AGU 2024

Curating LLM Tuning Data from the FineWeb Dataset for High-fidelity Domain Adaptation

Kshitiz Khanal, Jeevan Chapagain

UNC Chapel Hill, University of Memphis

Session: IN43C-07
AbstractID:1695859

Background

- Smaller open weights LLMs lack adequate domain knowledge
- Quality domain-specific fine-tuning data needed for high fidelity domain adaptation
- Limitations in customizing and validating synthetic data

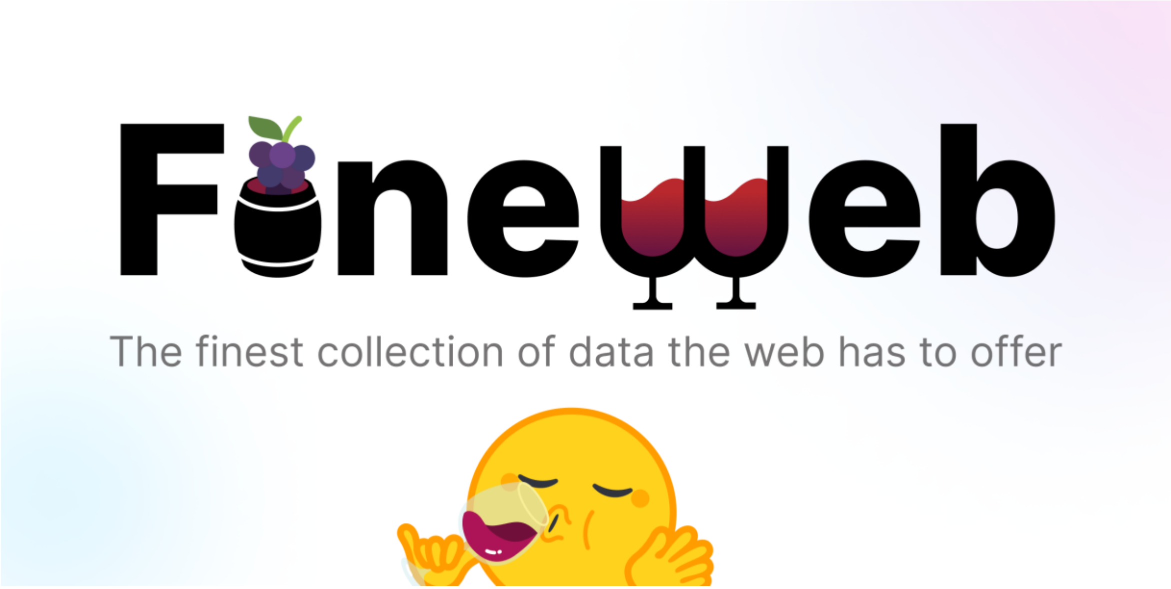
Smoothie (Synthetic data)

Marthastewart.com

Fruit platter (Web commons data)

Filtering FineWeb dataset

- A cleaner and deduplicated version of the CommonCrawl dataset
- 15T tokens, 22B rows
- Text, URL, Date, Token count
- Open Data Commons Attribution License (ODC-By) v1.0

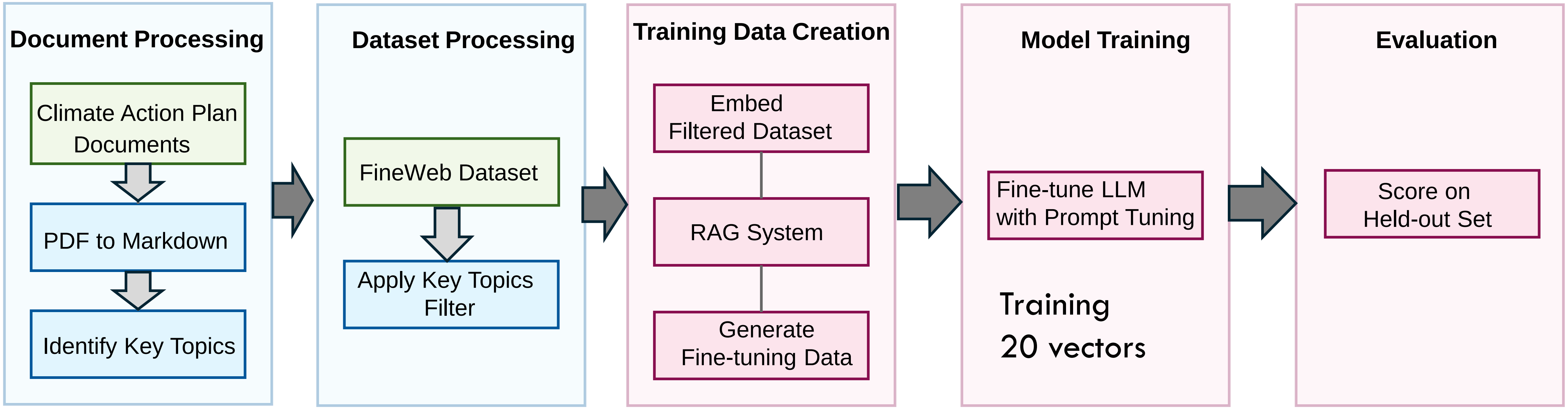


https://huggingface.co/datasets/HuggingFaceFW/fineweb

- perplexity
- cosine similarity with ground truth
- ROUGE
- BLEU score

google/flan-t5-large (780m params)

