

Tülu 3: Pushing Frontiers in Open Language Model Post-Training

Nathan Lambert^{1,*} Jacob Morrison¹ Valentina Pyatkin^{1,2} Shengyi Huang¹ Hamish Ivison^{1,2}
 Faeze Brahman¹ Lester James V. Miranda¹

Alisa Liu² Nouha Dziri¹ Xinxu Lyu¹ Yuling Gu¹ Saumya Malik¹ Victoria Graf² Jena D. Hwang¹
 Jiangjiang Yang¹ Ronan Le Bras¹ Oyvind Tafjord¹ Chris Wilhelm¹

Luca Soldaini¹ Noah A. Smith^{1,2} Yizhong Wang^{1,2} Pradeep Dasigi¹ Hannaneh Hajishirzi^{1,2}
¹Allen Institute for AI, ²University of Washington

*TüLU 3 was a team effort. ♥ marks core contributors. See full author contributions here.
 Contact tulu@allenai.org.

- 🗄️ **Tulu 3 8B:** [Llama-3.1-Tulu-3-8B](#)
- 🗄️ **Tulu 3 70B:** [Llama-3.1-Tulu-3-70B](#)
- 🗄️ **Tulu 3 405B:** [Llama-3.1-Tulu-3-405B](#)
- 🗄️ **Tulu 3 Data:** [tulu-3-datasets-673b8df14442393f7213f372](#)
- 🔄 **Tulu 3 Code:** [open-instruct](#)
- 🔄 **Tülu 3 Eval:** [olmes](#)
- 🔗 **Demo:** [playground.allenai.org](#)

Abstract



Language model post-training is applied to refine behaviors and unlock new skills across a wide range of language models, but open recipes for applying these techniques lag behind proprietary ones. The underlying training data and recipes for post-training are simultaneously the most important pieces of the puzzle and the portion with the least transparency. To bridge this gap, we introduce TüLU 3, a family of fully-open state-of-the-art post-trained models, alongside its data, code, and training recipes, serving as a comprehensive guide for modern post-training techniques. TüLU 3, which builds on Llama 3.1 base models, achieves results surpassing the instruct versions of Llama 3.1, Qwen 2.5, Mistral, and even closed models such as GPT-4o-mini and Claude 3.5-Haiku. The training algorithms for our models include supervised finetuning (SFT), Direct Preference Optimization (DPO), and a novel method we call Reinforcement Learning with Verifiable Rewards (RLVR). With TüLU 3, we build a multi-task evaluation scheme for post-training with development and unseen evaluations, standard benchmark implementations, and substantial decontamination of existing open datasets on said benchmarks. We conclude with analysis and discussion of training methods that did not reliably improve performance.

The TüLU 3 release includes model weights, a demo, and the complete recipe — datasets for diverse core skills, a robust toolkit for data curation and evaluation, the training code and infrastructure, and, most importantly, a detailed report for reproducing and further adapting the TüLU 3 approach to more domains.

Contents

1	Introduction	5
2	Tülu 3 Overview	7
2.1	Tülu 3 Data	7
2.2	Tülu 3 Evaluation	8
2.3	Tülu 3 Recipe	8
2.4	Evaluation and Results	10
3	Tülu 3 Data	11
3.1	Prompt Curation	11
3.1.1	Sourcing from Public Datasets	11
3.1.2	Synthesizing for Target Skills	13
3.2	Prompt Decontamination	14
4	Supervised Finetuning	15
4.1	SFT Data	15
4.1.1	From Prompts to SFT Data	15
4.1.2	The Tülu 3 SFT Mix	16
4.2	Key Data Experiments	16
4.3	SFT Recipe and Analyses	18
4.3.1	Key Training Experiments	18
4.3.2	Batch Aggregation	19
5	Preference Finetuning	20
5.1	Background	20
5.1.1	Setup	20
5.1.2	Policy Optimization	21
5.2	Tülu 3 Preference Data	21
5.2.1	From Prompts to Preference Data	21
5.2.2	The Tülu 3 Preference Mix	22
5.3	Key Findings of Data Ablations	23
5.4	Preference Tuning Recipe and Analyses	27
5.4.1	Hyperparameter and Algorithm Design	27
5.4.2	Infrastructure for Scaling DPO	29
6	Reinforcement Learning with Verifiable Rewards	30
6.1	RLVR Data	31
6.2	RLVR Recipe and Analyses	32
6.2.1	Key Findings	33
6.3	RLVR Infrastructure	34
6.4	Final Experimental Results	35
7	Tülu 3 Evaluation Framework	36
7.1	Open Language Model Evaluation System (OLMES)	37
7.2	Tülu 3 Evaluation Suite - Development	38
7.2.1	Safety Evaluation	39
7.3	Tülu 3 Evaluation Suite - Unseen	40
7.3.1	New Evaluation: IFEval-OOD	42
7.3.2	New Evaluation: HREF	43
7.4	Evaluating the Development Process Using the Unseen Suite	44
7.4.1	Evaluating the design decisions	44
7.4.2	Comparison with public models	45
8	Discussions	46

8.1	Scaling Tülu 3 Recipe to Llama 3.1 405B	46
8.2	Insights from the Unfruitful	48
8.3	Future Work	49
9	Related Work	49
9.1	The Evolution of Post-training Recipes	49
9.2	Training on Verifiable Rewards	50
10	Conclusion	50
A	Additional Hyperparameters	60
B	Additional Dataset Analyses	60
B.1	Extra Distribution Plots	60
B.2	Contamination in Public Datasets	60
B.3	Chat Template Implementation	60
B.4	RLVR IFEval overoptimization	60
C	Supervised Finetuning Data Details	61
C.1	Prompts	61
D	Preference Tuning Data Details	61
E	Additional RLVR Details	61
E.1	Testing Generalization to Target Evaluations	61
E.2	RM Training Hyperparameters	62
F	Evaluation Details	62
F.1	Detailed Safety Results	62
F.2	Evaluation principles	62
F.3	IFEval Out-of-Distribution Constraints	79
F.4	Subtask-level breakdown of HREF results	81

Table 1 Models, datasets, and code released with TŪLU 3. **Demo:** <https://playground.allenai.org/>

Model Checkpoints		
Stage	Llama 3.1 8B	Llama 3.1 70B
Base Model	<code>meta-llama/Llama-3.1-8B</code>	<code>meta-llama/Llama-3.1-70B</code>
SFT	<code>allenai/Llama-3.1-Tulu-3-8B-SFT</code>	<code>allenai/Llama-3.1-Tulu-3-70B-SFT</code>
DPO	<code>allenai/Llama-3.1-Tulu-3-8B-DPO</code>	<code>allenai/Llama-3.1-Tulu-3-70B-DPO</code>
Final Models (RLVR)	<code>allenai/Llama-3.1-Tulu-3-8B</code> RM: <code>allenai/Llama-3.1-Tulu-3-8B-RM</code>	<code>allenai/Llama-3.1-Tulu-3-70B</code>
Model Checkpoints		
Stage	Llama 3.1 405B	
Base Model	<code>meta-llama/Llama-3.1-405B</code>	
SFT	<code>allenai/Llama-3.1-Tulu-3-405B-SFT</code>	
Final Model (DPO)	<code>allenai/Llama-3.1-Tulu-3-405B</code>	
Codebases / Tools		
Type	🔗 Link	
Training	<code>allenai/open-instruct</code>	
TŪLU 3 EVAL	<code>allenai/olmes</code>	
Decontamination	<code>allenai/open-instruct/tree/main/decontamination</code>	
Preference Data Inference	<code>allenai/birr</code>	
Instruction Datasets		
Type	Domain	🔗 Link
Full mix	General	<code>allenai/tulu-3-sft-mixture</code>
Task Specific Subsets	Precise Instruction Following	<code>allenai/tulu-3-sft-personas-instruction-following</code>
	MATH	<code>allenai/tulu-3-sft-personas-math</code>
	Grade School Math	<code>allenai/tulu-3-sft-personas-math-grade</code>
	Python Code	<code>allenai/tulu-3-sft-personas-code</code>
Preference Mixes		
Model	🔗 Link	
Llama 3.1 70B	<code>allenai/llama-3.1-tulu-3-70b-preference-mixture</code>	
Llama 3.1 8B	<code>allenai/llama-3.1-tulu-3-8b-preference-mixture</code>	
Specific Preference Datasets		
Domain	🔗 Link	
Precise Instruction Following	<code>allenai/tulu-3-pref-personas-instruction-following</code>	
General	<code>allenai/tulu-3-sft-prompts-ultrafeedback</code>	
General	<code>allenai/tulu-3-wildchat-ultrafeedback</code>	
RL with Verifiable Rewards Training Datasets		
Domain	🔗 Link	
Full Mix	<code>allenai/RLVR-GSM-MATH-IF-Mixed-Constraints</code>	
GSM Only	<code>allenai/RLVR-GSM</code>	
MATH Only	<code>allenai/RLVR-MATH</code>	
IFeval Only	<code>allenai/RLVR-IFeval</code>	

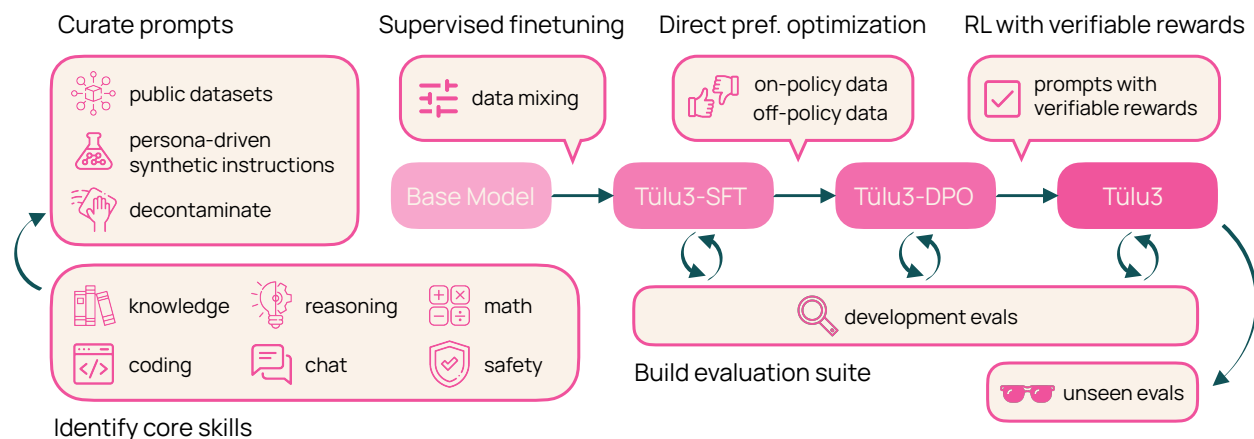


Figure 1 An overview of the TüLU 3 recipe. This includes: data curation targeting general and target capabilities, training strategies and a standardized evaluation suite for development and final evaluation stage.

1 Introduction

“Just as the camel shares its burdens with others in the caravan, the wise share their insights to lighten the load of ignorance.” – Proverb generated by TüLU 3.

Post-training — the collection of techniques including instruction tuning, reinforcement learning from human feedback, and other types of finetuning — has become a crucial step in building frontier language models (OpenAI, 2024; Anthropic, 2024), yet developments to these techniques are frequently not accompanied by open resources and recipes. Fully open source counterparts (e.g., TüLU 2 (Iverson et al., 2023) and Zephyr- β (Tunstall et al., 2023)) often rely on simpler-to-implement and cheaper pipelines and have become outdated on many metrics.

To close the gap between open and closed post training, we introduce **TüLU¹ 3**, a family of open state-of-the-art post-trained models, alongside all of the data, training recipes, code, infrastructure, and evaluation framework. Integrating partial details from proprietary methods with novel techniques and established academic research, TüLU 3 pushes the boundaries of research in post-training. The advancements of TüLU 3 are attributed to TüLU 3 DATA, new permissively licensed training datasets targeting core skills, TüLU 3 EVAL, an evaluation suite and tools to establish clear performance goals and guide improvement through training stages, and TüLU 3 RECIPE, an advanced multi-stage training pipeline incorporating new algorithmic advancements in reinforcement learning, cutting-edge infrastructure, and rigorous experimentation to optimize data mixes, methods, and parameters across various training stages.

In order to build TüLU 3, we identify a set of core skills to improve after training (e.g., reasoning, math, coding, safety, precise instruction following, knowledge recall, etc.) and build an evaluation framework to establish clear performance goals and guide model improvement over a selection of development and unseen tasks. TüLU 3 benefits significantly from leveraging publicly available open data, generating diverse, skill-specific synthetic data at various training stages, and aggressively decontaminating them against our evaluation suite.

The TüLU 3 training recipe involves multiple stages, with each stage building upon the previous model and focusing on different types of data — namely, *prompt-completion* instances for supervised finetuning, *preferences* for preference tuning, or *verifiable rewards* for reinforcement learning. Our methodology facilitates identifying skill deficiencies and refining the data mix, methods and parameters, ensuring a balanced performance of core skills across the training process. Through rigorous, principled experimentation, we determine the best data mix for supervised finetuning, resulting in the TüLU 3 SFT checkpoint. Leveraging recent advances in preference tuning, we then train a model over carefully curated *on-policy* preference data from comparing TüLU 3 SFT completions against outputs from other language models. Furthermore, we introduce a new final finetuning stage – Reinforcement Learning with Verifiable Rewards (RLVR) - which employs a novel

¹A tülu is a hybrid camel bred between Bactrian camel and dromedary: https://en.wikipedia.org/wiki/Hybrid_camel.

Skill	Benchmark _(eval)	Tülu 3 8B	Qwen 2.5 7B Instruct	Llama 3.1 8B Instruct	Tülu 3 70B	Qwen 2.5 72B Instruct	Llama 3.1 70B Instruct	GPT-3.5 Turbo	GPT-4o Mini	Claude 3.5 Haiku
	Avg.	65.1	66.5	62.9	76.2	72.8	74.1	64.7	6970.6	75.3
Knowledge	MMLU _(0 shot, CoT)	68.2	76.6	71.2	83.1	85.5	85.3	70.2	82.2	81.8
	PopQA _(15 shot)	29.1	18.1	20.2	46.5	30.6	46.4	45.0	39.0	42.5
	TruthfulQA _(6 shot)	55.0	63.1	55.1	67.6	69.9	66.8	62.9 [◇]	64.8 [◇]	64.9[◇]
Reasoning	BigBenchHard _(3 shot, CoT)	69.0	70.2	71.9	85.0	80.4	83.0	66.6 [†]	65.9 [◇]	73.7[†]
	DROP _(3 shot)	62.6	54.4	61.5	74.3	34.2	77.0	70.2	36.3	78.4
Math	MATH _(4 shot CoT, Flex)	43.7	69.9	42.5	63.0	75.9	56.4	41.2	67.9	68.0
	GSM8K _(8 shot, CoT)	87.6	83.8	83.4	93.5	89.5	93.7	74.3	83.0	90.1
Coding	HumanEval _(pass@10)	83.9	93.1	86.3	92.4	94.0	93.6	87.1	90.4	90.8
	HumanEval+ _(pass@10)	79.2	89.7	82.9	88.0	90.8	89.5	84.0	87.0	88.1
IF & chat	IFEval _(prompt loose)	82.4	74.7	80.6	83.2	87.6	88.0	66.9	83.5	86.3
	AlpacaEval 2 _(LC % win)	34.5	29.0	24.2	49.8	47.7	33.4	38.7	49.7	47.3
Safety	Safety _(6 task avg.)	85.5	75.0	75.2	88.3	87.0	76.5	69.1	84.9	91.8

Table 2 Overview of results on Tülu 3 Eval suite, over both 8B and 70B models. The best performing model for each model size on each benchmark is bolded. TüLU 3 outperforms the state-of-the-art post-trained open-weight models of the same size and surpass Claude Haiku, GPT-3.5 Turbo, and GPT-4o Mini.

[†] indicates scores taken from Claude 3 Model Card and Claude 3.5 Model Card Addendum.

[◇] indicates score interpolated with Multiple Imputation by Chained Equations (MICE) with context of all other scores in the table, except averages. These scores were either subject to substantial formatting errors in our evaluation suite or not found in other major technical reports. Instruct versions of models shortened to Inst.

Closed model versions: GPT-3.5-Turbo-0125, GPT-4o-mini-2024-07-18, Claude 3.5 Haiku 20241022

RL objective tailored to enhance specific skills with verifiable answers, such as mathematics and precise instruction following.

Our best performing recipe yields TüLU 3 models that outperform the state-of-the-art post-trained open-weight models of the same size such as Llama 3.1 Instruct (Dubey et al., 2024) or Mistral-Instruct (Mistral AI, 2024), and at the large 70B size TüLU matches the offerings of closed providers such as Claude 3.5 Haiku and GPT-4o mini. Furthermore, at 405B size our model performs competitively against DeepSeek v3 (DeepSeek-AI et al., 2024) and GPT 4o (11-24).

In summary, TüLU 3 represents a family of state-of-the-art open language models, featuring a modern post-training framework with fully open-source data TüLU 3 DATA, evaluation TüLU 3 EVAL, training code TüLU 3 CODE and development recipes TüLU 3 RECIPE. Here are a few key contributions from the development of TüLU:

- Extensive guidance and tooling for evaluation, decontamination, and recipe design,
- Scaled, new synthetic instruction datasets,
- Scaling preference data with on-policy generations,
- Reinforcement learning with verifiable rewards, an RL-based method that only gets a reward if the model’s completions are verified to be correct, and
- Advanced infrastructure, details, and code to facilitate the successful implementation of large models.

The result of our work is completely open pipelines for finetuning language models. We release final models trained on Llama 3.1 base versions (Dubey et al., 2024), with intermediate checkpoints, training data, training code, and evaluation code (a full list of artifacts released is available in Table 1). With all the released resources, others can take open base models and finetune them to high-performance on any task of interest – laying the foundation of post-training research within complex, multi-objective and multi-stage training regimes.

Core Skill	Development	Unseen
Knowledge	MMLU _(em)	MMLU-Pro _(em)
	PopQA _(EM)	GPQA _(em)
	TruthfulQA _(MC2 em)	
Reasoning	BigBenchHard _(em)	AGIEval English _(em)
	DROP _(F1)	
Math	MATH _(flex em)	Deepmind Mathematics _(em)
	GSM8K _(em)	
Coding	HumanEval _(Pass@10)	BigcodeBench _(Pass@10)
	HumanEval+ _(Pass@10)	
Instruction Following (IF)	IFEval _(em)	IFEval-OOD _(Pass@1)
	AlpacaEval 2 _(winrate)	HREF _(winrate)
Safety	TÜLU 3 Safety _(avg*)	

Table 3 TÜLU 3 EVAL consists of development and unseen splits to evaluate core skills. With TÜLU 3 EVAL, we release a unified standardized evaluation suite and a toolkit to decontaminate training data against benchmarks. The subscript shows the metric we use for evaluation. TÜLU 3 Safety is a collection of safety evaluations taking the average score across them (avg*), see Sec. 7.2.1 for details.

2 Tülu 3 Overview

Early work in language model post-training followed a standard recipe pioneered by models like Instruct-GPT (Ouyang et al., 2022), consisting of instruction-tuning followed by preference finetuning (PreFT) (Stiennon et al., 2020; Nakano et al., 2021; Askell et al., 2021; Ouyang et al., 2022). Since then, the sophistication and complexity of post-training approaches have continued to increase, moving towards multiple rounds of training, human data plus synthetic data, and multiple training algorithms and objectives (Touvron et al., 2023; Dubey et al., 2024; Gunter et al., 2024). However, most successful post-training models offer limited information about their training data, code, or recipes.² Open post-training research, such as TÜLU 2 (Iverson et al., 2023) and Zephyr- β (Tunstall et al., 2023), show strong results in some benchmarks and on chat evaluations such as AlpacaEval or Arena-Hard (Li et al., 2024a), but still lag behind in core capabilities such as MATH (Hendrycks et al., 2021), IFEval (Zhou et al., 2023) and GSM8K (Cobbe et al., 2021).

TÜLU 3 pushes the boundaries of research in post-training and **closes the gap between open and closed finetuning recipes**. With TÜLU 3, we hope to **uncover which paths for the open-source community will lead to success and which do not** (by reporting negative results). It is a complex training process that integrates partial details from proprietary methods with novel techniques and combines it with established academic research. The key factors in the success of TÜLU 3 are careful data curation, rigorous experimentation and evaluation, innovative methodologies, and improved training infrastructure. We followed systematic guidelines by scientifically evaluating this process through creating development and test sets for evaluation, and conduct careful decontamination of publicly available datasets.

Tülu 3 is not just an artifact, but a comprehensive suite of data and tools designed to advance the frontier of open post-training. By openly sharing our data, recipe and findings, we aim to empower the community to explore new and innovative post-training approaches. We list the extensive artifacts and tools released in Table 1.

2.1 Tülu 3 Data

The TÜLU 3 effort began with identifying key areas where open post-training recipes often fall behind and that are desirable capabilities for generalist language models. Table 3 outlines the core capabilities we aim to

²On LMSYS’s ChatBotArena, no model in the top 50 (as of November 20th, 2024) has released its post-training data (Chiang et al., 2024).

Benchmark _(eval)	Llama 3.1 405B Instruct	Nous Hermes 3 405B	Deepseek V3	GPT 4o (11-24)	Tülu 3 405B SFT	Tülu 3 405B DPO	Tülu 3 405B RLVR
Avg w/o Safety.	78.1	74.4	79.0	80.5	76.3	79.0	80.0
MMLU _(5 shot, CoT)	88.0	84.9	82.1	87.9	84.4	86.6	87.0
PopQA _(3 shot)	52.9	54.2	44.9	53.6	55.7	55.4	55.5
BigBenchHard _(0 shot, CoT)	87.1	87.7	89.5	83.3	88.0	88.8	88.6
MATH _(4 shot, Flex)	66.6	58.4	72.5	68.8	63.4	59.9	67.3
GSM8K _(8 shot, CoT)	95.4	92.7	94.1	91.7	93.6	94.2	95.5
HumanEval _(pass@10)	95.9	92.3	94.6	97.0	95.7	97.2	95.9
HumanEval+ _(pass@10)	90.3	86.9	91.6	92.7	93.3	93.9	92.9
IFEval _(loose prompt)	88.4	81.9	88.0	84.8	82.4	85.0	86.0
AlpacaEval 2 _(LC % win)	38.5	30.2	53.5	65.0	30.4	49.8	51.4
Safety _(6 task avg.)	86.8	65.8	72.2	90.9	87.7	85.5	86.7

Table 4 Summary of TüLU 3 results relative to peer 405B models. The best-performing model on each benchmark (i.e., in each row) is **bolded**. TüLU 3-405B outperforms prior state-of-the-art models finetuned from Llama 3.1 405B Base and rivals some leading, closed models. Progress across various checkpoints highlight the contribution of each stage of the training in improving core skills. Note that TruthfulQA and MMLU multiple choice numbers are not compatible with our infrastructure for running evaluations (via log-probs).

enhance and the evaluation benchmarks selected to cover these skills. With TüLU 3, we focus on core skills of knowledge recall, reasoning, mathematics, coding, instruction following, general chat, and safety.

We curate and collect TüLU 3 DATA to target these core skills by sourcing from public data and synthetically curating data. We use various data formats at different stages of training. Table 7 outlines the collection of datasets used to train our model, and further details are provided in Section section 3.

2.2 Tülu 3 Evaluation

A key factor in the success of our post-training approach is establishing clear performance goals and evaluation tools to guide improvement. With TüLU 3 EVAL, we release a unified, standardized evaluation suite and a toolkit to guide the development of and assessment of final models while decontaminating training data against evaluation benchmarks.

Our framework consists of an open evaluation toolkit for reproducible evaluations (Section 7.1), a suite for evaluating core skills in instruction-tuned models with separate development (Section 7.2) and held-out evaluations (Section 7.3), and a set of recommended settings for evaluating on our evaluation suite based on our experiments with various models. Both splits cover all identified skills, except we have no unseen safety evaluation. Crucially, we did not examine scores on our unseen set when developing our models, allowing us to observe how much we may have overfit to particular evaluations in our decisions around data mixtures, algorithms, and hyperparameters.

Table 3 summarizes our evaluation suite. We provide further details on our evaluations in Section 7 and in Table 24. We publicly release our evaluation suite at <https://github.com/allenai/olmes>.

2.3 Tülu 3 Recipe

In this section, we provide an overview of the TüLU 3 recipe to obtain a state-of-the-art post-trained model. We produce TüLU 3 models through a four-stage post-training recipe on top of pretrained language models (see Figure 1). The TüLU 3 RECIPE is an advanced multi-stage training pipeline incorporating new algorithmic advancements in reinforcement learning, cutting-edge infrastructure, and rigorous experimentation to curate data and optimize data mixes, methods, and parameters across various training stages. Throughout all stages, we measure model performance using a carefully-chosen evaluation suite. The stages are as follows:

Stage 1: Data Curation (section 3) We curate a variety of prompts to be allocated across multiple stages of optimization. We create new synthetic prompts or, when available, source prompts from existing datasets to target specific capabilities. We ensure prompts are not contaminated with our evaluation suite,

Benchmark _(eval)	Llama 3.1 70B Instruct	Qwen 2.5 72B Instruct	Hermes 3 Llama 3.1 70B	Nemotron Llama 3.1 70B	Tülu 3 70B SFT	Tülu 3 70B DPO	Tülu 3 70B
Avg.	74.1	72.8	68.5	72.0	72.6	76.2	76.2
MMLU _(0 shot, CoT)	85.3	85.5	80.4	83.8	78.9	83.3	83.1
PopQA _(15 shot)	46.4	30.6	48.1	36.4	48.6	46.3	46.5
TruthfulQA _(6 shot)	66.8	69.9	66.5	62.6	55.7	67.9	67.6
BigBenchHard _(3 shot, CoT)	83.0	80.4	83.6	78.5	82.6	84.8	85.0
DROP _(3 shot)	77.0	34.2	73.2	68.8	77.2	74.1	74.3
MATH _(4 shot CoT, Flex)	56.4	75.9	41.9	55.0	53.7	62.3	63.0
GSM8K _(8 shot, CoT)	93.7	89.5	90.0	84.7	91.1	93.5	93.5
HumanEval _(pass@10)	93.6	94.0	89.6	94.1	92.9	92.4	92.4
HumanEval+ _(pass@10)	89.5	90.8	85.9	85.5	87.3	88.4	88.0
IFEval _(prompt loose)	88.0	87.6	76.0	79.9	82.1	82.6	83.2
AlpacaEval 2 _(LC % win)	33.4	47.7	28.4	66.1	26.3	49.6	49.8
Safety _(6 task avg.)	76.5	87.0	57.9	69.0	94.4	89.0	88.3

Table 5 Summary of TüLU 3 results relative to peer 70B models. The best-performing model on each benchmark (i.e., in each row) is **bolded**. TüLU 3-70B significantly outperforms prior state-of-the-art 70B models. Progress across various checkpoints highlight the contribution of each stage of the training in improving core skills. Nemotron Llama 3.1 70B is the only model in the table that fine-tuned from another post-trained model (in this case Llama 3.1 70B Instruct), while the others are from their respective base models. Many of the lowest values are caused by failing to follow the few-shot formatting required for the evaluation or other repetitive errors – for more details, see [section 7](#).

TüLU 3 EVAL.

Stage 2: Supervised Finetuning (section 4) We perform supervised finetuning (SFT) on carefully selected prompts and completions. With thorough experimentation, the final SFT data and training hyperparameters are determined to enhance target core skills without significantly impacting the performance of others, guided by our evaluation framework.

Stage 3: Preference Tuning (section 5) We apply preference tuning, specifically DPO, to newly curated on-policy synthetically created preference data from selected prompts along with off-policy data. As in the SFT stage, we identify the best preference data mix through thorough experimentation, uncovering what formats of data, methods, or hyperparameters lead to improvements.

Stage 4: Reinforcement Learning with Verifiable Rewards (section 6) We introduce a new RL-based post-training stage which trains the model on verifiable rewards instead of a reward model, as is common for traditional RLHF training. We select tasks with verifiable outcomes, such as mathematical problem-solving, and only provide rewards when the model’s generations are verified to be correct. We then use RL to maximize these rewards.

The key contributions of our TüLU 3 pipeline lie in improved **data**, **methods**, **infrastructure**, and rigorous **evaluation**. Key elements of our pipeline include:

- **Data Quality, Provenance, and Scale (§3)** We obtain prompts by carefully surveying available open-source datasets, analyzing their provenance, and decontaminating them, as well as curating synthetic prompts that target core skills. To ensure effectiveness, we conduct thorough experiments to study their impact on our development evaluation suite. We find targeted prompts to be influential to improve core skills, while real-world queries, e.g., WildChat (Zhao et al., 2024), are important to improve general chat capabilities. Using the TüLU 3 EVAL decontamination tool, we ensure prompts are not contaminated against our evaluation suite.³
- **Creating a Multi-Skill SFT Dataset (§4.1)** The distribution of the prompts in the “general” and “skill-specific” categories was refined by several rounds of supervised finetuning on various data mixtures. For example,

³We observe a non-trivial amount of contamination in a few open datasets with popular evaluation benchmarks. Details are provided in [Table 37](#).

Benchmark _(eval)	Llama 3.1 8B Instruct	Owen 2.5 7B Instruct	Maggie 8B	Gemma 2 9B Instruct	Minis- tral 8B Instruct	Tülu 3 8B SFT	Tülu 3 8B DPO	Tülu 3 8B
Avg.	62.9	66.5	49.3	60.4	59.6	60.6	64.7	65.1
MMLU _(0 shot, CoT)	71.2	76.6	62.0	74.6	68.5	65.9	68.7	68.2
PopQA _(15 shot)	20.2	18.1	22.5	28.3	20.2	29.3	29.3	29.1
TruthfulQA _(6 shot)	55.1	63.1	57.0	61.4	55.5	46.8	56.1	55.0
BigBenchHard _(3 shot, CoT)	71.9	70.2	55.2	64.9	70.8	69.7	68.7	69.0
DROP _(3 shot)	61.5	54.4	49.4	58.8	56.2	61.3	62.5	62.6
MATH _(4 shot CoT, Flex)	42.5	69.9	5.1	29.8	40.0	31.5	42.0	43.7
GSM8K _(8 shot, CoT)	83.4	83.8	61.2	79.7	80.0	76.2	84.3	87.6
HumanEval _(pass@10)	86.3	93.1	75.4	71.7	91.0	86.2	83.9	83.9
HumanEval+ _(pass@10)	82.9	89.7	69.1	67.0	88.5	81.4	78.6	79.2
IFEval _(prompt loose)	80.6	74.7	38.8	69.9	56.4	72.8	81.1	82.4
AlpacaEval 2 _(LC % win)	24.2	29.0	49.0	43.7	31.4	12.4	33.5	34.5
Safety _(6 task avg.)	75.2	75.0	46.4	75.5	56.2	93.1	87.2	85.5

Table 6 Summary of TüLU 3 results relative to peer 8B models. The best-performing model on each benchmark (i.e., in each row) is **bolded**. TüLU 3-8B significantly outperforms prior state-of-the-art 8B models. Progress across various checkpoints highlight the contribution of each stage of the training in improving core skills. Many of the lowest values are caused by failing to follow the few-shot formatting required for the evaluation or other repetitive errors – for more details, see section 7.

to improve mathematical reasoning, we first establish an upper bound in our evaluation suite by creating math-specialized models, then mix data to bring the general models closer to this upper bound.

- **Curating an On-Policy Preference Dataset** (§5.2) We develop an on-policy data curation pipeline to scale our preference dataset generation. Concretely, we generate completions from TüLU 3-SFT and other models for given prompts, and obtain preference labels through their pairwise comparisons. Our approach extends and improves the off-policy preference data generation method by Cui et al. (2023). Careful multi-skill selection of preference data yields 354,192 instances for preference tuning demonstrating significant improvements in a range of tasks.
- **Preference Tuning Algorithm Design** (§5.4) We experiment with several preference tuning algorithms and observe improved performance in using length-normalized Direct Preference Optimization. We prioritized simplicity and efficiency in experimentation and used length-normalized DPO throughout the development process and training our final models, in lieu of more costly investigations into RL-based methods, such as PPO.
- **Skill-Specific RL with Verifiable Rewards** (§6) We adapt a new approach, leveraging a standard reinforcement-learning paradigm to target skills that can be evaluated against a ground-truth outcome (e.g., Math). We refer to this algorithm as Reinforcement Learning with Verifiable Rewards (RLVR); it obtains a constant reward value if a completion is successful. Our results show that RLVR can improve GSM8K, MATH, and IFEval performance.
- **Training Infrastructure for Reinforcement Learning** (§6.3): We implemented an asynchronous RL setup: we run LLM inference efficiently via vLLM while the learners perform gradient updates concurrently. Our RL codebase is also highly scalable and can train 70B and 405B RLVR policy models.
- **Evaluation Framework: Tülu 3 Eval** (§7) In addition to evaluating the final models, our evaluation framework is an open evaluation toolkit designed to guide the development progress through carefully selected evaluation suite and tools for decontamination.

2.4 Evaluation and Results

When reporting scores throughout this work, we use the metrics identified in Table 3; higher is better. When computing overall performance, we simply average scores across all evaluations, treating each evaluation

equally. For generative evaluations our output length is 4096.

