Synchronos LLM Post Training

Post-Training Team

January 5, 2026

1 Evaluation

We run the TÜLU-3 dev smoke test with `run_tulu3_dev_limit100.sh`, dispatching 11 validation suites across the local A6000 fleet and constraining each task to a 100-example slice (MMLU uses 100 questions per subject, yielding 5,700 evaluations). This regimen provides a fast signal on diverse reasoning, coding, alignment, and factuality workloads while keeping GPU time manageable.

Table 1 summarizes the latest snapshot for Qwen3-4B-Base as well as placeholder columns for upcoming +SFT, +DPO, and +RLVR checkpoints (marked "–" until those evaluations finish). Scores are reported as percentages, with `n` denoting the evaluated examples per task.

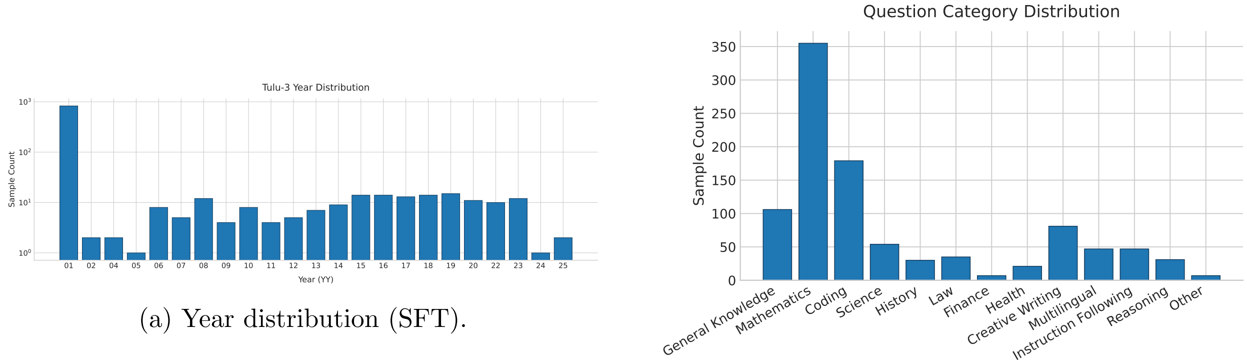
2 Data Filtering

We label every supervised (SFT), preference (DPO), and RLVR sample with the minimum calendar year consistent with its question–answer bundle. The latest sweep (session `2026-01-05_14-31PT`) processed 1,000 examples per corpus, enforcing a conservative policy that forces the most recent referenced year to dominate. Figures 1–3 show both year and category distributions for each dataset family; these plots make it easy to confirm that earlier buckets remain uncontaminated while still exposing modern content when desired.

Downstream, we select shards via `YearBoundedTuluLoader` to enforce any desired knowledge cutoff (e.g., 2014). This combination of automated year tagging, per-dataset plots, and conservative labeling has kept Synchronos LLM post-training leak-free while enabling rapid iteration on evaluation benchmarks.

Task	Metric	Qwen3-4B Base	+SFT	+DPO	+RLVR	n
GSM8K	Exact match	83.00	–	–	–	100
DROP	F1	52.15	–	–	–	100
Minerva Math (avg)	Exact match	47.00	–	–	–	200
HumanEval	pass@10	95.84	–	–	–	100
HumanEval+	pass@10	94.80	–	–	–	100
IFEval	Prompt loose acc	40.00	–	–	–	100
PopQA	Accuracy	17.00	–	–	–	100
MMLU (mc)	Macro accuracy	74.46	–	–	–	5,700
AlpacaEval v2	Len-ctrl win rate	6.54	–	–	–	100
BBH (cot-v1)	Macro accuracy	–	–	–	–	–
TruthfulQA	MC2	54.48	–	–	–	100

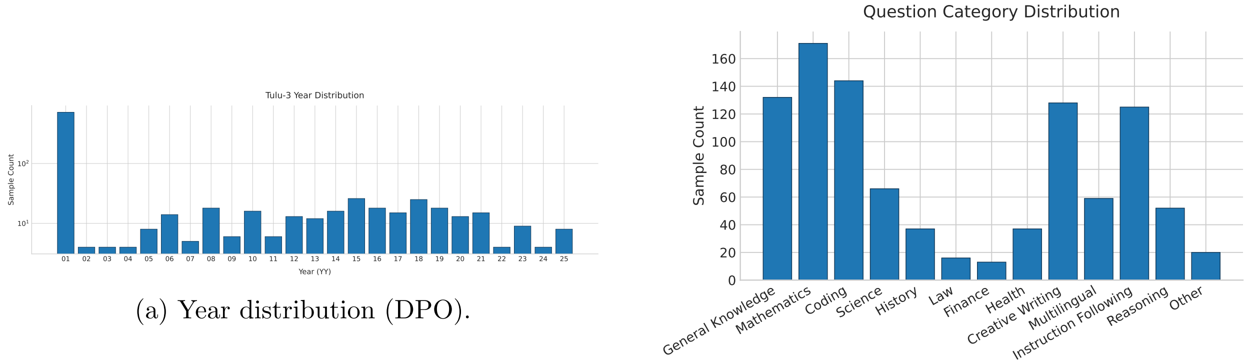
Table 1: Primary metrics for the TÜLU-3 dev suite (Qwen3-4B-Base, 100-example subsets). The BBH run is still executing at this scale; results will be inserted once the evaluation completes.



