A. Model and Dataset Details

A.1. Model Architecture and training configurations

Table 3: Comparison of model architectures and training setups.

	Pythia	**OLMo-2**	**GPT-2**
Position Embedding	Learned	Rotary (RoPE)	Learned
Norm Type	LayerNorm	RMSNorm	LayerNorm
Norm Position	Pre-layer	Pre-layer	Pre-layer
Dataset	The Pile (825 GB)	OLMoStack (4T tokens)	Fineweb (10BT)
Optimizer	AdamW	AdamW	AdamW
LR Scheduler	Cosine decay	Linear decay w/ warmup	Cosine decay
Loss Function	Cross-Entropy	Cross-Entropy	Cross-Entropy

Table 4: Tülu Model Architecture and Training Setup

Component	**Tülu**
Base Models	Llama 3 base models
Position Embedding	Inherited from base model
Normalization Type	Inherited from base model (LayerNorm)
Normalization Position	Pre-layer
Instruction Datasets	Tulu3 Mixture(FLAN V2, OpenAssistant, WildChat GPT-4)
Training Techniques	SFT, DPO, RLVR
Optimizer	AdamW
Learning Rate Scheduler	Linear decay with warmup
Loss Function	Cross-Entropy

A.2. Dataset Details

In this section, we provide an overview of the datasets used in our experiments.

FineWeb: The FineWeb dataset (Penedo et al., 2024) consists of more than 15T tokens of cleaned and deduplicated english text obtained from the web using CommonCrawl. While the full dataset contains 15T tokens, we use the smallest subset, i.e. a subsampled version of the dataset consisting of 10B tokens. The dataset is accessible on HuggingFace at https://huggingface.co/datasets/HuggingFaceFW/fineweb.

WikiText: The Wikitext dataset (Merity et al., 2016) is a collection of over 100 million tokens extracted from the set of verified Good and Featured articles on Wikipedia. We use only a subset of the dataset to perform early evaluations of `RankMe` and αReQ, before running our final experiments using FineWeb.

SciQ: The SciQ dataset (Welbl et al., 2017) contains over 13K crowdsourced science exam questions about physics, chemistry and biology, among many others. The questions are in multiple-choice format with 4 answer options each. The dataset is accessible on HuggingFace at https://huggingface.co/datasets/allenai/sciq.

TriviaQA: The TriviaQA dataset (Joshi et al., 2017) is a reading comprehension dataset containing over 650K question-answer-evidence triples. We use the TriviaQA dataset to evaluate a model's ability to

incorporate long-context information from the question in order to correctly answer it. The dataset is accessible on HuggingFace at https://huggingface.co/datasets/mandarjoshi/trivia_qa.

LAMBADA OpenAI: This dataset (Radford et al., 2019) is comprised of the LAMBADA test split, pre-processed by OpenAI, and contains machine translated versions of the split in German, Spanish, French and Italian. We use this dataset to evaluate the model's text understanding capabilities. The dataset is accessible on HuggingFace at https://huggingface.co/datasets/EleutherAI/lambada_openai.

Anthropic Helpful-Harmless (HH): The Anthropic-HH dataset provides human preference data about helpfulness and harmlessness, and is meant to be used for training preference models in a Reinforcement Learning with Human Feedback (RLHF) setting. However, we use a variant of this dataset for SFT. Specifically, we generate a human-assistant chat dataset of $\sim 161K$ samples by parsing the "chosen" responses for each instruction from the original dataset and using it to finetune a base model by treating the "chosen" response as the target (similar to (Springer et al., 2025)). While such a use of this dataset is discouraged in practical settings, we use this modified dataset as a testbed for our SFT experiments. The original dataset is accessible on HuggingFace at https://huggingface.co/datasets/Anthropic/hh-rlhf.

AlpacaFarm Human-ANN chat (AlpacaFarm): This dataset is created by following a similar procedure as mentioned above for the Anthropic-HH dataset, but for the Human Evaluation dataset of the AlpacaFarm evaluation set (Dubois et al., 2023). As a result, this dataset consists of $\sim 17.7K$ samples, and is used as a positive control in our SFT experiments. Models that are finetuned on this dataset are expected to perform well on the AlpacaEval chat task (see below), compared to models that are finetuned on a different dataset. This positive control is essential to disentangle the in-distribution vs out-of-distribution abilities of a SFT-model. The original dataset is accessible on HuggingFace at https://huggingface.co/datasets/tatsu-lab/alpaca_farm.

AlpacaEval: AlpacaEval is an LLM-based automatic evaluation setup for comparing chat models in a fast, cheap and replicable setting. We use AlpacaEval as a test bench to study the behavior of models after undergoing SFT. Models that are finetuned on the AlpacaFarm dataset are expected to produce better chat models and generate responses more aligned to human-preferred responses to instructions in the AlpacaEval setup. We defer the reader to the corresponding github repository for further details of the evaluation setup.

AMC23: The AMC23 benchmark refers to a specific set of evaluations based on the American Mathematics Competitions. This benchmark is designed to assess the mathematical reasoning capabilities of advanced AI models using problems characteristic of the AMC series. For the evaluation of AMC23, we utilize the resources and methodologies found in the Qwen2.5-Math repository. This repository is accessible at https://github.com/QwenLM/Qwen2.5-Math and provides the framework for our assessment process.

A.3. Compute and hyperparameter configuration details

Compute resources: All of our LLM inference experiments were run either on a single 80GB A100 or a 40GB L40S GPU. The finetuning experiments (SFT and DPO) were run on a single node consisting of 4 A100 GPUs.

Hyperparameter	Value
Dataset	FineWeb sample-10BT
Max sequence length	512
Number of sequences	15000
Batch size	16

Table 5: Hyperparameter configurations used for computing `RankMe` and αReQ in Figure 2.

Hyperparameter	Value
SFT dataset	Anthropic-HH or AlpacaFarm Human-ANN chat (train split)
Max sequence length	4096
Batch size	16
Gradient accumulation steps	16
Learning rate	1e-5
Learning rate schedule	Linear decay with 10% warmup
Number of epochs	2
Loss reduction	sum
Seeds	0, 7, 8, 42, 420

Table 6: Hyperparameter configurations used for Supervised FineTuning (SFT).

Hyperparameter	Value
Base model	OLMo2-1B
In-distribution dataset	Anthropic-HH (test split)
Out-of-distribution dataset	AlpacaFarm Human-ANN chat (train split)
Max sequence length	1024
Number of sequences	10000
Batch size	32

Table 7: Hyperparameter configurations used for ID and OOD loss eval.

A.4. Reproducing Tülu-3-8B SFT and DPO

We follow instructions from [https://github.com/allenai/open-instruct] for reproducing and gathering the intermediate stage checkpoints (both for SFT and DPO) without changing any hyperparamters.