去年年底我参加了中国未来金融分析师大赛的智能体搭建赛道，我们小组成员一起做了一个绿色金融问答系统。首先，我们在各个国家政策性网站上爬取大量绿色金融相关数据，并整理成结构化的知识图谱。然后我们基于知识图谱构建35000条问答对，用这些问答对基于DeepSeek-r1-7b的模型进行Lora微调来提高模型在绿色金融专业化领域的表现，同时还使用了RAG技术来对精选文本构成的知识库做检索增强。随后我们将成果制作成ppt参加比赛评比，获得了第四名的好成绩。

-微调时是否比较怎样选择batch size和上下文窗口长度效果最好？

-为什么选择DeepSeek-r1-7b作为基模？

-你们做RAG的知识库有多大？为什么不直接作为prompt给模型而是要做RAG？

-你们比赛的评判标准是什么？你觉得你们为什么脱颖而出？

-这个绿色金融问答系统服务B端还是C端用户，有哪些主要功能？

-可以讲讲Lora微调的原理与具体实践过程吗？

-你们的数据是怎么来的？怎么挑选信源？

2023年我参与实验室“古典文学AI”项目，负责开发支持多风格约束的古诗词生成引擎。用户输入主题和体裁，系统生成符合平仄、押韵规则且意境连贯的诗词。成果应用于某文化类App的“AI诗人”模块，上线首月用户生成诗词超2000首。首先我从《全唐诗》《宋词菁华》等典籍爬取9.6万首诗词，按朝代、流派标注，然后构建格律知识库：解析每首词的平仄模板、押韵位置、对仗规则，转化为结构化规则树。在模型方面我选择了GPT-2-medium，并且进行微调来实现标准语言建模。具体细节层面，我在Beam Search中引入韵律有限状态机，强制生成结果在指定位置押韵，并且使用动态Token Masking禁止生成重复意象。

-为什么选择GPT-2而非LLaMA等更大模型？推理延迟具体是多少？ 是否测试过量化/蒸馏的加速效果？

-Beam Search中如何实现韵律状态机？请举例说明《浣溪沙》词牌在解码时如何约束第2/4/6句押韵。

-动态Token Masking如何实现“禁止重复意象”？是否构建了意象冲突表？其数据来源是什么？

-辅助任务（平仄/风格分类）的损失函数权重如何设定？实验显示哪个任务对生成质量影响最大？  
-如何处理古汉语中的通假字/异体字？是否引入繁体转简体模块？  
-标注“风格”存在主观性，如何验证标注一致性？

-人工评估中“意境连贯性”的具体评分标准是什么？ 是否尝试过自动化评估指标？

-若用户要求生成“李白风格”的七言诗，如何实现诗人级细粒度控制？ 是否需要新增诗人编码向量？

去年我主导了公司新一代企业知识库系统的检索模块升级，核心目标是解决内部技术文档的精准检索问题。我们首先从Confluence、GitHub Wiki等平台爬取35万份技术文档，通过PDF解析和段落分割构建了结构化文本库。为了提升语义检索能力，我们基于BAAI/bge-large-zh预训练模型进行领域适应微调：使用对比学习框架，人工标注了2万组查询-段落匹配对，通过LoRA技术微调嵌入层，使Embedding适配企业技术术语。检索架构采用RAG方案，用Faiss建立256维向量的IVF-PQ索引，实现毫秒级响应。系统上线后内部工单解决率提升40%，平均检索时间从7分钟降至28秒。

-微调时为什么选择对比学习框架而非交叉熵？这对检索任务的指标提升有何具体影响？

-256维向量是否会造成信息损失？实验中有没有对比过384/512维的效果差异？

-Faiss索引选择IVF-PQ而非HNSW的关键考量是什么？参数调优时如何平衡召回率和延迟？

-遇到OOV（未登录词）技术术语时，微调后的Embedding模型如何处理？

-人工标注2万组数据时制定了哪些匹配标准？如何解决“部分相关”段落的标注歧义？

-RAG方案中重排序模块的设计逻辑？为什么没有直接使用纯向量检索？

-上线后如何持续评估检索效果？构建了哪些自动化反馈机制？

我在最近主导了LLM推理引擎的Speculative Decoding优化项目，核心解决GPT-4级别模型API响应延迟高的问题。我们基于Medusa框架设计三级加速方案：首先构建包含7个草稿头的轻量级并行预测树，通过知识蒸馏从主模型提取文本生成模式；其次创新动态头选择机制——实时分析历史token置信度分布，仅激活概率差异>0.2的头；最后实现安全验证-梯度回传联合优化：用CUDA内核融合技术将验证阶段的拒绝采样损失反向传播至草稿头，使其错误率从18%降至6.3%。在8×A100集群部署后，对代码生成任务实现平均3.8倍加速，吞吐量提升290%，同时保证输出完全匹配贪婪解码结果。

