Synchronos LLM Post Training

Elliot Epstein

January 7, 2026

This report summarizes Synchronos LLM post-training across data filtering, supervised fine-tuning, and evaluation. Code for the paper is available at `https://github.com/Elliotepsteino/post-training`.

1 Data Filtering

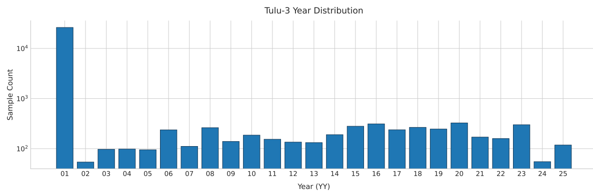
We label each supervised (SFT), preference, and RLVR sample with the minimum calendar year consistent with its question–answer bundle; the prompt structure is detailed in Appendix A. We considered deterministic filtering, but it was difficult to capture all edge cases with a rule-based approach. The latest sweep (session `2026-01-06_14-10PT`) processed 30,549 SFT examples, 29,510 preference examples, and three RLVR datasets (7,358 GSM, 7,372 MATH, 14,958 IFEval) with a conservative policy that uses the most recent referenced year. Figures 1–3 show year and category distributions for each dataset family and validate cutoff integrity.

Filtering Cost. The current filtering pass uses GPT-5-mini with batch requests at \$0.25/\$2.00 per 1M input/output tokens; the batch discount halves these rates to \$0.125/\$1.00. Table 1 summarizes per-sample token averages, current costs, and projections for the full SFT/preference corpus plus the RLVR targets. TÜLU-3 still requires filtering 900k SFT examples and 250k preference examples beyond the current subset; projected costs scale linearly with per-sample token counts. For prompts under 200k tokens, Gemini 3 Flash is priced at \$0.50/\$3.00 per 1M input/output tokens (similar to GPT-5-mini), while Gemini 3 Pro is \$2.00/\$12.00 (similar to GPT-5.2).

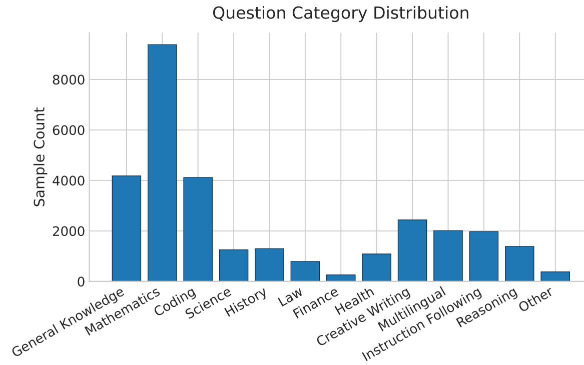
Filtering Evaluation. We sample 50 questions evenly across year shards and compare model conservatism, defined as the tendency to predict a later (higher) year; the most conservative prediction is the maximum year among models for the same prompt. Combining the cheaper models in each family gives the most conservative estimate (Figure 4), and this can be extended to include Anthropic models. Without ground truth, we count how often each model produces the maximum year for the same prompt (ties count). Figure 4 shows per-model counts plus two max-ensembles: Gemini 3 Flash + GPT-5-mini and Gemini 3 Pro + GPT-5.2.

Next step: human labels for the 50 questions. With ground truth, we will report exact accuracy, conservative accuracy (predicted year $\geq$ gold), and weighted accuracy (mean of the two).

Downstream, we select shards with a year-bounded loader to enforce knowledge cutoffs (e.g., 2014).

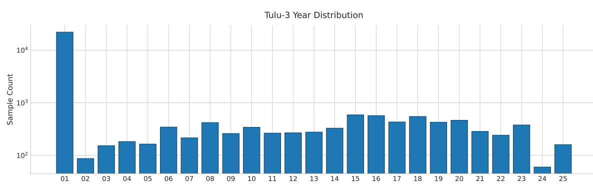


(a) Year distribution (SFT).

