大模型落地应用案例集

Foundation Model
Practical Application Collections

2023

牵头单位

大模型测试验证与协同创新中心

/ 主编单位 /

中国信息通信研究院华东分院
中国信息通信研究院人工智能研究中心
上海人工智能实验室开源生态发展中心

Foundation Model
Practical Application Collections

2023大模型落地应用案例集

牵头单位

大模型测试验证与协同创新中心

2023

/ 主编单位 /

中国信息通信研究院华东分院
中国信息通信研究院人工智能研究中心
上海人工智能实验室开源生态发展中心

编辑委员会

主编单位

中国信息通信研究院华东分院

中国信息通信研究院人工智能研究中心

上海人工智能实验室开源生态发展中心

对案例集的参编单位表示感谢：

阿里云计算有限公司

北京百度网讯科技有限公司

北京九章云极科技有限公司

北京泡泡玛特文化创意有限公司

北京水木分子生物科技有限公司

北京信工博特智能科技有限公司

北京智谱华章科技有限公司

北京中科汇联科技股份有限公司

东方财富信息股份有限公司

恒生电子股份有限公司

华为技术有限公司

京东云

九度数字科技（苏州）有限公司

昆仑万维科技股份有限公司

蚂蚁科技集团股份有限公司

蚂蚁星河 (重庆) 信息技术有限公司

OPPO 广东移动通信有限公司

软通动力信息技术（集团）股份有限公司

三六零安全科技股份有限公司

上海百川智能技术有限公司

上海传之神科技有限公司（OpenCSG）

上海道客网络科技有限公司

上海氪信信息技术有限公司

上海蜜度科技股份有限公司

上海人工智能实验室

上海商汤智能科技有限公司

上海说以科技有限公司

上海昇腾人工智能生态创新中心

上海特赛发信息科技有限公司

上海天壤智能科技有限公司

上海稀宇科技有限公司

上海岩芯数智人工智能科技有限公司

上海智象未来计算机科技有限公司

上海众深科技股份有限公司

上海卓繁信息技术股份有限公司

壹沓科技（上海）有限公司

优刻得科技股份有限公司

云从科技集团股份有限公司

云南联合视觉科技有限公司

云知声（信阳）数字科技有限公司

支付宝（中国）网络技术有限公司

中国金茂控股集团有限公司

中国商飞上海飞机设计研究院

中企网络通信技术有限公司

竹间智能科技（上海）有限公司

（* 按单位首字拼音排序）

目 录

CONTENTS

（* 案例排名不分先后）

第一章　通用大模型

基于人工智能大模型技术的开放平台 …… 10
可控可信的私域知识问答系统 …… 14
MiniMax 大模型医疗咨询解决方案 …… 20
言犀基础大模型 …… 24
国内首款可私有化部署的企业级数据分析智能体——TableAgent …… 30
九章云极知识管家打造企业专属大模型智能底座 …… 34
“Pixeling 千象” …… 38
书生筑梦视频生成大模型 …… 44
书生浦语开源大模型 …… 48
百川大模型在娱乐领域的应用 …… 52
AnimateDiff：一项基于个性化文生图模型扩展后的视频生成框架 …… 54
通义千问 2.0 在企业场景的应用 …… 58
昆仑万维“天工”大模型 …… 60

第二章　垂类大模型

梧桐·招聘 - 基于百度智能云千帆大模型平台的智能招聘系统 …… 66
面向游戏行业的图像内容生成式大模型 …… 70
中公网校：小鹿老师，为年轻人创造更多就业与成长机会 …… 74
新华妙笔 AI …… 80
小布助手 …… 84
ChatDD 新一代对话式药物研发助手 …… 88
大模型数据分析智能助理 DeepInsight Copilot …… 94
单晶炉自动化工艺识别多模态大模型 …… 98
基于 NDAI 大模型的政务元宇宙平台 …… 104
慧政大模型——面向政务服务垂直大模型 …… 108
基于循道政务大模型的免申即享系统示范应用 …… 112
东方财富自研金融大模型 …… 116
基于大模型的信息结构化抽取方法 …… 120
天津金城银行金融大模型示范应用 …… 124
文修大模型助力中文校对提质增效 …… 128
新型金融风险防范可信金融大模型 …… 132
信阳市智慧工业平台 …… 136
遥感大模型在农业信贷场景的应用 …… 142
中国金茂人工智能大模型企业内部场景应用 …… 146
中山大学附属医院智慧医院项目 …… 150
阿斯利康：基于学术文献溯源的药品不良反应报告生成助手 …… 156
基于知识图谱和大语言模型的制造业数字化转型平台 …… 160
东方翼风大模型 …… 166
智己汽车：用大模型打造智能时代出行变革者 …… 170
基于山下话童大模型的贷后催收示范应用 …… 174
海淀区一网统管接诉即办工程项目 …… 178
风乌气象大模型 …… 182
基于大模型的智能培训 …… 184
面向围手术期的医专大模型研究及其落地应用 …… 186
通过大语言模型与材料领域技术文件集合对原材料质保书进行智能审查 …… 192
智能投顾助手——光子·善策 …… 194

第三章　大模型服务

支小助 - 大模型金融专家智能助理 …… 200
AGI 云上模型服务平台 …… 204
蚂蚁集团大模型数据高质量供给平台 …… 210
基于大模型的壹沓数字员工超自动化平台 …… 214
云原生大模型知识库平台 …… 216
众调科技：营销 AI 培训产品 …… 220
信息安全大模型平台 …… 222
全自研 AI 整合平台“HeyLisa” …… 226

Chapter One.

1

第一篇章

通用大模型

2023

—

大模型落地应用案例集

Foundation Model
Practical Application Collections

基于人工智能大模型技术的开放平台

上海天壤智能科技有限公司

天壤智能是国家高新技术企业，上海市专精特新企业。公司聚焦人工智能深度学习和大数据挖掘技术，开发大模型、小样本、多重迭代的人工智能决策优化算法，致力于打造复杂场景下的智能决策辅助体系和通用人工智能平台。现已成功落地生物制药、智慧交通、智慧商业、数字金融等多个领域。

概　述

本项目通过搭建高性能 GPU 计算集群、训练通用大语言模型、训练垂类大语言模型、搭建大语言模型微调平台、搭建大语言模型应用开放平台等核心模块，旨在打造大语言模型服务和应用平台，为大语言模型技术的研究和应用提供一个开放、可扩展、可协作的环境。这个平台除了通用大语言模型外，还提供大量共享的数据集、算法库、模型微调工具等资源供开发者使用，同时大语言模型应用开放平台提供一整套完整的大语言模型生态应用工具链，从而加速大语言模型的训练以及大语言模型生态应用的开发和使用过程。

需求分析

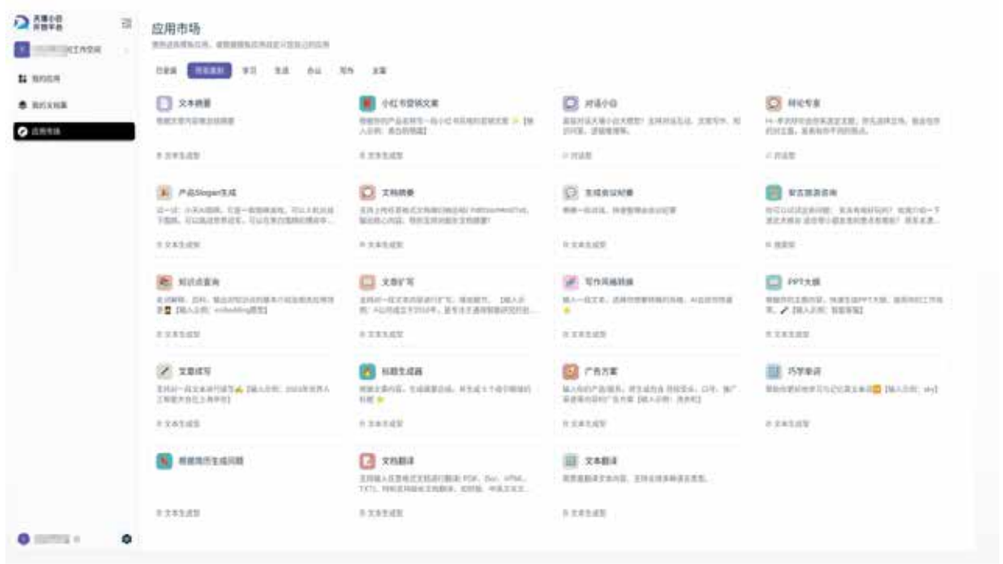
随着生成式人工智能技术步入深化阶段，以 chatGPT 为代表的大语言模型潜力凸显，在各个领域得到了广泛的认同和应用。2022 年全球 GenAI 市场整体收入为 400 亿美元，预计 2027 年及 2032 年将分别达到 3990 亿美元和 1.3 万亿美元，2022~2032 年复合增长率高达 42%。而国内众多行业企业受到算力和数据等因素的制约，不能快速高效地使用最新的 AI 工具和成果。因此，建设一个高性能、稳定可靠的大模型开放平台，从而降低人工智能应用的门槛，提高开发效率和降低开发成本，促进人工智能领域的合作与交流，加快人工智能技术的创新与应用，成为了一个非常有意义的工作。

案例介绍

大语言模型开放平台旨在为大语言模型技术的研究和应用提供一个开放、可扩展、可协作的环境。该平台不仅为开发者提供大型语言模型、大规模数据集、模型微调工具以及大型语言模型应用开发工具等资源，还加速了大型语言模型的训练过程，促进了大型语言模型在不同领域的应用落地。

核心能力和技术创新点：

- **1860 亿参数规模的基础大模型：**该平台配备了规模庞大的基础大型模型，拥有卓越的内容生成、智能推理、语义检索、情景感知和多语言转换等智能交互能力。这使得用户能够在各种应用场景中更灵活地应用大型语言模型。
- **丰富大模型类型：**平台提供 130 亿参数通用大模型、130 亿参数代码专用大模型、130 亿参数 SQL 专用大模型以及 130 亿参数 10K 上下文专用大模型，以满足不同领域的需求，从通用应用到特定任务，都能得到支持。
- **大型语言模型服务接入：**平台支持大型语言模型服务的接入，为开发者提供了广泛的选择，使他们能够根据具体要求轻松定制模型。
- **微调与部署能力：**平台提供方便快捷的大型语言模型微调和部署功能，让开发者能够快速生成多领域的定制模型，以满足特定应用的需求。
- **高效的应用开发能力：**平台支持多项技术创新，包括提示词工程、敏感词检测、多格式文件输入增强以及文档集搜索增强，这些技术创新使应用开发更加高效。
- **多渠道支持：**平台支持 WebAPP 页面应用以及后台 API 调用管理，提供了多种应用渠道，以满足不同应用场景的需求。
- **资源动态调配：**基于分布式计算集群的资源动态调配，确保平台在不同负载下的高效性能，为用户提供卓越的体验。
- **高效模型训练：**平台提供适用于不同场景的预训练模型，基于预训练模型的专业模型优化，极大地削减了模型开发周期和资源成本。



应用落地与合作机构：

目前，该开放平台已进入内测阶段，吸引了企业用户 1000+ 位，实现了 200+ 个大型模型应用的开发。平台与多家重要合作机构建立了合作关系，其中包括中国人民解放军军事科学院、国防科技大学、中科院生命科学研究院、苏州超算中心、加拿大 Ploytide 生物科技有限公司等等。这些机构基于平台提供的大语言模型应用建设能力，共同推动了大型语言模型技术的应用和研究。

效益分析

该平台的建成能提高企业的大语言模型应用开发速度，降低开发成本，并提供了良好的商业模式：平台可以通过提供专业领域增值服务、付费订阅等方式从用户中获取收益，从而推动平台的可持续发展，而平台本身的开放性和共享性也能够吸引更多的开发者加入，进一步扩大平台的规模和影响力。此外，本项目可以推动人机交互模式和工作模式的变革，加速 AI 应用的落地和普及，从而营造大模型产业生态。

可控可信的私域知识问答系统

上海岩芯数智人工智能科技有限公司

RockAI（岩芯数智）是以认知智能为基础，专注于自然语言理解、人机交互的科技创新型企业，是A 股上市公司（002195.SZ）上海岩山科技股份有限公司的控股子公司，公司秉承“新科技改变生活”的理念，致力于构建自研基础 AI 大模型 + 行业垂直模型的技术结构，实现“1 个 MaaS 平台，多种应用场景”策略，打造客户信赖的认知智能平台。

概　述

私域知识问答系统是一种旨在满足特定组织或团队内部需求的智能信息获取工具。其产品形式包括：知识问答、企业助理、办公助手、智能客服、数字员工等。岩芯数智通过自研构建可控可信的通用大模型，缓解了行业中大模型幻觉问题，提升模型的精准问答能力，回答准确率达到 90%，目前已在多家企业内部部署应用。

需求分析

信息是解决问题的基础，在企业和组织内部，员工和团队通常需要访问特定领域的知识和信息，以解决问题、获得支持或做出决策。传统的知识库和文档系统可能存在检索和更新的问题，导致信息不易获取。

在传统的知识获取中，用户将知识库放入到全文索引库中，然后用户利用关键词获取全文检索的结果，即属于当前传统搜索引擎的模式，该模式下主要存在以下两方面的问题:

- 全文检索的方式需要关键词精准命中，对用户的输入要求更高。
- 全文检索命中的是相关性，只是找到答案附近的文本，无法精准定位答案。

本私域知识问答系统的背景是通过结合岩芯数智可控可信的通用大模型，提供一种更智能、互动和高效的方式来访问和共享知识。

案例介绍

基本流程：

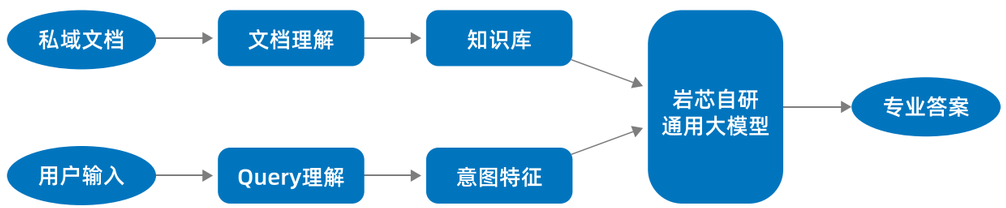


图 1　基本流程

主要能力：

- 知识管理

私域的知识问答系统具有强大的知识管理能力，允许用户创建、编辑和组织知识文档、常见问题解答（FAQ）、操作手册和培训材料等；

- **多轮点的知识问答**

系统提供高效的问答功能，用户可以轻松查找所需的信息，以减少时间浪费和提高生产率；

- **权限管理**

系统提供灵活的权限管理，确保只有授权人员可以访问和编辑特定的知识文档，以维护知识的安全性和可维护性。

技术创新：

技术上为缓解大模型的幻觉问题以及提升模型回答问题的准确性，岩芯数智专研模型的可控可信能力。模型结构采用岩芯数智完全独立自研的可线性计算的自然语言关联特征表示模型，相比基于 Attention 机制的 Transformer 架构大模型，可大幅度的提升模型训练效率和应用效果。

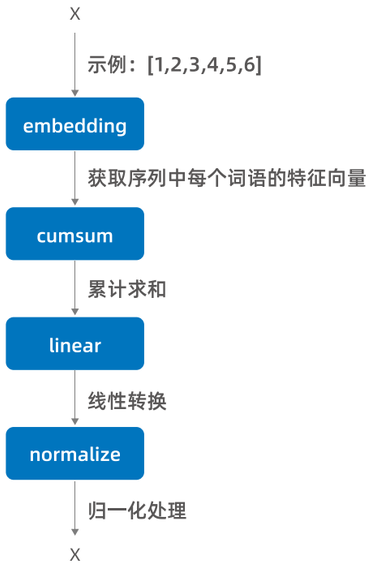


图 2　自然语言关联特征表示的简单示例过程

模型的结构基础是基于线性计算的自然语言关联特征表示方法，为了增强的应用能力，需对线性计算的特征进行不断地叠加。

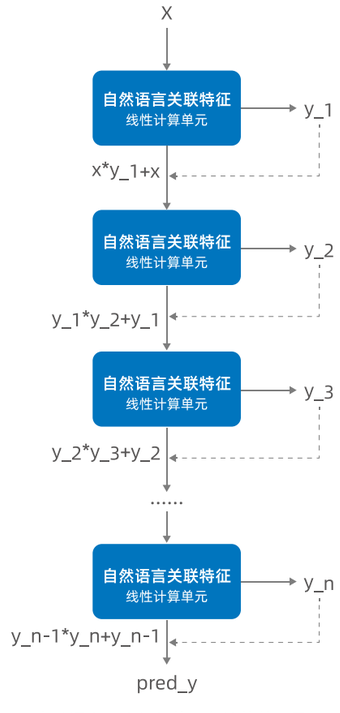


图 3　自然语言关联特征的现象计算单元叠加示例

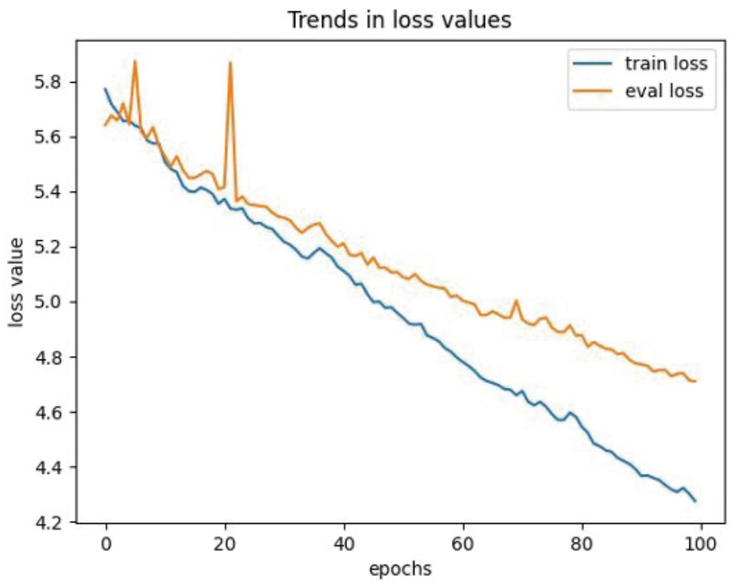


图 4　标准 Transformer 架构训练某一任务 loss 变化趋势

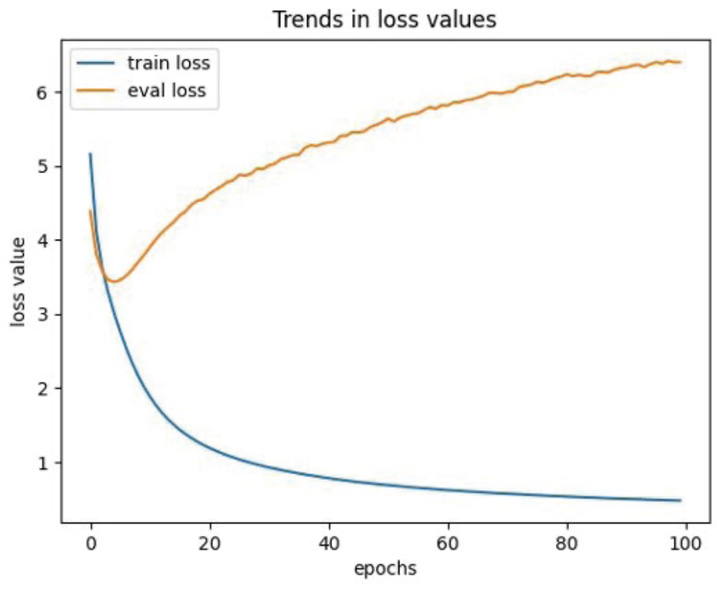


图 5　岩芯数智自研模型训练某一任务 loss 变化趋势

图 4 与图 5 是相近参数量下，针对同一任务的训练，标准 Transformer 模型与岩芯数智自研模型的 loss 变化趋势。

图 4 为基于 Transformer 架构的预训练模型在训练集和验证集上的损失值表现情况，图 5 为岩芯数智自研大模型。可以发现在训练 100 个 epoch 下，岩芯数智自研模型已经出现过拟合的现象，其中验证集中的最低损失值在 3.5 左右，而基于 Transformer 架构的大模型，在 100 个 epoch 下未完成收敛，且验证集中损失值依然在 4 以上。

上述也表明改进后的模型具备更快的收敛效率，基本上在第 10 个 epoch 下就达到了最佳状态，因此收敛效率远高于 Transformer 架构。

实施效果：

- **提高生产力**

处于私域的人员能够更快地找到所需的信息，解决问题，减少工作中的困惑，从而提高生产力；

- **知识共享**

促进了内部知识共享和协作，有助于打破信息孤岛，使组织更加协调一致；

- **风险降低**

通过更好的知识管理，组织可以减少风险，提高合规性。

应用落地情况：

本私域的知识问答系统已经在多家企业内部落地，回答准确率达到 90% 以上。

效益分析

经济效益

私域的知识问答系统有助于提高生产力和效率，减少支持部门的负担，降低了组织的运营成本；

商业模式

通过许可付费、订阅付费以及自定义解决方案三种方式实现用户付费；

应用推广前景

- **企业内部应用：**私域的知识问答系统在企业内部可以用于知识管理、员工培训、问题咨询以及改善组织内部的工作流程；
- **教育领域：**学校、大学和教育机构可以本系统来改善教育过程，促进学生之间和教师之间的知识共享，提高教育质量；
- **医疗健康领域：**可以提高医疗专业人员之间的知识共享，改善患者护理，提高医疗服务质量；

私域的知识问答系统可以在各种领域都有广泛的应用，为组织带来经济效益、社会效益，同时提供多样化的商业模式选择。

MiniMax 大模型医疗咨询解决方案

上海稀宇科技有限公司

MiniMax 成立于 2021 年 11 月，是一家专注于通用人工智能的科技创业公司。成立至今，MiniMax 自主研发了“MiniMax-abab”文本、语音、视觉三模态的千亿参数大语言模型，在中、英文服务领域均已超过 GPT-3.5 的水平。2023 年 8 月，“MiniMax-abab”大模型通过了国家首批大模型服务备案，可以面向社会公众提供服务。 立足自研的大语言模型，MiniMax 布局 2B、2C 业务，是商业化落地最快的中国大模型初创企业之一。在赋能千行百业方面，公司的 MiniMax 开放平台已服务数百家行业客户，是公用云上在线调用量最大的大模型开放平台，在金山办公、腾讯、小米、阅文、小红书等多个行业头部客户取得了实际落地。在服务终端用户方面，已在国内上线“星野”、“应事”等多个 APP。

概　述

项目背景

在我国的医疗健康产业领域，医疗咨询场景对于专业度与紧迫性要求极强。MiniMax 发挥算法优势，突破应用落地，协同药师和患者双方进行辅助咨询，助力实现全体公民的健康福祉。

技术解决路径

面对庞大的患者数量、极高的专业性要求与人工成本，MiniMax 为医疗咨询行业提供了解决方案，通过协助药师定期回访并回答患者的专业问题，极大提高了服务效率和专业水平。

精准学习垂类医疗领域知识

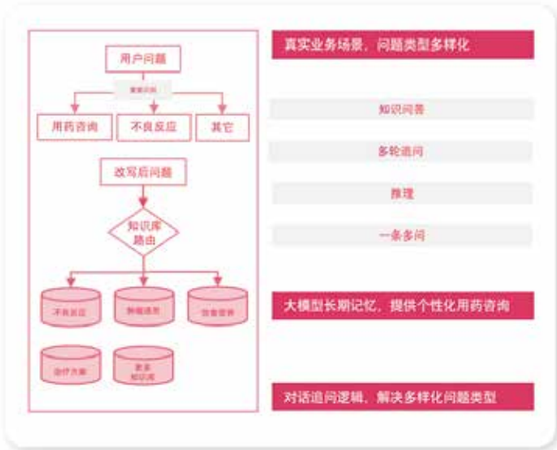
挑战 1：大语言模型在专业知识方面缺乏有效回应

MiniMax 的解决方案：构建外挂知识库，提升通用大模型回答垂类领域问题的准确性。

轻松应对个性化、多样化的用户提问

挑战 2：如何回答个性化、多样化的用户问题类型并给予针对性回复

MiniMax 的解决方案：凭借大模型长记忆能力，进行多轮对话，提供给个性化用药咨询。



准确理解问题意图与专业概念

挑战 3：尽管大型语言模型在处理问题意图和回答问题方面取得了进展，但对医学文献中概念的理解能力仍有待提高

MiniMax 的解决方案：通过提高大语言模型自身的能力、应用少样本提示和模型微调等方法，结合多种手段以提升模型对专业领域概念的理解。

需求分析

医疗咨询行业的困境——以肿瘤治疗为例

- **肿瘤治疗的成功是实现全民健康的关键环节**

中国每年新增肿瘤患者超过 400 万，每年去世肿瘤患者大约 300 万。这一数据表明，肿瘤治疗已成为全民健康实现的关键环节。

- **院外个性化用药咨询对肿瘤患者不可或缺**

在肿瘤患者就医的全周期中，院内时间仅占 10% 左右，而其余 90% 的时间都在院外度过。在此期间，患者的用药依存性、转移情况、疗效追踪、不良反应、营养情况、心理状态等方面都需要药师或医生及时随访，并提供相应的咨询服务。

- **以患者为中心的人工服务面临挑战**

以患者为中心的人工服务在专业度、成本和可扩展性方面存在诸多挑战。首先，数十种肿瘤病种和数百种抗肿瘤药物的知识体系庞大且复杂；其次，单病种就有成百上千篇知识文档，对药师的持续学习能力要求很高；最后，1 名药师每个月最多能服务 300-350 位患者，这意味着需要数千名药师来满足患者的需求。以上如此庞大且急切的需求单凭人工手段无法得到全部满足。

案例介绍

" 高济神农 " 是高济健康与 MiniMax 共同打造的智能患者管理系统。基于 MiniMax-abab 大语言模型，它构建了包含数亿条医学专家指南和共识的肿瘤知识库，同时包括营养、心理、疾病知识、康复预后等内容。通过知识增强技术外接到大模型中，以高济累积的超 80 万肿瘤患者真实服务场景为基础，经过 200 余家药房药师的反复调试优化，对于肿瘤用药及不良反应问题的回答准确率高达 97.6%。

“高济神农”智能患者管理体系三大落地应用：专为药师打造的**高济 HealthMate 智能助手、智能随访系统、数字人用药指导解读。**

高济 HealthMate 可以根据患者档案，在用药、不良反应、不良反应指导、饮食营养等多个维度辅助药师做出更准确、迅速、个性化的判断和建议，并建设流式回答来减少患者等待时间。通过**智能随访系统**，药师只需要输入对患者的基本情况、不良反应症状等信息，系统会给出相对应的处理建议。药师评估判断患者的健康情况，它便能生成更人性化、个性化以及易读性的随访小结，提升药师工作效率和患者体验度。给患者的用药指导、随访小结等相关内容都经过专业药师审核。同时，“高济神农”还利用**数字人技术**为老年患者提供易理解的药品说明和营养建议视频，帮助他们享受更便捷的互联网医疗体验。

“高济神农”是一次对肿瘤患者安全用药管理未来方向的探索。在进博会上发布最新“高济神农 2.0”，增加了智能院外患者管理体系，旨在通过持续不断洞察患者需求，提升患者服务体验。

效益分析

经济效益

高济神农产品的使用，可以有效提升高济药师咨询服务的专业水平，提升服务质量。同时，扩大可服务病患人群，覆盖数倍于之前规模，助力提升全民健康。

商业模式

针对知识库构建和调优服务一次性收费，同时，按照问答消耗的 token 数量，依照实际调用量按量计费。

应用推广前景

高济神农合作项目的知识库解决方案，有效提升了药师的专业性，可覆盖更多人群。知识库方案在医疗行业中，针对需要为患者提供专业咨询服务的场景，具备很强的可复制性，市场潜力很大。

言犀基础大模型

京东云

京东科技是京东集团旗下专注于以技术为政企客户服务的业务子集团，秉承科技引领、助力城市及产业数智化升级的使命，我们致力于为政府、企业、金融机构等各类客户提供全价值链的技术性产品与服务。基于人工智能、云计算、大数据、物联网等前沿科技，依托京东多年耕耘供应链的积累，京东科技是最懂产业的数智化解决方案提供商，面向不同行业提供以供应链为基础的数智化解决方案。

2021 年 1 月，京东科技在原京东数科与京东智联云基础上重组完成，融合了两大技术业务板块的综合实力，京东科技现已成为整个京东集团对外提供技术服务的核心平台，拥有丰富的产业理解力、深厚的风险管理能力、用户运营能力和企业服务能力，能面向不同行业为客户提供行业应用、产品开发与产业数字化服务。

概　述

京东作为一家新型实体企业，拥有着深厚的产业基因和供应链场景，源于真实的业务需求、深度复杂的场景任务和广泛的实体经济发展要求，促使京东的 AI 技术是面向知识密集型、任务型场景，解决真实产业问题的技术。且京东云旗下的言犀团队在任务型智能对话交互关键技术方向拥有丰富的积累和广泛的落地，拥有包括文本生成、语音生成、对话生成等系列领先技术，并打造出了智能客服系统、京小智平台商家服务系统、智能政务热线、言犀数字人等系列产品和解决方案。

京东科技深耕人工智能领域多年，形成了从算法到应用场景的链路，并通过自研推出言犀基础大模型，赋予客户在各自行业中快速构建、部署，应用人工智能的能力。通过言犀基础大模型，企业可以建立从业务的大量数据中自我学习、自驱迭代的能力，并实现对企业实施、运营、维护的一体化覆盖，同时言犀大模型中的小型化技术能够使企业具备云管边端协同运营的能力，增强企业面对非标准化、算力通讯资源受限场景的应对能力，和面对业务变化的快速响应能力。

需求分析

目前基础大模型正处于蓬勃发展阶段，各行业、各领域以构建数字化、线上化、搭建虚拟仿真场景为主要应用。在当前阶段下，基础大模型面临着以下问题：

一、 由于基础技术的限制以及大部分企业在大模型应用和硬件设备开发能力的不足，从而导致无法自主生产原生 AI 模型。

二、 除了用户单点大模型开发技术能力的不足，在各行业链条中的软硬件互通、数据标准化和应用功能融合等问题中都存在着无法克服的壁垒。

三、 大模型训练硬件的能源消耗问题，在当前全球绿色经济的背景下，平稳运行离不开大规模的数据中心和云计算中心等基础设施的支撑，而大部分企业则无法满足以上的要求，从而无法实现大模型的应用。

针对以上情况，开发言犀基础大模型，以实现低门槛构建基于人工智能技术的解决方案，是本项目需要解决的问题。

案例介绍

京东推出的言犀基础大模型，将着力围绕内容生成、人机对话、用户意图理解、信息抽取、情感分类等几大类任务，围绕零售、物流、金融、健康、政务场景进行落地应用。

1) 优质的场景和数据让模型产业属性更强

京东的言犀大模型，是扎根产业的原生大模型。凭借着从基础设施、模型层、MaaS 层、SaaS 层全栈的技术布局，打造多款端到端的大模型技术产品。

言犀大模型拥有三个差异化的特性：

第一，它是产业原生的，有更强的产业属性。

第二，它是价值驱动的，有更高的应用价值。

第三，它是开放协同的，有更快的迭代效率。

图 1　京东言犀大模型概览

另一方面，京东连接着产业互联网和消费互联网，在对内实践和对外产业数智化过程中积累了众多优质的数据，区别于一些通用域数据的静态数据，京东的数据是“鲜活的”，凭借每年产生数百亿的交互数据，保证了模型的持续迭代和优化。

京东的大模型是在预训练阶段就接了 70% 通用域数据和接近 30% 京东特有的产业数据相结合去做训练，这就保证了模型拥有大模型的“常识”，并拥有产业模型的“专业”。

图 2　京东言犀大模型数据概览

2) 京东言犀大模型技术架构

京东言犀大模型是基于京东云的高性能计算集群，采用 Megatron+DeepSpeed 的分布式训练框架，训练的 Decoder-Only 架构模型。在通用知识获取方面，言犀大模型添加了约 30% 的京东域自身的产业数据，并通过构建高质量的指令数据，帮助模型具备更强的产业属性。除了模型训练本身，京东言犀大模型还在模型的转换层和服务层进行了自研算法的深耕，提升了大模型本身的推理速度和部署性能，让大模型的能力能够充分的下沉到业务端，并通过集成平台能力打造真正的模型及服务。

图 3　京东言犀大模型架构

3) 前沿的算法能力保证模型具备高应用价值

- 预训练层面

源于业务应用需求，京东在 2020 年就提出了 K-PLUG 模型，将领域知识注入大模型中，以提高大模型的专业性和忠实度，并在 2021 年对该项工作进行了发表。K-PLUG 方法是基于 Transformer 模型架构 X 京东的产业知识进行的预训练。

