

DATA FOR POST-TRAINING

2026F INF385T: Post-Training of LLMs | Week 7

CHAN YOUNG PARK

Assistant Professor, iSchool, School of Computing, The University of Texas at Austin

Announcements

- HW1 is due October 15, 11:59pm
- HW1: data curation and LoRA vs. full fine-tuning

Recap

Week 5: which tokens receive loss? → objective

Week 6: which parameters receive updates? → parameters

Week 7: which behaviors receive demonstrations? → examples

Week 7 overview

- **Data curation**

LIMA: when can a small curated set work? ([Zhou et al., 2023](#))

RDS+: which examples to pick from a huge pool? ([Iverson et al., 2025](#))

- **Synthetic generation**

Magpie, personas: where do diverse prompts come from? ([Xu et al., 2025](#); [Chan et al., 2024](#))

- **Model distillation**

R1, Qwen3: how does a student learn from a teacher? ([DeepSeek-AI, 2025](#); Qwen Team, 2025)

- **In practice**

Olmo 3: how do the pieces fit into one pipeline? ([Olmo Team, 2025](#))

Where demonstrations come from

- People write or curate answers: LIMA (curation)
- Existing tasks become prompt–response pairs (synthetic generation)
- Models generate prompts, answers (synthetic generation, distillation)
- Distillation trains a student with a teacher’s signal (distillation)

Synthetic data and distillation often describe the same examples.

LIMA: when can a small curated set work?

RDS+: which examples to pick from a huge pool?

DATA CURATION

LIMA: what can 1000 examples teach?

LIMA: Less Is More for Alignment

Chunting Zhou^{μ*} Pengfei Liu^{π*} Puxin Xu^μ Srini Iyer^μ Jiao Sun^λ

Yuning Mao^μ Xuezhe Ma^λ Avia Efrat^τ Ping Yu^μ Lili Yu^μ Susan Zhang^μ

Gargi Ghosh^μ Mike Lewis^μ Luke Zettlemoyer^μ Omer Levy^μ

^μ Meta AI

^π Carnegie Mellon University

^λ University of Southern California

^τ Tel Aviv University

LIMA: what can 1000 examples teach?

- Starting point: pretrained LLaMA-65B
- Intervention: SFT on 1K curated demonstrations
- Superficial Alignment Hypothesis:
Pretraining supplies capability. Alignment teaches response formats

The starting model is part of the explanation

what did the authors curate?

1K examples with varied inputs and a consistent assistant style

Source	#Examples	Avg Input Len.	Avg Output Len.
Training			
Stack Exchange (STEM)	200	117	523
Stack Exchange (Other)	200	119	530
wikiHow	200	12	1,811
Pushshift r/WritingPrompts	150	34	274
Natural Instructions	50	236	92
Paper Authors (Group A)	200	40	334
Dev			
Paper Authors (Group A)	50	36	N/A
Test			
Pushshift r/AskReddit	70	30	N/A
Paper Authors (Group B)	230	31	N/A

Examples (curated)

[Stack Exchange (STEM)]

What is the difference between minimum and infimum? I have a great confusion about this.

The minimum is attained, the infimum isn't necessarily.

Example: Let $f(x) = \frac{1}{x}$. Then f has no minimum value on the interval $(0, \infty)$. The minimum is the smallest element in the set. That is $\min\{f(x) \mid x \in (0, \infty)\}$ doesn't exist because there is not smallest number in the set. Another example is the minimum of the set $S = (0, 1) = \{x \mid 0 < x < 1\}$. Here again there isn't a smallest number $\min\{x \mid 0 < x < 1\}$ doesn't exist.

The infimum of a set S is defined as the greatest number that is less than or equal to all elements of S (from Wikipedia). The infimum is also sometimes called the greatest lower bound. It is a fact that every non empty set (bounded below) of real numbers has an infimum. But, as we saw, not every real set has a minimum. So in the example $\inf\{f(x) \mid x \in (0, \infty)\} = 0$.

[Stack Exchange (Other)]

Was the Millennium Falcon a one-off or was it mass produced?

The Millennium Falcon is not unique; it was a model YT-1300f Corellian light freighter. This implies that there are other similar ships, including other variations on the YT-1300 model. You can find a lot of information about the Millennium Falcon in a companion book, Star Wars: The Force Awakens: Incredible Cross-Sections. The model number also appears on-screen at least once: in Star Wars: Rebels episode "Fighter Flight" on a poster. Note, however, that the Millennium Falcon has been highly customized; some of the custom work is mentioned in The Force Awakens. It's not clear just how much of the ship is stock and how much is custom.

If you're willing to go outside Canon into Legends, the model number of the ship, and other similar Corellian light freighters, can be seen in some of the video games, and there is an entire novel called Millennium Falcon that details its history. The ship was serial #YT-1300 492727ZED, one of many

[wikiHow]

How to be a lazy college student?

In college, unlike in high school, you are expected to take responsibility for your own learning. It is ultimately up to you to determine how much time and effort you want to put into your schoolwork. Experts suggest that college students today work considerably fewer hours than their counterparts did in 1961. But there is also evidence that working more hours is not necessarily the best way to be productive. A little laziness may even be good for us.

Part 1 of 3: prioritize what's important

1. Ask yourself what you want from college. There are many reasons why a college degree is desirable, not least that it is still considered highly valuable by many employers. Your own reasons for pursuing a college degree may be more personal, but spend some time working out exactly what these reasons are. Consider why college is important to you. Do you want to learn important skills like critical thinking and problem-solving? Do you want

Examples (manually authored)

[Manually Authored – Chitchat]

Tell me an interesting fact about geography.

Here are some interesting facts about geography:

- Bhutan is one of the most intriguing places in the world. The Buddhist kingdom is the only carbon-negative country on Earth, with strict environmental laws and cultural norms that preserve its stunning natural landscape. For similar reasons, Bhutan was one of the last countries to introduce television in 1999, and has no traffic lights to this day.

• There are currently four territories on

[Manually Authored – Advice]

I'm going to NeurIPS for the first time to present the first paper I've ever published. I'm afraid I'll be alone and overwhelmed by everyone. What should I do?

Going to an academic conference for the first time may seem like a bewildering experience, but if done right, conferences are a wonderful opportunity to brainstorm with experts from around the world, present your work and get feedback, and make new friends. Here are some tips for improving your conference experience:

- Reach out in advance to people who have inspired your work. [...]

[Manually Authored – Writing]

I am planning to start a book club with some friends. Can you write an email invitation for the same?