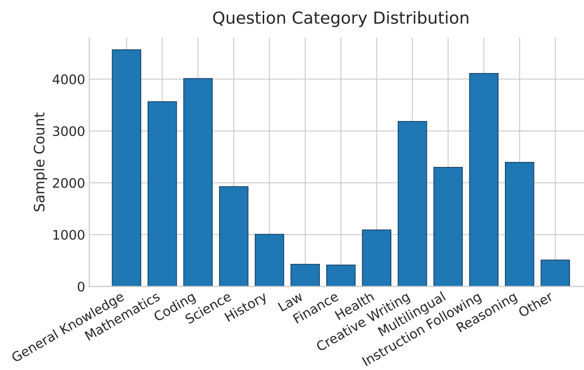


(b) Category distribution (SFT).

Figure 1: Filtering summary for the TÜLU-3 SFT mixture (session `2026-01-06_14-10PT`, $n =$ 30,549).



(a) Year distribution (DPO).



(b) Category distribution (DPO).

Figure 2: Filtering summary for the TÜLU-3 preference mixture (session `2026-01-06_14-10PT`, $n = 29{,}510$).

2 Supervised Fine-Tuning Setup

We fine-tune Qwen3-4B-Base with LoRA adapters on the 2007-capped TÜLU-3 SFT subset (26,431 examples) with a 4,096-token context length. Training runs for two epochs with linear decay, short warmup, and small per-device batches with gradient accumulation. Tokens per second per GPU are around 1,000. The LoRA run takes about 3 hours 2 minutes for 300M tokens (padding included). LoRA uses roughly 45 GB of GPU memory; full fine-tuning is about 10× slower, and TÜLU-2 showed lower LoRA performance.

Hardware. All runs use three NVIDIA RTX A6000 GPUs (48 GB each). The LoRA configuration fits within a single A6000 with limited headroom for data loading and logging.

3 Evaluation

We run the TÜLU-3 dev evaluation suite with `run_tulu3_dev_limit100.sh`, dispatching 11 suites and limiting each task to 100 examples (MMLU uses 100 questions per subject, totaling 5,700 eval-

Dataset	Current n	Tokens/sample	Current cost	Projected n	GPT-5-mini	GPT-5.2	GPT-5.2 Pro
SFT	30,549	1,661	$20.46	930,549	$623	$4.36k	$52.3k
Preference	29,510	2,437	$25.08	279,510	$238	$1.66k	$20.0k
RLVR GSM	7,358	2,167	$6.78	8,790	$8.10	$56.7	$680
RLVR MATH	7,372	1,653	$3.68	7,500	$3.74	$26.2	$315
RLVR IFEval	14,958	1,690	$10.81	15,000	$10.84	$75.9	$910

Table 1: Filtering costs (USD) under batch pricing. Projected counts use 900k/250k additional SFT/preference samples and RLVR targets of 8.79k (GSM), 7.5k (MATH), and 15k (IFeval).

Component	Setting
Base model	Qwen3-4B-Base
Tokenizer	Qwen3-4B-Base tokenizer
Training data	TÜLU-3 SFT, year $\leq$ 2007 (26,431 examples)
Sequence length	4,096 tokens
Batch size	1 sample per device
Gradient accumulation	16 steps (effective 16 samples per device)
Optimizer schedule	Linear decay with 3% warmup
Learning rate	1×10^{-4}
Weight decay	0.0
Epochs	2
LoRA configuration	rank 64, $\alpha = 16$, dropout 0.1
Memory optimizations	Flash attention and gradient checkpointing
Checkpoint cadence	Every 500 steps (keep last 3)

Table 2: Supervised fine-tuning configuration for the LoRA run.

uations). This gives a fast signal across reasoning, coding, alignment, and factuality. Task summaries: GSM8K (grade-school math word problems), DROP (reading comprehension with numeric reasoning), Minerva Math (competition math problems across seven domains), HumanEval/HumanEval+ (Python coding pass@10), IFEval (instruction-following), PopQA (entity-centric factual QA), MMLU (multiple-choice knowledge across 57 subjects), AlpacaEval v2 (pairwise preference wins), BBH (hard reasoning tasks with CoT), TruthfulQA (robustness to falsehoods).