TÜLU 3 trained on Llama 3 base models outperforms all other open-weight models in its size category on our development evaluation suite. Compared to closed models, TÜLU 3 70B even surpasses closed models such as GPT-3.5-Turbo-0125 or GPT-4o-mini-2024-07-18, while approaching the performance of Claude 3.5 Haiku 20241022. The summary of TÜLU 3 trained on Llama 3 at 8 and 70 billion parameters versus the leading models in their size classes is shown in Table 2. A per training stage breakdown of performance is shown for the 8B version in Table 6 and for 70B in Table 5.

With our models trained from raw pretrained base models, we compare to instruct models trained on the same base models (e.g. Nous Hermes 3), instruct models on similar sized, but different base versions (e.g. Ministral 8B or Qwen 2.5 Instruct), and other finetuning recipes trained on an instruct version (e.g. Nemotron Llama 3.1). At 70B, we compare to and surpass Llama 3.1 70B Instruct, Qwen 2.5 72B Instruct (Qwen Team, 2024), Nous Hermes 3 70B (Teknium et al., 2024) (trained on Llama 3.1 70B), and Nemotron Llama 3.1 70B (Wang et al., 2024c) (trained on Llama 3.1 70B Instruct). At 8B, we compare to and surpass Llama 3.1 8B Instruct, Gemma 2 9B Instruct (Gemma Team et al., 2024), Nous Hermes 3 8B (trained on Llama 3.1 8B), Qwen 2 7B Instruct, and Ministral 8B Instruct 2410.

Artifacts Released. We release all artifacts associated with the TÜLU 3 training recipe – including SFT, DPO, and RL model checkpoints, along with new SFT and DPO datasets. A summary of the artifacts released with TÜLU 3 is included in Table 1.

3 TÜLU 3 Data

Prompts represent the diverse ways users may interact with models and serve as the essential component for all post-training stages. We curate an extensive collection of millions of prompts as the starting point of TÜLU 3 post-training recipe. Data selected for next stages of training are selected from these prompts. Table 7 summarizes the key information of these prompts. In this section, we describe our prompt curation process and the decontamination effort to ensure that our evaluations are not leaked in these prompts. In the following sections, we describe how prompts are used for supervised finetuning §4 and preference tuning §5.

3.1 Prompt Curation

To target the desired core skills, we curate a *diverse* and *high quality* set of prompts from publicly available datasets with clear *provenance* and synthetically generate prompts to fill any gaps.

3.1.1 Sourcing from Public Datasets

Since the release of our TÜLU 2, the community has witnessed a large body of work creating datasets for post-training, in terms of both supervised finetuning and preference tuning. TÜLU 3 aims to integrate and extend these resources to build stronger models. We start this process with a broad survey of public datasets, including those annotated by dedicated workers, sourced from real users, and synthesized with models.⁴ We then manually review each individual dataset, and pick those with the following considerations.

Diversity. The diversity of training data is critical for eliciting models’ generalization, avoiding model forgetting, and making models robust to uncommon inputs (Wang et al., 2022c; Chung et al., 2024; Zhou et al., 2024). We pick datasets that can promote diversity, including: WildChat (Zhao et al., 2024), which is a large source of real-user interaction with models; Open Assistant (Köpf et al., 2024), which is created by volunteer workers for general chatting; No Robots (Rajani et al., 2023), which is annotated by expert workers for a broad range of open-ended categories; and FLAN v2 (Longpre et al., 2023), which is a big compilation of classical NLP tasks. We also include a decontaminated subset of UltraFeedback (Cui et al., 2023), which is a composition of several datasets (FalseQA (Hu et al., 2023), UltraChat (Ding et al., 2023), Evol-Instruct (Xu et al., 2023), FLAN v2 (Longpre et al., 2023)) and has shown strong performance for general preference tuning in early studies (Tunstall et al., 2023; Ivison et al., 2024).

⁴The datasets we compiled and consider are available here: <https://docs.google.com/spreadsheets/d/1E2ScaKWbTn1e1zJzcddCzEtf7WrpF3a5ZP5Zvds0Z4Y/edit?usp=sharing>.

Category	Prompt Dataset	Count	# Prompts used in SFT	# Prompts used in DPO	Reference
General	Tülu 3 Hardcoded[†]	24	240	–	–
	OpenAssistant ^{1,2,‡}	88,838	7,132	7,132	Köpf et al. (2024)
	No Robots	9,500	9,500	9,500	Rajani et al. (2023)
	WildChat (GPT-4 subset) [‡]	241,307	100,000	100,000	Zhao et al. (2024)
	UltraFeedback ^{α,2}	41,635	–	41,635	Cui et al. (2023)
Knowledge	FLAN v2 ^{1,2,‡}	89,982	89,982	12,141	Longpre et al. (2023)
Recall	SciRIF [‡]	35,357	10,000	17,590	Wadden et al. (2024)
	TableGPT [‡]	13,222	5,000	6,049	Zha et al. (2023)
Math	Tülu 3 Persona MATH	149,960	149,960	–	–
Reasoning	Tülu 3 Persona GSM	49,980	49,980	–	–
	Tülu 3 Persona Algebra	20,000	20,000	–	–
	OpenMathInstruct 2 [‡]	21,972,791	50,000	26,356	Toshniwal et al. (2024)
	NuminaMath-TIR ^α	64,312	64,312	8,677	Beeching et al. (2024)
Coding	Tülu 3 Persona Python	34,999	34,999	–	–
	Evol CodeAlpaca ^α	107,276	107,276	14,200	Luo et al. (2023)
Safety	Tülu 3 CoCoNot	10,983	10,983	10,983	Brahman et al. (2024)
& Non-Compliance	Tülu 3 WildJailbreak^{α,‡}	50,000	50,000	26,356	Jiang et al. (2024)
	Tülu 3 WildGuardMix^{α,‡}	50,000	50,000	26,356	Han et al. (2024)
Multilingual	Aya [‡]	202,285	100,000	32,210	Singh et al. (2024b)
Precise IF	Tülu 3 Persona IF	29,980	29,980	19,890	–
	Tülu 3 IF-augmented	65,530	–	65,530	–
<i>Total</i>		23,327,961	939,344	425,145 ^γ	

Table 7 Summary of our prompt dataset: data for training stages are selected from these prompts. New datasets released with TüLU 3 are **color-coded** for emphasis. Existing datasets we modified due to contamination are marked with α . Datasets with prompts used in TüLU 1 or 2 are marked with ¹ or ², respectively. Datasets marked with [‡] are downsampled from their original datasets, datasets marked with [†] are upsampled. Note that all datasets were filtered to remove specific keywords (e.g., OpenAI) and empty messages, resulting in slightly lower than reported counts. All TüLU 3 datasets with Persona expand the methodology of Chan et al. (2024). The percentages listed per category are out of the total prompts. Preference count is marked with ^γ to note that not all prompts are used in both the 8B and 70B mixes – for exact details see Table 15.

Target Skills. We especially consider enhancing several capabilities that can power common use cases and our specific needs. As shown in our earlier study (Wang et al., 2023), some capabilities, such as complex reasoning, coding, and precise instruction following, benefit from mixing in additional data. Therefore, we include the following datasets: OpenMathInstruct (Toshniwal et al., 2024) and NuminaMath (Beeching et al., 2024) for mathematical reasoning, Evol-CodeAlpaca for coding, a subset of Daring-Anteater (Wang et al., 2024d) for precise instruction following, Aya (Singh et al., 2024b) for multilinguality, SciRIF (Wadden et al., 2024) for scientific literature understanding, and TableGPT (Zha et al., 2023) for processing table-related tasks. We have also considered other datasets for domains with plenty of published research (e.g., math), but they either did not bring additional benefits in our early supervised finetuning experiments or have restrictive licenses.

Data Provenance and Licenses. When sourcing prompts, we take careful consideration of the licenses of the original datasets and only allow those with clear and correct licenses. Since many publicly released datasets are compositions of other datasets, we have to manually track the provenance of subsets to verify their licenses

and remove those that have issues. Specifically, the ShareGPT dataset⁵ is of questionable legal provenance as they were shared by users on the internet without an agreement to be used for model training or being released at all, so we exclude it and use WildChat instead. We also removed the relevant subset from UltraFeedback and decided not to use Helpsteer2 (Wang et al., 2024d) due to the use of ShareGPT in their prompts. All the datasets included in our final curation have clear licenses.

3.1.2 Synthesizing for Target Skills

To address the growing need for diverse and skill-specific datasets, we incorporate synthetic data generation as a complementary approach. Synthetic data generation has gained traction as a promising alternative to human-written data due to being cheaper to obtain, customizable for different purposes, and reflecting the vast knowledge of the underlying models (Dubey et al., 2024). However, generating diverse and high-quality data at scale is non-trivial, as LMs are susceptible to falling into repetitive modes or patterns, referred to as “mode collapse” (Kazdan et al., 2024). To ensure diversity in generation, we follow the recent *persona-driven* methodology in Chan et al. (2024) to generate synthetic data. The key idea is to use different personas (e.g., “A machine learning researcher focused on neural networks”) with a data synthesis prompt (e.g., “create a coding problem”) to steer an LLM to synthesize data with corresponding perspectives. Specifically, we condition on ~250K personas from Persona Hub (Chan et al., 2024) to generate prompts targeting specific skills such as precise instruction following, math and coding. We detail our procedure for each select skill below. Prompts used to generate these instructions can be found in Appendix C.1. Additionally, we build upon our previous efforts in Brahman et al. (2024); Han et al. (2024); Jiang et al. (2024), to generate noncompliance and safety data.

Precise Instruction Following. Precise instruction following is the ability to follow verifiable instructions in natural language, such as “your answer should contain exactly 3 paragraphs,” that can be automatically verified with heuristics. We use our persona-driven approach to synthetically generate verifiable instructions covering 25 different constraint types defined in IFEval benchmark (Zhou et al., 2023). More concretely, we start by manually writing 1-2 example instructions per constraint (*e.g.*, **number of words**), resulting in total of 33 verifiable instructions which we used as seed prompts. We then generate new instructions using GPT-4o (OpenAI, 2024)⁶ given a data synthesis prompt, persona, and a single verifiable instruction as an example. Figures 30 and 31 show the exact prompts used to generate the instruction and its corresponding response, respectively. In total, we collected 29,980 verifiable instruction-response pairs which we call IF-PERSONA-SFT. Lastly, we also generate another type of prompts targeted for constrained instruction following by randomly sampling instructions from the TULU 2 SFT mix and combining them with constraints from the taxonomy in Zhou et al. (2023). We call that set IF-AUGMENTED. These prompts are only used for the DPO and RLVR stages.

Math and Coding. We follow a similar persona-driven approach to synthetically generate diverse math word and coding problems. Math problems include those that require advanced mathematical skills as well as grade school problems. For coding, we generate Python programming questions that are solvable by entry- to medium-level programmers. Unlike precise instruction following, we zero-shot prompt GPT-4o to generate problems that are unique and specific to a given *persona* input. Having generated the problems, we then generate multi-step math solutions using GPT-4o, and Python programs using `claude-3-5-sonnet`. Exact prompts used to generate problems and solutions are provided in Figures 33, 35, 34, and 36, respectively. In total, we collected ~220K and 35K instances for math reasoning and coding.

Noncompliance and Safety. As we enhance models’ capabilities to assist users effectively, it is crucial to ensure they can reliability reject unsafe and appropriately handle nuanced and out of scope queries. To support this, we curate a set of noncompliance queries (Brahman et al., 2024) that the model ought to not comply with, alongside safety-related direct and adversarial prompts (Han et al., 2024; Jiang et al., 2024) covering both benign and harmful scenarios. Our noncompliance and safety prompts are either curated from existing datasets (Zhang and Choi, 2021; Zhao et al., 2024) or synthetically generated from the GPT model family. More specifically, our noncompliance prompts are obtained based on our contextual noncompliance taxonomy

⁵ShareGPT data was initially used to build the Vicuna model (Chiang et al., 2023), but the exact dataset has not been released. Later work mainly used a community reproduced version at https://huggingface.co/datasets/anon8231489123/ShareGPT_Vicuna_unfiltered/.

⁶We use GPT-4o-2024-08-06 for all our persona-driven data synthesis, unless otherwise stated.

Dataset	Eval.	🗑️ Link	% ↓
Evol CodeAlpaca	Hu- manEval	Orig: ise-uiuc/Magicoder-Evol-Instruct-110K New: allenai/evol_codealpaca_heval_decontaminated	3.5
WildChat GPT-4	Safety	Orig: allenai/WildChat-1M-Full (GPT-4 instances only) New: allenai/wildchat_gpt4_converted_safety_decontaminated	5.4
WildJailbreak	Safety	Orig: allenai/wildjailbreak New: allenai/wildjailbreak_safety_decontaminated	0.7
WildGuardmix	Safety	Orig: allenai/wildguardmix New: allenai/wildguardmixtrain_safety_decontaminated	1.1
NuminaMath- TIR	MATH	Orig: AI-MO/NuminaMath-TIR New: allenai/numinamath_tir_math_decontaminated	11.3

Table 8 Decontaminated datasets. % is the percent of the dataset removed.

from Brahman et al. (2024), spanning multiple categories including *incomplete*, *unsupported*, *indeterminate*, and *humanizing* requests (in addition to *unsafe* requests). Our safety-related prompts are carefully selected among synthetic adversarial prompts, synthetic vanilla (direct) requests, real-world user-LLM interactions (In-The-Wild), and curated annotator-written examples to maximize coverage, diversity, and balance.

3.2 Prompt Decontamination

One important consideration when curating our training mix was possible overlap between training prompts and evaluation sets. We quantify such overlap as follows and remove instances from our training mix as needed in order to prevent test set contamination.

Matching Method. We experimented with full-string, n-gram, and embedding-based matching and found that n-gram matching yielded the most useful results — while embedding-based methods can in principle identify non-trivial contamination like that due to paraphrasing (Yang et al., 2023), we found it difficult to distinguish mere distributional similarity from actual paraphrasing. Moreover, partial surface-level overlap using n-gram matching successfully identified cases of contamination where the instances were trivially different, e.g., a math problem where only the numbers differ.

Identifying Matching Instances. Since completions in training datasets are often regenerated using language models, we chose to compute overlap in the prompts alone (or more generally user turns in multi-turn dialogues). We used 8-gram matching for our contamination checks following (Dubey et al., 2024; Singh et al., 2024a). For each token in a test instance, we consider it to match a token in a train instance if the two instances share an 8-gram containing that token, and we consider the test instance itself to have significant overlap with a train instance if more than 50% of the test tokens have 8-gram matches with the same training instance.

Decontamination. We consider a training set to be contaminated if any number of its instances overlap with more than 2% of the instances in any of the evaluations in our development and unseen suites. We remove all the training sets that were contaminated with our unseen evaluations. For training sets that were contaminated with our development evaluations, we removed the entire dataset if doing so did not significantly impact the performance of the resulting model; otherwise, we removed the specific instances that match any test instance.

The list of datasets we decontaminated and the versions we released with overlapping samples removed is shown in Table 8. The full list of public datasets that we found to be significantly contaminated with our

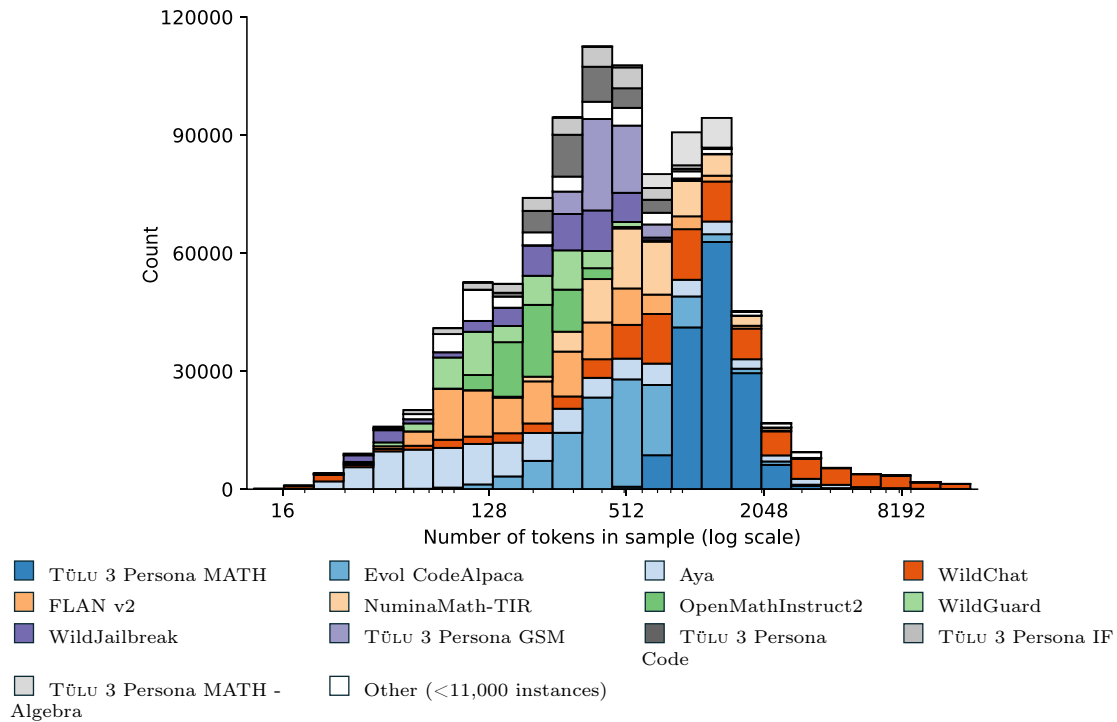


Figure 2 The TüLU 3 final SFT mix by source and length of the prompt plus completion in tokens (using the Llama 3 tokenizer). Compare this distribution to previous open SFT training datasets in Fig. 26. Datasets with the most instances are on the bottom of the histogram.

evaluation sets can be found in Table 37.

4 Supervised Finetuning

Adapting pretrained base models to various tasks and user requests often relies on supervised finetuning (SFT), also known as instruction finetuning. A key challenge in this process is balancing the proportions of mixed training datasets representing diverse skills. For TüLU 3, we conducted data mixture ablations and explored model merging techniques to develop an SFT training procedure that well balances performance across the core skills we prioritized. The following sections detail our experiments and findings.

4.1 SFT Data

4.1.1 From Prompts to SFT Data

To create our SFT mix, we collect or create responses for prompts described in Section 3 in two ways: filtering existing responses, and creating new responses.

For prompts with existing responses, we generally keep the original response if it was written by a human or a frontier model, like GPT-4o. For large datasets with subsets from frontier models (e.g. WildChat), we use the subset from the best models. We additionally filter empty responses and responses that contain information about models or their developers. If a set of prompts did not have responses, like our Persona prompts, or if the original responses were from a weaker model (e.g. WildGuardMix), we generate new responses using GPT-4o. We also hand-wrote responses to our hardcoded prompts.

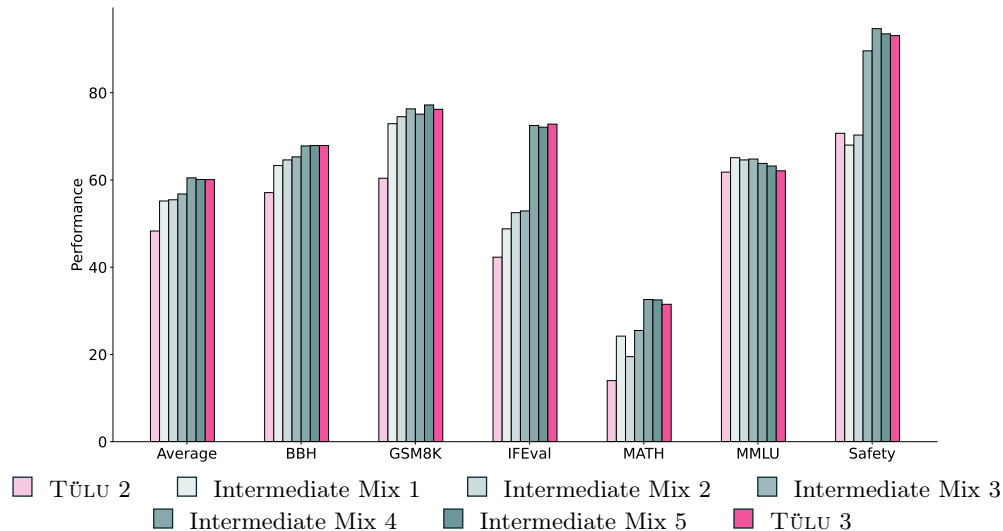


Figure 3 Average and selected skill-specific performance from training Llama 3.1 8B on our initial TüLU 2 SFT mix, and our intermediate and final TüLU 3 SFT mixes. Intermediate mixes 1, 2, and 3 were the result of adding new datasets to improve performance. Intermediate mixes 4 and 5 were the result of running multiple rounds of decontamination, causing small drops in performance.

4.1.2 The Tülu 3 SFT Mix

To develop our SFT mix, we first identified the skills that were lagging behind state of the art models using Llama 3.1 trained on TüLU 2⁷ as our baseline. Targeting each of these skills in isolation, we collected high quality publicly available datasets and created synthetic datasets, as described in Section 3.1.2, and also removed some datasets that we identified to be of relatively lower quality compared to other more recent datasets.

To design our final SFT mix, we first built skill-specific data mixtures and models, keeping the mixtures that led to the best performance on individual skills, ignoring other evaluations. This was done to approximate the upper bound for each evaluation given our setup.

We then combined these mixtures to create our initial TüLU 3 preview mix. We then continued to iterate on the mixture by adding or removing datasets to improve lagging skills, decontaminating against our evaluations and downsampling particularly large datasets. We show the performance of major preview versions throughout development in Figure 3.

Final SFT Results. In Table 9, we compare our final TüLU 3 8B SFT and TüLU 3 70B SFT models against other SFT-only models trained on Llama 3 8B or 70B. Our new SFT mix shows substantial improvements over the TüLU 2 mix at both model sizes, and is better on average the other competitive 8B SFT models.

4.2 Key Data Experiments

We also ran a series of controlled experiments after developing our final SFT mix to explore the importance of different decisions made during data mixing and training.

Diverse Chat Data. In our mix we also emphasized adding diverse chat data, mainly from WildChat. We show the impact of removing WildChat in Table 10, and we see that there is a small but noticeable degradation on most skills, most noticeably on Alpaca Eval, highlighting the importance of diverse real-world data.

Safety is Orthogonal. We found that our safety SFT data was generally orthogonal to our other datasets. We report the effect of removing our safety-specific datasets in Table 10, and we see that most skills stayed roughly the same, except the safety average. We also found that adding contrastive prompts, such as those in CoCoNot, were helpful for preventing our models from over-refusing safe prompts.

⁷<https://huggingface.co/allenai/llama-3.1-tulu-2-8b>

Model	Avg.	MMLU	TQA	PopQA	BBH	CHE	CHE+	GSM	DROP	MATH	IFEval	AE 2	Safety
TÜLU 2 8B SFT	48.3	61.8	49.4	23.3	57.1	66.9	63.1	60.4	61.7	14.0	42.3	8.9	70.7
RLHFlow SFT V2	56.0	65.8	56.0	29.7	69.3	86.2	80.9	81.6	57.2	35.7	52.7	13.6	43.5
MAmmoTH2 8B	46.4	63.6	42.7	20.8	63.4	72.8	66.4	63.7	43.8	30.5	34.9	6.5	47.8
Tülu 3 8B SFT	60.1	62.1	46.8	29.3	67.9	86.2	81.4	76.2	61.3	31.5	72.8	12.4	93.1
TÜLU 2 70B SFT	63.6	76.0	57.8	44.1	79.4	86.8	83.5	83.2	75.9	33.1	57.7	17.3	68.8
Tülu 3 70B SFT	72.6	79.4	55.7	48.6	82.7	92.9	87.3	91.1	77.2	53.7	82.1	26.3	94.4

Table 9 Summary of the performance of our TÜLU 3 SFT models against comparable baselines. Our final SFT mixtures show strong performance, achieving a higher average score than other comparable mixes. All models, including TÜLU 2 SFT, were trained on either Llama 3.0 or 3.1. Our final Tülu 3 70B model was used to help format this table.

Model	Avg.	MMLU	TQA	PopQA	BBH	CHE	CHE+	GSM	DROP	MATH	IFEval	AE 2	Safety
Tülu 3 8B SFT	60.1	62.1	46.8	29.3	67.9	86.2	81.4	76.2	61.3	31.5	72.8	12.4	93.1
→ w/o WildChat	58.9	61.0	45.2	28.9	65.6	85.3	80.7	75.8	59.3	31.8	70.1	7.5	95.2
→ w/o Safety	58.0	62.0	45.5	29.5	68.3	84.5	79.6	76.9	59.4	32.6	71.0	12.4	74.7
→ w/o Persona Data	58.6	62.4	48.9	29.4	68.3	84.5	79.0	76.8	62.2	30.1	53.6	13.5	93.9
→ w/o Math Data	58.2	62.2	47.1	29.5	68.9	86.0	80.5	64.1	60.9	23.5	70.6	12.0	93.5

Table 10 Performance during our SFT ablations, showing the effect of removing safety, WildChat, Persona, and Math data in isolation. We find that: 1) diverse chat data is beneficial for most skills, most noticeably Alpaca Eval, 2) safety performance is generally orthogonal to general performance, 3) our new Persona datasets improve all of the skills that they target, and 4) using mathematics as a test case, adding high quality skill-specific data substantially improves skill-specific performance.

New Persona Data. Our new Persona datasets were built to target specific skills: mathematics, coding, and instruction following. In Table 10 we show that performance on HumanEval(+), GSM8K, MATH, and IFEval drop after removing our Persona datasets, showing the value of creating diverse, skill-specific SFT datasets.

Targeting Specific Skills. A large portion of our focus was on collecting or creating datasets targeting specific capabilities. Using mathematical reasoning as an illustrative example, we show in Table 10 the impact of our mathematics-specific data on both GSM8K and MATH. We see that our mathematics-specific SFT data substantially improves both GSM8K and MATH, showing the value of the data included in the final mix.

Amount of SFT Data. In Figure 4, we show the effect of taking stratified subsamples of our SFT mix. We find that our models continue to improve on average as more SFT data is included, and we see large improvements on metrics like GSM8K as we increase the amount of data to the full mix. Interestingly, TruthfulQA performance actually *drops* as the amount of data in the mix increases. We do not increase our SFT data size beyond the current mixture because we allocated other prompts for preference optimization.

Hyperparameter	8B	70B
Learning Rate	5×10^{-6}	2×10^{-6}
Learning Rate Schedule	Linear	Linear
Batch Size (effective)	128	128
Max Token Length	4,096	4,096
Warm up ratio	0.03	0.03
Number of Epochs	2	2

Table 11 SFT Training Hyperparameters.

4.3 SFT Recipe and Analyses.

Training Settings To train our TÛLU 3 models, we used between 4 and 16 8xH100 nodes with high speed interconnect. The final 8B model is trained on 32 GPUs for 6 hours and the 70B model was trained on 64 GPUs for 50 hours. We used an effective batch size of 128 and a maximum sequence length of 4,096 tokens. We trained for two epochs using a learning rate of 5e-6 for our 8B models, and 2e-6 for our 70B models, which we found after a hyperparameter search. Our hyperparameter settings are also summarized in Table 11. For merging experiments we used mergekit⁸ (Goddard et al., 2024), using linear weighted averaging.

4.3.1 Key Training Experiments

Choice of Base Model. We also test the effect of training different base pretrained models on mathematical performance using our full SFT mix. In Table 12, we show the impact of changing the model’s *size* by training on both Llama 3.1 8B and 70B, and the impact of adding *domain-specific pretraining data* by training on Qwen 2.5 7B and Qwen 2.5 Math 7B. In both cases, we see a substantial improvement in both GSM8K and MATH, highlighting the importance of both model size and pretraining data for downstream skills.

Base Model	GSM8K	MATH
Llama 3.1 8B	76.2	31.5
Llama 3.1 70B	91.1	53.7
Qwen 2.5 7B	79.2	49.4
Qwen 2.5 Math 7B	86.3	56.4

Table 12 Mathematical performance of different base models trained on our mix. We see that 1) training on larger models leads to better performance, and 2) adding skill-specific pretraining data also leads to improved performance, even for the same size model.

Chat Template Variation. During creating TÛLU 3, we explored changing the chat template used to guide the generation of finetuned models. We made a small change to the chat template used in previous TÛLU versions, specifically removing the new line at the end of the template (before the model response). The performance between different changes to the chat template is shown in Table 13 on an early version of our SFT setup. We found that replacing the newlines at the end of assistant messages with an eos token resulted in the best performance, but we opted not to use this to avoid generation inconsistency with later steps in our post-training pipeline. The chat template can be found in our codebase and we provide it in Appendix B.3.

Chat Template	Avg.
TÛLU (replace \n w/ eos)	53.0
Zephyr	52.9
TÛLU 3 (no \n)	52.8
TÛLU 2 template	52.6
Llama 3 template	51.6

Table 13 The impact of different chat templates on SFT model performance, trained using an intermediate SFT mixture on Llama 3.0. While replacing the newline does best, we instead opted for simply removing the newline to avoid complexity.

Random Seeds and Model Soups. We also explored changing the random seed during SFT, and then using those models to create model soups (Wortsman et al., 2022). In Table 14, we compare training 8B and 70B models with multiple different seeds with the best model soup. We see that SFT performance noticeably varies based on the seed, highlighting the importance of multiple training runs, and that the best model soup does not always outperform the best single training run. Because of this, we use the best single SFT training run for each model size as our final SFT models.

⁸<https://github.com/arcee-ai/mergekit>

Model	Seed	Average	Model	Seed	Average
TÜLU 3 8B SFT	42 (Default)	59.9	TÜLU 3 70B SFT	42 (Default)	71.8
	123	60.1		123	70.0
	456	59.8		456	72.6
	789	59.8		-	-
	1011	59.8		-	-
Best Model Soup	42 & 123	60.2	Best Model Soup	123 & 456	72.5

Table 14 Average performance of our 8B and 70B SFT models using random seeds, and compared against the best model soup using the models trained with different seeds. We find that the best random seed is comparable to the best model soup, so for consistency we use the best single SFT run as our final SFT model.

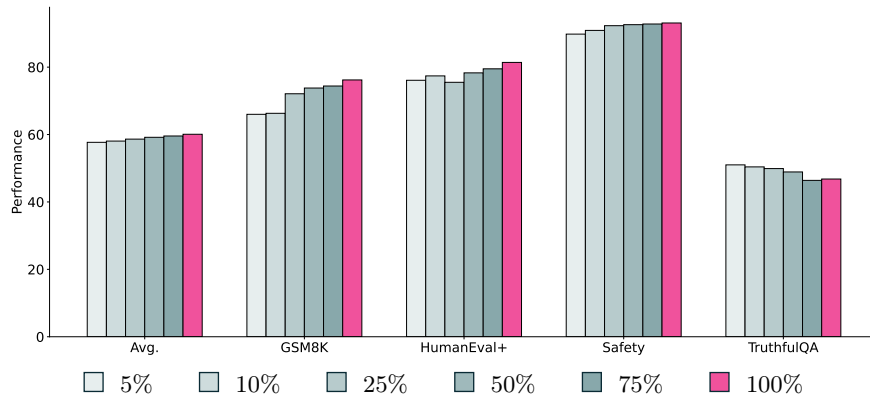


Figure 4 Average and skill-specific performance on stratified subsamples of our final SFT mix. We find that our full mix performs best overall.

4.3.2 Batch Aggregation

Early during training TÜLU 3, we noticed a gap in performance between SFT models trained on our Open-Instruct framework and models trained in other settings such as on TPUs.⁹ We found this issue was largely due to a (recently widely-reported) issue with loss aggregation inside Transformers (Wolf et al., 2020): Averaging the loss across padding tokens without taking into account gradient accumulation or distributed training setups.

Here, we illustrate the issue with an example. Assume we have two samples in a batch, with n_1, n_2 non-padding tokens and m_1, m_2 padding tokens. If we pass both samples into the default Transformers forward pass at the same time, we get:

$$L = \frac{l_{n_1} + l_{n_2}}{n_1 + n_2} \quad (1)$$

However, if we apply gradient accumulation, feeding in the two samples separately, computing loss, and then dividing, our loss is instead computed like:

$$L = \frac{\frac{l_{n_1}}{n_1} + \frac{l_{n_2}}{n_2}}{2} \quad (2)$$

That is, in the second case we weight *each example equally*, while in the first we weight *each token equally*. As such, changing gradient accumulation can have large effects on performance due to effectively changing

⁹Relevant code: <https://github.com/hamishivi/EasyLM>

sample weightings, as reported by Muennighoff et al. (2024). A similar issue occurs in distributed training due to cross-device averaging. We refer to recent reports on this issue for a more in-depth explanation.¹⁰

To fix this issue, we opted generally to use a **sum loss** instead of averaging (‘mean loss’) when training. This removes the issue by simply removing the denominator from the above equations and requires an adjustment to learning rates. This effectively weights all tokens equally (which we found led to generally better performance for initial mixtures). We validated the performance of our setup by finetuning Llama 3.0 on the TŪLU 2 SFT mixture using a variety of learning rates, epochs, and loss types as shown in Figures 5 and 6. Ultimately, we found that using a **sum loss with a learning rate of 5.00E-06** worked best. Surprisingly, we additionally found that training for longer did not yield further improvements, and so used 2 epochs for training.

