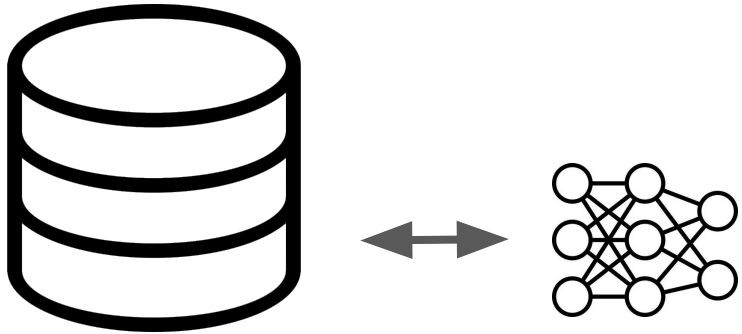
Test-Time Training on Nearest Neighbors for Large Language Models

Moritz Hardt
Max Planck Institute for Intelligent Systems
Tübingen AI Center

Yu Sun
Stanford University

Motivation

Question: How good would a language model be with *access to a large database at test time*?



Lots of work on *retrieval augmentation*, e.g., Deepmind's RETRO

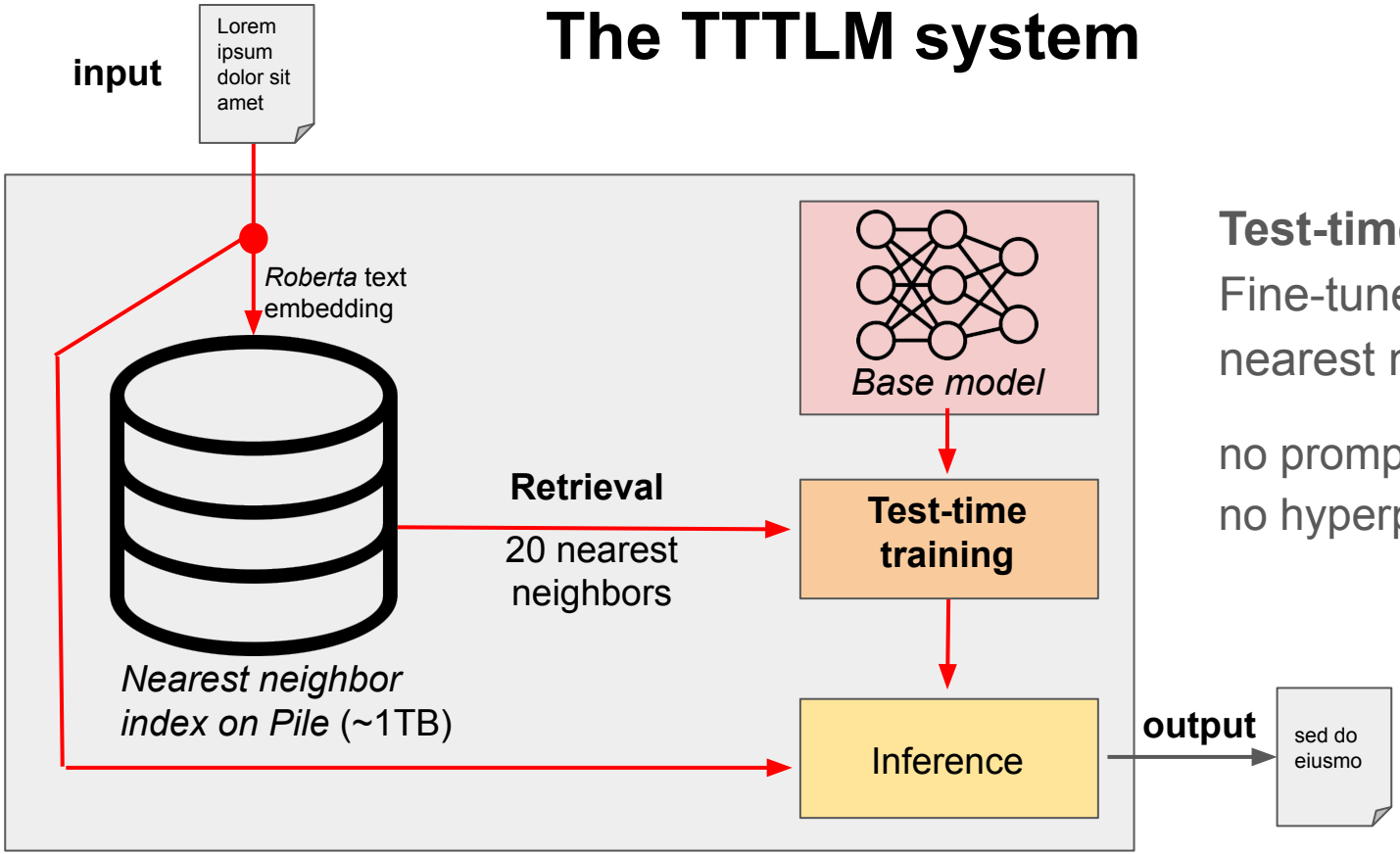
Problems:

