FineInstructions: Scaling Synthetic Instructions to Pre-Training Scale

Ajay Patel[1] Colin Raffel[2,3,4] Chris Callison-Burch[1]

[1] University of Pennsylvania [2] University of Toronto [3] Vector Institute [4] Hugging Face

huggingface.co/fineinstructions / arxiv.org/abs/2601.22146

COLM 2026

Motivation: The Mismatch in Pre-Training

- **Pre-training (99.9% of compute):** next-token prediction on unstructured web text for unsupervised language representation learning and knowledge.
- **Instruction-tuning (<0.1% of compute):** Teaches the downstream use case format with supervision.
- Most compute is wasted on language fitting and predict the next word (a proxy task), which is not the downstream use case.

FineInstructions: We synthetically transform pre-training documents into the downstream use case format: diverse instruction-answer pairs that are in-domain with real user queries.

We then pre-train models entirely on the instruction-tuning objective alone. No next-token prediction objective. Pre-training ablation experiments controlled for same source pre-training corpora and same # of tokens.

$$P(\text{next word}|\text{context}) \rightarrow P(\text{answer}|\text{instruction})$$

TYPICAL PRE-TRAINING MIX

99.9% next-token

<0.1% instruction-tuning

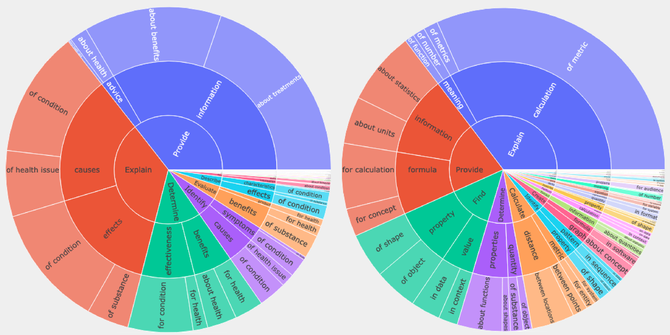
FINEINSTRUCTIONS PRE-TRAINING MIX

100% synthetic instruction-tuning

Compared Baselines and Other Synthetic Pre-Training Methods

- **Standard Pre-Training:** next-token prediction on the raw documents
- **Microsoft's IPT:** train on synthetic reading comprehension Q&A over the document (Cheng et al., 2024)
- **Apple's WRAP:** synthetically rephrase the document (Maini et al., 2024)
- **NVIDIA's Nemotron-CC:** mix of synthetic rephrase, synthetic Q&A, knowledge extraction (Su et al., 2024)
- **NVIDIA's Nemotron Q&A:** Nemotron's synthetic Q&A-only split (closest format to FI) (Su et al., 2024)