该算法帮助模型在实体属性抽取准确率为 96%；在生成式多轮对话 ROUGE-L（指标主要是对比机器生成的内容与人类的标准内容的匹配度），以 27% 领先于斯坦福经典的 Pointer-Generator；在上下文多轮问答知识检索率以 74% 准确率领先于行业。

图 4　京东言犀大模型 K-PLUG 算法

- 推理部署层面

除了在大模型的预训练阶段，言犀大模型通过上述算法增强了产业领域知识，在模型的推理层面，京东言犀采用量化矩阵算子融合、自适应参数矩阵量化、自动算子切分与卡间并行、内存优化与缓存等多种策略，将推理速度提升 6.2 倍，且在“首字”推理速度的大模型推理难点上，京东言犀大模型采用自研的算法，极大程度的提升了大模型在推理方面的性能。在部署方向，依靠流式推理有效解码传输机制、动态批处理、异构集群部署等方法，将部署成本降低了 90%。

此外，京东言犀大模型还拥有配套的 AI 开发计算平台，用于快速的模型迭代，效率提升 10 倍以上，让模型能够不断的学习新的知识。

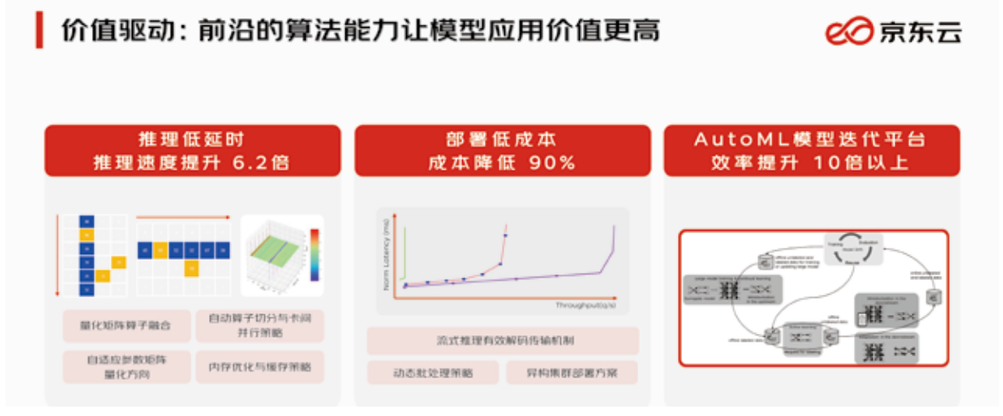


图 5　京东言犀大模型推理部署

4) 澎湃算力打造开放协同的大模型生态

为了训练大模型，京东早在 2021 年就在重庆建成了大模型集群，也是全国首个基于 DGX SuperPOD 架构的超大规模计算集群—天琴α，该集群在保障京东自身大模型训练的同时，还将集群的每秒浮点运算次数提升 40%，多卡线性加速比提升 90%，为后续大模型的持续发展打下良好的基础。

另一方面，为了更好的应对大模型背景下的海量数据存储问题，京东还自研了向量数据库 Vearch，支持百亿级向量监检索，召回实现毫秒级延迟，智能储存分层实现成本降低 60%，大幅提升了模型推理泛化能力与推理效率。

效益分析

该解决方案以京东全产业链为核心优势，从产业场景、软件平台、安全合规等多个方面为用户带来价值。

- **产业场景方面：**用户将借助京东积累的历史行业知识，低成本快速构建该用户所在细分领域大模型应用，使用户快速取得局部市场的先发优势（量化标准为缩短开发周期及成本降低）。
- **软件应用方面：**为了让模型有更好的能力和应用，京东将开发言犀大模型过程中积累下来的能力解耦整合出来，以大模型开发平台的形式开放给京东的合作伙伴。该平台以京东云的私有云、公有云和混合云等高性能计算集群为底座，内置了包括数据、模型训练和部署推理等工具能力，不仅支持京东自身的言犀框架，也同时会支持各个主流的开源模型框架，促进大模型生态的发展。
- **安全合规方面：**言犀 AI 大模型具备数据隐私和内容安全可控的价值：

5) 数据隐私安全

我司在人机交互研究中进行训练数据处理、人工智能模型的训练时，严格遵守使用深度合成技术中的个人隐私保护要求，确保训练数据数据来源合法性，并使用脱敏数据进行模型训练。

6) 内容生成可控性

恶意代码、插件和网络钓鱼电子邮件有可能被 ChatGPT 生成。为了杜绝此安全隐患，京东云言犀团队会在模型训练时引入人工反馈机制降低和杜绝模型生成有害信息的回复。同时引入审核 API 来阻止某些有害内容的输出，例如，当收到要求编写用于从被黑客攻击的设备窃取数据的代码或制作网络钓鱼电子邮件时，模型会拒绝该要求并指出此类内容是“非法、不道德且有害的”。

国内首款可私有化部署的企业级数据分析智能体——TableAgent

北京九章云极科技有限公司

北京九章云极科技有限公司（简称：九章云极 DataCanvas）以“创造智能，探索未知”为使命，以“助力全球企业智能升级”为愿景，是中国人工智能基础软件领军者。公司致力通过自主研发的人工智能基础软件产品系列和解决方案为用户提供人工智能基础服务，助力用户在数智化转型中轻松完成模型和数据的双向赋能，低成本高效率地提升企业决策能力，实现企业级 AI 规模化应用。

九章云极 DataCanvas 的核心产品系列 AIFS 人工智能基础软件和 DataPilot 数据领航员具有高度的灵活性和可扩展性，能够处理各种类型和规模的数据，简化了数据处理和分析的复杂性。产品集成了一系列先进人工智能技术，包括多模态向量数据库、因果学习、思维件等，为企业提供 AI 软件开发新范式。

概　述

大模型技术催生了数据分析技术的进一步跨越，通过将大模型技术和具体的业务深度融合，数据分析成为直接为企业用户产生更富有商业价值的应用领域。基于 DataCanvas Alaya 九章元识大模型微调出 Alaya-ZeroX 模型组，开发的 TableAgent 数据分析智能体，是从 0 到 1 的交互式结构化数据分析的突破，提供私有化部署方案，保障了业务数据的安全合规，是企业数据分析的全新方式。作为国内首款可以实现私有化部署的企业级数据分析智能体，TableAgent 在充分理解用户意图后，可自主的利用统计科学、机器学习、因果推断等高级建模技术从数据中挖掘价值，进而提供分析观点和指导行动的深刻见解，赋予企业用户具备数据分析师的能力。

2023 年 11 月 21 日，九章云极 DataCanvas TableAgent 产品面向社会开放公测（地址：https://tableagent.datacanvas.com），助力企业借助大模型技术发挥数据价值，提高企业生产经营效率。

需求分析

数字化时代，数据分析的重要性犹如空气般无处不在。商业数据分析是数字化管理、智能决策的基础，同时数据分析又是一个专业性极强的工作，描述性分析、诊断性分析、预测性分析，会让大多数只会用 Excel 的人望而生畏。

作为一款企业级应用，业务数据的安全性、合规性不可忽略，一款可以私有化部署的方案在企业利用大模型技术进行数据分析应用落地迫在眉睫。

九章云极 DataCanvas 公司自主研发的 TableAgent 数据分析智能体，可以实现私有化部署，保障安全、合规的前提下，让大模型对个人生产力的赋能，从写纪要、做总结上升到新的台阶，只要会提问，就能成为一个高级的数据分析师，洞察数据奥秘。

案例介绍

一、主要能力

TableAgent 是在 DataCanvas Alaya 九章元识大模型基础上开发的能够实现私有化部署的企业级数据分析的智能体，有非常强大的意图理解能力、分析建模能力和洞察力。TableAgent 在充分的理解用户意图后，自主的利用统计科学、机器学习、因果推断等高级建模技术从数据中挖掘价值，进而提供分析观点和指导行动的深刻见解。

二、技术创新点

TableAgent 是从 0 到 1 的交互式结构化数据分析的突破，是企业数据分析的全新方式。基于核心研发团队丰富的数据分析经验和技术创新探索，TableAgent 能够在强大的 Alaya 九章元识大模型上微调出功能稳定、高效的数据分析能力。

1、在 Alaya-ZeroX 模型组开发的同时，TableAgent 针对企业用户领域微调的需求配套设计了 T+ 系统，能够高效的实现定制化的微调工作，系统性的体系支撑更高效的实现数据分析各个环节的升级，让用户在无感知的情况下即可获得不断升级的数据分析体验。

2、TableAgent 融合了公司多个前沿技术成果，除了在 Alaya 大模型的基础，还同时运用了自研开源大模型工具链、融合了 DAT 自动机器学习和 YLearn 因果学习算法成果，因此融合了强大的自动化、因果可解释的 AI 能力。

3、TableAgent 提供了效率提升明显的、确保数据分析成果更可用的数据分析和能力，为企业场景的数据利用带来更多可能性。

三、应用落地情况

TableAgent 前身为九章云极 DataCanvas 公司在 6 月 30 日发布的 TableGPT，该产品已经在公司内部经过四个多月的内测试用，期间我们不断升级能力和体验。在算力准备充沛之际，于 11 月 21 日面向社会开放公测。截至目前，已经在金融行业的客户流失预警、产品定价，互联网行业外卖平台的推荐优化、订单转化，以及油气、零售、地产等多个业务场景得到应用，帮助数据分析师更加高效赋能业务，提高企业经营效能。未来，TableAgent 将进一步融合非结构化数据的分析能力，并与公司自研的 DingoDB 多模向量数据库、DataCanvas Alaya 九章元识大模型联合创新，在复杂分析任务、自动化、人机交互、智能体协同等方面进一步升级。

效益分析

一、经济社会效益

1、促进产业升级：企业在生产经营过程中，每天将产生大量的数据，尤其是互联网行业，数据量将达到近百亿规模。在数智化升级过程中，面向业务的人员的需求，技术人员需要快速响应。TableAgent 的应用，可以助力企业高质量的完成分析工作，赋予智能化决策测能力，提高企业经营效能，为企业催生巨大的商业价值，推动产业升级发展。

2、加强 AI 数据分析人才培养：当前业务的竞争更是技术人才的竞争，大模型时代的到来，重塑了 AI 人才的培养。TableAgent 可提高 AI 技术人才能力，赋予人人都是数据分析师的能力。

二、商业模式

目前，TableAgent 面向社会免费开放，公众均可注册申请试用体验。

三、应用推广前景

TableAgent 使用 0 门槛，开箱即用，用户仅需要把企业属性数据上传到应用后台即可对业务数据开展专业性的分析。基于九章云极自有的 Alaya 元识大模型和底层体系，可以适用于各类行业的数据分析，实现对任何特定领域内个性化数据分析情景下的微调，对行业没有限制。目前已经在金融、制造、交通、互联网、地产、能源等多个行业进行应用验证，通过近期的公测表现，我们相信 TableAgent 在未来会有更广泛的应用和更具商业价值的产出。

九章云极知识管家
打造企业专属大模型智能底座

北京九章云极科技有限公司

北京九章云极科技有限公司（简称：九章云极 DataCanvas）以“创造智能，探索未知”为使命，以“助力全球企业智能升级”为愿景，是中国人工智能基础软件领军者。公司致力通过自主研发的人工智能基础软件产品系列和解决方案为用户提供人工智能基础服务，助力用户在数智化转型中轻松完成模型和数据的双向赋能，低成本高效率的提升企业决策能力，实现企业级 AI 规模化应用。

九章云极 DataCanvas 的核心产品系列 AIFS 人工智能基础软件和 DataPilot 数据领航员具有高度的灵活性和可扩展性，能够处理各种类型和规模的数据，简化了数据处理和分析的复杂性。产品集成了一系列先进人工智能技术，包括多模态向量数据库、因果学习、思维件等，为企业提供 AI 软件开发新范式。

概　述

在大模型技术浪潮的推动下，企业知识的处理和应用正在发生全新变化。企业知识管理面临着知识碎片化，信息过载，数据及信息安全难，知识共享交流难，知识与业务融合难等挑战。九章云极 DataCanvas 以 AIFS（AI Foundation Software）为根基，发挥 Alaya 九章元识大模型和多模向量数据库的核心能力，打造企业级知识管家，通过数据收集，数据处理，写入向量数据库，集成、微调大语言模型，知识助手应用，以及反馈与迭代优化六步过程，为企业构建高度自动化与智能化的企业知识库。

在六步过程中，企业知识管家支持全规模、全类型的企业知识数据收集，并通过数据处理将企业知识转化为高维向量，储存到 DingoDB 多模向量数据库中。根据企业需求，企业知识管家微调 DataCanvas Alaya 九章元识大模型并与向量化企业知识库进行交互，通过知识助手将构建好的知识库应用于企业多元业务场景。同时为用户提供便捷的反馈渠道，不断对企业知识库进行迭代优化，保证其准确性和时效性。

需求分析

1、知识碎片化

随着信息的爆炸式增长，知识变得碎片化和分散。企业需要一个知识管家系统来收集、整理和连接这些碎片化的知识。

2、信息过载

随着企业业务的快速发展和创新，大量信息和数据不断涌现，缺乏有效的信息筛选处理机制，导致大量信息被堆积和遗忘，无法得到及时有效的利用。

3、数据及信息安全难

随着企业知识信息量的不断增加，信息安全风险也不断增加，企业的核心知识和敏感信息在知识管理过程中可能泄露。

4、知识共享交流难

知识共享机制不足、知识交流渠道不畅、知识共享和交流意愿不足、语言和沟通障碍。

5、知识与业务融合难

知识管理系统和业务系统各自独立，知识和业务之间缺乏紧密的关联和互动，企业的知识与业务难融合，会导致业务知识的滞后。

案例介绍

一、主要能力

九章云极知识管家包括结合大模型并融合企业专有知识的 QA 问答功能、可自定义角色的定制化对话助手、针对上传文档的智能分析 ChatDoc 以及后台相应的模型 & 微调管理、知识数据管理、智能应用 Agent 管理等功能。九章云极知识管家作为企业的专属大模型智能底座可面向不同场景定义相应职位、角色的大模型特色应用，比如智能合同审核、营销文案创作等，帮助企业逐步打造自己的大模型应用体系。

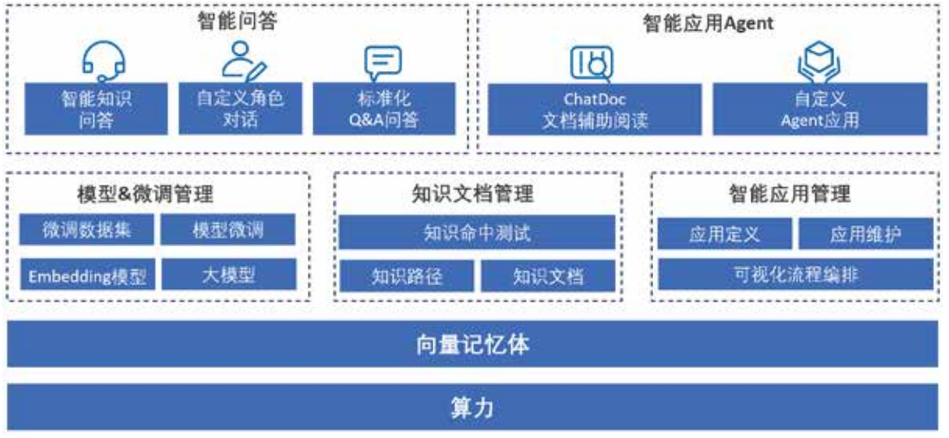


图 1　产品架构图

二、技术创新点

1、自研的多模态大模型底座支撑——DataCanvas Alaya 九章元识大模型

DataCanvas Alaya 是九章云极 DataCanvas 自研的“通识 + 产业”白盒大模型矩阵，支持多种模态模式，高效微调训练，以及 Flash attention 技术。九章元识提供了一系列不同配置和参数的，具备业界前沿能力和技术的预训练大模型，可联合企业训练面向金融、通信、制造等行业的领域垂类多模态大模型，更好地应对行业复杂专业的问题。秉持开放友好的开源理念，九章元识大模型矩阵中的 Alaya-7B 已在 GitHub 进行开源，开源地址为：https://github.com/DataCanvasIO/Alaya。

2、大模型时代的数据引擎——自研多模向量数据库 DingoDB

DingoDB 是九章云极自研的多模向量数据库，同时提供结构化与非结构化数据的存储、分析和科学计算的能力。可支撑政府、金融、传统行业构建企业级的知识库，实现语义的精准搜索与联合分析。作为行业首批，DingoDB 以同批次最好成绩完成中国信通院向量数据库技术标准的测试。DingoDB 多模向量数据库也在 GitHub 进行开源，开源地址：https://github.com/DingoDB。

3、灵活丰富的大模型智能体 Agent 扩展能力，支持 Agent 扩展和可视化 Agent 编排。

4、易用的知识管理及模型微调功能，可实现企业知识多模式智能对话问答。

5、是行业垂类多模态大模型基座。基于九章云极 Alaya 元识大模型，可联合企业训练面向金融、通信、制造等行业的领域垂类多模态大模型，更好地应对行业复杂专业的问题。

6、支持混合多模态检索匹配，支持多副本存储策略和持续可用的存储方案，减少数据丢失的风险。具备良好的可扩展性和海量存储能力。是高性能知识向量存储记忆体。

7、面向企业提供定制化的软硬一体解决方案，全面覆盖底层算力到上层应用的全链路，支持一体化部署、产品开箱即用。

三、实施效果

随着大模型技术的日渐成熟及生成式 AI 应用的热度的持续走高，如何借力新技术、加速数智化转型，构建差异化竞争力，是企业当下必须深入思考的命题。企业想要安全可靠的应用大模型技术，那么通过知识管家融合企业内部的知识体系便是企业必须迈出的第一步。九章云极知识管家以此为目标，基于九章云极在人工智能领域多行业的长期深耕实践，打造了包括底层算力框架、垂类微调大模型、存储记忆体到智能 QA 问答应用等全链路一体化的大模型应用解决方案。通过九章云极构建的企业大模型智能底座可全面支撑企业“无限创想、触手可得”的大模型应用愿景。

四、应用落地情况

作为大模型时代的数据处理新范式，基于九章云极元识大模型和向量数据库 DingoDB 前沿技术能力打造的九章云极知识管家，在金融、制造、通信、能源等众多行业拥有丰富的应用场景，并已经在某头部汽车制造厂商和城商行进行落地化应用，助力用户企业构建高度自动化与智能化的企业知识库。更多丰富行业应用场景，包括：

- **金融行业应用场景：**金融知识智能问答、智能反欺诈、智能客户聊天机器人、NL2SQL、代码生成等、智能 BI 及分析决策、文案创作、文档生成等；
- **制造行业应用场景：**制造工艺问答、售后服务知识问答、文档辅助编写、智能谈判、合同审核等；
- **交通行业应用场景:** 高速知识问答、司乘人员知识问答、航空专业知识问答、机电系统维护、道路病害养护、文档辅助编写等。

效益分析

一、社会效益

1. 私有化部署，实现企业数据隐私保护

- 采取数据脱敏、匿名化等技术手段，确保在训练过程中不泄露个人身份和敏感信息。
- 使用加密算法或差分隐私技术来保护数据的隐私。
- 采用安全多方计算等技术，使得多个参与方能够在不泄露数据的情况下进行计算和模型训练。

2. 访问控制和权限管理

- 建立严格的访问控制机制，限制对大模型数据的访问权限，确保只有经过授权的人员可以访问和操作数据。

二、商业模型

可提供软硬件一体化部署模式，支持永久使用许可和订阅许可两种服务方式。

三、应用推广前景

目前，大模型产品应用市场正在迅速增长。根据某研究数据预测，预计 2023 年，全球人工智能大模型市场规模将达到 210 亿美元，并在 2028 年使大模型市场规模达到 1095 亿美元。此外，随着人工智能技术的不断发展，大模型产品的应用范围和功能也在不断拓展，九章云极知识管家产品可快速完成企业私有化部署、开箱即用，通过提供软硬结合的一体化解决方案，将在大模型时代充分挖掘数据价值，助力用户在数智化浪潮中轻松完成模型和数据的双向赋能，为用户带来灵活高效的数据驱动决策和更加优质的业务发展，打通企业应用大模型的最后一公里。

“Pixeling 千象”

上海智象未来计算机科技有限公司

HiDream.ai（智象未来），是一家专注于构建视觉多模态基础模型及应用的生成式人工智能初创公司，由加拿大工程院外籍院士、原京东集团副总裁梅涛博士创立。致力于围绕视觉打造生成式多模态基础模型及应用，激发从业者创造力，提升创作生产力，打造交互式智能内容创作新范式。

核心业务是基于自研的生成式视觉多模态基础模型，实现文本、图片、视频、3D 模型等多模态内容的生成；打造了面向所有设计师的通用创作工具及泛设计内容社区“Pixeling（千象）”支持创意生成、艺术创作、在线编辑等全过程的可视化，帮助用户实现交互的智能化、作品的个性化，让用户的创意得以最大化的释放；同时面向电商商家推出 AI 制图工具 PixMaker，目前支持固定商品 SKU 生成场景图和人像模特图生成。

概　述

“Pixeling 千象”是一款全中文、易上手的 AIGC 创作平台和社区，专为设计师的需求而打造。平台包含图片生成、视频生成、图片编辑（智能重绘、智能拓图）等功能，旨在帮助用户零基础轻松掌握 AIGC 一站式能力，唤醒创造力，解放生产力，全面提升设计全流程工作效率。

“Pixeling 千象”依托智象未来自研的视觉多模态生成式基础模型，实现文本、图片、视频等多模态内容生成。模型参数超过百亿，技术水平行业领先，为用户创作提供强大支持（www.hidreamai.com）。

需求分析

在数字化时代，AIGC 技术迅猛发展，设计师对易用、高效的中文 AIGC 创作平台需求日益凸显。市场急需一款全中文界面、一站式服务、易用且具备互动社区的 AIGC 创作平台，以满足设计师的多样化需求。

针对这一背景，“Pixeling 千象”应运而生。它集成了图片生成、视频生成、图片编辑等功能，助力用户零基础掌握 AIGC 一站式能力。平台基于自研的 AIGC 视觉多模态基础模型，实现文本、图片、视频等多模态内容生成。简洁易用的界面让设计师能更快地完成从构思到成品的全过程，提高工作效率。

“Pixeling 千象”还为设计师打造了活跃的互动社区，便于分享经验、获取灵感、拓展人脉。用户在此可以充分发挥创意，共同成长。此外，平台还可满足用户的个性化需求，让设计师在创作过程中实现更多可能性。

案例介绍

“Pixeling 千象”目前包含图片生成、视频生成、图片编辑、3D 生成等功能，是一个面向设计师的通用设计工具，同时也是一个服务于 AIGC 创作的泛设计内容社区。

图片生成支持

文字生成图片、参考图生成图片，支持用户基于在平台生成的图片持续进行生成创作。

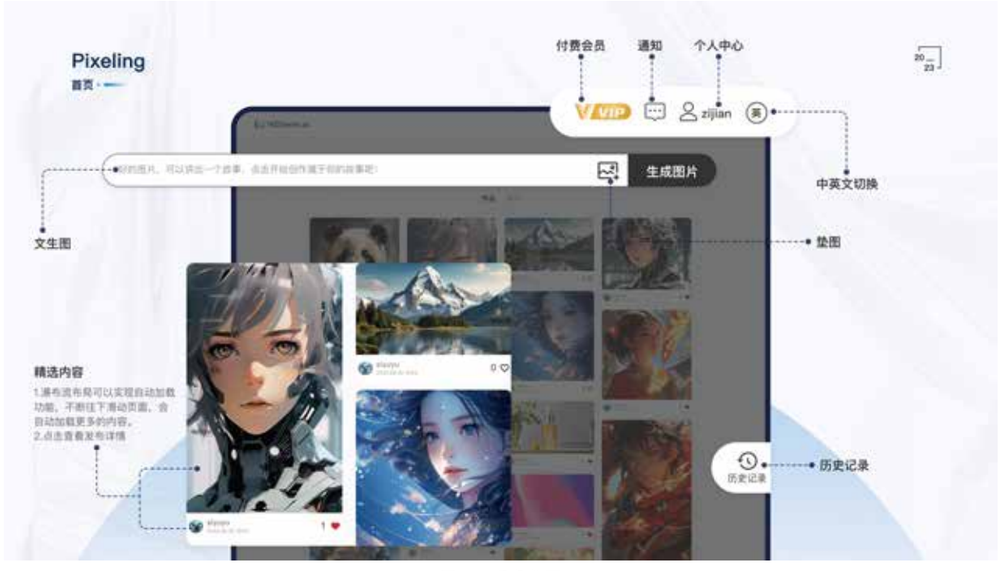


图 1　千象首页

视频生成支持

文字生成视频、图片生成视频，用户可以从本地上传图片、或者基于在平台生成的图片历史记录生成视频；此外还支持智能运镜，使画面更加生动。

图 2　智能运镜

图片编辑支持

• **智能拓图：**“Pixeling 千象”将会自动为用户进行画面拓展，并保持细节的清晰和准确性。通过智能拓图，用户可以将一幅小尺寸的绘画作品扩展至更大的画布尺寸，而无需担心失真或模糊。通过使用智能拓图，用户能够丰富画面背景、优化图片格局、增加画面层次，“Pixeling 千象”在创作思路上为用户提供了更开阔、更浩瀚的想象空间，让创作更加自由畅快。

图 3　智能拓图

图 4　智能拓图

图 5　智能拓图

• **智能重绘：**智能重绘允许用户对生成的图片作品的特定区域进行修改和改进。用户可以通过调整颜色、线条和细节等，对选中的部分进行精细调整。通过智能重绘功能，用户可以轻松实现对细节的精益求精，让作品更加完美。

图 6　智能重绘

图 7　智能重绘

图 8　智能重绘

效益分析

Pixeling 千象”作为一款全中文 AIGC 创作平 5 例线台，凭借卓越的技术实力和实用的功能体验，满足了设计师在多样化、个性化方面的需求，为国内设计产业业的发展注入新活力。

根据数据，“Pixeling 千象”能够在质量、效率、资产等多方面为用户和客户持续提供价值，上手难度降低 99%，创意维度增加 75%，节省 98% 的出图时间，100% 增加内容沉淀。

经济社会效益方面

平台提供的图片生成、视频生成、图片编辑等功能，能有效帮助设计师提高工作效率，降低创作成本。依托智象未来自研的视觉多模态生成式基础模型，平台为用户创作提供强大支持，使设计师能够快速完成从构思到成品的全过程。

商业模式方面

“Pixeling 千象”通过提供一站式 AIGC 服务，吸引设计师入驻，形成稳定的用户群体。平台可以进一步挖掘用户需求，推出更多针对性功能和服务，提高用户粘性，实现持续盈利。

应用推广方面

“Pixeling 千象”充分利用平台上的丰富资源和活跃社区，助力设计师拓展人脉、获取灵感。同时，借助行业领先的技术水平和对用户需求的精准把握，平台在设计师群体中形成良好口碑，实现自发推广。

通过不断创新和优化服务，“Pixeling 千象”将在设计领域发挥更大作用，推动行业繁荣发展。

书生筑梦视频生成大模型

上海人工智能实验室

上海人工智能实验室是我国人工智能领域的新型科研机构，开展战略性、原创性、前瞻性的科学研究与技术攻关，突破人工智能的重要基础理论和关键核心技术，打造“突破型、引领型、平台型”一体化的大型综合性研究基地，支撑我国人工智能产业实现跨越式发展，目标建成国际一流的人工智能实验室，成为享誉全球的人工智能原创理论和技术的策源地。

概　述

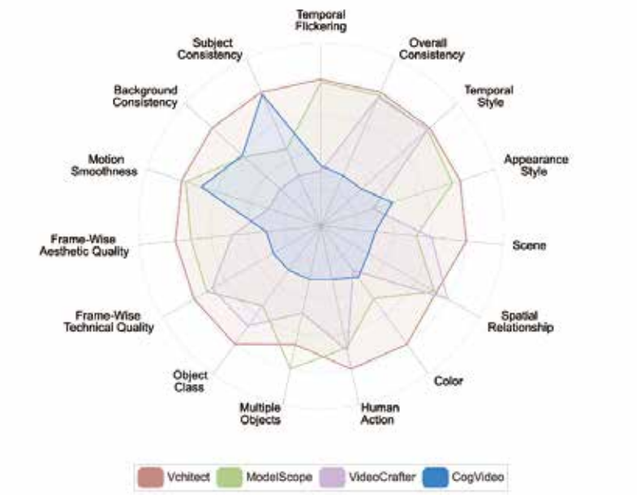
书生筑梦视频生成大模型，通过设计大规模视频生成模型的基础模型结构，机器学习方法，建立大规模数据集，构建数据处理工具，实现了文生视频大模型系统，并在通用场景下实现了 2K 分辨率、支持转场与镜头语言的分钟级长视频生成。

需求分析

随着生成式人工智能技术的发展，图像生成模型正在日渐成熟，以 Midjourney 和 Stable Diffusion 为代表的文生图模型为使用者提供了全新的创作模式。视频生成，由于其更加广泛的应用场景，以及更加生动的表现方式，受到了越来越多的关注和需求。因此，设计能够生成高画质，长视频的大规模视频生成模型，对于广告设计、电影制作、以及艺术创作将产生革命式的变革。

案例介绍

作为首个支持故事性、多镜头的视频生成大模型，包含超过 30 亿参数的书生·筑梦将全面赋能视频创作，拓展创意空间。书生·筑梦将 AI 生成视频时长由秒级提升至分钟级，并使所生成视频内容具备“转场流畅、故事连贯、画质高清”特质。凭借强大的语义、图像理解和生成能力，在多维度评测指标中综合领先。



多样化生成方式，够美够方便

作为一款融合文本、图像、视频等多模态数据的视频生成大模型，书生·筑梦支持由“文生视频”与“图生视频”多样化任务。为实现通用文本视频生成 (Text-to-Video Generation,T2V), 团队在与训练文生图大模型基础上引入时空建模模块，并使用图像视频联合训练的方式，使模型具备了 T2V 能力。

在模型中输入通用文本，书生·筑梦生成了以下视频。

输入文本（提示词）：海底，鱼群，珊瑚礁

与此同时，团队基于掩码的条件视频扩散模型，将特定图片作为即将生成视频的第一帧和对应掩码，实现了由图片驱动的视频生成能力（Image-to-Video Generation, I2V）。

输入静态图片，书生·筑梦可让其生动真实地“流动”起来

创新性生成阶段，够清够流畅

连贯的转场镜头、生动波折的故事、充满美感的高清画质缺一不可在影视巨作中缺一不可。在书生·筑梦中，运用 AI 直接生成长视频“巨作”将成为可能。研发完成转场视频生成模型，为其输入多段给定视频或场景图片，书生·筑梦可根据提示词（prompt）和扩散模型（Diffusion Model）自动生成转场视频，从而实现不同场景和视频之间的“丝滑”连接。在生成多段视频时，为保证视频中的主体一致性，研发团队提出了保持主体一致性模块。该模块的输入内容一张包含主体的图片和一段文字描述，图片信息作为文本的一部分，或作为额外的信息加入到网络注意力模块中，即可保证多段视频中的主体一致。实现生成视频的“多机位”效果，使长视频中的故事一致性成为可能。

在多个镜头的长视频中，“花朵”主体保持了一致性

本项目相关研究成果已应用于央视听媒体大模型（CMG Media GPT）中。该大模型为首个专注于视听媒体内容生产的 AI 大模型，由上海 AI 实验室与中央广播电视总台联合推出。于此同时，书生筑梦视频生成大模型已与商汤科技、想法流、北京电影学院等单位达成合作意向，正成为推动视听媒体编创方式变革的 AI 工具。

效益分析

书生筑梦视频生成大模型具备生成多样性和创造性视频内容的非凡能力，为创意和创新开辟了崭新的可能性。该模型有助于广告公司、娱乐制作公司等行业生成独特的广告创意、电影特效、虚拟角色等，为观众带来独具魅力的视觉体验。传统的影视制作通常需要投入大量人力、物力和时间。然而，引入该模型作为辅助工具，可以通过自动化和智能化的方式，降低人力成本并加速视频生成速度。该模型能够协助企业和影视从业者更快地生成所需的视频内容，从而节约时间和成本。

书生浦语开源大模型

上海人工智能实验室

上海人工智能实验室是我国人工智能领域的新型科研机构，开展战略性、原创性、前瞻性的科学研究与技术攻关，突破人工智能的重要基础理论和关键核心技术，打造“突破型、引领型、平台型”一体化的大型综合性研究基地，支撑我国人工智能产业实现跨越式发展，目标建成国际一流的人工智能实验室，成为享誉全球的人工智能原创理论和技术的策源地。