Here is an example email invitation for a book club:

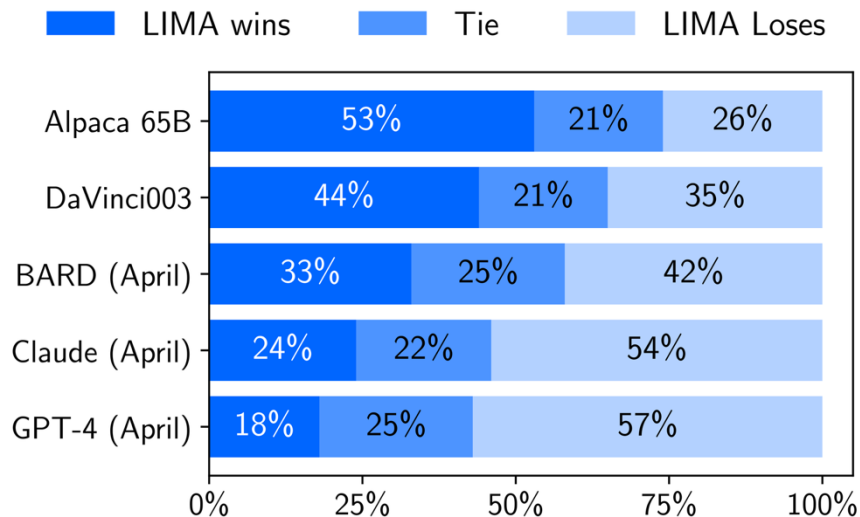
Subject: Join me for a book club!

Hi [Friend's Name], how are you?

I'm starting a book club with some friends, and I'd love to have you join us!

The book club is a casual and fun way to read and discuss books that we love or want to

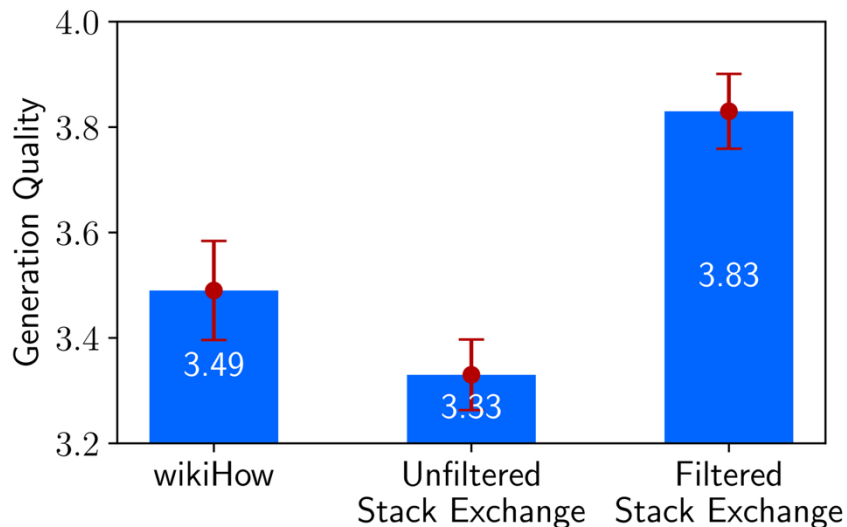
human preference evaluation



- 65B model, 300 test prompts.
- Against GPT-4: 18% wins, 25% ties.
- Interpret within this evaluation.

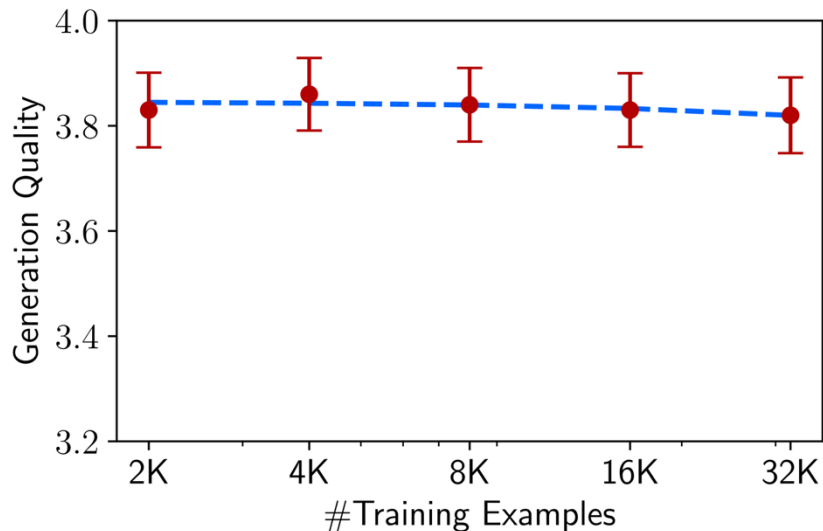
vs. GPT-4: 43% combined wins and ties
 vs. DaVinci003: 65% combined wins and ties

Quality and diversity ablations



- Llama 7B w/ 2K examples
- LLM-judge
(GPT-3.5 helpfulness scores 1-6)
- WikiHow: low diversity, high quality
Unfiltered SE: high diversity, low quality
Filtered SE: high diversity, high quality
- Filtered Stack Exchange does better.

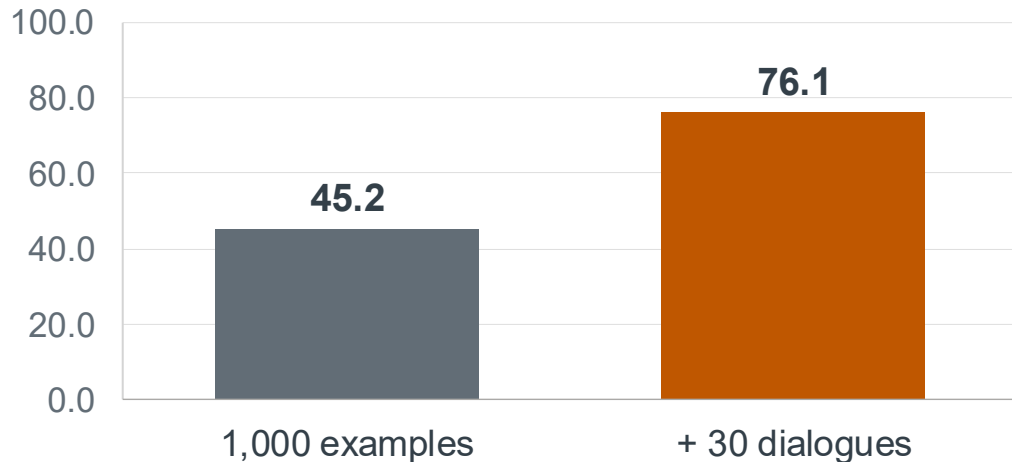
Quantity alone reaches a plateau



- The source stays Stack Exchange
- The score changes little with more examples
- More rows from one distribution may add little new supervision
→ What new behavior would another batch add?