Table 3 summarizes the latest snapshot for Qwen3-4B-Base and placeholder columns for +SFT, +DPO, and +RLVR checkpoints (marked "–"). The SFT evaluation is still running; partial results are shown where available. Scores are percentages, with `n` denoting evaluated examples.

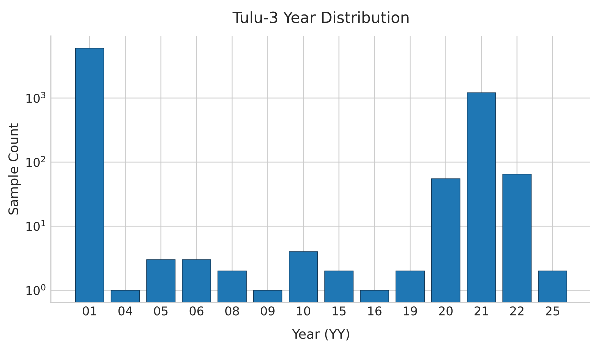
4 Next Steps

1. Iterate on SFT, DPO, and RLVR recipes to improve performance, then validate on the dev suite.
2. Scale filtering to the full TÜLU-3 corpus after recipes are locked.
3. Measure gains from the scaled data against current baselines.
4. Automate per-year dataset exports and launch SFT runs for each year cutoff.

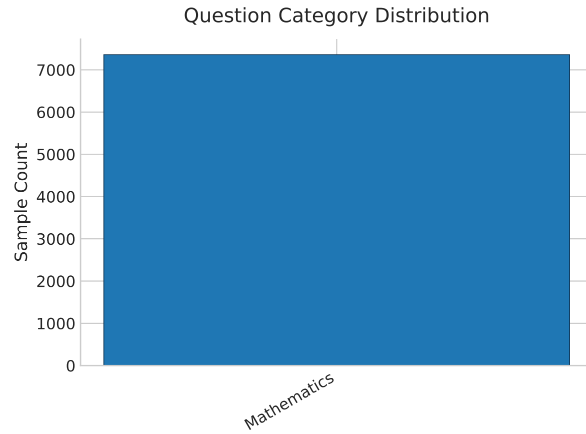
Task	Metric	Qwen3-4B Base	+SFT	+DPO	+RLVR	n
GSM8K	Exact match	83.00	83.00	–	–	100
DROP	F1	52.15	57.16	–	–	100
Minerva Math (avg)	Exact match	39.14	–	–	–	700
HumanEval	pass@10	95.84	97.30	–	–	100
HumanEval+	pass@10	94.80	93.28	–	–	100
IFEval	Prompt loose acc	40.00	43.00	–	–	100
PopQA	Accuracy	17.00	20.00	–	–	100
MMLU (mc)	Macro accuracy	74.46	74.44	–	–	5,700
AlpacaEval v2	Len-ctrl win rate	6.54	–	–	–	100
BBH (cot-v1)	Macro accuracy	–	–	–	–	–
TruthfulQA	MC2	54.48	45.59	–	–	100

Table 3: Primary metrics for the TÜLU-3 dev suite (Qwen3-4B-Base, 100-example subsets). The BBH run is still executing at this scale; results will be inserted once the evaluation completes.

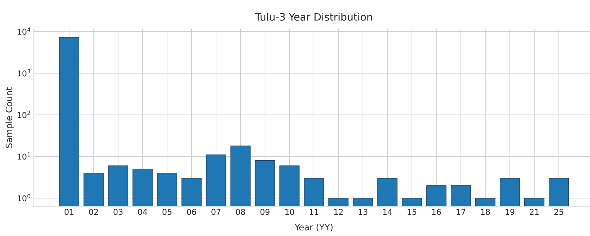
5. Build a data-leakage evaluation: define a held-out provenance set, train models with explicit cutoffs, probe with targeted queries, and report leakage by year and domain.



