

张配天

✉ namespace.pt@gmail.com

🗣 namespace-Pt

🎓 Google Scholar

📍 北京

教育经历

中国人民大学 - 人工智能专业 - 学术型硕士 2022 – 2025
中国人民大学 - 计算机科学与技术专业 - 本科 2018 – 2022

实习 & 项目经历

北京智源人工智能研究院 2023.6 – 至今
知识检索与计算组, 实习生

• FlagEmbedding

- (介绍) 一系列精准、通用的向量表征模型, 适用于用于一般检索任务和检索增强大语言模型, 其中包括:
 - * BGE: state-of-the-art 的通用表征模型;
 - * BGE-M3: 支持多种语言、多种检索方式、多种检索粒度的表征模型;
 - * LLM-Embedder: 支持 LLM 多样化的检索增强场景的表征模型。
- (角色) 模型构思, 数据建设, 模型训练, 模型评测。
- (成效) 这些模型在 MTEB 等多个评测基准上达到最好效果; 在 Huggingface 上获得了 20M+ 总下载及 1M+ 月均下载。我们的开源仓库在 Github 上获得了 5K+ 标星。

• Long-Context LLM

- Activation Beacon

- * (介绍) 一个强大、高效、兼容、低成本的 LLM 上下文扩展技术。
- * (角色) 模型构思, 数据建设, 模型训练, 模型评测。
- * (成效) Activation Beacon 能够近乎无损地对 LLM 上下文激活进行压缩, 从而显著改进了 Llama-2、Mistral 模型对长上下文的理解和利用能力, 同时其保持极高的运行效率。

- Long-LLM QLoRA

- * (介绍) 揭示了 LLM 自身拥有强大的上下文扩展潜力, 可以通过在很少的长上下文数据上进行 QLoRA 训练激发。
- * (角色) 模型构思, 数据建设, 模型训练, 模型评测。
- * (成效) 我们使用该方法高效地将 Llama-3 的上下文长度从 8K 拓展到 80K, 整个过程仅依靠 3.5K 合成数据, 耗时 8 小时。得到的模型在一系列长上下文基准上获得了极佳的效果。

中国人民大学智能类案检索系统 2022.8 – 2022.9
个人项目

- (介绍) 该系统能够基于关键词匹配和语义向量相似度完成对超过 10M 裁判文书的检索, 同时支持分片搜索、解释搜索结果等高级功能。
- (角色) 数据收集, 模型训练, 前后端开发, 系统部署。
- (成效) 该系统是中国人民大学第一届法律大数据分析挑战赛的基座, 同时被中国人民大学的教师和学生广泛使用。

微软亚洲研究院 2021.6 – 2022.4
社会计算组, 实习生

• Hybrid Inverted Index

- (介绍) 一个使语义向量聚类 and 关键词倒列表协同工作的近似最近邻搜索 (ANN) 索引。
- (角色) 模型构思, 数据建设, 模型训练, 模型评测。
- (成效) 该索引无需监督训练即达到和 HNSW 达到接近的精度和时延, 而其空间占用仅有十分之一; 经过监督训练后, 其精度能够显著超过 HNSW。

发表论文

- [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) (CCF B) Hybrid Inverted Index is A Rubust Accelerator for Dense Retrieval
Peitian Zhang, Zheng Liu, Shitao Xiao, Zhicheng Dou, Jing Yao
- [4] (SIGIR'24) (CCF A) 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) (CCF A) 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

技能

编程技能
专业技能

Python, C++, HTML, CSS
PyTorch, Transformers, Faiss, Elasticsearch, Django