张泽铭

年龄：24　　政治面貌：党员

籍贯：河北省邢台市　　求职意向：计算机视觉/LLM 应用开发工程师

电话：15811516045　　电子邮箱：23021211861@stu.xidian.edu.cn

教育背景及成果

西安电子科技大学（本科），电子信息工程　　2019.09 - 2023.06

- 奖项荣誉：星火杯校级一等奖、两次校级奖学金

西安电子科技大学（硕士），电子科学与技术　　2023.09 - 2026.06

- 实验室：所在 VIPSL 实验室隶属于高新波教授（现西安电子科技大学校长、中国高等教育学会常务理事，原重庆邮电大学校长）团队，主要研究方向：图像超分辨率算法研究
- 荣誉：校级一等奖学金、校级二等奖学金、实验室一等奖学金、2024 学年‘优秀学生干部’
- 成果： Neurocomputing（一作论文在投）Efficient Dynamic Selective-Attention Network for Image Super-Resolution

专业技能

- **深度学习 &Transformer 核心：**熟练掌握 pytorch 及主流大模型（GPT、DeepSeek、llama 等）原理与训练流程，能结合 LoRA/PEFT 等微调方法完成模型的开发与迭代。对 CV、NLP 领域有所了解，掌握手动搭建、魔改深度学习网络模型技术。
- **LLM 工程落地：**精通 Prompt Engineering、CoT、多轮对话管理等技术，善于在复杂场景（RAG 问答、Chatbot、Agent 等）中定制并优化大模型应用。
- **检索与微调优化：**深入理解向量检索（Faiss/Milvus）、BM25 等检索机制，具备 Embedding 微调、Cross-Encoder Rerank、RLHF 等能力，有效提升相关性与回答精准度。
- **其他技能：**熟练使用 Git、Maven、Docker、Latex、figma 等工具，有丰富的 Linux 服务器项目部署经验。有一定英语阅读、写作、交流能力（CET-6 通过）。

项目经历

AI 投标辅助配置系统　　**2024.5 - 2024.10**

项目简介： 负责构建面向企业招标方的 AI 辅助配置系统，

主要工作：

1. 离线解析与知识库构建: 通过 olmOCR 将标书文件解析为结构化数据，引入锚点机制，提升了结构化 PDF 解析中的表格检测精度和单元格跨页识别准度，保障标书内容提取的完整性和准确性。
2.Query 模块优化: 融合规则、BERT 分类与 PromptEngineering 实现科研意图识别，并据此动态调整检索策略；针对缩写、拼写误差与模糊提问 (如”效果对比”、”实验设计”) 进行标准化与术语扩展；引入 HyDE 生成假设文段以丰富查询语义，提升短查询命中率与长尾问题的召回能力。
3. 检索召回模块: 采用 BM25 与向量召回融合的混合检索策略，兼顾专业术语的关键词精确匹配与语义泛化能力；在标书数据上微调 BGE/SimCSE 等嵌入模型，提升对领域术语的理解与召回效果；引入微调后的 Cross-Encoder 进行结果重排，提升相关性排序；基于人工标注测试集，采用 MRR、Precision@K、NDCG@K 等指标评估检索性能，P@3 提升约 30%，首条命中率显著改善。

4. 生成阶段优化: 引入多轮对话上下文管理机制，通过核心实体注入与代词指代消解提升语义连贯性；在 Prompt 中启用答案出处标注机制，增强生成配置内容的可解释性与用户信任度。

基于 Qwen2.5-VL 的 Arxiv 论文问答系统 **2025.2 - 至今**

项目简介： 本项目旨在构建一个高效、准确的本地 Arxiv 论文问答系统，以克服传统 RAG 在科研 PDF 处理中的局限。基于 Qwen2.5-VL 模型的多模态问答系统，支持 PDF 截图的嵌入-召回-回答流程，避免了传统 OCR+ 文本分块所带来的处理冗余、语义缺失与 embedding 局限。

主要工作：

1. 为了解决传统 RAG 系统 OCR+ 文本分块带来的处理时间长、召回结果差、只能利用文本 embedding 召回的问题，我们构建了基于 Qwen2.5-VL 的多模态能力的原生 PDF 截图的嵌入-召回-回答 pipeline。
2. 系统在 PDF 缩略图的多模态向量库基于 Late Interaction 计算召回相似度 (比 cosine 相似度更精细)，并且利用 Qwen2.5VL 的原生分辨率进行召回图像动态分辨率的 QA 问答。
3. 为了提升 Qwen2.5-VL 在 PDF 问答上的图标理解能力，我们在 pdfvqa/chartQA 上进行 SFT。并且搭建了基于 DeepSeek-Chat 的 Agent 评测系统，自动化评测 3k+ QA。

8K 超高清画质提升芯片，海信——西电联合实验室项目 **2022.9 - 2023.12**

项目简介： 与海信信芯微联合研发国内首颗全自研 8K 画质 AI 芯片，负责核心画质增强 IP 模块算法设计与轻量化优化。项目聚焦超分辨率重建、动态插帧、感知增强与芯片部署技术，在 4K/8K@60fps 下实现了画质优化与实时推理的兼顾。

主要工作：

1. 设计多任务渐进式视频画质提升模型，构建“失真复原—结构增强—感知增强”三阶段子网络，提升视频的色彩、清晰度与纹理质量。
2. 主导自建数据集的清洗与样本筛选，提出场景标签驱动的数据预处理策略，解决 Repeat、Databroken 等场景问题。
3. 参与设计基于特征金字塔的运动视频插帧模块，优化大尺度运动下帧间对齐与融合精度。
4. 负责模型剪枝、特征级知识蒸馏与分级量化策略设计，压缩模型参数至 13K−25K，显著提升硬件部署效率。