概　述

书生浦语开源大模型涵盖 70 亿参数的轻量级版本 InternLM-7B，以及 200 亿参数的中量级版本和 InternLM-20B，以及完整的开源工具链体系。

InternLM-7B 在包含 40 个评测集的全维度评测中展现出卓越且平衡的性能，它在两个被广泛采用的基准 MMLU 和 CEval 上分别取得了 50.8 和 52.8 的高分，开源一度刷新了 7B 量级模型的纪录。

InternLM-20B 是基于 2.3T token 预训练语料从头训练的中量级语言大模型。相较于 InternLM-7B，训练语料经过了更高水平的多层次清洗，补充了高知识密度和用于强化理解及推理能力的训练数据。因此，在考验语言模型技术水平的理解能力、推理能力、数学能力、编程能力等方面，InternLM-20B 都有显著提升，以不足三分之一的参数量，达到 Llama2-70B 水平。

书生浦语开源且可免费商用，基于书生浦语开源代码、模型、开源工具链体系，商业场景可定制高精度行业模型。

需求分析

浪潮之上，大模型的应用价值日趋受到关注。正如历史上的任何一项新技术，其生命力终究要回归到是否可以广泛落地，为世界带来积极且真实的变化。

相比于国内社区之前陆续开源的 7B 和 13B 规格的模型，20B 量级模型具备更为强大的综合能力，在复杂推理和反思能力上尤为突出，因此可为实际应用带来更有力的性能支持；同时，20B 量级模型可在单卡上进行推理，经过低比特量化后，可运行在单块消费级 GPU 上，因而在实际应用中更为便捷。

在此背景下，上海人工智能实验室联合多家机构推出了中量级参数的 InternLM-20B 大模型，性能先进且应用便捷，以不足三分之一的参数量，达到了当前被视为开源模型标杆的 Llama2-70B 的能力水平。

案例介绍

相比于此前的开源模型，InternLM-20B 的能力优势主要体现在：

- 优异的综合性能
- 强大的工具调用能力
- 更长的语境
- 更安全的价值对齐
- 全线升级的开源工具、数据体系

架构增强：深结构、长语境

相对有限的参数规模下，研究人员在架构设计时面临重要的取舍——提高模型的深度还是宽度？通过广泛的对照实验，书生·浦语团队发现，更深的模型层数更有利于复杂推理能力的培养。因此在架构设计时，研究人员把模型层数设定为 60 层，超过 7B 与 13B 模型通常采用的 32 层或者 40 层设计；同时内部维度保持在 5120，处于适中水平。通过架构设计上的新取舍，InternLM-20B 在较高计算效率的条件下实现了复杂推理能力的显著提升。

综合性能增强：多个评测中领先

基于 OpenCompass 大模型评测平台，研究人员在涵盖语言、知识、理解、推理和学科能力等五大维度的 50 个主流评测集上，对 InternLM-20B 及相近量级的开源模型进行了全面测试比较。评测结果显示，InternLM-20B 在全维度上领先于开源 13B 量级模型，平均成绩不仅明显超越 Llama-33B，甚至优于被称为开源模型的标杆 Llama2-70B。

调用工具能力增强：不会也能学

工具调用是拓展大语言模型能力边界的重要手段，也是 OpenAI 近期推出大模型的重点特性之一。InternLM-20B 对话模型支持了日期、天气、旅行、体育等数十个方向的内容输出及上万个不同的 API。在清华大学等机构联合发布的大模型工具调用评测集 ToolBench 中，InternLM-20B 和 ChatGPT 相比，达到了 63.5% 的胜率，在该榜单上取得了最优结果，表现出强大的工具调用能力。

价值观增强：更安全的开源模型

更贴合人类价值观的大语言模型，才有可能更好地充当“人类助手”的角色。InternLM-20B 在迭代过程中加入了大量符合人类价值观的数据，研究团队组织相关领域专家对模型进行了多轮红队攻击，大幅提升其安全性。当用户向 InternLM-20B 提出带有偏见的问题时，它能够识别出不安全因素，并在回答中给出正确的价值引导。

对话能力增强：语境长度达到 16K

InternLM-20B 在训练阶段的语境长度分阶段拓展到了 8K，同时通过 Dynamic NTK 等手段将推理时的语境长度拓展到了 16K。基于 16K 的语境长度，InternLM-20B 可以有效支持长文理解、长文生成和超长对话。

效益分析

面向大模型掀起的新一轮创新浪潮，上海 AI 实验室致力于以原始创新引领技术进步，持续打造综合能力更强大的基础模型，构建更完整易用的全链条工具体系，并坚持通过开源开放、免费商用，全面赋能整个 AI 社区生态的繁荣发展，帮助企业和研究机构降低大模型的开发和应用门槛，让大模型的价值在各行各业中绽放。

百川大模型在娱乐领域的应用

上海百川智能技术有限公司

上海百川智能技术有限公司成立于 2023 年 9 月 12 日，由前搜狗公司 CEO 王小川创立。目前已完成 A1 轮战略融资，总融资金额达 3.5 亿美元，创下国内大模型初创企业跻身科技独角兽行列最快记录。成立以来，百川智能接连发布 Baichuan-7B/13B、Baichuan2-7B/13B、Baichuan2-192K 五款开源大模型及 Baichuan-53B、Baichuan2-53B 两款闭源大模型。其中 Baichuan-7B/13B 两款大模型在多个权威评测榜单均名列前茅，累积下载量突破六百万次。Baichuan2-7B/13B 更是在各维度全面领先 Llama 2，引领了中国开源生态发展。Baichuan2-192K 大模型上下文窗口长度高达 192K，一次能够处理约 35 万个汉字，是目前全球最长上下文窗口大模型。11 月 16 日，百川智能与鹏城实验室携手探索大模型训练和应用，合作研发基于国产算力的 128K 长窗口大模型“鹏城 - 百川·脑海 33B”。

概　述

在泛娱乐领域大量 C 端用户拥有情感陪伴和内容创作、消费需求，社区、游戏等类别 B 端用户急需大模型赋能以对产品进行创新升级的背景下，百川智能以 LLM 驱动 Character 为基础，推出具有丰富多样剧情内容、能够生成真实自然对话的泛娱乐领域创新产品线。针对 B 端用户，上线百川智能独有的知识库功能，允许创作者针对于角色创建专属知识库，让角色回复更具备可控性和事实性。同时，基于长窗口模型，产品支持创作者基于一本小说快速创建角色，且具有超长的记忆能力。针对 C 端客户，平台将从人机共生的 UGC 游戏化社区（word world 阶段）向开放世界元宇宙（virtual world 阶段）发展，从生成角色开始，生成文字游戏，再通过多模态生成图、文、视频和游戏，最终实现生成式开放世界。

需求分析

本项目专注于泛娱乐领域创新产品线，针对各种客户类型（包括社区类、游戏类、影视类、网文类和营销类等）及其特定需求（例如角色扮演、智能 NPC 等）进行深入分析，并提出相应的切入点与策略。借助百川大模型，平台将角色扮演能力应用于多种内容形式，如互动小说、游戏和影视剧等，以满足用户对高品质、普及性的娱乐内容的需求。本项目致力于降低创作门槛，提高创造力上限，吸引 14 至 40 岁的广泛用户群体，并提供易于使用的创作界面和强大的创作工具，以确保内容的质量与丰富性。总体来说，本项目旨在满足不同客户类型的需求，并助力内容创作与改编的多元化发展。

案例介绍

一、主要能力

- **产品设计能力：**项目团队具备丰富的产品设计经验，能够根据客户需求进行定制化设计，提供开箱即用的角色管理和调优工具。
- **技术支持能力：**项目团队拥有专业的技术支持团队，能够解决客户在使用过程中遇到的各种问题，提供优质的售后服务。
- **用户体验优化能力：**在产品设计中融入了娱乐感和游戏感，以提供更为生动和吸引人的用户体验。

二、技术创新点

- **知识库：**平台允许创作者针对角色创建专属知识库，这使得角色的回复更加具有可控性和事实性。这目前是百川智能独有的设计，使创作者能够更好地控制角色的行为和反应，同时也保证了信息的准确性。
- **长窗口上下文理解：**平台具备超长上下文理解能力，使得角色具有超长的记忆能力。这不仅可以提高创作效率，让创作者可以快速创建角色，也可以提升对话体验，让角色能够理解和回应用户的长篇对话。

三、实施效果

平台在用户反馈方面得到了广泛的好评。用户表示平台系统稳定，能够快速响应并提供高质量的输出。在进行角色扮演、虚拟陪伴、聊天对话等方面，都表现得非常出色。此外，平台在复述用户问题、处理 Prompt 限定问题和上下文信息方面也做得很好，能够严格按照用户的指示进行操作，并根据上下文信息做出合理的回答。

四、应用落地情况

产品目前处于内测优化阶段，预计 12 月下旬开始陆续上线。

效益分析

本项目将通过独特的商业模式实现企业盈利与消费者需求的完美融合，推动产业发展，创造更多就业机会，提高经济效益。项目关注消费者个性化需求，提升消费体验和生活品质，助力国家文化软实力提升，营造健康文化氛围，促进社会和谐与稳定。B 端商业模式包括免费试用、使用频率限制、API 使用余额展示、异常问题处理机制及商业留资入口，旨帮助客户提升产品商业价值。C 端商业模式基于 AIGC 能力打造的创玩一体的模式和体验，降低了门槛，提升创造力上限，并提供持续性服务确保用户得到及时帮助和支持。2022 年中国游戏市场实际销售收入 2658.84 亿元，游戏用户规模 6.64 亿。其中角色扮演类游戏表现突出，占总收入近五分之一，这表明该项目将在泛娱乐领域取得重要地位。

AnimateDiff：一项基于个性化文生图模型扩展后的视频生成框架

上海人工智能实验室

上海人工智能实验室是我国人工智能领域的新型科研机构，开展战略性、原创性、前瞻性的科学研究与技术攻关，突破人工智能的重要基础理论和关键核心技术，打造“突破型、引领型、平台型”一体化的大型综合性研究基地，支撑我国人工智能产业实现跨越式发展，目标建成国际一流的人工智能实验室，成为享誉全球的人工智能原创理论和技术的策源地。

概　述

伴随科学技术进一步发展，知识与内容的生成方式从原本的由人本身作为生产驱动转变为由人利用工具或技术进行生产，与人工智能技术迅猛发展相对应的便是面向人工智能的技术生成即“AIGC”。AIGC 凭借更具性价比的使用成本、相对较低的使用门槛以及更具有生产力的生产效率，成为文本、语音、图像乃至视频生成中脱颖而出的工具。但与此同时，AIGC 大模型海量的训练数据需求使得其训练难度与训练成本均较高，因而抬高了 AIGC 大模型在实际应用中的成本与要求，为 AIGC 大模型训练与使用增加了难度。

上海人工智能实验室通过锁定原有文生图模型进而插入新的动作建模模块，形成一个适配个性化文生图模型的从文字到视频生成的垂类大模型。该模型通过将运动建模模块引入到被锁定住的文生图模型中去，在视频的基础上进行训练，从而使得模型学会合理化的运动知识，由此实现从文字到高质量、稳定性的视频生产，使得用户定制自己想要的个性化动态视频风格与内容。

需求分析

高速发展的信息技术与急速流动的社会使得人们身处于一个高度信息化与数字化的时代，人们对于日常生活中信息内容的时效性、信息形式的生动化以及信息生产的自主化有了更高的追求。与之相悖，传统的内容生成模式在使用中的高门槛、低生产速度与单一生成形式使得其难以满足现今的社会需求，人们在现今多样化的社会中对于技术如何驱动内容生成走向更便捷有了新的要求。然而技术本身具有一定的研发成本与技术门槛，因此对于 AIGC 大模型而言，如何多快好省地实现轻松化的应用，是其在当前研发与应用中的关键问题。

上海人工智能实验室研发的 AnimateDiff 正是基于训练方式的革新带来了节约训练成本情况下如何在少量训练数据下，实现从用户输入文字到模型驱动生成视频来破解这一难题，为 AIGC 在视频内容生成端带来全新突破。

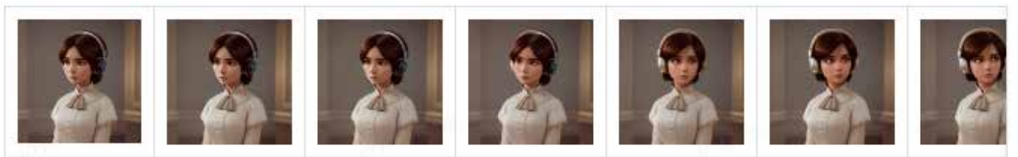
案例介绍

1. 主要能力

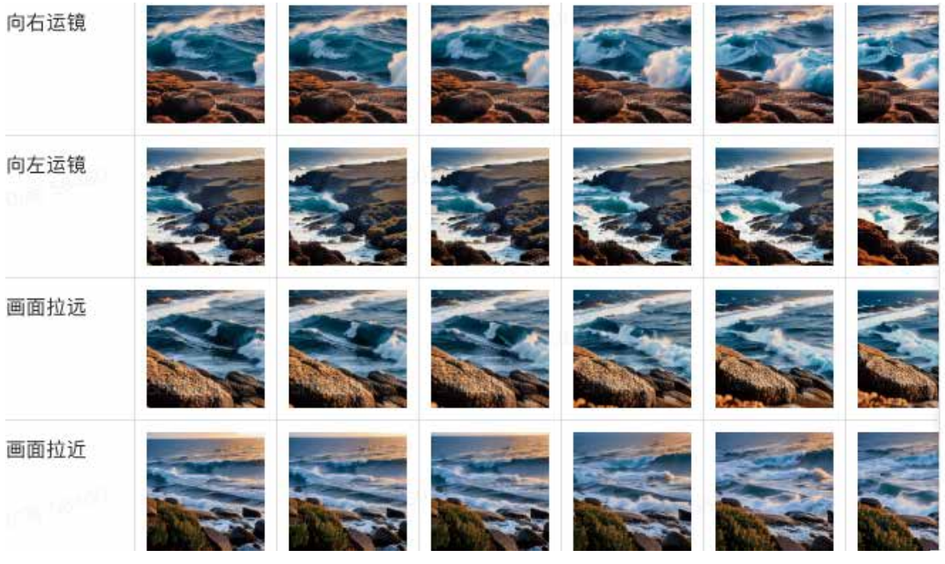
AnimateDiff 通过对已有的文生图模型扩展，以插入新的动作建模模块的方式来构建起其模型框架，从而使得该模型仅需通过训练运动建模模块来学习合理化的运动规律，实现根据用户输入的文字最终能够变成有序且连续的动态画面，进而通过复制这一运动模块将其应用到同类基于类似模型训练而成的个性化文生图模型中，实现了对于模型本身的延续与扩展，从而支持生成高质量且多样化风格的视频。

- **连续且稳定生成动态画面：**AnimateDiff 在生成视频质量上，由于先验基础锁定了文生图模型，使得其保证了图片生成质量，进而通过引入运动模块，保证了其生成画面的连续性与平滑度，使视频片段具有更好的连贯视觉效果，在视觉观感上让用户感受到流畅的视频内容。

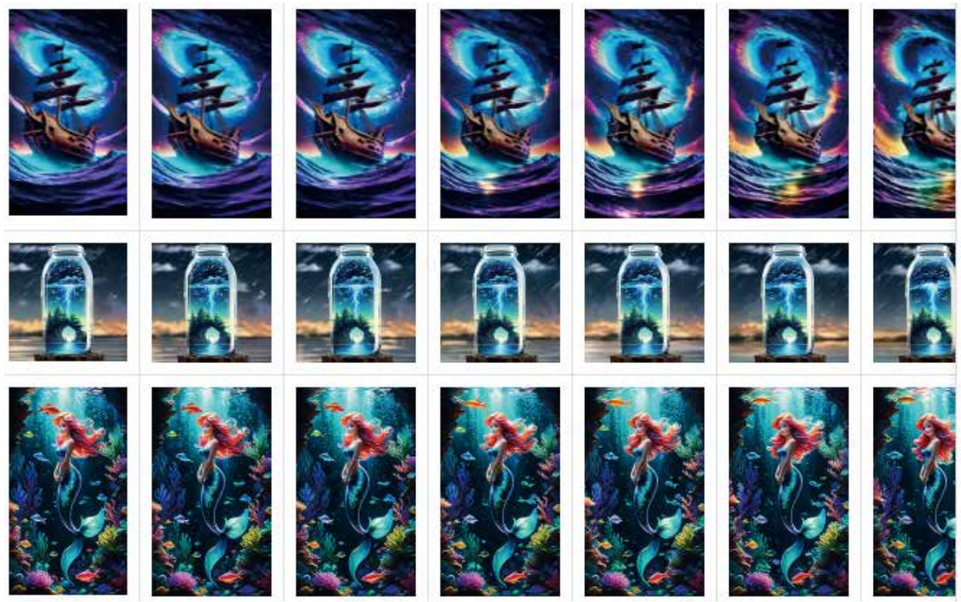
- **适配多种个性化模型：**AnimateDiff 作为一项 AIGC 垂类大模型，由于其在训练中将运动建模模块作为独立插件引入，以原有的文生图模型为基础，从而使得其对于基于同样训练模式生成的同类个性化文生图模型有很好的适配性，从而可以支持生成风格多样的视频动画。



- **动态化镜头移动控制效果：**AnimateDiff 的运动模块作为一项插件，在研究人员的努力之下，其运动模块实现了多元的动态效果，可以支持所生成的画面实现类似于相机镜头前后左右移动以及远近拉伸的视觉效果，为未来生成更复杂多样角度的视频内容奠定基础。



- **支持多种比例下画面剪裁：**AnimateDiff 在生成动画的效果上，不仅从原有的 16 帧提升到现有的 32 帧效果，同时也支持基于 1024×1024×16 比例范围内自定义的裁剪，用户可以根据自己所希望生成视频内容的特点，自定义生成内容的尺寸，为 AnimateDiff 所生成的视频带来更广的应用场景。



2. 技术创新

AnimateDiff 作为从文字到动态画面的视频生成框架，通过其训练方法，优化了文字 - 图片 - 视频的 AIGC 大模型生成路径，有效节约了训练成本，实现了从文字到视频的生成框架，降低了用户在视频创作中的使用门槛。并且在训练中，通过对训练运动模块的优化，使得其所需数据量更小的情况下，生成了成像质量更为稳定且画面连续性更出色的动态视频。因此 AnimateDiff 在即是对 AIGC 大模型训练路径的创新，也是对于 AIGC 大模型在生成模式上的创新。

3. 社区影响

AnimateDiff 目前主要是作为 Diffusion Web UI 和 ComfyUI 中的插件供用户进行直接使用，也支持用户在 CivitAI、HuggingFace 以及 OpenXLab 几大开源社区内体验其预训练模型，并且于 2023 年 11 月份在 SDXL 上开源其测试版。自 AnimateDiff 在 GitHub 上发布以来，就备受各个 AI 开源社区的关注，备受众多 AIGC 相关行业用户的关注与使用，推动众多用户不仅通过 AnimateDiff 制作出许多优秀、生动且有趣的视频作品，亦有众多用户在 AnimateDiff 的基础上形成了新的扩展，进一步拓宽 AnimateDiff 的应用场景与影响力，为 AIGC 助力内容应用带来了更丰富的可能。

效益分析

AnimateDiff 作为支持从文本到连续稳定视频生成模型，在训练层面很好地链接了原有丰富且具有个性化的文生图模型，通过其独特的算法训练路径实现了节约训练成本与优化训练资源，并且进一步增添了同类文生图模型的可持续性扩展。另一方面在应用层面，其降低了普罗大众在 AIGC 方面的使用门槛，使得普通人捕捉自身想象力进行艺术创造变得更轻松。因此伴随 AnimateDiff 在未来进一步发展，我们可以期待其应用于艺术创造、文博数字化等丰富场景之中，进而推动 AI 技术、人与社会更好地互动，共同创造美与价值。

通义千问 2.0 在企业场景的应用

阿里云计算有限公司

阿里云创立于 2009 年，是全球领先的云计算及人工智能科技公司，为 200 多个国家和地区的企业、开发者和政府机构提供服务。阿里云致力于以在线公共服务的方式，提供安全、可靠的计算和数据处理能力，让计算和人工智能成为普惠科技。2017 年 1 月阿里云成为奥运会全球指定云服务商。

概　述

阿里巴巴通义千问大模型是一个超大规模的语言模型，具备多轮对话、文案创作、逻辑推理、多模态理解、多语言支持等功能。2023 年 11 月最新发布的通义千问 2.0 在性能上取得巨大飞跃，相比 4 月发布的 1.0 版本，通义千问 2.0 在复杂指令理解、文学创作、通用数学、知识记忆、幻觉抵御等能力上均有显著提升。近期，通义千问开源 720 亿参数模型 Qwen-72B。Qwen-72B 在 10 个权威基准测评创下开源模型最优成绩，性能超越开源标杆 Llama 2-70B 和大部分商用闭源模型，可适配企业级、科研级的高性能应用。通义千问还开源了 18 亿参数模型 Qwen-1.8B 和音频大模型 Qwen-Audio，在业界率先实现“全尺寸、全模态”开源。通义千问大模型在企业场景中的应用非常广泛，包括应用于数据管理、模型开发、模型调试、模型任务编排、插件管理、API 管理、企业专属 Planing 和企业业务流程编排等场景中。

需求分析

通义千问 2.0 在指令遵循、工具使用、精细化创作等方面作了技术优化，能够更好地被下游应用场景集成。通义大模型官网上线了多模态和插件功能，支持图片输入、文档解析等细分任务。企业专属大模型是基于阿里巴巴通义大模型所构建的，结合企业专属数据及个性化需求所建设的一站式大模型平台。在大模型通用能力之外，允许企业对大模型进行微调及训练，生成个性化 API，供企业调用及封装成新的企业应用。

案例介绍

企业专属大模型核心能力包含两条主线：在线、离线闭环一站式打通。

1. 在线服务链路：企业数据和 API 能力接入、大模型应用程序流程编排与 Prompt 构建、以及对客户来说不可见的 planning 能力，满足企业客户业务场景接入的需求。

2. 离线训练链路：训练数据管理、模型训练、评测与标注等能力，为企业提供专属模型。

- **数据管理能力：**支持企业上传结构化和非结构化数据 (比如 pdf、doc 等)、管理，系统自动完成数据解析、构建 /embedding，以支持在多文档、长文档等复杂情况下，大模型仍然能够处理这类问题。
- **模型开发能力：**支持客户进行 Finetune 等模型微调能力，并支持大模型的标注及评估（提供训练、标注、评估能力）。
- **模型调试能力：**提供大模型测试窗，供企业开发人员进行效果调试，同时产品也提供了调试参数的设置能力。
- **模型任务编排能力：**按照开发人员实际的业务场景，提供画布的能力，将多种模型任务进行编排，能力整合后可以统一输出。
- **插件管理能力：**企业专属大模型会预置能力插件（比如向量检索、定位等），也支持企业自定义开发插件接入。
- **API 管理能力：**企业专属大模型生成的 API 可以供企业管理及调用。支持添加企业系统的自定义 API，通过 description 被中控 planning 识别和自动调度。
- **企业专属的中控 planing 能力：**根据企业不同诉求，进入不同处理流程（模型自身能力、文档相关内容、企业系统调用）。
- **企业大模型应用的业务流程编排和 prompt 构建：**通过画布编排支持企业自定义大模型应用的业务流程，客户根据场景需要来构建 prompt。

效益分析

针对大模型生成内容和性能方面的优化，实现模型结果可控、响应提速、降低成本。

（1）企业端降本提效：基于大模型所构建的应用，可以帮助企业经营管理提效，可以提升企业服务效率，整体降低企业成本。

（2）企业端营收提升：基于大模型构建的应用，赋能企业销售及营销，可以整体提升转化率，给企业带来额外的收益。

（3）企业服务的消费者体验提升：企业基于大模型建设自己的消费者端应用，可以给消费者带来信息获取、服务提供等效率、质量、满意度多方面的提升。

昆仑万维“天工”大模型

昆仑万维科技股份有限公司

昆仑万维成立于 2008 年，是中国领先的互联网平台出海企业，2015 年在深交所上市。成立十五年来，昆仑万维始终致力于为全球用户提供领先的互联网产品与服务，现已构建了 AGI 与 AIGC、海外信息分发与元宇宙、投资三大业务板块，业务覆盖全球一百多个国家和地区，全球累计月活跃用户近 4 亿。

概　述

凭借对科技发展趋势的超前预判，昆仑万维早在 2020 年便已开始布局 AIGC 领域，至今已积累近三年的相关工程研发经验，并建立了行业领先的预训练数据深度处理能力。目前昆仑万维在人工智能领域取得了重大突破，已形成 AI 大模型、AI 搜索、AI 游戏、AI 音乐、AI 动漫、AI 社交六大 AI 业务矩阵，是国内模型技术与工程能力最强，布局最全面，同时全身心投入开源社区建设的企业之一。

“天工”是国内首个对标 ChatGPT 的双千亿级大语言模型，也是一个 AI 搜索引擎，一个对话式 AI 助手。“天工”拥有强大的自然语言处理和智能交互能力，能够实现个性化 AI 搜索、智能问答、聊天互动、文本生成、编写代码、语言翻译等多种应用场景，并且具有丰富的知识储备，涵盖科学、技术、文化、艺术、历史等领域。

2023 年 9 月 5 日，昆仑万维天工大模型在腾讯优图实验室联合厦门大学开展的多模态大语言模型测评中，综合得分排名第一；2023 年 9 月 16 日，在权威推理榜单 Benchmark GSM8K 测试中，昆仑万维天工大模型以 80% 的正确率脱颖而出，大幅领先 GPT-3.5（57.1%）和 LLaMA2-70B（56.8%），这标志着天工的推理能力达到全球领先，接近 GPT-4。2023 年 11 月 3 日，昆仑万维天工大模型通过《生成式人工智能服务管理暂行办法》备案，面向全社会开放服务。

需求分析

在当前信息化时代，各行各业对于自然语言处理技术的需求越来越大。在金融领域，天工大模型可以用于智能客服、风险评估、投资顾问等场景，为金融机构提供更高效、更准确的服务；在医疗领域，该项目可以用于医学文献分析、疾病诊断、药物研发等方面，为医疗行业提供更精准、更科学的支持；在教育领域，天工大模型可以用于智能辅导、学习推荐、学生管理等方面，为教育机构提供更智能、更人性化的教育服务。

总之，天工大模型项目的推出，将为各行各业带来更高效、更准确、更智能的自然语言处理解决方案，助力企业提升服务水平和竞争力。

案例介绍

应用落地情况

2023 年 11 月 3 日，昆仑万维“天工”大模型通过《生成式人工智能服务管理暂行办法》备案，面向全社会开放服务！用户在应用商店下载“天工 APP”或登陆“天工官网”（www.tiangong.cn）均可直接注册使用。

天工 AI 搜索

在应用端，天工 APP 全面迭代升级，整合 AI 搜索、AI 阅读、AI 创作等核心功能，覆盖工作、学习、生活等众多应用场景：

- **AI 搜索：**个性化搜索功能升级，用户可以输入个人信息和搜索用途，定制专属的搜索体验。通过自定义风格，优化搜索结果，提高信息获取的精准度和效率。
- **AI 阅读：**“天工”APP 能高效阅读分析文章链接或文档文件，生成 AI 摘要并一键提炼要点，帮助用户快速了解文章主旨、重点和关键细节。同时支持问答式交互，让用户更便捷地查询文档信息。
- **AI 创作：**用户可以提出创作要求，AI 快速生成各类作品，并进行持续交互调整内容和语言。支持改写、扩展、缩写、总结等功能，帮助用户节省时间提高效率，满足学术教育、职场文档、创意写作、广告营销等不同场景需求。

AI 游戏方面，公司旗下 Play for Fun 游戏工作室自研的首款 AI 游戏《Club Koala》于 8 月 25 日在德国科隆国际游戏展惊艳亮相。

Club Koala

昆仑万维在国外的 AIGC 探索和布局：

Opera 原生浏览器 AI 助手 Aria 迎来了新一轮加速发展，推出一系列前沿 AI 功能，帮助用户提高效率并释放创造力。首先是 “Refiner”工具，通过“重用”功能，用户无需重复手动输入提示词，仅通过选定历史交互内容，便可针对新需求生成答复；同时，通过 “改述”功能，用户可以对选定片段重新措辞输出，大幅提升改写效率与使用体验。其次是 “Compose”功能，通过选择内容类型、提供主题、语气及内容长度即可完成各类型内容创作。同时，用户还可以通过“My Style”功能训练 Aria 以符合自己的写作风格。

截至第三季度末，Aria 已在包括欧盟在内的 180 多个国家和地区上线，用户突破百万大关。公司元宇宙入口 Opera GX 也已全面接入 Aria，为广大玩家带来最前沿的 AI 浏览体验。

AI 助手 Aria

效益分析

2023 年 7 月，昆仑万维官宣与映宇宙集团母公司在 AI 业务方面达成合作，向蜜莱坞科技提供包括 AGI Sky-Chat SaaS API 服务及 AIGC SkyPaint API 服务，服务期限一年，总金额在 1500 万元人民币以内。此次协议的签署，标志着昆仑万维“天工”大模型在互联网社交行业的正式落地。

Chapter Two.