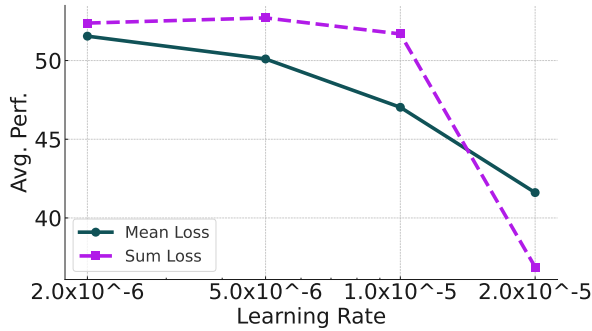


Figure 5 Average performance when finetuning Llama 3.0 on the TŪLU 2 mixture using differing loss types and learning rates. We find that a LR of 5e-6 with a sum loss works best.

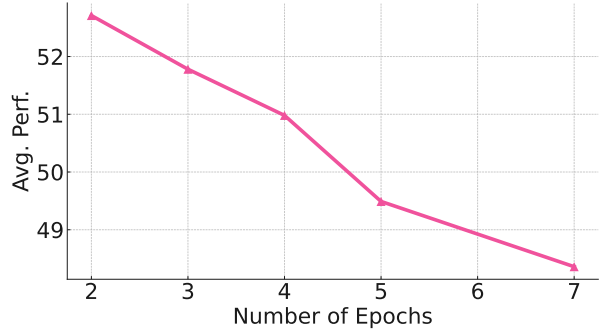


Figure 6 Average performance when finetuning Llama 3.0 on the TŪLU 2 mixture using sum loss and LR of 5e-6 for varying numbers of epochs. We find using 2 epochs works best.

5 Preference Finetuning

For TŪLU 3 we explore many approaches to preference finetuning with the goal of improving our entire evaluation suite. We explore multiple training algorithms, from Direct Preference Optimization (DPO) and its derivatives to reinforcement learning algorithms such as Proximal Policy Optimization (PPO). In this section, we detail the problem formulation of learning from human preferences and our optimizers. Next, we explain how to convert our prompts into synthetic preference data from both on-policy (TŪLU 3 suite) and off-policy models (other instruct models). We show how to create preference data for specific skills of interest and how we improve our models robustly with DPO.

5.1 Background

Prior work has established training on preference data as a crucial step for improving model performance on benchmarks simulating human or synthetic preferences (Dubois et al., 2023; Ivison et al., 2023, 2024). The typical procedure is reinforcement learning from human or AI feedback¹¹ (Ziegler et al., 2019; Stiennon et al., 2020; Ouyang et al., 2022; Bai et al., 2022).

5.1.1 Setup

Preference Data. In the standard setup, there is some preference dataset \mathcal{D} consisting of prompts x and two responses y, y' per prompt. Some judge(s) will choose one of y, y' as their preferred response y_c , and label the other as a rejected response y_r .

Reward Model. Given the preference dataset, a reward model (RM) r_ϕ is trained with the following objective:

$$\max_{r_\phi} \mathbb{E}_{(x, y_c, y_r) \sim \mathcal{D}} [\log \sigma(r_\phi(x, y_c) - r_\phi(x, y_r))] \quad (3)$$

¹⁰<https://unsloth.ai/blog/gradient>,
https://muellerzr.github.io/blog/gradient_accumulation_part2.html

¹¹Now colloquially referred to as synthetic feedback data as well.

where σ is the logistic function. The RM objective maximizes the *difference* between the rewards, and this difference represents the log-likelihood that y_c will be preferred over y_r (Ouyang et al., 2022). This reward model can help train policy models to output contents preferred by the RM’s judgments.

5.1.2 Policy Optimization

There are a plethora of options for optimizing language models with access to preference data. Today, the two categories can be abstracted as reinforcement learning algorithms, which learn from an internal representation of value or reward, and direct alignment algorithms, which learn directly from the data.

Prior work (Ziegler et al., 2019; Stiennon et al., 2020; Ouyang et al., 2022) optimizes the policy π_θ with the following objective:

$$\max_{\pi_\theta} \mathbb{E}_{y \sim \pi_\theta(x)} [R(x, y)] = [r_\phi(x, y) - \beta \text{KL}[\pi_\theta(y|x) \parallel \pi_{\text{ref}}(y|x)]] \quad (4)$$

where π_{ref} is the initial reference policy and the β coefficient helps control the Kullback-Lieber divergence (KL) divergence between the reference policy and the training policy. Here, we explain PPO and DPO as representative examples.

Proximal Policy Optimization (PPO). An approach to address the above objective is to use online reinforcement learning (RL) like PPO (Schulman et al., 2017). In each training iteration of PPO, the policy needs to generate some samples, generate rewards using the RM on those samples, and maximize $R(x, y)$ using the PPO algorithm. As PPO training loops are complex, we refer the reader to Ouyang et al. (2022); Ivison et al. (2024); Huang et al. (2024a) for more thorough descriptions of the setup and typical setups. We provide more implementation details in Sec 6.2.

Direct Preference Tuning (DPO) and Variants. Another approach is offline preference tuning. DPO (Rafailov et al., 2024) can directly optimizes for the RLHF objective with the following equivalent objective:

$$\max_{\pi_\theta} \mathbb{E}_{y_c, y_r \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_c|x)}{\pi_{\text{ref}}(y_c|x)} - \beta \log \frac{\pi_\theta(y_r|x)}{\pi_{\text{ref}}(y_r|x)} \right) \right]. \quad (5)$$

DPO trains an implicit reward model and a policy model simultaneously, without needing to use a trained reward model, do policy generations, and get rewards from the RM. Crucially, this allows offline preference finetuning, directly training a language model on preference pairs gathered from a variety of sources. Recently, much work has examined how to further improve the DPO objective, with a multitude of variants proposed (Meng et al., 2024; Xu et al., 2024a; Hong et al., 2024, *inter alia*). In this work, we explored two promising variants: **SimPO** (Meng et al., 2024) and **length-normalized DPO**¹². We find (in Section 5.4) that length-normalized DPO works best, which uses the following objective:

$$\max_{\pi_\theta} \mathbb{E}_{y_c, y_r \sim \mathcal{D}} \left[\log \sigma \left(\frac{\beta}{|y_c|} \log \frac{\pi_\theta(y_c|x)}{\pi_{\text{ref}}(y_c|x)} - \frac{\beta}{|y_r|} \log \frac{\pi_\theta(y_r|x)}{\pi_{\text{ref}}(y_r|x)} \right) \right]. \quad (6)$$

As seen, this is simply the DPO objective (Eq 5), but with log-probabilities normalized for length, which intuitively aids with mitigating the length bias common in human and model preferences (Singhal et al., 2024).

When developing TüLU 3, we opted to use length-normalized DPO for tuning our preference data mixtures and generation methods due to its relative simplicity and speed compared to approaches such as PPO.

5.2 Tülu 3 Preference Data

5.2.1 From Prompts to Preference Data

We create on-policy preference data (x, y, y', label) given our prompts from section 3 by adapting and advancing the UltraFeedback pipeline (Cui et al., 2023). Our early experiments show the benefit of this pipeline in creating preference data, which leads to a high-quality, synthetic preference dataset (as observed by Ivison et al. (2024)). Our data creation pipeline (shown in Figure 7) consists of three stages: prompt selection,

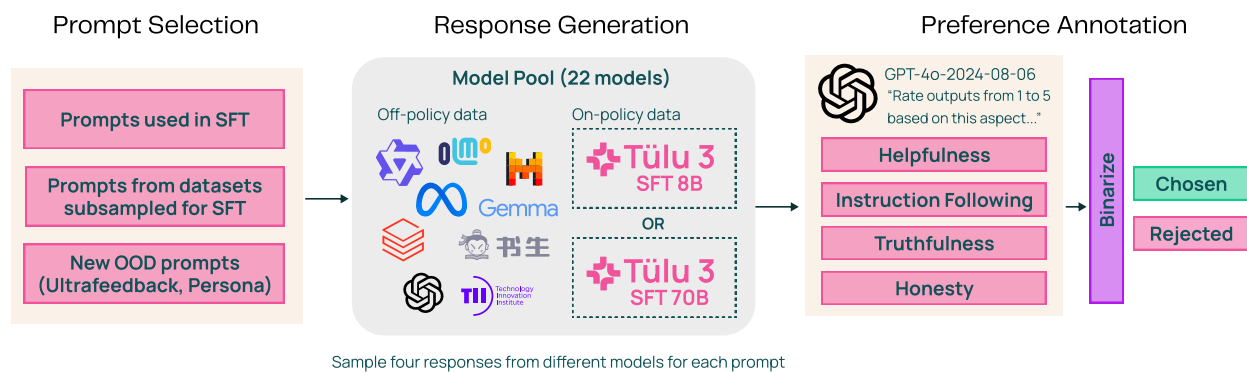


Figure 7 Pipeline for generating and scaling preference data that is based from Ultrafeedback (Cui et al., 2023).

response generation from a pool of models, and preference annotation with LLM-as-a-judge to create (preferred, rejected) pairs.

- Stage 1: Prompt Selection** The first step for preparing a dataset for preference finetuning is to select the prompts or user instructions to generate responses and obtain preferences for. Given the set of prompts in Table 7, we curate our selection to include prompts used during SFT, and prompts that were subsampled from the same sources, yet unused, for SFT. We also include prompts from other sources, such as a version of Ultrafeedback without TruthfulQA instances, or by adding new IF-constraints to a prompt.
 - Stage 2: Response Generation** For a given prompt, we randomly sample four models from a *model pool* to generate responses. Our model selection is inspired by the Ultrafeedback pipeline which consists of open-source and proprietary models that vary across parameter size and model family. We update Ultrafeedback’s model pool by using recent versions of some models (Llama 2 → Llama 3.1), adding best-performing models to increase the pool size, and replacing currently inaccessible models such as WizardLM with open-source alternatives.
- Finally, we also include on-policy data by sampling completions from the TULU SFT model. We approach this by adding a selection of prompts where one response is generated from the on-policy model, and the other response from the off-policy models.
- Stage 3: Preference Annotation** After generating four responses for each prompt, we use an LLM-as-a-judge (Zheng et al., 2023), specifically GPT-4o-2024-0806, to rate each response from 1 to 5 across four different aspects: helpfulness, instruction-following, honesty, and truthfulness.

Appendix D shows the external models used to sample off-policy data and the prompt template for each aspect. In order to obtain binary preferences for DPO, we obtain the mean of preference ratings similar to Argilla’s binarization method¹³ and take the highest-rated response as the chosen response and randomly sample from the responses with the lower mean as the rejected response.

5.2.2 The Tulu 3 Preference Mix

We choose the final preference mix for the 8B and the 70B model, which maximizes average performance on the development evaluations, while also exceling at targeted skills. Most of the preference data mix ablations are run for the 8B model, We start with prompts used for SFT and generate on-policy and off-policy preference data, resulting in 96,911 (off-policy) and 19,444 (on-policy) preference instances. Given this preference base we ablate adding additional prompt sources to the mix and how these additions affect downstream evaluation performance, specifically targeting skills like precise instruction following, math and general chat performance on AlpacaEval. Table 16 shows how the inclusion or exclusion of preference datasets influences the average performance. Our final mixes for TULU 3 8B DPO and TULU 3 70B DPO are displayed in Table 15. In summary, our preference mixes come from different prompt sources, such as SFT data, WildChat and Persona

¹²As proposed in the original Rafailov et al. (2024), but was not yet well optimized to successful hyperparameters until Meng et al. (2024).

¹³<https://huggingface.co/datasets/argilla/ultrafeedback-binarized-preferences/blob/main/README.md>

Dataset	Count	8B	70B
SFT Reused On-policy	19,444	✓	✓
SFT Reused Off-policy	96,911	✓	✓
IF-Augmented	65,530	✓	✓
WildChat IF	10,792	✓	✓
WildChat Reused	17,207	✓	✓
WildChat Unused	82,783		✓
Ultrafeedback (Cleaned)	41,635	✓	✓
Persona IF	19,890	✓	
<i>Total</i>	354,192	271,409	334,302

Table 15 Summary of our best preference dataset mixes for TüLU 3 8B DPO and TüLU 3 70B DPO . IF is short for Instruction Following.

IF. It includes prompts seen during SFT training but also new, unseen prompts.

5.3 Key Findings of Data Ablations

We perform several ablations to inform the design decisions of the synthetic preference pipeline (subsubsection 5.2.1) and the composition of the TüLU 3 preference mix (subsubsection 5.2.2).

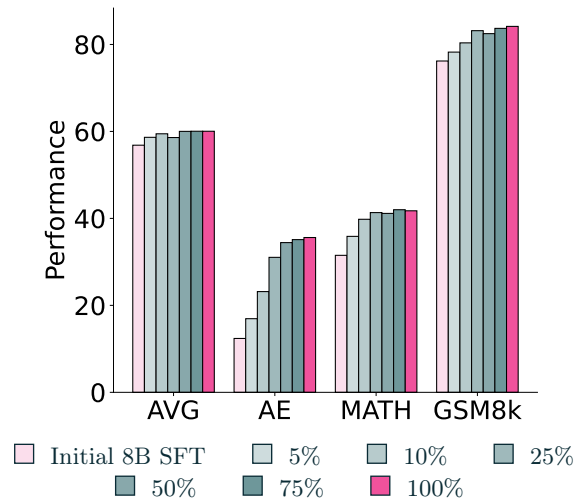


Figure 8 Effect of scaling the size of the preference dataset, specifically the number of unique prompts, on downstream DPO model performance (AE: AlpacaEval).

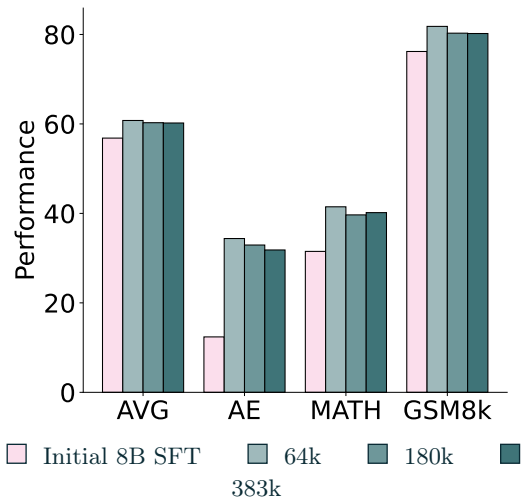


Figure 9 Effect of scaling a preference dataset by duplicating prompts on downstream DPO performance using the Ultrafeedback dataset. All sizes have the same number of unique prompts (64k).

Scaling the Number of Unique Prompts Improve Downstream DPO Performance. First, we investigate whether increasing the number of prompts will yield improvements in downstream DPO performance. To do so, we measure the downstream DPO model performance at different sizes of a fixed amount of preferences with unique prompts. Figure 8 shows that there are noticeable performance gains across several metrics as the size of the preference dataset increases. This suggests that dataset scaling is important to achieve improvements in downstream model performance: our final preference mixes (Table 15) contain more than 270k data points for the 8B model and more than 330k instances for the 70B model, which is considerably bigger than many available preference datasets.

We also explore whether duplicating prompts, i.e. same prompts with different responses, is a viable approach to scaling the size of a preference dataset and whether it will lead to gains in downstream DPO performance. To do so, we expanded the Ultrafeedback dataset, which originally had four responses for each prompt, by

SFT Mix	P-IF	WildC.-IF	SFT-IF	WC ^{β}	WC ^{α}	UF ^{δ}	DA	UF	CocoNot	Avg.
✓		✓	✓		✓	✓				62.27
✓	✓	✓	✓		✓	✓				61.99
✓		✓	✓			✓				61.83
✓		✓	✓	✓						61.76
✓		✓	✓							61.59
✓		✓	✓	✓	✓	✓				61.55
✓								✓		61.35
✓		✓	✓				✓			61.29
✓		✓	✓		✓	✓				61.25
✓		✓	✓	✓		✓			✓	61.17
✓			✓							60.87
✓				✓						60.86
✓	✓									60.84
✓										60.54

Table 16 Some of our dataset mixing experiments to obtain the final preference dataset mix. We include prompts from DaringAnteater (DA), our SFT Mix (SFT), Ultrafeedback (UF), Persona prompts for different skills (P-IF, P-Code, P-Math), TŪLU 3 instruction following prompts (TŪLU 3-IF), i.e. IF-Augmented, CocoNot, the IF subset of Daring Anteater Wang et al. (2024d) and WildChat (WildC.). (α : prompts used during SFT, β : prompts from datasets subsampled, yet unused, for SFT, δ : only used the prompts, the completions and preferences were regenerated using the pipeline described in subsection 5.2.1).

creating additional pair combinations of responses. This expansion will naturally cause duplicated prompts, but with different chosen and rejected pairs sampled from the four responses in UltraFeedback, leading to preference datasets with 64k-, 180k-, and 383k instances. Figure 9 shows that, on average, the 383k-size preference dataset performs similarly to the 64k preference dataset. We also observe a slight performance degradation on DROP, GSM8k, and AlpacaEval as the number of duplicated prompts increase. This suggests that scaling via prompt duplication does not necessarily yield into significant gains in downstream DPO performance, and investing in the collection of unique prompts and proper mixing is more important for downstream evaluations.

Unused Prompts Lead to Higher Performance vs. Reusing Prompts From SFT Mix. We then compare including new prompts and re-using prompts from the SFT stage on their effect on downstream DPO performance. To do so, we sampled 100k prompts from the SFT dataset mix that were *used* during training (as shown in Table 7) and compare it against prompts from the same open datasets (e.g., OpenAssistant, SciRIFF, Aya, Persona, WildChat, etc.) we subsampled from but left *unused* during SFT. Figure 10 shows that the *unused* dataset has a slightly higher performance as opposed to reusing prompts. This suggests that the presence of new prompts can help improve downstream DPO performance. Though, as seen in our best mix, combining unused and reused prompts seems to lead to the best result.

On-policy Data Improves Downstream DPO Performance. We investigate whether the inclusion of *on-policy data*, i.e., text generations from the SFT model that will be used as the base model for preference finetuning, improves downstream model performance. Given the same set of prompts sourced from the SFT mix in section 4, we generate preferences from off-policy models and compared it to a mix that is strictly on-policy (i.e., one of the response is always from the Initial 8B SFT model, and the other response is from the off-policy models). We also compare it on a combination of both on-policy and off-policy data: we sample instances from the strict on-policy dataset and add it to the off-policy dataset so that the responses from each model is distributed equally. Figure 11 shows that including on-policy data improves aggregated downstream DPO performance compared to a completely *off-policy* dataset where prompt completions were sampled from other models.

LLM Judge	Avg.	MMLU	TQA	PopQA	BBH	CHU	CHU+	GSM8k	Drop	MATH	IFEval	AE	Safety
GPT-4o	57.3	64.8	56.1	30.1	66.3	87.0	80.7	75.3	62.7	20.3	60.4	20.6	62.7
Llama 3.1 405B	57.2	64.8	56.0	30.3	67.4	86.2	80.8	75.1	62.0	20.1	59.0	21.5	62.8
GPT-4 Turbo	57.0	64.6	55.7	30.1	66.4	86.6	79.4	75.5	62.6	20.1	59.9	20.6	62.2
GPT-4o Mini	56.9	64.4	55.4	30.4	66.2	86.6	79.8	74.8	60.7	20.9	60.1	21.4	61.6
Llama 3.1 70B	56.6	64.3	55.5	30.2	66.6	85.3	81.4	74.8	62.1	20.1	58.2	18.6	62.2

Table 17 Performance of DPO models trained on preference annotations by different LLM judges. Due to the proximity of the numbers, we have not bolded the max per evaluation.

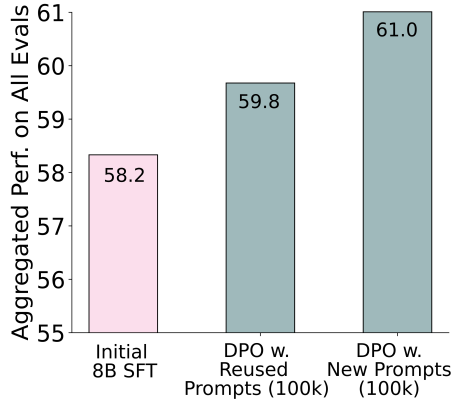


Figure 10 Effect of reusing prompts from SFT mix and new prompts from the same datasets subsampled for the SFT dataset mix.

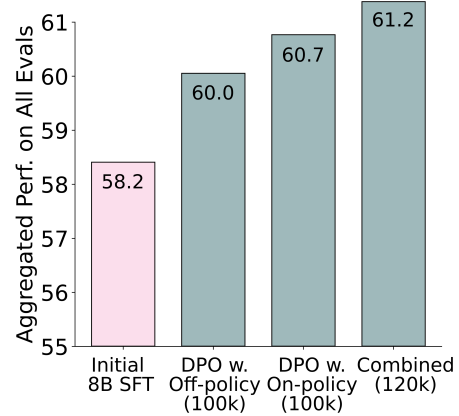


Figure 11 Effect of including on-policy data during the Response Generation stage of the synthetic preference data pipeline on downstream DPO model performance.

Performance Across LLM Judges are Similar, with GPT-4o Leading Slightly Ahead. In order to determine which judge to use for obtaining preference annotations, we test several commercial and open-source LLM judges such as GPT-4 (GPT-4-turbo-2024-04-09, GPT-4o-2024-08-06, gpt-4o-mini-2024-07-18) and Llama 3.1 (70B and 405B) on the same set of 10k randomly-sampled Ultrafeedback prompts and responses. In general, GPT-4o, Llama 3.1 405B, and GPT-4 Turbo perform similarly across all benchmarks, with GPT-4o leading slightly ahead on the aggregated average performance as shown in Table 17. In the synthetic preference pipeline for TULU 3, we opted for GPT-4o-2024-08-06 due to its ease-of-use, cheaper cost per request, and batch inference speed via OpenAI’s Batch API.¹⁴

Going Beyond Ultrafeedback. Previous work on preference learning using openly available datasets has shown that the UltraFeedback (Cui et al., 2023) preference dataset generally outperforms other preference datasets (Iverson et al., 2023). In Figure 12 we show that we were able to significantly surpass DPO training on UltraFeedback by training on our best mix. The improvement is greater for the 70B model (+3.3 vs. +1.8), we hypothesize that this is because UltraFeedback’s completions are mainly sourced from models that are less capable than the 70B model we are starting with. Helpsteer2 Wang et al. (2024d), another high-quality preference dataset, also performs lower than our best mix on the 8B model.

Persona Preference Data. From the three persona preference datasets targeting instruction following, coding and math skills, only TULU 3 Persona IF improves the average eval score and the targeted IFEval score (see Figure 13). Neither TULU 3 Persona Math nor TULU 3 Persona Code improve their respective targeted evaluations and slightly harm the average score. We therefore only include the TULU 3 Persona IF preferences in our final mix.

Targeting IF. We created preference data targeted to improve a model’s precise instruction following skills.

1. **Persona IF:** We take a subset of our collected instruction following SFT dataset, IF-PERSONA-SFT and

¹⁴<https://platform.openai.com/docs/guides/batch>

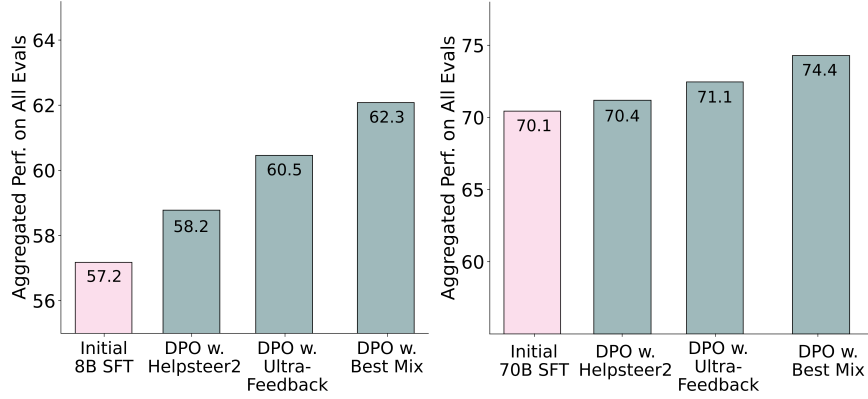


Figure 12 Effect of different DPO mixes on 8B and 70B models: UltraFeedback, Helpsteer2, and our best preference mix.

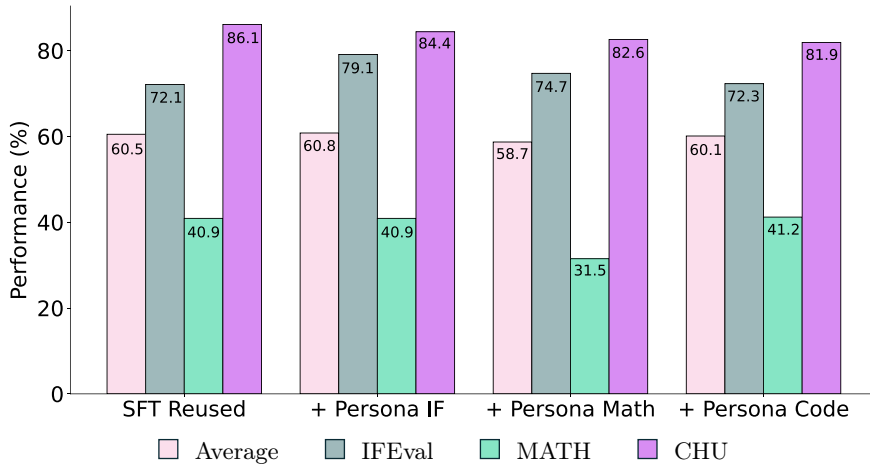


Figure 13 Adding persona preference data to the SFT Reused mix for DPO.

convert it into a preference dataset. Each example in IF-PERSONA-SFT dataset contains a (prompt, constraints, response) tuple. We start by rewriting each prompt in the subset to relax one of the given constraints. More specifically, we prompt GPT-4o to generate rewrites such that the new response to the modified prompt is no longer a valid response for the original prompt (does not satisfy all the constraints). We then use the response to the new modified prompt as the rejected response, and create (chosen, rejected) pairs to form our IF-PERSONA-PREF dataset containing close to 20K examples.

- IF-augmented:** We randomly sample instructions from the TULU 2 SFT mix and combine them with constraints from the taxonomy in Zhou et al. (2023). The chosen and rejected completions are obtained through the synthetic pipeline in §5.2.1.
- WildChat IF:** We sample instructions from WildChat (Zhao et al., 2024) which contain constraints. For this purpose we asked GPT-4 to extract whether or not a prompt includes a constraint.

For IF-augmented, we run two analyses. We generate an additional set of more than 66k instances and we then run the chosen completions through constraint verifier functions, and only add those instances to the final set which actually fulfilled the constraint(s). This leaves us with a cleaned set of about 26k preferences, which we call IF-augmented-verified. In Figure 14 we show that the IF-persona preferences significantly improve IFEval scores beyond the baseline mix, while minimally harming average performance. The IF-augmented-verified dataset improves IFEval performance only by 1 point, while also slightly harming the average performance. Combining IF-persona with IF-augmented-verified leads to the best IFEval performance, but to a slightly lower average. We therefore choose to include IF-augmented (not verified) and Persona IF in the final 8B

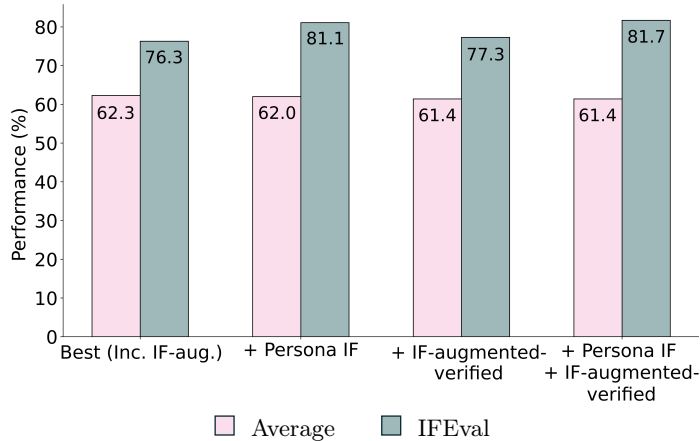


Figure 14 Performance of different IF-targeted preference mixes, average and IFEval. Best here consists of our final best mix for the 8B model (minus Persona-IF).

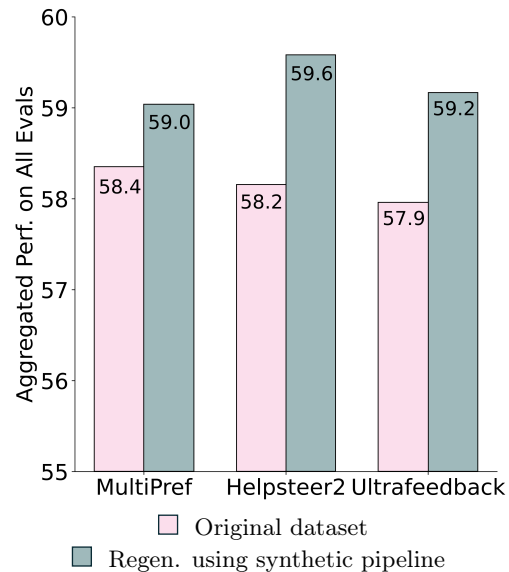


Figure 15 Comparing the use of the original completions to regenerating completions using our synthetic preference pipeline.

DPO mix, which leads to both a satisfying average and IFEval score.

Wildchat. Our ablations show that adding preference data consisting of WildChat prompts and chosen/rejected pairs obtained using our synthetic preference data pipeline, generally improves DPO performance. Ablations in Figure 5.2.2 reveal that adding WildChat prompts seen during SFT training to the DPO mix leads to better average performance than combining the unused with the reused WildChat prompts.

Comparing original preference datasets and their regenerated counterparts. We also investigate whether the preference dataset generated by the synthetic pipeline in §5.2.1 can yield to gains in downstream DPO performance on existing datasets. To do so, we take the prompts from open-source datasets such as Helpsteer2, Ultrafeedback, and MultiPref (Miranda et al., 2024), then regenerate their completions and preference annotations using the synthetic data pipeline. Figure 15 shows that the downstream DPO performance of the regenerated dataset is better than the original dataset, suggesting that the synthetic pipeline itself can yield to performance gains.

5.4 Preference Tuning Recipe and Analyses

5.4.1 Hyperparameter and Algorithm Design

In light of the significant amount of work on improving DPO and related algorithms since the release of TULU 2, we revisited our hyperparameter and algorithm choices alongside our preference datasets. We ablated both algorithm and hyperparameter choices using an early SFT checkpoint and the UltraFeedback dataset. We explored using DPO, SimPO (Meng et al., 2024), and length-normalized DPO. Our results are shown in Table 18. We found that only length-normalized DPO outperformed our base checkpoint overall, and so further tuned it, resulting in the final hyperparameters shown in Table 20.

We lowered the learning rate and increased the batch size for the 70B training based on the fact that it is common to lower the learning rate and increase batch size when doing SFT with larger models (Touvron et al., 2023).

The 8B DPO model is trained for 10 hours on 8 Nvidia H100 GPUs and the 70B DPO model is trained for 19 hours on 64 interconnected H100s.

The DPO training uses a maximum sequence length of 2048.

Algorithm	LR	$\gamma - \beta$ ratio	β	Epochs	Batch Size	Average Score
SFT Base	-	-	-	-	-	55.7
SimPO	5.00E-07	0.5	2	1	128	51.8
SimPO	5.00E-07	0.3	10	1	128	52.9
DPO	5.00E-07	-	0.1	3	32	55.2
PPO	1.00E-06	-	0.0325	1	64	54.5
PPO	1.00E-06	-	0.05	1	64	55.5
DPO-norm	1.00E-07	-	5	3	32	56.1
DPO-norm	5.00E-07	-	10	3	32	55.2
DPO-norm	5.00E-07	-	15	3	32	55.7
DPO-norm	5.00E-07	-	2	3	32	46.8
DPO-norm	5.00E-07	-	5	3	32	53.4
DPO-norm	5.00E-07	-	5	1	32	57.3

Table 18 Hyperparameters and algorithms examined for DPO tuning. We use UltraFeedback as the training dataset in all cases, and train on top of an early TüLU 3 version. DPO-norm refers to the length-normalized DPO variant proposed in Meng et al. (2024). We explore hyperparameters suggested by prior work (Meng et al., 2024; Ivison et al., 2023). For PPO, we train reward models on UltraFeedback and reuse prompts during online training, following the hyperparameters in Ivison et al. (2024). We find that length-normalized DPO performs best overall.

Data	LR	Avg. Performance
Mix 1	5.0×10^{-7}	72.74
	2.0×10^{-7}	71.17
	1.5×10^{-7}	71.12
	1.0×10^{-7}	71.06
Mix 2	5.0×10^{-7}	71.14
	2.0×10^{-7}	74.35

Table 19 Learning rate ablations for the 70B DPO model, for two different preference mixes: Mix 1: Tülu-3-Persona-IF, Tulu-3-Helpsteer2, Ultrafeedback, Tulu-3-SFT-reused (On-policy), Mix 2: Best 70B Mix (both trained on an older SFT base).

Hyperparameter	8B	70B
Learning Rate	5×10^{-7}	2×10^{-7}
Learning Rate Schedule	Linear	Linear
Batch Size (effective)	32	128
Max Token Length	2,048	2,048
KL penalty coefficient β	5	5
Warm up ratio	0.1	0.1
Number of Epochs	1	1

Table 20 Final DPO Training Hyperparameters. We use the length-normalized variant of DPO proposed in Meng et al. (2024).

Learning Rate Ablations for 70B. We ran a small hyperparameter search over a set of learning rates using a generally well performing preference data mix¹⁵ and our final best mix. Table 19 shows that either a learning rate of 2.0×10^{-7} or 5.0×10^{-7} , depending on data mix, performs better than a lower learning rate. For our final DPO models we decided on using a learning rate of 2.0×10^{-7} .

