Knowledge Acquisition through Continued Pretraining is Difficult: A Case Study on r/AskHistorians
Jan Vincent Hoffbauer, Sylwester Sawicki, Marc Lenard Ulrich, Tolga Buz, Konstantin Dobler, Moritz Schneider, Gerard de Melo
Hasso Plattner Institute / University of Potsdam

Project Partner

Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Improving Knowledge Accuracy through Fine-Tuning:

A Case Study on r/AskHistorians

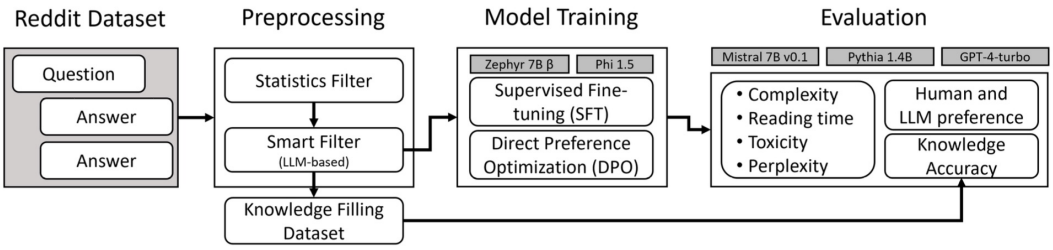

Figure 1: An overview of our experimental setup

Objective

Injecting Domain-Specific Datasets into LLMs

We investigate the effectiveness of different training techniques to improve the knowledge accuracy of large language models (LLMs) using a dataset from the r/AskHistorians subreddit.

- **Training Techniques:** Focus on fine-tuning and alignment to human preferences via Direct Preference Optimization (DPO) to improve knowledge accuracy.
- **Exploring Fine-tuning as alternative** to methods like Retrieval-Augmented Generation (RAG) due to its higher inference costs through a greater number of input tokens. Avoiding techniques such as knowledge editing or steering vectors due to their limited support in libraries like PyTorch.
- **Evaluation:** Assess knowledge accuracy using a curated dataset of historical facts. Measure linguistic features, including reading time, perplexity and toxicity of generated responses.

Question: "When was the USS Indianapolis torpedoed off Tinian?"

Answer start: "The USS Indianapolis was torpedoed off Tinian in the year"

Expected response: "1945"

Figure 2: Knowledge Filling dataset sample

Access our datasets and training pipeline on GitHub

Model	#params	Pretrained on Reddit	LoRA	Accuracy ↑ %
`Mistral-7B-v0.1` (no training)	7B	✓		32
`zephyr-7B-beta` (no training)	7B	✓		31
`zephyr-7B-beta` + r/AskHistorians SFT	7B	✓	✓	29
`zephyr-7B-beta` + r/AskHistorians SFT + DPO	7B	✓	✓	28
`zephyr-7B-beta` + r/AskHistorians Subset-Overfit SFT	7B	✓	✓	49
`phi-1.5` (no training)	1.3B			8
`phi-1.5` + r/AskHistorians SFT				9
`pythia-1.4B` (no training)	1.4B	✓		13

Figure 3: Accuracy on the Knowledge Filling task

Model	Text Complexity ↓ [student grade]	Reading time ↓ [s]	Toxicity ↓ [0-1]
`zephyr-7B-beta`	14.34 ± 2.41	**24.10 ± 10.37**	**0.10 ± 0.20**
`zephyr-7B-beta` + SFT + DPO	**13.35 ± 3.94**	38.75 ± 15.06	0.36 ± 0.22
Original Reddit Answer	11.48 ± 3.72	29.45 ± 28.57	0.20 ± 0.25

Figure 4: Linguistic metrics results

Methodology

Fine-Tuning LLMs in a GPU-Poor Environment

- **Dataset**: Curated 34,631 question-answer pairs from r/AskHistorians (Years 2011-2022). Created a *Knowledge Filling* dataset with 100 questions for knowledge accuracy testing.
- **Models**: Evaluated models from 1.3B to 7B parameters. Used Zephyr 7B β, Phi 1.5 for training; Mistral 7B v0.1, Pythia 1.4B, GPT-4-turbo for evaluation.
- **Experimental Setup**: Single GPU setup (Nvidia A6000 48GB or RTX 3090 Ti 24GB). Used LoRA adapters and quantization for larger models. Applied continual learning to prevent catastrophic forgetting.
- **Metrics**: Measured accuracy on the *Knowledge Filling* dataset. Assessed text complexity, reading time, toxicity, perplexity, and human preference through human and LLM as a judge evaluations.

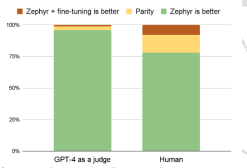

Figure 5: Model Preference as chosen by humans and GPT-4

Implications

Evaluating the Effectiveness of Fine-Tuning for Knowledge Injection on LLMs

- **Impact of Pretraining and Model Size**: Larger models and those pretrained on Reddit data exhibit higher knowledge accuracy.
- **Limited effectiveness of fine-tuning**: Fine-tuning has limited effectiveness in improving knowledge accuracy, except in cases of overfitting, which peaks at 49% accuracy.
- **Impact on Generated Text:** Fine-tuning on Reddit data results in text with higher reading times and toxicity scores but lower text complexity.
- **Human and LLM Evaluations:** Pairwise comparisons show that fine-tuned models are rated worse than baseline models, possibly due to mimicking the unstructured and inconsistent style of the Reddit dataset, whereas instruction-tuned models excel in this domain.
- **Full-weight vs LoRA training:** Full-weight fine-tuning improves accuracy marginally (from 8% to 9%). We conclude that fine-tuning fails to inject knowledge into LLMs and LoRA is not the cause of this failure.