Index size and cost of retrieval

Retrieval quality

How to use retrieve data points (prompt tinkering or prompt training)

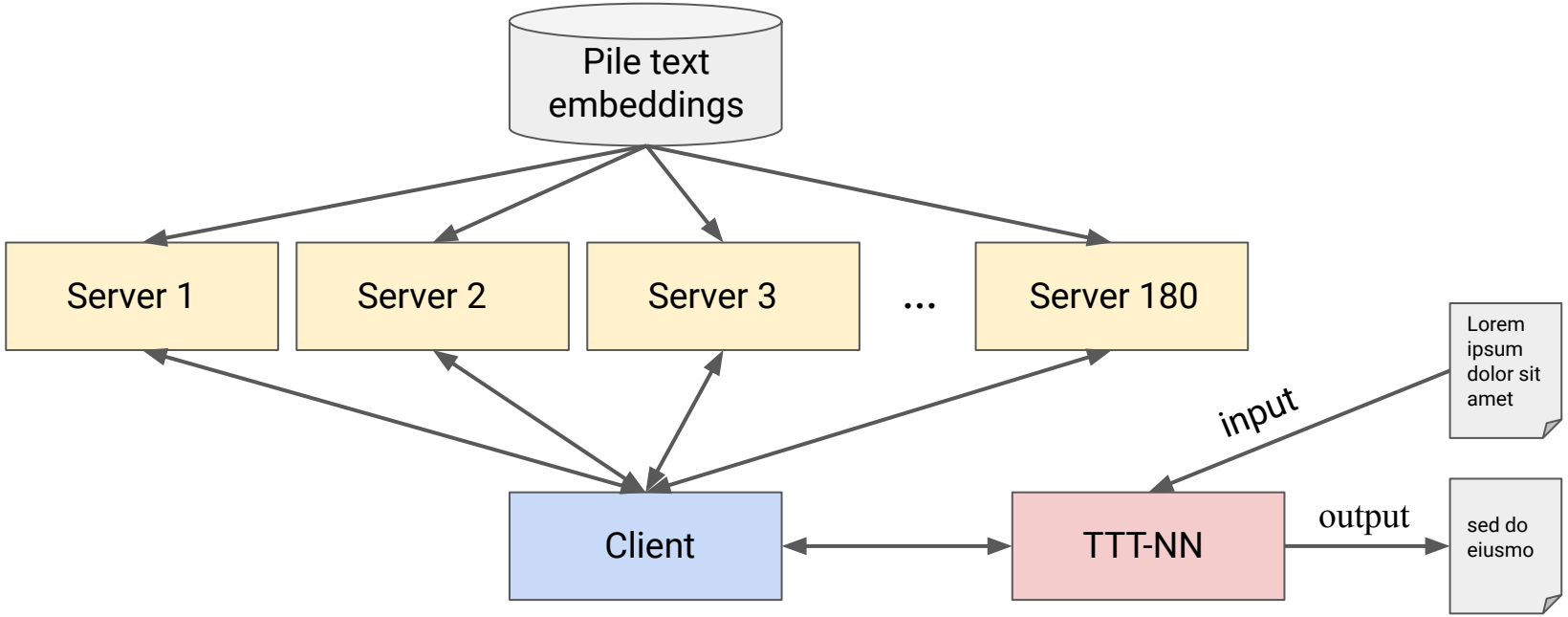
The TTTLM system



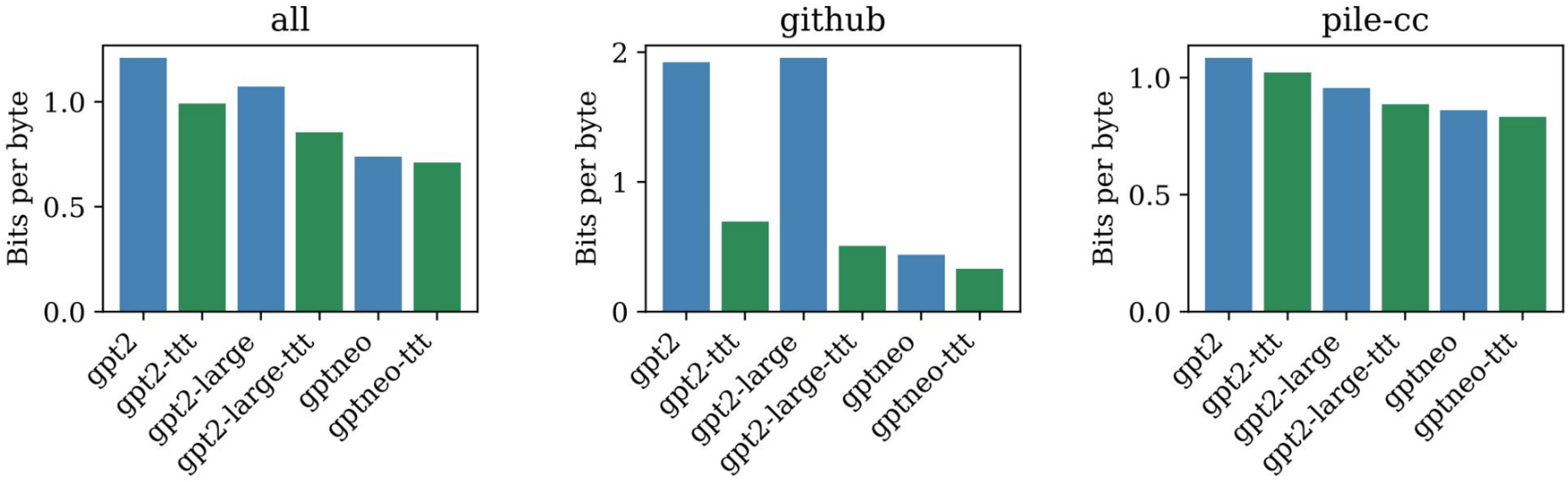
Test-time training (TTT):
Fine-tune model on nearest neighbors

no prompt tinkering/training
no hyperparameters

Fast distributed nearest neighbor index based on FAISS



Queries all of Pile (1TB) in < 1 sec.



Evaluation on all 22 Pile language modeling tasks with GPT2, GPT2-XL, GPTNeo

Strong improvements when model has *not* been pre-trained on database

Strongest improvements for code generation

Limitations: Very long sequences (books/articles), cost of test-time training

Poster
Wed 8 May 10:45 a.m.

Code

Including:

- Code for distributed NN index
- Full Pile index files
- Pretrained models

https://github.com/socialfoundations/tttlm