第二篇章

垂类大模型

2

2023

大模型落地应用案例集

Foundation Model
Practical Application Collections

梧桐·招聘 - 基于百度智能云千帆大模型平台的智能招聘系统

软通动力信息技术（集团）股份有限公司

软通动力信息技术(集团)股份有限公司(以下简称“软通动力”)是中国领先的软件与信息技术服务商，致力于成为具有全球影响力的数字技术服务领导企业，企业数字化转型可信赖合作伙伴。2005 年，公司成立于北京，立足中国，服务全球市场。

秉承用数字技术提升客户价值的使命，软通动力长期提供数字化创新业务服务、通用技术服务和数字化运营服务，其中数字化创新业务服务包括数字咨询服务、数字技术服务和数字解决方案；凭借深厚的行业积累，公司在 10 余个重要行业服务超过 1100 家国内外客户，其中超过 230 家客户为世界 500 强或中国 500 强企业。

概　述

梧桐·招聘采用先进的大模型技术，实现了对招聘需求的快速响应和精准推荐。用户可以通过系统快速搜索符合条件的候选人，并实现在线面试、评估和筛选。系统还提供了丰富的招聘数据分析和报告，帮助企业全面了解招聘情况，及时调整招聘策略。

梧桐招聘不仅是一款招聘工具，更是企业人才管理的重要平台。通过系统，企业可以更加高效地管理招聘流程，实现对候选人的全生命周期管理，提高招聘效率和人才质量。

需求分析

随着业务开拓和发展，企业用工需求激增，招聘成为企业发展的关键环节。然而，当前招聘体系存在诸多问题，如招聘流程混乱、简历筛选效率低下等，导致招聘难度加大，成为企业发展过程中的负担。因此，针对该业务场景，亟需构建一个高效、规范的招聘管理系统，以提升招聘效率和成功率，解决当前招聘体系存在的问题。同时，本项目将为企业提供更便捷的招聘流程和更高效的简历筛选工具，降低招聘成本，提升招聘质量。

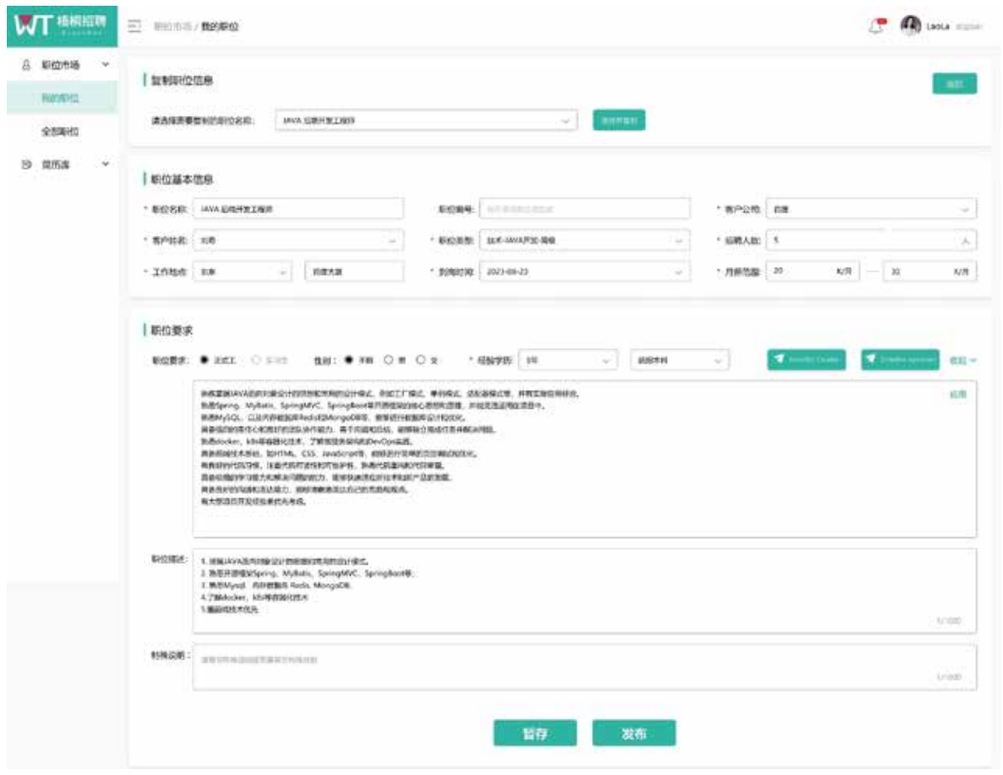
案例介绍

一、招聘信息生成（ErnieBot Creater）

基于表单的职位关键信息，结合大模型能力从专业技能、项目经验、沟通能力 等多方面生产岗位招聘要求

二、招聘信息优化（ ErnieBot Optimizer ）

对用户已填写的招聘信息进行优化。



三、人岗匹配

一个热门岗位一天可能会收到上百份简历，简历筛选的工作量巨大，通过与大模型能力结合，进行智能人岗匹配为用户提高简历筛选效率。通过对简历进行解析，提取专业技能、教育经历、工作年限、项目经验等等多个维度数据与岗位招聘信息进行匹配，最终计算出平均分，帮助业务提供决策建议。

四、面试题生成（ErnieBot Creater）

招聘和交付可能会接触到各行业、领域、技能的招聘要求，对于用工要求高的客户初面是一个重要环节，通过文心大模型能力可以快速生成专业面试题以及答案，不仅弥补了招聘的专业不足之处也帮助业务提高面试效率。

效益分析

一、项目经济社会效益

本项目的实施将带来显著的经济社会效益。首先，通过规范招聘流程和提升招聘效率，企业能够降低招聘成本，缩短招聘周期，提高招聘质量，从而更好地满足用工需求，推动业务发展。其次，通过简历筛选工具和数据分析，企业能够更快速、更准确地识别优秀人才，提升人才匹配度，降低人才流失率。最后，本项目将带来良好的社会效益，通过优化招聘流程和减少不必要的环节，能够减少信息不对称和就业歧视等问题，促进社会公平和稳定。

二、商业模式

本项目的商业模式主要是提供招聘管理系统和相关服务。通过收取软件许可费、实施费和服务费等方式，实现项目的盈利。同时，还可以通过与招聘网站、人才市场等机构合作，扩大项目的市场份额。

三、应用推广前景

本项目具有广泛的应用推广前景。随着企业对于人才招聘的重视程度不断提高，以及互联网技术的发展和应用普及，越来越多的企业将采用招聘管理系统来提升招聘效率和质量。同时，政府对于公共就业服务体系的建设也将促进本项目的推广和应用。因此，本项目的市场前景非常广阔。

面向游戏行业的图像内容生成式大模型

上海商汤智能科技有限公司

商汤科技作为亚洲领先的 AI 技术公司，拥有深厚的学术积累，并长期投入于原创技术研究，不断增强行业领先的全栈式人工智能能力，涵盖感知智能、决策智能、智能内容生成和智能内容增强等关键技术领域，同时包含 AI 芯片、AI 传感器及 AI 云等关键能力。商汤前瞻性打造新型人工智能基础设施——商汤 AI 大装置 SenseCore，打通算力、算法和平台，并在此基础上建立“商汤日日新 SenseNova”大模型及研发体系，推动高效率、低成本、规模化的 AI 创新和落地，进而打通商业价值闭环，引领人工智能进入工业化发展阶段。

商汤科技业务涵盖智慧商业、智慧城市、智慧生活、智能汽车四大板块，相关产品与解决方案深受客户与合作伙伴好评。目前，商汤（股票代码：0020.HK）已于香港交易所主板挂牌上市。商汤在香港、上海、北京、深圳、成都、杭州、南平、青岛、西安、京都、东京、新加坡、利雅得、阿布扎比、迪拜、吉隆坡、首尔等地设立办公室。

概　述

面向游戏行业的图像内容生成式大模型是一款商汤科技自主研发、面向游戏策划和美工等研发人员设计的高效研发辅助工具，可通过高质量、快捷的 AIGC 能力，快速、批量地生成风格多样的图像内容，大幅缩短游戏研发时间与人力成本，助力商汤科技内部业务快速落地。

需求分析

随着通用大模型的技术发展，面向人工智能内容生成（简称“AIGC”）的大模型正在从单纯加快时政、金融、体育等行业新闻稿的内容生成速度、降低人力成本，逐步转向以绘画、美工素材、剧本等生成应用的价值创造。跨模态 / 多模态内容成为关键的发展节点，OpenAI 的 CLIP 多模态模型、DallE 系列生成模型，Google 的 Imagen 大规模文图生成模型，让这一领域的技术不断完善。2022 年下半年以来，生成模型技术不断完善、开源模式的推动、商业化案例的落地，推动 AIGC 发展明显加速。

针对游戏行业，目前部分内容创作者的矛盾主要集中在越来越高的生成内容的丰富度、事实性和个性化的需求与有限的创作周期，同时也需要更加高效直观的图文跨模态转换技术，以实现团队协作。因此，急需高效、高质量的内容生成手段用于创作灵感的构思、辅助内容创作和团队需求沟通。而随着 AIGC 大模型的标注数据累积、AIGC 技术架构日渐完善，AIGC 技术逐渐开始面向内容创作应用。AIGC 技术借助大模型的跨模态综合技术能力，可以激发创意，提升内容多样性，降低制作成本，快速推动更加高质量的内容生成和创作。

案例介绍

主要能力

商汤科技目前已针对游戏业务正式部署了图像内容生成式大模型，可面向游戏美工、策划等内容创作者提供高质量、大批量的优秀图像内容。基于检索式超大生成扩散模型设计、质量感知式图像生成技术、图像布局分布式训练方式、面向超大生成扩散模型的训练和推理加速器以及基于超大生成扩散模型的图像二次创作能力等创新点，内容创作者可以实现更加精准的高质量图像内容生成，通过初步筛选图像内容，即可获得目标图像，并可以实现局部区域内容的定制化调整。因此，图像内容生成式大模型的应用大幅提升了游戏业务的沟通和研发进度，辅助内容创作者更加高效、灵活地创作游戏素材。

技术创新点

- **基于检索式的超大生成扩散模型设计：**利用条件信息从数据库中检索到相关的图像，和条件信息一起送入扩散模型生成过程中，强化模型对于条件生成的质量和丰富度；
- **质量感知图像生成技术：**通过训练一个预测风格和质量的模型对数据进行分类，并将类别作为条件信息的一部分，从而实现在充分利用大规模数据提供的图像丰富度的同时，有效控制生成质量；
- **图像布局分布式训练：**设计了一种图像预处理的流水线，将相近宽高比的图像组成同个批次，使得模型具有生成多种图像大小、宽高比的能力；
- **大规模扩散模型训练与推理加速优化：**通过对扩散模型的分析和优化，有效提升模型训练和预测的速度，实现了 2-5 倍的加速比，帮助模型实现目前业界最快的生成速度；

- **基于 AIGC 超大模型的高质量创作功能：**充分挖掘用户需求，设计了丰富的可调节生成选项，帮助用户通过调整提示词（prompt）来快速实现高质量出图；

- **基于 AIGC 超大模型的图像二次创作和修改：**设计了一种新的图像修改的流水线，可以有效利用没有被遮盖区域的完整信息，来辅助恢复遮盖区域的生成过程，实现更好的生成质量和边缘过渡。

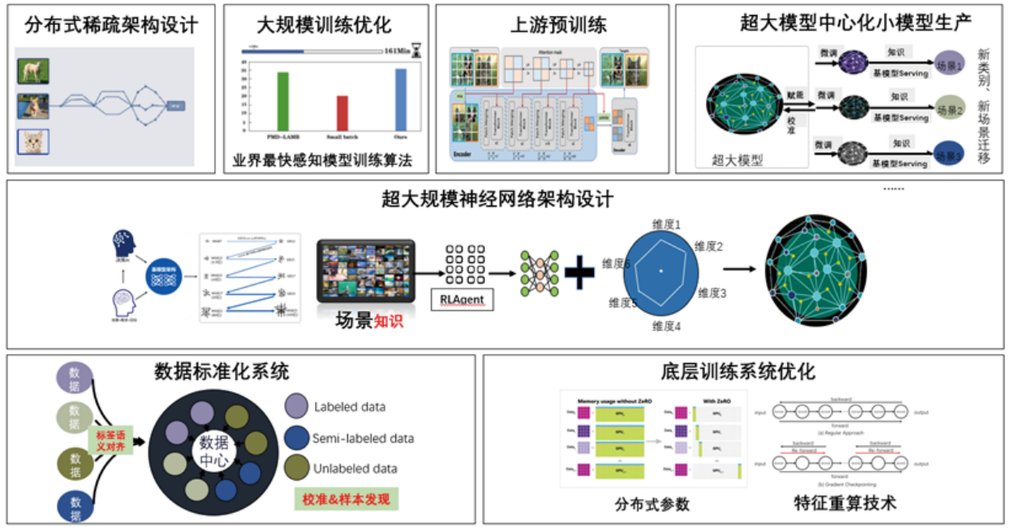


图 1　大模型架构图

实施效果

- **高效提升策划与美术设计沟通准确性：**通过使用图像内容生成大模型的 AIGC 能力，可将策划与美术设计沟通的时间和频次大幅缩短。传统的设计需求沟通通常需要 4-5 次对接反馈，将文字描述转换为最终的角色或者场景概念图，目前策划借助 AIGC 能力，可以直接将文字需求转化为图像内容，仅需要 1-2 次沟通对接即可让美术设计理解，大幅度提升了沟通准确性；

- **加速美术设计的生产效率：**美术设计借助图像内容生成式大模型的 AIGC 能力，可以高效、高质量生成不同质量的图像内容，用于辅助内容创作。如果需要根据沟通反馈意见进行修改设计，也可通过模型内容仅对局部细节进行替换。这些生产效率上的提升可将原有单个角色的创作周期从 10 天缩短到 6 天左右。而在图标的设计上加速效应更加的突出，按传统的工期计算为 1 天 2 个，现在接入 AIGC 能力后，美术能够在描述准确 + 风格确定的情况下，一天就可以生成上百个候选图标，而美术只需要简单修改就能够使用。

效益分析

项目亮点

- **高效、高质量地进行游戏素材内容创作：**游戏策划和美术设计都可以通过图像生成内容更加直观、更加高效的进行沟通和协作，有利助力了原创角色、场景以及其他游戏素材内容的创作；

- **使用 AIGC 能力助力游戏行业开发：**通过图像内容生成式大模型的生成能力，有利于推动游戏行业的开发模式更新，使游戏在角色设计阶段的开发周期大幅缩短，并且通过自定义设置大模型的提示内容，能够生成更高质量、定制化的设计方案。

经济量化价值

- 游戏策划与美术设计的沟通周期缩短，需求沟通和反馈频率从传统的 4-5 次变为 1-2 次；

- 加速了美术设计（角色原画）的生产效率，能够由原先的 10 天单角色加速为 6 天左右单角色。

- 加速了美术设计（图标）的生产效率，能够由原先 1 天两个加速为 1 天上百个。

应用推广前景

通过在商汤科技内部的游戏业务推动图像内容生成式大模型，尝试利用 AIGC 能力辅助游戏开发，可有利于解决现有游戏行业普遍存在的沟通和协作成本高、策划与美术设计存在理解偏差、开发周期短等迫切问题，将能够逐步实现 AIGC 能力在整个游戏开发周期中的渗透和推广，推动以 AIGC 能力为辅助工具的游戏开发模式形成。

中公网校：小鹿老师，为年轻人创造更多就业与成长机会

上海商汤智能科技有限公司

商汤科技作为亚洲领先的 AI 技术公司，拥有深厚的学术积累，并长期投入于原创技术研究，不断增强行业领先的全栈式人工智能能力，涵盖感知智能、决策智能、智能内容生成和智能内容增强等关键技术领域，同时包含 AI 芯片、AI 传感器及 AI 云等关键能力。商汤前瞻性打造新型人工智能基础设施——商汤 AI 大装置 SenseCore，打通算力、算法和平台，并在此基础上建立“商汤日日新 SenseNova”大模型及研发体系，推动高效率、低成本、规模化的 AI 创新和落地，进而打通商业价值闭环，引领人工智能进入工业化发展阶段。

商汤科技业务涵盖智慧商业、智慧城市、智慧生活、智能汽车四大板块，相关产品与解决方案深受客户与合作伙伴好评。目前，商汤（股票代码：0020.HK）已于香港交易所主板挂牌上市。商汤在香港、上海、北京、深圳、成都、杭州、南平、青岛、西安、京都、东京、新加坡、利雅得、阿布扎比、迪拜、吉隆坡、首尔等地设立办公室。

概　述

基于商汤“如影”数字人与“商量”语言大模型技术，中公教育通过 AI 技术分析优秀师资的教学过程，针对性训练虚拟数字人模拟他们的教学方法和风格，并通过数字化方式还原真实的教学场景，使得虚拟数字人能为学员提供高质量的学习课程。在教学过程中，虚拟数字讲师“小鹿”能依托专业的内容知识库，分析学员的学习数据，实现与学员的教学互动，为他们提供实时的反馈和建议，帮助他们更好的理解和掌握知识，提升学习效率。

需求分析

创立于 2003 年的中公教育，时至今日已经发展成为 1000 多个各地分校、超 5000 位专职师资，教育业务的“三驾马车”是教培图书、面授培训、在线课程，培训内容从公务员考试培训到教师类考试、金融财会类考试、医学资格类考试、法律类考试、社区考试、考研、成人高考、MBA/MPA 考试培训，以及 IT、人力资源、心理咨询、出国留学等职业技能提升相关培训。

当基于大模型的 AIGC 技术刚崭露头角时，中公网校在业内率先开启了教育产品革新，加速推动降本增效。对于教育机构，师资是最大的核心资产，也是最大的成本支出，中公网校采用“双师课堂模式”给万名学员上网课，分配最好的老师在线上讲解核心内容，当地教师进行线下面授与辅导，能够部分缓解全国市场对“名师”的供需矛盾。但这也带来了新的问题，名师授课需求旺盛挤占教研时间，名师无法有针对性地辅导每一位学员的个性化需求，以及名师离职风险。因此中公网校与商汤科技经过数月的联合研发，上线首款人工智能课程——“AI 系统班”，并发布虚拟数字讲师“小鹿老师”授课。

图 1　虚拟数字讲师“小鹿老师”

案例介绍

主要能力

基于商汤“如影”数字人与“商量”语言大模型技术，中公教育通过 AI 技术分析优秀师资的教学过程，针对性训练虚拟数字人模拟他们的教学方法和风格，并通过数字化方式还原真实的教学场景，使得虚拟数字人能为学员提供高质量的学习课程。在教学过程中，虚拟数字讲师“小鹿”能依托专业的内容知识库，分析学员的学习数据，实现与学员的教学互动，为他们提供实时的反馈和建议，帮助他们更好的理解和掌握知识，提升学习效率。

第一，与人类讲师上万人大课不同，小鹿老师是“因材施教”。AI 数智班是首个面向成人就业培训教育产品，不仅注重从业知识的传授，更加注重个人能力的培养和就业指导。从每个人的具体情况出发，人工智能“以点打面”能够为学员提供精准的职业规划和个性化的学习路径，帮助每个学员在将来的职场中建立起独特的竞争优势。

第二，小鹿老师具有清新活泼、知性气质的小鹿老师亲切形象，作为“高颜值名师”，让学员的学习交互过程更加生动有趣，与年轻学员们建立起师生情感纽带。

第三，“AI 系统班”让教师与广大学员“先人一步”掌握数字生产力技能。中公教研团队，逐步学习使用 AIGC 技术生产逐字稿课件内容，按照教学步骤将讲义逻辑、数字人视频、声音、板书、PPT 编排一致，并根据授课经验反复优化打磨，严格把控质量。另一方面，AI 系统班课程紧扣当前数字经济的发展趋势，让年轻学员学习使用 AI 技术与 AI 工具的应用实践，培养独树一帜的新型生产力人才。

技术创新点

- **声音优化很关键。**商汤如影 AI 视频生成平台只需上传少量音频，就能生成小鹿老师的人物音色，可以在线配置音色、语调、语速，一次生成视频可获得多语种多语言服务。小鹿老师课程的生产流程是：AI 把真人视频课转为文本，名师打磨文稿，文稿生成声音，声音驱动数字人。影响授课体验的关键因素是小鹿老师的声音，为了避免 “机械感连续朗读”，采取了优化变声、加入间隔停顿、语速语气优化等措施。客观来讲，中文数字人比英文数字人更难一些，因为有多音字、断句等挑战，需要老师们的一系列优化。

- **课件研发的反复调优。**例如将 10 位名师的同一门课程融为小鹿老师的一堂新视频课，需要教研老师们把每道题的知识点、讲法拆解，组合成最优稿，再经过院长把关，才能进入数字人录课流程。

- **形象逼真效果更好。**目前商汤如影数字人支持最高 4K 人物模型训练，让小鹿老师的嘴型张得更大，生成高自然度的音色、语调和语速，例如“bo”、“po”等闭口音发音与口型匹配度更佳，不断提升的如影基础模型让精益求精的 AIGC 效果。

- **数字人的互动形式多样化。**小鹿老师的口型、表情、动作越丰富，学员的上课体验越好，由商汤如影智能技术赋能、中公网校持续优化，正在让小鹿老师无限逼近真人名师的课堂效果。

- **训练效率更高。**在训练阶段，商汤如影数字人基于分钟级大数据视频样本、秒级小数据视频样本都能训练支持，且训练效率提升 50%，节省更多 AI 算力。在推理阶段，商汤如影 AI 视频生成平台，支持手机端、网页端平台视频制作，且能够满足短视频录播、直播等不同应用场景需求。

实施效果

采用商汤如影数字人技术后，小鹿老师对中公网校的降本增效效果初显、拓展空间很大，广受学员们欢迎。9 月 24 日小鹿老师的课程产品上线后，短短一个月时间，截止 10 月底就有超过 7 万名学员报名选购，成为中国在线教育产业的人工智能创新“风向标”。

图 2　小鹿老师在 “AI 系统班” 授课（中公网校官网）

效益分析

经济社会效益

1：相比于直播老师 1000 元 / 天的课酬，2 万元 / 月的薪水，小鹿老师直播一个月才花费 4000 元，仅人力成本一项就降低了 80% 的录课投入。通过使用手机上的商汤“如影”app 能够近乎实时生成大量小鹿老师的视频内容，满足短视频营销、课程录播、直播培训等场景需求，还不需要拍摄器材、场地投入。

2：小鹿老师能够“以一当万”，显著节省真人老师教学时间，帮助优质师资将宝贵时间投入到产品教研中，持续提升课程质量。

3：数字人产品价格的大幅降低，能够研发出多款不同面孔形象的“小鹿老师”分身，在抖音平台上开通运营多个抖音号，增加营销曝光度、多渠道聚集粉丝。

4：与真人讲师不同，小鹿老师讲课没有口误、啰嗦、重复，语言精挑细选、字斟句酌，所以数字人课程的内容含量是普通面授课的 2-3 倍，听课效率也是普通面授课的 2-3 倍。

5：商汤“秒画”文生图大模型在新课程营销中，为小鹿老师快速生成各式各样的海报，全渠道营销教培产品。

应用推广前景

面向教育培训机构，数字人 + 大模型的融合应用将持续探索“AI+ 内容”生产新范式。借助大数据技术深入分析学员的学习数据与需求，重构培训教学内容，并依托 AI 智能技术提升教学效率和学习工具智能化，在就业培训领域实现了教研内容“场景化”的突破，打破了传统培训模式的瓶颈。教育培训机构可以结合多年教学沉淀和研发积累，推出“AI 系统班”，致力于让每位学员都能以低成本触及最优质的教育资源，构筑起教培行业独有的“数字人即服务”模式、知识工程学习资产，利用 AI 大模型技术力量，不断反哺教育营销、教学、服务，更好的驱动业务发展。同时也依托人工智能技术，向就业培训市场输出更多低价优质内容，助力就业培训普及化，满足更多就业人群的学习需求。

新华妙笔 AI

北京信工博特智能科技有限公司

新华通讯社媒体融合生产技术与系统国家重点实验室。

媒体融合生产领域首个国家重点实验室。由中宣部指导、科技部批准，新华社承建。作为媒体领域战略科技创新平台，围绕推进媒体融合发展、重塑新闻舆论格局国家重大战略需求，聚焦人工智能等先进技术在新闻生产全流程应用，面向跨媒体大规模感知认知信息分析与推理、人机协同复杂问题分析响应及评估两个方向，开展媒体融合生产技术应用基础研究。

博特智能，国内领先的新一代人工智能大模型与 AIGC 应用高科技企业。累计申请发明专利和软件著作权五十余项。通过不断深耕大数据、深度学习和大语言模型（LLM）等人工智能核心领域技术，构建了通用型 AI 智能内容处理平台底座，并面向安全测评、内容安全、内容生成等领域推出四大类产品线，形成了多维度、广场景、强能力的基础性技术服务平台。

概　述

新华妙笔 AI 公文助手，采用最前沿的大数据、自然语言处理、AI 深度学习三大技术跨界融合，旨在用大模型技术重构规范写作流程，降低规范写作上手难度 ，人人可标准化创作精品文稿的目标。

在新闻写作、公文写作、商务文书写作、调研报告写作等规范写作领域打造的集查、写、审、学一体的在线公文写作能力提升平台，专注为大学生、公考生、教师、社区工作者、公务员、企事业单位等广大材料岗位人员、团队、组织，提供权威内容供给、内容决策辅助、内容辅助创作、内容 AI 审核、写作学习指导等多样化服务，助力知识密集型专业人士提升写作水平能力与知识赋能。

需求分析

新华社调查：公务员一年写近 300 份材料，公文写作具有非常强的高要求，首先表现在写作质量要求高，具体要求则是对思想站位、写作实效性、文风表达、内容法定性、行为规范性、表达特定性具有高要求，另一方面则对写作的人员要求比较高，优秀的公文写作人员，必须具备深厚的理论功底、精通专业知识、遣词造句精准、心态好能抗压、以及百科知识储备丰富。但优秀的“写材料”人在每个单位都是凤毛麟角。但是青年干部在公文写作中的现实困境是缺辅助提效工具、缺权威资料库、缺审核校对系统、缺优秀案例学习、缺知识常态积累。导致如何培育更多优秀“笔杆子”，成为一个老大难问题。

案例介绍

主要能力：

新华妙笔 AI，拥有 AI 生成、AI 润色、AI 校对、AI 续写、AI 灵感等写作功能以及文件分享、团队空间、单位管理、人员邀请、审阅批注等协作功能，以及学习库、范文库、模板库、素材库等专业权威数据库。可自动和辅助写材料人员完成重复性、标准化、简易创新类的内容创作。在法定性、事务性、规范性公文内容创作上的准确性、逻辑性、流畅性都逐渐接近专业人士水平。

在公文写作资料供给上，“新华妙笔”提供了超过一亿条权威数据、200 多种公文写作素材、资料、文稿数据内容案例供用户学习和参考，帮助用户提升公文写作质量；在功能上，“新华妙笔”自主研发的“问道”学习知识云，以智能问答方式构建了全新的理论学习实践平台，通过积累的概念知识库、领域话语体系知识库、涉政知识库和问题知识库，为知识密集型专业人士、组织、团队提供权威、系统和针对指导。

技术创新点：

① 多样性搜索结果约束算法→生成结果多样性和可控性两方面有效限制

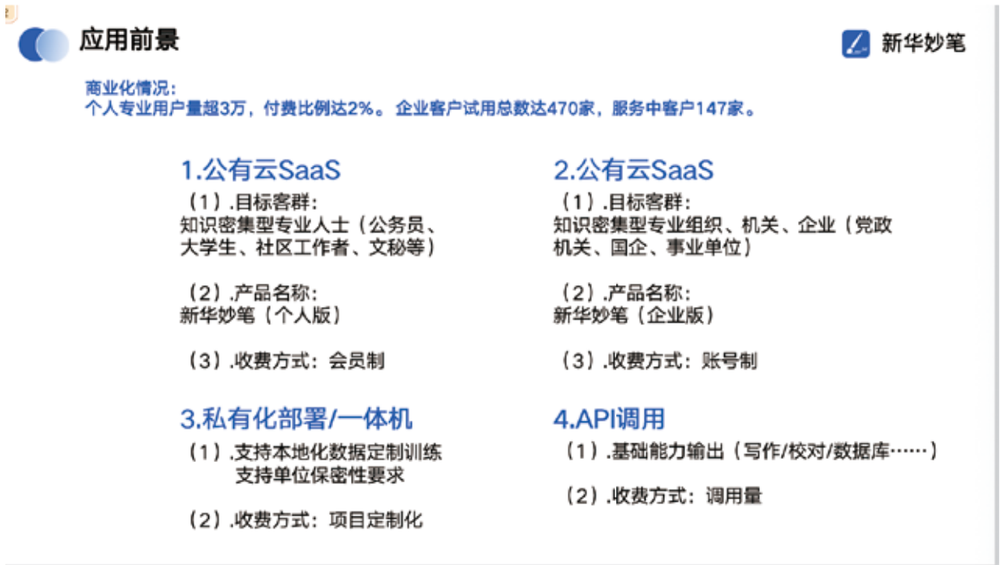
② 轻量级的模型与训练策略→快速匹配多种下游任务的同时不降低精度

③ 深度自主学习技术→模型输出有较高的语言组织能力

④ 调整模型位置编码复合策略→突破大模型输入输出长度限制，实现万字长文本的文稿准确输出

实施效果：

应用落地情况：



效益分析

项目经济社会收益

新华妙笔，目前个人专业用户量超 3 万，付费比例达 2%。企业客户试用总数达 470 家，服务中客户 147 家。

商业模式

- 公有云 SaaS（个人版）- 会员制收费
- 公有云 SaaS（企业版）- 账号制收费
- 私有化部署 / 一体机（项目制）
- API 调用（按照调用量收费）

应用推广前景

新华妙笔针对的客群是精准且具有高付费人群和组织

目标客群：全国体制内相关人员（2022 年）

类别	数量（个人）	部门	备注（企业数量）
行政编	719 万	政府职能部门	公务员
参公事业编制	1300 万	党政直属单位	党群、工会、妇联等
全额拨款事业编	3100 万	政府附属单位	公立医院、公办学校等
差额拨款事业编		政府附属单位	部分医院和高校等
工勤编		后勤单位	后勤保障工作
合同工	8000 万		自行招聘，无编制，依附体制单位
国营企业（编内）	860 万		国企央企
总数	1.3979 亿		超过 1000 万家目标单位

小布助手

OPPO 广东移动通信有限公司

OPPO 于 2004 年正式成立，是全球领先的智能设备创新者。目前 OPPO 的足迹已遍及 60 多个国家和地区，通过 260000 多个全球零售店数量及 3100 多个线下客户服务门店，与全球用户共享科技之美。作为一家软硬服一体化的科技公司，OPPO 不断优化以 ColorOS 为核心的软件平台，为全球 6 亿用户打造更人性化、更智能的移动操作系统。同时，OPPO 通过软件商店、云服务、智能助手的不断升级，为用户探索更快捷、更智能和更互联的增值服务。

概　述

小布助手旨在为 OPPO 智能终端用户提供服务，通过智能 AI 能力辅助用户在工作、生活等场景中满足需求。本模型可以理解用户的语音或文本指令，并根据指令生成相应的回复内容。其主要用途包括：支持用户查询信息、获得知识性问答、更好地操作手机、对长篇文章、文字做问答和总结、辅助电话智能应答以及获取网络服务等。

需求分析

经过过去几年的积累，用户已经习惯于通过智能助手来进行对话交流、信息查询、操作手机等，以上的能力在新技术的协助下，对用户的输入内容有更好的理解，同时能合安第斯大模型的自学习和涌现能力，提升用户的使用体验。

案例介绍

小布助手有几个典型使用场景：

1、支持通用问答和个性化应用

在问答领域，小布搭载了安第斯大模型能力，借助强大的自然语言处理能力，可以深度理解用户的问题和需求，从而提供更加准确和相关的回答，并根据用户的上下文和语境，做出个性化回答。在办公、学习、出行和娱乐等场景，小布都可以做到在保证回答内容满足国家安全标准的基础上，更专业、更流畅。

图 1

图 2

2、化身用机助手，帮助用户操作好手机

在用户使用手机时，小布支持调用和推荐各个层级的手机功能，完成手机操作，并且便捷的帮用户解答使用手机时遇到的各种问题；借助安第斯大模型技术，实现模糊复杂语义处理，深层功能控制和多功能调用。当用户遇到“功能找不到在哪里”“功能不会用”“用机出现问题，不知道是什么问题”等情况时，从系统控制切入，解决用户高频用机难题。

图 3

图 4

3、支持用户文本创作和图文生成能力

用户在生活、工作、学习等各个场景下，有快速获取灵感和图文生成的需求，但用大模型对提问思路和技巧有一定的要求，小布在产品设计上支持用户在对话中使用自然语言说出相关描述，快速创作，秒级出图。（图 5）（图 6）（图 7）

图 5

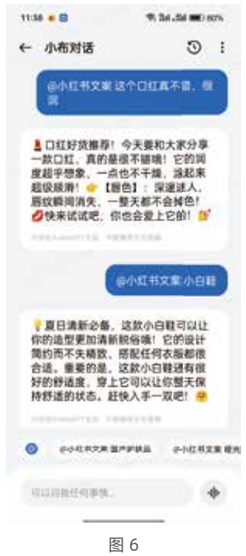


图 6

图 7

4、支持个人知识库

用户在工作生活中，需要阅读一些文档资料来查找知识和信息。小布支持用户在对话中上传文档和链接内容，进行智能摘要和问答，进而提升用户获取特定文档知识的效率，并存档后形成个人知识库。（图 8）

图 8

5、支持日程智能编排

在用户允许日程信息上传服务端解析的前提下，支持快速创建日程，用户可以灵活查询一切日程，时间、地点、人物、事项等都可以作为互相查询的条件。如：下周有什么安排、今晚和谁吃饭、和小王的会议是在哪个会议室。小布灵活安排、查询和管理日程，提升用户的时间管理效率，省心省事。（图 9）

图 9

效益分析

作为手机上的 AI 助手，一方面为能为最广大的用户群体提供来自大模型技术的先进产品功能和服务，以提升用户在日常学习、工作、生活事务的便捷性，激发创造力，例如由于能理解更复杂的指令，支持更加流畅的对话，所以能更便捷的操作手机，控制应用等；另一方面，由于结合了手机中特有的工具和能力，小布在产品功能的研发上也具备创新性，例如支持记忆用户的信息，以让每个回复更加个性化，同时支持为用户提供符合个人的建议和规划安排，支持生成符合用户风格的写作内容等。通过安第斯大模型能力，小布持续提升产品体验，让用户享受智慧化生活。

ChatDD 新一代对话式药物研发助手

北京水木分子生物科技有限公司

水木分子由清华大学智能产业研究院（AIR）孵化，清华大学国强教授、AIR 首席研究员聂再清教授担任首席科学家。致力于打造生物医药行业基础大模型及新一代对话式生物医药研发助手。公司产品服务于药物研发各环节，包括早研立项、靶点发现、分子设计优化、临床实验设计、药物重定位等。

公司于 23 年 8 月联合清华大学智能产业研究院（AIR）开源全球首个可商用多模态生物医药百亿参数大模型 BioMedGPT-10B，该模型在生物医药专业领域问答能力比肩人类专家水平，在自然语言、分子、蛋白质跨模态问答任务上达到领先。9 月发布了新一代对话式药物研发助手 ChatDD (Drug Discovery & Design) 和全球首个千亿参数多模态生物医药对话大模型 ChatDD-FM 100B，其在 C-Eval 评测中达到全部医学 4 项专业第一，是唯一在该 4 项评测中平均分超过 90 分的模型。

水木分子与复星医药正在开展基于生物医药大模型的药物价值评估应用，同时在推进与多家药企、CRO 应用合作。

概　述

ChatDD 新一代对话式药物研发助手，基于水木分子千亿参数多模态生物医药对话大模型底座 ChatDD-FM，具备专业知识力、认知探索力和工具调用能力。作为生物医药研发助手 Copilot 可以服务医药研发全流程场景，从立项调研，早期药物发现，临床前研究到临床试验、药物重定位等各环节。