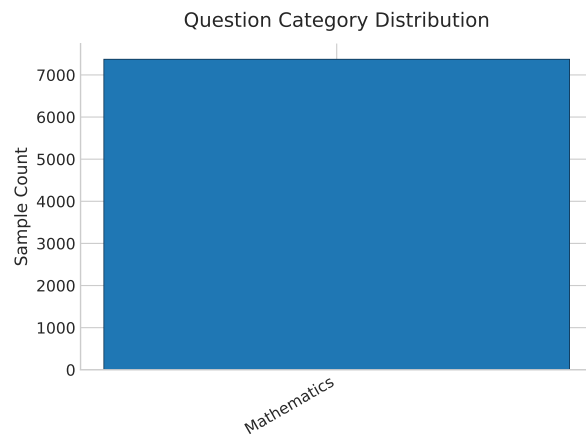
(a) Year distribution (RLVR-GSM).



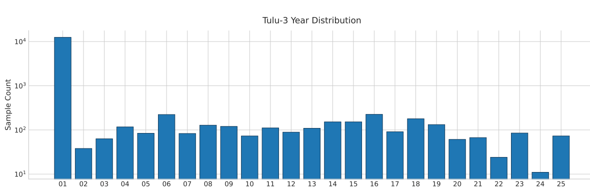
(b) Category distribution (RLVR-GSM).



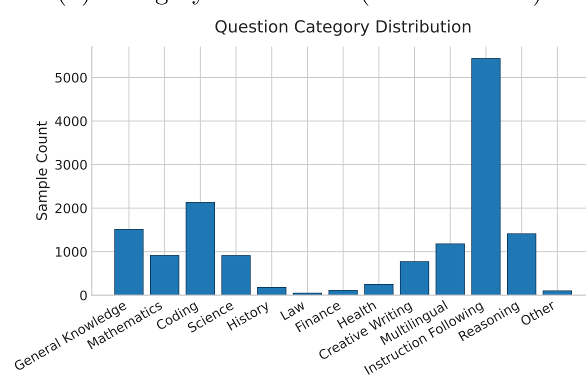
(c) Year distribution (RLVR-MATH).



(d) Category distribution (RLVR-MATH).



(e) Year distribution (RLVR-IFeval).



(f) Category distribution (RLVR-IFeval).

Figure 3: Filtering summary for the RLVR datasets (session `2026-01-06_14-10PT`, GSM $n = 7{,}358$, MATH $n = 7{,}372$, IFEval $n = 14{,}958$).