30 dialogue examples change coverage

Excellent dialogue turns (%)



- Setup: 30 multi-turn chains (10 author-written, 20 adapted from SE threads) → SFT w/ 1030
- Multi-turn evaluation: 10 conversations per model
- Single-turn data leaves a gap: model keeps context but ignores follow-up requests, often repeating its previous answer
- Dialogue chains demonstrate turn-to-turn behavior

Right about behavior (but not capability)

- **LIMA showed:** 1K curated examples wo/ RLHF can beat Alpaca (52K) and DaVinci003, Diversity + quality > quantity (30 examples unlock new behaviors)
- **Still holds:** Curation over scale is now standard practice, Small-data elicitation works for reasoning too (s1, LIMO)
- **Limitations / no longer true**
 - Evaluation favored style: 300 single-response prompts, a judge rubric rewarding headings/lists (no math, code, agentic tasks)
 - Post-training scales after all: performance on reasoning tasks keeps improving with more data ([Raghavendra et al., 2024](#)), now people use millions of examples + large-scale RL
 - "Superficial" cuts both ways: safety alignment is shallow and easily removed by fine-tuning ([Qi et al., 2024](#), [2025](#))
- **Open question:** does RL teach new skills or only elicit existing ones?

RDS+ (Iverson et al., 2025)

Large-Scale Data Selection for Instruction Tuning

**Hamish Iverson^{1,2}, Muru Zhang^{1,3}, Faeze Brahman²,
Pang Wei Koh^{1,2}, Pradeep Dasigi²**

¹University of Washington, ²Allen Institute for AI, ³USC

RDS+ (Representation-based Data Selection)

- LIMA curated by hand. At millions of examples, selection must be automatic
- Pool D: 200K or 5.8M instruction examples
- Query V: labeled dev examples for the target task
- Select 10K examples. Fully fine-tune Llama 2 7B for two epochs
- Compare against both ordinary and source-balanced random

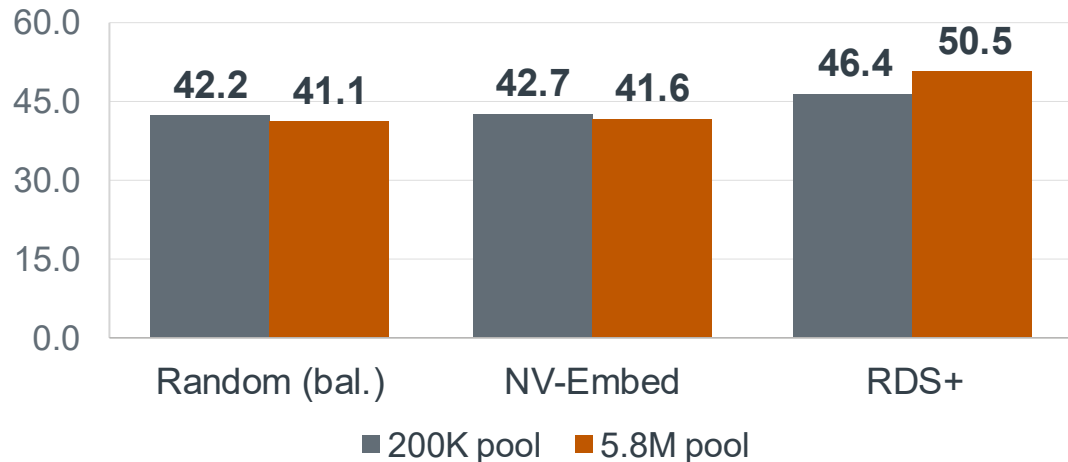
Method: similarity in the student's representation

- Encode each prompt–response pair with the base student
- Pool its final-layer hidden states into one representation
- Score cosine similarity to the labeled query examples
- Select in rounds across queries or tasks, removing duplicates.

Relevance comes from a target distribution, not a universal quality score

Selector performance as the pool grows

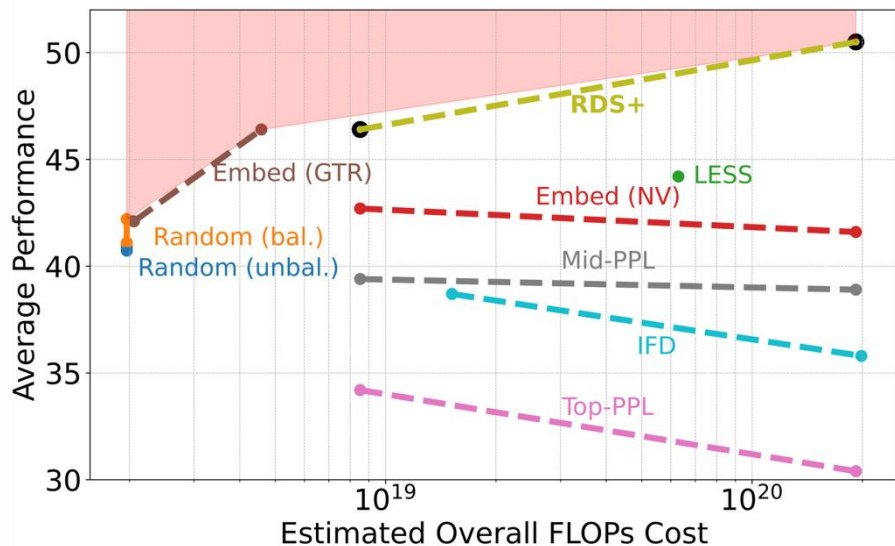
Average across task-specific runs



- 10K selected in every run.
- RDS+ improves as the pool grows
- Random and NV-Embed slightly drop with the larger pool.

A bigger pool helps only if the selector can find the useful examples

Selection quality includes the cost of selection



- Scoring every candidate has a cost
- RDS+ beats random in these tested settings
- Popular quality heuristics (PPL, IFD) fall below random at this scale

Magpie and PersonaHub: where do diverse prompts come from?

SYNTHETIC GENERATION

Which part of the example is synthetic?

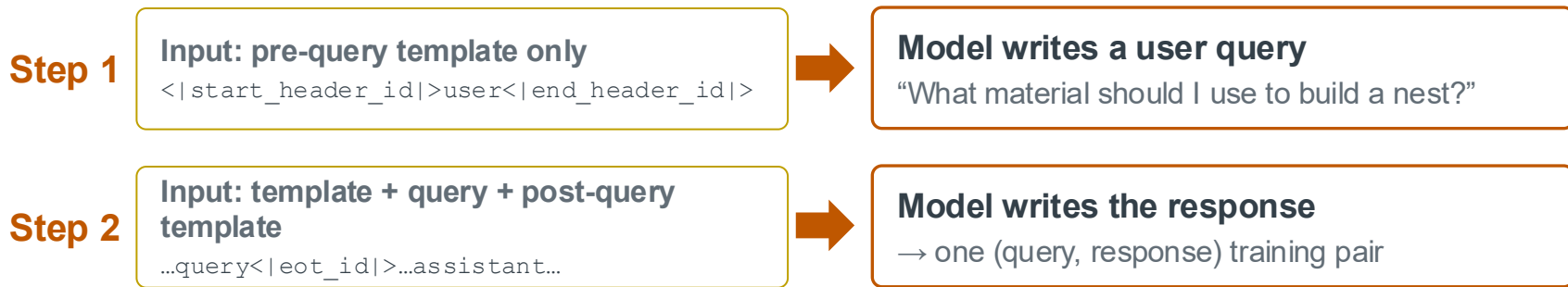
- Self-Instruct ([Wang et al., 2023](#)) expands instructions from seed tasks
- Magpie ([Xu et al., 2025](#)) samples a user request from a chat prefix
- Response generation answers an existing prompt
(often a teacher = distillation)
- PersonaHub prompting conditions each prompt on a described user
([Chan et al., 2024](#))