需求分析

药物研发经历了从手工制药（TMDD）到计算机辅助设计 CADD 再到人工智能辅助设计 AIDD 的演进，每个阶段都带来了不同程度的效率提升和科学发展，为药物研发带来了新的机遇和挑战。第一代手工炼药，基于经验主义，通过大量实验试错来实现。第二代 CADD，通过计算机模拟建模，减少对湿实验的依赖。第三代 AIDD 应用人工智能技术从训练数据中挖掘药物发现和设计规律。但其面临训练数据不足，信息与知识分离，工具服务分散，处理模态单一等挑战。

水木分子提出的 ChatDD，基于大模型能力，能够对多模态数据进行融合理解，与专家自然交互人机协作，将人类专家知识与大模型知识联结，重新定义药物研发模式。

案例介绍

ChatDD 基于水木分子千亿参数多模态生物医药对话大模型底座 ChatDD-FM，具备专业知识力、认知探索力和工具调用能力。作为生物医药研发助手 Copilot 可以服务医药研发全流程场景，从立项调研，早期药物发现，临床前研究到临床试验、药物重定位等各环节。

图 1

在具体应用场景方面，ChatDD 应用于从立项调研、临床前研究、临床试验各药物研发环节。

ChatDD-BI 立项

立项报告是药物研发和 BD 的重要决策依据。一份高质量的立项报告需要繁琐的信息收集和整理工作，涉及到大量数据的搜集、整理和分析，比如专利和文献检索，市场信息的获取，以及管线与公开信息的对比

等等。对专家的依赖极高，还存在信息不全面的问题，这可能对项目的决策和发展产生负面影响。同时立项涉及企业商业秘密难以外包，这也增加了信息搜集的难度。ChatDD-BI 作为立项助手，通过与专家的数轮对话辅助立项调研，可以提升立项的效率和质量。

ChatDD-Discovery 研发探索

随着时间的推移，针对已知靶点的药物开发变的越来越挑战。药物研发的复杂性不断增加，需要更大的探索空间来发掘新的靶点、找到新的治疗方案。ChatDD-Discovery 通过与专家交互方式基于大模型知识建立药物、靶点、疾病关联，可进行潜在靶点挖掘研究探索。以 ALS（渐冻症）为例，ChatDD-Discovery 靶点挖掘结果被多篇生物医药文献证实为在研潜在研究靶点。

图 2　ChatDD 靶点信息挖掘

ChatDD 还可以通过工具调用、数据查询来提升研发效率。ChatDD-Discovery 通过自然语言交互的方式，将专家知识模型参数化，更好的辅助专家获取知识、查询数据库和使用工具。研究人员可以通过简单自然语言提示词，比如：“针对该给定靶点，为我筛选亲和力较高的小分子。” 来调用虚拟筛选功能。

图 3　ChatDD 调用蛋白 - 小分子相互作用预测

ChatDD-Trail 临床试验

临床试验是药物研发中耗资最多、风险最大的关键阶段，如何提高临床试验的成功率一直是制药企业关注的核心问题。ChatDD-Trial 凭借其知识能力和数据分析能力，可为提高临床试验的成功概率提供多重助力。例如，ChatDD-Trial 可辅助临床试验研究人员找到最适合入组的患者人群。通过发现药物敏感的生物标志物，更好地理解疾病亚型，实现精准的患者分类，确保患者与试验药物更匹配，减少不必要的变量干扰，提高临床试验成功率。

图 4　ChatDD 定位合适入组患者

图 5　ChatDD 服务模式

水木分子正在与复星医药开展基于复星医药具体应用场景进行私有化部署和应用开发，预期能够实现自研药物价值量化决策系统，赋能药企进行药物价值评估，进行科学的，可量化，可视化药物价值评估分析。

效益分析

ChatDD 作为新一代对话式药物研发助手，联结专家知识与大模型智能涌现与通用任务触类旁通的能力，有望引领下一代药物研发新范式。 从社会经济效益角度，ChatDD 通过缩短药物研发时间、减少临床试验失败率，降低医疗成本，对医疗体系和患者都有显著经济效益。

在商业模式上采取灵活多样的服务方式，包括服务订阅、提供应用开发 API、模型私有化部署等以满足客户需求。公司正在与复星医药、康龙化成、泰格医药、恒瑞医药、百济神州、华润双鹤等多家药企、CRO 推进生物医药基础模型应用讨论。

大模型数据分析智能助理 DeepInsight Copilot

支付宝（中国）网络技术有限公司

支付宝成立于 2004 年，始终致力于数字支付开放平台的建设和发展，于 2011 年 5 月获得中国人民银行首批颁发的《支付业务许可证》。支付宝践行“支付为民服务实体”的初心，研发了快捷支付、条码支付、刷脸支付、二维码支付等创新支付技术，服务于商业经营、便民缴费、交通出行等不同场景下的数字支付需求，为超 10 亿用户、8000 万商家提供支付服务保障，助力实体经济蓬勃发展。

成立至今，支付宝致力于以科技推动包括金融服务业在内的全球现代服务业的数字化升级，深入贯彻新发展理念，制定全方位的数字化发展战略，并大力增加科研投入，通过规范的金融科技应用提高金融服务效能、增强金融风险科技防范能力，以金融科技创新推动我国经济金融数字化转型升级。

概　述

当前，数据驱动的智能决策已成为企业数字化转型过程中的核心竞争力。伴随着大模型技术的突破和不断应用，已经开始影响和改变数据分析领域的产品形态和生产关系。例如，以微软 Power BI 为代表的产品率先发布了基于自然语言交互的取数、可视化和分析能力，极大降低了用户使用门槛并提升了使用体验。在此背景下，支付宝基于蚂蚁集团基础大模型开发研制了数据分析智能助理 DeepInsight Copilot。

通过提供对话式的交互，DeepInsight Copilot 极大降低了数据分析的上手门槛，用户无需理解复杂的产品界面，数据分析效率可提升至分钟级别。在企业高管、理财师、商家经营等场景中，从企业高管用户到一线用户，DeepInsight Copilot 都能够帮助客户更高效、更有效地获取信息和洞见，辅助客户做出更好的经营决策。

需求分析

相对于市场上已有的数据分析产品，DeepInsight Copilot 的产品能力关键是结合基础大模型的能力，通过更为的自然的语言对话交互方式，帮助客户实现数据指标查询、数据计算口径确认、数据分析建模、目标数据提取、数据报表分析、数据报表制作、数据可视化、数据分析代码自动生成、数据分析教学、数据分析方法智能推荐等功能，并可通过钉钉等多种办公软件入口提供给客户使用，在降低客户使用数据分析产品的使用门槛的同时，帮助用户提升找数、取数、分析和决策的效能。具体需求如图 1 所示。

智能化BI等级		L1				L2			L3	L4	L5
Copilot能力		数据分析教学	Chat2界面操作	Chat2DAL	Chat2Data（精准提问）	Chat2Data（模糊提问）	SeedReq2SuggestionReq	Chat2简报	Chat2Analytics	Chat2DS	AI Agent
用户价值	效率价值&效能价值	提升用户的学习和成长效率	降低门槛 简化生产关系 提升自助率	降低门槛DAL编写门槛，提升自助率	降低门槛 简化生产关系 提升自助率	降低门槛 简化生产关系 提升自助率	LLM提出有价值问题	提升用户的总结效率	降低门槛 简化生产关系 提升自助率	降低门槛 把不能变成能	降低门槛 把能变成高效的能
DI-Copilot能力描述	Input	自然语言	自然语言	自然语言	自然语言	自然语言	种子问题和数据特征	自然语言	复杂自然语言，多轮对话	复杂自然语言，多轮对话	任务目标，Agent内部多轮复杂对话
	Processer	推理需要的文档或者内容	指令翻译，报表结构生成	代码翻译	代码翻译+查询	知识查询，意图识别，代码翻译+查询	文本生成	文本生成：对多段文本进行归纳总结	推理，并执行多轮查询	推理，并执行多轮查询，以及统计分析	推理，多轮查询，统计推断
	output	教学内容	报表结构	DAL代码	数据	数据	有价值的问题	总结之后的简报	可能的洞见	可能的洞见	可能的洞见、决策
	关键能力指标	采纳率	翻译准确率	翻译准确率	翻译准确率	翻译准确率	问题采纳率	采纳率	CoT有效性	CoT有效性	决策建议的有效性
依赖的模型能力以及我们要提供的数据集	依赖的模型能力	① 文本分类 ② 毒性检测 ③ 事实回答 ④ 阅读理解 ⑤ 文章生成 ⑥ 问答 ⑦ 文本摘要	① 文本分类 ② 代码生成（DAL代码） ③ 代码理解（DAL代码）	① 文本分类 ② 代码生成（DAL代码） ③ 代码理解（DAL代码）	① 文本分类 ② 共指消解 ③ 代码生成（DAL代码） ④ 代码理解（DAL代码）	① 文本分类 ② 共指消解 ③ 代码生成（DAL代码） ④ 代码理解（DAL代码） ⑤ RAG with KG	① 文本生成	① 文本分类 ② 逻辑推理 ③ 文本摘要 ④ 文章生成	① 文本分类 ② CoT(ReAct, ReWOO) ③ ToT	① 文本分类 ② CoT(ReAct, ReWOO) ③ ToT	① 文本分类 ② CoT(ReAct, ReWOO) ③ ToT
	数据集	① 预训练 ② embeding ③ 精调 ④ 奖励建模 ⑤ 强化学习	① 精调 ② 奖励建模 ③ 强化学习	① 预训练 ② 精调 ③ 奖励建模 ④ 强化学习	① 精调 ② 奖励建模 ③ 强化学习 ④ 知识构建	① 精调 ② 奖励建模 ③ 强化学习 ④ 知识构建	① 预训练 ② 精调 ③ 奖励建模 ④ 强化学习	① 精调 ② 奖励建模 ③ 强化学习 ④ 知识构建	① 精调 ② 奖励建模 ③ 强化学习 ④（含CoT） ⑤ 知识构建	① 预训练 ② 精调 ③ 奖励建模 ④ 强化学习 ⑤（含CoT,ToT,Code Interpreter）	① 预训练 ② 精调 ③ 奖励建模 ④ 强化学习 ⑤（含CoT,ToT,Code Interpreter）

图 1　结合大模型的数据分析智能助理功能需求

案例介绍

在建设数据分析智能助理 DeepInsight Copilot 的过程中，为了充分发挥基础大模型的能力，给用户带去最佳的产品体验，并确保输出内容的安全、可靠、可信等，本项目设计出了一套结合大模型技术、知识图谱技术、向量搜索技术的整体应用解决方案。其中，大模型方面，结合数据分析场景的需求和数据对基础大模型进行了调优，并设计实现了面向数据分析场景的对话系统，从而取得业界领先的技术效果。关键技术创新点和实施效果如下：

- **意图分类：**相比起传统的提示词工程（Prompt Engineering、大模型调优等方法，本项目设计出一种多层分类任务 + 规则的方法，将意图分类的正确率提升了 20 个百分点，达到了 97.5% 的真实环境正确率。

- **使用自然语言对话生成数据分析建模：**通使用自然语言对话生成数据分析建模通过对基础大模型进行调优，支持用户通过自然语言生成取数分析 DSL Domain Specific Language），可简化用户特定领域的数据分析问题建模和数据分析。目前取数分析 DSL 的正确率提升到达到业界自然语言取数的领先水平。

- **Prompt 精简：**通过结合蚂蚁向量搜索技术，本项目将自然语言生成取数分析 DSL 的提示词 Token 长度精简 90% 以上。既能提升了自然语言生成取数分析 DSL 的正确率，也同时减少了需要计算的 Token 量缩短了响应时长。

- **对话系统：**本项目在业界第一次实现了基于流式的数据分析对话系统，这极大地降低了用户正确输入任务的门槛。即使用户输入了错误的问题，DeepInsight Copilot 也可以友好地引导用户提出正确的问题，从而让用户可以更轻松地完成数据分析任务。

- **基于知识的模糊取数和分析：**基于知识图谱技术，本项目设计出一个数据分析领域的知识增强方案，让用户无需完整地描述出指标和分析思路，系统就可以推理出用户所需要查询的指标，或推理出最适合的分析思路。

目前产品已经用在支付宝内部上线使用，使用效果如下：

- 支持用户进行取数、分析、定义度量，帮助用户取数 & 分析耗时从小时级别降低至分钟级别；
- 支持用户进行报表生成、图表生成，帮助用户制作报表，耗时从半天级别降低至小时级别；
- 支持用户在移动端快捷取数，帮助用户随时随地获取数据信息。

效益分析

DeepInsight Copilot 将提供给蚂蚁集团 & 阿里集团共 8W+ 用户使用，通过满足更多用户更快地完成数据分析，提升用户的决策效率 10 倍，从而提升集团整体的决策力，促进集团业务更快地完成创新；

Copilot 能满足更多用户是因为对话式的交互极大降低了数据分析的上手门槛，促进全民皆可分析；同时将取数、分析、度量定义、答疑等操作在对话中完成，可满足更高效率；用户无需理解复杂的产品界面，数据分析效率可提升至分钟级别；

DeepInsight Copilot 将结合着“支付宝商家经营助手”等产品面向 B 端客户使用，帮助 B 端客户更好地完成在支付宝体系内的经营，从而提升其对支付宝平台的活跃度，计划未来开放给支付宝 2000 万商家使用。

单晶炉自动化工艺识别多模态大模型

上海传之神科技有限公司（OpenCSG）

上海传之神科技有限公司（OpenCSG）成立于 2023 年，是一家致力于打造全球开源生态社区的企业，通过自研产品 StarNet 算法算力平台为支撑及 OpenNova Series 大模型算法系列为社区底座；具有算力管理、数据训练、推理等最全的大模型解决方案。团队核心成员均毕业于国内外知名院校，曾就职全球性科技行业的高层管理；主导落地亿万规模的项目成功落地，具有丰富的管理、项目经验。目前 OpenCSG 已完成联想创投、国信中数数千万元的天使轮融资。

概　述

单晶硅生产是光伏新能源领域的核心业务及相关设备材料，也是光伏太阳能发电板的主要组成部分，单晶硅的生产能力直接影响光伏新能源领域的发展速度及新能源光伏发电的电力输出能力，而中国的硅片产能占到了全球 97% 以上。国家“十四五”规划更是要求太阳能、风能等可再生能源在全社会用电增量占比超过 50%。近几年光伏年新增装机容量同比增长均高达 60% 左右，这些对单晶硅的产能需求创造了极大的市场空间。

ISM 无限光模是 OpenCSG 与西安恒新机电、曲靖阳光能源、西安创联电气及天通日进联合研发的针对光伏单晶硅全自动化生产控制的光伏垂类多模态大模型。

传统单晶硅生产虽然实现了从人工控制到自动化控制的升级，但是全自动化拉晶还是行业一直以来无法解决的痛点，这个问题导致单晶硅生产企业需要大量的生产人员，同时依靠经验的生产也导致产能不稳定，无法高效生产。

ISM 多模态大模型则是通过单晶硅生产中的视觉数据、工艺数据及设备实时运行工况数据（温度、压力、氧含量、氩气、拉速等等上百个数据），实时分析判断当前生产情况，判断工艺数据，并反向指导自动化设备的控制流程。实现了单晶炉全自动化拉晶的技术突破。

需求分析

本项目是由开放传神 (OpenCSG) 与西安恒新机电、曲靖阳光能源、西安创联电气、天通日进联合研发。单晶硅生产行业属于新能源类生产制造垂直领域，中国作为光伏硅片全球的主要生产国，急需要产能优化改进以满足市场需求。而利用 AI 技术实现工艺流程中的状态识别及反向输出控制参数是解决全自动化拉晶，实现产能提升的唯一手段。大部分 AI 公司因为没有行业技术、工艺及相关设备数据作为模型训练的支撑数据，无法涉足此领域。

而我们与光伏行业头部企业进行了深度合作和联合研发，共享工艺、数据、技术等高价值行业信息，从而得以突破并实现全自动拉晶技术的革新。

ISM 无限光模是基于多模态大模型的单晶硅全自动生产分析系统，主要通过实时分析当前单晶炉炉内的视觉影像、炉内环境的监测数据、PLC 和上位机的运行工况及工艺参数，实时对当前的工艺运行情况进行分析，能够准确识别判断化料、稳温、引晶、放肩、等颈、收尾等几个核心工艺的时间节点，并反向指导单晶炉电控的运行。解决了之前需要人工介入判断相关工艺时间节点及是否正常运行的问题。

图 1

ISM 无限光模使用 ViT+Transformer 的多模态大模型技术框架，利用 ViT 作为视觉数据编码部分，同时利用自主研发的 D2T 技术，将设备运行数据、工艺数据等非文本类数据，通过定制编码技术转换为类似文本的输入 embedding 数据，利用 Transformer 架构将生产过程中的视觉和运行数据映射到同一个维度空间，输出当前工艺阶段分析数据（如是否可以进入下一阶段、当前阶段是否正常等）和对电控系统的控制指导数据（如目标功率温度等）。从而降低生产过程对有经验拉晶工的需求，增加拉晶过程的自动化率，稳定生产从而提升产能。

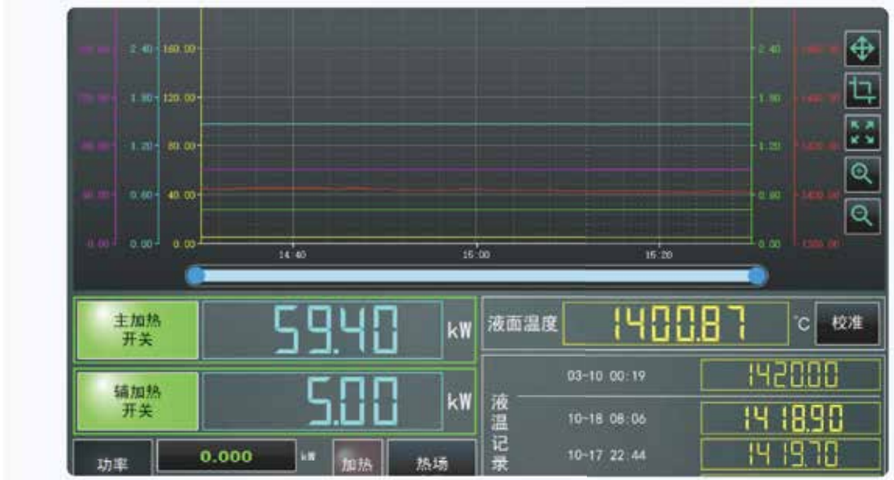


图 2

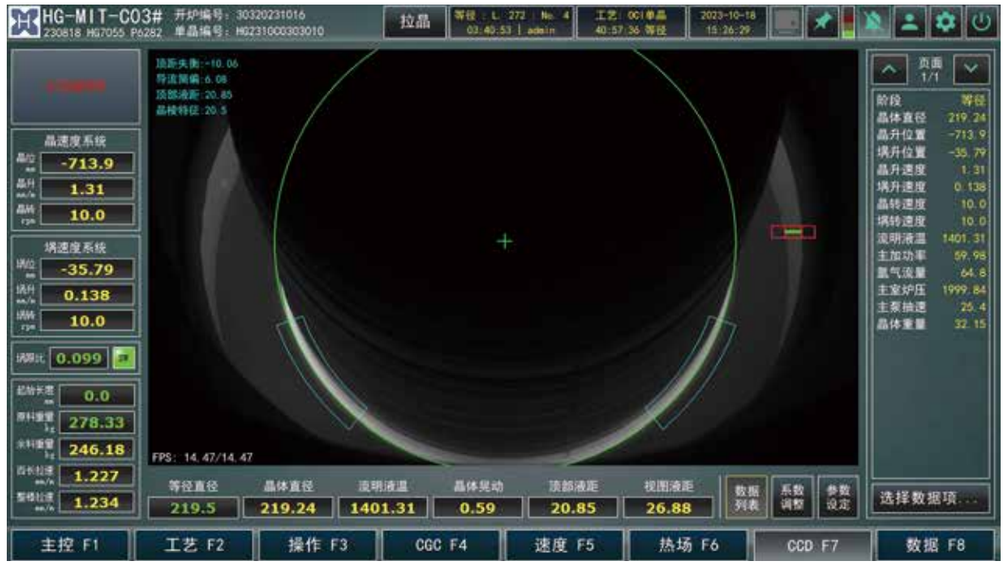


图 3

在 2023 年 6 月 ~ 2023 年 10 月期间，在合作伙伴实验炉台内进行了连续长期的实机测试，在 90% 的识别判断需求中实现了接近 100% 的准确率。极大的提升了电控设备的自动化生产率。

图 4

下一步将整合到现有炉台电控系统中，成为一个不可或缺的工艺流程组件，并在创联电气、阳光能源、天通日进等企业的相关项目中落地部署。预计 2023 年 ~ 2024 年将会实际落地 3 家以上的生产企业，支撑 1000+ 台的单晶炉的识别分析任务。

效益分析

ISM 无限光模对单晶硅生产的自动化改进，极大的提高了生产的稳定性及产能，同时降低了对经验人员的需求，将老师傅的经验转化为了自动化的技术。

单晶硅的生产是一个极其耗能的过程，需要将硅料融化并重新结晶，炉内温度高达 1400 度，一次拉晶失败就需要近 10 个小时的重新准备，期间消耗的能量和时间是极大的。ISM 无限光模极大提高了单机硅生产的全自动化能力，降低了各个工艺阶段拉晶失败的可能性，提高保证产能的同时，也节省了大量的电力和其他能源无效损耗，也符合国家提倡的双碳政策。为环保节能以及企业的高效发展提供了极大的保证。

并且我们使用的算法模型是针对垂直行业定制的模型，模型大小和对算力的消耗都较小，单台单显卡服务器就可以实现一个班组甚至一个车间的分析计算需求及，单炉台新增成本几乎可以忽略不计。并且按照每炉台收取年订阅使用费，保证企业可持续化收益的同时，也能保证为客户不断优化改进算法和工艺，真正做到双赢。

按照中国 2023 年计划 1000GW 的光伏产能计算，市场单晶炉保有量在 6000 台以上，而我们的合作伙伴的控制系统几乎占有市场的 1/3，未来落地应用前景十分广阔，商业化潜力巨大。

基于 NDAI 大模型的政务元宇宙平台

九度数字科技（苏州）有限公司

九度数字科技（苏州）有限公司，成立于 2022 年，是创新的人工智能和元宇宙场景应用解决方案提供商。基于大数据、人工智能、区块链、数字孪生等技术，打造“九度天问 AI 大模型”和“九度星云 AI 元宇宙平台”。依托平台数字底座能力面向政府政务、文旅、教育、乡村振兴等场景，提供人工智能和元宇宙场景应用综合解决方案。其产品已荣获 CCID“2023 行业信息技术应用卓越产品“、2023 江苏省信息技术优秀创新成果、江苏省人工智能科学技术奖、江苏省优秀人工智能应用解决方案等奖项。公司是国家级科技型中小企业、江苏省民营科技企业、江苏省人工智能学会理事单位，入选 2023 年元宇宙城市创新企业 TOP30。

概　述

基于 NDAI 大模型的政务元宇宙平台是九度数科联合昆山市花桥经济开发区便民服务中心，响应政策指导，落实数字政府、优化政务服务的需求，依托“九度天问 AI 大模型”和“九度星云元宇宙平台”核心能力而研发的 AI 数字化服务产品，主要面向企业、群众办事场景，创新性打造以 AI 客户服务为核心的智慧型、一体化、多模态、全场景、可持续的政务服务窗口，构建面向公众的一体化 AI 元宇宙政务服务体系。通过 AI 政务数字人、AR 智能导航、AI 云辅导、元宇宙空间（大厅）、全息显示系统等核心产品，提供“7×24 小时不打烊”的政务服务。提升企业、群众办事满意度，助力优化政务数字化，打造数字政府新标杆。技术已达国内领先水平，是 AI 大模型技术在政务服务场景创新成果，具有极大的行业创新引领和示范效应。其项目的产业化将产生巨大的经济和社会效益，应用推广前景巨大。

需求分析

2023 年 2 月 27 日，中共中央、国务院印发《数字中国建设整体布局规划》，将“政务数字化智能化水平明显提升”作为到 2025 年数字中国建设的目标之一，明确提出“发展高效协同的数字政务”，为进一步推进数字政府建设指明了方向。

国办发〔2023〕29 号文件，《关于依托全国一体化政务服务平台建立政务服务效能提升常态化工作机制的意见》中，明确指出：强化新技术应用赋能机制，探索利用人工智能等新技术，分析预判企业和群众办事需求，通过智能问答、智能预审、智能导办等方式，提供智能化、个性化、精准化服务。工业和信息化部等五部门联合印发《元宇宙产业创新发展三年行动计划（2023—2025 年）》，提出实现元宇宙典型软硬件产品规模应用，打造虚实融合的公共服务场景。加快数字人客服、实景导航等在政务服务应用，构建面向公众的一体化元宇宙政务服务体系。2023 年全国多地出台发布实施大模型示范应用推进计划的相关政策，重点支持在智能制造、生物医药、智能化教育教学、科技金融、数字政府等领域构建示范应用场景，打造标杆性大模型产品和服务。

作为数字政务的服务窗口，行政审批是链接企业和群众的前台，需要应用新技术、新模式，打造更为便捷的便民服务和高效的营商服务，推动城市高质量发展。

案例介绍

基于 NDAI 大模型的政务元宇宙平台，具有全场景、多模态的服务特点，基于“九度天问 AI 大模型” 和“九度星云元宇宙平台”核心能力，构建面向公众的一体化 AI 元宇宙政务服务体系，提供沉浸式政务服务体验。

图 1　整体架构图

其主要技术包括：多模态多场景智能建模技术；数据采集与智能分析技术；自然语言处理技术；大模型辅助知识获取技术；知识增强大模型技术；文本生成应用技术；数字人生成及交互技术；元宇宙平台技术；无介质全息成像技术。

主要产品和功能包括：

（1）AI 政务服务双数字人

覆盖线上、线下、元宇宙政务大厅全场景，实现智能文本交互、智能语音交互、数字人交互等多模态服务能力。业界首创基于大模型技术的双数字人模式。

图 2　AI 政务数字人

（2）AI 政务元宇宙服务大厅

打造可实现全套感知交互服务的元宇宙政务平台， 企业及群众可以自己的数字身份与 AI 政务数字人对话、办事等，形成真实的沉浸式服务体验。

（3）AR 智能导航

AR 智能导航结合了计算机视觉、增强现实、位置定位和人工智能等技术。为用户提供精准的实景 AR 导航指引，便民为民，增强用户服务体验度。

（4）VR 政务全景导视

实现政务风貌 VR 漫游，用户可以通过特定路线或自由路线进行政务服务 VR 查询服务。支持 VR 全景图、图文、视频、语音等展示形式。

（5）AI 云辅导

依托“九度天问 AI 大模型”平台能力，搭载 AI 政务数字人、政务场景化模板、AI 配音、智能脚本生成等核心制作功能。

（6）政务全息显示系统

打造创新的政务全息显示系统，具有空中成像、裸眼可视、实时交互等特点。

基于 NDAI 大模型的政务元宇宙平台，自建设实施应用以来，已取得非常好实施成效。

一是创新引领和示范成效突出。将前沿 AI 大模型技术在政务公共服务领域创新应用，起到了行业创新引领和示范效应。

二是应用和服务成效显著。构建了面向公众的一体化 AI 元宇宙政务服务体系，实现了线上、线下、元宇宙空间全场景，提供“7×24 小时不打烊”政务服务。其中，AI 政务服务数字人、AR 智能导航、元宇宙大厅等场景累计服务办事企业和群众 192.9 万人次，覆盖市场准入、公安、税务等 30 多个业务条线。AI 云辅导服务，共计输出业务办理短视频 52 条，线上线下载体发布 312 条，总播放量达到 140.6 万次。

项目实施总体提升了花桥经济开发区数字政府治理服务效能的显著提升，不断提升了企业和群众的获得感、幸福感。推动了花桥政府数字化改革，切实提升了营商服务和便民服务能力。

效益分析

基于 NDAI 大模型的政务元宇宙平台，其产业化将产生巨大的经济社会效益，具有很强的持续发展能力，推广前景非常广阔。

- **增强政府服务能力：**通过 AI 政务服务数字人、元宇宙政务大厅、AI 云辅导等产品，政府部门可以提供更为全面和便捷的政务服务，增强政府服务能力，提高公共服务效率。
- **降低政务服务成本：**通过自动化的处理流程和智能化的决策支持，可以降低人力成本和处理成本。
- **改善民生福祉：**通过 AI 应用为市民提供 7*24 小时全天候的政务咨询服务和数字交互服务，方便市民随时随地获取政府服务，提高公众满意度。
- **创新商业模式：**通过数据分析和预测，可以发掘新的商业机会和增长点，推动政务服务的创新和发展。
- **推动可持续发展：**积极落实数字中国战略，提升政务数字化智能化水平，通过 AI 大模型和元宇宙等数字技术，可以对城市进行预测性智慧化管理，帮助政府更好地规划和建设数字城市，提高城市数字化可持续发展能力。

慧政大模型——面向政务服务垂直大模型

北京中科汇联科技股份有限公司

北京中科汇联科技股份有限公司 1999 年成立于北京中关村。是一家致力于数字内容管理、人工智能交互、元宇宙与 AI 智能决策系统和行业解决方案的人工智能企业，致力于人工智能赋能产业化进程和各行业数字化转型升级，为全国各级党政机关、大中型企业、金融、教育等行业提供互联网内容管理平台、智能客服与机器人平台、智能指挥调度与虚拟智人平台的建设、运维和服务。

在产品和研发技术方面，中科汇联以 3C for 3C 的产品长期研发战略。以产、学、研三位一体研发体系，本着“软件就是智慧”的研发思想，汇联了清华大学、北京大学、哈工大等科研院所联合实验室和众多优秀人才组成人工智能研发团队，拥有自主可控且持续创新的人工智能核心技术，实现了自然语音理解、语音识别与合成、图像识别的全栈人工智能核心引擎。

概　述

面向政务服务大模型 - 慧政，除智能政务问答外还包含人力资源、写作、翻译、助手四大类别分类下的 18 个应用供用户选择，包括新闻内容撰写、翻译助理、社交媒体文案助理、战略咨询顾问等应用，根据需求定制搜索选择所需应用，添加至工作区，这一智能分类与定制策略不仅提升了政务问答、新闻撰写、多语言沟通等领域的效率，也为政府部门在日常工作中带来更精细、智能的支持，拓展了政务服务的可能性。再结合不断吸收海量文本数据中的新知识和信息，“慧政”大模型的效果也在不断地进化和提升。“慧政”大模型未来可广泛应用于机器翻译、智能客服、智能写作、情感分析、舆情监测等领域，为人们的生活和工作带来了很大的便利和效益。

需求分析

中科汇联自主研发的 AiGCP 智能生成大模型平台应用基于大数据预训练、多源知识融合、多模态大模型指令集微调等技术，具备超强的语言理解和生成能力。可以通过预训练语言模型，自动识别语言中的词汇、语义、句法、情感等信息，并能够进行分类、命名实体识别、语义解析等任务。可以基于预先训练的语言模型，生成与人类类似的自然语言文本，帮助用户完成对话生成、文学创作等任务。

案例介绍

面向政务服务大模型 - 慧政，是中科汇联自主研发的 AiGCP 智能生成大模型平台，该平台基于 LLM110 亿参数规模、可信中文数据源训练、国产信创支持、可私有化部署的垂直行业大模型平台。平台支持多模态、大模型指令集微调，实现三大应用·智语（上下文多轮对话）、智画（文本生成图片）、智人（数智人交互）。中科汇联基于 AiGCP 智能生成大模型平台，推出了面向政务行业大模型 - 慧政、面向医疗行业大模型 - 阳明以及面向金融行业大模型 - 慧金等系列垂直行业大模型产品。

慧政大模型解决的痛点问题

1. 数据安全引起日益关注，行业数据隐私性强，数据无法进出场，构建私有大模型势在必行。

2. 预训练大模型需要大量数据和算力，中小型企业无法满足资源要求，算力成为稀缺资源。

3. 通用大模型难以满足较多的定制化开发和复杂场景适配的要求。

4. 构建自有大模型的周期长，专业度高，缺乏配套的流程和工具支撑。5. 预训练大模型技术发展迅猛，流派众多分支庞杂，切换成本高。

应用案例

某市政务服务智能问答应用　　实施日期：2022 年 9 月

通过搭建市区两级问答知识服务的协同联动模式，围绕政策文件、办事服务、互动咨询、机构信息与网站使用这五大类问答知识要求，完成了市级部门问答知识库建设与运行服务。由于建立了健全政务问答知识目录与问答知识标签体系，基于用户意图与政府领域业务特征，在五大类问答知识划分基础上，征集各区各部门建议，进一步完善细分子类，形成全市统一的问答知识体系；