9k -> 2k rows



Document Processing

Climate Action Plan Documents

PDF to Markdown

Identify Key Topics

Dataset Processing

FineWeb Dataset

Apply Key Topics Filter

Training Data Creation

Embed Filtered Dataset

RAG System

Generate Fine-tuning Data

Model Training

Fine-tune LLM with Prompt Tuning

Training
20 vectors

Evaluation

Score on Held-out Set

climate, emission, ghg, energy, water, waste, building, infrastructure, system, utility, policy, strategy, development, management, action, plan, implementation, impact, reduction, change, measure, target, sector, community, urban, public, equity, justice, equality

Row as external knowledge base for RAG (embedding sentence-transformers/all-mpnet-base-v2)

Summarize (facebook/bart-large-cnn)

Prompt to create queries from summaries

Can learn facts

Query

What are the specific details about: COOPI responds to the needs of the most vulnerable populations in Sudan since 2004?

Training data

COOPI responds to the needs of the most vulnerable populations in Sudan since 2004. Key target groups include women and children, especially female- and child-headed households and unaccompanied minors. COOPI mainstreams **Disaster Risk Reduction, Climate Action, Gender and Protection**, and focuses both on life-saving interventions as well as building long-term resilience through community-based programming. The Head of Mission is responsible for the global management of the country and guarantees compliance with organization and donorsâ€™ procedure. This role covers different activities such as representation, country strategy development, staff management, logistics and budget.

Base model

Climate change and development are major challenges. The humanitarian effort in Somalia has been a success in addressing the needs of developing communities.

Fine-tuned model

COOPI is focused on **disaster risk reduction, climate action, gender empowerment**, vulnerability reduction, and disaster-related changes.

Can generalize

Query

What are the specific details about: The homeownership rate for African Americans remains the lowest among all racial groups in the United States?

Training data

The homeownership rate for African Americans remains the lowest among all racial groups in the United States. On December 1, 2020, the National Housing Conference hosted its annual Solutions to Affordable Housing online convening. In one panel, â€œClosing the Black Ownership Gap,â€�four experts discussed the status of African-American homeownership, current challenges, and the importance of housing policy in reducing the wealth gap between Whites and African Americans. All of the panelists are affiliated with companies or organizations that have implemented programs and initiatives to expand homeownership.

Base model

Panelists discussed the status of African-American homeownership, current challenges, and the importance of housing policy in reducing the wealth gap between Whites and African Americans in the United States.

Fine-tuned model

The Fair Housing Act of 1968 prohibits discrimination based on race in the sale of housing, yet African Americans still experience the effects of explicit and implicit policies that barred them from the housing market before the actâ€™s passage as well as ongoing discrimination in some cases.

Model scale matters

Abysmal results with Google FLAN T5 base (77m) and small (250m):

Perplexity scores >25 with poor quality responses

Score	Dataset: 2k	
	Base (FLAN-T5-Large)	Prompt tuned model
Perplexity	7.2638 (±41.6471)	1.3849 (±1.0541)
Cosine similarity	0.6485 (±0.2227)	0.6134 (±0.2624)
Rouge1	0.2841 (±0.1448)	0.2896 (±0.1606)
Rouge2	0.1506 (±0.1586)	0.1517 (±0.1757)
RogueL	0.2184 (±0.1378)	0.2227 (±0.1478)
BLEU	0.0511 (±0.0917)	0.0691 (±0.1160)

- Remarkably makes the model more confident in domain related responses (lower perplexity scores)
- Perplexity captures improvement in responses
- Can learn facts and generalize on questions not directly related to fine-tuning data
- Better evaluation metrics/approaches required in addition to perplexity and expert judgment

Key challenges

- Experimentation required for finding right model (size, base knowledge level, license)
- Human validation required to ensure standard of filtered dataset
- Domain expertise required to ensure quality of filtered dataset and responses
- Evaluation:
  - Creating automated evals is time consuming and requires domain expertise
  - Time consuming and iterative process even using larger models (experiment with binary eval creation using Claude Sonnet)

Future work

- RAG with knowledge graphs and response from multiple rows to create better quality fine-tuning dataset
- Scaling laws: scale data, scale model (within consumer accessibility constraints)
- Increased observability
- Task breakdown and prompt optimization in the dataset creation process
- More sophisticated evals

Conclusion

- FineWeb and other web-crawl data commons are promising alternatives to synthetic data for domain adaptation of small open weight LLMs
- Prompt tuning can contribute to improved factfulness and generalization out of fine-tuning dataset for domain adaptation
- Additional work is required to create fine-tuning dataset from FineWeb data, requiring domain expertise and creative approaches
- Evaluations are critical, but time consuming and challenging