-为何选择Medusa而非Lookahead/Assisted Decoding方案？在长文本生成场景下Medusa的注意力计算复杂度如何量化？

-草稿头蒸馏时采用KL散度损失而非MSE的理论依据？实验中有否观察到蒸馏偏差导致代码补全错误率上升？

-动态头选择机制中“概率差异>0.2”的阈值如何确定？是否测试过基于信息熵的自适应阈值？

-推导拒绝采样阶段的期望加速比公式。

-如何证明梯度回传优化未破坏Medusa的零误差保证？  
-轻量级草稿头结构设计：为何用单层Transformer而非LSTM？多头共享Embedding是否导致预测同质化？  
-动态头选择机制的具体实现：是每token选择还是按segment选择？选择延迟如何控制在50μs内？  
-梯度回传优化细节：回传时冻结主模型参数的理论依据？联合训练是否引发草稿头过拟合主模型历史错误？  
-是否尝试过树状结构剪枝策略（如删除置信度<0.3的分支）？剪枝后如何维持k=7的并行度？  
-验证阶段采用Token并行验证还是序列验证？如何避免并行验证的内存爆炸？

-对长距离依赖（如函数跨行调用）的预测失败率高达42%，如何改进架构？

我在实验室主导了「DiffuText」项目，核心解决大模型长文本生成中的连贯性衰减问题。我们创新性地将离散扩散模型与LLM解码过程融合：首先训练一个条件化扩散桥接器——通过变分推断将LLM的隐空间映射到扩散过程的Markov链，用Gumbel-Softmax重构离散文本分布；其次设计双阶段生成架构：LLM先生成核心语义骨架，扩散模型再执行多轮去噪填充，通过分类器指导强化事实一致性，在每轮扩散中动态注入检索增强的实体约束。在arXiv学术生成任务中，该方法使1000+token长文本的实体连贯性提升57%（，人类评估显示逻辑流畅度得分达4.32/5.0，推理速度仅比标准自回归慢1.8倍。该工作入选ACL 2024 Industry Track，并正与医学期刊合作构建学术写作助手。

-为何选择离散扩散而非连续扩散？如何解决文本离散空间与高斯噪声的兼容性问题？

-推导桥接器的变分下界如何包含LLM隐变量与扩散状态的互信息？

-分类器指导中使用的事实一致性信号具体如何量化？为何不用RLHF？

-当扩散模型生成与LLM骨架冲突时，误差传播机制如何建模？

-Gumbel-Softmax的温度调度策略如何影响生成多样性？  
-语义骨架抽取：为何用关键词重要性排序而非简单截断？是否尝试过Seq2Seq骨架压缩？  
-扩散填充的条件注入机制：如何将LLM骨架信息编码为扩散模型的Attention K/V？  
-实体约束的动态注入：为何选择知识图谱嵌入而非实体标签拼接？如何避免约束过度导致文本僵化？  
-迭代去噪中的早期停止策略：基于困惑度还是方差阈值？  
-是否测试过非自回归扩散替代标准迭代扩散？相对收益如何？  
-如何解决扩散模型对专业术语（如“拓扑绝缘体”）的生成匮乏问题？  
-骨架生成与扩散填充的联合训练方案：是否共享参数？梯度如何反向传播？

最近我主导了新一代跨平台实时竞价系统的开发，核心解决Meta/Google/TikTok等多平台广告预算分配效率低下的痛点。系统采用联邦学习架构：在本地部署的边缘节点聚合各平台历史竞价数据，通过差分隐私保护生成联合特征；算法层创新设计三阶分配模型——先用Transformer时序编码器预测平台级流量波动，再用组合拍卖理论计算边际ROI，最后通过对抗鲁棒优化（ARO）动态调整出价策略以应对平台规则突变。部署时自研分布式信用拍卖协议，实现亚秒级跨平台预算调度。上线半年服务200+头部广告主，平均获客成本降低37%，预算浪费减少52%，其中某跨境电商客户年度ROI从1.8提升至3.2。

-联邦学习中如何解决平台数据异构性？Meta的CPM数据与TikTok的CPC数据如何归一化？

-差分隐私噪声注入导致ROI预测方差增加40%，如何补偿模型精度损失？

-如何检测平台虚假流量（如TikTok的bot点击）？特征工程中哪些指标最具判别力？

-边缘节点数据回传频率设定依据？如何平衡实时性与网络开销？  
-Transformer时序编码器为何选择LogSparse注意力而非全注意力？计算复杂度降低多少？  
-组合拍卖中的VCG机制适用性分析：为何最终选择GSP+预算补偿的混合方案？  
-出价策略的多目标优化：如何权衡点击率/转化率/平台服务费？帕累托前沿求解方法？  
-当预测流量波动与实际偏差>30%，动态出价如何紧急纠偏？  
-模型迭代中在线学习与批量更新的频率决策依据？  
-为何拒绝使用LLM生成出价策略？确定性保障与可解释性如何实现？