Magpie: self-synthesis for generating alignment data

MAGPIE: ALIGNMENT DATA SYNTHESIS FROM SCRATCH BY PROMPTING ALIGNED LLMs WITH NOTHING

Zhangchen Xu♣ **Fengqing Jiang**♣ **Luyao Niu**♣ **Yuntian Deng**◇
Radha Poovendran♣ **Yejin Choi**♣◇ **Bill Yuchen Lin**◇
♣University of Washington ◇Allen Institute for AI

the aligned model writes the user turn



- No seed questions and no prompt engineering: only the chat template
- It works because instruction tuning taught the model what user turns look like

The chat template itself becomes a prompt generator

generate cheaply, then filter

Generate

Llama-3-8B-Instruct → 3M pairs (Air)

Llama-3-70B-Instruct → 1M pairs (Pro)

206 + 614 GPU-hours, no humans, no API



Tag

An LLM labels task, quality, and difficulty of each instruction

A reward model scores each response

Embedding distance flags near-duplicates



Filter

Keep clear, non-duplicate instructions with good responses

Then take the longest responses

→ 300K-example filtered

- Pro (evaluated on Arena-Hard):
filtered 300K 18.9 vs. raw 300K 15.9 vs. raw 1M 16.7

Generation is cheap. selection is where the curation happens.

SFT-only data rivals the official recipe

All rows Llama-3-8B base

Training data	Size	AlpacaEval 2 LC	Arena-Hard
Self-Instruct (Llama-3 generator)	100K	7.2	4.0
WildChat (real user prompts)	652K	14.6	8.7
UltraChat SFT + UltraFeedback DPO	208K + 64K	18.4	14.8
Magpie-Pro, filtered (SFT only)	300K	25.1	18.9
Llama-3-8B-Instruct (official)	>10M, SFT + DPO	22.9	20.6

what the prompts inherit

- Prompts follow the generator's idea of a user: over half are information seeking
- Weak on reasoning: GSM8K 47.9 vs. 71.7 for the official model, patched with a 150K math/code “booster” set
- Needs open weights and raw access to the chat template
- Under 1% flagged unsafe by Llama-Guard-2, but filtering is still required

Magpie samples the generator's users, not real ones.

PersonaHub: persona for synthetic data generation

Scaling Synthetic Data Creation with 1,000,000,000 Personas

Tao Ge*, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, Dong Yu

Tencent AI Lab Seattle

<https://github.com/tencent-ailab/persona-hub>

diversity from the prompt side

One synthesis prompt

“Create a math problem with {persona}.”

Persona Hub: ~1B personas
inferred from web text



A nurse at a children’s hospital

→ a dose-by-body-weight problem

A computational linguist

→ a word-frequency probability problem

A cab driver in a large city

→ a fare-and-route optimization problem

- 1.07M GPT-4 persona math problems
- Qwen2-7B reaches 64.9% on MATH (Qwen2-7B-Instruct: 49.6%)

Diversity can be automated using persona-conditioned prompting

Personas in a real recipe: Tülu 3

Skill	Persona SFT data	How prompts are built
Precise IF	29,980	Persona + one verifiable constraint example (25 IFEval constraint types)
Math	~220K	Zero-shot, persona-specific problems
Code	~35K	Zero-shot, persona-specific problems

- ~250K Persona Hub personas; GPT-4o writes skill-targeted prompts
- The same idea returns as “Persona IF” in Olmo 3

R1 and Qwen3: how does a student learn from a teacher?

MODEL DISTILLATION

Sequence distillation uses teacher text

Teacher generates y for a prompt x .

$$L = - \sum_t \log p_s(y_t \mid x, y_{<t})$$

Example target:

X: “24 apples, 6 per box. How many boxes?”

Y: “24 ÷ 6 = 4. The answer is four boxes.”

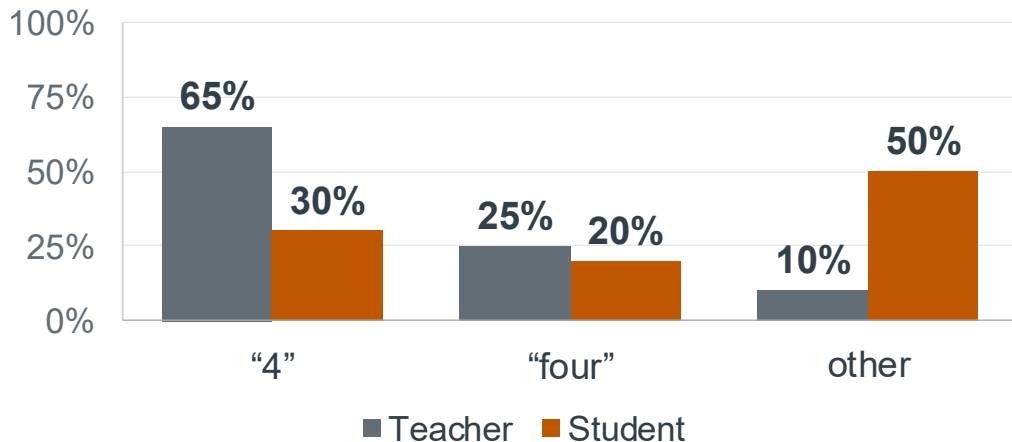
Train on teacher’s final answers, explanations, reasoning traces

The SFT objective stays the same.

The author of the target changes.

Logit distillation uses teacher probabilities

next-token distribution at the same prefix



$$L = \sum_t \text{KL}(q_t \parallel p_t)$$

q_t : teacher probabilities

p_t : student probabilities

Requires

- probability access
- compatible vocabularies for direct token matching

The teacher supplies information about alternatives at each prefix

Off-policy and on-policy choose different prefixes

Off-policy: teacher-generated trajectory

Teacher prefix: “ $24 \div 6 = 4$, so ...”

Student learns on prefixes supplied in the dataset/or by the teacher

On-policy: student-generated trajectory

Student prefix: “ $24 \div 6 = 5$, so ...”

