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

Beijing Academy of Artificial Intelligence Jul. 2023 – Present

Research 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: state-of-the-art general embedding model;
    - BGE-M3: 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 received 20M+ total downloads, 1M+ monthly downloads on Huggingface. 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 activation compression.
    - *(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 (yet largely underestimated) 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.

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

Research 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] *(Arxiv)* 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] *(Arxiv)* 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] *(Arxiv)* LM-Cocktail: Resilient Tuning of Language Models via Model Merging
Shitao Xiao, Zheng Liu, **Peitian Zhang**, Xingrun Xing

[8] *(Arxiv)* 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

SKILLS

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