Peitian Zhang

namespace.pt@gmail.com namespace-Pt Google Scholar Beijing

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

Projects

Retrieval Models & Indexes

- **Dense Retrieval: 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)* Participate in training BGE/BGE-M3; Lead the LLM-Embedder project.
  - *(Outcome)* Our models achieved state-of-the art performance on MTEB/C-MTEB benchmark. They are **the most downloaded AI models on Huggingface throughout China**, and have been integrated into popular LLM frameworks and cloud services such as LlamaIndex and Azure. Our open-source project earns **6K+** stars on Github. *Three corresponding papers are accepted by ACL 2024.*
- **Generative Retrieval: TSGen**
  - *(Description)* A novel generative retrieval framework where each document is identified by a set of key terms, and these terms can be generated in any permutation. The model is learned to explore the optimal generation order on its own.
  - *(Role)* Lead the project.
  - *(Outcome)* TSGen overcomes the falsely pruning problem in generating conventional sequential DocIDs, thereby significantly improving the retrieval quality and the generalizability for new documents. *The corresponding first-author paper is accepted by SIGIR 2024.*
- **ANN Index: Hybrid Inverted Index**
  - *(Description)* An ANN index where embedding clusters and salient terms collaborate to accelerate dense retrieval.
  - *(Role)* Lead the project.
  - *(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. *The corresponding first-author paper is accepted by EMNLP 2023.*
- **Engineering Practice: Case Retrieval System of Renmin University of China**
  - *(Description)* A legal case retrieval system that supports keyword retrieval, similar case retrieval, faceted retrieval, and interpretation of search results over 10M+ documents.
  - *(Role)* Lead the project.
  - *(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.

Long-Context LLMs

- **Context Window Extension: LongLLM-QLoRA**
  - *(Description)* Establish the long-context capability for Llama-3 by position extrapolation and synthesized long-dependency data.

  - *(Role)* Lead the project.
  - *(Outcome)* Extend the context length of Llama-3 from 8K to 80K using merely 4.5K high-quality SFT data. The model significantly outperforms concurrent works in the community.

- **Efficient Computation of Long Context: Activation Beacon**
  - *(Description)* Compress the long context into shorter yet more compact KV activations, hence enables the LLM to perceive longer context with higher efficiency.
  - *(Role)* Lead the project.
  - *(Outcome)* The context compression capability can be quickly established with lightweight training on short context. Experiments on modern LLMs demonstrate minimal information loss with **x4 and x8 compression rate**. The method is compatible with context window extension and KV compression from other dimensions. *The corresponding first-author paper is under review.*

Internships

Baidu (Ernie Lite Team) 2024.6 – Present

Topic: Research and application of long-context LLMs.

Beijing Academy of Artificial Intelligence (Knowledge Retrieval and Computing Team) 2023.6 – 2024.6

Topic: Research and application of embedding models; Research of long-context LLMs.

Microsoft Research Asia (Social Computing Team) 2021.6 – 2022.4

Topic: Research of efficient ANN indexes; Research of efficient news recommendation systems.

Publications

[1] *(Under Review)* 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] *(Under Review)* 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++, HTML, CSS
Professional Knowledge	PyTorch, Transformers, Faiss, Elasticsearch