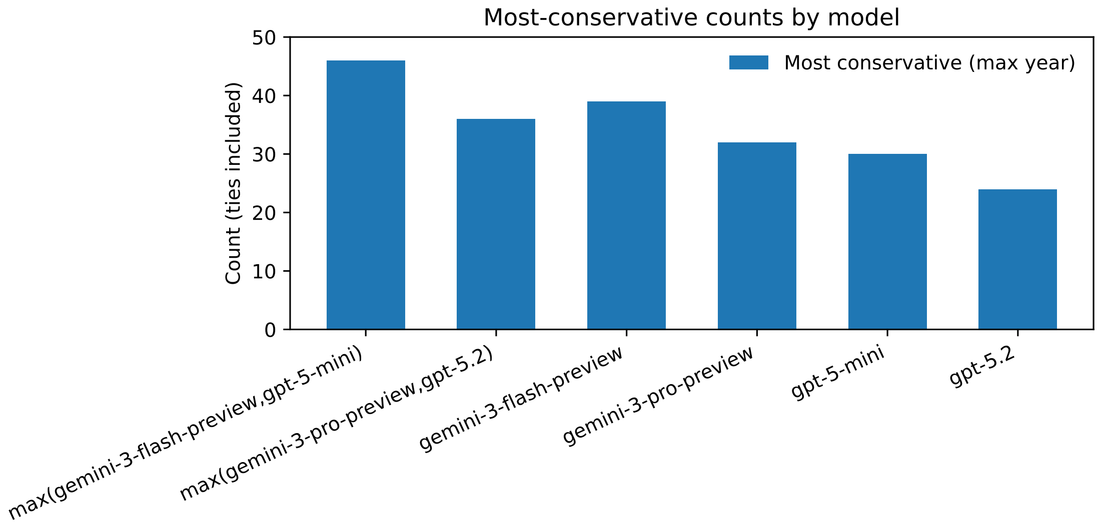


Figure 4: Most-conservative counts over 50 questions, with max-ensembles of Gemini 3 Flash + GPT-5-mini and Gemini 3 Pro + GPT-5.2 (ties count).

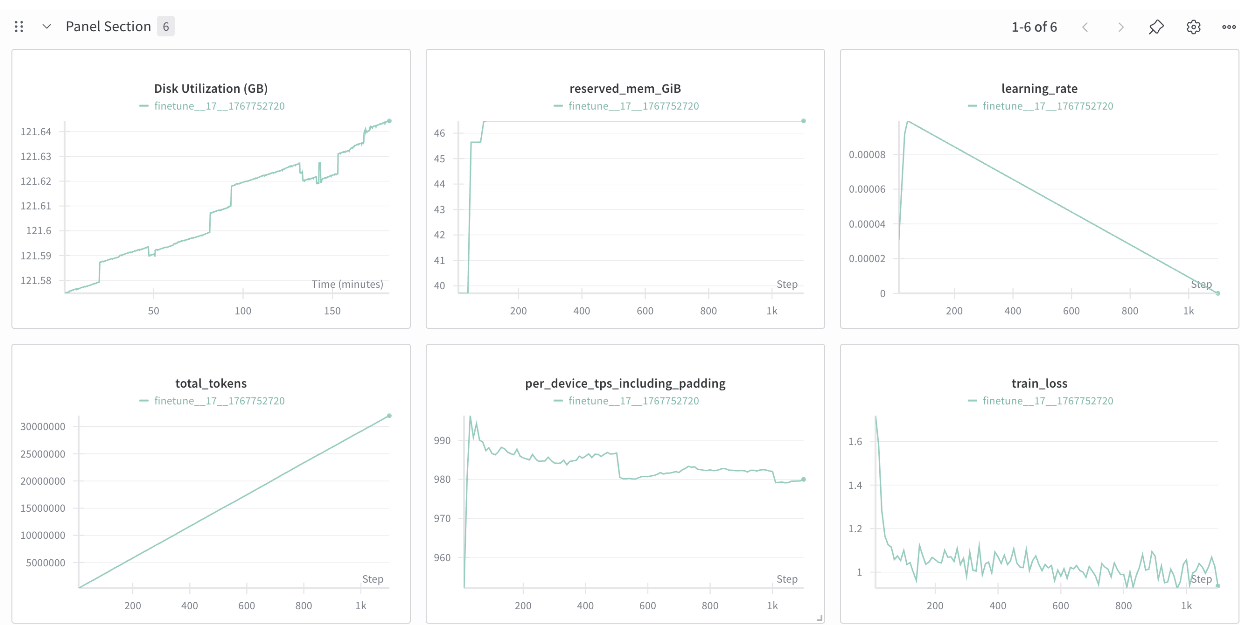


Figure 5: LoRA SFT training metrics: disk utilization (GB), reserved GPU memory (GB), learning rate, total tokens, per-device tokens per second, and train loss.