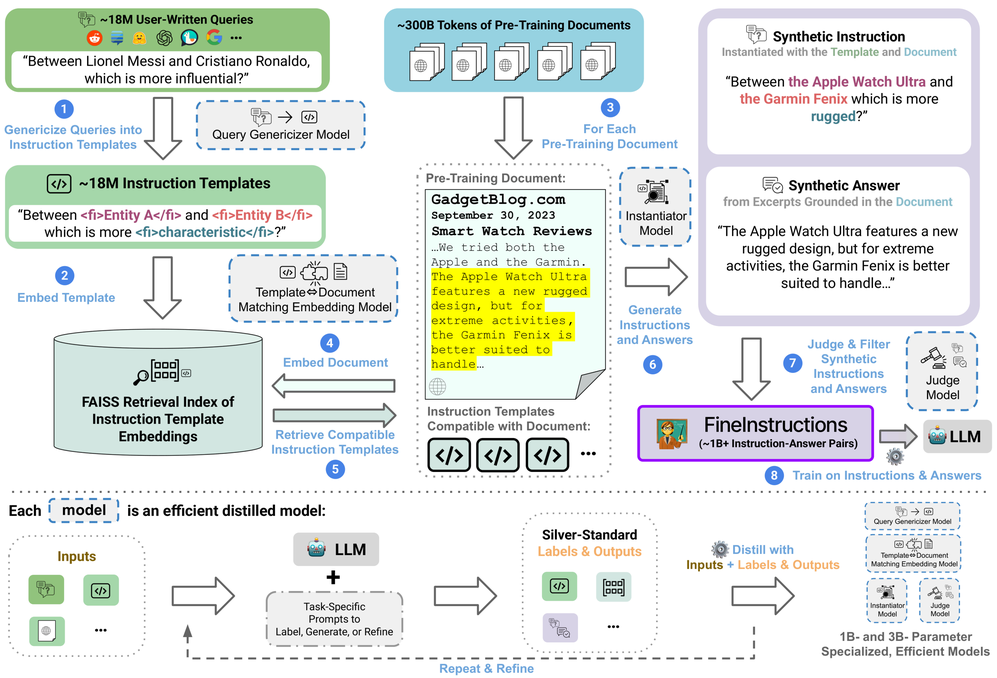
Sample of FineInstructions



Sample of instruction topics related to medicine
(~110M instructions / 1.08B total instructions)

Sample of instruction topics related to math
(~6M instructions / 1.08B total instructions)

FineInstructions Synthetic Pre-Training Pipeline



1.08B synthetic instruction-answer pairs

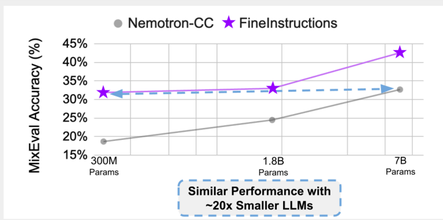
350B tokens of instruction-tuning data

~18M instruction templates

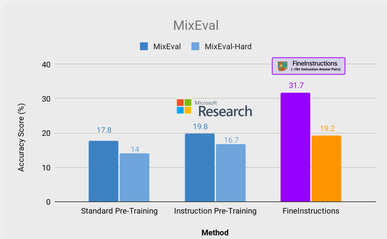
4.3M unique templates used

~3 instruction-answer pairs per document

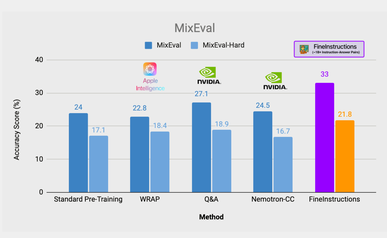
Results



A **300M** model on FineInstructions matches **7B Nemotron-CC** on MixEval. Same tokens, ~20× fewer parameters.



+69% relative MixEval vs Standard Pre-Training on IPT Corpus
(17.8 → 31.7)



+39% relative MixEval vs Standard Pre-Training on Nemotron-CC Corpus
(24.0 → 33.0)

Method	MixEval		MT-Bench-101	AlpacaEval
	Standard	Hard		FI Win Rate %
IPT corpus · 23B tokens				
Standard Pre-Training	17.8	14.0	1.9	73.6%
IPT	19.8	16.7	2.4	68.2%
FineInstructions	31.7	19.2	2.8	—
Nemotron-CC corpus · 300B tokens				
Standard Pre-Training	24.0	17.1	3.5	63.6%
WRAP	22.8	18.4	3.6	65.1%
Nemotron Q&A	27.1	18.9	3.4	76.1%
Nemotron-CC	24.5	16.7	3.6	65.9%
FineInstructions	33.0	21.8	3.9	—

Takeaways

- **Weak supervision:** bootstrap a large, diverse set of pseudo-supervised instruction-answer examples from unlabeled pre-training corpora and in-domain user queries.
- **A path to scale on two independent unlabeled data sources:** scale synthetic supervision with unlabeled pre-training documents and unlabeled natural user queries.
- Token-for-token, **FineInstructions beats standard next-token prediction PT and prior synthetic PT** on MixEval, MT-Bench-101, and AlpacaEval.
- Compute win for **small and efficient** models: 300M FI ≈ 7B Nemotron-CC on MixEval.
- **Resources:** 1B+ pairs, code, 1.8B param models (huggingface.co/fineinstructions)