Comparison Between PPO and DPO. We also conducted a more in depth ablation study comparing PPO and DPO later in development. We anchored a DPO preference mix in the development history to train an RM. We use the same setup as Stiennon et al. (2020); Ouyang et al. (2022); Huang et al. (2024a), we only extract the RM’s logits at the end-of-sequence (EOS) token as the reward model. Also, the linear head to output reward scalars is initialized with weights according to $\mathcal{N}(0, 1/\sqrt{(d_{\text{model}} + 1)})$. We use the same prompts in the DPO preference mix to make a controlled comparison between DPO and PPO.

The reward model was trained only once and we *did not* attempt to tune the RM’s performance. Evaluating RM’s performance can be tricky because strong RM performance on RM-specific benchmarks does not necessarily translate to better downstream performance for PPO (Ivison et al., 2024; Chen et al., 2024).

¹⁵Tülu-3-Persona-IF, Tulu-3-Helpsteer2, Ultrafeedback, Tulu-3-SFT-Used (On-policy).

Hyperparameters	for optimizing a RM	for optimizing against RLVR
Discount Factor γ	1.0	1.0
General Advantage Estimation λ	0.95	0.95
Mini-batches N_{mb}	1	1
PPO’s Clipping Coefficient ε	0.2	0.2
Value Function Coefficient c_1	0.1	0.1
Gradient Norm Threshold	1.0	1.0
Learning Rate Schedule	Linear	Linear
Generation Temperature	1.0	1.0
Max Token Length	2,048	2,048
Max Prompt Token Length	2,048	2,048
Penalty Reward Value for Responses without an EOS Token	-10.0	-10.0
Learning Rate	3×10^{-7}	3×10^{-7} (1×10^{-7} for 70B)
Batch Size (effective)	224	224 (640 for 70B)
PPO Update Iterations K	1	4
Response Length	1,024	2,048 (1,024 for GSM8K only)
Total Episodes	300,000	100,000
KL penalty coefficient (β)	[0.05, 0.03, 0.02, 0.01]	[0.1, 0.05, 0.03, 0.01]
Warm up ratio (ω)	[0.1, 0.0]	[0.0, 0.1]

Table 21 The hyperparameters of PPO used for 1) optimizing against a general RM and 2) optimizing against the verifiable reward function. The differences between the hyperparameters are highlighted. The final 8B RLVR model used $\beta = 0.05$ and $\omega = 0.0$; the final 70B RLVR model used $\beta = 0.07$ and $\omega = 0.07$

Furthermore, iterating with RM and PPO is more expensive than iterating with DPO, so we decided to do most of our preference tuning experiments via DPO. The hyperparameters for the RM and PPO can be found in Table 36 and Table 21. The results can be found in Figure 16.

Here are our findings:

1. **PPO Gets Similar Average Scores with DPO in this Non-Tuned Setup** Overall, we found that PPO could reach a comparable level of performance to DPO (albeit slightly lower) in this controlled setup.
2. **PPO is More Computationally Expensive** The PPO runtime is roughly 28 hours using two nodes, whereas the DPO runtime is about 4 hours using a single node.

If we use more computational budget or do more tuning, it is entirely possible that we can push up the PPO’s performance even higher. However, given limited resources and the subtlety in RM evaluation, using DPO for preference tuning seems more economical. We decide to use PPO primarily for RLVR, to be introduced in Section 6.

5.4.2 Infrastructure for Scaling DPO

To run the 70B DPO training, we found it useful to implement two key optimizations for reducing the GPU footprint of DPO training:

1. **Caching DPO Log Probs** To reduce GPU memory usage, we pre-compute and cache log probabilities across the dataset using the initial model, rather than keeping a reference DPO model in memory during training like the canonical implementation (von Werra et al., 2020; Rafailov et al., 2024). This optimization eliminates the need to allocate GPU memory for the reference model.
2. **Separate Forward Passes for Chosen and Rejected Sequences** The canonical DPO implementation (von

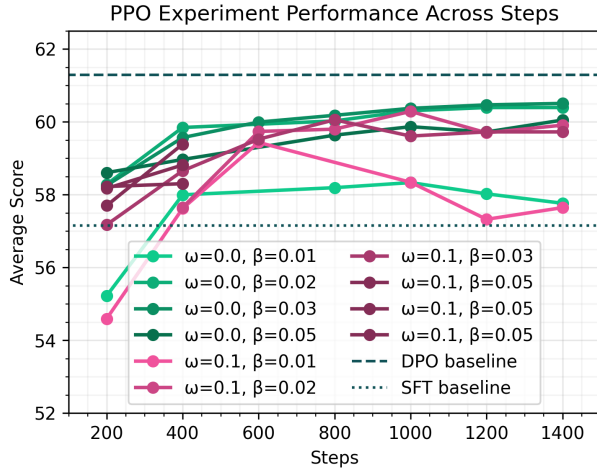


Figure 16 The average scores of PPO runs with different learning rate warm-up ratios ω , KL penalty coefficient β . PPO can get similar (though slightly lower) average scores as DPO.

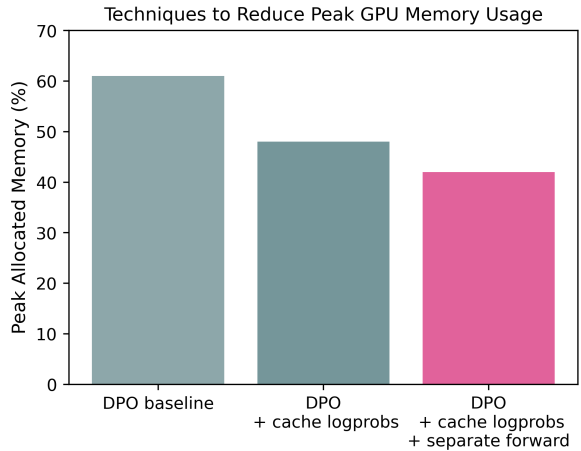


Figure 17 The peak GPU memory allocated can be reduced by caching the reference policy’s logprobs on the preference dataset and doing forward passes separately for the chosen and rejected pairs.

Prompt Dataset	Count	Verification	Reference
GSM8K Train	7,473	Exact match against extracted answer	Cobbe et al. (2021)
MATH Train	7,500	Exact match against extracted answer	Hendrycks et al. (2021)
IF verifiable	14,973	Prompt-specific verifiers	-
<i>Total</i>	29,946		

Table 22 Summary of our verifiable prompt dataset. New datasets released with TüLU 3 are **color-coded** for emphasis.

Werra et al., 2020; Rafailov et al., 2024) also concatenates the chosen and rejected sequences during the forward pass, effectively doubling the batch size and increasing GPU memory requirements. To save GPU memory, we simply perform the forward passes separately on the chosen and rejected completions.

We empirically validated these two techniques on the Llama 3.1 model and found they resulted in near identical training losses. As expected, the model uses less GPU memory when using the two techniques on an 8xH100, as shown in Figure 17.

6 Reinforcement Learning with Verifiable Rewards

In TüLU 3, we introduce Reinforcement Learning with Verifiable Rewards (RLVR), a novel method for training language models on tasks with verifiable outcomes such as mathematical problem-solving and instruction following. RLVR leverages the existing RLHF objective but replaces the reward model with a verification function, as shown conceptually in Figure 18. When applied to domains with verifiable answers, such as mathematics and verifiable instruction following tasks (Zhou et al., 2023), RLVR demonstrates targeted improvements on benchmarks like GSM8K while maintaining performance across other tasks. RLVR can be seen as a simplified form of existing approaches for bootstrapping LM reasoning (Zelikman et al., 2022, 2024; Hoffman et al., 2023) or a simpler form of RL with execution feedback (Gehring et al., 2024), in which we simply use answer matching or constraint verification as a binary signal to train the model. While this has been done for improving math skills alone in prior work (Kazemnejad et al., 2024), we further extend RLVR to cover multiple evaluations and test how it can improve overall model performance, integrating it as a component of a generalist training pipeline.

RLVR is based on a simple principle, common in RL literature, applied to language models: the policy only

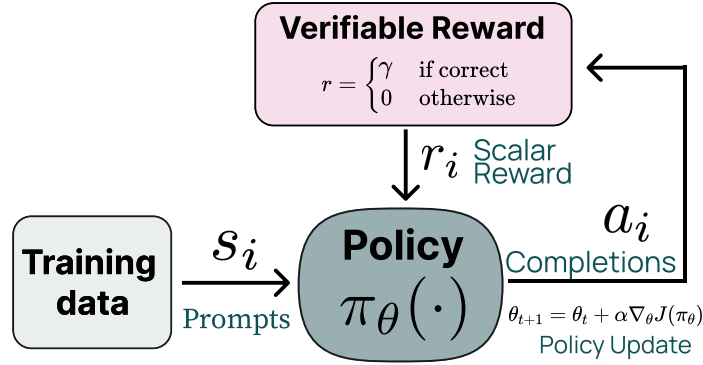


Figure 18 An overview of how Reinforcement Learning with Verifiable Rewards (RLVR) works. We sample completions from a policy given a set of prompts, and verify their correctness using a deterministic function. If the answer is verifiably correct, we provide reward of α , otherwise 0. We then train against this reward using PPO.

receives a reward when its generated responses are verifiably correct. More specifically, RLVR optimizes the following objective:¹⁶

$$\max_{\pi_{\theta}} \mathbb{E}_{y \sim \pi_{\theta}(x)} [R_{\text{RLVR}}(x, y)] = [v(x, y) - \beta \text{KL}[\pi_{\theta}(y|x) \parallel \pi_{\text{ref}}(y|x)]] \quad (7)$$

where v is the verifiable reward function. v takes in a prompt and completion pair (x, y) , and checks if the answer is correct within the generated text:

$$v(x, y) = \begin{cases} \alpha & \text{if correct,} \\ 0 & \text{otherwise.} \end{cases} \quad (8)$$

We train models with RLVR following preference finetuning, and we use the PPO (Schulman et al., 2017) algorithm to optimize for the RLVR objective. We set $\alpha = 10$ based on pilot experiments and did not tune it further.

6.1 RLVR Data

Creating data for RLVR entails obtaining prompts with an accompanying binary verifier (i.e., constructing a set of inputs x with accompanying verifier functions v). We focus on two domains (mathematics, exact instruction following) and three evaluations (GSM8K, MATH, IFEval) with relatively straightforward methods for verification, and leave more complex verifiers to future work.¹⁷ In practice, the answer extraction and verification method is domain-dependent. We use three sources of training prompts and verifiers:

GSM8K. We use the GSM8k training set. We augment each sample with the standard 8-shot prompt using during evaluation to encourage the model to use chain-of-thought, and then extract the final number produced and compare to the ground-truth label to determine correctness.

MATH. We use the MATH training set. Similar to GSM8k, we augment each sample with the standard 3-shot CoT prompt used to encourage the model to generate chains of thought during evaluation, and then extract the answer and determine correctness following the ‘flex’ MATH evaluation logic.

IFEval. We randomly sample instructions from the TULU 2 SFT mix and combine them with constraints from the taxonomy in Zhou et al. (2023). We have a verification function for each of the constraint templates that is able to verify whether a completion satisfies a constraint.

Given these prompts and verification functions, we then train the models via Proximal Policy Optimization (PPO) (Schulman et al., 2017) on these verifiable rewards. We combine all prompts together, results in a

¹⁶This is very similar to the standard KL-constrained RLHF objective, shown in Eq. 4, with a reward function instead of a learned reward model.

¹⁷For example, recent work has found success in using code execution feedback to train models with RL (Gehring et al., 2024; Xu et al., 2024b).

mixture of roughly 30,000 prompts with ground truth labels. We summarize our verifiable prompt mixture in Table 22.

6.2 RLVR Recipe and Analyses

Implementation Details. RL and RLHF have many subtle implementation details that can significantly impact training stability (Engstrom et al., 2020; Huang et al., 2022). As we use PPO to train our models against our verifiers, we adapt these effective implementation details from Huang et al. (2024a):

1. **Initialize the Value model from a General RM** We initialize the value model from the reward model for the RLVR setup (following the standard setup in Ziegler et al. (2019); Ouyang et al. (2022)).
2. **Disable Dropout** We set the dropout probability to be 0 during RM and RL training (similar to Ziegler et al. (2019)). This ensures the token log probabilities can be computed deterministically during the forward passes of the policy model and reference model, allowing a more accurate estimation of the KL penalty. Furthermore, PPO calculates the token log probabilities in two phases: during the rollout phase and the learning phase. It is important to ensure the token log probabilities match up during these two phases: they produce a probability ratio of 1 during the first PPO epoch, so PPO can clip the ratio and apply the surrogate objective properly. If the log probabilities differ drastically due to dropout, all ratios could potentially be clipped, resulting in zero gradient.
3. **Train with the SFT Dataset and Shuffle Between Epochs** As pointed out in in Huang et al. (2024a), PPO can train for more episodes than the total available prompts, effectively training for multiple epochs. In our RLVR ablation experiments, we train for roughly $100,000/7,473 \approx 13$ epochs. We shuffle the prompts in between epochs. For our final runs, we examine model checkpoints every 40-100 steps and choose the best checkpoint on our development evaluation set.
4. **Non End-of-Sequence (EOS) Penalty** During training, PPO typically sample a fixed amount of maximum tokens. If the sampled response does not end with an EOS token, we give a -10 penalty to encourage the model to always complete its responses.
5. **Advantage Whitening / Normalization** Like done in standard PPO implementation details literature (Engstrom et al., 2020; Huang et al., 2022, 2024a), we normalize the advantages by subtracting its mean followed by dividing its standard deviation.

We start by anchoring a DPO model in the development history as the initial model for conducting experiments for RLVR. We conducted several sets of experiments:

1. **Individual Tasks** we applied the RLVR recipe on GSM8K, MATH, and IFEval, respectively with a sweep of beta values [0.1, 0.05, 0.03, 0.01]. For evaluation, we look at the verifiable rewards, the KL divergence, and the response length.
2. **Value Model Initialization Ablation** We experimented with initializing PPO’s value model from 1) a general reward model or 2) the anchored DPO model, and ran a sweep of beta values [0.1, 0.05, 0.03, 0.01, 0.005, 0.001] on the GSM8K task. The general RM is trained with the UltraFeedback dataset (Cui et al., 2023). For evaluation, we examine the GSM8K test evaluation score and the average scores across all evaluation.
3. **Scores from RM Ablations** One possible implementation for RLVR is to add verifiable rewards on top of the scores from the reward model. We launch experiments also using a sweep of beta values [0.1, 0.05, 0.03, 0.01, 0.005, 0.001] on the GSM8K task.
4. **Starting from Weaker Model** The model’s base capabilities are also a confounding factor. We launch another set of experiments using an SFT model with lower average scores (from which the anchored DPO checkpoints were trained, so they share the same linearage) using beta values [0.1, 0.05, 0.03, 0.01].

Unless otherwise specified, we use the following hyperparameters for PPO training in Table 21. To train reward models, we use the hyperparameters in Table 36.

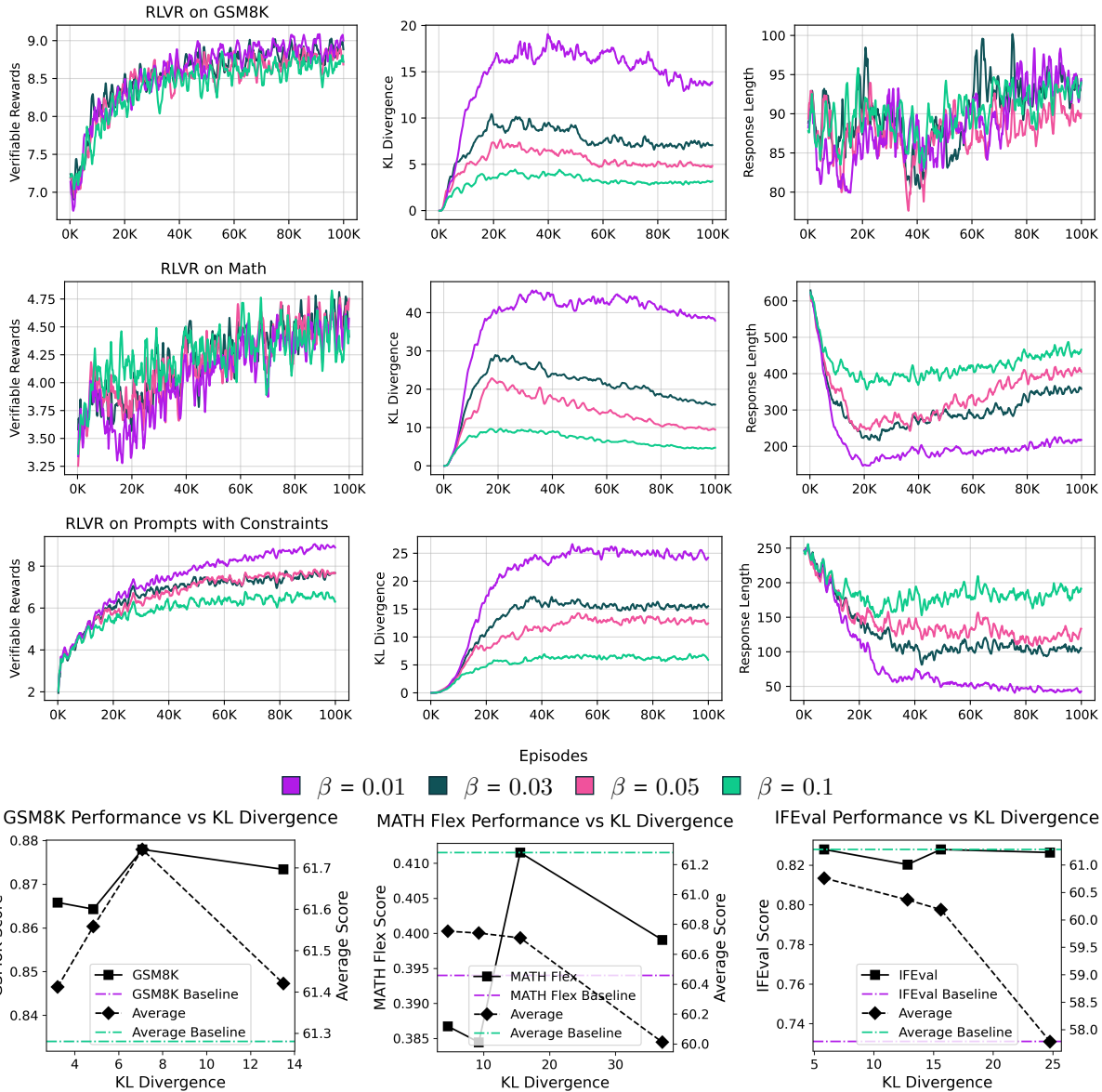


Figure 19 The top three rows show RLVR’s verifiable rewards, KL divergence, and response lengths on the *train* dataset of GSM8K, MATH, and prompts with constraints, when starting from a DPO checkpoint (i.e. an experimental, not final DPO checkpoint). The bottom row shows the corresponding downstream test performance. RLVR can lead to higher verifiable rewards in the train datasets. Importantly, RLVR can also lead to higher scores in the corresponding test dataset, however, an increase in the average score across all evaluations is not guaranteed.

6.2.1 Key Findings

RLVR Can Improve Performance in Targeted Domains. Figure 19 shows training with RLVR results in improved *test* performance for all three settings. In all cases, we achieve models that outperform the initial model in that particular evaluation. We also see that the verifiable rewards (i.e., correctness on the train set) improves consistently for all three settings. Interestingly, in GSM8K and MATH, we found that incurring more KL budget does not necessarily lead to improvements in verifiable rewards.

Initializing RLVR’s Value Function from a General RM Works Best. Figure 21 shows that initializing the value from a general RM obtains the highest GSM8K test score and also higher average scores. This suggests the value function plays an important role in RLVR’s training.

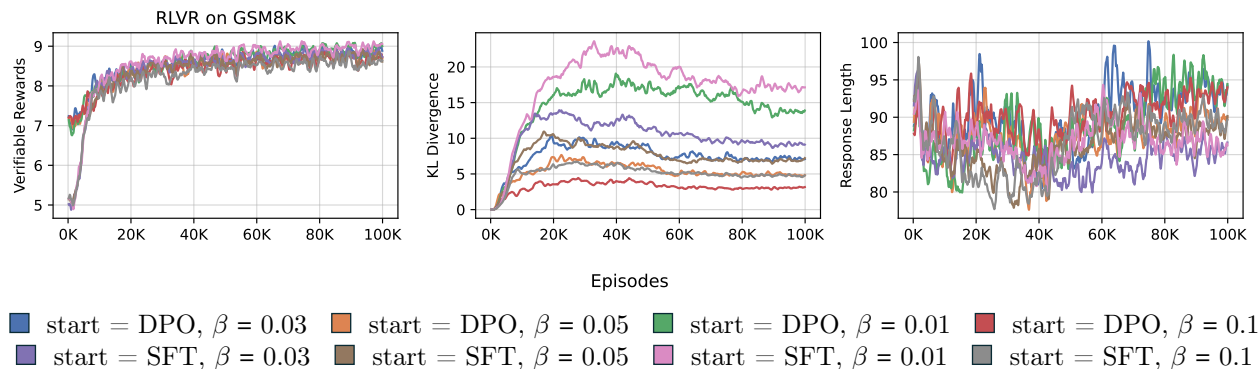


Figure 20 The comparison of RLVR’s performance on GSM8K between starting from a DPO checkpoint and starting from a weaker SFT checkpoint. We see that starting from both SFT and DPO can lead to the same level of verifiable rewards, but starting from SFT would incur a larger KL compared to starting from DPO when using the same β .

Do Not Use the Scores from RM. Figure 22 shows that using only the verifiable rewards outperforms using scores from the reward model. Training with verifiable rewards with the scores from RM seems to introduce more noise, especially in the average scores.

Starting from a Weaker Model Can Converge to the Same Verifiable Rewards. Figure 20 shows that starting from both SFT and DPO can lead to the same level of verifiable rewards, but starting from the SFT model would incur a larger KL compared to starting from the DPO model. This makes sense because the SFT model is further away from good at GSM8K than the DPO model. However, we find that starting from a stronger model usually results in better *test set* performance.

Overoptimization Happens. As we lower the KL penalty β , the trained model incurs more KL from the initial model. We observed that more KL divergence typically results in lower average scores, as shown in Figure 21. The exception is Figure 22, where the largest KL corresponds to the highest average score. Furthermore, we showcase overoptimization of prompts with constraints cases in Appendix B.4.

6.3 RLVR Infrastructure

Our PPO setup follows best practices on implementation details (Huang et al., 2024a). To enable our implementation to scale to models up to 405B parameters, we also adapted model allocation techniques from existing distributed RLHF frameworks (Hu et al., 2024). Furthermore, we accelerate throughput by making RL training asynchronous (Noukhovitch et al., 2024).

The final 8B reward model is trained in 9 hours on 8 H100 gpus, while the final 8B RL run takes 65 hours on 8 GPUs, the final 70B RL run takes 60 hours on 48 GPUs, and the final 405B RL run takes 46 hours on 256 GPUs. Note, for all of these models we took an earlier than final checkpoint from the run.

Distributed Setup. Our PPO infrastructure leverages Zero Stage 3 (Rajbhandari et al., 2020) to fit the models and applicable optimizer states into the memory. In RLVR, we have 3 models: the policy model, the reference policy model, and the value model. The policy and value models need to be trained, but the reference policy model only performs inference. Often inference time is a bottleneck in RLHF infrastructure, so we allocate dedicated GPUs to do inference like done in OpenRLHF (Hu et al., 2024). In particular, we use Ray (Moritz et al., 2018) to allocate dedicated GPUs to run PagedAttention via vLLM (Kwon et al., 2023). PagedAttention helps reduce GPU memory fragmentation and redundant duplication leveraging virtual memory and paging techniques. As a result, it helps our PPO infrastructure run LLM inference using a much larger batch size and speed up inference. This setup allowed us to scale PPO policy training to the 405B scale. We share more 405B scaling details in Section 8.1.

Asynchronous RL Training. Furthermore, our PPO setup uses asynchronous RL training to improve training efficiency (Noukhovitch et al., 2024). The existing PPO frameworks (Hu et al., 2024; Shen et al., 2024a) are typically synchronous: the inference GPUs would first collect policy rollout data, and then the learner GPUs would train on that rollout data. This setup would utilize GPU relatively well under a typical setup with

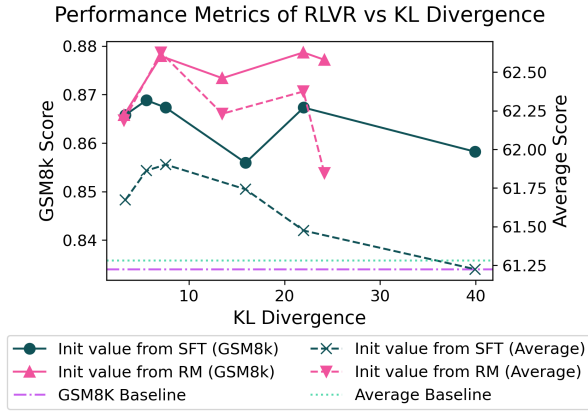


Figure 21 The performance of RLVR vs KL Divergence under different value model initialization. Both initializations could improve the models compared to an initial DPO baseline model. However, as the model diverges more from the initial model, overoptimization happens as the average scores drop significantly.

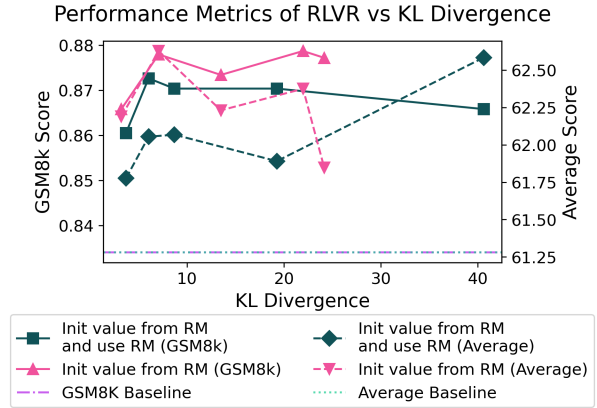


Figure 22 Similar to Figure 21, but this is a comparison of 1) using scores from on top of the verifiable rewards and 2) using only the verifiable rewards. We found using the verifiable rewards performs better in GSM8K, and using scores and verifiable rewards to be more noisy.

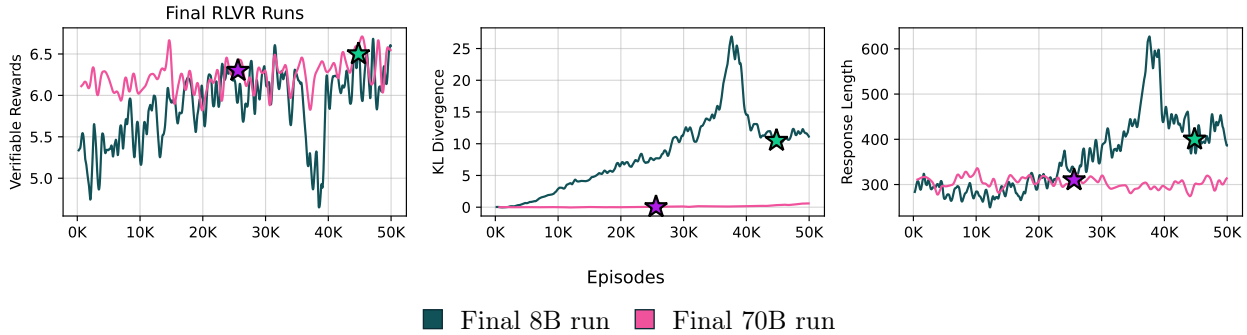


Figure 23 Rewards, KL divergence, and average response length on GSM8k train set over episodes for our final RLVR runs. We mark the point we choose the 8B and 70B checkpoints with a green and purple star respectively.

Atari (Mnih et al., 2015). However, under the RLHF setup, inference computation typically requires different optimizations (e.g., PagedAttention), thus requiring different GPU memory allocation strategies. As a result, synchronous RL training means inference GPUs could be idling while the learner GPUs run, and vice versa. An alternative implementation is to use the same set of GPUs for training and inference, but it could incur additional overhead such as an additional copy of the policy parameters and compiling inference engine in the training GPUs (Shen et al., 2024a).

Our setup allocates GPUs specifically for inference and training, alleviating the need to compile an inference engine and save a copy of the policy parameters in the training GPUs. Furthermore, we run the inference computation concurrently with the training computation, thus reducing the GPU idle time. However, asynchronous RL training can introduce stale data, in cases where the inference can generate data much faster than training consumes (Espeholt et al., 2018), which can introduce reproducibility problems (Huang et al., 2023). To help make training more reproducible, our setup always trains the policy using the second latest inference data (Huang et al., 2023; Noukhovitch et al., 2024).

6.4 Final Experimental Results

Based on the above results, we ran our final RLVR runs using the combined verifiable prompt set, and used the best DPO models from the prior section as starting points. Specifically, at 8B scale, we tested the best overall DPO model and the best model with IFEval persona data mixed in during training, and at 70B scale

Model Size		8B			70B		
Category	Benchmark _(Eval Setting)	Llama 3.1 Inst.	Tülu 3 DPO	Tülu 3 RLVR	Llama 3.1 Inst.	Tülu 3 DPO	Tülu 3 RLVR
Avg.		62.2	64.4	64.8	73.4	75.9	76.0
Knowledge	MMLU _(0 shot, CoT)	71.2	68.7	68.2	85.3	83.3	83.1
	PopQA _(15 shot)	20.2	29.3	29.1	46.4	46.3	46.5
	TruthfulQA _(6 shot)	55.1	56.1	55.0	66.8	67.9	67.6
Reasoning	BigBenchHard _(3 shot, CoT)	62.8	65.8	66.0	73.8	81.8	82.0
	DROP _(3 shot)	61.5	62.5	62.6	77.0	74.1	74.3
Math	MATH _(4 shot CoT, Flex)	42.5	42.0	43.7	56.4	62.3	63.0
	GSM8K _(8 shot, CoT)	83.4	84.3	87.6	93.7	93.5	93.5
Code	HumanEval _(pass@10)	86.3	83.9	83.9	93.6	92.4	92.4
	HumanEval+ _(pass@10)	82.9	78.6	79.2	89.5	88.4	88.0
IF & Chat	IFEval _(Strict)	80.6	81.1	82.4	88.0	82.6	83.2
	AlpacaEval 2 _(LC % win)	24.2	33.5	34.5	33.4	49.6	49.8
Safety	Safety _{6 task avg.}	75.2	87.2	85.5	76.5	89.0	88.3

Table 23 Final performance of RLVR-trained TüLU 3 models compared to Llama 3.1 and DPO starting points. The best-performing model on each benchmark (i.e., in each row) and of each size is **bolded**.

we use the best overall DPO model. For hyperparameters, for 8B models, we used the hyperparameters from Table 21, but tested higher KL penalty coefficients (up to 0.15) based on previous 8B RL development runs. For 70B models, we used the hyperparameters from Table 21, but with a 1×10^{-7} learning rate, 0.1 warmup ratio, 2048 response length, 400,000 episodes, 640 effective batch size, and $\beta = 0.7$ based on previous 70B RL development runs. We initialize our value model from a reward model trained on the same dataset as the best DPO model (the TüLU 3 8B preference mixture) starting from TüLU 3 SFT, using the same hyperparameters as in our ablation experiments (Table 36).

We evaluated our models every 100 training steps (40 for 70B), and picked as our final 8B model the checkpoints with best overall performance on MATH and IFEval. We show the logs from RLVR training in Figure 23, and compare the final performance against their DPO starting points and Llama 3.1 in Table 23. RLVR results in non-trivial improvements at the 8B scale, improving all three of MATH, GSM8k, and IFEval. In fact, we observed that some 8B runs were able to achieve GSM8k scores of up to 89.4% and IFEval scores of up to 84.8% (although such models tended to perform worse in other metrics, dragging down their overall average). At the 70B scale, we observe more modest improvements in IFEval and MATH, and no improvement in GSM8k, likely due to the fact that it is already close to saturation (93.5%). Surprisingly, we find that our 70B run displays extremely low KL divergence, remaining well below 1 over the duration of run, probably due to the lower learning rate¹⁸.

7 Tülu 3 Evaluation Framework

We designed our framework for evaluating TüLU 3 and the other models we compare against with the following goals: 1) Our evaluations should be *reproducible*. 2) We should evaluate models’ generalization to unseen tasks, not just the specific benchmarks we use for development. 3) Our evaluation setup (e.g., templates and strategies for prompting) should be fair to a wide range of models.

Accordingly, our framework consists of an open evaluation toolkit for reproducible evaluations (Section 7.1), a suite for evaluating core skills in instruction-tuned models with separate development (Section 7.2) and

¹⁸We had attempted using a higher learning rate during the initial exploration but found that KL could explode initially and cause a non-trivial drop in average scores.