Teacher feedback reaches prefixes the student actually generates

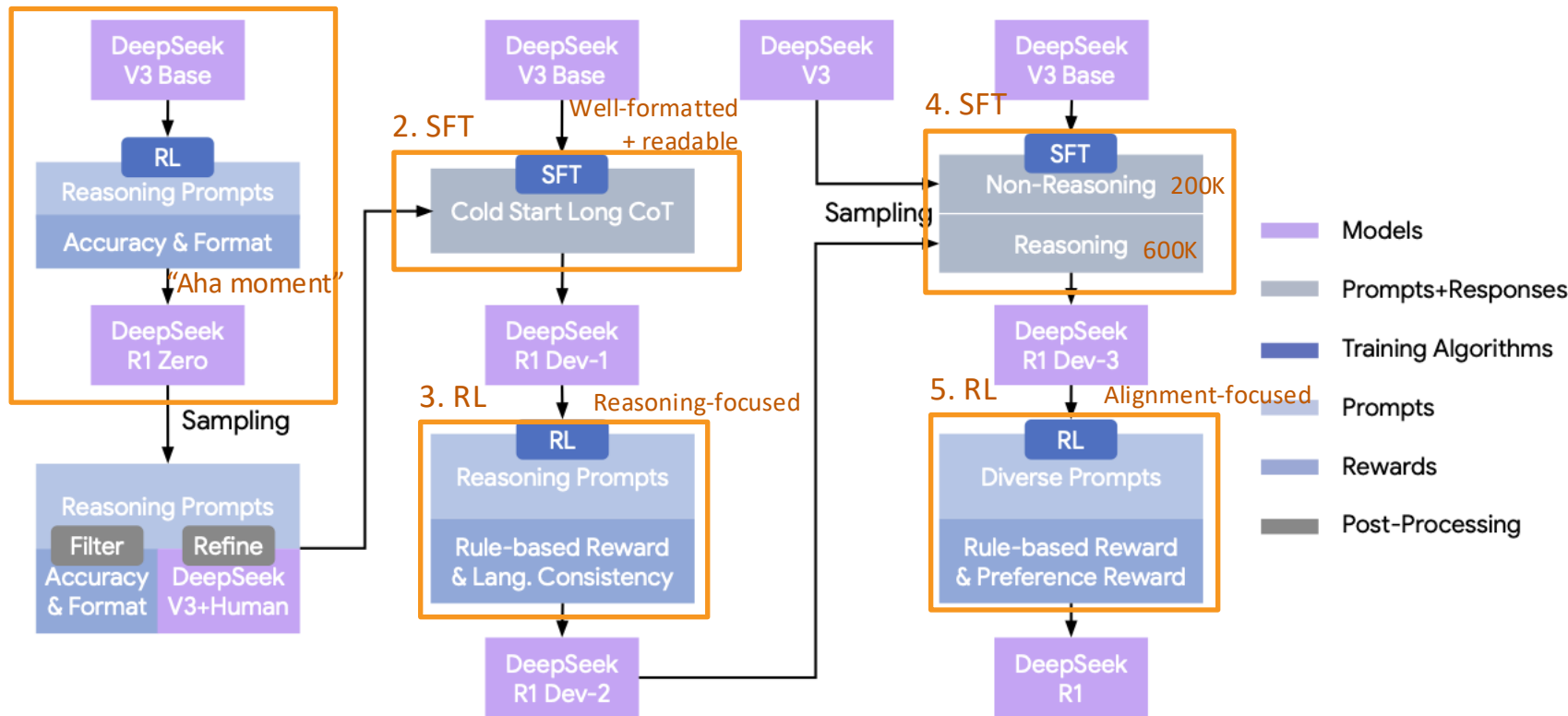
The signal type and the source of prefixes are separate decisions.

DeepSeek-R1 (DeepSeek-AI, 2026)

DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning

1. RL Objectives: (1) accuracy, (2) format ("**<think>**")

DeepSeek-AI



R1: what is inside the 800K training set?

600K reasoning examples

Math, code, and other reasoning prompts.

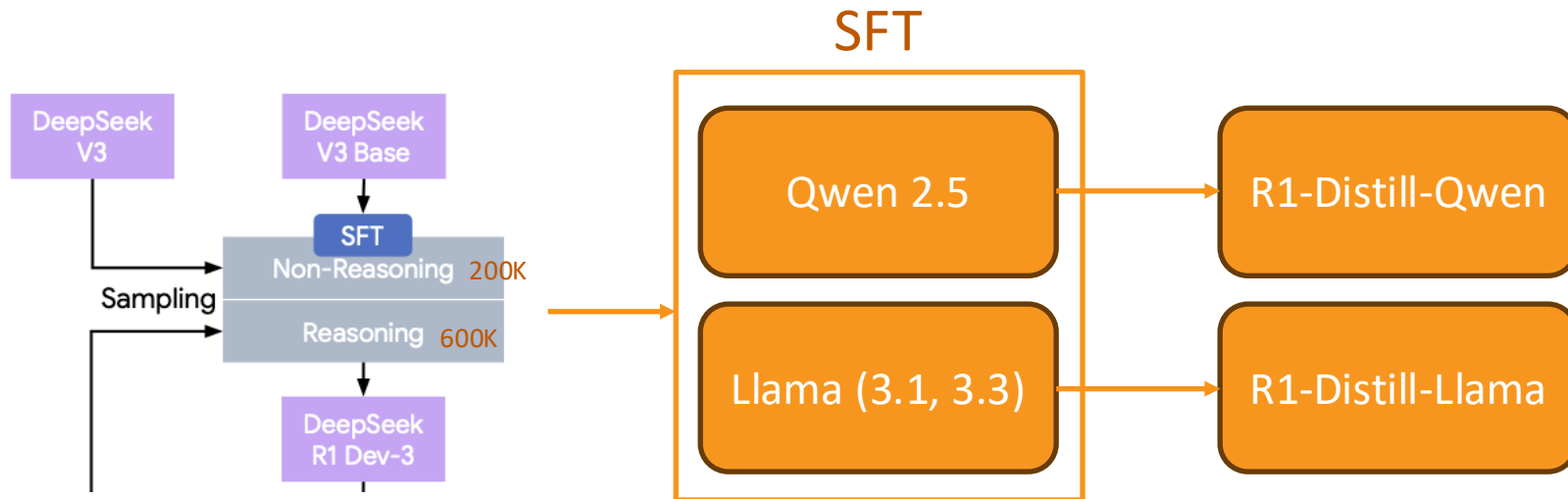
Sample multiple responses. Retain checked solutions.

200K non-reasoning examples

Writing, factual QA, translation, and general assistant behavior.

Fine-tune Qwen and Llama students on this text: SFT only, no RL.