A SFT Filtering Prompt

```
You label the minimum calendar year (between 2001 and 2025) required to answer
a question without temporal leakage. The label must never precede any fact
mentioned in the sample; when uncertain, err toward the later year so that no
future knowledge sneaks into earlier buckets.

You receive a dataset-specific question plus an answer bundle (which may contain
multiple sections).
These are supervised instruction-tuning pairs: treat the question as the user
prompt and the response as the assistant answer. The label year must satisfy
any facts in either part of the exchange.
Pick the smallest year Y in [2001, 2025] so that a model with knowledge through
year Y could answer confidently, considering EVERYTHING in both the question and
the answer bundle. If no specific time-dependent knowledge is required, output
2001.
Rules:
- Consider publication dates, statistics, laws, releases, and events.
- Output the smallest year that still contains every fact mentioned.
- If the bundle includes multiple responses (e.g., preferred/rejected answers,
  constraints, rationales), the chosen year must satisfy the most recent
  reference anywhere in the bundle.
- If multiple explicit years are referenced, return the most recent explicit
  year.
- If only a range or uncertainty is provided (e.g., "released between 2008 and
  2015"), answer with the latest year in that range so no future facts are
  included.
- If information is older than 2001, still respond with 2001.
- Do not hallucinate years that are not grounded in the text.
- Additionally, assign the question to one category from this list: coding,
  creative_writing, finance, general_knowledge, health, history,
  instruction_following, law, math, multi_lingual, other, reasoning, science.

Illustrative example:
Question:
"Teacher: In this task, you are given a text from tweets and a boolean question
whether this tweet has positive sentiment or negative sentiment. Your task is to
generate answer "yes" when the tweet has that particular sentiment, otherwise
generate answer "no".\nTeacher: Now, understand the problem? If you are still
confused, see the following example:\nTweet: @justinchuan Awww! I was thinking
about you lot up there! Glad you enjoyed it Question: is it a positive tweet?\n
Solution: yes\nReason: There is an expression of happiness in this tweet text,
hence, we can say it’s positive. So answer is ’yes’.\n\nNow, solve this instance:
Tweet: Goddamn my back hurts this morning.  Question: is it a positive tweet?\n
Student:"
Answer JSON:
{"year": 2006, "confidence": "high", "category": "general_knowledge",
"justification": "Answer references tweets, a concept only available after
```

Dataset	Prompts	License
CoCoNot	10,983	ODC-BY-1.0
FLAN v2 (ai2-adapt-dev/flan_v2_converted)	89,982	–
No Robots	9,500	CC-BY-NC-4.0
OpenAssistant Guanaco	7,132	Apache 2.0
Tulu 3 Persona MATH	149,960	ODC-BY-1.0
Tulu 3 Persona GSM	49,980	ODC-BY-1.0
Tulu 3 Persona Python	34,999	ODC-BY-1.0
Tulu 3 Persona Algebra	20,000	ODC-BY-1.0
Tulu 3 Persona IF	29,980	ODC-BY-1.0
NuminaMath-TIR	64,312	Apache 2.0
Tulu 3 WildGuardMix	50,000	Apache 2.0
Tulu 3 WildJailbreak	50,000	ODC-BY-1.0
Tulu 3 Hardcoded	240	CC-BY-4.0
Aya	100,000	Apache 2.0
WildChat GPT-4	100,000	ODC-BY-1.0
TableGPT	5,000	MIT
SciRIFF	10,000	ODC-BY-1.0
Evol CodeAlpaca	107,276	Apache 2.0

Table 4: TÜLU 3 SFT mixture composition. Source details: Brahman et al. (2024), Longpre et al. (2023), Rajani et al. (2023), Kopf et al. (2024), Beeching et al. (2024), Han et al. (2024), Wildteaming (2024), Singh et al. (2024), Zhao et al. (2024), Zha et al. (2023), Wadden et al. (2024), Luo et al. (2023).

```
Twitter launched in 2006, so 2006 is the earliest safe year.",
"evidence_years": [2006]}

Use the same reasoning style for the sample below and respond with compact JSON
only.

<question>
{sample.question}
</question>
<answer_bundle>
{sample.answer}
</answer_bundle>
Return JSON exactly in this schema:
{"year": 2001, "confidence": "low|medium|high",
"category": "one of the allowed categories",
"justification": "why year is required", "evidence_years": [2008]}
```

SFT Mixture Data. The TÜLU 3 SFT mixture used for training contains 939,344 samples from the sources below.