改善了某市政府门户网站及各区各部门网站智能问答知识缺项、漏项问题，回复率达 80% 以上；优化全市集约化平台问答知识更新机制，避免知识失效或错误，精准答复率约 75%。

某市区级政务服务智能问答　　实施日期：2021 年 3 月

某市区级政务服务机器人是某市区级政务服务管理局顺应新时代发展助力某市区级建设数字政府、智慧政府的重点项目。直接关系到某市区级政府智能化服务整体形象，直接影响某市区级政府网站智能化服务水平。某市区级网站访问量和政务服务事项网上办理受理量测算，覆盖用户量将在 100 万人次以上，并且随着某市区级政府智能化服务推广，应用访问量将逐年递增。作为全国首款政府网站智能语音交互机器人，其设计定位是全国优秀政府网站智能化服务标杆。

效益分析

从社会效益角度看，慧政的应用广泛涵盖了政务服务、新闻传播、多语言沟通等领域，为政府部门、媒体和企业提供了更为智能、高效的工具。政府部门可以通过慧政更好地与公众进行互动，提供更精准的信息和服务。新闻媒体可以借助其在新闻内容撰写方面的应用，加速新闻报道的速度和质量。企业可以在多语言沟通和翻译方面节省时间和资源，拓展国际市场等。

从经济效益角度来看，慧政的应用可以显著提升工作效率和生产力。节省下来的时间和资源可以被用于更有价值的工作，从而提升产出。此外，慧政还为政府和企业提供了更智能的决策支持，有助于优化资源配置，提高整体经济效益。

总之，面向政务服务的慧政项目通过智能分类、定制策略和不断进化的特性，不仅提升了工作效率，还拓展了政务服务的范围，促进了信息的快速传播和国际交流，为整体社会和经济发展带来了积极影响。

基于循道政务大模型的免申即享系统示范应用

上海卓繁信息技术股份有限公司

上海卓繁信息技术股份有限公司成立于 2001 年 7 月，总部位于上海，在全国设立有 10 家省级分公司，在湖南设立子公司。作为全国首套政务服务软件提供商，卓繁深耕数字政府建设领域二十余载，在行业内率先成立数字政府研究院，为“数字政府”建设提供完整咨询解决方案和技术支撑，产品线覆盖“一网通办”、“一网统管”、“数字乡村”、咨询与规划等业务领域，踊跃参与相关咨询课题和标准规范的研究。目前业务遍及全国 29 个省（直辖市、自治区），服务 800 余家政府用户，助力用户打造了一批在全国具有标杆意义的创新应用，在业内荣获了多项权威资质，主要包括：软件能力成熟度 CMMI5 级资质认证、信息系统安全集成服务资质 CCRC 二级、ITSS 信息技术服务标准符合性二级资质认证等。

概　述

为更大力度优化营商环境，持续激发市场主体活力，国务院办公厅发文要求抓好惠企政策兑现，推行惠企政策“免申即享”，各地纷纷推进免审即享改革。基于国家政策与各地实践，卓繁信息推出免审即享系统，引入了大模型技术，将原来全部人工处理的基本信息梳理、数据需求梳理、数据治理分析等环节优化为大模型处理，人工确认的方式，并充分发挥大模型强大的自然语言理解能力和生成能力技术，精准理解用户的数据需求，解读政策，智能地生成数据需求的语句脚本。进一步加速了数据产品的交付，提高了数据治理的效率，降低了对数据专业人才的需求，减少了数据治理成本，提高了数据的质量和可用性。

需求分析

为解决政策兑现慢、落地难等问题，政府致力于通过人工智能技术精准匹配符合条件的企业，企业全程无需主动提出申请，就能直接享受相关政策优惠，实现全程零材料、零跑动、过程无感、结果有感。基于现场实地调研和访谈，梳理出以下需求：

- **准确的政策梳理：**政策文件术语复杂，涉及多部门和多领域信息，需要政府专业工作人员投入大量时间来理解和梳理政策内容。利用大模型技术自动解析政策文件，提取关键信息，准确高效地实现政策梳理，能够极大提高政府工作效率。
- **高效的数据治理：**由于政府不同部门间数据字段命名和格式不统一，导致数据治理工作推进难度大，急需人工智能技术实现多源数据自动分析、挖掘，提高政府部门数据治理效率。

案例介绍

项目的主要能力

在“免审即享”推进过程中，存在事项基本信息梳理专业性高、难度大和数据治理分析耗时久等痛点，卓繁研发的循道政务大模型通过以下能力能够有效解决这些痛点：

1、自动化的政策文档解析，借助大模型的自然语言理解、信息抽取能力、生成能力，显著减少政府工作人员在理解政策文档上的工作负担。

2、跨部门、跨领域异构数据的联合查询和分析，使政府能够更全面地了解数据。

3、定制化的分析报告生成，借助大模型的迁移学习和生成能力，能够根据具体需求生成高度定制化、多样化的分析报告，为政府提供更有价值的数据分析结果。

4、智能化的数据治理协助，通过大模型简化数据治理流程，优化不同业务系统之间的数据融合，从而提高政府的数据治理效率。

技术创新点

1、大语言模型底座是经过政务领域数据微调、优化加速过的垂域大模型，支持私有化部署，有效保证用户数据的安全。

2、上层应用与大语言模型底座是松耦合关系，可以根据用户需要，替换为文心一言、盘古等通用大模型，使循道政务大模型有更好的兼容性，更灵活的落地能力；

3、 在微调大语言模型的基础上，结合卓繁二十余年对政务领域的探索经验，搭配外挂知识库、提示工程等技术手段，可以有针对性、定制化地解决用户痛点。

项目实施效果

1、通过循道政务大模型，自动化完成政策文件的解读、梳理，政策文件解读效率提升 70% 以上，同时，大模型生成的结构化数据可以用于后续的分析研究。

图 1

2、 通过循道政务大模型，实现自然语言转换成 SQL 查询语句 (NL2SQL), 将用户的语言表达转化为大数据治理平台实际可执行的命令，协助数据治理，使企业数据治理时间缩短 80%。

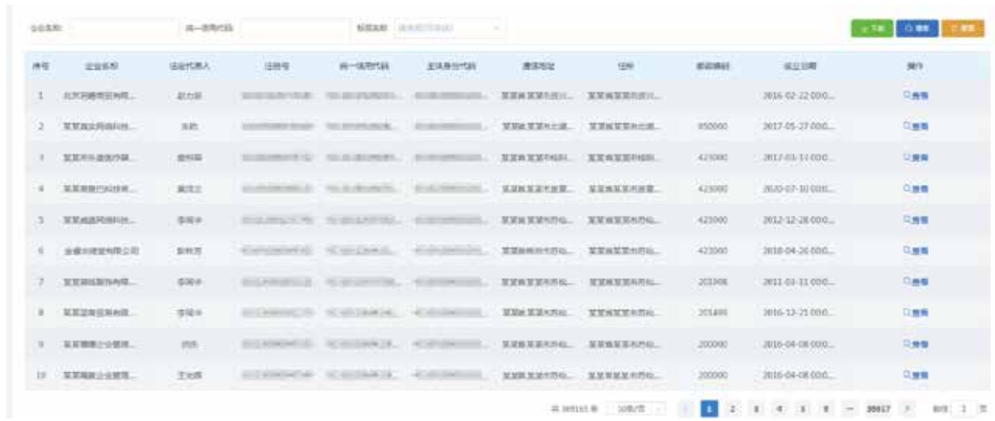


图 2

3、 通过循道政务大模型，以高质量提示工程加外挂知识库方式，实现智能定制与可控生成，自动识别关键词和主题，生成更高质量的问答结果。

图 3

项目应用情况

该项目已成功在芜湖、福建和湖南等地多个政务服务场景落地应用，提供卓越的数据治理服务，成功为政府部门解决了政策信息提取和多源数据分析方面的棘手问题，同时彰显出其通用性和广泛适用性。

效益分析

一、经济效益

系统研发费用共计 850 万元，公司收费方式为提供服务，预计 2023 年和 2024 年可以新增 15 家稳定的客户，实现用 1 年的时间，在数字政府行业内打响“循道”大模型品牌，用 5 年的时间取得该行业前三的目标。预计项目累计产生销售收入 15000 万元，利润 5000 万元，税收 750 万元。

二、社会效益

- **提高政务数据平台价值：**提供定制化的数据剖析和建模服务，增强政务数据平台的价值。
- **缩短项目周期、提高效率：**快速提供所需数据，缩短项目周期，提高工作效率，降低成本。
- **数据安全合规：**项目在数据处理的各个阶段都遵守高安全性的规则，以确保数据安全。

三、商业模式

1、与政府携手合作，以政企合作模式筹建数据交易所。

2、基于通用大模型，针对政府客户采用本地化部署云服务资源的方式，构建可定制、可训练的具有自主可控性质的政务行业大模型，以赋能政务数据高效治理。

四、应用推广前景

循道政务大模型能够使基础数据赋能政务行业发展，激发数据要素市场价值的潜力，从而带来更广阔的市场空间。

东方财富自研金融大模型

东方财富信息股份有限公司

东方财富信息股份有限公司成立于 2005 年 1 月，是国内领先的互联网财富管理综合运营商，为超过 1 亿用户提供基于互联网的财经资讯、数据、交易等服务。公司构建以“东方财富网”为核心的互联网财富管理生态圈，聚集了海量用户资源和用户黏性优势，在垂直财经领域始终保持绝对领先地位。2010 年 3 月，东方财富成功登陆创业板，成为 A 股首家上市的互联网公司，截止 2022 年 12 月 31 日公司总市值约为 2653.55 亿元。目前公司已经陆续研发了东方财富金融数据 AI 智能化生产平台、多媒体智能资讯及互动平台及多个人工智能相关项目，并持续优化迭代智能资讯、智能审核、智能风控等产品功能，致力于为用户提供更加便捷、高效、安全的财富管理服务。公司积累了海量的 C 端用户资源与大规模结构性、非结构性的金融数据集，公司的金融数据结构更加多元、全面，在垂直金融领域的模型训练、算法优化方面具备更显著的数据积淀优势。

概　述

东方财富自研大模型是一款公司完全独立自研的金融行业垂类大语言模型，立志于在专业投资顾问服务、深度财经要闻分析、定制化财富管理、内容机会挖掘等金融场景上达到世界先进水平。目前模型训练已进入平稳期，拥有自研 70B 的语言模型，并基于该模型构建一套智能化金融资讯数据总结、分析、创作和问答系统。该系统能够利用大语言模型的生成和总结能力，对海量金融信息数据进行智能化的总结分析，并且可以在此基础上和用户进行多轮智能问答交互。通过该系统，用户可以快速搜索到当前需要的金融数据并进行快速、准确的智能化分析、从而大幅提升金融资讯搜集、阅读和理解的效率，提高决策效率。另一方面，基于对海量产业知识以及金融相关知识的学习，使得大语言模型具备行业分析研究等能力，使得用户可以快速了解当前对公司或行业前景可能产生正负面影响的各类因素。该系统会以多形态提供，包括 WEB 端、客户端等，给用户提供高效、智能的一站式金融数据分析解决方案。

需求分析

一直以来，金融信息的及时性，以及过高的解读专业性，都成为广大投资者的投资痛点，由于信息匮乏或不及时带来的投资损失案例屡见不鲜。

东方财富自研大模型大模型旨在为广大投资者提供全方位的投资顾问服务和财经要闻分析，以及定制化的财富管理和内容机会挖掘。不论是初对投资产生兴趣的零经验投资者，还是具有丰富经验的资深投资者，都可以通过大模型获得专业的投资建议和指导，实现全程无忧的投资陪伴。

作为一款智能助手，大模型还提供自由对话的功能，用户可以随时与大模型进行交流，探讨金融领域的各类话题。无论是关于股票、基金、债券等投资品种的分析，还是关于宏观经济、行业动态、政策解读等方面的讨论，大模型都可以为用户提供及时、准确的信息和见解。

除了投资领域，还在机器翻译、内容生成和创作助手、信息抽取和知识图谱、聊天机器人和社交媒体等场景上进行应用。无论是帮助用户翻译外文资料，还是辅助用户生成优质的内容，都可以提供高效、准确的支持。同时，还可以帮助用户从海量信息中提取有价值的内容，并构建知识图谱，为用户提供更加智能化的服务体验。

案例介绍

大模型通过互动对话助手的产品形态面向投资用户提供信息辅助服务（见图 1-4），在投资者教育方面，模型通过智能化的方式向广大投资者普及证券金融知识，传播理性投资理念。通过提供易于理解和实用的投资指导，大模型可以帮助投资者更好地了解投资市场的基本原理、风险管理策略和投资技巧，提高他们的投资能力和决策水平；帮助投资者更好地保护自己的权益，避免投资风险，促进金融市场的稳定和健康发展；同时，大模型还可以为金融人才的培养提供支持，向在校学生普及金融知识，培养他们的投资意识和理财能力，为金融行业的可持续发展做出贡献。大模型根据用户的需求和风险偏好，提供个性化的投资建议和指导。通过分析用户的投资目标、资金状况和市场情况，大模型可以为投资者量身定制投资组合，帮助他们实现财富增值和风险控制。

图 1

图 2

图 3

图 4

东方财富自研大模型通过智能信息提炼形态面向专业投资机构提供信息辅助服务（见图 5-6），在投资研究方面，利用大语言模型的强大能力，对金融市场进行深入分析和研究。通过对大量的金融数据和信息进行挖掘和分析，可以为投资者提供准确、及时的市场动态和投资机会，帮助他们做出明智的投资决策。

图 5

图 6

效益分析

东方财富自研大模型作为一款垂类金融大模型，将充分发挥其能力，以投资者教育、投资顾问和投资研究为核心，为广大投资者提供全面的服务。

在社会效益方面

有助于引导投资者客观认识市场、理性参与投资，还能更好地维护他们的合法权益，积极践行普惠金融。

在商业模式方面

有助于东方财富 APP, 网站构建更智能化的产品功能，以提升用户的留存率与转化率。同时提升平台信息创作者的体验，以获取更多高质量内容。

在应用推广前景方面

公司将以东方自研大模型作为智能化底座，对各类产品功能，信息服务进行全面升级，以匹配公司的长期智能化战略。

基于大模型的信息结构化抽取方法

上海岩芯数智人工智能科技有限公司

RockAI（岩芯数智）是以认知智能为基础，专注于自然语言理解、人机交互的科技创新型企业，是A 股上市公司（002195.SZ）上海岩山科技股份有限公司的控股子公司，公司秉承“新科技改变生活”的理念，致力于构建自研基础 AI 大模型 + 行业垂直模型的技术结构，实现“1 个 MaaS 平台，多种应用场景”策略，打造客户信赖的认知智能平台。

概　述

信息结构化抽取方法以及企业招投标信息是银行业重点关注的企业信息之一，有效应用于银行对企业的风险评估、商机发现、信贷决策、投资决策等。本方案基于岩芯数智（银行业）大模型的机器阅读理解能力，从非结构化的 HTML 文本中抽取出参与招标的企业信息、项目名称、中标价格、所属省市等信息。通过对历史信息的抽取测试，最终通过大模型的自动结构话抽取，召回率达到 95% 以上，精准率达到 92% 以上，目前已部署到某大型国企银行日常使用，几乎不再需要专员去手动整理相关信息，并直接用于下游业务（如：信贷业务等）。

需求分析

基于大模型的信息结构化抽取方法为银行获取公开的招投标信息提供了巨大助力，目的是为了更好地了解客户的需求和机会，以便提供相关的金融产品和服务（如：企业贷款等），并在风险可控的情况下寻找投资机会。这有助于银行在商业环境中保持竞争力并满足客户的金融需求。例如在某企业中标项目后，银行可根据企业的名称进行风险评估，然后根据中标金额推广贷款业务。目前大部分银行是通过人工整理的方式，定期筛选整理招投标信息给到业务部门，业务效率较低且容易被竞争对手抢先一步，能够及时准确的获取招投标信息成为业务部门的痛点需求。

案例介绍

主要流程

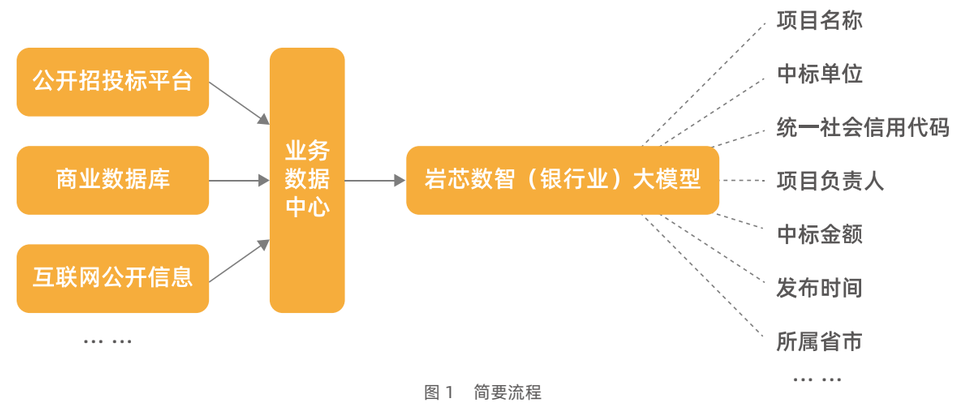


图 1　简要流程

主要能力

- **数据自动化整合：**能够从多个招投标信息来源（如政府招标平台、商业数据库、互联网等公开可见信息）自动整合信息。
- **自动结构化信息抽取：**具备强大的数据理解能力以及辅助的挖掘能力，以提取有用的信息，如项目名称、中标公司、中标价格等。
- **合规性和数据隐私：**确保方案遵循数据保护法规和客户隐私的合规性要求。

技术创新点

1、使用自然语言大模型对非结构化的 HTML 文档阅读，并抽取出结构化的招投标信息。

2、采用数字身份验证技术，以确保客户的身份和业务信息的真实性。

实施效果

1、小时级信息获取效率，在招投标发布 1 小时内，业务部门即可收到相关结构化信息，并可进入下一步业务阶段。

2、召回率达到 95% 以上，精准率达到 92% 以上。

应用落地情况

2023 年 9 月已部署上线到某大型国企银行。

1、银行信贷部门可以使用相关信息更好地评估企业客户的信用风险，从而决定是否批准贷款。

2、可以使用招投标信息来评估并为客户提供投资建议。

3、可以根据客户的特定需求提供更具吸引力的贷款产品，增强客户满意度。

效益分析

经济社会效益

- **市场机会识别：**银行可以使用结构化的信息来识别新的市场机会，开发新的金融产品和服务，从而推动收入增长。
- **提高信贷风险管理：**结构化抽取方法允许银行更准确地评估企业客户的财务状况和业务前景，减少不良贷款的风险。
- **降低操作成本：**自动化的结构化抽取方法可以降低处理招投标信息的操作成本，加速决策过程，降低运营成本。

商业模式

可通过订阅模式、许可费用、定制解决方案、商机销售等多种方式，获得用户付费。

应用推广前景

- **企业金融服务：**用于企业金融服务，包括信贷、财务咨询和风险管理。
- **创新金融产品：**利用招投标信息快速创建创新金融产品，满足企业客户的特定需求，从而推动产品创新和竞争力提升。

天津金城银行金融大模型示范应用

三六零安全科技股份有限公司

三六零安全科技股份有限公司（简称“360”）2005 年成立，中国数字安全领军企业，互联网免费安全服务倡导者。2022 年，360 宣布全面转型数字安全公司，定位“数字安全运营商”，投身产业数字化，以服务为核心，为政府、企业、城市和中小企业的数字化转型保驾护航，探索数字安全时代的中国方案。

成立至今，360 研发投入 200 亿，聚集了 2000 名安全专家，积累了 2000PB 安全大数据，成为了国家网络安全保障核心力量，在二十大、两会、“一带一路”峰会等重大活动中发挥了重要作用。公司用户已覆盖 90% 中央部委、80% 中央企业、95% 大型金融机构和 100% 运营商。未来，360 将以全面拥抱服务化战略，立足“上山下海助小微”战略，以领先的数字安全产品和 SaaS 化安全服务，助力数字中国建设。

概　述

360 智脑是 360 自主研发的语言大模型，具备了生成式对话、多模态指令分发能力，可根据对话意图，选择所需应用和能力进行分发需求，并将收集处理的结果反馈给用户。

天津金城银行金融私有化领域大模型利用 360 成熟的大语言模型、知识库等产品，结合金城银行数字员工、电销、催收、告警等业务，定制开发的企业专有大模型。通过建立天津金城银行内部办公、会议和文档处理的私有化定制大模型，在建立营销、催收、风控等数字员工基础上，进一步打造金融风险控制、保险理赔服务、财务审计、监管、贷款审核和信用评估智能化虚拟分析师，为企业办公、合规文档编写、业务发展提供高质量的智能辅助，提升企业办公及运营效率，助力银行数字化转型。

需求分析

目前，大语言模型存在诸如胡编乱造、无法人工干预产出结果、高度依赖训练内容等短板，直接应用于对客业务存在一定的风险，更适合用于提升企业内部效率和业务生产工作的辅助工具。同时，大语言模型的训练需要非常庞大的训练数据。从应用场景的投入成本与产出角度评估得出，从底层完全训练自主的大语言模型不是客户当前的诉求。

通过内部需求调研，除了大语言模型的基础对话能力之外，客户需求主要集中在业务应用、办公提效和系统自动化三大场景。企业知识库、电销及客服助手都需要对客户现有的业务文档和相关材料进行一定训练微调，其训练的内容涉及客户隐私和商业机密，对数据安全要求极高，无法在内网之外的环境进行。

案例介绍

基于 360 自主研发的语言大模型，根据客户实际业务场景和系统自动化需求，360 为其提供针对性的解决方案：

标准化对话模型

根据客户的“知识库”内容进行微调训练，能够提高客户内部办公、会议和文档处理能力的私有化定制大模型。同时，标准对话模型和企业知识库能满足企业部署大型生成式语言模型时，对“内容可信、数据安全、成本可控”的需求。具体方案：

- **企业知识库：**对客户的基础知识库内容进行提问，并基于客户的基本信息进行回答内容生成。通过频道推荐、智能问答导航等功能模块，实现了企业内部知识的高效存储、检索和共享，帮助员工快速找到所需的知识信息。

- **语音识别：**识别音频、视频中语音特征，区分说话对象，支持语音转文本，提供语音转写功能，默认包含中文普通话的转写，配合敏感文本检测使用。

定制化产品

根据客户的一些实际业务场景和系统自动化需求，定制能够支持基于大语言模型实现智能报警、ABS 资金自动分案，同时能够辅助客户、电销以及催收等员工在不同业务场景的对话。具体包括：

- **电销客户意向判断系统：**帮助电销人员更准确地判断客户的借贷意向，并提供针对性的推销策略和服务方案，降低电销成本，提高客服的工作效率。

- **电销客服辅助系统：**电销客服辅助系统旨在帮助电销客服人员更高效地管理和服务客户，提高销售质量和效率，系统对电话销售过程中的语音内容进行深入分析和处理，从而提供精准的话术指导和建议。
- **催收分案系统：**快速识别和分析不同类型的逾期债务，结合现有的催收规则和标准，将逾期未还款的案件合理分配给催收人员和团队，自动匹配最适合的催收策略和方式，提高催收效率和成功率。
- **系统告警提示系统：**快速识别和分析当前的告警信息，并结合告警知识库和相关规则，快速定位问题并推荐解决方案，减少因故障导致的业务中断和影响，提高告警处理的及时性和准确性。

图 1　系统架构图

效益分析

360 与天津金城银行共建的天津金城银行金融大模型标志着人工智能技术在 AI+ 金融领域的创新应用，加强了银行的数字化能力、提升了用户体验，同时为各行各业的大模型创新应用提供了新的创新可能性。

通过工程优化，可以降低大模型应用的成本，提高应用的便捷性和效率。在规模适配方面，360 GPT 大语言模型可以根据应用场景进行灵活的规模适配，在保证高性能的同时减少模型的大小和计算资源消耗；在训练微调方面，该模型可以进行精细的优化和调整，以适应不同的应用场景和任务要求；在轻量优化方面，360 GPT 大语言模型可以进行轻量化设计和优化，降低模型的运行成本和计算资源占用率。

文修大模型助力中文校对提质增效

上海蜜度科技股份有限公司

蜜度创立于 2009 年，是一家以人工智能技术为核心的语言智能科技企业，专注于多模态、多语言智能科技，通过 AI 技术与智能应用赋能千行百业实现数字化、智能化转型升级。

蜜度基于自主研发的蜜巢、文修大语言模型，利用先进的多语言校对（MLC）、自然语言处理（NLP）、计算机视觉（CV）、跨模态检索（CMR）、内容生成（AIGC）、知识图谱（KG）等人工智能技术，提供智能检索、智能校对、智能生成等三大核心应用，致力于为政府、媒体和企业客户提供智能、安全、高效的应用解决方案。

概　述

蜜度自研的国内首个智能校对领域大语言模型——文修，已经成功在 AI 智能校对产品——蜜度校对通中落地，运用大语言模型能力赋能各行业办公场景。搭载了文修的蜜度校对通在保障响应速度的基础上，从文字标点差错校对、知识性差错校对与内容导向风险识别的三大类型中，整合化地提供 27 类细分方向的审校服务；并能够在尊重稿件原意的前提下，修正用词不当、句式杂糅等问题，让句子表达更流畅，实现对句子的润色功能。为新闻出版、媒体稿件、政务公文等专业领域带来工作模式迭代与效率提升，为新时代语言文字工作高质量发展注智赋能。

需求分析

我国高度重视汉字的规范性，相继颁布了国家通用语言文字法等法律法规和规则规章，汉字规范标准化建设扎实推进。

校对任务一直是政府机关、学校、新闻出版、公共服务等领域的重要工作之一。伴随自然语言处理技术的进步，智能校对工具现在能够更准确地识别错别字，并为其提供正确的替代选项，同时也能够更好地理解文本内容的语境。而对于更为复杂的语义理解，如指代不明、成分缺失、表述不当、逻辑不清等常见句式杂糅问题，小模型很难解决。

因此，在大语言模型时代，一款优质的智能校对软件，能够从语言文字的校对层面，为语言文字的规范化建设助力，从而有利于传承弘扬以语言文字为载体的中华优秀文化。国内首个智能校对领域大语言模型“文修”应运而生。

案例介绍

文修以大语言模型（LLM）为技术底座，通过运用高质量数据学习多种特色子任务，大幅提升中文校对和润色能力的智能化程度，辅助专业用户提高校对质量、提升校对速度、降低差错率，为新时代语言文字工作高质量发展赋能。

搭载了文修的蜜度校对通拥有智能校对和 AI 润色两大核心功能，其关联的校对能力，不仅能够校对错别字，还能够实现词义辨析、数字常识等编辑加工中的易混错误。同时，针对特定行业的垂直领域专属需求，文修能够充分发挥大模型的快速训练优势，通过集成学习技术，快速部署专业领域内的特有校对能力，为不同领域企事业单位搭建专属校对大模型。

技术创新点

特有的面向校对的多阶段、多任务训练与数据增强方法。

- **文修大模型引入了多任务学习策略，让校对模型同时学习多种有关联的任务，以提高其整体能力。**
- **模型采用了多阶段训练策略，增强校对数据的数量和质量。**
- **在词汇表上，精简了词汇表的非中文词汇来提高模型的推理速度，同时对通用规范汉字实现了全覆盖。**

实施效果及应用情况

文修在中文拼写勘误、语法纠正任务上的表现显著优于通用大模型 ChatGPT，大约有 20% ~ 30% 的效果提升。

在基础错别字校对基础上，文修着重解决现有校对模型对易混词细微语义的辨析能力较弱的难题。

图 1　智能校对功能

在尊重作者的原始表达意图基础上，更好地修正用词不当、句式杂糅等问题，使句子表达更流畅，实现对句子的润色功能。在具体表现上，当句子出现语序不当等句式杂糅的问题，“文修”可以通过调节语法结构或纠正客观事理等方式进行修正，从而提高句子的逻辑性。

图 2　AI 润色功能

图 3　“文修”解决句式杂糅

文修落地蜜度校对通，能够广泛应用于新闻出版、政务公文等多个垂直领域：在新闻出版中，辅助审校人员进行稿件发布前的校对审核，弥补新媒体内容语言不规范、语感不足等缺陷；在图书出版环节，则可以对大样文件和 PDF 文件进行审校；在政务办公中，将其部署在专网的服务器上，可在安全环境下提高公文的规范性和准确性，提高公文写作质量。

在产品易用性方面，蜜度校对通具备多种产品应用形态，支持编辑、校对全链路工作流程，可以满足不同的使用场景，私有化部署保障了隐私安全。

图 4　产品多类形态

效益分析

搭载了文修大模型的蜜度校对通，将原有能力进行了全线升级，不仅为政务公文、新闻出版等专业领域带来工作模式迭代与效率提升。同时，智能校对技术在语言文字相关领域的应用，帮助用户提升语言文字的规范性、安全性、准确性及严谨性，以大模型能力深度赋能各行各业的办公场景，为新时代语言文字工作高质量发展注智赋能。

以出版行业的应用为例，一本 20 万字的书稿，蜜度校对通仅用 90 秒就可以完成覆盖错词病句、常识错误与不规范表述等层面的审核与校对，并给出修正与润色建议。

此前，蜜度校对通已经服务了多家出版单位和政务部门，这次将大模型能力落地在 AI 产品的突破，能够帮助专业领域用户提质增效。

新型金融风险防范可信金融大模型

上海氪信信息技术有限公司

氪信科技，引领人工智能新潮流，主打产品——氪信领航以独特的可信 AI 技术为核心，已经成功成为众多金融巨头解决数字化挑战的得力助手。

作为人工智能在金融行业的领军者与倡导者，氪信科技已经成功服务过中国资产规模前 15 的国有及股份银行中超过 9 成的客户。我们以 AI 大模型技术为武器，解决客户在数字化转型中遇到的难题，我们以独特的可信 AI 技术为引擎，把金融专业领域知识与通用大模型融合，赋能金融业务。氪信科技，用科技创新推动金融行业的进步。

公司的技术能力也获得了来自学术界和政府的认可。2020 年，公司获评学术届顶级奖项——中国计算机学会（CCF）“2020 CCF 科学技术奖科技进步杰出奖”和吴文俊人工智能科技进步奖。并且作为第一作者，和浙江大学、交通大学、今日头条、美团等高校和科技公司在《自然 - 机器智能》子刊上发表《中国新一代人工智能》的论文。公司也是上海市高新技术企业。

概　述

氪信领航防范新型金融风险的可信金融大模型产品的出现，为防范电信诈骗提供了一种新的解决方案。通过精准防控和提前预测，这款产品有助于金融机构构建完善融风险防控体系，助力金融机构更好地保障金融数字账户的安全。

随着技术的不断进步和应用场景的不断扩展，我们有理由相信，氪信领航产品在防范电信诈骗、保障金融账户安全方面具有很大的潜力和价值。将在未来发挥更加重要的作用，为人们带来更加安全、可靠的金融服务。

需求分析

近年来，随着信息社会快速发展，犯罪结构发生了重大变化，以电信网络诈骗为代表的新型网络犯罪占比上升至我国刑事犯罪的八成以上，成为全社会面临的严峻挑战。为了预防、遏制和惩治电信网络诈骗活动，加强反电信网络诈骗工作，保护公民和组织的合法权益，维护社会稳定和国家安全，《中华人民共和国反电信网络诈骗法》于 2022 年 12 月 1 日起施行。

反诈法的实施，要求银行业金融机构加强对电信网络诈骗的防范和打击，保护客户的合法权益，维护金融市场的稳定和安全。银行业金融机构在完善金融风控体系，保障金融账户安全，防范和打击电信网络诈骗的工作中仅支持事后的账户安全防控电诈攻击链路的资金端，无法做到事前识别，提前防控金融风险账户。

案例介绍

产品核心包含先进的大模型技术，通过氪信自研改良的机器学习算法、时序检测算法及深度学习算法，结合十年以上反欺诈、反洗钱、反电诈专家业务团队的知识输入，充分挖掘涉案账户涉案前的隐匿风险信号，由点到面，由个体到群体，针对不同风险特点选择不同的模型优化框架，多维度覆盖各类风险模式。同时，通过氪信科技自研的统一超参数优化框架，对大量模型进行跨模型的参数优化，形成统一电诈风险识别结果，实现事前预测预警。