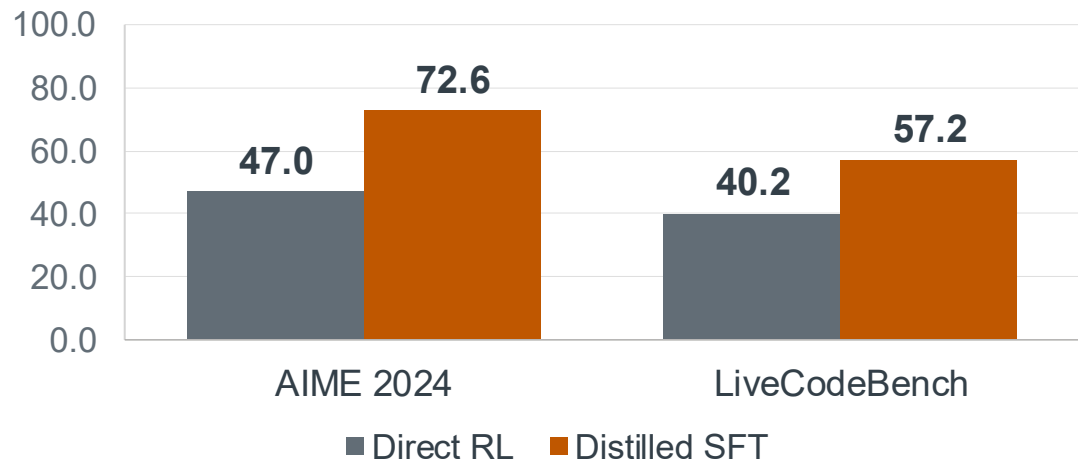
R1-Distill series



Qwen 1.5B model surpassed GPT-4o, Claude 3.5 Sonnet in Math

SFT distillation vs. RL on a 32B base

Published pass@1 scores



- Same Qwen2.5-32B base
- RL: >10K steps
Distilled: SFT only
- Caveat: teacher training cost is not included

(1) model distillation is very effective

(2) Can save a lot of compute compared to relying on RL (from base)

(3) however, if you still need to be the “best model”, you still need large-scale RL

Superficial Alignment Hypothesis vs R1-Distill

- SAH: SFT only imitates the style
vs. R1-distill: models genuinely become better
- Seems to contradict. Why?
- There are three differences
 - What you transfer: R1 not only trains the student model on the “answer” but also the entire “reasoning traces”. Students can learn teacher’s logic/process/reasoning patterns
 - Filtering: the data contains “verified” traces
 - Evaluated on reasoning tasks. Less reliance on “style”

Trends after R1

- Small high-quality distillation with reasoning traces (s1, LIMO)
- “Small models struggle to learn from strong reasoners”
(Li et al., 2025)
- ignited a fierce global debate centered around the legal, ethical, and contractual boundaries of using LLM outputs to train competing models
- On-policy distillation → Qwen 3

Qwen 3

Qwen3 Technical Report

Qwen Team



<https://huggingface.co/Qwen>

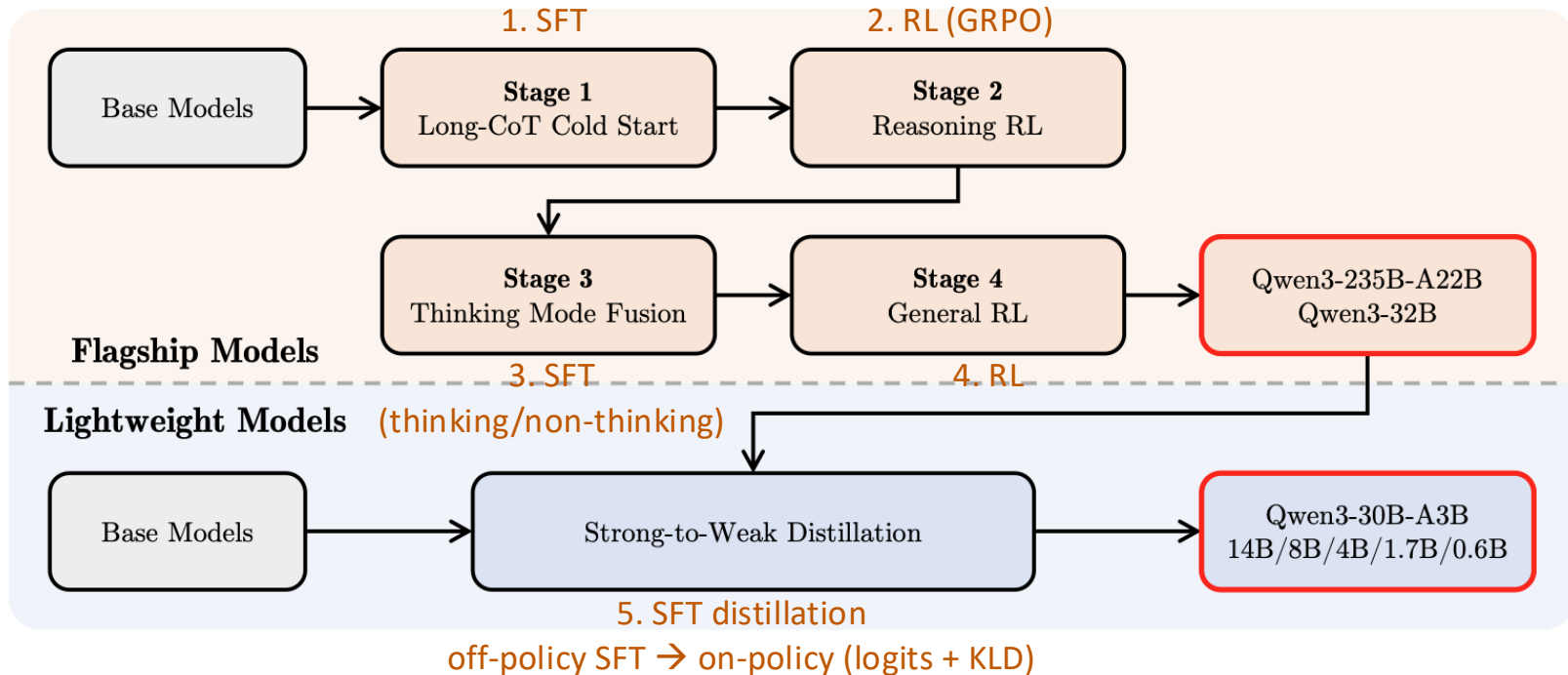


<https://modelscope.cn/organization/qwen>



<https://github.com/QwenLM/Qwen3>

off-policy, then on-policy distillation



Why off-policy $\rightarrow$ on-policy?

- Off-policy: students learn from teacher-generated texts
- On-policy: Student generates a text and the teacher scores its prefixes
- If you do only on-policy, the student's responses are too noisy/bad and teacher's guidance gets wasted
- Off-policy supplies a starting point for on-policy learning

compare the same 8B starting point

Additional training	AIME '24	LCB v5	GPU-hours
None	55.0	42.0	—
RL	67.6	52.9	17,920
On-policy distillation	74.4	60.3	1,800

Both interventions start from the same off-policy-distilled checkpoint

Reported stage cost: 1800 vs. 17920 GPU-hours

Distillation choices for your student

- Text-only teacher access supports response SFT
- Logic access enables distribution matching
- Student-generated trajectories expose its own mistakes
- Compare transfer gains, inference cost, teacher cost.

Olmo 3: how do the pieces fit into one pipeline?