AI 恋爱大师智能体 **2022.9 - 2023.12**

项目简介： 本项目开发的智能体，可以为用户提供感情指导服务。支持多轮对话、记忆持久化、RAG 知识库检索等能力。底层设计框架基于 ReAct 模式，能够主动思考并调用工具来完成复杂任务。

技术选型： Spring AI + LLM + RAG + Tools Calling + MCP + OpenManus + Docker

主要工作： 1. 大模型集成：利用 Spring AI 框架快速接入 AI 大模型，并封装统一的调用接口，能够实现大模型的灵活切换。
2. 多轮记忆对话：实现对话上下文记忆功能，并解决了服务重启后对话记忆丢失的问题。
3.RAG 知识库构建：实现文档的切片处理、向量存储、文档检索以及检索增强等功能
4. 工具调用：实现多种工具调用功能，包括文件操作、联网搜索和 PDF 生成等，扩展了 AI 的能力边界。
5. 集成高德地图 MCP：利用 MCP 集成地图定位功能，让 AI 能够准确地基于地理位置推荐约会地点。
6. 自主规划智能体：构建具备自主规划能力的智能体，能够分解任务、选择工具、循环执行直至完成复杂任务。
7. 服务部署: 实现基于 SSE 的流式输出, 利用 Docker 实现后端项目容器化, 并通过云托管平台进行部署。

AI 小伴教学助手平台——陕西智瞳科技有限公司合作项目 **2024.6 - 至今**

技术栈：

数据库：MongoDB、Milvus、Redis、MySQL；后端：SpringBoot、Maven、Tomcat、RabbitMQ；对象存

储：MinIO；LLM：graphrag-local-ollama、Qwen7B

项目简介： 本项目旨在为教育领域提供自动化的辅助功能，主要应用于学生与教师的教学互动。项目以 Java 为后端开发语言，结合多种深度学习大模型技术，旨在通过自然语言处理、大规模语言模型、数据存储与分发等功能，提升教学效率和用户体验。

主要工作：

1. 后台权限管理：用户分级为学生、教师、管理员。使用 MySQL 存储管理用户数据。
2. 读取学生用户端上传的文本内容或文件内容存入 MongoDB 并调取服务器端 LLM API，返回 LLM 输出结果。
3. 教师端上传的教案文件存入 MinIO, 并实现提供用户可编辑白板区域, 为教师提供课程研发服务。
4. 将所有用户上传数据进行编码嵌入，存入向量数据库 Milvus，作为知识库进行 graphrag-local-ollama 模型构建微调。
5. 使用 RabbitMQ 实现用户请求的异步处理，避免上课时多用户在线卡顿问题。

导引头低光增强 & 检测算法项目—— 研究所合作 **2024.6 - 至今**

技术栈： FPGA + RK3588 处理器（ARM 架构）+ CameraLink + MIPI-CSI + RKNN

项目简介： 开发和优化一套低光环境下的导引头低光增强与目标检测系统，通过创新的算法设计与硬件架构提升导弹系统在复杂环境下的目标识别精度和实时反应能力，显著增强作战性能。

主要工作：

1. 主导 FPGA 与 RK3588 双处理架构的设计与集成，结合 CameraLink 和 MIPI-CSI 协议，成功实现 FPGA 对图像数据的并行加速处理，提高了系统在高速图像传输与处理中的效率，显著提升了系统处理能力与实时响应速度。
2. 提出并实现了一种基于 VPSS 的图像处理流水线，有效地将图像去噪、裁剪、缩放与锐化处理结合，优化了低光环境下图像的清晰度和精度。主导低光增强与目标检测模型的开发，结合 YOLO 目标检测算法，在低光或夜间环境下的目标识别精度提升了 30% 以上，通过 PSNR 和 SSIM 指标验证了图像质量的显著提升。
3. 负责将 PyTorch 训练的深度学习模型转换为 ONNX 格式，并通过 RKNN-Toolkit 将其优化为适用于 Rockchip NPU 的 RKNN 模型，提升了模型的推理速度，并减少了系统的内存占用和计算资源消耗。
4. 精确优化了量化过程，采用 KL-Divergence 和 MMSE 量化算法，保证在推理性能提升的同时，量化后的精度损失控制在 3% 以内，有效平衡了性能和精度。
5. 深入优化 YOLO 模型，提升了系统在复杂低光环境中的目标检测和跟踪精度，目标检测精度（mAP）提升了 20%。结合 F1-score 和 Precision-Recall 分析，优化了目标追踪和定位算法，成功实现了目标的实时追踪，确保导引头在复杂战场环境中的高精度跟踪。
6. 对 RKNN 模型推理进行了全面的性能评估，成功将推理延迟降低至 30ms 以内，并通过优化计算资源，系统吞吐量提升了 40%，满足实时作战需求。

个人介绍

- 兴趣爱好广泛，喜欢运动。多次参加校级、院级羽毛球比赛，并获得混合团体赛冠军、亚军。
- 作为学生干部，有较强的语言表达能力、组织协作能力、分析和解决实际问题的能力。
- 勤于学习，乐于思考，善于发现，做事情有较强的计划性，对问题有独到的领悟和理解。