

基于图书馆馆藏资源的大 模型中文语料库建设经验 分享

——同伴客数据

High quality data helps AI create unlimited possibilities

Contents

- 01 公司简介
- 02 经验与案例分享
- 02 前台交互方案展示



公司简介



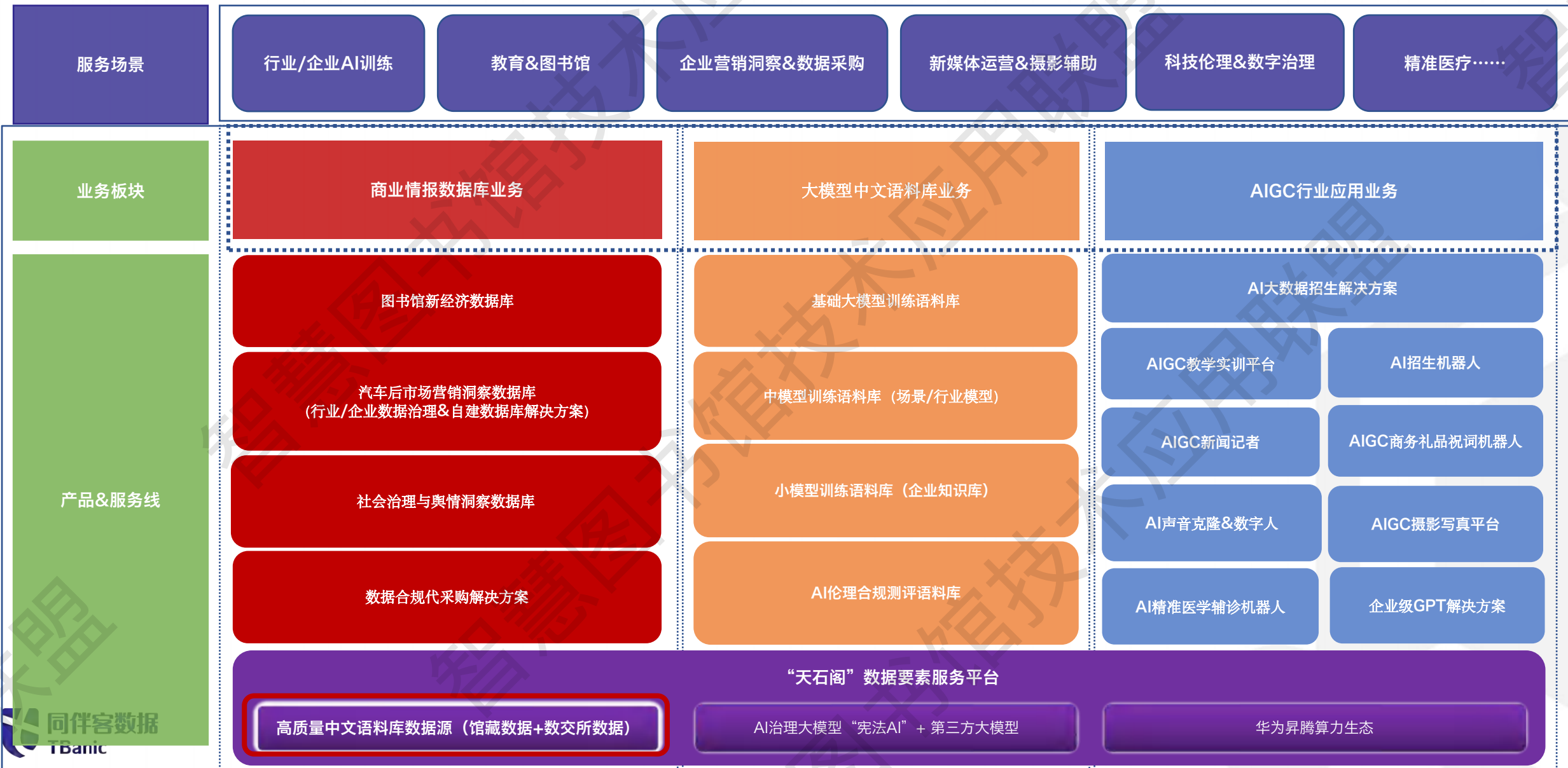
公司简介

- 同伴客数据是一家专注于以高质量数据要素助力人工智能创造无限可能的科技公司，由数据科技与AI治理领域的资深数据专家团队创立，总部位于北京。我们以数据要素服务（DAAS）平台为依托，不断研发和推出创新的产品和服务，助力人工智能行业发展。
- 同伴客是国家高新技术企业和中关村高新技术企业，也是首批入驻北京国际大数据交易所、福建大数据交易所等数据交易机构的数据服务商，并承接了多项国家级数字经济课题的研究工作，深厚的数据要素积淀是同伴客能够提供持续推出卓越产品的重要保障。
- 我们的业务足迹已经覆盖遍布全球的270余家企业、高校和政府客户，其中包括埃克森美孚、壳牌、中国林业、清华大学等世界500强企业和知名研究机构。