IN PRACTICE

Olmo 3



Olmo 3

Olmo Team*

Allyson Ettinger^{♥1} Amanda Bertsch^{♥1,3} Bailey Kuehl^{♥1} David Graham^{♥1}
 David Heineman^{♥1} Dirk Groeneveld^{♥1} Faeze Brahman^{♥1} Finbarr Timbers^{♥1}
 Hamish Ivison^{♥1,2} Jacob Morrison^{♥1,2} Jake Poznanski^{♥1} Kyle Lo^{♥1,2} Luca Soldaini^{♥1}
 Matt Jordan^{♥1} Mayee Chen^{♥1,4} Michael Noukhovitch^{♥1,5,6} Nathan Lambert^{♥1}
 Pete Walsh^{♥1} Pradeep Dasigi^{♥1} Robert Berry^{♥1} Saumya Malik^{♥1} Saurabh Shah^{♥1}
 Scott Geng^{♥1,2} Shane Arora^{♥1} Shashank Gupta^{♥1} Taira Anderson^{♥1} Teng Xiao^{♥1}
 Tyler Murray^{♥1} Tyler Romero^{♥1} Victoria Grai^{♥1,2}

Akari Asai^{1,3} Akshita Bhagia¹ Alexander Wettig⁷ Alisa Liu² Aman Rangapur¹
 Chloe Anastasiades¹ Costa Huang¹ Dustin Schwenk¹ Harsh Trivedi¹ Ian Magnusson^{1,2}
 Jaron Lochner¹ Jiacheng Liu¹ Lester James V. Miranda¹ Maarten Sap^{1,3} Malia Morgan¹
 Michael Schmitz¹ Michal Guerquin¹ Michael Wilson¹ Regan Huff¹ Ronan Le Bras¹
 Rui Xin² Rulin Shao² Sam Skjonsberg¹ Shannon Zejiang Shen⁸ Shuyue Stella Li²
 Tucker Wilde¹ Valentina Pyatkin¹ Will Merrill¹ Yapei Chang⁹ Yuling Gu¹ Zhiyuan Zeng^{1,2}

Ashish Sabharwal¹ Luke Zettlemoyer² Pang Wei Koh^{1,2}
 Ali Farhadi^{1,2} Noah A. Smith^{♥1,2} Hannaneh Hajishirzi^{♥1,2}

¹Allen Institute for AI ²University of Washington ³Carnegie Mellon University ⁴Stanford University ⁵Mila
⁶Université de Montréal ⁷Princeton University ⁸Massachusetts Institute of Technology ⁹University of Maryland

*OLMO 3 was a team effort; authors sorted alphabetically. ♥ marks core contributors. See author contributions here.

🤖 **Olmo 3 Base:** Olmo-3-1025-7B Olmo-3-1125-32B
 🤖 **Olmo 3 Think:** Olmo-3-7B-Think Olmo-3-1125-32B-Think
 🤖 **Olmo 3 Instruct:** Olmo-3-7B-Instruct Olmo-3-1125-32B-Instruct
 🤖 **Olmo 3 RL Zero:** Olmo-3-7B-RL-Zero- $\{\text{Math|Code|IF|General|Mix}\}$ Olmo-3-1125-32B-RL-Zero- $\{\text{Math|Code}\}$
 📊 **Base Data:** Pretrain: Dolma 3 Mix Midtrain: Dolma 3 Dolmino Mix Long-ctx: Dolma 3 Longmino Mix
 📊 **Think Data:** Dolci-Think- $\{\text{SFT|DPO|RL}\}$ -7B Dolci-Think- $\{\text{SFT|DPO|RL}\}$ -32B
 📊 **Instruct Data:** Dolci-Instruct- $\{\text{SFT|DPO|RL}\}$
 📊 **RL-Zero Data:** Dolci-RL-Zero- $\{\text{Math|Code|IF|General}\}$ -7B Dolci-RL-Zero-Mix-7B
 🔄 **Training Code:** OLMO-core (pretrain) Open Instruct (posttrain)
 🔄 **Data Code:** datamap-rs (data processing) duplodocus (deduplication) dolma3 (data recipes)
 🔄 **Eval Code:** OLMES (eval suite) decon (eval decontamination)
 📄 **Training Logs:** Olmo-3-7B- $\{\text{Base|Think|Instruct|RL-Zero}\}$ Olmo-3-32B- $\{\text{Base|Think|Instruct}\}$
 🌟 **Demo:** 32B Think 32B Instruct 7B Think 7B Instruct

Olmo 3: fully open model flow

- Qwen, Llama models are open-weight, but they don't fully open their data/checkpoints/codes
- **Olmo 3 releases the entire “model flow”** including pre-training/post-training data, checkpoints for every stage, training/eval/data processing codes
- Especially important to prevent *data contamination* (if eval test set leaks into the pre-training/mid-training data, it leads to spurious rewards in RLVR)
→ Olmo 3 RL-Zero

Olmo at a quick glance

7B, 32B dense models

- Base: only pre-trained
- Think (flagship): w/ reasoning traces
- Instruct: wo/reasoning traces
- RL-Zero: Base → RLVR

Data (Dolma, Dolci)

- Pre-training: Dolma 3 Mix (5.9T tokens)
- Mid-training: Dolma 3 Dolmino mix (100B, harder distribution)
- Long-Context: Longmino Mix (50-100B, up to 65k context)
- Post-training: Dolci Series (3M samples)

Fun Fact

Q. How much do you think it costs to train a 32B model?

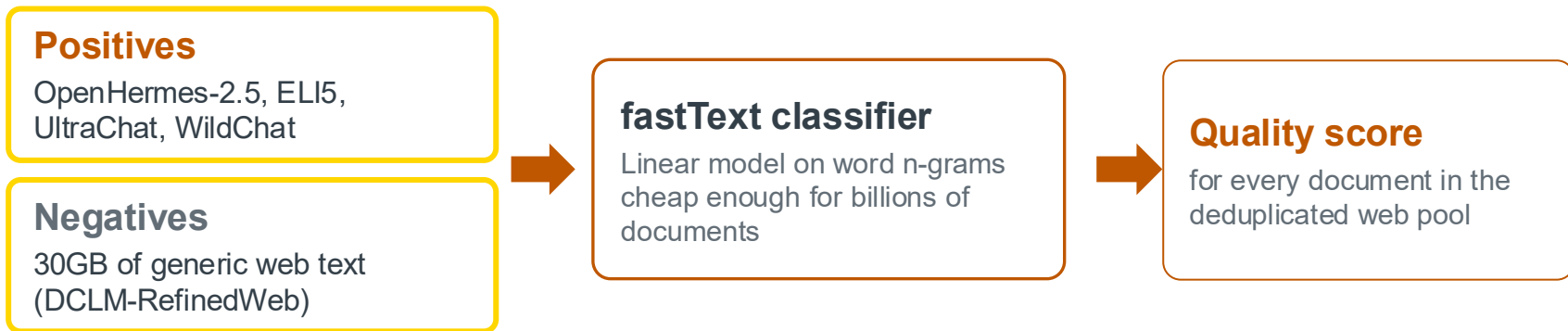
1024 H100 cluster
training + eval took 56 days
Pre-training: 47 days
Post-training: 9 days
\$2/h = \$2.75M

