Peitian Zhang

namespace.pt@gmail.com Google Scholar

Education

Renmin University of China (RUC), Beijing, China 2022 – 2025

M.E. in Artificial Intelligence

Renmin University of China (RUC), Beijing, China 2018 – 2022

B.E. in Computer Science and Technology

Experiences & Projects

Baidu Jun. 2024 – Present

Intern in ErnieBot Group

- **Long-Context LLM**
  - **Industrial Extension of Activation Beacon**
    - *(Description)* Improving Activation Beacon towards industrial application.
    - *(Role)* Proposition, model training, model evaluation.
    - *(Outcome)* We coordinate Activation Beacon with modern acceleration techniques like FlashAttention-2, increase its parallelism, and apply it on newly released LLMs such as Qwen-2 and Llama-3, resulting in improved running efficiency and compression quality.

Beijing Academy of Artificial Intelligence Jun. 2023 – Jun. 2024

Intern in Knowledge and Computing Group

- **FlagEmbedding**
  - *(Description)* A series of effective and versatile embedding models for general retrieval and retrieval augmentation of LLMs, including:
    - BGE: a series state-of-the-art general embedding model;
    - BGE-M3: a multi-lingual, multi-functionality, and multi-granularity embedding model;
    - LLM-Embedder: a unified embedding model to support LLM's diverse retrieval augmentation needs.
  - *(Role)* Proposition, data curation, model training, model evaluation.
  - *(Outcome)* Our models are the most downloaded AI models on Huggingface throughout China, and have been integrated into popular LLM frameworks and cloud services such as LlamaIndex, Azure, and Volcengine. Our open-source project earned 5K+ stars on Github.
- **Long-Context LLM**
  - **Activation Beacon**
    - *(Description)* An effective, efficient, compatible, and low-cost method to extend the context length of LLMs through KV compression along sequence dimension.
    - *(Role)* Proposition, data curation, model training, model evaluation.
    - *(Outcome)* Activation Beacon significantly improves the long-context utilization of Llama-2 and Mistral owing to the nearly lossless context compression effect, meanwhile achieving high running efficiency.
  - **Long-LLM QLoRA**
    - *(Description)* Revealing LLM's inherent potential in context extension can be unlocked via QLoRA training over a few synthetic data.
    - *(Role)* Proposition, data curation, model training, model evaluation.
    - *(Outcome)* The context length of Llama-3 is extended from 8K to 80K using only 3.5K synthetic data and 8 hours training, while the model achieves remarkable performance on various long-context benchmarks with little compromise on short-context tasks.

Case Retrieval System of Renmin University of China Aug. 2022 – Sep. 2022

Individual Project

- *(Description)* A legal case retrieval system that supports keyword retrieval, similar case retrieval, faceted retrieval, and interpretation of search results over 10M+ documents.
- *(Role)* Data curation, model training, backend/frontend development, and system deployment.
- *(Outcome)* The system is a fundamental backbone of the first Legal Data Analysis Challenge of RUC and is actively used by students and teachers in RUC.

Microsoft Research Asia Jul. 2021 – Apr. 2022

Intern in Social Computing Group

- **Hybrid Inverted Index**
  - *(Description)* An ANN method where embedding clusters and salient terms collaborate to accelerate dense retrieval.
  - *(Role)* Responsible for proposition, model training, and evaluation.
  - *(Outcome)* The method achieves on par performance against HNSW with 10x smaller index size without supervised training, and significantly outperforms it with end-to-end optimization.

Selected Publications

[1] *(Arxiv)* Soaring from 4K to 400K: Extending LLM's Context with Activation Beacon
Peitian Zhang, Zheng Liu, Shitao Xiao, Ninglu Shao, Qiwei Ye, Zhicheng Dou

[2] *(ACL'24)* Retrieve Anything To Augment Large Language Models
Peitian Zhang, Shitao Xiao, Zheng Liu, Zhicheng Dou, Jian-Yun Nie

[3] *(EMNLP'23)* Hybrid Inverted Index is A Rubust Accelerator for Dense Retrieval
Peitian Zhang, Zheng Liu, Shitao Xiao, Zhicheng Dou, Jing Yao

[4] *(SIGIR'24)* Term-Sets Can Be Strong Document Identifiers For Auto-Regressive Search Engines
Peitian Zhang, Zheng Liu, Yujia Zhou, Zhicheng Dou, Zhao Cao

[5] *(SIGIR'24)* C-pack: Packaged Resources to Advanced General Chinese Embedding
Shitao Xiao, Zheng Liu, **Peitian Zhang**, Niklas Muennighof

[6] *(ACL'24)* BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation
Jianlv Chen, Shitao Xiao, **Peitian Zhang**, Kun Luo, Defu Lian, Zheng Liu

[7] *(ACL'24)* LM-Cocktail: Resilient Tuning of Language Models via Model Merging
Shitao Xiao, Zheng Liu, **Peitian Zhang**, Xingrun Xing

[8] *(ACL'24)* INTERS: Unlocking the Power of Large Language Models in Search with Instruction Tuning
Yutao Zhu, **Peitian Zhang**, Chenghao Zhang, Yifei Chen, Binyu Xie, Zhicheng Dou, Zheng Liu, Ji-Rong Wen

[9] *(Arixv)* Are Long-LLMs A Necessity For Long-Context Tasks?
Hongjin Qian, Zheng Liu, **Peitian Zhang**, Kelong Mao, Yujia Zhou, Xu Chen, Zhicheng Dou

Skills

Programming	Python, C/C++, HTML, CSS
Professional Knowledge	PyTorch, Transformers, Faiss, Elasticsearch, Django