generation	Given a set of context and supporting facts, answer the question asked.

Continued on next page

Task Name	Summary
task133 winowhy reason plausibility detection	Detect if a reason that explains an answer to a pronoun coreference resolution question is correct or not.
task004 mctaco answer generation event duration	Answering questions that involve commonsense understanding of event duration.
task019 mctaco temporal reasoning category	Verifying the temporal reasoning category of a given question.
task229 arc answer generation hard	Given a hard science question, provide the answer based on scientific facts and reasoning.
task106 scruples ethical judgment	Given two actions choose the one that is considered less ethical.
task178 quartz question answering	Given a question, select the correct answer from the given options using an explanation.
task1152 bard analogical reasoning causation	Given an analogy that relates actions with their consequences, give the appropriate consequence of the given action.
task090 equation learner algebra	Answer the given equation.
task850 synthetic longest palindrome	Given a string find the longest substring that is a palindrome.
task1422 mathqa physics	Given a problem on physics and options to choose from, find the correct option that answers the problem.
task393 plausible result generation	Given a sentence, write another sentence that is a likely result of it.
task085 unnatural addsub arithmetic	Performing arithmetic with swapped operator symbols.
task1529 scitail1.1 classification	Determining if there is entailment between hypothesis and premise.
task867 mawps multiop question answering	Given a mathematical question involving multiple operations, find the most suitable numerical answer.
task211 logic2text classification	Given a command and corresponding interpretation, classify whether it is the right interpretation or not.
task1548 wiqa binary classification	Binary classification based on steps in wiqa.
task966 ruletaker fact checking based on given context	Fact checking based on given context.
task935 defeasible nli atomic classification	Given a premise, hypothesis and an update, identify whether the update strengthens or weakens the hypothesis.
task116 com2sense commonsense reasoning	Decide whether a sentence is plausible and matches commonsense.
task087 new operator addsub arithmetic	Performing arithmetic with newly defined operator symbols.
task206 collatz conjecture	Given a list of integers, compute the next number in the 3n+1 problem.
task970 sherliic causal relationship	Determine if A and B share a causal relationship.
task086 translated symbol arithmetic	Performing arithmetic with translated operator symbols.
task270 csrg counterfactual context generation	Given a premise, initial context with ending, and new counterfactual ending, generate counterfactual context which supports the new story ending.
task392 inverse causal relationship	Given two sentences, decide whether the first sentence can be the result of the second one.
task105 story cloze-rocstories sentence generation	Given four sentences, predict the next coherent sentence.
task1507 boolean temporal reasoning	Given a statement about date and time values, deduce whether it is true or false.

Continued on next page

Task Name	Summary
task1404 date conversion	Given a date in a particular format, convert it into some other format.
task1153 bard analogical reasoning affordance	Given an analogy that signifies affordances, give the appropriate affordance of the given action.
task069 abductivenli classification	Choosing text that completes a story based on given beginning and ending.
task062 bigbench repeat copy logic	Generating text that follows simple logical operations such as repeat, before, after etc.
task1088 array of products	Given an integer array, return an array such that its element at each location is equal to the product of elements at every other location in the input array.
task190 snli classification	Given two sentences choose whether they agree, disagree, or neither with each other.
task1333 check validity date ddmmyyyy	Given a date in dd/mm/yyyy format, check if it is a valid date or not.
task016 mctaco answer generation frequency	Answering questions that involve commonsense understanding of event frequency.
task1208 atomic classification xreason	Given a tuple, determine whether the Tail is the reason for the Head or not.
task1386 anli r2 entailment	Given a premise and hypothesis, determine if the hypothesis entails, contradicts, or is neutral to the premise.
task1516 imppres naturallanguageinference	Classify a given premise and hypothesis pair.
task199 mnli classification	Given 2 sentences, determine if they clearly agree or disagree with each other or if they cannot be answered.
task1210 atomic classification madeupof	Given a tuple, determine whether the Head is made of the Tail or not.
task217 rocstories ordering answer generation	Given a five sentence story in shuffled order and the title, put the story in the correct order.
task1155 bard analogical reasoning trash or treasure	Given an analogy that relates items to whether they are trash or treasure, determine if the given item is trash or treasure.
task218 rocstories swap order answer generation	Given a five sentence story and the title, determine which two sentences must be swapped so that the story makes complete sense.
task1211 atomic classification hassubevent	Given a tuple, determine whether the Head includes an event or an action in the Tail or not.
task213 rocstories correct ending classification	Given the title and the first four sentences of a five sentence story, choose the correct story ending.