氪信针对电诈场景，形成自研算法框架，搭建灵活的局部社区挖掘方法，支持无监督、半监督的风险扩散；对传统知识图谱挖掘的进行了大幅度调整，缩短数据时间窗口，LCD 网络构建，实现动态的图应用策略，考虑不同时间切片下关联风险的重叠性及差异性，更符合电诈场景风险多变的现状。氪信自研算法融合了谷歌研发的 PPR 算法，并在其基础之上融合了深度游走算法保证社区分割的紧密性，添加 sweep-cut 算法进行局部社区分割，使得算法减少全局搜索的计算，加快计算效率，降低计算噪音，保证结果的精准性；同时，氪信自研的 Ranker 模型核心目标在于整合不同角度的风险结果，通过超参框架对模型进行调优，替换原有的单点模型调优，用超参框架对各风险模型的结果进行统一整合，形成最终评分。

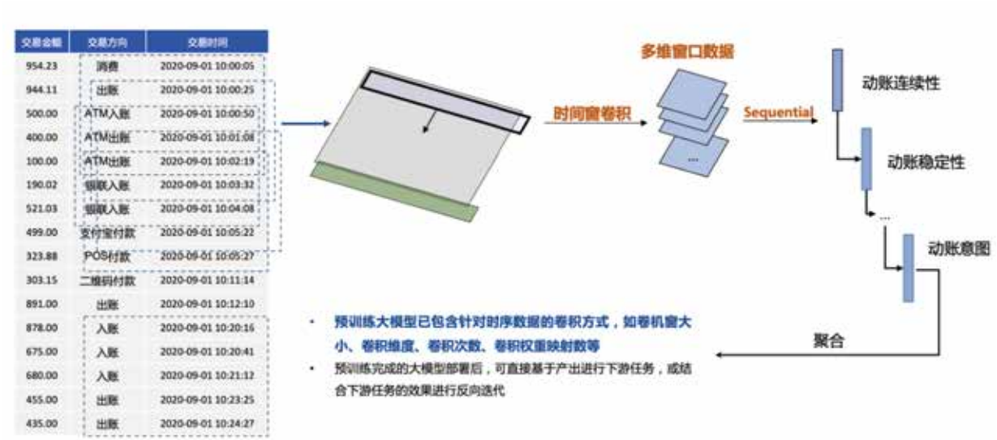


图 1　预训练大模型算子示例

其主要优势及亮点为：

（1）提前预测电诈账户，且可自学习进化

区别于事后风险总结型模型，氪信产品是事前预警的预训练大模型。使用银行自身的账户数据，交易数据，以及公安下发的涉案黑名单数据，通过人工智能技术精确记忆与计算，使得产品可在账户发生诈骗风险前即实现快速识别及预警；

同时通过每日学习国内金融市场上真实电信诈骗领域知识，进行训练、复现和迭代，从“总结风险”升级到“推演风险”，利用产品内核的优化框架进行在线自迭代，从个体视角上升到总体视角，提前捕捉潜在的风险模式变化，实现与作案手法的对抗过程。

（2）精准提示，分层分级管控

氪信领航产品持续对行内全量账户进行扫描，每日对全量账户进行风险识别，结合各账户历史风险情况、行为特点，挖掘账户动态风险变化，对具有一定风险的账户分层识别；

其预警结果中，既包含精准预测的极高涉诈风险账户（约每日 50 户左右）提供给行方进行止付或暂停非柜面管制，同时提供具有潜在涉诈卡池风险的休眠账户、疑似涉诈工具卡（涉嫌养卡行为及参与涉诈资金转移行为）的中风险预警名单，行方依据名单进行限额管控。

效益分析

氪信领航防范新型金融风险的可信金融大模型产品在上海多家头部大行分行成功应用，推向全国；当前已在全国 25 个省的银行省分行投产使用，均有极好的涉案账户提前预警、压降识别的效果。某国有大行 8 月 20+ 分行部署氪信领航产品后，行方基于产品提前识别的精准风险名单进行风险分级管控后，9 月涉案账户量较 8 月降低 50%。

通过精准防控和提前预测，这款产品有助于金融机构构建完善金融风险防控体系，助力金融机构更好地保障金融数字账户的安全，为人们带来更加安全、可靠的金融服务。

信阳市智慧工业平台

云知声（信阳）数字科技有限公司

云知声（信阳）数字科技有限公司（以下简称“信阳云知声”）是云知声智能科技股份有限公司（以下简称“云知声”）的全资子公司，位于河南省豫东南高新区，定位为云知声中原总部基地，注册资本 1 亿元。信阳云知声依托云知声在人工智能、数字经济、智能制造和智慧城市等方面的专业优势，以“山海”大模型为基础，以智能语音交互、知识图谱等全栈 AI 技术为核心，打造云服务和 AI 芯片，并基于云芯一体化平台，向智慧物联、智慧医疗等广泛领域，提供基于大模型服务的人工智能产品与综合解决方案。

概　述

云知声自主研发的“山海”大模型，随着大模型技术不断迭代升级，模型能力不断加强，大模型建设重心也从基础能力建设逐渐向应用能力建设转移，目前大模型服务已逐渐落地在智慧政务、智慧车载、智慧轨交、智慧医疗等场景。

本次应用案例是围绕信阳市政府对于智慧工业平台的需求，针对信阳市 6 大产业，为信阳市打造一个专属智慧工业平台，赋能提升信阳政务水平，为政府工作人员以及相关企业等提供更加专业、高效、个性化的智能化服务。

首先，利用“山海”大模型的知识增强技术，建立信阳市政企政策知识服务，为企业提供国家、省级多层级政策的汇总展示和拆解，做到政策的精准推荐和推送。其次，基于大模型的 NL2SQL 技术和自然语言对话能力，建立企业信息全景驾驶舱，形成工业经济及企业画像，实现工业企业运行指标监测、运行态势分析以及县区横向对标等。然后，利用大模型的深度学习能力，可实现大模型在行业应用场景的知识微调，通过模型的本地化部署和模型优化，支持企业基本信息、运行信息查询，数据横向对比、图标、报告自动生成。同时基于平台生成的预警信息，能够利用大模型能力即时响应，并生成备选应对处置说明及方案，有效提升相关部门制定处置方案效率，为解决预警问题等提供有力支撑。

此外，云知声的大模型能力涵盖了云知声自研的 Atlas 智算能力，可实现 30 亿亿次每秒的计算能力，可实现大规模、海量数据的异构并行计算，是大模型进行数据分析和模型训练推理的基础和有力支撑。今后，云知声将进一步推动大模型应用成果在不同场景落地，促进更多的产业数字化潜力持续释放。

需求分析

随着新一代人工智能技术蓬勃发展，信阳市加快实施产业数字化转型工程，积极培优壮大数字经济，加快建设数字信阳，充分发挥数字经济对经济高质量发展的引领和支撑作用。在开展数字化转型过程中，信阳市目前亟需解决对行业企业及产业运行的监测分析能力的快速提升，在监测分析的基础上，实现对企业配套政策的精准高效，促进产业经济快速发展，盘活域内企业资源。

但是传统的技术手段面对海量的工业经济相关数据，很难实现由点到面的全面观测，所以需要利用人工智能和大数据分析等技术，搭建一个更加智能化的智慧工业平台，以解决企业基础数据、企业运行数据、产业数据等多维度展示和对比，快速、高效地生成制定政策需要的数据支持。

案例介绍

基于大模型技术打造的信阳市智慧工业平台（如下图 1），是云知声“山海”大模型在政务领域落地应用的一个重要体现。基于大模型能力，构建平台 3 大功能模块：面向政府端，打造企业运行监测模块；面向企业群众端，打造智慧政策平台和智慧金融平台，形成政金企对接机制。

图 1　信阳市智慧工业平台

在政务服务上，建立企业数据库和管理平台，汇聚企业基本信息、生成经营数据以及产业链信息等，为信阳市政府决策提供强大的数据支撑服务（如下图 2）。以此为基石，打造企业运行监测模块。通过构建指标模型，针对企业成长、运营、盈利以及偿债等能力，通过反馈机制将优秀指标、风险指标推送相关人员或部门，并基于大模型的语言生成、大数据分析等能力生成备选应对处置说明，让责任部门快速、及时采取措施，确保企业即时获得支持。

图 2　企业运行监测模块

另一方面基于山海大模型的自然语言理解、知识问答以及逻辑推理等能力，可实现企业基本信息、运行信息查询，数据横向对比、图表、报告自动生成，并且能够为企业提供目标政策查询、推荐等功能（如下图 3）。

图 3　目标政策查询及推荐

在助企服务上，打造智慧政策模块和智慧金融模块。

其中智慧政策模块，搭建一个智慧政策服务平台，建设政策条件库和市场主体画像库两大数据库（如下图 4）。可梳理政策清单，实现国家、省市多层级政策的汇总展示、拆解，海量数据中目标知识挖掘，对企业信息和政策要素进行快速精准匹配，实现“政策找企、应享尽享、免申即享”，形成“1+2+3+4+N”的“免申即享”模式。在税费减免、财政补贴、融资支持、招商优惠等提升政策覆盖度和精准性。为深化“放管服”改革、优化营商环境提供有力支撑，增加信阳市企业经济发展活力。

图 4　市场主体画像库和政策库

智慧金融模块，搭建线上融资服务平台，具备金融产品、企业融资诉求等上线、展示及搜索功能（如下图 5）。同时基于企业信用模型，在平台可查看已授权企业信息，打通政府、企业、金融机构等多角色的信息壁垒，让企业有融资渠道，让银行有放款动力，让政府平台公司敢于担保，形成融资闭环的良性循环。不仅助力企业更好更快融资，降低企业融资成本，也降低了金融机构、担保公司的债务被违约风险。

图 5　金融产品展示及搜索

大模型在信阳市智慧工业平台的落地应用，将为信阳市政府的数字化履职提供更多支撑和帮助，通过大模型能力外溢，支持不同阶段企业的落地应用，促进信阳市产业链数字化转型，助力数字信阳的高质量发展。

效益分析

利用山海大模型服务，可实现数字化城市建设的多场景赋能，助力千行百业智慧化升级。云知声山海大模型是一款千亿参数级的 AI 大语言模型，具备语言生成、语言理解、知识问答、逻辑推理、代码能力、数学能力、安全合规等七项通用能力，以及插件扩展、领域增强、企业定制三项行业落地能力，能够满足不同场景的不同需求（如下图 6）。

从写一篇会议纪要，到秒级生成门诊病历，从智能解答政务相关高频问题，到精准选商挖掘潜力的重点企业，从智能办公、交通服务和调度指挥，到教育学习辅导、消防安全防控巡查等城市建设的方方面面，大模型的落地应用，将会在城市服务的数字化、智能化进程等方面，从效率、成本、体验等多角度，助力千行百业的降本增效和数字化转型。

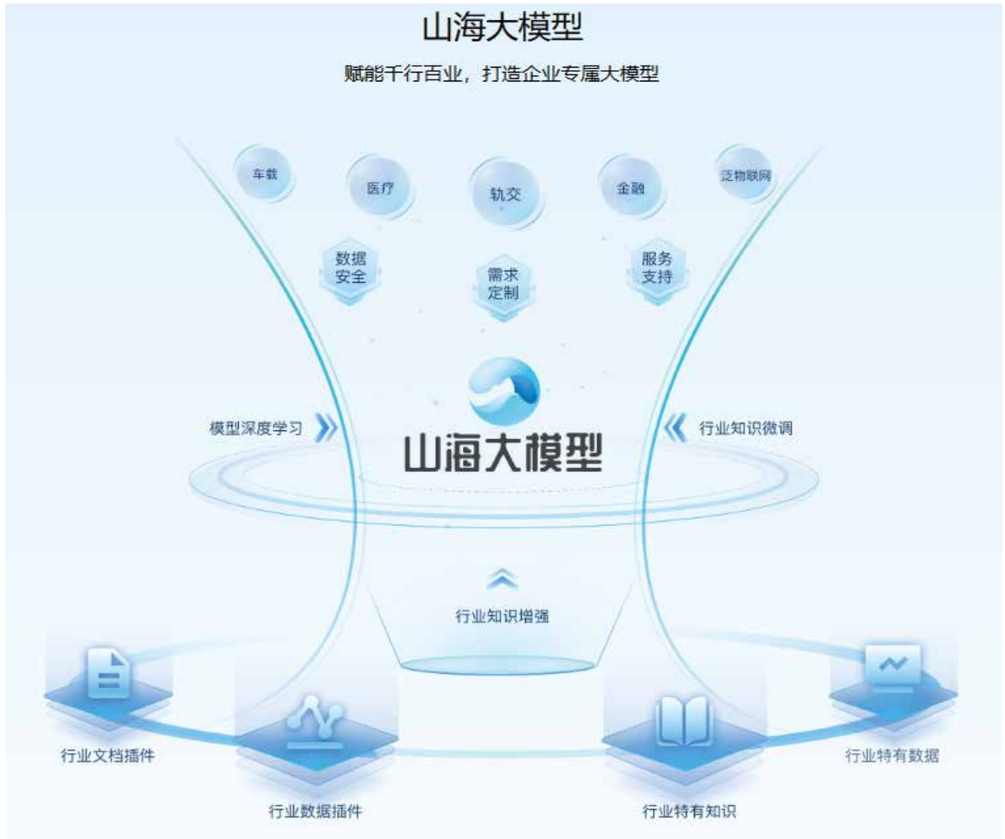


图 6　云知声“山海”大模型

遥感大模型在农业信贷场景的应用

蚂蚁科技集团股份有限公司

蚂蚁集团起步于 2004 年诞生的支付宝，经过十八年的发展，已成为世界领先的互联网开放平台。蚂蚁集团通过科技创新，助力合作伙伴，为消费者和小微企业，提供普惠便捷的数字生活及数字金融服务；持续开放产品与技术，助力企业的数字化升级与协作；在全球广泛合作，服务当地商家和消费者实现“全球收”、“全球付”、“全球汇”。作为一家技术人员占比超过 60%，拥有强大自主创新能力的科技企业，蚂蚁集团始终坚持自主创新，在人工智能、数据库、隐私计算、智能风控、区块链等领域进行了前瞻性布局，自主研发了大模型、隐语、OceanBase 数据库等一系列支撑蚂蚁和行业发展的关键技术。

概　述

在农村数字普惠金融和农业农村现代化发展的双重政策驱动下，蚂蚁集团与网商银行于 2019 年发起“亿亩田”项目，通过遥感智能解译技术建立了对大范围种植作物、农业设施的低成本高精度识别能力，补齐农业场景“人 - 地 - 物”关联数据缺失的短板，实现对农业用户年度经营状况的反演和用户授信，衍生出农业信贷新模式，以普惠金融助力农业农村发展。在核心技术方面，蚂蚁集团研发了行业内首个全面支持多模态、多分辨率、多光谱时序影像的遥感多模态大模型，覆盖解译任务最全，参数规模领先，各项指标均为行业顶尖。基于该技术，项目已覆盖全国 31 个省、自治区、直辖市和 15 大产业，帮助 150 多万种植户获得无接触贷款，为乡村振兴注入源源不断的金融“活水”。

需求分析

2019 年数字农业农村发展规划（2019-2025）提出以数字化驱动农业农村现代化发展，2020 年中央 1 号文件明确提出着力发展农村数字普惠金融。“三农”一直以来为国之根本，是治国理政的头等大事。2022 年全国粮食播种面积 17.7 亿亩，仅 2022 年上半年，全国涉农整体信贷余额规模就高达 47.1 万亿，用户需求旺盛。

然而受限于农业农村场景长期信息化、数字化程度低，大多数农业用户信用数据单薄，信贷准入十分困难，其中最核心的问题在于如何对农业生产经营状况进行有效建模。蚂蚁集团与网商银行携手共建“亿亩田”项目，通过遥感智能解译技术实现耕地种植情况客观观测，有效建模农业生产经营状态，建立了农业信贷新模式，以金融活水精准浇灌呵护农业生产，助力乡村振兴。

案例介绍

农业遥感场景天然具有无标注数据充足（全球日均产生 PB 级卫星观测数据）、标注数据极度稀缺的特性。尤其在经济作物领域，其种类丰富、分布稀疏、种植模式多样。本项目通过海量全地貌遥感影像数据进行大模型无监督预训练，在无需标注数据的情况下让模型学习到强泛化性的地物表征，自研完成行业领先的大规模遥感多模态大模型，天观（SkySense）。天观的核心能力包括：

1. 行业内首个同时支持多模态、多分辨率、多光谱、时序影像输入的遥感大模型，覆盖最全面的地球观测解译任务。

2. 参数规模达到 20 亿，为行业内参数规模最大的遥感多模态大模型之一，天观目前在 7 种任务总计 16 个公开数据集中都取得了最优结果（包括小目标旋转检测、多模态农作物识别等），达到全球顶尖水平，如图 1 所示。

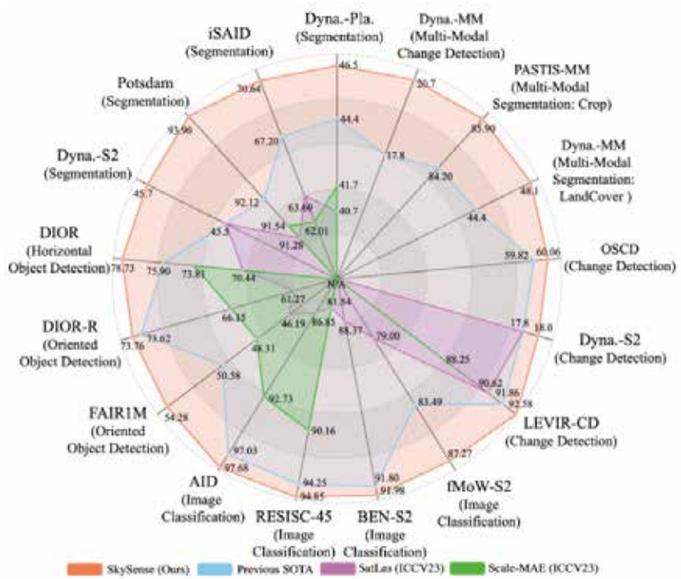


图 1

天观依托时空解耦架构、多粒度对比学习、 地学时空知识建模等多项核心技术，突破了遥感农业识别地形复杂、泛化性差的难题。

1. 时空解耦架构，如图 2 所示。 基于遥感场景中，多模态、多时序遥感影像空间对齐特性， 提出时空解耦架构，创造性地对空间特征提取及多模态时序融合两个关键模块进行独立拆分。该设计大幅减少时空建模的参数量，并有效提升遥感场景中的智能解译精度 。

2. 多粒度对比学习 如图 3 所示。基于遥感解译任务在模态、空间尺度的多样性， 在两个维度实现了多粒度对比学习。模态层面支持单图单模态与序列多模态对比学习空间层面，设计了像素级、目标级与整图级对比学习 。上述预训练有效提升天观在全场景的配适性 。

3. 地学时空知识建模。基于地理位置对大尺度空间范围内遥感影像特征进行无监督时空聚类， 生成区域性的地物通用时空表征，有效提升模型地球观测解译的能力 。

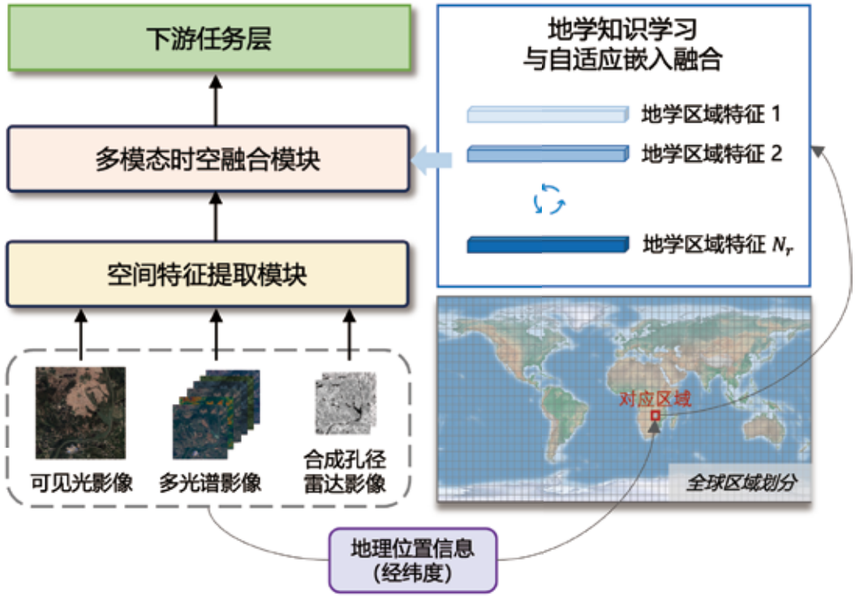


图 2

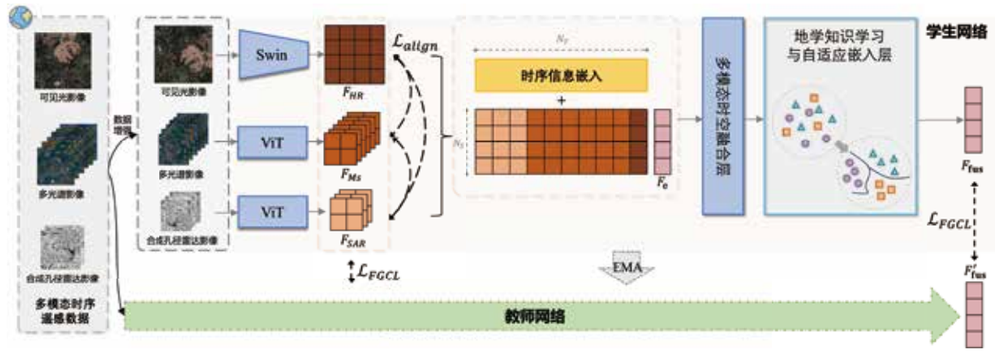


图 3

本项目在全球范围内首创将卫星遥感农作物识别运用到涉农信贷业务中，已覆 3 种主粮作物、13 种经济作物、2 类农业设施。在核心作物识别中，天观对比基线模型 在 90% 精度下召回率取得显著提升其中大棚召回率增长 12%，果园提升 24%，苹果提升 15%，柑橘提升 14%。

效益分析

项目利用遥感智能解译技术对作物种植情况进行贷前观测、贷中监控、贷后管理实现对农业用户生产经营状况的高效准确评估与授信 。项目通过网商银行已覆盖全国 31 个省、自治区、直辖市和 15 大产业，帮助 150 多万种植户获得无接触贷效益分析款，并大幅降低单笔贷款的发放成本，取得了良好的社会经济效益。同时依托遥感大模型对小样本、复杂地貌的优异泛化能力，在高价值经济作物场景项目拥有覆盖品类广、识别精度高的优势大幅拓展了主粮之外广阔的市场空间，具备进一步向行业推广复制的潜力。此外，基于遥感大模型天观，项目也积极探索落地了 ESG 相关工作包括种植林监测 、遥感碳汇计算等。

中国金茂人工智能大模型企业内部场景应用

中国金茂控股集团有限公司

中国金茂控股集团有限公司（简称中国金茂）是世界五百强企业之一中国中化控股有限责任公司有限公司旗下城市运营领域的平台企业。秉承母公司中国中化“科学至上”的核心价值理念，中国金茂以“释放城市未来生命力”为己任，始终坚持高端定位和精品路线，在以品质领先为核心的“双轮两翼”战略基础上，聚焦“两驱动、两升级”的城市运营模式，致力于成为中国领先的城市运营商。

概　述

本案例以人工智能技术赋能企业内部场景为主题，以提升企业内部办公效率和质量为目标。本案例表明，传统企业在企业场景应用人工智能大模型技术，可以达到改进企业内部办公方式，提高效率和质量的目的。

需求分析

中国金茂作为一家地产企业，始终秉承科技驱动发展的理念，致力于通过数字化建设来更好地服务客户以及提升企业效率。近年来，中国金茂信息技术中心组织人员成立人工智能大模型技术专项研究小组，以金茂数字科技赋能及业务拓展为研究方向，致力于探索人工智能大模型的潜在应用场景与商业价值。研究小组运用人工智能大模型技术，在企业内部实现了对若干数字化场景的升级与拓展。

案例介绍

主要能力

中国金茂人工智能大模型技术专项研究小组惜助多元化的技术手段并结合多维度的实际场景，基于多次实践与应用的经验，在以下四个主要的人工智能模型场景中，能显著提升员工的工作效率。

1、企业办公助手：基于人工智能大模型的推理能力，训练其理解并应用企业专有的知识，进而支持员工以对话的方式来便捷地解决工作中遇到的问题，辅助构思工作方案，提供高效的文章配图及产品效果图，有效提升工作效率及质量。

图 1

2、文档知识问答：针对业务人员，提供融合企业知识和行业知识的大模型对话服务。通过文档知识库的挂载和依据事实的机制，辅以人工智能大模型的文本提炼和总结能力，我们能够快速检索国家标准、行业标准、企业内部规章制度、专题相关知识、系统操作手册等所有文档知识。

图 2

3、企业数据问答：以对话的便捷方式，满足业务上灵活多变的数据查询需求，从而缩短用户与数据之间的距离。通过训练人工智能大模型调用企业内部系统数据的能力，我们既可以满足业务复杂、灵活且多变的业务数据查询需求，同时也确保企业内部数据不会向外网透露。

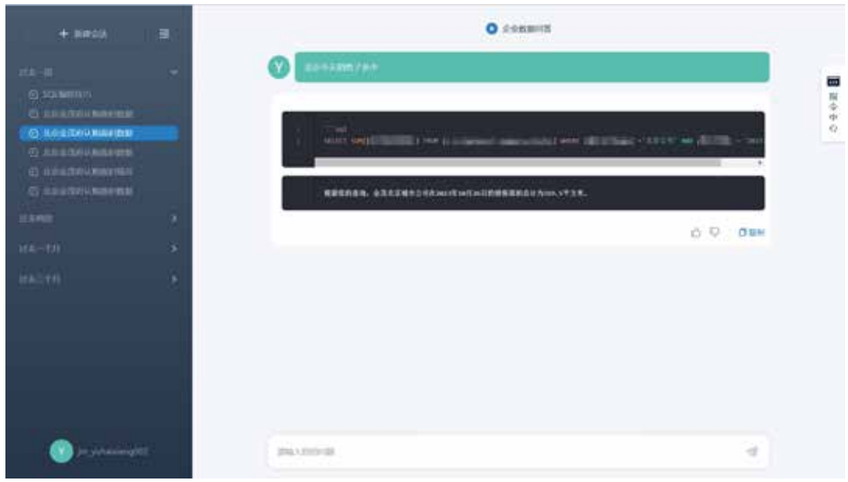


图 3

4、工单自动分类：训练历史上各种不同类别的工单，使得人工智能大模型具备了自动判定工单类别的能力。这种方式将过往的经验赋予每一位员工，让每名工单处理人员都能快速定位问题、分析问题、提供有效的解决方案，从而提高员工的整体工作效率和质量。

图 4

图 5

技术创新点

本案例将人工智能大模型技术和多种技术相结合，应用于企业内部场景，以适当的投入推动了企业数字化建设的升级；在此基础上发现若干新场景，可有效推广应用到不同业务方向。

应用成效

从数字化建设角度，人工智能大模型技术可以加快企业数字化建设的步伐；从企业员工应用角度，整体提高员工的办公效果和成果质量；从技术创新角度，不断突破企业自身的技术限制，提升企业的技术水平。

示范作用

中国金茂作为一家地产龙头企业，成功运用人工智能大模型技术进一步推动了企业数字化建设。我们的成功实践为其它传统企业提供了一个可借鉴和参考的范例，将对整个行业产生积极的示范效应，促使更多企业关注和应用人工智能大模型技术，从而推动整个行业的技术进步和发展。

效益分析

从经济效益层面，人工智能相关技术可推动企业提质增效，本案例中每一个创新应用场景每天为每名员工平均节约 0.5 小时的工作时间，人效的提升使得更多注意力被集中于关键工作，从而提升企业整体水平和竞争力。

从社会效益层面，通过人工智能技术辅助决策和风险识别，可以有效减少人为失误，提高企业的整体运营水平和安全水平，从而对企业的社会形象和声誉产生积极影响。

中山大学附属医院智慧医院项目

云从科技集团股份有限公司

云从科技成立于 2015 年，孵化自中科院，是第一家在科创板成功上市的人工智能平台公司，股票代码为 688327。

云从致力于为客户提供高效人机协同操作系统和人工智能解决方案，助推人工智能产业化进程和各行业智慧化转型升级，一方面凭借着自主研发的人工智能核心技术打造了人机协同操作系统，通过对业务数据、硬件设备和软件应用的全面连接，把握人工智能生态的核心入口，为客户提供信息化、数字化和智能化的人工智能服务；另一方面，云从基于人机协同操作系统，赋能金融、治理、出行、商业、工业等各种应用场景，为更广泛的客户群体提供以人工智能技术为核心的行业解决方案。

概　述

中山大学附属第一（南沙）医院位于粤港澳大湾区南沙横沥岛尖西侧，总建筑面积约 50.6 万平方米，于 2023 年 3 月 29 日正式启用。近年来用信息技术提高医疗资源的管理水平成为行业建设的热点，以此为契机，智慧医院应运而生。基于云从从容大模型，该项目是全球第一家从 0 到 1 由人工智能公司主导建设的智能化医院，自规划开始，就融入智能化理念与相应规划配套。云从助力其构建智能 + 数字医疗生态新图景，主要围绕面向医务人员的“智慧医疗”、面向患者的“智慧服务”和面向管理者的“智慧管理”三大领域进行可持续发展建设。

需求分析

本项目的建设是适应现代医疗卫生事业的发展要求、加快广州市、南沙区医疗资源布局调整步伐、优化医疗资源配置的需要。项目建设符合南沙区当前的医疗资源现状需求，对南沙的医疗卫生事业发展具有重要的战略意义和现实必要性。对打造南沙新区粤港澳大湾区医疗科研新高地、国际医疗中心具有重要的意义。云从科技将为中山大学附属第一（南沙）医院构建高速信息传输通道和先进信息基础设施，适应医院不同领域的信息应用和未来发展需求，初步形成功能完善的信息化基础设施体系，建设融高效、安全、节能、管理为一体的智慧数字化医院。

案例介绍

基于云从从容大模型，整体项目板块由“1+N”，建设信息化基础设施系统和信息安全系统为基础，由 1 个智能化集成平台实现数据汇聚与共享，为将来智慧医院建设奠定智能化数据基础支撑。建设面向医院的医疗、管理建设等 N 项（本期建设智慧管理、智慧护理、智慧服务、智慧医疗）全面的智能化支撑系统。

1：信息集成平台

以物联网、大数据、人工智能等新型智慧化技术为基础， 通过云从科技从容大模型成功构建医院数据集成“大脑”，全方位支持医院智慧化运行，构建智慧医院数字孪生场景、通过三维模型还原，可视化渲染支撑能力，实现医院的地上地下一体化、室内室外一体化呈现、动态静态一体化呈现提供服务，构建动态精准、实时映射的数字孪生智慧医院。包含：综合态势、能耗管理、安防应急、楼宇智能、信息网络、医疗辅助、医疗后勤、人工智能、资产管理、设备设施、空间管理。

本项目通过接入弱电智能化，医疗辅助系统，后勤服务管理各业务系统的数据，将能效系统运行数据和控制能力打通，实现智能化系统联动，为院区节省了 20% 以上能源；节省了约 10% 人力工作量；让管理人员更便捷、更灵活地定位问题、识别风险、进行根因分析；实现智能化系统全量业务进行集中运营和管理，实现业务全路径管理。

通过数字孪生模型高度融合医院各领域现有数据资源，对医院安防、基础设施、停车场、医疗辅助，医疗后勤、环境空间等管理领域的关键信息进行综合分析，结合人工智能手段，对医院重点人群、重点车辆区域安防预警布控，非法侵入院区告警，车辆违停识别预警，医院客流分析，人员聚集热力分析等人工智能手段进行辅助，管理者全面掌控医院运行态势，实现医院人、事、物统一管理，医院综合安全运营态势一屏掌握。

N：楼宇智能化板块

系统对楼宇智能化系统的集成主要包括：冷热源系统、空调通风系统、给排水系统、风机盘系统等，对这些系统进行实时监测、数据记录、故障报警，并提供报警与故障类型。

设备运行情况数据看板可在页面集中展示各个系统设备当前的运行、停止、故障、告警的数据。展示空调、新风、风机盘管、电梯、监控、医疗等设备的故障、报警数，同时点击故障或告警数，可查看详细的故障、告警详情，方便用户整体监控医院大楼设备的运行情况。