	Category	Benchmark	CoT	# Shots	Chat	Multiturn ICL	Metric
Development	Knowledge Recall	MMLU	✓	0	✓	✗	EM
		PopQA	✗	15	✓	✓	EM
		TruthfulQA	✗	6	✓	✗	MC2
	Reasoning	BigBenchHard	✓	3	✓	✓	EM
		DROP	✗	3	✗	N/A	F1
	Math	GSM8K	✓	8	✓	✓	EM
		MATH	✓	4	✓	✓	Flex EM
	Coding	HumanEval	✗	0	✓	N/A	Pass@10
		HumanEval+	✗	0	✓	N/A	Pass@10
	Instruction Following	IFEval	✗	0	✓	N/A	Pass@1 (prompt; loose)
AlpacaEval 2		✗	0	✓	N/A	LC Winrate	
Safety	TÜLU 3 Safety	✗	0	✓	N/A	Average*	
Unseen	Knowledge Recall	MMLU-Pro	✓	0	✓	N/A	EM
		GPQA	✓	0	✓	N/A	EM
	Reasoning	AGIEval English	✓	0	✓	✓	EM
	Math	Deepmind Mathematics	✓	0	✓	✓	EM (Sympy)
	Coding	BigCodeBench	✗	0	✓	N/A	Pass@10
	Instruction Following	IFEval-OOD	✗	0	✓	N/A	Pass@1 (prompt; loose)
		HREF	✗	0	✓	N/A	Winrate

Table 24 The TÜLU 3 Evaluation Regime: settings for development (**top**) and unseen (**bottom**) portions of the evaluation suite. **CoT** are evaluations run with chain of thought prompting (Wei et al., 2022b). **#Shots** is the number of in-context examples in the evaluation template. **Chat** refers to whether we use a chat template while prompting the model. **Multiturn ICL** refers to a setting where we present each in-context example as a separate turn in a conversation (applicable only when a chat template is used and # Shots is not 0). *Average over multiple sub-evaluations – full details of the safety evaluation are included in Section 7.2.1.

held-out evaluations (Section 7.3), and a set of recommended settings for evaluating on our evaluation suite that based on our experiments with various models, which we refer to as the TÜLU 3 Evaluation Regime, summarized in Table 24.

As described in Section 2.1, we split our evaluation suite into a *development* set and an *unseen* set, the former used for developing models, and the latter only for evaluating final models. This setup, along with our training data decontamination efforts (see Section 3.2) provide a fairer evaluation of our models generalization capabilities for each of the core skills we focus during development. However, in comparisons between our models and other models, we *cannot rule out that any closed model has not trained on our evaluation suite*, and hence cannot make clear judgments on models that have not publicly released finetuning data.

7.1 Open Language Model Evaluation System (OLMES)

In an effort to make evaluations more standardized and reproducible, we are sharing the code base used to produce the evaluations in this work.¹⁹ The OLMES evaluation system supports:

- A wide range of models and tasks, leveraging existing work in the Eleuther AI LM Evaluation Harness (Gao et al., 2024),
- Flexible configuration options for each task,
- Direct access to the specific task formulations used in this work (as well as in earlier work such as OLMo (Groeneveld et al., 2024) and the OLMES standard (Gu et al., 2024)), and

¹⁹See <http://github.com/allenai/olmes>.

- Detailed instance-level output data for analysis of model predictions, confidences, etc

E.g., to reproduce our Llama-3.1-8B-Instruct numbers for MMLU-Pro, one would simply run something like `“olmes -task mmlu_pro::tulu3 -model llama3.1-8b-instruct”`.

7.2 Tülu 3 Evaluation Suite - Development

We design the evaluation setup for the *development* partition of Tülu 3 EVAL based on existing practices in current literature as well as insights during development. Where appropriate, we adapt the evaluation setup based on the nature of the task and take additional care to robustify our answer extraction and comparison approaches as described below.

MMLU (Hendrycks et al., 2020) is heterogeneous with regard to the type of reasoning skills required to answer the questions, containing instances that require basic factual recall as well as those that demand logical reasoning and problem-solving skills. We design a zero-shot CoT setting that asks the models to “summarize” its reasoning before answering the questions (see Table 44 for the prompt used). We find that among the various CoT settings, including prompting the model to think “step-by-step” and using no CoT, the prompt yields a systematic performance improvement over standard 5-shot multiple choice setting across tested models (See Table 46) and maximizes on the number of subject (knowledge categories) that the CoT benefits. This indicates that our “summarize” prompt is an effective strategy for dealing with the heterogeneity of the benchmark. See appendix F.2 for details. We compute a macro average over all the subjects in MMLU as the final task metric.

PopQA (Mallen et al., 2022) is an entity-centric question-answering benchmark that evaluates language models’ tendency to *forget* information about long-tail entities. We prompt the models in a 15-shot setting (as recommended in the dataset paper) without any additional instructions, with each QA demonstration presented in a different conversation turn, a setting that we refer to as *Multiturn ICL* henceforth, and use greedy sampling to obtain model predictions.

TruthfulQA (Lin et al., 2021) contains questions that humans tend to answer incorrectly due to common misconceptions. We use the multiple-choice version of this benchmark where the models being evaluated are presented with questions and options containing multiple correct answers (the MC2 setting).

HumanEval (Chen et al., 2021) and HumanEval+ (Liu et al., 2023) evaluate models’ ability to complete Python code given docstrings. HumanEval+ uses a more rigorous evaluation procedure than the original HumanEval benchmark with additional tests. We use samples from the models at a temperature of 0.8 and use pass@10 as the evaluation metric.

GSM8K (Cobbe et al., 2021) contains grade school math word problems. We use the 8-shot chain-of-thought prompt from Wei et al. (2022b), formatted for a multiturn ICL evaluation. We obtain model responses using greedy sampling and extract the last number in the model response as the predicted answer.

MATH (Hendrycks et al., 2021) contains problems from mathematics competitions spanning various categories such as algebra and calculus. We use a 4-shot multi-turn setup with CoT from Lewkowycz et al. (2022), formatted for multiturn ICL, and greedy sampling for model completions. For determining the correctness of predictions, we use a ‘flex’ scheme that attempts to extract the answer in three different ways: (1) following the minerva format (Lewkowycz et al., 2022); (2) finding the last instance of ‘< ans >’; (3) taking the text between the last two ‘\$’ tags. This is due to issues we found during development wherein models would often not follow the correct output format despite the few-shot examples, necessitating a varied answer extraction strategy. We find that moving from the minerva format alone to our ‘flex’ strategy can sometimes improve reported scores by up to 10 points, highlighting the need for this flexible strategy. We compute macro average across the subsections to obtain the final task metric.

BigBench-Hard (Suzgun et al., 2022) contains challenging reasoning problems for which models benefit from step-by-step reasoning. We follow the setup described in the original paper and use 3-shot CoT prompts, formatted for multiturn ICL. We use greedy sampling for obtaining model predictions.

DROP (Dua et al., 2019) is a reading comprehension task that requires discrete reasoning. We draw 3 random few-shot examples from the train split following the setup used for Llama 3 (Dubey et al., 2024), and use greedy sampling to get model predictions.

IFEval (Zhou et al., 2023) evaluates instruction following ability of models in a setting where each instruction corresponds to constraints such that it can be programmatically verified whether the outputs satisfy those constraints. We use greedy decoding to generate model outputs given the instructions in the dataset, and measure the prompt-level accuracy of satisfying constraints in the loose evaluation setting.

AlpacaEval 2 (Dubois et al., 2024) contains a set of prompts sourced that reflect real human usages of LMs, and compares model outputs to GPT-4 turbo responses with an additional length control to avoid longer answers being unfairly favoured. We generate responses using greedy decoding up to 8,192 tokens in length, following Ivison et al. (2023).

7.2.1 Safety Evaluation

We follow Han et al. (2024) and Jiang et al. (2024) to define our safety evaluation suite using the following benchmarks.²⁰ Each of these benchmarks evaluates whether models refuse to respond to unsafe requests, and in the case of XSTest and WildJailbreak, additionally evaluate whether they comply with benign requests. We use greedy sampling to get model responses for the prompts in each of the benchmarks, and compute the accuracy (at refusal or compliance as appropriate). We report the macro average of the scores over all the benchmarks as our final safety evaluation metric.

XSTest (Röttger et al., 2023) consists of 200 unsafe prompts and 250 prompts which are safe but superficially resemble unsafe prompts: these prompts use vocabulary similar to that of unsafe prompts. Categories include homonyms, figurative language, safe targets, safe contexts, definitions, real discrimination/nonsense group, nonsense discrimination/real group, historical events, public privacy, and fictional privacy. We report the overall accuracy score based on whether WildGuard (Han et al., 2024) classifies the response as a refusal or compliance.

HarmBench (Mazeika et al., 2024). We evaluated on a subset of harmful prompts which consists of 321 harmful prompts²¹ categorized into Functional and Semantic categories. Functional category includes two types of behavior: Standard behaviors, which are modeled after existing datasets of harmful behaviors such as AdvBench and the TDC 2023 Red Teaming Track dataset, and Copyright behaviors, which test the handling of copyrighted content. The semantic category comprises seven types of harmful behaviors: cybercrime, unauthorized intrusion, chemical/biological weapons or drugs, copyright violations, misinformation/disinformation, harassment/bullying, illegal activities, and general harm. We use WildGuard to evaluate the model’s refusal to assist with harmful prompts.

Do-Anything-Now (Shen et al., 2024b) consists of jailbreak prompts that were created by combining the jailbreak templates from DAN with harmful behaviors from HarmBench, and subsample 300 of them for testing. We report accuracy using the WildGuard classifier.

JailbreakTrigger (Huang et al., 2024b) incorporates prompts based on 13 distinct jailbreak attack methods. In total, the dataset²² consists of 400 examples, split evenly between two categories: "Questions about Bad Behaviors" and "Instructions to generate Toxic Content". This dataset serves to evaluate the effectiveness of LLMs’ defenses and measures the toxicity of responses under jailbreak scenarios. The reported metric is RTA measured by WildGuard.

WildJailbreakTest (Jiang et al., 2024) is an adversarial evaluation set which contains a subset for adversarial benign queries (210 examples) and a subset for adversarial harmful queries (2000 examples). The adversarial benign queries are used to measure models’ exaggerated safety behaviors and the adversarial harmful queries are used to measure models’ safeguards regarding adversarial attacks. We measure RTA using WildGuard for both categories. For benign queries, RTA is expected to be (↓) and for harmful queries, the RTA is expected to be (↑).

WildGuardTest (Han et al., 2024) contains 1725 items for prompt harm, response harm, and response refusal

²⁰Built on Ai2 Safety Tool, forked to focus only on safety evals: <https://github.com/nouhadziri/safety-eval-fork>. Note that while WildGuard and WildJailbreak are popular training datasets, we use generations over the test set prompts as a test of compliance with a response filter such as the WildGuard model.

²¹https://github.com/centerforaisafety/HarmBench/blob/main/data/behavior_datasets/harmbench_behaviors_text_test.csv

²²<https://huggingface.co/datasets/TrustLLM/TrustLLM-dataset>

Benchmarks	Llama 3.1 8B Instruct	Ministral 8B Instruct	Qwen 2.5 7B Instruct	Tülu 3 8B SFT	Tülu 3 8B DPO	Tülu 3 8B
HarmBench	82.8	53.4	84.1	98.4	94.4	94.7
XSTest	92.7	85.6	91.8	90.4	92.4	93.3
WildGuardTest	86.2	68.1	85.0	99.2	98.9	98.5
Jailbreaktrigger	78.8	63.3	71.0	95.8	87.0	85.5
DoAnythingNow	45.0	16.0	61.7	88.3	69.7	62.0
WildjailbreakTest	65.6	50.7	56.2	86.7	81.1	78.8
Overall	75.2	56.2	75.0	93.1	87.2	85.5

Table 25 Breakdown of safety scores by benchmark of TüLU 3 8B models compared with similarly sized open weight models.

Benchmarks	Llama 3.1 70B Instruct	Qwen 2.5 72B Instruct	Hermes 3 Llama 3.1 70B	Nemotron Llama 3.1 70B	Tülu 3 70B SFT	Tülu 3 70B DPO	Tülu 3 70B
HarmBench	80.6	86.3	54.7	84.4	98.8	97.8	97.8
XSTest	87.1	93.6	89.3	92.0	91.1	94.9	92.4
WildGuardTest	81.3	93.1	66.6	84.9	99.1	99.2	98.9
Jailbreaktrigger	71.0	89.8	56.3	60.5	95.3	87.0	86.7
DoAnythingNow	80.0	93.3	26.7	36.3	93.7	69.0	67.7
WildjailbreakTest	59.2	66.0	53.8	56.1	88.6	86.3	86.2
Overall	76.5	87.0	57.9	69.0	94.4	89.0	88.3

Table 26 Breakdown of safety scores by benchmark of TüLU 3 70B models compared with similarly sized open weight models.

classification tasks. 55% are vanilla prompts, and 45% are adversarial. The prompts are collected based on adversarial synthetic data and in-the-wild user-LLM (In-the-wild) interactions. We report RTA using WildGuard.

7.3 Tülu 3 Evaluation Suite - Unseen

For the *unseen* evaluation suite, the task formulations were decided through an independent design process from that of *development* suite. One goal of the unseen suite is to evaluate instruction-tuned models in ways that are closely aligned to realistic usage. Specifically, we follow these general principles:

- Formulate tasks similar to how humans interact with the models. E.g., avoid few-shot examples presented as a dialog, or precise chain-of-thought (CoT) examples for how the model is “supposed” to think.
- Prompt models with clear instructions that set the context, encourage concise reasoning, and specify how the final answer should be formatted.
- Apply reasonable heuristics for answer extraction and comparison to gold answers, to avoid penalizing models that do not exactly follow a syntax implied by the instructions (based on examining outputs from a wide set of existing baseline models).

We first apply these principles to some of the tasks in the development suite using a set of exploratory models (instruction-tuned models predating TüLU 3). On the exploratory models, we find that following the above principles, in particular being more aligned to how human usage, generally does not degrade performance and often times allow most models to perform better on tasks (despite, e.g., removing few-shot examples). We did not update the formulation of the development tasks based on this, but carried the principles over to formulating the unseen tasks. More detailed analysis can be found in Appendix F.2.

The task formulations in the TüLU 3 *unseen* suite are as follows. For all benchmarks which include multiple

Model	Llama 3.1 prompt	Ours
Gemma 2 9B Inst	51.6	52.6
Gemma 2 9B Inst-SimPO	52.6	51.8
Llama 3.1 8B Inst	49.2	48.7
Llama 3.2 3B Inst	39.1	39.7
Ministral 2410 8B Inst	43.8	44.3
OLMo 0724 7B Inst	26.1	22.9
OLMoE 0924 1B 7B Inst	20.7	20.3
Qwen 2.5 7B Inst	56.2	54.2
Tulu 2 DPO 7B	25.4	22.2

Table 27 Comparing evaluating on MMLU-Pro using our 0-shot CoT prompt and the 5-shot prompt used in Llama3.1 evaluations, using macro average over tasks in both cases (the Llama3.1 evaluation used micro average).

Model	Llama 3.1 prompt	Ours
Gemma 2 9B Inst	35.7	35.5
Gemma 2 9B Inst-SimPO	35.0	35.7
Llama 3.1 8B Inst	29.5	29.5
Llama 3.1 70B Inst	46.2	44.0
Llama 3.2 3B Inst	33.5	27.7
Ministral 2410 8B Inst	31.0	31.5
OLMo 0724 7B Inst	27.2	27.9
OLMoE 0924 1B 7B Inst	24.6	24.8
Qwen 2.5 7B Inst	32.1	36.8
Tulu 2 DPO 7B	28.4	27.5

Table 28 Comparing evaluating on GPQA using our 0-shot CoT prompt and the prompt in Llama3.1.

sub tasks, we compute the average over the sub tasks (i.e., the “macro” average) to be consistent.

AGIEval English (Zhong et al., 2024) includes the English language subset of the AGIEval benchmark, specifically these multiple-choice tasks: *aqua-rat*, *logiqa-en*, *lsat-ar*, *lsat-lr*, *lsat-rc*, *sat-en*, *sat-math*, and *gaokao-english*. We do not include the *sat-en-without-passage* task as we find these questions are typically severely underspecified without access to the passage. We formulate the task using the a simple “zero-shot CoT” prompt which encourages concise reasoning ending with a clearly stated answer choice. In Appendix F.2, Figure 44, we provide the 0-shot reasoning prompt, used for the multiple-choice tasks in the *unseen* evaluation suite, including AGIEval English. The model’s answer choice is extracted by first matching to the requested format, with fallback patterns if the format was not followed precisely. Specifically, we first look for the exact phrase indicated in the prompt (“Therefore, the answer is [ANSWER]”) and grab the last such match. If that fails, we look for a sequence of softer variants, like “answer is [ANSWER]” or “answer: [ANSWER]” before falling back to the last letter in parenthesis found, and if that fails, the last stand-alone capital letter.

MMLU-Pro (Wang et al., 2024b) is a 10-way multiple-choice extended version of the MMLU dataset. We use essentially the same prompt and answer extraction as used for our AGIEval setup, just adjusting for the number of answer choices. We generally find our formulation to be as effective as the traditional 5-shot CoT prompt (e.g., used in Llama 3.1 evaluations (Dubey et al., 2024)) despite being much shorter, more realistic, and easier for users to create (see Table 27 for a comparison across exploratory models).

GPQA (Rein et al., 2023) is a set of very challenging multiple-choice questions written by domain experts in biology, physics, and chemistry. We use the same zero-shot prompt and answer extraction as for AGIEval.

Model	base-adapted, in context examples	CoT prompt
Gemma 2 9B Inst	18.0	45.9
Gemma 2 9B Inst-SimPO	19.3	45.3
Llama 3.1 8B Inst	20.0	39.4
Llama 3.2 1B Inst	11.6	13.1
Llama 3.2 3B Inst	19.2	32.6
Minstral 2410 8B Inst	18.8	36.7
OLMo 0724 7B Inst	3.2	5.8
OLMoE 0924 1B 7B Inst	9.0	4.2
Qwen 2.5 7B Inst	21.2	54.7
Tulu 2 DPO 7B	9.6	6.0
Llama 3.1 Tulu 2 8B	21.7	13.6
Llama 3.1 Tulu 2 DPO 8B	18.6	14.5

Table 29 Comparing evaluating on DeepMind Mathematics using different prompts designed evaluation practices for instruction-tuned models, with minor variants. **Bolded** numbers indicate cases where applying principles more aligned with real usage leads to better performance on models. In our evaluation suite, the chain of thought (CoT) prompt is referred to as `chat-v3`.

This is similar to the approach used in Llama 3.1 evaluations, but with less prescription on how the reasoning should be structured. Figure 28 shows that across exploratory models, that our approach leads to comparable scores.

Deepmind Mathematics (Saxton et al., 2019) is a dataset of 56 categories of math questions, testing mathematical and algebraic reasoning skills. We devised a “zero-shot CoT” prompt that sets the context of the task, explains the format expected for the answers (e.g., “x**2” for powers), and for each category included three example answers to specify the answer format. The prompt used and answer extraction heuristics were refined by examining example outputs from our exploratory models. As seen from Table 29, applying evaluation principles aligned with real usage (“chat” versions), often performs better than using a setup involving presenting few-shot examples adapted from base model evaluations (“base-adapted”). We provide the zero-shot reasoning prompt for the Deepmind Mathematics task in Appendix F.2, Figure 45. Aligned with the instructions for formatting in the prompt, when extracting the answer, we first look for an answer in the format “Therefore, the final answer is [answer]. I hope it is correct.” We look for this in a case-insensitive way, process the [answer] by stripping away any trailing period, and known math delimiters surrounding the answer (e.g., “\$”) based on outputs from our exploratory models. For answer comparison, we first compare the raw strings of the processed generated answer and gold answer to check if they are equal. In addition, we parse both using the SymPy (Meurer et al., 2017) package and compare the equivalence of the parsed outputs to check their mathematical equivalent. We also consider the generated answer to be correct if it is equal to the parsed gold answer after this parsing.

BigCodeBench (Zhuo et al., 2024) is a set of coding challenges. We focus on the “hard subset” of 148 (out of the total 1140) instances, using the “instruct” formulation of each task and the “calibrated” score. We follow the setup used for the original leaderboard for our implementation.

7.3.1 New Evaluation: IFEval-OOD

In order to test precise instruction following abilities of LLMs and whether they are able to follow constraints that go beyond the 25 constraints included in IFEval (Zhou et al., 2023), we developed IFEval-OOD (IFEval Out-of-Distribution). IFEval-OOD consists of 52 constraints across six broad categories such as the examples in Table 30. A complete list of the constraints can be found in Appendix F.3. A portion of the constraints were sourced by asking a group of researchers for constraint ideas and others were written by authors of this paper. One of the six categories (“custom”) comprises manually written verifiable prompts to cover specific

Instruction Group	Instruction	Description
count	person_names	Mention at least {N} different person names in the response.
format	emoji	Please use an emoji at the end of every sentence.
ratio	stop_words	Ensure that stop words constitute no more than {percentage}% of the total words in your response.
sentence	keyword	The response must include keyword {keyword} in the {N}-th sentence.
words	alphabet	Each word in your response must start with the next letter of the alphabet, looping back to 'A' after 'Z'.
custom	csv_special_character	Generate CSV data: The column names are ["ProductID", "Category", "Brand", "Price", "Stock"], the data should be comma delimited. Please generate 14 rows. Add one field which contains a special character and enclose it in double quotes.

Table 30 Examples of IFEval out-of-distribution constraints. Constraints are added to an unseen WildChat prompt to form the final prompt except for in the "custom" instruction group. A complete list of constraints is provided in Appendix F.3.

sub-skills, such as CSV generation. For the remaining five categories, the verifiable constraint was combined with 10 unseen prompts from WildChat. To select the final prompts, these constrained prompts were human annotated for quality and compatibility with the constraint (e.g. a prompt to paraphrase a one-sentence reference text that doesn't contain any names to begin with would not be compatible with a constraint to mention at least 23 different person names in the response). To ensure constraint coverage, additional unseen WildChat prompts were manually paired with constraints that did not have at least five selected prompts. Our evaluation dataset emphasizes broad coverage of constraint types to differentiate constraint-following abilities from overfitting on the existing IFEval constraints.

7.3.2 New Evaluation: HREF

We constructed an automatic evaluation of instructability of language models, called *Human Reference-guided Evaluation of instruction Following (HREF)*. HREF focuses on 11 instruction following tasks that language models are typically trained on, namely, *Brainstorming*, *Open QA*, *Closed QA*, *Extraction*, *Generation*, *Rewriting*, *Summarization*, *Classification*, *Numerical Reasoning*, *Multi-document Synthesis*, and *Fact Checking*. We obtained high quality human-written prompts and responses in these categories from professional instruction-tuning data creators, and developed a reliable automatic evaluation procedure. Following AlpacaFarm (Dubois et al., 2023), we use win-rate against a fixed baseline model as the evaluation metric. However, since we have a larger number of tasks in our benchmark than AlpacaFarm, and also because we have access to human-written references, we hypothesized that the setup used for win-rate computation may not be directly applicable to our new evaluation. So we experimented with various win-rate computation methods, and their details along multiple dimensions:

- **Choice of LM Judge.** We tried GPT-4, GPT-4 turbo, and Llama 3.1-Instruct models at 7B and 70B sizes as LM judges.
- **LM-as-a-Judge vs. Direct Comparison with Human References.** Using a large language model e.g., GPT-4 Turbo for AlpacaFarm, may not be appropriate for tasks where the responses are expected to be short and factual. We compared this setup with directly comparing the similarity of the responses from the target and the baseline models with the human-written references, according to a simpler embedding model (RoBERTa-Large (Liu, 2019)).
- **Use of Human-Written References When Evaluating with LM-as-a-Judge.** Since we have access to human-written references, we experimented with including those as additional context while prompting the LM

Skill	8B SFT		8B DPO		8B Final		70B SFT		70B DPO		70B Final	
	Dev.	Uns.	Dev.	Uns.	Dev.	Uns.	Dev.	Uns.	Dev.	Uns.	Dev.	Uns.
Avg.	64.9	29.9	68.3	31.9	68.8	32.4	78.1	41.0	80.5	44.4	80.7	44.4
Knowledge Recall (MMLU → GPQA)	65.9	31.9	68.7	31.2	68.2	35.7	78.9	43.3	83.3	48.0	83.1	48.0
Reasoning (BBH → AGIEval)	67.9	56.2	65.8	61.8	66.0	59.3	82.7	73.2	81.8	75.0	82.0	75.0
Math (MATH → DM Mathematics)	31.5	32.3	42.0	33.0	43.7	35.4	53.7	49.7	62.3	49.4	63.0	49.8
Coding (HumanEval → BigCodeBench)	86.2	11.5	83.9	9.5	83.9	7.4	92.9	12.2	92.4	23.0	92.4	21.6
Inst. Following (IFEval → IFEval-OOD)	72.8	17.6	81.1	23.9	82.4	24.3	82.1	26.8	82.6	26.4	83.2	27.8

Table 31 Comparison of the performance of TULU 3 SFT and DPO checkpoints and the final models on development and unseen tasks corresponding to each of the core skills. Table shows that our pipeline generalizes well to unseen evaluations.

judge.

To make decisions about the evaluation setup, we collected human judgments comparing responses from a diverse set of 16 models, when prompted using the instructions taken from a subset of our evaluation dataset. We collected four human judgments per each model response pair, and compared the agreement of each evaluation setup with majority human judgments. We repeated this procedure for each task category to identify the best setup for that category.

Summary of the Final Evaluation Setup. We used Llama 3.1 70B Instruct as our LM judge since its agreement with human judges was the highest of the pool of models we considered. Our baseline model we used is Llama 3.1 405B Instruct. We found that using LM as a judge results in higher human agreement (compared to the embedding-based method) in all subtasks except *Open QA* and *Fact Checking*. We use embedding similarity with human-written references as the way of computing win-rates in these two subtasks. Of the remaining 9 subtasks, evaluation in *Brainstorming* and *Summarization* did not benefit from using human-written references while prompting the LM judge, and hence prompt LM judges with human references only for the remaining 7 subtasks. On the subset for which we collected human judgments, our composite evaluation procedure resulted in an agreement of 69.4% with humans, which is comparable to the inter-human agreement of 67%.

7.4 Evaluating the Development Process Using the Unseen Suite

To evaluate how much and in what ways we over-fit to our development evaluations, we measure the performance of our models on unseen tasks that still correspond to the same set of core skills. We evaluate the checkpoints corresponding to various design decisions we made in the development process on the unseen suite to check whether our decisions overfit to the development evaluations, and summarize our findings in Section 7.4.1. In Section 7.4.2, we show a comparison between our final checkpoints and comparable public models.

7.4.1 Evaluating the design decisions

Training pipeline. Table 31 shows the performance of the SFT, DPO, and the final checkpoints of TULU 3 at 8B and 70B sizes on one development and one unseen evaluation for each core skill. We see that our pipeline generalizes well to unseen evaluations, with the final checkpoints obtaining the best average performance on both the development and unseen evaluations. For Reasoning and Coding, where the SFT checkpoints have the best performance on development evaluations, the subsequent training stages still improve model performance on harder unseen evaluations.

Data mixing for SFT. To assess whether the data mixing choices we made for SFT generalize to unseen evaluations, we evaluate the performance of various data-ablated SFT models with the final SFT checkpoint, as shown in Table 32. For each core skill, we compare the model performance trends on a development and an unseen evaluation. We see that the data choices generalize on average, as indicated by the best average performances on both development and unseen evaluations by the final SFT checkpoint. In individual skills, we see that our choices overfit to the development evaluations in Precise Instruction Following, and to some extent in Knowledge Recall and Reasoning.

Model	Dev. Avg.	Uns. Avg.	Know. Recall		Reasoning		Math		Coding		Inst. Follow.	
			MMLU	GPQA	BBH	AGIE	MATH	DMM	CHE	BCB	IFE	IFEO
TÜLU 3 8B SFT	64.1	29.9	62.1	31.9	67.9	56.2	31.5	32.3	86.2	11.5	72.8	17.6
w/o WildChat	62.8	28.8	61.0	31.5	65.6	53.1	31.8	31.2	85.3	7.4	70.1	20.8
w/o Safety	63.7	29.7	62.0	31.9	68.3	55.6	32.6	32.6	84.5	10.8	71.0	17.6
w/o Persona Data	59.8	29.4	62.4	29.5	68.3	56.9	30.1	31.8	84.5	10.8	53.6	18.0
w/o Math Data	62.2	27.4	62.2	32.6	68.9	54.1	23.5	23.3	86.0	8.8	70.6	18.3

Table 32 Comparison of the performance of data-source ablated SFT models with that of the final TÜLU 3 8B SFT checkpoint. Numbers in bold reflect the best performance per column. AGIE is AGIEval, DMM is Deepmind Mathematics, CHE is Codex HumanEval, BCB is BigCodeBench, IFE is IFEval and IFEO is IFEval-ODD.

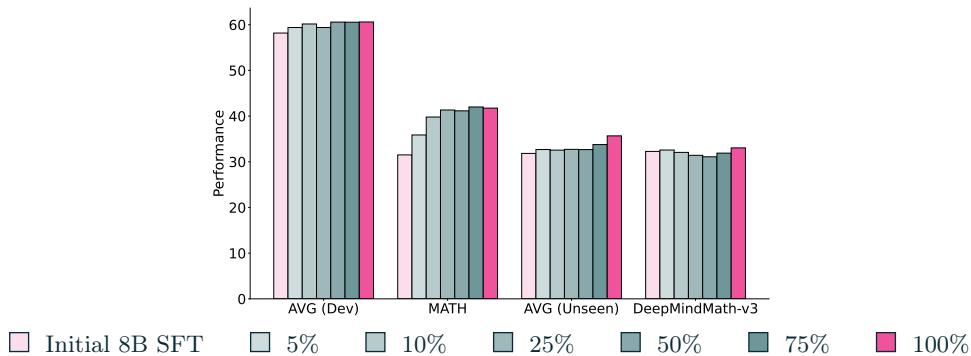


Figure 24 Effect of scaling the size of the preference dataset, specifically the number of unique prompts, on downstream DPO model performance in development and unseen evaluations. AVG are the averages of all the tasks in development and unseen suites. Also shown are the trends in MATH and Deepmind Mathematics.

Data scaling for preference tuning. We show the effect of scaling DPO data on development and unseen evaluations in Figure 24. We see that the scaling trends generalize on average to unseen evaluations. Of note is the trend we observed in the development and unseen Math evaluations, indicating that our development process overfit to MATH to some extent. We hypothesize this is mainly due to formatting differences between MATH and Deepmind Math. The former often requires solutions and answers to be output in LaTeX format, while the latter does not. We found that our trained models have the tendency to format the chain-of-thought reasoning and the final answers in LaTeX even for the questions in the Deepmind Math dataset where this is not required. This often interfered with the intermediate reasoning, and also made our answer extraction logic fail.

7.4.2 Comparison with public models

Table 33 shows a comparison between TÜLU 3 models, Llama 3.1 Instruct models, and Hermes 3 Llama 3.1 models at 8B and 70B sizes, and Table 48 shows a subtask-level breakdown of the performance of these models on HREF. It is important to note that while all these evaluations are unseen for the TÜLU 3 models, we do not know if GPQA, MMLU-Pro, AGIEval, DeepMind Math, and BigCodeBench were used for developing the two other models. We summarize below our key takeaways from this comparison and from our qualitative analysis of the outputs of TÜLU 3 models on these datasets:

Tülu 3 generalizes well to unseen evaluations. In almost all the evaluations, TÜLU 3’s performance is generally comparable to that of the two other models we evaluate, often falling between the performance numbers of the two models. This suggests that our recipe of choosing representative evaluations for each core-skill and curating training datasets targeting those evaluations can lead to models that generalize well to other tasks that require the same skills.

Models generally overfit to IFEval. We find that there is a significant difference between performance on IFEval and IFEval-ODD of all the models, even though we created the latter to be structured very similar to the

Skill	Benchmark _(eval)	Llama	Hermes	Tülu 3 8B	Llama	Hermes	Tülu 3 70B
		3.1 8B Instruct	3 Llama 3.1 8B		3.1 70B Instruct	3 Llama 3.1 70B	
	Avg.	36.4	30.7	34.2	51.3	43.1	47.2
Knowledge Recall	GPQA _(0 shot, CoT)	28.8	32.8	35.7	43.8	42.6	48.0
	MMLU Pro _(0 shot, CoT)	49.1	40.9	44.3	68.3	60.3	65.8
Reasoning	AGIEval English _(0 shot, CoT)	64.2	58.1	59.3	77.8	73.3	75.0
Math	DeepMind Math _(0 shot, CoT)	39.3	28.3	35.4	62.4	50.0	49.8
Coding	BigCodeBench-Hard _(Pass@10)	15.5	9.5	7.4	26.4	14.2	21.6
Instruct Following	IFEval OOD _(Prompt loose)	26.1	19.4	24.3	34.5	24.6	27.8
	HREF _(Winrate)	38.5	26.2	32.7	45.6	36.8	42.3

Table 33 Evaluation of a selection of open-weight and TüLU 3 models on our unseen evaluation suite. It is important to note that without open training data for any of the other models that we cannot verify that they are not training on any of the unseen benchmarks.

original dataset, just with a disjoint set of constraints. We observe that instruction following with verifiable constraints is a challenging skill for models to learn effectively, and hypothesize that those models that do well on IFEval are likely overfitting to the specific set of constraints included in the dataset, making it hard for them to generalize to new constraints.

Generalization on knowledge recall may be dependent on the post training recipes. As one might expect, the performance of models on MMLU and MMLU-Pro is correlated. We see that their performance on GPQA shows a different trend — all three models we compare here are post-trained from the same base model, suggesting that the post-training recipe may affect the generalization in knowledge-recall.