同伴客数据业务矩阵





服务全球一流客户，打造全球一流产品

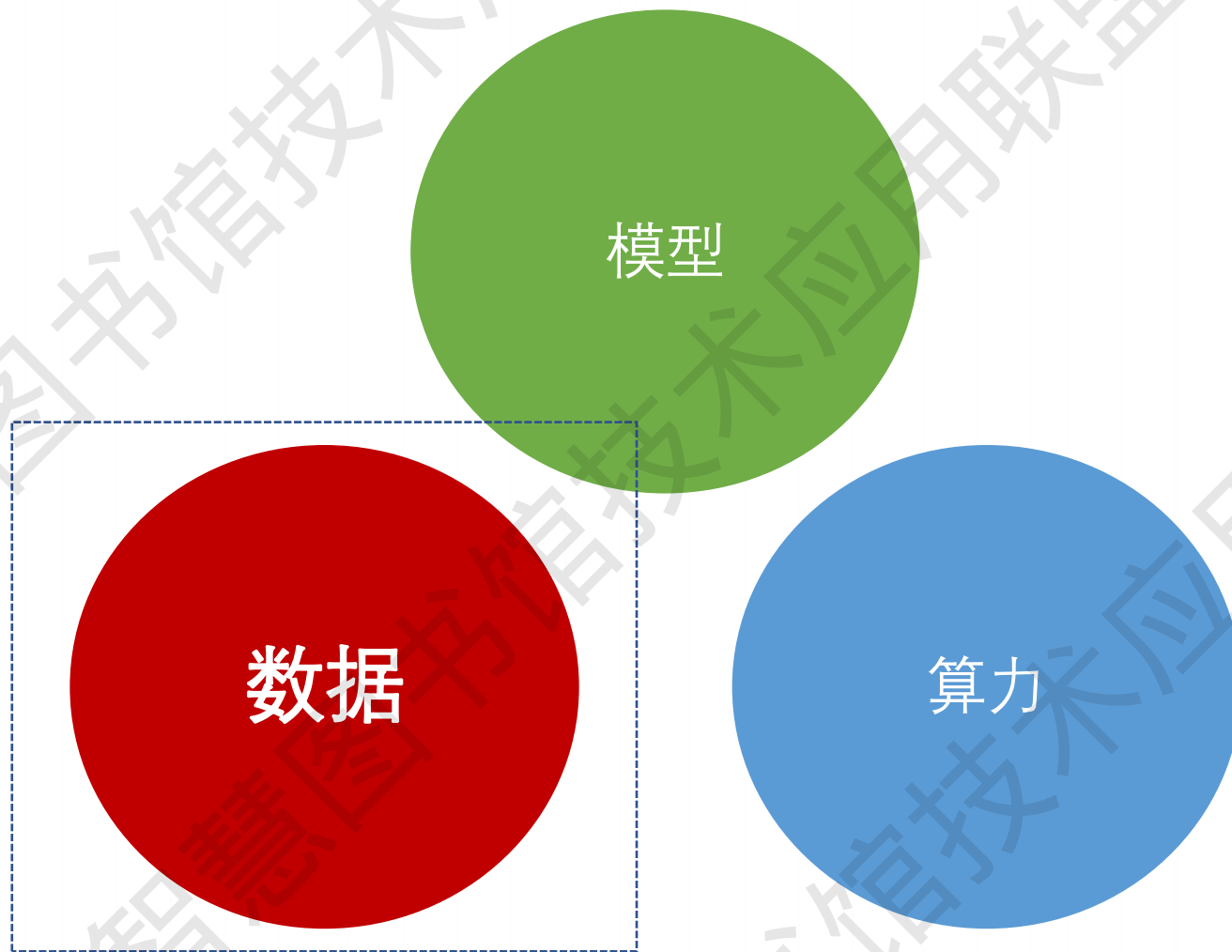




经验与案例分享



行业三要素生态





数据决定了人工智能的上限



传统智能换装结果

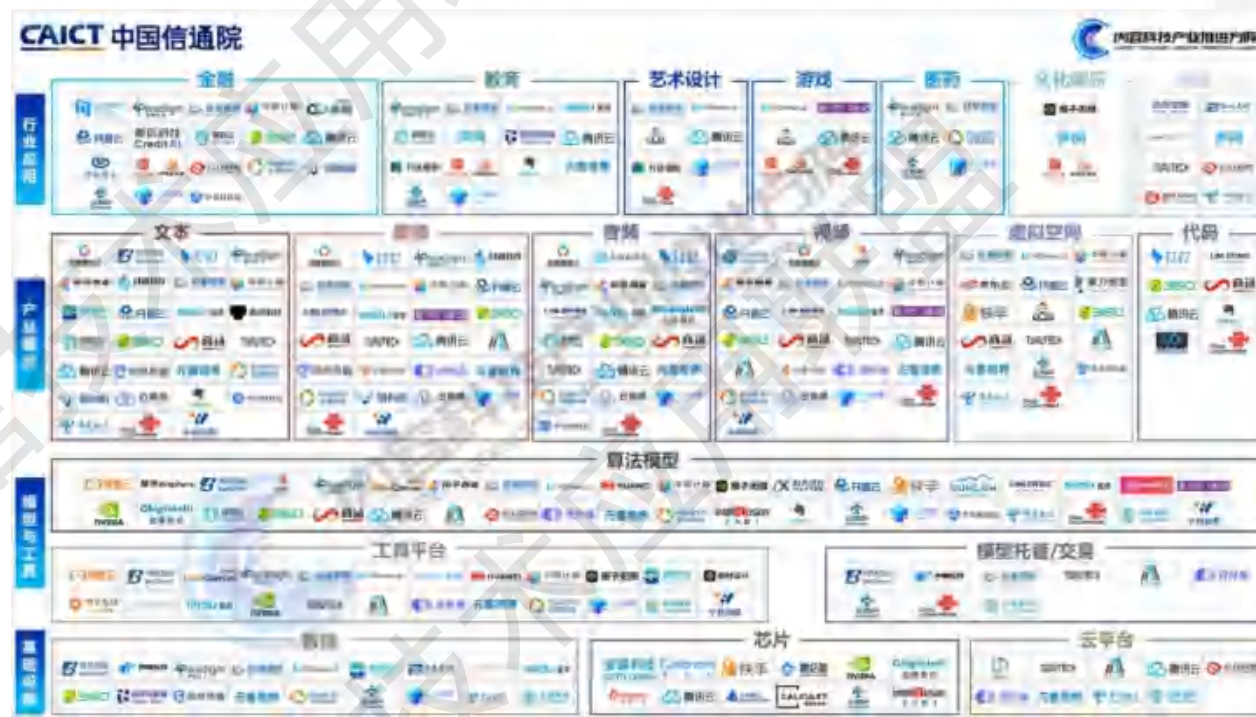


基于线下门店大数据训练的AI换装结果



“卡脖子”问题：高质量中文语料库

- 数据、算力、算法是人工智能行业的三大发展要素，而国内高质量中文语料库的缺失，成为我国通用人工智能“卡脖子”的关键问题。而**图书馆正是国内专业中文知识的重要储藏基地，只是目前馆藏资源的数字化水平、以及图书馆与人工智能行业信息差，使得图书馆与AIGC行业未能实现有效协同。**
- 同伴客数据是一家领先的AIGC行业数据要素提供商，也是中国信通院“AIGC产业链图谱”中，AIGC数据基础设施领域唯一专注于图书馆行业的数据企业。我们目前正在垂直大模型调优与建设领域拥有诸多技术积累与应用案例。





实践方案

数据要素服务平台

高质量中文语料库数据源（馆藏数据+数交所数据）

AI治理大模型“宪法AI”+ 第三方大模型

华为昇腾算力生态

云服务器部署

本地化部署

专有云部署

公有云部署

机房建设



案例：联合顶尖高校图书馆、研究院定制医学垂类大模型

方向一：基于人工智能与大数据技术的垂类专病大模型开发与临床评价

研究内容：1) 使用大数据和人工智能技术，构建专病大模型。2) 在临床实践中应用该专病大模型。3) 建立专病的多模态标准化数据库，并制定临床评价体系。

考核指标：创建与专病大模型相关的新技术和方法 3-5 项；开发 1 套基于大数据和人工智能的专病大模型系统；在不少于 2000 人的队列中进行临床验证，并提供临床评价报告 1 份；申请国家级发明专利 2-3 项；在相关领域的重要期刊发表学术论文 3-5 篇。