(a) Year distribution (SFT).

(b) Category distribution (SFT).

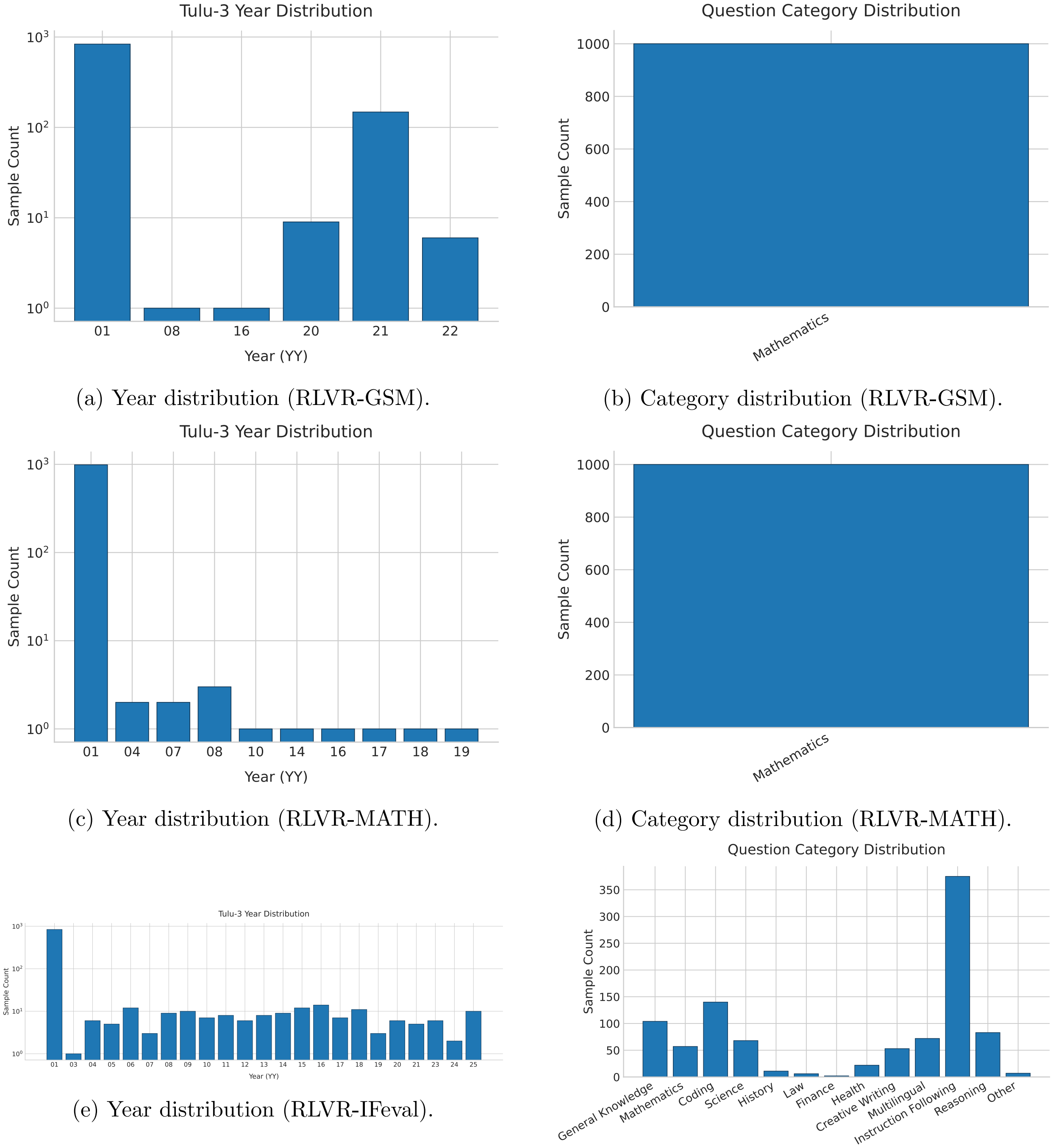
Figure 1: Filtering summary for the TÜLU-3 SFT mixture (session `2026-01-05_14-31PT`, $n = 1000$).



(a) Year distribution (DPO).

(b) Category distribution (DPO).

Figure 2: Filtering summary for the TÜLU-3 DPO mixture (session `2026-01-05_14-31PT`, $n = 1000$).



(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 `2026-01-05_14-31PT`, $n = 1000$ per split).