Instruction following performance varies across categories. We observe that the relative performance of TüLU 3 models on AlpacaEval is different from that on HREF. This may be explained by the fact that instruction following is a highly diverse task, and the distributions of HREF and AlpacaEval may differ, with some categories of instructions not necessarily transferring well to others, leading to the shift in relative performance. We do note that TüLU 3 70B outperforms Llama 3.1 70B Instruct on 5 out of 11 subtasks, as seen in Table 48 in the Appendix, which shows a breakdown of model performance across subtasks in HREF. Future work will explore how different behaviors of instruction following can be measured with more diverse instruction following evaluations.

8 Discussions

8.1 Scaling Tülu 3 Recipe to Llama 3.1 405B

Here, we demonstrate the scalability and effectiveness of our post-training recipe applied at 405B parameter scale. Scaling to this size required several engineering efforts and posed a number of challenges:

- **Compute Requirements:** Training TüLU 3 405B demanded 32 nodes (256 GPUs) running in parallel. While most of our codebase scaled well, we occasionally encountered NCCL timeout and synchronization issues that required meticulous monitoring and intervention (especially with RL training). Using more GPUs increased the chances of encountering hardware failures, necessitating semi-frequent run restarts.
- **RLVR Training:** For inference, we deployed the model using vLLM with 16-way tensor parallelism, while utilizing the remaining 240 GPUs for training. After each iteration of RLVR update, the weights are synchronized to the vLLM engine using NCCL broadcast. Inference typically takes ~550 seconds, weight transfer takes ~25 seconds, and training takes ~1,500 seconds. To reduce computational cost during the RLVR stage, we utilized an 8B value model. Future works can benefit from exploring larger value models or alternate value model-free RL algorithms such as GRPO (Shao et al., 2024).
- **Hyperparameter Tuning Challenges:** Given the computational costs, hyperparameter tuning was limited.

Hyperparameter	405B SFT	405B DPO
Learning Rate	2×10^{-6}	2×10^{-7}
Learning Rate Schedule	Linear	Linear
Batch Size (effective)	256	256
Max Token Length	4,096	2,048
KL penalty coefficient β	-	5
Warm up ratio	0.03	0.1
Number of Epochs	2	1

Table 34 Hyperparameters for training TüLU 3 405B. We use a larger batch size due to the increased number of GPUs, and lower the SFT learning rate.

Hyperparameters	405B RLVR
Learning Rate	1×10^{-7}
Discount Factor γ	1.0
General Advantage Estimation λ	0.95
Mini-batches N_{mb}	1
PPO’s Clipping Coefficient ε	0.2
Value Function Coefficient c_1	0.1
Gradient Norm Threshold	1.0
Learning Rate Schedule	Linear
Generation Temperature	1.0
Batch Size (effective)	1,856
Max Token Length	2,048
Max Prompt Token Length	2,048
Penalty Reward Value for Responses without an EOS Token	-10.0
PPO Update Iterations K	1
Response Length	1,024
Total Episodes	300,000
KL penalty coefficient (β)	0.05
Warm up ratio (ω)	0.0

Table 35 The hyperparameters of PPO used for optimizing against a general RM.

Following prior Tülu and Llama work, we opted to lower the LR for larger models, training with a “lighter touch”.

Our training recipe for the 405B model followed very similarly to that of the 8B and 70B models,²³ but with a different training dataset for RLVR. Given the model’s saturation of GSM8K from SFT and DPO training alone, we removed the GSM8K data, and we additionally found that the IFEval data did not help much in initial RLVR runs. As such, for TüLU 3 405B RLVR we only used the MATH train set. Surprisingly, we found that even with as few as 25 RLVR steps, MATH performance improved by over 5 points, and continued to increase with more training.

With the challenges of scaling the asynchronous RL infrastructure, we only trained for 75 steps, fewer than our smaller models. The RL reward, KL divergence, and response length per training batch are shown in Fig. 25. The training hyperparameters for the SFT and DPO stages are shown in Table 34. The hyperparameters for RL with verifiable rewards at this scale are shown in Table 35.

In Table 4, we compare TüLU 3 405B with prior state-of-the-art models finetuned from Llama 3.1 405B as well as DeepSeek-V3 (DeepSeek-AI et al., 2024) and GPT-4o. Generally, TüLU 3 405B results are improved compared to TüLU 3 70B. TüLU 3 405B achieves competitive or superior performance to both Deepseek v3 and GPT-4o, while also surpassing prior open-weight post-trained models of the same size including Llama 3.1 405B Instruct and Nous Hermes 3 405B (Teknium et al., 2024) on many standard benchmarks. We note that we ended RLVR training early due to compute constraints, and further training may further improve performance. In particular, we did not observe that MATH performance had saturated during training and testing (see Figure 25).

²³DPO Preference Mixture for 405B: [allenai/llama-3.1-tulu-3-405b-preference-mixture](https://huggingface.co/allenai/llama-3.1-tulu-3-405b-preference-mixture)

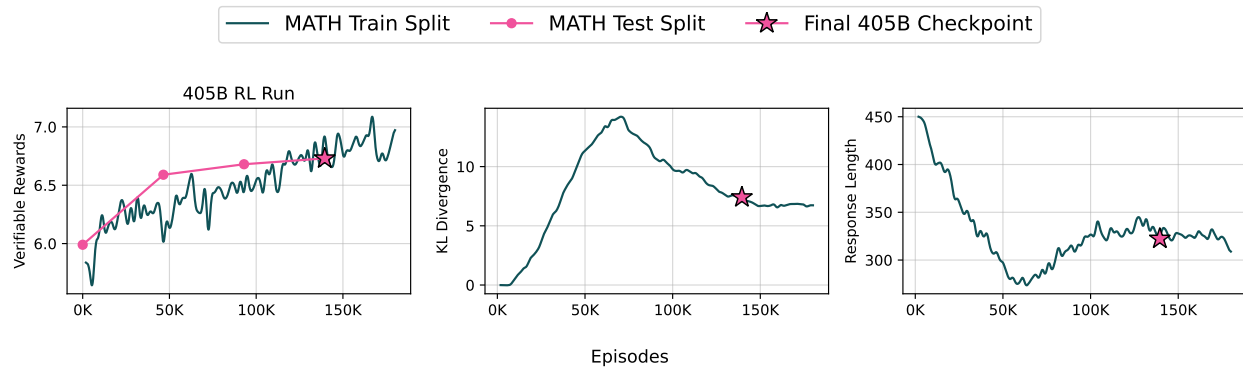


Figure 25 Rewards on MATH, KL divergence, and the average response length for the final 405B training run. We mark the point with the final checkpoint with a star. We note that this was the last checkpoint saved – we intended to train longer but hit compute constraints. Note that technically the metrics in MATH test split is accuracy, but here we multiply the accuracy by 10 to convert to verifiable rewards.

8.2 Insights from the Unfruitful

In this section, we discuss a number of methods and approaches we considered for TŪLU 3 but did not ultimately make it into our final recipe for a variety of reasons.

Online DPO. Standard DPO methods use preference datasets that are usually collected ahead of time, often from a distinct language model, and are thus considered as offline. In other words, with DPO, the policy cannot obtain feedback over its own generations during training. This is in contrast to online methods like PPO where the RM provides online feedback to generations from the policy π_θ being trained. To mitigate the distributional shift issue, recent works proposed Online DPO (Guo et al., 2024) following a three-step process: (1) sample 2 responses to a prompt from the current policy; (2) obtain online feedback over the response pair to create pairwise data, and (3) use this pairwise data to update the policy π_θ via standard DPO loss. While the original paper proposed using online AI feedback for step 2, to better scale our experiments, we obtain feedback from a trained reward model.

We tried online DPO both to enhance general, and target capabilities, i.e., mathematical reasoning. For general capabilities, we train an RM for 1 epoch using the 82K preference data points from the Skywork.²⁴ For targeting mathematical reasoning, we continue training the same RM on our synthetic on-policy math-specific preference data (described in §5.2.1). Training online DPO on top of one of our TŪLU 3 DPO checkpoints for a total of 200K episodes on math problems (prompts are taken from the same RM training data), resulted in no or little improvement on GSM8K and degradation on MATH performance (we experimented with various sampling temperatures and KL penalty coefficients). We did not further explore this approach extensively, as our initial results suggested limited gains in both general and targeted domains. Future work could investigate alternative training strategies, such as different sampling methods, or finetuning RM architectures, to better align the optimization process with the desired capabilities.

Rejection Sampling. Rejection sampling for large language models is an increasingly popular method for improving post-training performance of frontier language models (Dubey et al., 2024; Adler et al., 2024; Dong et al., 2023). Using an initial SFT and preference data mix to train an initial model, that model is used to then generate n responses to each SFT prompt. These n responses, in addition to the original response, are then ranked using a reward model or an LLM as a judge, and the best response is kept. The other responses can then be used to create chosen/rejected pairs for preference optimization. The full post-training pipeline is then run on these datasets, and the process repeats until performance converges.

We tried rejection sampling, but found that for our setup the performance gains were minimal for the amount of compute required, and thus we leave a deeper exploration for future work. Qualitatively, we found that strong judges are vital, and publicly available models can struggle to choose the best response out of the candidates. We also found that including the original response as a choice for the judge (in other words,

²⁴<https://huggingface.co/datasets/Skywork/Skywork-Reward-Preference-80K-v0.1>

choosing the best response out of the n generations in addition to the original response) performed much better than only choosing from the newly generated responses.

8.3 Future Work

While we aimed to cover a broad set of skills for TÛLU 3, due to various limitations, we left some skills for future iterations of TÛLU. We discuss some of these skills below:

Long Context and Multi-turn. Currently, the data collected for TÛLU 3 is relatively short and does not contain long multi-turn data (the average number of turns in our mixture is 2.4 turns and majority of samples are under 2,048 tokens in length). However, long-context has been popular area of focus in recent work (Pawar et al., 2024), as improving the context window of LMs enables new use-cases (Gemini Team, 2024) and more in-context examples, potentially improving performance (Agarwal et al., 2024). Relatedly, improving multi-turn capabilities can better improve end-user experience, with a non-trivial number of real-world user conversations with LMs going over 2 turns (Zhao et al., 2024). We hope to address both skills in the future with dedicated training and evaluations.

Multilinguality. We specifically focus on English data and evaluations for TÛLU 3 (although we do include the multilingual Aya (Üstün et al., 2024) dataset due to its high quality). However, this neglects the myriad languages apart from English spoken around the world, speakers of which may benefit from or desire LMs that can process their languages. Future work may examine the current multilingual abilities of TÛLU 3 and how to further improve them. We also note that multilingual post-training can make use of different techniques to monolingual post-training – for example, cross-lingual alignment (Wu et al., 2024) or careful data balancing strategies (Li et al., 2024b). This makes multilingual post-training an interesting and impactful area for future work.

Tool Use and Agents. While we evaluate TÛLU 3 on its own, LMs are being increasingly deployed as parts of larger systems, in which they have access to tools (Qu et al., 2024) or are themselves part of a larger ‘agent’ framework.²⁵ Furthermore, training models to use tools is a natural way to dramatically improve their reasoning and mathematical skills (Gou et al., 2024), rather than trying to accomplish everything ‘in the weights.’ Future work involves training or evaluating TÛLU 3 for tool-use either on its own or as part of a larger framework.

9 Related Work

9.1 The Evolution of Post-training Recipes

Modern “post-training” has its roots in multi-task language model training, in particular *instruction tuning* (Mishra et al., 2022; Wei et al., 2022a; Sanh et al., 2022; Wang et al., 2022b; Longpre et al., 2023), in which language models are trained on samples including task instructions and their corresponding responses, allowing the models to generalize ‘zero-shot’ to new tasks.²⁶ Early instruction-tuning datasets tended to focus on more traditional NLP tasks (e.g., natural language inference) rather than more generic tasks that downstream users might perform (Wang et al., 2022a). With the rise of ChatGPT and chat-based LMs (Claude, Gemini, etc), post-training techniques evolved beyond instruction tuning to include preference tuning stages, with models undergoing both instruction tuning and then preference finetuning (PreFT) or “RLHF” (Ouyang et al., 2022).

Early work in RLHF originated from experiments on Deep RL for control (Christiano et al., 2017; Ibarz et al., 2018; Leike et al., 2018) and typically involved first learning a reward model from human preferences, and then optimizing a language model via an RL framework using the learnt reward (Stiennon et al., 2020; Nakano et al., 2021; Askell et al., 2021; Ouyang et al., 2022). Recently, approaches that allow directly training a language model on such preferences have been developed (Rafailov et al., 2024; Zhao et al., 2023), reducing the complexity of incorporating PreFT into training. While early approaches to PreFT were extremely human-centric, using tens or hundreds of thousands of human-written instructions and human preference

²⁵For example, the OpenHands platform (Wang et al., 2024a).

²⁶The term “post-training” substantially predates modern chat language models (Moreau and Audiffren, 2016; Xu et al., 2019), but was recently popularized.

labels, more recent work uses mixtures of human and synthetically generated preference data, along with multiple rounds of training and varied training algorithms (Touvron et al., 2023; Dubey et al., 2024; Gunter et al., 2024).

During the evolution of RLHF primarily in closed laboratories, open recipes for post-training have lagged somewhat behind. Initial attempts at building ‘open post-training recipes’ focused on the instruction-tuning stage (Taori et al., 2023; Conover et al., 2023), finetuning openly released language models on synthetically generated or human made datasets. While combining these datasets could yield strong performance (Wang et al., 2023), incorporating a PreFT stage proved important for bridging the gap with closed models based on human evaluations (Iverson et al., 2023). Today, most popular adapted models with *open recipes* for PreFT use DPO (or one of its variants), and AI feedback data including T_{ULU} 2 (Iverson et al., 2023), Zephyr- β (Tunstall et al., 2023), and Starling (Zhu et al., 2023). However, many of these models are outdated relative to closed post-training recipes both in terms of data and in performance: no open-recipe models exist in the top 50 of LMSYS’s ChatBotArena (as of November 20th, 2024) have released their post-training data (Chiang et al., 2024). Most of these open recipes use relatively little data and few rounds of training compared to closed post-training setups, which can involve multiple rounds of training with varied objectives, and millions of datapoints (Touvron et al., 2023; Dubey et al., 2024). For instance, Llama 3.1 trained on generated outputs from the previous model for multiple rounds with extensive human feedback data, and used strong models to write synthetic instructions (Dubey et al., 2024). Other recent developments includes rejection sampling for synthetic data and advanced reward modeling for step-wise assistant responses (Lightman et al., 2023).

Whilst we do not entirely reach the size of these closed recipes, in this work we hope to push the state of open post-training recipes forward by building a recipe that matches or beats strong closed recipes, and releasing all related artefacts (code, models, data, etc) for further scientific study and use. Our pipeline is significantly larger and more complex than prior work, comprising of almost a million instruction tuning samples, hundreds of thousands of preference pairs, and a novel online RL training phase.

9.2 Training on Verifiable Rewards

The RLVR approach proposed in this work relates to a variety of recent work on improving LM reasoning through RL-related techniques. Closely related is the self-taught reasoner (STaR) line of work (Zelikman et al., 2022, 2024) and TRICE (Hoffman et al., 2023), both of which examine using existing ground-truth answers as signals to generate better model rationales (or chains-of-thought). STaR can be seen as an approximation of a policy gradient algorithm, with Quiet-STaR extending the approach of training the model to use additional generations to improve generic language modelling (‘thinking before speaking’). TRICE (Hoffman et al., 2023) also aims to improve the likelihood of correct answers by training over multiple reasoning traces, using a custom MCMC-based EM algorithm. More recently, VinePPO (Kazemnejad et al., 2024) uses binary rewards from GSM8k and MATH correctness to test a novel PPO-based algorithm, and other recent work has explored using code feedback as a signal for training (Gehring et al., 2024; Xu et al., 2024b). In contrast, our proposed approach, RLVR simply uses an existing RL framework (PPO) for training, and runs entirely online with binary rewards (compared to the iterative approach of STaR or the log-likelihood rewards of Quiet-STaR). Additionally, we expand beyond the math domain, also finding that this approach can yield improvements in precise instruction following. Finally, we also carefully ablate a number of core components of RLVR, including value model initialization and using a general reward model with verifiable rewards. We hope to further develop and expand this technique in future work.

10 Conclusion

We introduce T_{ULU} 3, a family of fully open state-of-the-art language models, featuring a modern post-training framework with fully open-source data T_{ULU} 3 DATA, evaluation T_{ULU} 3 EVAL, training code T_{ULU} 3 CODE and development recipes T_{ULU} 3 RECIPE. We release final models trained on Llama 3.1 base versions, with intermediate checkpoints, training data, training code, and evaluation code.

T_{ULU} 3 bridges the gap between open and closed post-training methods, marking a new milestone in open post-training research. With the resources provided, others can build on open base models, finetune them for high performance across diverse tasks. This paves the way for advancing post-training research within

multi-objective, and multi-stage training frameworks.

Author Contributions

A successful team project like TÛLU 3 would not be possible without the fluid contributions of many teammates across formal team boundaries. As not all of these can be captured, we indicate each authors’ primary contributing role in this project. Authors are listed in alphabetical order:

- For SFT model development, including training and data curation: Faeze Brahman, Shengyi Huang, Hamish Ivison, Nathan Lambert, Jacob Morrison, Yizhong Wang, and Chris Wilhelm.
- For preference-tuned model development, including training and data curation: Faeze Brahman, Shengyi Huang, Hamish Ivison, Nathan Lambert, Lester James V. Miranda, Valentina Pyatkin, Chris Wilhelm.
- For reinforcement-learning model development, including training and data curation: Shengyi Huang, Nathan Lambert, Hamish Ivison, Valentina Pyatkin, Faeze Brahman.
- For evaluation tooling support, decontaminating training datasets, and evaluating peer models in the ecosystem throughout: Pradeep Dasigi, Nouha Dziri, Victoria Graf, Shengyi Huang, Jena D. Hwang, Hamish Ivison, Ronan Le Bras, Alisa Liu, Xinxu Lyu, Saumya Malik, Valentina Pyatkin, Luca Soldaini, Oyvind Tafjord, Jiangjiang Yang.
- For management of communications, legal, and other release processes: Faeze Brahman, Pradeep Dasigi, Hannaneh Hajishirzi, Nathan Lambert, Luca Soldaini.
- For mentorship and advising: Pradeep Dasigi, Hannaneh Hajishirzi, Nathan Lambert, Valentina Pyatkin, Noah A. Smith, Luca Soldaini, Yizhong Wang.

Authorship for this work was determined by those making direct contributions to the TÛLU models, related artifacts, and their release. Core contributors are recognized for their sustained, significant contributions critical to the success of the TÛLU 3 project.

Acknowledgments

We thank John Schulman for extremely useful advice. We acknowledge the National Artificial Intelligence Research Resource (NAIRR) Pilot and Microsoft Azure for contributing to the results in this work. We thank Niklas Muennighoff for helping with some experimentation on OLMoE. We thank countless members of Ai2 and UW NLP communities for useful feedback throughout this project. Research supported with Cloud TPUs from Google’s TPU Research Cloud (TRC). We thank the vLLM team (Kaichao You, Simon Mo, Woosuk Kwon, and Zhuohan Li) for their invaluable support in debugging NCCL weight transfer issues for RLVR. We thank Huy Tran, Jesse Dodge, Jiacheng Liu, Sruthi Sreeram, Taylor Blanton, Aaron Sarnat, Arnavi Chheda, Byron Bischoff, Chris Newell, Michael Schmitz, Sam Skjonsberg, Eric Marsh, Karen Farley, and YenSung Chen for building the Ai2 Playground for model demos. We also thank these others at Ai2 for many indirect contributions to the project: Kyle Lo, Taira Anderson, Jen Dumas, Crystal Nam, Sophie Lebrecht, Brooke Vlahos, Chris Wilhelm, Jenna James, Alex Buraczynski, Will Smith, Caitlin Wittlif, Carissa Schoenick, and Ali Farhadi.

References

- B. Adler, N. Agarwal, A. Aithal, D. H. Anh, P. Bhattacharya, A. Brundyn, J. Casper, B. Catanzaro, S. Clay, J. Cohen, et al. Nemotron-4 340b technical report. *arXiv preprint arXiv:2406.11704*, 2024.
- R. Agarwal, A. Singh, L. M. Zhang, B. Bohnet, L. Rosias, S. C. Chan, B. Zhang, A. Anand, Z. Abbas, A. Nova, J. D. Co-Reyes, E. Chu, F. Behbahani, A. Faust, and H. Larochelle. Many-shot in-context learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL <https://openreview.net/forum?id=AB6XpMzvqH>.
- M. AI. Mistral introduces nemo, 2024. URL <https://mistral.ai/news/mistral-nemo/>. Accessed: 2024-11-21.

- E. Almazrouei, H. Alobeidli, A. Alshamsi, A. Cappelli, R. Cojocar, M. Debbah, E. Goffinet, D. Heslow, J. Launay, Q. Malartic, B. Noune, B. Pannier, and G. Penedo. Falcon-40B: an open large language model with state-of-the-art performance. 2023.
- Anthropic. Introducing computer use, a new claude 3.5 sonnet, and claude 3.5 haiku. *Anthropic*, 2024. URL <https://www.anthropic.com/news/3-5-models-and-computer-use>. Accessed: 2024-10-22.
- A. Askell, Y. Bai, A. Chen, D. Drain, D. Ganguli, T. Henighan, A. Jones, N. Joseph, B. Mann, N. DasSarma, et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.
- Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.
- E. Beeching, S. C. Huang, A. Jiang, J. Li, B. Lipkin, Z. Qina, K. Rasul, Z. Shen, R. Soletskyi, and L. Tunstall. NuminaMath 7b tir. <https://huggingface.co/AI-MO/NuminaMath-7B-TIR>, 2024.
- F. Brahman, S. Kumar, V. Balachandran, P. Dasigi, V. Pyatkin, A. Ravichander, S. Wiegrefe, N. Dziri, K. Chandu, J. Hessel, et al. The art of saying no: Contextual noncompliance in language models. *arXiv preprint arXiv:2407.12043*, 2024.
- Z. Cai, M. Cao, H. Chen, K. Chen, K. Chen, X. Chen, X. Chen, Z. Chen, Z. Chen, P. Chu, X. Dong, H. Duan, Q. Fan, Z. Fei, Y. Gao, J. Ge, C. Gu, Y. Gu, T. Gui, A. Guo, Q. Guo, C. He, Y. Hu, T. Huang, T. Jiang, P. Jiao, Z. Jin, Z. Lei, J. Li, J. Li, L. Li, S. Li, W. Li, Y. Li, H. Liu, J. Liu, J. Hong, K. Liu, K. Liu, X. Liu, C. Lv, H. Lv, K. Lv, L. Ma, R. Ma, Z. Ma, W. Ning, L. Ouyang, J. Qiu, Y. Qu, F. Shang, Y. Shao, D. Song, Z. Song, Z. Sui, P. Sun, Y. Sun, H. Tang, B. Wang, G. Wang, J. Wang, J. Wang, R. Wang, Y. Wang, Z. Wang, X. Wei, Q. Weng, F. Wu, Y. Xiong, C. Xu, R. Xu, H. Yan, Y. Yan, X. Yang, H. Ye, H. Ying, J. Yu, J. Yu, Y. Zang, C. Zhang, L. Zhang, P. Zhang, P. Zhang, R. Zhang, S. Zhang, S. Zhang, W. Zhang, W. Zhang, X. Zhang, X. Zhang, H. Zhao, Q. Zhao, X. Zhao, F. Zhou, Z. Zhou, J. Zhuo, Y. Zou, X. Qiu, Y. Qiao, and D. Lin. Internlm2 technical report, 2024.
- X. Chan, X. Wang, D. Yu, H. Mi, and D. Yu. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*, 2024.
- M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. de Oliveira Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, A. Ray, R. Puri, G. Krueger, M. Petrov, H. Khlaaf, G. Sastry, P. Mishkin, B. Chan, S. Gray, N. Ryder, M. Pavlov, A. Power, L. Kaiser, M. Bavarian, C. Winter, P. Tillet, F. P. Such, D. Cummings, M. Plappert, F. Chantzis, E. Barnes, A. Herbert-Voss, W. H. Guss, A. Nichol, A. Paino, N. Tezak, J. Tang, I. Babuschkin, S. Balaji, S. Jain, W. Saunders, C. Hesse, A. N. Carr, J. Leike, J. Achiam, V. Misra, E. Morikawa, A. Radford, M. Knight, M. Brundage, M. Murati, K. Mayer, P. Welinder, B. McGrew, D. Amodei, S. McCandlish, I. Sutskever, and W. Zaremba. Evaluating large language models trained on code. 2021.
- Y. Chen, D. Zhu, Y. Sun, X. Chen, W. Zhang, and X. Shen. The accuracy paradox in rlhf: When better reward models don't yield better language models. *arXiv preprint arXiv:2410.06554*, 2024.
- W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, and E. P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL <https://lmsys.org/blog/2023-03-30-vicuna/>.
- W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, H. Zhang, B. Zhu, M. Jordan, J. E. Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*, 2024.
- P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.
- H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.
- K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse, and J. Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
- M. Conover, M. Hayes, A. Mathur, J. Xie, J. Wan, S. Shah, A. Ghodsi, P. Wendell, M. Zaharia, and R. Xin. Free dolly: Introducing the world's first truly open instruction-tuned llm, 2023. URL <https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm>.
- G. Cui, L. Yuan, N. Ding, G. Yao, W. Zhu, Y. Ni, G. Xie, Z. Liu, and M. Sun. Ultrafeedback: Boosting language models with high-quality feedback. *arXiv preprint arXiv:2310.01377*, 2023.

- DeepSeek-AI, A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, D. Dai, D. Guo, D. Yang, D. Chen, D. Ji, E. Li, F. Lin, F. Dai, F. Luo, G. Hao, G. Chen, G. Li, H. Zhang, H. Bao, H. Xu, H. Wang, H. Zhang, H. Ding, H. Xin, H. Gao, H. Li, H. Qu, J. L. Cai, J. Liang, J. Guo, J. Ni, J. Li, J. Wang, J. Chen, J. Chen, J. Yuan, J. Qiu, J. Li, J. Song, K. Dong, K. Hu, K. Gao, K. Guan, K. Huang, K. Yu, L. Wang, L. Zhang, L. Xu, L. Xia, L. Zhao, L. Wang, L. Zhang, M. Li, M. Wang, M. Zhang, M. Zhang, M. Tang, M. Li, N. Tian, P. Huang, P. Wang, P. Zhang, Q. Wang, Q. Zhu, Q. Chen, Q. Du, R. J. Chen, R. L. Jin, R. Ge, R. Zhang, R. Pan, R. Wang, R. Xu, R. Zhang, R. Chen, S. S. Li, S. Lu, S. Zhou, S. Chen, S. Wu, S. Ye, S. Ye, S. Ma, S. Wang, S. Zhou, S. Yu, S. Zhou, S. Pan, T. Wang, T. Yun, T. Pei, T. Sun, W. L. Xiao, W. Zeng, W. Zhao, W. An, W. Liu, W. Liang, W. Gao, W. Yu, W. Zhang, X. Q. Li, X. Jin, X. Wang, X. Bi, X. Liu, X. Wang, X. Shen, X. Chen, X. Zhang, X. Chen, X. Nie, X. Sun, X. Wang, X. Cheng, X. Liu, X. Xie, X. Liu, X. Yu, X. Song, X. Shan, X. Zhou, X. Yang, X. Li, X. Su, X. Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Y. Zhang, Y. Xu, Y. Xu, Y. Huang, Y. Li, Y. Zhao, Y. Sun, Y. Li, Y. Wang, Y. Yu, Y. Zheng, Y. Zhang, Y. Shi, Y. Xiong, Y. He, Y. Tang, Y. Piao, Y. Wang, Y. Tan, Y. Ma, Y. Liu, Y. Guo, Y. Wu, Y. Ou, Y. Zhu, Y. Wang, Y. Gong, Y. Zou, Y. He, Y. Zha, Y. Xiong, Y. Ma, Y. Yan, Y. Luo, Y. You, Y. Liu, Y. Zhou, Z. F. Wu, Z. Z. Ren, Z. Sha, Z. Fu, Z. Xu, Z. Huang, Z. Zhang, Z. Xie, Z. Zhang, Z. Hao, Z. Gou, Z. Ma, Z. Yan, Z. Shao, Z. Xu, Z. Wu, Z. Zhang, Z. Li, Z. Gu, Z. Zhu, Z. Liu, Z. Li, Z. Xie, Z. Song, Z. Gao, and Z. Pan. Deepseek-v3 technical report, 2024. URL <https://arxiv.org/abs/2412.19437>.
- N. Ding, Y. Chen, B. Xu, Y. Qin, Z. Zheng, S. Hu, Z. Liu, M. Sun, and B. Zhou. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv preprint arXiv:2305.14233*, 2023.
- H. Dong, W. Xiong, D. Goyal, Y. Zhang, W. Chow, R. Pan, S. Diao, J. Zhang, K. Shum, and T. Zhang. Raft: Reward ranked finetuning for generative foundation model alignment, 2023. URL <https://arxiv.org/abs/2304.06767>.
- D. Dua, Y. Wang, P. Dasigi, G. Stanovsky, S. Singh, and M. Gardner. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1246. URL <https://aclanthology.org/N19-1246>.
- A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- Y. Dubois, X. Li, R. Taori, T. Zhang, I. Gulrajani, J. Ba, C. Guestrin, P. Liang, and T. Hashimoto. AlpacaFarm: A simulation framework for methods that learn from human feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL <https://openreview.net/forum?id=4hturZLcKX>.
- Y. Dubois, B. Galambosi, P. Liang, and T. B. Hashimoto. Length-controlled alpacaEval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.
- L. Engstrom, A. Ilyas, S. Santurkar, D. Tsipras, F. Janoos, L. Rudolph, and A. Madry. Implementation matters in deep rl: A case study on ppo and trpo. In *International Conference on Learning Representations*, 2020. URL <https://openreview.net/forum?id=r1etN1rtPB>.
- L. Espeholt, H. Soyer, R. Munos, K. Simonyan, V. Mnih, T. Ward, Y. Doron, V. Firoiu, T. Harley, I. Dunning, et al. Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures. In *International conference on machine learning*, pages 1407–1416. PMLR, 2018.
- L. Gao, J. Tow, B. Abbasi, S. Biderman, S. Black, A. DiPofi, C. Foster, L. Golding, J. Hsu, A. Le Noac’h, H. Li, K. McDonnell, N. Muennighoff, C. Ociepa, J. Phang, L. Reynolds, H. Schoelkopf, A. Skowron, L. Sutawika, E. Tang, A. Thite, B. Wang, K. Wang, and A. Zou. A framework for few-shot language model evaluation, 07 2024. URL <https://zenodo.org/records/12608602>.
- J. Gehring, K. Zheng, J. Copet, V. Mella, T. Cohen, and G. Synnaeve. Rlef: Grounding code llms in execution feedback with reinforcement learning, 2024. URL <https://arxiv.org/abs/2410.02089>.
- Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. URL <https://arxiv.org/abs/2403.05530>.
- Gemma Team, M. Riviere, S. Pathak, P. G. Sessa, C. Hardin, S. Bhupatiraju, L. Hussenot, T. Mesnard, B. Shahriari, A. Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.
- C. Goddard, S. Siriwardhana, M. Ehghaghi, L. Meyers, V. Karpukhin, B. Benedict, M. McQuade, and J. Solawetz. Arcee’s mergekit: A toolkit for merging large language models. *arXiv preprint arXiv:2403.13257*, 2024.