项目立项

Cumulative Reasoning With Large Language Models

Yifan Zhang^{*1} Jingqi Yang^{*1} Yang Yuan^{1,2} Andrew Chi-Chih Yao^{1,2}

¹ILIS, Tsinghua University

²Shanghai Qizhi Institute

{yuanyang, andrewcyao}@tsinghua.edu.cn

Abstract

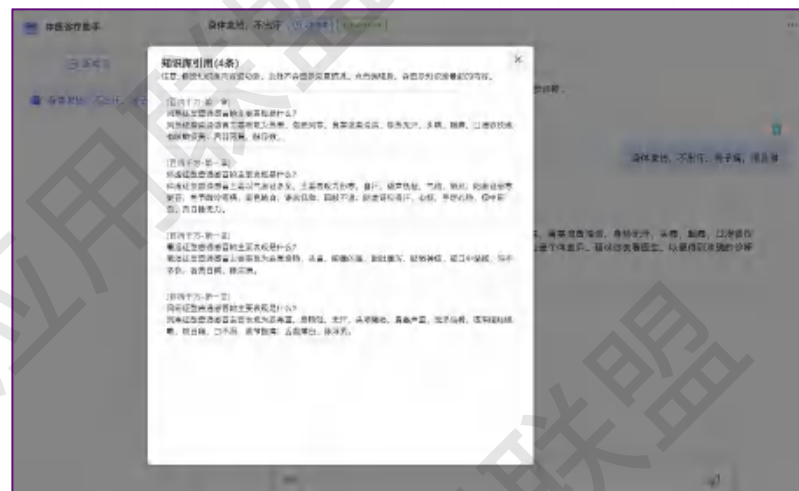
While language models are powerful and versatile, they often fail to address highly complex problems. This is because solving complex problems requires deliberate thinking, which has been only minimally guided during training. In this paper, we propose a new method called Cumulative Reasoning (CR), which employs language models in a cumulative and iterative manner to emulate human thought processes. By decomposing tasks into smaller components, CR streamlines the problem-solving process, rendering it both more manageable and effective. For logical inference tasks, CR consistently outperforms existing methods with an improvement up to 9.3%, and achieves the astonishing accuracy of 98.04% on the curated FOLIO wiki dataset. In the context of the Game of 24, CR achieves an accuracy of 98%, which signifies a substantial enhancement of 24% over the previous state-of-the-art method. Finally, on the MATH dataset, we establish new state-of-the-art results with 58.0% overall accuracy, surpassing the previous best approach by a margin of 4.2%, and achieving 43% relative improvement on the hardest level 5 problems (22.4% $\rightarrow$ 32.1%).

垂类语料库建设



案例：联合顶尖高校图书馆、研究院定制医学垂类大模型

- 基于海量药学知识和专家经验,研发行业前沿的用药咨询语言大模型,将为患者提供更智慧便捷高效的线上用药咨询服务和体验。
- 借助大模型辅助进行特色病、疑难病的诊疗,有效推动医疗长尾问题的解决。
- 可基于语言大模型打造的智慧就医平台,可提供智能导诊、智能预问诊、诊前检验、智能随访、院内导航等多种功能,实现了诊前、诊中、诊后多个环节的数字化升级,重塑患者就诊流程,实现门诊环节全流程智能化,缩短患者在院停留时间,优化患者就医体验。





案例：定制高校AI机器人



招生机器人个性化设置



高校招生机器人24小时人性化沟通



前台交互方案展示



支持自主模型选择

全部模型

公开模型

与我相关

自然语言处理

文本分类 分词分类 表格问题回答

回答 零样本分类 文本翻译

文本总结 对话能力 文本生成

文本生成文本 完形填空 语句相似性

文本修正 情感分析 文本比对

语义理解

语音

文本生成语音 自动演讲识别

音频生成音频 音频分类 音频识别

语音识别 语音合成 音频分析

语音翻译 语音评测 语音增强

多模态

特征抽取 文本生成图片

图片生成文本 文本生成视频

可视化问题回答 文档问题回答

图机器学习 组合能力

公开模型 共 4 个

请输入模型名称关键字

chinese_llama_alpaca_2

中文alpaca模型 (ymcui)

文本生成 框架: torch2.0-tf2.11.0-cuda11.7-ubuntu20.02-llm

AICloud 更新于2023-09-01 17:06:22 浏览量: 395

Notebook微调

零代码微调

chatglm2_6b

chatglm2_6b

文本生成 框架: torch2.0-tf2.11.0-cuda11.7-ubuntu20.02-llm

AICloud 更新于2023-08-30 18:14:04 浏览量: 31

Notebook微调

零代码微调

llama2_7b

llama2_7b

文本生成 框架: torch2.0-tf2.11.0-cuda11.7-ubuntu20.02-llm

AICloud 更新于2023-08-16 14:23:05 浏览量: 1348

Notebook微调

零代码微调

chatglm_6b

ChatGLM-6B是一个具有62亿参数，并且支持中英问答、多轮对话的中英双语语言模型，并支持在单张 2080Ti 上进行推理使用。

文本生成 框架: torch2.0-tf2.11.0-cuda11.7-ubuntu20.02-llm

AICloud 更新于2023-08-16 13:58:47 浏览量: 244

Notebook微调

零代码微调

共 4 项数据

< 1 >



可视化模块设计





无代码指令式操作方式





无代码指令式操作方式



谢谢观看

