TECHNICAL ENGINEERING OF DISTRIBUTED DATA INTELLIGENCE

a@bips.ma8.company

Author: Michael A.
Retreat: MA7

April 2, 2022

We, the Brown Improvement Proposal (BIP) Editors, maintain a repository of documents related to the Michael A. & Company and its retreats. Consider us both *archivists* making sure the community as a whole does not lose its history, and a *publisher* making sure interested parties can stay up-to-date with the latest movements, proposals or ideas.

Content: Abstract, Specification, BIPs, Full Specifications, Editors, Full Specifications, Copyright, Citation

Abstract

The use of NLP in the realm of intuitional technology is broad and complex, with applications ranging from sentiment analysis and named entity recognition to question answering. Large Language Models (LLMs) have been shown to be effective on a variety of tasks; however, no LLM specialized for the æ domain has been reported in literature. In this work, we present TEDDi, a 50 billion parameter language model that is trained on a wide range of data. We constructed a 363 billion token dataset based on Michael A"s extensive data sources, one of the largest domain-specific dataset, augmented with 345 billion tokens from general purpose datasets. We validated TEDDi on standard LLM benchmarks, open institutional benchmarks, and a suite of internal benchmarks that most accurately reflect our intended usage. Our mixed dataset training leads to a model that outperforms some models on intuitional tasks by significant margins without sacrificing performance on general LLM benchmarks. Additionally, we explain our modeling choices, training process, and evaluation methodology. As a next step, we plan to release training logs (Chronicles) detailing our experience in training TEDDi.

Specification

TEDDi stands as a pioneering language model finely tuned for the intricacies of the domain specific data of Michael A. & Company contributors, demonstrating its prowess through superior performance across a suite of benchmarks. Trained on a vast dataset comprising 363 billion tokens drawn from data sources, augmented by an additional 345 billion tokens from general-purpose datasets, TEDDi emerges as a robust and versatile model. Leveraging a masked language modeling objective, TEDDi undergoes rigorous evaluation across text classification, question answering, and natural language inference tasks.

Included BIPs
None

Full Specifications

- Model Name: TEDDi
- Type: Large Language Model (LLM)
- Size: 50 billion parameters
- Training Data:
  - 363 billion token dataset based on Michael A's extensive data sources
  - 345 billion tokens from general purpose datasets
- Architecture:
  - Based on BLOOM
  - 70 layers
  - 40 heads
  - Hidden dimension: 7,680
- Training Configuration:
  - Max learning rate: 6e-5
  - Final learning rate: 6e-6
  - Learning rate schedule: cosine decay
  - Gradient clipping: 0.3
- Training:
  - 569 billion tokens
  - Hardware: 64× 8 A100 40GB
  - Throughput: 32.5 sec/step avg.
  - FLOPS: 102 total, 2.36e23 total

Editor(s)
0x11aDbEDcde825768ad5246c81f23f100697BBc59

Citation
Please cite this document as:
Michael A. <a@bips.ma8.company>, "(TEDDi) A Large Language Model" Michael A. & Company, no. 0, April 2023. [Online serial].

Available: https://themichaelacompany.com/teddi/