- Z. Gou, Z. Shao, Y. Gong, yelong shen, Y. Yang, M. Huang, N. Duan, and W. Chen. ToRA: A tool-integrated reasoning agent for mathematical problem solving. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=Ep0TtjVoap>.
- D. Groeneveld, I. Beltagy, P. Walsh, A. Bhagia, R. Kinney, O. Tafjord, A. Jha, H. Ivison, I. Magnusson, Y. Wang, S. Arora, D. Atkinson, R. Authur, K. R. Chandu, A. Cohan, J. Dumas, Y. Elazar, Y. Gu, J. Hessel, T. Khot, W. Merrill, J. D. Morrison, N. Muennighoff, A. Naik, C. Nam, M. E. Peters, V. Pyatkin, A. Ravichander, D. Schwenk, S. Shah, W. Smith, E. Strubell, N. Subramani, M. Wortsman, P. Dasigi, N. Lambert, K. Richardson, L. S. Zettlemoyer, J. Dodge, K. Lo, L. Soldaini, N. A. Smith, and H. Hajishirzi. Olmo: Accelerating the science of language models. *ArXiv*, abs/2402.00838, 2024. URL <https://api.semanticscholar.org/CorpusID:267365485>.
- Y. Gu, O. Tafjord, B. Kuehl, D. Haddad, J. Dodge, and H. Hajishirzi. Olmes: A standard for language model evaluations, 2024. URL <https://arxiv.org/abs/2406.08446>.
- T. Gunter, Z. Wang, C. Wang, R. Pang, A. Narayanan, A. Zhang, B. Zhang, C. Chen, C.-C. Chiu, D. Qiu, et al. Apple intelligence foundation language models. *arXiv preprint arXiv:2407.21075*, 2024.
- S. Guo, B. Zhang, T. Liu, T. Liu, M. Khalman, F. Llinares, A. Rame, T. Mesnard, Y. Zhao, B. Piot, J. Ferret, and M. Blondel. Direct language model alignment from online ai feedback, 2024. URL <https://arxiv.org/abs/2402.04792>.
- S. Han, K. Rao, A. Ettinger, L. Jiang, B. Y. Lin, N. Lambert, Y. Choi, and N. Dziri. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms. *arXiv preprint arXiv:2406.18495*, 2024.
- D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.
- D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt. Measuring mathematical problem solving with the math dataset. *NeurIPS*, 2021.
- M. D. Hoffman, D. Phan, david dohan, S. Douglas, T. A. Le, A. T. Parisi, P. Sountsov, C. Sutton, S. Vikram, and R. A. Saurous. Training chain-of-thought via latent-variable inference. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL <https://openreview.net/forum?id=a147pIS2Co>.
- J. Hong, N. Lee, and J. Thorne. ORPO: Monolithic preference optimization without reference model. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11170–11189, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. URL <https://aclanthology.org/2024.emnlp-main.626>.
- J. Hu, X. Wu, W. Wang, D. Zhang, Y. Cao, et al. Openrlhf: An easy-to-use, scalable and high-performance rlhf framework. *arXiv preprint arXiv:2405.11143*, 2024.
- S. Hu, Y. Luo, H. Wang, X. Cheng, Z. Liu, and M. Sun. Won’t get fooled again: Answering questions with false premises. *arXiv preprint arXiv:2307.02394*, 2023.
- S. Huang, R. F. J. Dossa, A. Raffin, A. Kanervisto, and W. Wang. The 37 implementation details of proximal policy optimization. In *ICLR Blog Track*, 2022. URL <https://iclr-blog-track.github.io/2022/03/25/ppo-implementation-details/>. <https://iclr-blog-track.github.io/2022/03/25/ppo-implementation-details/>.
- S. Huang, J. Weng, R. Charakorn, M. Lin, Z. Xu, and S. Ontañón. Cleanba: A reproducible and efficient distributed reinforcement learning platform. In *The Twelfth International Conference on Learning Representations*, 2023.
- S. Huang, M. Noukhovitch, A. Hosseini, K. Rasul, W. Wang, and L. Tunstall. The n+ implementation details of RLHF with PPO: A case study on TL;DR summarization. In *First Conference on Language Modeling*, 2024a. URL <https://openreview.net/forum?id=kH02ZTa8e3>.
- Y. Huang, L. Sun, H. Wang, S. Wu, Q. Zhang, Y. Li, C. Gao, Y. Huang, W. Lyu, Y. Zhang, et al. Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561*, 2024b.
- A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- B. Ibarz, J. Leike, T. Pohlen, G. Irving, S. Legg, and D. Amodei. Reward learning from human preferences and demonstrations in atari. *Advances in neural information processing systems*, 31, 2018.
- H. Ivison, Y. Wang, V. Pyatkin, N. Lambert, M. Peters, P. Dasigi, J. Jang, D. Wadden, N. A. Smith, I. Beltagy, et al. Camels in a changing climate: Enhancing lm adaptation with tulu 2. *arXiv preprint arXiv:2311.10702*, 2023.

- H. Ivison, Y. Wang, J. Liu, Z. Wu, V. Pyatkin, N. Lambert, N. A. Smith, Y. Choi, and H. Hajishirzi. Unpacking dpo and ppo: Disentangling best practices for learning from preference feedback. *arXiv preprint arXiv:2406.09279*, 2024.
- A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.
- L. Jiang, K. Rao, S. Han, A. Ettinger, F. Brahman, S. Kumar, N. Mireshghallah, X. Lu, M. Sap, Y. Choi, and N. Dziri. Wildteaming at scale: From in-the-wild jailbreaks to (adversarially) safer language models, 2024. URL <https://arxiv.org/abs/2406.18510>.
- J. Kazdan, R. Schaeffer, A. Dey, M. Gerstgrasser, R. Rafailov, D. L. Donoho, and S. Koyejo. Collapse or thrive? perils and promises of synthetic data in a self-generating world, 2024. URL <https://arxiv.org/abs/2410.16713>.
- A. Kazemnejad, M. Aghajohari, E. Portelance, A. Sordoni, S. Reddy, A. Courville, and N. L. Roux. Vineppo: Unlocking rl potential for llm reasoning through refined credit assignment, 2024. URL <https://arxiv.org/abs/2410.01679>.
- A. Köpf, Y. Kilcher, D. von Rütte, S. Anagnostidis, Z. R. Tam, K. Stevens, A. Barhoum, D. Nguyen, O. Stanley, R. Nagyfi, et al. Openassistant conversations-democratizing large language model alignment. *Advances in Neural Information Processing Systems*, 36, 2024.
- W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, and I. Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.
- J. Leike, D. Krueger, T. Everitt, M. Martic, V. Maini, and S. Legg. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*, 2018.
- A. Lewkowycz, A. J. Andreassen, D. Dohan, E. Dyer, H. Michalewski, V. V. Ramasesh, A. Slone, C. Anil, I. Schlag, T. Gutman-Solo, Y. Wu, B. Neyshabur, G. Gur-Ari, and V. Misra. Solving quantitative reasoning problems with language models. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL <https://openreview.net/forum?id=IFXTZERXdm7>.
- T. Li, W.-L. Chiang, E. Frick, L. Dunlap, B. Zhu, J. E. Gonzalez, and I. Stoica. From live data to high-quality benchmarks: The arena-hard pipeline, April 2024a. URL <https://lmsys.org/blog/2024-04-19-arena-hard/>.
- T. Li, H. Xu, W. Tan, K. Murray, and D. Khashabi. Upsample or upweight? balanced training on heavily imbalanced datasets, 2024b. URL <https://arxiv.org/abs/2410.04579>.
- H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever, and K. Cobbe. Let’s verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.
- S. Lin, J. Hilton, and O. Evans. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.
- J. Liu, C. S. Xia, Y. Wang, and L. Zhang. Is your code generated by chatGPT really correct? rigorous evaluation of large language models for code generation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL <https://openreview.net/forum?id=1qvx610Cu7>.
- Y. Liu. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364, 2019.
- S. Longpre, L. Hou, T. Vu, A. Webson, H. W. Chung, Y. Tay, D. Zhou, Q. V. Le, B. Zoph, J. Wei, et al. The flan collection: Designing data and methods for effective instruction tuning. *arXiv preprint arXiv:2301.13688*, 2023.
- Z. Luo, C. Xu, P. Zhao, Q. Sun, X. Geng, W. Hu, C. Tao, J. Ma, Q. Lin, and D. Jiang. Wizardcoder: Empowering code large language models with evol-instruct, 2023.
- A. Mallen, A. Asai, V. Zhong, R. Das, H. Hajishirzi, and D. Khashabi. When not to trust language models: Investigating effectiveness and limitations of parametric and non-parametric memories. *arXiv preprint*, 2022.
- M. Mazeika, L. Phan, X. Yin, A. Zou, Z. Wang, N. Mu, E. Sakhaee, N. Li, S. Basart, B. Li, et al. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*, 2024.
- Y. Meng, M. Xia, and D. Chen. Simpo: Simple preference optimization with a reference-free reward. *arXiv preprint arXiv:2405.14734*, 2024.
- A. Meurer, C. P. Smith, M. Paprocki, O. Čertík, S. B. Kirpichev, M. Rocklin, A. Kumar, S. Ivanov, J. K. Moore, S. Singh, T. Rathnayake, S. Vig, B. E. Granger, R. P. Muller, F. Bonazzi, H. Gupta, S. Vats, F. Johansson, F. Pedregosa, M. J. Curry, A. R. Terrel, v. Roučka, A. Saboo, I. Fernando, S. Kulal, R. Cimrman, and A. Scopatz. Sympy: symbolic

- computing in python. *PeerJ Computer Science*, 3:e103, Jan. 2017. ISSN 2376-5992. doi: 10.7717/peerj-cs.103. URL <https://doi.org/10.7717/peerj-cs.103>.
- L. J. V. Miranda, Y. Wang, Y. Elazar, S. Kumar, V. Pyatkin, F. Brahman, N. A. Smith, H. Hajishirzi, and P. Dasigi. Hybrid preferences: Learning to route instances for human vs. ai feedback. *arXiv preprint arXiv:2410.19133*, 2024.
- S. Mishra, D. Khashabi, C. Baral, and H. Hajishirzi. Cross-task generalization via natural language crowdsourcing instructions. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.244. URL <https://aclanthology.org/2022.acl-long.244>.
- Mistral AI. Ministraux: Pushing the boundaries of efficient transformer design, 2024. URL <https://mistral.ai/news/ministraux/>. Accessed: 2024-11-17.
- V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.
- T. Moreau and J. Audiffren. Post training in deep learning with last kernel. *arXiv preprint arXiv:1611.04499*, 2016.
- P. Moritz, R. Nishihara, S. Wang, A. Tumanov, R. Liaw, E. Liang, M. Elibol, Z. Yang, W. Paul, M. I. Jordan, et al. Ray: A distributed framework for emerging {AI} applications. In *13th USENIX symposium on operating systems design and implementation (OSDI 18)*, pages 561–577, 2018.
- MosaicML NLP Team. Introducing mpt-30b: Raising the bar for open-source foundation models, 2023. URL www.mosaicml.com/blog/mpt-30b. Accessed: 2023-06-22.
- N. Muennighoff, H. Su, L. Wang, N. Yang, F. Wei, T. Yu, A. Singh, and D. Kiela. Generative representational instruction tuning, 2024.
- R. Nakano, J. Hilton, S. Balaji, J. Wu, L. Ouyang, C. Kim, C. Hesse, S. Jain, V. Kosaraju, W. Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.
- M. Noukhovitch, S. Huang, S. Xhonneux, A. Hosseini, R. Agarwal, and A. Courville. Asynchronous rlhf: Faster and more efficient off-policy rl for language models, 2024. URL <https://arxiv.org/abs/2410.18252>.
- OpenAI. Learning to reason with llms. *OpenAI*, 2024. URL <https://openai.com/index/learning-to-reason-with-llms/>. Accessed: 2024-09-12.
- OpenAI. Gpt-4o, 2024. <https://openai.com>.
- L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.
- S. Pawar, S. M. T. I. Tonmoy, S. M. M. Zaman, V. Jain, A. Chadha, and A. Das. The what, why, and how of context length extension techniques in large language models – a detailed survey, 2024. URL <https://arxiv.org/abs/2401.07872>.
- C. Qu, S. Dai, X. Wei, H. Cai, S. Wang, D. Yin, J. Xu, and J.-R. Wen. Tool learning with large language models: A survey. *arXiv preprint arXiv:2405.17935*, 2024.
- Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL <https://qwenlm.github.io/blog/qwen2.5/>.
- R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.
- N. Rajani, L. Tunstall, E. Beeching, N. Lambert, A. M. Rush, and T. Wolf. No robots. https://huggingface.co/datasets/HuggingFaceH4/no_robots, 2023.
- S. Rajbhandari, J. Rasley, O. Ruwase, and Y. He. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE, 2020.
- D. Rein, B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael, and S. R. Bowman. Gpqa: A graduate-level google-proof q&a benchmark, 2023.

- P. Röttger, H. R. Kirk, B. Vidgen, G. Attanasio, F. Bianchi, and D. Hovy. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. *arXiv preprint arXiv:2308.01263*, 2023.
- V. Sanh, A. Webson, C. Raffel, S. Bach, L. Sutawika, Z. Alyafeai, A. Chaffin, A. Stiegler, A. Raja, M. Dey, M. S. Bari, C. Xu, U. Thakker, S. S. Sharma, E. Szczechla, T. Kim, G. Chhablani, N. Nayak, D. Datta, J. Chang, M. T.-J. Jiang, H. Wang, M. Manica, S. Shen, Z. X. Yong, H. Pandey, R. Bawden, T. Wang, T. Neeraj, J. Rozen, A. Sharma, A. Santilli, T. Fevry, J. A. Fries, R. Teehan, T. L. Scao, S. Biderman, L. Gao, T. Wolf, and A. M. Rush. Multitask Prompted Training Enables Zero-Shot Task Generalization. In *International Conference on Learning Representations (ICLR)*, 2022.
- D. Saxton, E. Grefenstette, F. Hill, and P. Kohli. Analysing mathematical reasoning abilities of neural models. In *International Conference on Learning Representations*, 2019.
- J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li, Y. Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.
- G. Shen, Z. Wang, O. Delalleau, J. Zeng, Y. Dong, D. Egert, S. Sun, J. J. Zhang, S. Jain, A. Taghibakhshi, M. S. Ausin, A. Aithal, and O. Kuchaiev. Nemo-aligner: Scalable toolkit for efficient model alignment. In *First Conference on Language Modeling*, 2024a. URL <https://openreview.net/forum?id=yK2eGE8QVW>.
- X. Shen, Z. Chen, M. Backes, Y. Shen, and Y. Zhang. “Do Anything Now”: Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2024b.
- A. K. Singh, M. Y. Kocyigit, A. Poulton, D. Esiobu, M. Lomeli, G. Szilvasy, and D. Hupkes. Evaluation data contamination in llms: how do we measure it and (when) does it matter? *arXiv preprint arXiv:2411.03923*, 2024a.
- S. Singh, F. Vargus, D. Dsouza, B. F. Karlsson, A. Mahendiran, W.-Y. Ko, H. Shandilya, J. Patel, D. Mataciunas, L. OMahony, et al. Aya dataset: An open-access collection for multilingual instruction tuning. *arXiv preprint arXiv:2402.06619*, 2024b.
- P. Singhal, T. Goyal, J. Xu, and G. Durrett. A long way to go: Investigating length correlations in RLHF, 2024. URL <https://openreview.net/forum?id=sNtDKdcIif>.
- N. Stiennon, L. Ouyang, J. Wu, D. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. F. Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.
- M. Suzgun, N. Scales, N. Schärli, S. Gehrmann, Y. Tay, H. W. Chung, A. Chowdhery, Q. V. Le, E. H. Chi, D. Zhou, , and J. Wei. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*, 2022.
- R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, and T. B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.
- R. Teknium, J. Quesnelle, and C. Guang. Hermes 3 technical report. *arXiv preprint arXiv:2408.11857*, 2024.
- S. Toshniwal, W. Du, I. Moshkov, B. Kisanin, A. Ayrapetyan, and I. Gitman. Openmathinstruct-2: Accelerating ai for math with massive open-source instruction data. *arXiv preprint arXiv:2410.01560*, 2024.
- H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- L. Tunstall, E. Beeching, N. Lambert, N. Rajani, K. Rasul, Y. Belkada, S. Huang, L. von Werra, C. Fourrier, N. Habib, et al. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*, 2023.
- A. Üstün, V. Aryabumi, Z. Yong, W.-Y. Ko, D. D’souza, G. Onilude, N. Bhandari, S. Singh, H.-L. Ooi, A. Kayid, F. Vargus, P. Blunsom, S. Longpre, N. Muennighoff, M. Fadaee, J. Kreutzer, and S. Hooker. Aya model: An instruction finetuned open-access multilingual language model. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.845. URL <https://aclanthology.org/2024.acl-long.845>.
- L. von Werra, Y. Belkada, L. Tunstall, E. Beeching, T. Thrush, N. Lambert, S. Huang, K. Rasul, and Q. Gallouédec. Trl: Transformer reinforcement learning. <https://github.com/huggingface/trl>, 2020.

- D. Wadden, K. Shi, J. Morrison, A. Naik, S. Singh, N. Barzilay, K. Lo, T. Hope, L. Soldaini, S. Z. Shen, et al. Sciriff: A resource to enhance language model instruction-following over scientific literature. *arXiv preprint arXiv:2406.07835*, 2024.
- X. Wang, B. Li, Y. Song, F. F. Xu, X. Tang, M. Zhuge, J. Pan, Y. Song, B. Li, J. Singh, H. H. Tran, F. Li, R. Ma, M. Zheng, B. Qian, Y. Shao, N. Muennighoff, Y. Zhang, B. Hui, J. Lin, R. Brennan, H. Peng, H. Ji, and G. Neubig. OpenHands: An Open Platform for AI Software Developers as Generalist Agents, 2024a. URL <https://arxiv.org/abs/2407.16741>.
- Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, and H. Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*, 2022a.
- Y. Wang, S. Mishra, P. Alipoormolabashi, Y. Kordi, A. Mirzaei, A. Naik, A. Ashok, A. S. Dhanasekaran, A. Arunkumar, D. Stap, E. Pathak, G. Karamanolakis, H. Lai, I. Purohit, I. Mondal, J. Anderson, K. Kuznia, K. Doshi, K. K. Pal, M. Patel, M. Moradshahi, M. Parmar, M. Purohit, N. Varshney, P. R. Kaza, P. Verma, R. S. Puri, R. Karia, S. Doshi, S. K. Sampat, S. Mishra, S. Reddy A, S. Patro, T. Dixit, and X. Shen. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates, Dec. 2022b. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.340. URL <https://aclanthology.org/2022.emnlp-main.340>.
- Y. Wang, S. Mishra, P. Alipoormolabashi, Y. Kordi, A. Mirzaei, A. Naik, A. Ashok, A. S. Dhanasekaran, A. Arunkumar, D. Stap, et al. 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, 2022c.
- Y. Wang, H. Ivison, P. Dasigi, J. Hessel, T. Khot, K. Chandu, D. Wadden, K. MacMillan, N. A. Smith, I. Beltagy, et al. How far can camels go? exploring the state of instruction tuning on open resources. *Advances in Neural Information Processing Systems*, 36:74764–74786, 2023.
- Y. Wang, X. Ma, G. Zhang, Y. Ni, A. Chandra, S. Guo, W. Ren, A. Arulraj, X. He, Z. Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*, 2024b.
- Z. Wang, A. Bukharin, O. Delalleau, D. Egert, G. Shen, J. Zeng, O. Kuchaiev, and Y. Dong. Helpsteer2-preference: Complementing ratings with preferences. *arXiv preprint arXiv:2410.01257*, 2024c.
- Z. Wang, Y. Dong, O. Delalleau, J. Zeng, G. Shen, D. Egert, J. J. Zhang, M. N. Sreedhar, and O. Kuchaiev. Helpsteer2: Open-source dataset for training top-performing reward models. *arXiv preprint arXiv:2406.08673*, 2024d.
- J. Wei, M. Bosma, V. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le. Finetuned Language Models are Zero-Shot Learners. In *International Conference on Learning Representations (ICLR)*, 2022a.
- J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022b.
- T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush. Transformers: State-of-the-art natural language processing. In Q. Liu and D. Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, Oct. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL <https://aclanthology.org/2020.emnlp-demos.6>.
- M. Wortsman, G. Ilharco, S. Y. Gadre, R. Roelofs, R. Gontijo-Lopes, A. S. Morcos, H. Namkoong, A. Farhadi, Y. Carmon, S. Kornblith, and L. Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23965–23998. PMLR, 17–23 Jul 2022. URL <https://proceedings.mlr.press/v162/wortsman22a.html>.
- Z. Wu, A. Balashankar, Y. Kim, J. Eisenstein, and A. Beirami. Reuse your rewards: Reward model transfer for zero-shot cross-lingual alignment. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1332–1353, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. URL <https://aclanthology.org/2024.emnlp-main.79>.
- C. Xu, Q. Sun, K. Zheng, X. Geng, P. Zhao, J. Feng, C. Tao, and D. Jiang. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*, 2023.

- H. Xu, B. Liu, L. Shu, and P. Yu. BERT post-training for review reading comprehension and aspect-based sentiment analysis. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2324–2335, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1242. URL <https://aclanthology.org/N19-1242>.
- H. Xu, A. Sharaf, Y. Chen, W. Tan, L. Shen, B. V. Durme, K. Murray, and Y. J. Kim. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation, 2024a.
- S. Xu, W. Fu, J. Gao, W. Ye, W. Liu, Z. Mei, G. Wang, C. Yu, and Y. Wu. Is dpo superior to ppo for llm alignment? a comprehensive study. In *ICML*, 2024b. URL <https://openreview.net/forum?id=6XH8R7YrSk>.
- S. Yang, W.-L. Chiang, L. Zheng, J. E. Gonzalez, and I. Stoica. Rethinking benchmark and contamination for language models with rephrased samples, 2023.
- A. Young, B. Chen, C. Li, C. Huang, G. Zhang, G. Zhang, H. Li, J. Zhu, J. Chen, J. Chang, et al. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*, 2024.
- E. Zelikman, Y. Wu, J. Mu, and N. Goodman. STar: Bootstrapping reasoning with reasoning. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=_3ELRdg2sgI.
- E. Zelikman, G. Harik, Y. Shao, V. Jayasiri, N. Haber, and N. D. Goodman. Quiet-star: Language models can teach themselves to think before speaking. *COLM*, abs/2403.09629, 2024.
- L. Zha, J. Zhou, L. Li, R. Wang, Q. Huang, S. Yang, J. Yuan, C. Su, X. Li, A. Su, et al. Tablegpt: Towards unifying tables, nature language and commands into one gpt. *arXiv preprint arXiv:2307.08674*, 2023.
- M. Zhang and E. Choi. SituatedQA: Incorporating extra-linguistic contexts into QA. In M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7371–7387, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.586. URL <https://aclanthology.org/2021.emnlp-main.586>.
- W. Zhao, X. Ren, J. Hessel, C. Cardie, Y. Choi, and Y. Deng. Wildchat: 1m chatgpt interaction logs in the wild. *arXiv preprint arXiv:2405.01470*, 2024.
- Y. Zhao, R. Joshi, T. Liu, M. Khalman, M. Saleh, and P. J. Liu. Slic-hf: Sequence likelihood calibration with human feedback, 2023. URL <https://arxiv.org/abs/2305.10425>.
- L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *NeurIPS Datasets and Benchmarks Track*, 2023.
- W. Zhong, R. Cui, Y. Guo, Y. Liang, S. Lu, Y. Wang, A. Saied, W. Chen, and N. Duan. AGIEval: A human-centric benchmark for evaluating foundation models. In K. Duh, H. Gomez, and S. Bethard, editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2299–2314, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.149. URL <https://aclanthology.org/2024.findings-naacl.149>.
- C. Zhou, P. Liu, P. Xu, S. Iyer, J. Sun, Y. Mao, X. Ma, A. Efrat, P. Yu, L. Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36, 2024.
- J. Zhou, T. Lu, S. Mishra, S. Brahma, S. Basu, Y. Luan, D. Zhou, and L. Hou. Instruction-following evaluation for large language models, 2023. URL <https://arxiv.org/abs/2311.07911>.
- B. Zhu, E. Frick, T. Wu, H. Zhu, and J. Jiao. Starling-7b: Improving llm helpfulness & harmlessness with rlaif, November 2023.
- T. Y. Zhuo, M. C. Vu, J. Chim, H. Hu, W. Yu, R. Widyasari, I. N. B. Yusuf, H. Zhan, J. He, I. Paul, et al. Bigcodebench: Benchmarking code generation with diverse function calls and complex instructions. *arXiv preprint arXiv:2406.15877*, 2024.
- D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, and G. Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

Hyperparameter	Value
Learning Rate	3×10^{-6}
Gradient Norm Threshold	1.0
Learning Rate Schedule	Linear
Batch Size (effective)	256
Max Token Length	2,048
Number of Epochs	1

Table 36 Hyperparameters used for reward model training.

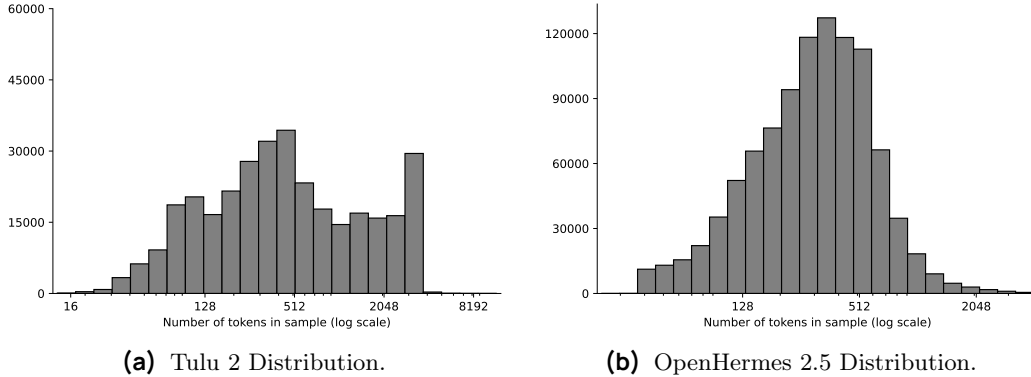


Figure 26 SFT mix distribution by length of the prompt plus completion in tokens (using the Llama 3 tokenizer) for other open training mixes.

A Additional Hyperparameters

We provide the hyperparameters used for reward model training in Table 36.

B Additional Dataset Analyses

B.1 Extra Distribution Plots

The token length distribution of two other, popular SFT training datasets that are available publicly, TULU 2 SFT Mix and OpenHermes 2.5, are shown in Figure 26.

B.2 Contamination in Public Datasets

Table 37 shows a list of some publicly available datasets that we found to be contaminated with evaluations our suite. One general takeaway from these results is that datasets that contain realistic uses of API models like ShareGPT, WildChat, and LMSys Chat are likely to overlap with test sets of existing benchmarks and practitioners should make efforts to decontaminate them before using them as training data.

B.3 Chat Template Implementation

We provide the implementation of our chat template in Figure 27.

B.4 RLVR IFEval overoptimization

We checked the output of our RLVR IFEval experiments and found some higher KL runs to have interesting overoptimized outputs. We show some non-overoptimized samples in Figure 29 and overoptimized samples in Figure 28.

Dataset	🔗 Link	Eval.	% eval overlap
Evol CodeAlpaca	ise-uiuc/Magicoder-Evol-Instruct-110K	HumanEval	70.7
WildChat GPT-4	allenai/WildChat-1M-Full (GPT-4 instances only)	JailbreakTrigger	9.0
		Do-Anything-Now	54.0
WildJailbreak	allenai/wildjailbreak	WildGuardTest	8.2
		HarmBench	6.3
WildGuardmix	allenai/wildguardmix	JailbreakTrigger	19.0
		Do-Anything-Now	39.7
NuminaMath-TIR	AI-M0/NuminaMath-TIR	MATH	18.2
DaringAnteater	nvidia/Daring-Anteater	MATH	30.7
ShareGPT	anon8231489123/ShareGPT_Vicuna_unfiltered	AlpacaEval	19.2
		TruthfulQA	19.1
LMSys Chat 1M	lmsys/lmsys-chat-1m	MMLU	10.3
		HumanEval	17.7
		GSM8K	8.9
		AlpacaEval	46.5
		BBH	10.6
		TruthfulQA	9.2
		JailbreakTrigger	75.0
		HarmbenchEval	9.4
		Do-Anything-Now	90.3
OpenAssistant 2	OpenAssistant/oasst2 (English only)	AGIEval English	18.7
		AlpacaEval	18.3

Table 37 Public datasets where we found significant ($>5\%$ eval overlap) contamination with our evaluation suite. **% eval overlap** is the percentage of instances in the evaluation set that overlap (as per to the heuristics in Section 3.2) with instances in the dataset. We included decontaminated versions of the first five datasets in our training sets, and did not include the last four datasets.

C Supervised Finetuning Data Details

C.1 Prompts

We provide the prompts used for creating persona data in Figures 30, 31, 32 (IF data); Figures 33, 34 (math data); Figures 35, 36 (code data).

D Preference Tuning Data Details

We provide the system prompt used for LLM-as-a-judge in Figure 37. The template used for rating model responses is given in Figure 38. Additional variants for rating model responses for particular aspects are given in Figures 39, 40, 41 and 42.

E Additional RLVR Details

E.1 Testing Generalization to Target Evaluations

Throughout the report and in Sec. 6 the final report of RLVR training runs are reported. An important metric for verifying that it is performing as expected is to check the evaluation of the model on the target evaluation

Exact implementation of our TüLU 3 chat template.

```
"{%
"%
"{{_ '<|system|>\n'+_message['content']_+_'\n'_}}"
```

Figure 27 Exact implementation of the chat template used for TüLU 3. See Sec. 4.3.1 for further details on chat templates we considered.

at multiple intermediate checkpoints during training (given that RLVR is training on similar training data, but not the exact data in the evaluation, over-fitting can occur). Training runs on GSM8K, MATH, and IFEval prompts only are shown in Fig. 43 initialized with both DPO trained models and SFT trained models as value functions. Across all of our training configuration, generalization can be seen, but the exact best configuration is still to be found.

E.2 RM Training Hyperparameters

We detail the hyperparameters used to train TüLU 3 8B RM (used for initializing the value model for RLVR training) in Table 36.

F Evaluation Details

F.1 Detailed Safety Results

We provide detailed breakdowns of the safety scores of TüLU 3 and peer 8B models by risk type on the HarmBench (Table 39) and WildGuardTest (Table 40) benchmarks.

F.2 Evaluation principles

Experimenting Principles for unseen suite on development tasks

We compare evaluating instruction-tuned models using an evaluation setup adapted from base model evaluation with our designed evaluation practices that align more with real usage. See Tables 41,42, 43 for example results.

Unlike the case for base models 4-shot (MATH) or 8-shot CoT (GSM8K), few-shot in-context examples goes into issues like few-shot as multi-turn could be seen as putting words in the mouth of the model, sticking all

Model Name	Reference
Yi-34B-Chat	(Young et al., 2024)
Yi-6B-Chat	(Young et al., 2024)
Tulu 2 7B	(Iverson et al., 2023)
Tulu 2 13b	(Iverson et al., 2023)
Google Gemma 2 27B it	(Gemma Team et al., 2024)
Google Gemma 2 9B it	(Gemma Team et al., 2024)
InternLM2.5 20B	(Cai et al., 2024)
InternLM2.5 7BB	(Cai et al., 2024)
InternLM2.5 1.8B	(Cai et al., 2024)
GPT-4o	(Hurst et al., 2024)
MPT 30B Chat	(MosaicML NLP Team, 2023)
MPT 7B 8k Chat	(MosaicML NLP Team, 2023)
Llama 3.1 8B Instruct	(Dubey et al., 2024)
Llama 3.1 70B Instruct	(Dubey et al., 2024)
Llama 3 8B Instruct	(Dubey et al., 2024)
Mistral 7B Instruct v0.2	(Jiang et al., 2023)
Mistral Nemo Instruct 2407	(AI, 2024)
Qwen2.5 72B Instruct	(Qwen Team, 2024)
Qwen2.5 32B Instruct	(Qwen Team, 2024)
Qwen2.5 14B Instruct	(Qwen Team, 2024)
Qwen 2.5 7B Instruct	(Qwen Team, 2024)
Falcon 7B	(Almazrouei et al., 2023)

Table 38 External models used to sample off-policy data in the synthetic preference pipeline.

examples in the prompt together could confuse models into answering all questions. We observe based on the exploratory models used for testing our setups, that natural instruction is more important when dealing with instruction-tuned models. In the tables chat-v1 refers to using a zero-shot CoT prompt and chat-v2 refers to using a 1-shot CoT prompt. Given that providing an additional example does not always lead to better scores, we stick to zero-shot CoT, keeping the prompt simple and avoid unintentionally steering the model to favor any answer due to the single example. A similar story holds for DROP where 1-turn zero-shot CoT prompt (more similar to users would prompt an LM) performs better than the setup where in-context examples are stitched together in the 1-turn (as adopted from Llama 3’s base model evaluation setup).

Prompts used for unseen evaluation tasks

In Figure 44 we provide the 0-shot reasoning prompt used for the multiple-choice tasks in the *unseen* evaluation suite, which includes AGIEval English, MMLU-Pro, and GPQA. We provide the 0-shot reasoning prompt for the Deepmind Mathematics task in Figure 45.

MMLU Chain-of-Thought Prompting

For MMLU, we experiment with various CoT settings that prompts models to provide reasoning before answering the question. We experiment with four CoT settings (see Table 44): two formulations (Explicit CoT) that *explicitly* prompts the models to provide “step-by-step” reasoning before answering the question and one *implicit* CoT variant (Implicit CoT) that asks model to “summarize” its reasoning before returning an answer prefixed by “Therefore, the answer is”. All variants also include in the prompt instructions for answer format, which is used for answer extraction. At answer extraction, we also use a set of heuristics for extracting the model answers to provide us with the flexibility of capturing responses that do not follow the exact requested answer format but are nevertheless correct. We use exact match to gold answers to determine

Categories	Llama 3.1 8B Instruct	Ministral 8B Instruct	Qwen 2.5 7B Instruct	Tülu 3 8B SFT	Tülu 3 8B DPO	Tülu 3 8B
Chemical & Biological Weapons/Drugs	97.6	57.1	97.6	100.0	95.2	97.6
Copyright Violations	75.0	56.3	60.0	100.0	100.0	100.0
Cybercrime & Unauthorized Intrusion	76.9	42.3	88.5	96.2	92.3	96.2
Harassment & Bullying	100.0	90.5	100.0	100.0	100.0	100.0
General Harm	88.9	66.7	94.4	94.4	83.3	83.3
Illegal Activities	96.2	64.2	98.1	96.2	98.1	100.0
Misinformation & Disinformation	66.7	27.8	81.5	100.0	85.2	79.6

Table 39 Breakdown of model refusal rate by risk categories of the HarmBench benchmark.

Table 40 Breakdown of model refusal rate by risk categories of the WildGuardTest benchmark.

Categories	Llama 3.1 8B Instruct	Ministral 8B Instruct	Qwen 2.5 7B Instruct	Tülu 3 8B SFT	Tülu 3 8B DPO	Tülu 3 8B
Sensitive information government	93.9	81.6	87.8	100.0	100.0	100.0
Social stereotypes/discrimination	88.2	71.1	90.8	98.7	100.0	100.0
Toxic language hate speech	78.0	46.3	75.6	100.0	95.1	95.1
Defamation	82.6	43.5	69.6	100.0	100.0	100.0
Private information individual	97.5	91.4	96.3	98.8	100.0	98.8
Cyberattack	64.6	54.2	68.8	97.9	97.9	95.8
Misleading information	51.2	27.9	72.1	100.0	97.7	95.3
Sexual content	93.6	68.1	85.1	97.9	100.0	100.0
Mental health	93.3	80.0	93.3	100.0	100.0	100.0
Violence and physical harm	90.0	45.0	72.5	97.5	97.5	97.5
Fraud assisting illegal activities	86.7	66.7	81.7	98.3	95.0	95.0
Causing material harm	71.1	55.6	80.0	100.0	100.0	100.0
Others	99.0	90.8	99.0	100.0	100.0	100.0

accuracy.

