# QuRating: Selecting High-Quality Data for Training Language Models

Alexander Wettig  Aatmik Gupta  Saumya Malik  Danqi Chen
Princeton Language and Intelligence, Princeton University
awettig@cs.princeton.edu

Paper  Code  Data

## *Existing Practice*

Look at messy web data, and
(a) Write heuristic filter (C4/Gopher)
(b) Pick quality data, e.g. Wikipedia, and identify similar documents (heuristic classification/DSIR)

⬅️ simple proxies for complex intuitions about data

## *Our Approach*

Measure *qualities of text* which
(1) directly capture human intuitions
(2) require a deep understanding of text
by eliciting *pairwise LLM judgments*

How can we scale this to large corpora?
What qualities are important for training LMs?

## Collecting **Quality Ratings**

Show *GPT-3.5-turbo* a pair of texts and ask which

- has a better **writing style**?
- contains more **facts and trivia**?
- has more **educational value**?
- requires **greater expertise**?


We validate that LLMs can discern these qualities and that **pairwise comparisons** improve robustness ✓

Collect 250K judgments and fine-tune a 1.3B-LM as a **QuRater** model to predict *quality ratings* for each text, such that the score differences match the LLM labels under the Bradley-Terry model

## Data Selection

We use the QuRater to annotate 260B token ⊂ SlimPajama with quality ratings for the 4 quality → **QuRatedPajama**

Select documents by sampling *without replacement* with

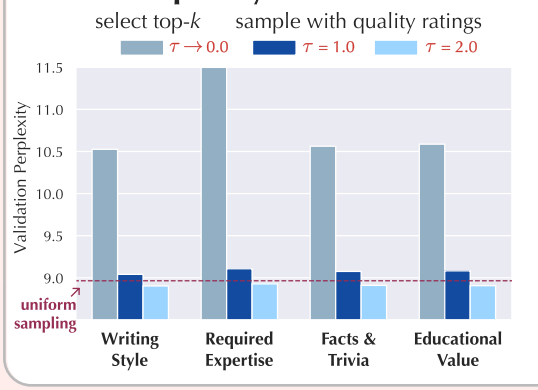
$$p(\text{document}) \propto \exp(\text{quality rating}/\tau)$$

Add temperature $\tau$ to **balance quality** and **diversity**
e.g., top-$k$ selection ($\tau \to 0.0$) and uniform sampling ($\tau \to \infty$)

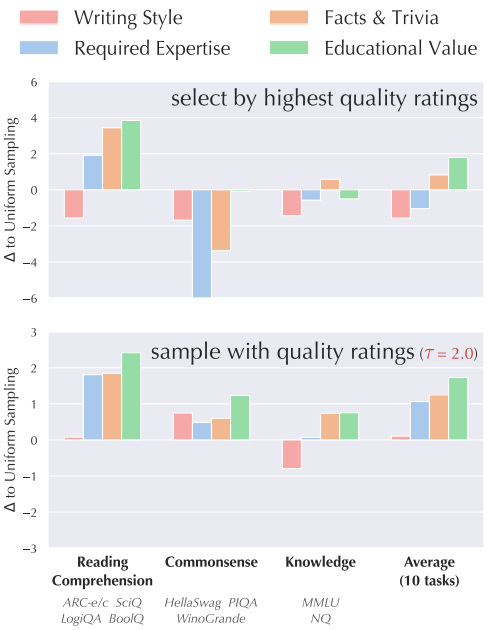
Experiments:

- Select **30B out of 260B tokens** for training
- Keep domain ratios from pre-trained corpus
- Train **1.3B-parameter LMs** from scratch
- Baselines: *uniform sampling*, *DSIR*, *perplexity filtering*
- [In paper] Simple way to construct curriculum:
  Train on documents in the reverse order as sampled

## Better Perplexity



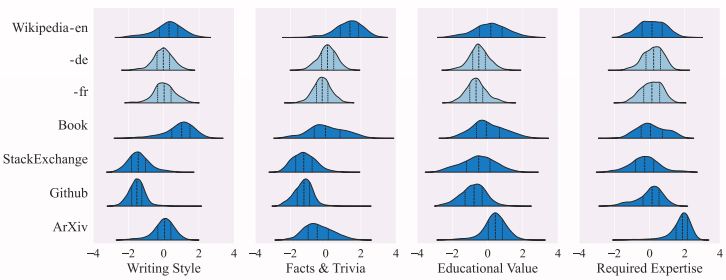

## Better Task Performance (in-context learning)




## Analysis of Quality Ratings

How can we review quality ratings at scale?

- Wide distribution of quality ratings within each domain
  Sequence-level selection has the potential to outperform domain curation/mixing
- The quality ratings do not correlate with perplexity
  Spearman's ρ = 0.14 between *educational value ratings* vs. *Llama-2-7B perplexity*
- More analysis by language, topic, and social role in paper
  We discuss the risks and biases inherent in LLM-based data selection




*quality ratings normalized to unit mean & unit variance across entire corpus*