垂直大模型技术研究与商业落地  
宗云兵  
金现代信息产业股份有限公司

通用大模型与垂直大模型  
金现代

通用大模型就是大家熟知的ChatGPT、通义千问等，它们能够解决教育、军事、医疗等多领域的问题。

垂直大模型，是用专业领域知识训练通用大模型得到的，用于解决专业领域问题的大模型。

通用大模型  
训练集  
微调训练

垂直大模型  
专业大模型、行业大模型、领域大模型、私有大模型  
LORA

大模型对算力的要求

从零训练通用大模型、微调通用大模型、使用大模型，三种情景对算力的要求完全不同。

训练通用大模型  
大模型：ChatGPT-3.5  
微调通用大模型  
大模型：ChatGLM3-6B、ChatGLM3-68B  
使用大模型  
参数量：1750亿（ChatGPT-3.5）、60亿（ChatGLM3-6B/68B）  
显卡数量：3000张（训练ChatGPT-3.5）、1张（微调及使用ChatGLM3-6B/68B）  
显卡品牌：英伟达 A100  
显卡成本：15万元（A100）、2万元（英伟达4090）  
训练语料：3.5TB（ChatGPT-3.5训练所需）、45TB（假设值，用于说明训练数据量之大）  
训练成本：6千万元（训练ChatGPT-3.5）、5千元（微调ChatGLM3-6B/68B，假设值）  
训练耗时：数月（训练ChatGPT-3.5）、半天~1周（微调ChatGLM3-6B/68B）

检索增强生成技术  
检索增强生成(RAG, Retrieval-Augmented Generation)

示例：“林黛玉倒拔垂杨柳?”（避免生成不符合常识的内容）  
信息过时：企业文档资料更新迭代快，无法每次更新资料时就训练一次大模型。  
文档资料（1500万条）：通过分块处理，提高检索效率。  
输入问题：“重要保电事件分几级?”  
相关块（30个）：从资料块中筛选出与问题相关的块。  
基于Qwen-14B训练的电力应急垂直大模型输出答案：“根据已知信息中的内容，重要保电事件分4级，分别为特别重大、重大、较大及一般。”

构建垂直大模型示例  
以某省电力公司应急业务垂直大模型为例，说明构建垂直领域大模型的过程。  
大模型：Qwen-14B  
获取方式：开源、免费  
硬件算力：1张A100显卡  
训练集：5000条数据（假设值）  
准备时间：1周  
训练时间：6小时  
问题示例：  
问：专项应急演练的周期?  
答：三年为一周期。

Qwen-14B电力应急  
通用大模型（用于微调的基础模型）  
LORA（用于微调的技术）  
微调训练（过程）  
问题示例：  
问：重要保电事件等级包括哪些?  
答：特别重大、重大、较大及一般。

电力应急智能问答  
垂直大模型（应用实例）

RAG的演进  
原生：检索质量低，问答效果差  
进阶：检索丰富，问答效果较好  
高级：模块化设计，功能模块化，使用方式灵活

技术演进：  
① 动态切分（提高检索效率）  
② 多路召回（增加检索结果的多样性）  
③ 问题改写（优化查询语句，提高检索准确性）  
④ 检索重排（对检索结果进行排序，提高问答质量）

Advanced RAG（高级RAG，表示技术演进后的RAG）

Noduiar RAO

RAG 提升之四:检索重排

对检索结果进行重排，可以减少噪声，提升大模型总结生成的效果。

问题：近5年的电力紧急事件，有哪些?

多路召回结果：top50片段

用户问题提出后，经过重排模型处理，得到top10片段：

1. 2024年电力紧急事件...
2. 2023年电力紧急事件...
3. 2022年电力紧急事件...
4. 2021年电力紧急事件...
5. 2020年电力紧急事件...

多模态RAG（文本）

通过优化分段和检索方式，模型的跨文档信息整合能力得到了提高。

查询示例：近5年的电力紧急事件，有哪些?

回答示例：

* 2020年的电力紧急事件有...
* 2021年的电力紧急事件有...
* 2022年的电力紧急事件有...
* 2023年的电力紧急事件有...
* 2024年的电力紧急事件有...

垂直大模型的六大落地方式

1. 智能问答
2. 智能理解
3. 智能填写
4. 智能配置
5. 智能调用接口（如Text -> SQL、Text -> DOC、Text -> Configuration、Text -> API）

自然语言处理 NLP

BERT、LSTM等传统自然语言处理模型具有速度快、专业性强的特点。在智能体中使用大小模型结合的策略，可以提高任务执行的准确率。

智能问答示例：

请假管理

用户输入：“我明天因病请两天假”

大模型处理：BERT进行意图识别，生成请假流程或请假条。

用车管理示例：

信息：“公司总部，山东数字人科技股份有限公司地址：济南市高新区天辰大街1188号。”  
“山东数字人科技股份有限公司成立于2002年，是一家专注于数字人体研发与应用的高新技术企业。”

研发架构（以山东数字人科技股份有限公司为例）

DICTHUMAN（公司品牌或项目名称）

* 医学教学部
  + 解剖产品部
  + 临床培训部
* 医学研发中心
  + 形态产品部
  + 定制项目部
  + 数据制作部（视频拍摄、影像处理）

公司简介（以道普信息技术有限公司为例）

道普信息技术有限公司是一家专注于数字化第三方保障服务的机构。公司致力于帮助客户全面控制数字化风险，提供数字化战略方向风险、项目建设过程风险、网络安全风险等相关的检测、咨询服务。

数字化项目类别主要风险

（以下列举了数字化项目在不同阶段可能面临的风险）

项目预算风险、项目可行性风险、项目需求风险、项目进度风险、项目成本风险、项目质量风险、安全保密风险、政策合规风险、项目验收和交付风险、项目决算风险等。

服务产品

1. 可研规划设计：根据组织、项目的需求，制定项目规划方案、建设方案、应用方案等。
2. 信息工程造价评估：对信息工程项目进行造价评估，出具权威、公正的信息工程造价评估报告。
3. 信息工程监理2.0：对信息工程项目的立项、采购、实施、验收、运行等阶段进行全程管控。
4. 智能系统检测：使用先进检测设备，对系统功能、性能、施工质量等进行全面测试。
5. 软件系统检测：对软件功能、性能、安全、可靠性、兼容性等多个层面进行测试，出具第三方专业权威的测试报告。