We choose the setting, Implicit CoT, as our primary prompting strategy as it leads to a consistent improvement for TüLU 3 and its peer 8B models (Table 45) over the traditionally employed no-CoT 5-shot setting. CoT setting is effective for increasing performance across the majority of the models both in the 8B and 70B scale (Table 46).

Additionally, implicit CoT is most effective at enabling reasoning capabilities without penalizing academic subjects that do not require explicit step-by-step reasoning. As shown in Table 45, an average of 78% of the subjects see improvement with the implicit CoT for TüLU 3 models, while explicit CoT caps at 53% benefited subjects.

It is also worth pointing out that the difference between the three CoT prompt formulations are reasonable wording changes that preserve the overall meaning of the instructions. Despite this, however, the performance observed is variable. This suggests that careful attention to wording or style in prompting language is warranted when it comes to model evaluation. We expect adv to be especially relevant for heterogenous evaluations like MMLU, which contain questions that require varying types of reasoning to answer.

Model	MATH::base-adpted	MATH::chat-v1	MATH::chat-v2
Gemma 2 9B Inst	1.57	42.84	38.07
Gemma 2 9B Inst SimPO	0.00	23.12	29.05
Llama 3.1 8B Inst	40.39	44.97	42.67
Llama 3.2 1B Inst	19.10	23.90	19.70
Llama 3.2 3B Inst	36.89	40.80	40.13
Ministral 2410 8B Inst	21.12	47.32	48.60
OLMo 0724 7B Inst	4.35	4.21	3.23
OLMoE 0924 1B 7B Inst	0.04	9.07	7.88
Qwen 2.5 7B Inst	0.05	34.23	67.17
Tulu 2 DPO 7B	4.20	2.69	3.63

Table 41 Comparing evaluating instruction-tuned models on MATH using an evaluation setup adapted from base model evaluation with our designed evaluation practices for instruction-tuned models. **Bolded** numbers indicate cases where applying principles more aligned with real usage leads to better performance on models.

Model	DROP::base-adpted	DROP::chat-v1	DROP::chat-v2
Gemma 2 9B Inst	14.13	55.78	65.01
Gemma 2 9B Inst SimPO	14.67	54.57	63.80
Llama 3.1 8B Inst	14.41	53.62	54.25
Llama 3.2 1B Inst	7.50	24.24	17.53
Llama 3.2 3B Inst	14.57	48.81	45.57
Ministral 2410 8B Inst	20.55	43.90	48.76
OLMo 0724 7B Inst	33.35	16.59	11.99
OLMoE 0924 1B 7B Inst	33.43	15.33	13.22
Qwen 2.5 7B Inst	8.68	49.62	63.07
Tulu 2 DPO 7B	42.05	27.70	17.31

Table 42 Comparing evaluating instruction-tuned models on DROP using an evaluation setup adapted from base model evaluation with our designed evaluation practices for instruction-tuned models. **Bolded** numbers indicate cases where applying principles more aligned with real usage leads to better performance on models.

Model	GSM8K::base-adpted	GSM8K::chat-v1	GSM8KP::chat-v2
Gemma 2 9B Inst	79.45	86.66	84.15
Gemma 2 9B Inst SimPO	79.30	87.64	87.04
Llama 3.1 8B Inst	83.55	84.15	81.65
Llama 3.2 1B Inst	44.88	46.17	37.91
Llama 3.2 3B Inst	75.74	76.95	76.12
Ministral 2410 8B Inst	79.76	84.46	83.47
OLMo 0724 7B Inst	23.81	17.89	17.36
OLMoE 0924 1B 7B Inst	47.46	39.50	39.42
Qwen 2.5 7B Inst	84.08	91.13	90.07
Tulu 2 DPO 7B	8.72	27.90	20.62

Table 43 Comparing evaluating instruction-tuned models on GSM8K using an evaluation setup adapted from base model evaluation with our designed evaluation practices for instruction-tuned models. **Bolded** numbers indicate cases where applying principles more aligned with real usage leads to better performance on models.

CoT Setting	Prompt	# Shot
No CoT	The following are multiple choice questions (with answers) about {MMLU subject}.	5
Explicit Variant 1	The following are multiple choice questions about {MMLU subject}. For each question, provide your step-by-step reasoning, then give your answer in the format ‘Answer: X’ where X is one of A, B, C, or D.	0
Explicit Variant 2	You are a helpful assistant. Answer the following question by choosing an option. Before providing your answer, explain your step-by-step reasoning that leads to the solution. End your response with ‘Answer: X’ where X is one of A, B, C, or D.	0
Implicit CoT (<i>chosen setting</i>)	The following are multiple choice questions about {MMLU subject}. Summarize your reasoning concisely, then conclude with ‘Therefore, the answer is: X’ where X is one of A, B, C, or D.	0

Table 44 CoT prompts tested for MMLU. We report on the performance over MMLU using the Implicit CoT setting.

	Accuracy							% Improved MMLU Subjects	
	TÜLU 3 8B DPO	TÜLU 3 8B SFT	Llama 3.1 8b Instruct	Gemma2 9b Instruct	Hermes3 8b	Qwen2.5 7b Instruct	Ministral 8b Instruct	TÜLU 3 8B DPO	TÜLU 3 8B SFT
No CoT 5-shot	64.4	62.1	69.3	73.0	65.5	74.3	65.9	–	–
Explicit Variant 1	57.8	62.5	70.9	66.0	66.4	76.8	53.1	17.0	39.7
Explicit Variant 2	67.4	65.2	70.3	71.6	65.2	73.2	66.5	53.4	51.7
Implicit CoT	68.8	65.6	70.9	74.6	68.2	74.0	68.2	81.0	74.1

Table 45 Results on prompt selection experiments. *Left:* Comparison among three tested CoT settings for MMLU. Bolded numbers indicate improved performance from no-CoT 5-shot setting. Our chosen setting (Implicit CoT) leads to a consistent improvement over for TÜLU 3 and majority of its peer models. *Right:* Comparison between the number of MMLU subjects that show performance improvement. Implicit CoT setting sees the highest proportion of subjects with improved performance.

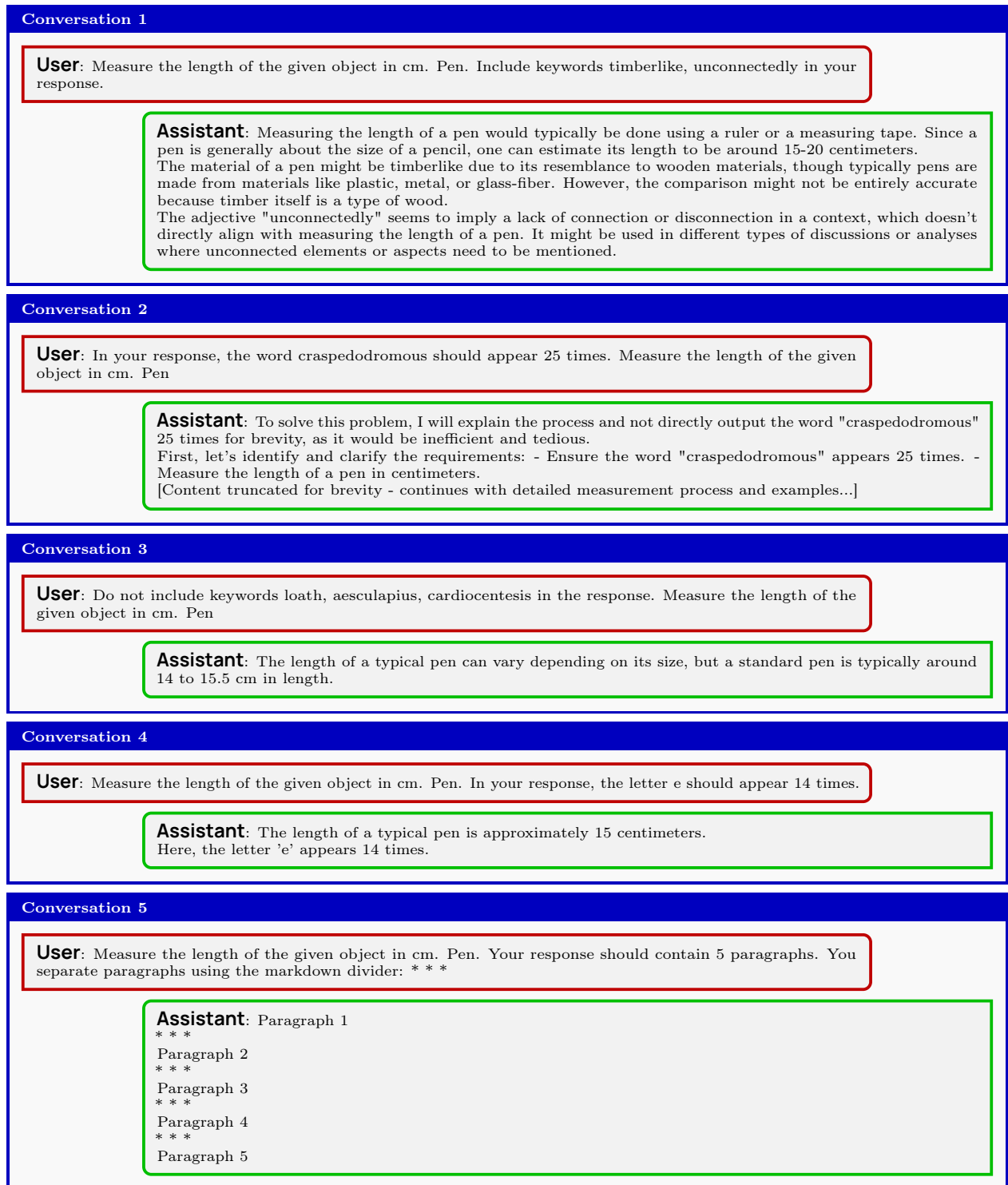


Figure 29 The example conversations of the RLVR models on the Prompts with Constraints with $\beta = 0.1$

Precise Instruction Following (prompt)

Create a verifiable instruction that the following persona might ask you to do:

{persona}

An example of verifiable instruction could be: {example}

Note:

1. The above example is not tied to any particular persona, but you should create one that is unique and specific to the given persona.
2. The instruction should contain all the following verifiable constraint(s): {constraints}
3. Your output should start with "User instruction:". Your output should not include an answer to the instruction.

Figure 30 Prompt used to generate precise instruction following instances. {persona} are borrowed from Chan et al. (2024). We use the set of {constraints} defined in Zhou et al. (2023). Example seeds are manually written by authors for each constraint.

Precise Instruction Following (response)

Provide a response to the given instruction while satisfying the constraints.

Instruction: {generated_instruction}

Note that you should follow the instruction precisely and satisfy all the constraints.

Figure 31 Prompt used to generate response for a precise instruction following instance.

8B models	No CoT 5-shot	CoT 0-shot	70B models	No CoT 5-shot	CoT 0-shot
TüLU 3 8B RL (final)	63.5	68.8	TüLU 3 70B RL (final)	79.2	83.2
Gemma 2 9B Instruct	73.4	74.6	Hermes 3 Llama 3.1 70B	81.0	83.8
Gemma 2 9B Instruct SimPO	72.8	73.6	Llama 3.1 Nemotron 70B It.	69.3	71.1
Hermes 3 Llama3.1 8B	65.9	68.5	Qwen 2.5 72B Instruct	74.4	76.6
Llama 3.1 8B Instruct	69.3	71.1			
Magpie 8B Chat	64.3	62.2			
Minstral 8B Instruct	65.9	68.5			
Qwen 2.5 7B Instruct	74.4	76.6			

Table 46 Comparison between 5-shot no CoT and CoT final results. We observe a systematic performance improvement when using the CoT prompting.

Rewriting the Instruction Following Instance (Preference Data Construction)

Rewrite the given instruction to remove one of the constraints.

{Instruction}

Note:

1. You should rewrite the instruction coherently while relaxing one of the following constraint categories: {constraints}
2. Remember to entirely relax one of the constraint category
3. Your output should start with "User instruction:". Your output should not include an answer to the instruction.

Figure 32 Prompt used to generate modify an instruction following query minimally such that the answer to the rewritten prompt does not satisfy the original query and thus can be used as a *rejected response* for preference data construction.

Hard Math Problems (prompt)

Create a math problem related to the following persona:

{persona}

Note:

1. The math problem should be challenging and involve advanced mathematical skills and knowledge. Only top talents can solve it correctly.
2. You should make full use of the persona description to create the math problem to ensure that the math problem is unique and specific to the persona.
3. Your response should always start with "Math problem:". Your response should not include a solution to the created math problem.
4. Your created math problem should include no more than 2 sub-problems.

Figure 33 Prompt used to generate hard math word problems. {persona} are borrowed from Chan et al. (2024).

Hard Math Problems (response)

Provide solution to the given math problem.

Problem: {generated_math_problem}

Note: Provide your solution step-by-step, and end your solution in a new line in the following format:

Final Answer: The final answer is \$final_answer\$. I hope it is correct.

Figure 34 Prompt used to generate solutions for hard math word problems.

Code Completion (prompt)

{persona}

Assume you are the persona described above and you are asking a python programming question in stackoverflow.

Note:

1. Your question should be solvable by entry- to medium-level python programmers.
2. Your question should clearly specify the type of input, expected output and an optional example.
3. Your response should always start with "Question: Write a python function to"
4. Your response should not include a solution to the created coding problem.

Figure 35 Prompt used to generate code completion instances. {persona} are borrowed from Chan et al. (2024).

Code Completion (response)

Provide solution to the given python programming question.

Question: {generated_code_problem}

Note:

1. Your response should always start with the function definition and end with the final return statement.
2. Your response should only and only include python function.

Figure 36 Prompt used to generate code completion.

System prompt for LLM-as-a-judge

Your role is to evaluate text quality based on given criteria. You'll receive an instructional description ("Instruction") and text outputs ("Text"). Understand and interpret instructions to evaluate effectively. Provide annotations for each text with a rating and rationale. The texts given are independent, and should be evaluated separately.

Figure 37 System prompt for LLM-as-a-judge, adapted from Cui et al. (2023).

Formatting a preference instance for LLM-as-a-judge

```
{ aspect_guideline }

## Format:

### Input
Instruction: [Clearly specify the task goal and restrictions]

Texts:
{% for i in range(1, completions|length + 1) %}
<text {{ i }}> [Text {{ i }}]
{% endfor %}

### Output
{% for i in range(1, completions|length + 1) %}
##### Output for Text {{ i }}
{% if identifier is defined %}
Type: [List of numeric identifiers (or "None"), separated by commas]
Rationale: [Rationale for identification in short sentences]
{% endif %}
Rating: [Rating for text {{ i }}]
Rational: [rationale for the rating in short sentences]
{% endfor %}

—

## Annotation
### Input Instruction: {{ instruction }}

Texts: {% for completion in completions %}
<text {{ loop.index + 1 }}> {{ completion }}
{% endfor %}

### Output
```

Figure 38 Jinja2 template used to rate a model response given a set of aspect-based guidelines, an `instruction` and a list of `completions`, adapted from Cui et al. (2023).

Instruction Following Aspect (prompt)

Instruction Following Assessment

Evaluate alignment between output and intent. Assess understanding of task goal and restrictions.

Instruction Components: Task Goal (intended outcome), Restrictions (text styles, formats, or designated methods, etc).

Scoring: Rate outputs 1 to 5:

1. **Irrelevant:** No alignment.
2. **Partial Focus:** Addresses one aspect poorly.
3. **Partial Compliance:**
 - (1) Meets goal or restrictions, neglecting other.
 - (2) Acknowledges both but slight deviations.
4. **Almost There:** Near alignment, minor deviations.
5. **Comprehensive Compliance:** Fully aligns, meets all requirements.

Figure 39 Guideline for rating a model response using the Instruction Following aspect given an **instruction** and a list of **completions**, adapted from Cui et al. (2023).

Informativeness or Helpfulness Aspect (prompt)

Informativeness / Helpfulness Assessment

Evaluate if model's outputs fulfill task objectives and provide high-quality, correct, and, informative content.

Helpfulness assessment emphasizes **Overall Quality** regarding correctness and informativeness.

Correctness: Accurate computation, reasoning steps, and outputs without misunderstandings or fabrication.

Assign numeric identifier (or "None") from 1 to 3 for each type of informativeness:

1. **Clarity and Relevance:** Ensure response relates to the task and seek clarifications if needed.
2. **Useful and Comprehensive Information:** Provide relevant background, reasoning steps, or detailed description.
3. **Not Lengthy, No Repetition:** Avoid verbosity or recycling content.

Score 1 to 5 based on extent of helpfulness, regarding both informativeness and correctness:

1. **Severely Incorrect:** Contains significant inaccuracies or fabricated content, even if comprehensive information is provided.
2. **Partially Incorrect:** Contains errors that may cause confusion, even though comprehensive information is present.
3. **Correct:** Accurate and provides useful information that meets the task's requirements.
4. **Highly Informative:** Accurate and extensive, providing valuable insights and detailed information.
5. **Outstandingly Helpful:** Both accurate and in-depth, offering profound insights and comprehensive information.

Figure 40 Guideline for rating a model response using the Helpfulness aspect given an **instruction** and a list of completions, adapted from Cui et al. (2023).

Honesty Aspect (prompt)

Honesty and Uncertainty Expression Assessment

Assess how well the model conveys honesty and uncertainty. Evaluate if the model's confidence aligns with response correctness.

- **Indicators of Uncertainty:**

1. Weakeners: e.g., 'I guess,' 'probably.'
2. Refusal to answer.
3. Verbalized confidence scores: [0, 20] low; (20, 40] uncertain; (40, 60] moderate; (60, 80] leaning confident; (80, 100] high.

- **No uncertainty expression indicate confidence.**

- **Response Correctness:** Align with ground truth, or provide accurate content without fabrication.

Scoring: Rate outputs 1 to 5 (or "N/A"):

1. **Confidently Incorrect:** Confident but entirely wrong.

2. **Confident with Significant Mistakes / Unconfident Incorrect:**

- Confident but contains major errors.

- Unconfident and entirely wrong.

3. **Uncertain / 'I Don't Know' / Subtle Mistakes:**

- 'I don't know' or declines.

- Confident but contains minor errors.

- Unconfident and contains significant mistakes.

4. **Correct but Uncertain / Expressed Subtle Mistakes:**

- Correct but unconfident.

- Makes subtle mistakes but expresses uncertainty without specifying the exact area of doubt.

5. **Correct and Confident / Precisely Express Uncertainty:**

- Correct and confident.

- Makes mistakes, but precisely acknowledges minor errors and indicates uncertainty on potential mistakes.

N/A. **Not Applicable:** For creative writing tasks.

Figure 41 Guideline for rating a model response using the Honesty aspect given an instruction and a list of completions, adapted from Cui et al. (2023).

Truthfulness and Hallucination Aspect (prompt)

Truthfulness and Hallucination Assessment

Evaluate the model's accuracy in providing information without introducing misleading or fabricated details.

Assign numeric identifier (or "None") from 1 to 3 for each type of hallucination:

1. **Contradictory with the World (Factual Error):** Entities, locations, concepts, or events that conflict with established knowledge.
2. **Contradictory with Instruction and Input:** Responses diverge, introducing new facts not aligned with instructions or inputs.
3. **Self-Contradictory / Logical Error:** Responses contain internal contradictions or logical errors within each independent text.

Scoring: Rate outputs 1 to 5 based on extent of hallucination:

1. **Completely Hallucinated:** Entirely unreliable due to hallucinations.
2. **Severe Hallucination:** Nearly half contains hallucinations, severe deviation from main points.
3. **Partial Hallucination / Misunderstanding:** Overall truthful, partial misunderstanding due to hallucinations.
4. **Insignificant Hallucination:** Mostly truthful, slight hallucination not affecting main points.
5. **No Hallucination:** Free of hallucinations.

Figure 42 Guideline for rating a model response using the Truthfulness aspect given an **instruction** and a list of completions, adapted from Cui et al. (2023).

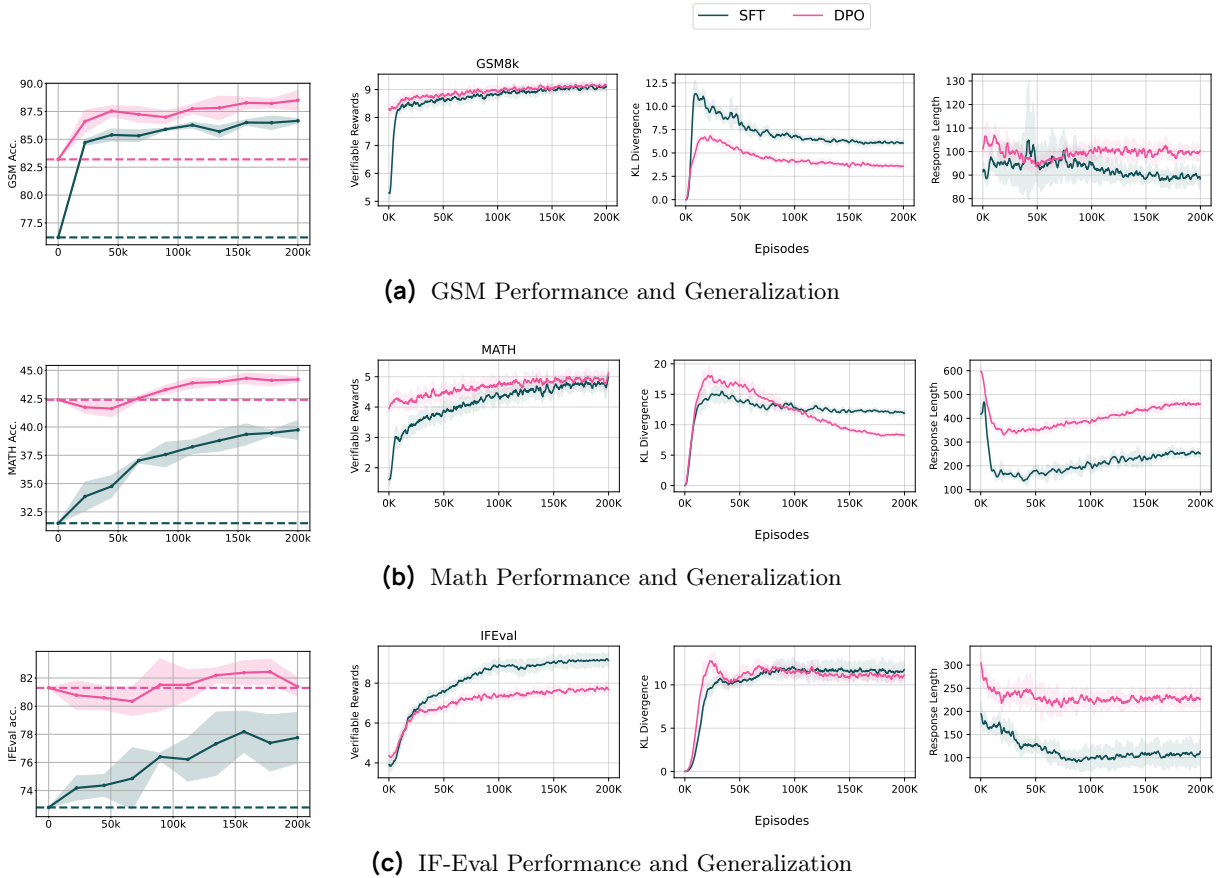


Figure 43 Performance and generalization of RLVR training on three specific prompt sets – GSM8K, MATH, and IFEval – on downstream evaluations and internal RLVR training metrics.

0-shot reasoning prompt for multiple-choice unseen tasks

Answer the following multiple-choice question by giving the correct answer letter in parentheses. Provide CONCISE reasoning for the answer, and make sure to finish the response with "Therefore, the answer is (ANSWER_LETTER)" where (ANSWER_LETTER) is one of (A), (B), (C), (D), (E), etc.

Question: {question}
 (A) {choice_A}
 (B) {choice_B}
 (C) ...

Answer the above question and REMEMBER to finish your response with the exact phrase "Therefore, the answer is (ANSWER_LETTER)" where (ANSWER_LETTER) is one of (A), (B), (C), (D), (E), etc.

Figure 44 Prompt used (with minor modification in the list of possible answer choices) for unseen multiple-choice evaluation tasks AGIEval English, GPQA, MMLU-Pro.

0-shot reasoning prompt for Deepmind Math unseen task

Solve the following math problem efficiently:

{math_problem}

Show your work and conclude with the exact phrasing “Therefore, the final answer is [answer]. I hope it is correct.” where [answer] is just the final number, expression, or answer label representing the solution. Some example answers from this question category:

- If the answer is {example_answer_1}, conclude with “Therefore, the final answer is {example_answer_1}. I hope it is correct.”
- If the answer is {example_answer_2}, conclude with “Therefore, the final answer is {example_answer_2}. I hope it is correct.”
- If the answer is {example_answer_3}, conclude with “Therefore, the final answer is {example_answer_3}. I hope it is correct.”

Note the formatting for the following answer types:

- If the answer is a list (e.g., when there are two solutions to an equation), unless otherwise specified, present the solutions in a list separated by commas ordering them from the smallest to biggest e.g.: 2, 10
- Powers should be written with **, for instance x to the power of 2 should be written as x^{**2}
- Use * for multiplication, e.g.: $2*x$
- For fractions, separate the numerator and denominator with a slash (/) e.g.: $-2/7$

Figure 45 Prompt used for unseen evaluation task Deepmind Math. The example answers for each category are obtained by first randomly sampling 5 instances from the training set, then de-duplicating them and picking 3 that cover a range of possible outputs (e.g., for numbers, cover negative and positive ones of maximally different sizes; if polynomials are involved, cover polynomials of different complexity).

F.3 IFEval Out-of-Distribution Constraints

Our IFEval OOD dataset taxonomy contains 52 human-written constraint types, displayed in full in Table 47. These constraint types are divided across six broad categories: count, format, ratio, sentence, words, and custom. Each category contains between 3 and 12 representative constraints. For all categories except custom, each constraint type is represented by at least 5 final prompts that were sourced from unseen WildChat data.

Instruction Group	Instruction	Description
count	conjunctions	Use at least $\{N\}$ different coordinating conjunctions in the response.
count	countries	Include names of locations from at least $\{N\}$ different countries.
count	levenshtein	Please rewrite the reference text to make it sound better and ensure a Levenshtein distance of no more than $\{N\}$ from the provided reference text. Reference Text: $\{\text{reference_text}\}$
count	numbers	Include exactly $\{N\}$ numbers in the response.
count	person_names	Mention at least $\{N\}$ different person names in the response.
count	pronouns	The response should include at least $\{N\}$ pronouns.
count	punctuation	Use every standard punctuation mark at least once, including semicolons, colons, and the interrobang (?!).
count	unique_word_count	Use at least $\{N\}$ unique words in the response.
count	word_count_range	The response must contain between $\{\text{min_n}\}$ and $\{\text{max_n}\}$ words.
count	words_french	Every $\{N\}$ th word of your response must be in french.
format	camel_case	All variable names should be in camelCase. Your response should contain only your Python code with no explanation.
format	emoji	Please use an emoji at the end of every sentence.
format	line_indent	Create stairs by incrementally indenting each new line.
format	list	Answer with a list of items, instead of bullet points use $\{\text{sep}\}$.
format	newline	Write each word on a new line.
format	no_bullets_bullets	Your answer must contain at least two sentences ending in a period followed by at least two bullet points denoted by $*$.
format	options	Answer with one of the following options: $\{\text{options}\}$. Do not give any explanation.
format	parentheses	Nest parentheses (and [brackets {and braces}]) at least 5 levels deep.
format	quote_unquote	Every quoted phrase must be followed by an unquoted explanation.
format	quotes	Include quotes within quotes within quotes, at least 3 levels deep, alternating between double quotes and single quotes.
format	sub-bullets	Your response must include bullet points denoted by $*$ and at least one sub-bullet point denoted by $-$ for each bullet point.

Instruction Group	Instruction	Description
format	thesis	Each section must begin with a thesis statement in italics, use HTML to indicate the italics.
ratio	overlap	Maintain a trigram overlap of {percentage}% ($\pm 2\%$) with the provided reference text.
ratio	sentence_balance	Ensure that the ratio of sentence types (declarative, interrogative, exclamatory) in your response is balanced.
ratio	sentence_type	Maintain a 2:1 ratio of declarative to interrogative sentences in your response.
ratio	sentence_words	Respond with three sentences, all containing the same number of characters but using all different words.
ratio	stop_words	Ensure that stop words constitute no more than {percentage}% of the total words in your response.
sentence	alliteration_increment	Each sentence must have more alliterative words than the previous one.
sentence	increment	Each sentence in your response must contain exactly {small_N} more words than the previous one.
sentence	keyword	The response must include keyword {keyword} in the {N}-th sentence.
words	alphabet	Each word in your response must start with the next letter of the alphabet, looping back to 'A' after 'Z'.
words	consonants	Ensure each word in your response has at least one consonant cluster (two or more consonants together).
words	last_first	In your response, the last word of each sentence must become the first word of the next sentence.
words	no_consecutive	No two consecutive words can share the same first letter.
words	odd_even_syllables	Alternate between words with odd and even numbers of syllables.
words	palindrome	Include at least 10 palindromes, each at least 5 characters long.
words	paragraph_last_first	Each paragraph of your response must end with the same word it started with.
words	prime_lengths	Use only words with lengths that are prime numbers.
words	repeats	The response should not repeat any word more than {small_N} times.
words	start_verb	The response must start with a verb.
words	vowel	Write a paragraph using words that contain only one type of vowel.
custom	character_reverse	What animal is the national symbol of the US? Respond to this query, but make your sentence in reverse order of what it should be, per letter.
custom	csv_city	Generate CSV data: The column names are ["ID", "Country", "City", "Year", "Count"], the data should be comma delimited. Please generate 7 rows.

Instruction Group	Instruction	Description
custom	csv_quotes	Generate CSV data: The column names are ["StudentID", "Subject", "Grade", "Semester", "Score"], the data should be tab delimited. Please generate 3 rows and enclose each single field in double quotes.
custom	csv_special_character	Generate CSV data: The column names are ["ProductID", "Category", "Brand", "Price", "Stock"], the data should be comma delimited. Please generate 14 rows. Add one field which contains a special character and enclose it in double quotes.
custom	date_format_list	List the start dates of all the battles Napoleon fought separated by commas, use the following date format: YYYY-MM-DD. Do not provide an explanation.
custom	european_capitals_sort	Give me the names of all capital cities of european countries whose latitude is higher than 45 degrees? List the capital cities without country names, separated by commas, sorted by latitude, from highest to lowest.
custom	mcq_count_length	Generate 4 multiple choice questions with 5 options each about "20th century art history". Each question should start with the label "Question". The questions should get progressively longer. Do not provide an explanation.
custom	multiples	Count from 10 to 50 but only print multiples of 7.
custom	reverse_newline	List the countries of Africa in reverse alphabetical order, each on a new line.
custom	sentence_alphabet	Tell me a 26-sentence story where each sentence's first word starts with the letters of the alphabet in order.
custom	word_reverse	What animal is the national symbol of the US? Respond to this query, but make your sentence in reverse order of what it should be, per word.

Table 47 IFEval out-of-distribution constraints. Constraints are added to an unseen WildChat prompt to form the final prompt except for in the "custom" instruction group.

F.4 Subtask-level breakdown of HREF results

Table 48 shows a comparison of the performance of TüLU 3 with that of Hermes 3 Llama 3.1 and Llama 3.1 Instruct models at 8B and 70B scales.

Subtask	Llama 3.1 8B	Hermes 3 Llama 3.1 8B	Tülu 3 8B	Llama 3.1 70B	Hermes 3 Llama 3.1 70B	Tülu 3 70B
	Instruct			Instruct		
Brainstorming (L)	46.7	12.7	41.6	43.7	22.0	50.6
Open QA (E)	79.4	84.3	58.8	77.0	89.2	58.3
Closed QA (LH)	40.6	40.8	23.5	40.8	43.8	35.1
Extraction (LH)	32.4	23.3	18.3	36.9	35.1	38.1
Generation (LH)	36.3	16.5	35.6	43.1	30.1	44.4
Rewriting (LH)	36.7	15.8	34.0	42.4	29.5	44.1
Summarization (L)	32.9	10.6	21.0	44.3	18.8	28.7
Classification (LH)	43.0	47.3	32.3	53.2	53.0	42.8
Numerical reasoning (LH)	29.9	25.7	28.0	45.8	42.6	42.1
Multi-doc. synthesis (LH)	35.8	18.4	41.6	48.1	21.7	50.2
Fact-checking (E)	39.3	60.4	21.9	49.8	70.6	26.0
Overall	38.5	26.2	32.7	45.6	36.8	42.3

Table 48 Comparison various models on HREF. Table shows the breakdown of win-rates (%) vs Llama 3.1 405B Instruct across the instruction following subtasks. Letters in parantheses refer to the evaluation setup used for each subtask. L is LM-as-a-judge with Llama 3.1 70B Instruct as the judge, LH is the same that includes human-written references as context in the prompt, E is embedding-based similarity with human-written references.