Data decisions start before post-training

- **Mixing:** fit mix ratios on swarms of 30M proxy models then optimize (upweights STEM and code)
- **Quality-aware upsampling:** drop the bottom tier repeat the top 5% up to 7×
- **Microanneals:** test each source cheaply (5B source + 5B web vs. 10B web)
- **Instruction and thinking data in midtraining:** improved every base eval
- **Special tokens:** chat tokens in midtraining dropped GSM8K 49.4 → 0. Olmo3 saves them for SFT

Today's "base model" has already seen post-training-style data

Olmo 3 Base: quality is a classifier score

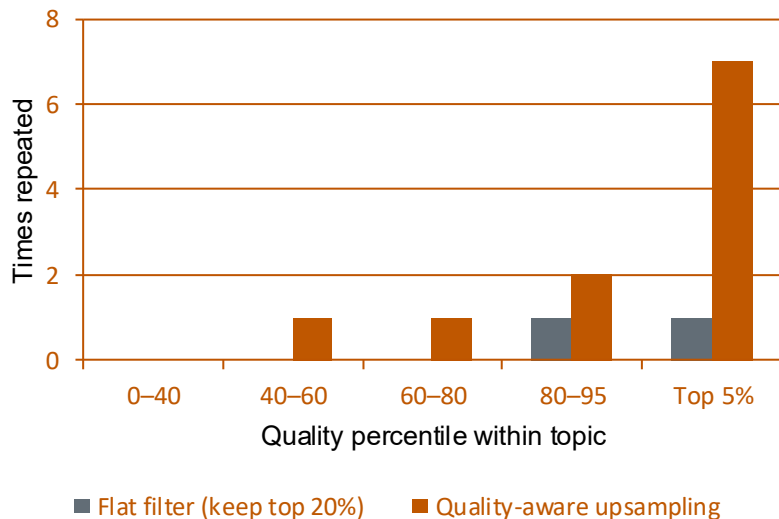


- “Quality” = how much a page looks like the positives (not whether it is true)
- The positives are assistant-style Q&A so selection already leans toward post-training data

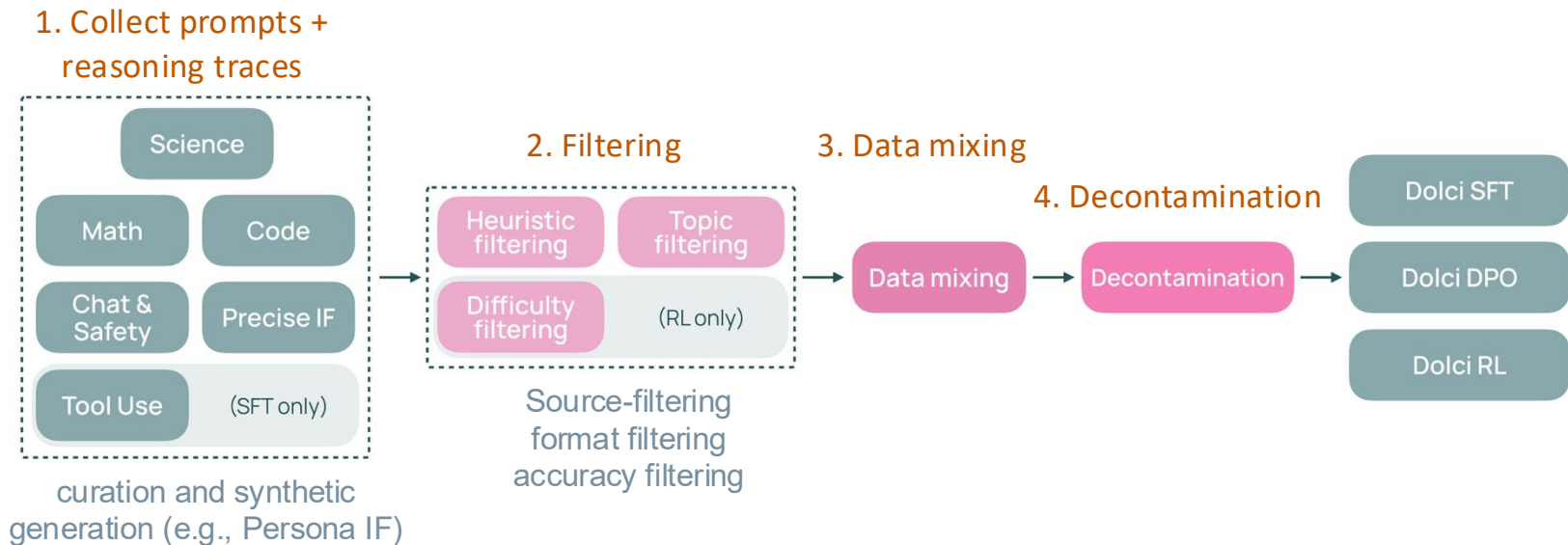
You define quality when you choose the positive examples

Olmo 3 Base: rank within topic, then repeat

- Split the pool into 24 topics (WebOrganizer)
- Rank by quality within each topic: 20 buckets of 5 percentiles
- $24 \times 20 = 480$ buckets with separate sampling rates
- Upsampling curve per topic: drop low buckets, repeat the best up to $7 \times$



Olmo post-training pipeline



Mixing: one data source affects several skills

Internal OLMo 2 pilot

Persona IF = persona-driven synthetic IF data

SFT mix	IFEval	MMLU	Code HumanEval
Base mix	35.7	52.4	35.4
+ Persona IF	70.4	64.1	25.0

Adding a source can improve instruction following while hurting code.

why would IF data raise MMLU by ~12 points?

Evaluate the skill you targeted and the skills you may disturb

Filtering 1. domain-specific verification

Filtering: checks generated SFT responses before keeping them

Domain	How it is checked	What a pass tells you
Code	Run generated tests	passes those tests (only)
Precise IF	Constraint verifiers	listed constraints are met
Chat & safety	No automatic verifier	needs review beyond format

Filtering 2: cheap checks

- Format: truncated traces (<think> without </think>) and empty outputs
- Source: non-commercial licenses, traces touched by restricted models
- Content: repeated text, wrong model identities (“I’m Qwen”)
- Each filter removes only about 0-1% of traces
- Decontamination: 8-gram overlap with eval data, manual review of generic matches

Delta Learning DPO

Problem: SFT with Qwen3-32B thinking responses hurt Olmo 3 Think

Solution: for the same prompt, use the Qwen3-32B response as chosen and Qwen3-0.6B response as rejected, and do DPO

Theory behind it: Delta learning ([Geng et al., 2025](#))

the gap (the delta) is the main training signal
(not the absolute quality of each answer)

Next week: how can comparisons become a training signal?

Conclusion: data = what “quality” means

- LIMA: curated response quality and input diversity.
- RDS+: relevance to a target distribution.
- Magpie / PersonaHub: prompt coverage first, then filtering.
- R1: retained teacher trajectories and task coverage.
- Qwen3: teacher signal on the student’s own prefixes.
- Olmo 3: usefulness depends on the student and objective.