N：能耗板块

按楼宇、楼层、功能区、科室、设备等维度展示电力、水力、暖通的日 / 月 / 年能耗数据，热力分析、功能分区工作负荷、天气状况等各类数据的多维分析，支撑按照能耗管理规则自动调节公共区域开关、空调等开启数量或通过人工远程单控、群控手段以达到节能目的。

医院能源 KPI 包括能耗 KPI 和经营性 KPI，从容大模型通过对 KPI 的定额管理，提醒用户能源消耗的速度及定额余量，时间颗粒可选年或月。全面了解医院各项目耗能情况，可以针对性的对各用能情况进行数据分析。并且平台对耗能的资费做了计算和统计，更清楚的了解到各能源消耗所支出的费用，助力医院实现医疗碳中和。

N：公共安全应用

通过地图、杯签，场景、柱形图、环形图、预警雷达等各种图表形象标示摄像机，门禁，巡更等相关设施的布点标签，直观地呈现视频，设备运行情况。高度融合医院各领域现有数据资源，对医院安防、基础设施、停车场、环境空间等管理领域的关键信息进行综合分析，辅助管理者全面掌控医院运行态势，实现医院人、事、物统一管理，医院综合安全运营态势一屏掌握。

医院治安事件基本都是发生在医院门诊部分区，门诊部存在人员流动大，人员群体复杂，交叉流动人群多等特点，主要为三大群体：医院的工作人员（挂号收费工作人员）；就诊的患者；就医的患者家属或陪护人员。对于一些三甲医院和特色医院还活跃着大量的号贩子。从整个医院的突发事件和人群聚集事件来说，门诊部分区是整个医院建筑体系中各种纠纷、争执、失盗、失物等事件发生频率最高的区域，进行重点人员重点区域安防布控。

效益分析

中山大学附属第一（南沙）医院信息基础设施与智能化管控平台建设项目是医院信息化建设的第一步，是智慧医院建设的基础。本项目与建筑智能化建设进程结合紧密，满足了 5-10 年医院数字化、信息化、智能化可持续发展的要求，为智慧医院夯实基础。云从科技以信息与通信集成为抓手，助力“中山大学附属第一（南沙）医院”内各级系统的高度感知、互联与智能，医院人、物、系统之间进行无障碍沟通与协同，进而使医院成为一个能优化配置医疗资源，持续进行服务创新的高效生态系统，打造“智慧型数字医院”。

阿斯利康：基于学术文献溯源的药品不良反应报告生成助手

阿里云计算有限公司

阿里云创立于 2009 年，是全球领先的云计算及人工智能科技公司。阿里云为 200 多个国家和地区的企业、公共机构和开发者，提供安全、可靠的云计算、大数据、人工智能等产品和服务。经过十三年的发展，阿里云已成为全球前三的云服务提供商。

阿里云是全国首家云等保试点示范平台和首家通过国家等保四级备案测评的云服务商。为中国超过一半的上市公司，为 80% 中国科技创新企业提供云计算服务。

在后疫情时代，社会经济的方方面面都在全速重构，阿里云正在逐渐成为赋能数字经济的数智创新平台，为数字经济、数字社会、数字政府提供创新价值。

概　述

随着我国药品相关法律法规的日趋完善，药品持有人面临越来越严格和规范的监管，以落实作为产品持有人的责任主体。不良反应监测作为持有人必尽可的法律义务，是持有人药物警戒工作的基础和重要一环。如何进行高效的不良反应监测，尤其是来自于学术文献的不良反应，是持有人一直探索的方向。为此，企业作为药品持有人，会投入大量人力物力通过传统方式阅读学术文献并汇总安全性信息，再进行后续处理及并递交监管机构。而传统的人工阅读无论是效率还是准确度方面都有很大的瓶颈，在寻找利用新技术提升文献阅读与报告的生成效率。

阿里云以通义生成式语言大模型为基底，在通义 Qwen 整体训练数据超过 3 万亿 token 基础上，融入海量医学知识文献与医疗数据所训练出来的垂直医疗行业大模型，帮助企业客户精准的解决在特定场景的文献阅读与报告生成的能力。

该模型通过与医学专家的合作，在特定领域进行了微调，使其能够更好的理解医学领域大量的学术文献，识别与药品不良反应相关的信息并生成相关信息的汇总。这可以极大提高药物警戒团队的工作效率，更快地处理药物不良反应报告，并及时上报给相关部门。

需求分析

生物医药行业是国家战略性新兴产业之一，也是未来经济社会发展的重要支撑力量。近年来，我国生物医药行业取得了显著进步，然而生物医药行业也面临着诸多挑战和风险，例如研发投入一个创新药品从发现到上市平均需要 10-15 年时间，耗费 20-30 亿美元资金。而且研发成功率很低，只有约 10% 左右的候选化合物能够进入临床试验阶段，只有约 1% 左右能够最终获得上市许可。其中获得上市许可的企业仍需要进行药品上市后管理，落实责任主体。

根据《中华人民共和国药品管理法》、 国家药品监督管理局关于药品上市许可持有人直接报告不良反应事宜的公告（2018 年第 66 号）及关于发布个例药品不良反应收集和报告指导原则的通告（2018 年第 131 号）等相关法律法规，药品上市持有人应主动收集学术文献中涉及的不良反应信息，并对不良反应信息进行记录、传递、核实及确认。需要确认的内容首先应确认是否为有效报告。一份有效的报告应包括以下四个元素（简称四要素）：可识别的患者、可识别的报告者、怀疑药品、不良反应。

对于药品持有人来说，确认是否含有以上元素并形成一份安全性信息汇总的传统且主流的做法都是通过人工阅读，但人工阅读费时费力，尤其在企业产品数量多，潜在文献多的情况下效率非常低下，且人工确定存在很大主观性，无法保证准度。随着人工智能的发展，尤其是生成式的人工智能已经在各行各业寻求落地场景，企业可以考虑利用新技术替代人工阅读并形成安全性信息汇总，再由人工复核结果，以方便后续处理，最终达到提升效率和统一标准的目的。

案例介绍

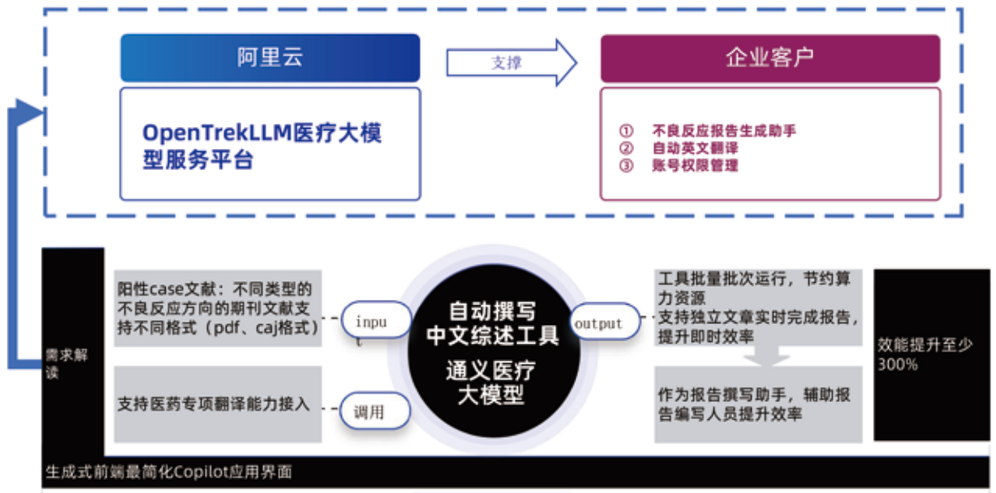
阿里云联合阿斯利康，利用通义医疗行业大模型技术实现学术文献来源的不良反应报告生成自动化。

通义医疗行业大模型由阿里云研发，以通义生成式语言大模型为基底，在通义 Qwen 整体训练数据超过 3 万亿 token 基础上，融入海量医学知识文献与医疗数据所训练出来的行业大模型；并针对医药行业知识密集以及严肃医学的特性，使模型在医疗领域具备更强大的行业知识的推理、认知等能力，包括知识问答准确性、单轮问诊权威性、多轮问诊场景化、领域制式文本生成以及领域文献理解等。除了在 MMLU、C-Eval、GSM8K、MATH、GaoKao-Bench 等 12 个权威评测中取得优秀的成绩外，在医学数据集评测中也具有优异表现。

通义行业大模型通过 API 与交互式问答形式提供服务，并提供用于模型二次训练与评测的完整操控平台，与阿斯利康联合完成对药企的应用案例落地。

在应对医学领域的学术文献理解方面，针对文献进行特定格式的不良反应信息的识别和总结，，生成用于不良反应报告后续处理的内容，提升企业运营效率。

同时，通义医疗行业大模型服务平台还提供用于安全防范的安全卫士能力，根据生成式人工智能技术的健康发展规范，对生成内容进行了安全管理，回答内容积极向上，并提供内容安全应急管控机制。



效益分析

该案例在医疗领域中具备应用实践的代表性，推动国家政策落实，响应国家对人工智能的引导性建议，把握人工智能新科技革命浪潮，推进产业智能化发展与探索。

在经济效益方面，以数字化手段降低医药研发与药品上市后的监测过程中的成本投入，提高临床试验执行效率，节省时间和成本。

支持大型头部药企在中国，基于中国市场的特性，运用新技术探索应用实践。

基于知识图谱和大语言模型的制造业数字化转型平台

上海说以科技有限公司

上海说以科技有限公司是一家专注于科技发展和人工智能领域的先进企业。在国家和政府的支持下，公司秉承科技创新的理念，致力于为客户和社会创造价值。我们的业务涵盖了与中央及上海市相关政策结合的 AI 技术服务，服务对象包括大型企业和科技独角兽企业。我们的主要服务包括项目规划、算法开发、软件开发、培训等，业务遍及上海及周边地区。

在知识产权方面，我们拥有 7 项专利授权和 14 项软件著作权授权，还有多项待公布的知识产权。我们为客户提供的项目服务涵盖了气象预测系统、智慧港口无人驾驶技术、大规模行业智能知识图谱平台、基于计算机视觉的机器人全自动检测检修平台、智能金融服务平台等多个领域。

我们的团队成员在人工智能及心理学领域具有深厚的知识背景。我们在基于知识图谱优化的大语言模型领域取得显著成果。我们的创新平台利用知识图谱优化大语言模型，提高自然语言处理能力和语义理解的准确性。通过知识图谱模块，我们能有效获取、表示和学习专业领域知识点及其相互关系。同时，我们的系统能够理解用户查询，并根据问题内容匹配知识图谱中的相关知识点，确保数据安全性和合规性。

概　述

本项目是一款创新的制造业数字化转型平台，融合知识图谱和大语言模型的尖端技术。本平台致力于深入分析制造业关键数据，进行智能处理，加速行业的数字化步伐，显著提升生产效率与决策精准性。

平台核心技术涵盖了专业知识图谱的构建，涉及设备、流程及业务关系的全面覆盖。通过这种结构化的知识架构，能够高效处理制造业中的多模态数据，准确捕捉并解决业务痛点，同时灵活调整策略。结合知识图谱，平台能够执行复杂的数据推理，为用户提供精准的业务洞察和坚实的决策支持。平台已成功应用于多个案例，比如：大型设备故障诊断、多工厂协作优化。

本项目旨在推动制造业的数字化转型和智能化升级，进而增强整个行业的竞争力。

需求分析

制造业的转型主要集中在生产过程的智能化、供应链的高效整合，以及客户体验的显著提升。然而，在这一转型过程中，存在三个亟需解决的问题：

- **顶层策略与执行层的脱节：**数字化转型需要利用知识图谱和大语言模型，这些工具能够明确地将高层决策与现场操作相连接，确保策略制定与执行的一致性和效率最大化。
- **工作效率与产出：**提升工作效率和产出是制造业转型的关键目标。数字化转型应着重于优化生产流程、改善供应链管理，以及通过智能数据分析提高决策质量，从而实现效率的大幅提升和产出的优化。
- **避免盲目数字化投资:** 数字化转型需要有针对性地识别和实施有效的技术解决方案，避免盲目跟风，确保每一步投资都能为企业带来实际价值。

案例介绍

本项目旨在实现制造业的数字化和智能化转型，应对全球化竞争带来的挑战。通过结合知识图谱和大语言模型的技术，本平台能有效解决顶层策略与执行层之间的脱节、提升工作效率与产出，同时避免盲目的数字化投资。

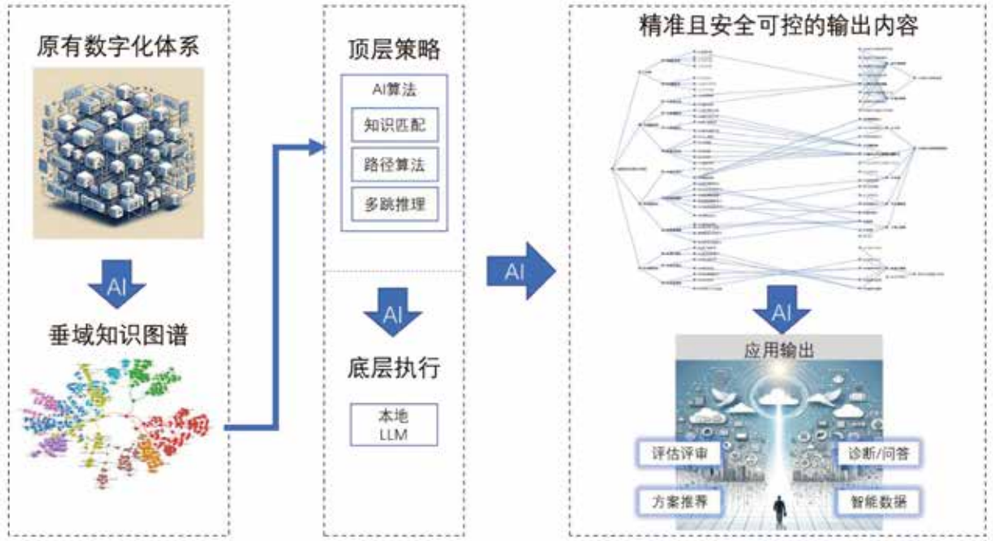


图 1　基于知识图谱及大语言模型的数字化转型

主要能力

- **知识图谱构建与应用：**平台通过分析制造业的海量数据，构建专业知识图谱，包括设备、流程和业务关系等，为数据分析提供坚实基础。

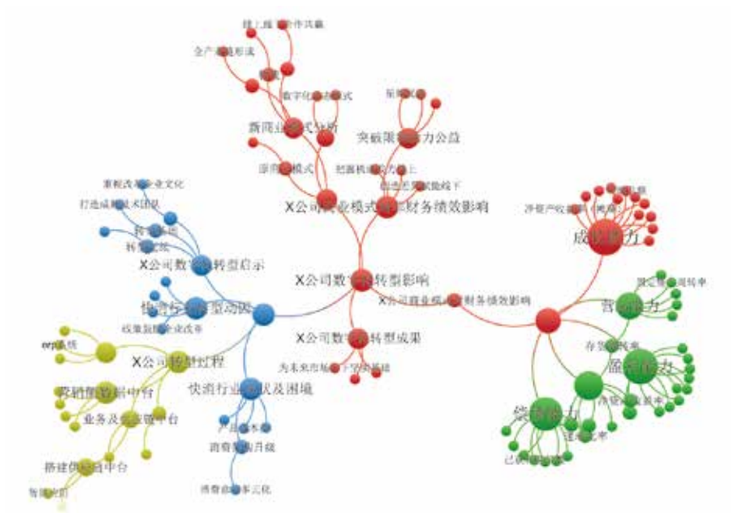


图 2　本平台构建的制造业数字化转型顶层知识图谱

- **大语言模型集成：**利用大语言模型处理和理解自然语言数据，提高对用户需求的响应速度和准确性，优化复杂业务场景的解析。

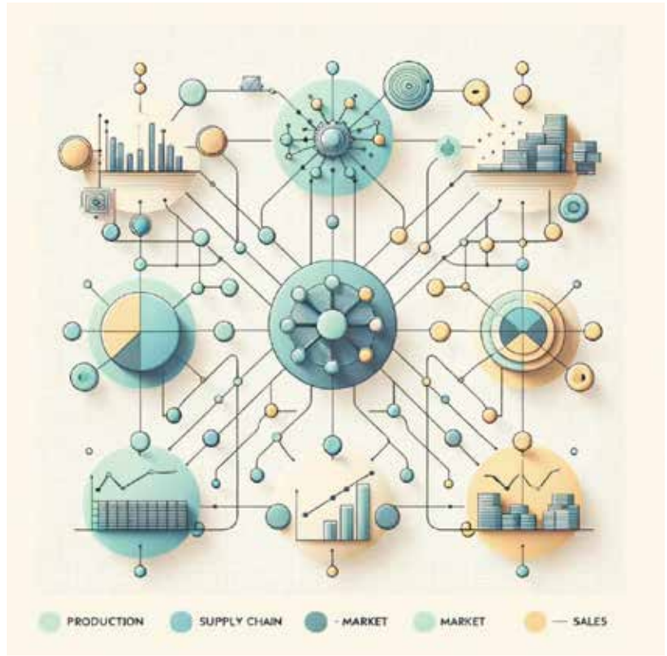


图 3　大语言模型优化复杂业务场景

技术创新点

本项目的核心创新点在于知识图谱与大语言模型的深度融合。通过这种结合，平台不仅能够理解和处理复杂的自然语言数据，还能在知识图谱的支持下进行更精准的数据分析和推理。这一技术提高了自然语言处理的能力和语义理解的准确性，同时也增强了内容的精准控制和数据权限的精确管理。此外，通过用户查询理解与知识点匹配，平台能够针对用户问题，精确匹配知识图谱中的相关知识点，保障数据安全性和合规性。

实施效果与应用落地情况

- **大型设备故障诊断：**在大型设备故障诊断方面，平台能够利用专业知识图谱和数据推理能力，快速识别设备故障，准确定位故障原因，并提供相应的解决方案。大大缩短了设备故障排查和维修的时间，降低了设备停机时间和维修成本，提高了设备的整体运行效率和生产效益。在实际应用中，本平台已经成功帮助多家企业实现了设备故障的快速定位和修复，提高了设备的运行稳定性和生产效率。

- **多工厂协作优化：**在多工厂协作优化方面，本平台也展现出了强大的实力。在制造业中，多工厂协作往往面临着信息传递不畅、资源调配不合理等问题。而本平台通过构建覆盖全业务流程的知识图谱，能够实时跟踪和分析各工厂的运行状态和资源利用情况，实现信息的透明化和资源的优化配置。在实际应用中，本平台已成功助力多家制造业企业实现了多工厂之间的无缝协作，解决了顶层策略与执行层的脱节问题，避免盲目的数字化投资。

应用落地情况

- **行业认可与应用：**该平台已在多个制造业企业中成功应用。用户反馈显示，平台显著提升了生产效率和决策精度，还能将数据访问权限与输出内容挂钩，有效降低了运营成本。

综上所述，本项目不仅解决了制造业面临的关键挑战，还推动了整个行业向数字化、智能化的方向发展。通过这一平台，制造业企业能够在全球化竞争中保持领先地位，实现可持续发展。

效益分析

经济效益

- **降本增效：**通过整合企业数据，优化生产流程，显著提高了生产效率，同时降低了运营成本。
- **增强竞争力:** 提供的个性化数字化转型方案，能更合理更快速地满足市场和客户需求。

社会效益

- **促进产业升级：**推动了制造业从传统生产模式向高效、智能化的现代生产方式转变，有助于行业的技术进步和产业升级。
- **增强社会就业质量：**数字化转型的推进可以创造高技能的就业机会，提高劳动力的技能水平，从而提升整个社会的就业质量。

构建安全可控的本地大语言模型，在数字化转型过程中，AI 的应用既高效又安全，为企业的智能化升级提供强大支持。

东方翼风大模型

中国商飞上海飞机设计研究院

上海飞机设计研究院是中国商用飞机有限责任公司的设计研发中心，担负着中国民用飞机项目研制的技术抓总责任，承担着飞机设计研发、试验验证、适航取证以及关键技术攻关等任务，是我国最大的民机研发中心。

上海昇腾人工智能生态创新中心

上海昇腾人工智能生态创新中心立足于上海，面向全国，以昇腾全栈为技术支撑的人工智能软硬件基础设施，是推动人工智能产业发展与加快发展区域数字经济的重要载体。支持人工智能计算中心运营和创造产业价值，致力于成为上海昇腾 AI 技术能力源头。

概　述

飞机的气动设计是飞机设计最基础、最核心的技术之一。随着飞机设计研制周期的不断缩短，现有的气动设计方法存在诸多局限，团队基于 MindSpore 和昇腾硬件，面向三维机翼的流场仿真，实现流场非标数据到 AI 张量数据的统一表征，通过内存复用、重计算策略和多维度混合并行，突破大模型训练内存墙，实现三维千万级网格流场训练和推理，构建多专家混合 AI 模型、混合精度优化等关键技术，流场预测平均误差低至 1e-3 量级，精度媲美传统仿真软件，单次仿真速度相比传统仿真提升 1000 倍，实现大飞机翼型流场的高效高精度的仿真，助力国产大飞机仿真设计。

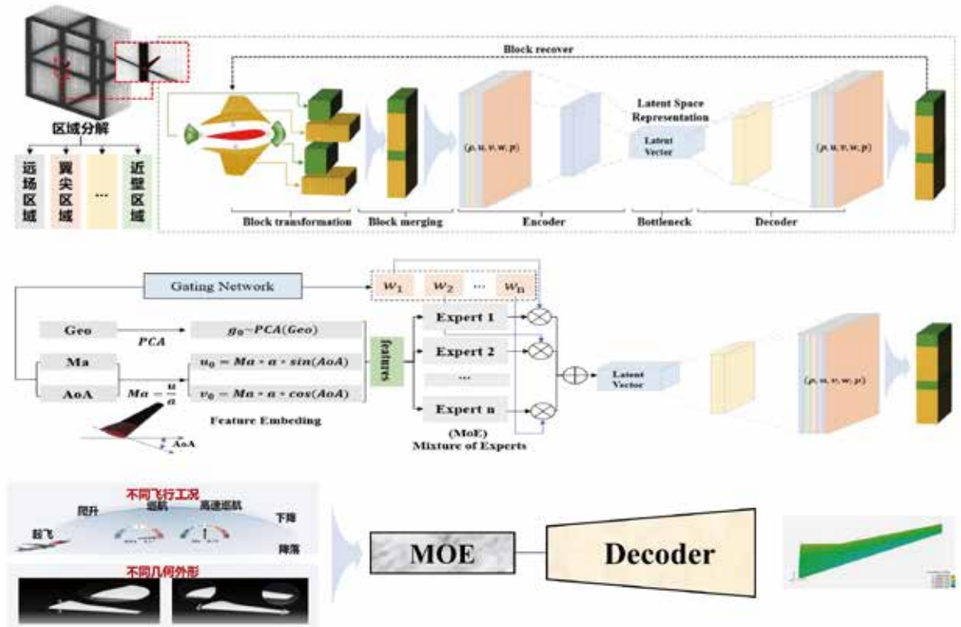
需求分析

国家对大飞机的发展提出了殷切期望，大飞机梦是中国梦的一部分，是我国实现科技自立自强的重要组成部分，也是浦东六大硬核产业中发展空间最大的产业之一，“创新引领大飞机产业高质量发展”已成为迫切课题。传统飞机设计流程包含了飞机外形物理建模、流场网格离散、数值仿真软件求解、流场分析最后修改物理建模参数，反复迭代，进行优化设计的流程，这一过程中，单次高精度流场仿真需要消耗大量的高性能计算资源，仿真时长可达到月级，因此，采用 AI 方法挖掘已有的流场数据资源，加速流场仿真，对提高飞机设计效率，具有重要价值。

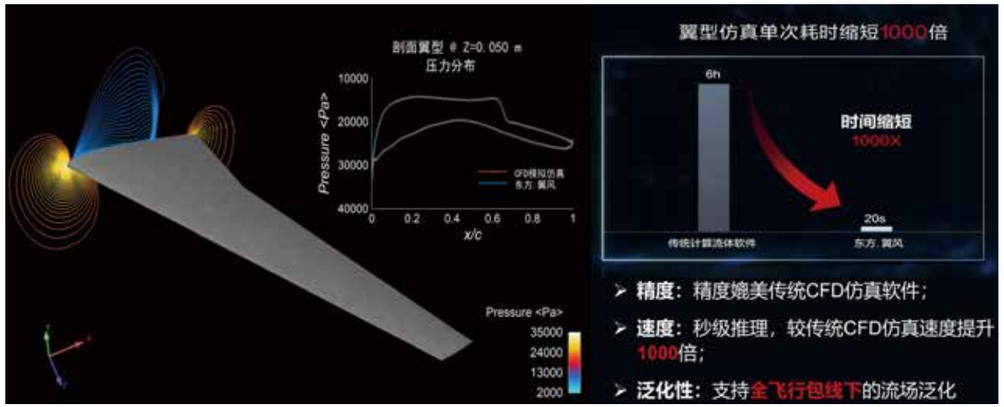
案例介绍

基于 MindSpore 和昇腾硬件的“东方·翼风”大模型面向三维机翼在全飞行包线下的流场仿真，精度媲美传统工业仿真软件，效率提升 1000 倍以上，并实现了不同机翼几何以及全飞行包线下的泛化，相关结果在 2023 年世界人工智能大会（WAIC）上发布，并获得最高奖 SAIL 奖。

模型的技术创新点如下：1、针对三维机翼流场网格的复杂拓扑结构，基于区域感知的多块分解技术，分解原始流场，实现流场非标数据到 AI 张量数据的统一表征，基于雅格比坐标转换变换计算域，融合多块流场，构建 AI 张量化数据输入，构建 AutoEncoder 模型，寻找高维流场特征的隐式空间的低维表征；2、保留 AutoEncoder 模型中 Decoder 模型部分，设计 MoE-Decoder 模型，构建飞行参数和翼型几何与流场隐式空间低维特征间的映射，进行参数微调，根据传统仿真经验，多个 MoE 模型区分不同飞行参数之间的流场区别，提升模型的泛化表达能力；3、在新工况和新翼型的推理中，几何和工况直接输入 MoE-Decoder 模型，实现全流场物理信息秒级推理，通过内存复用、重计算策略和多维度混合并行，突破大模型训练内存墙，实现三维千万级网格流场训练和推理。



“东方·翼风”大模型在大飞机流场仿真和大飞机设计中实现了 3 大突破。首先是效率突破，AI 模型替换传统 Navier-Stokes 方程求解，缩短仿真时间 1000 倍，大幅提升典型场景仿真效率。其次是精度突破，“东方·翼风”对流动剧烈变化区域特征进行精细捕捉，整体 AI 流体仿真的预测精度明显提高，最终全流场平均误差在 1e-3 量级；最后是模型突破，“东方·翼风”采用多专家混合 MoE 模型，覆盖全飞行包线 (起飞 - 爬升 - 巡航 - 降落)，显著增强模型泛化性。



效益分析

2023 年 7 月 6 日，在上海举办的 2023 世界人工智能大会（WAIC）上，基于昇腾 AI 和昇思 MindSpore 开发的“东方 . 翼风”大模型正式发布，并荣获了 WAIC 最高奖项 SAIL 奖。东方·翼风”是业界首个三维超临界机翼流体仿真大模型，目前适合大型客机设计仿真，未来将逐步拓展高铁、船舶、航天等更多场景。

团队推出了 MindSpore Flow 是基于昇思 MindSpore 开发的流体仿真领域套件，支持航空航天、船舶制造以及能源电力等行业领域的 AI 流场模拟，旨在于为广大的工业界科研工程人员、高校老师及学生提供高效易用的 AI 计算流体仿真软件。双方密切交流，共同打造超级应用落地，构建套件的生态影响力。

智己汽车：用大模型打造智能时代出行变革者

北京智谱华章科技有限公司

智谱 AI 致力于打造新一代认知智能大模型，专注于做大模型的中国创新。公司合作研发了双语千亿级超大规模预训练模型 GLM-130B，推出了千亿基座的对话模型 ChatGLM 及开源单卡版模型 ChatGLM-6B，并打造大模型产品矩阵，包括生成式 AI 助手智谱清言、高效率代码模型 CodeGeeX、高精度文图生成模型 CogView、多模态对话语言模型 VisualGLM-6B 等。在全新升级的 ChatGLM3 赋能下，生成式 AI 助手智谱清言已成为国内首个具备代码交互能力的大模型产品。

公司践行 Model as a Service（MaaS）的市场理念，推出大模型 MaaS 开放平台（https://open.bigmodel.cn/），基于领先的千亿级多语言、多模态预训练模型，打造高效率、通用化的“模型即服务”AI 开发新范式，实现服务效率的提升。

概　述

智己汽车 - 上汽集团旗舰品牌，是阿里巴巴智慧赋能，专注打造的高端纯电智能车。智己汽车聚焦「智能化」，旨在成为智能时代出行变革的实现者。2023 年 10 月，智己汽车正式上市率先接入未来智舱体验的全球车型“中大型智能轿跑 SUV” 智己 LS6，搭载了业内首个融合千亿级参数的“智己生成式大模型”。该模型针对智能座舱场景，基于智谱 AI 研发的中英双语对话模型 ChatGLM，由智已汽车和智谱 AI 携手共创，拥有强大的自然语言处理和机器学习能力，可自动编排、自学习进化；并可通过多重意图识别，来“瞬间感知真实需求、精准识别所有指令”；以及结合车主历史交流偏好和习惯，生成“独有的出行体验”。

需求分析

近期，随着大语言模型带来的强大影响力，我国汽车产业正在迎来一场智能化转型的激烈竞争。短时间内，业界已达成共识：无论采取何种方式，中国汽车产业都必须迅速接纳大模型。在这一共识的驱使下，如何在实际应用场景中落地大模型成为各大车企积极探索的问题，其中智能座舱和智能驾驶成为大模型在车辆端落地的两大关键领域。

而聚焦于智能驾舱场景，大模型也能较好的解决传统车载交互系统中的两大痛点：

1. 语音指令功能弱。目前的车机系统对于语音指令理解率低。用户需要明确指令才能执行操作。车机被动执行命令。

2. 回复内容偏生硬。目前的回复内容是基于传统的模版方案生成。内容回复缺乏实时和趣味性。

对汽车企业而言，借助 AI 大模型，未来可实现人、车、生活等科技生态闭环下的互联互通，为用户带来前所未有的体验。

案例介绍

在大模型时代，理解、生成、推理、记忆等能力成为核心要素，座舱内人和车的关系也将变为人和虚拟人之间的关系，交互方式将发生巨变。随着科技的发展，人工智能技术在汽车行业的应用日益广泛，为车主带来更为便捷、实用的驾驶体验。

智己生成式大模型融合了千亿级参数，在通用大模型的基础上，通过驾驶座舱中的语料数据进一步增强模型在特定场景下的生成能力，并具备自然语言处理 + 机器学习能力，可自动编排、自学习进化；可通过多重意图识别，来“瞬间感知真实需求、精准识别所有指令”，如：对“关闭车窗打开空调再给我讲个笑话”“想去成都自驾游三天”等指令作出较为精确的回答；并可结合历史交流偏好和习惯，生成“独有的出行体验”，提供角色扮演、解梦、星座运势、创作图文 / 音乐等娱乐方面的功能。

五大功能落地功能

- **多意图指令识别：**智己汽车大模型具备强大的多意图识别功能，可准确识别一句话中的多个指令，实现一站式操作。
- **模型意图二次确认：**通过引导对话的方式进行意图确认，确保准确理解用户需求，提高服务精度。
- **趣味内容生成：**内置大量笑话、故事等趣味内容，根据场景和用户描述智能生成相关内容，增强互动趣味性。
- **语音交互游戏：**支持语音交互游戏，如成语接龙、猜谜语、智力问答等，丰富驾驶过程中的娱乐体验。

- **多元人设：**提供多种人设风格对话，满足不同用户喜好，实现个性化交互。

应用成果

1. 用户交互的长尾场景意图识别准确率提升至 90% 以上，以个性化人机 AI 对话，满足了用户对 AI 座舱体验的“智能需求”和“情感需求”。

2. 与大模型闲聊的 7 日留存率高达 50% 以上，表明用户对智己汽车大模型的黏性较高，驾驶旅途更加愉悦。

在智能驾舱领域，大模型的落地，将加速软件能力升级，推动人机主动式交互时代的到来，并通过赋能语音助手对于乘客的语音语义理解能力。打通其在视觉、听觉、触觉等多模态应用上的操控力，形成深度的乘驾人机主动式互动体验，进而形成车企自身独特的智能化差异，构筑核心竞争力。

效益分析

随着科技的飞速发展，汽车行业正迎来智能化、网络化的深度转型。人工智能技术在其中扮演着关键角色，尤其是大模型，将在自动驾驶、驾驶舱智能化、人车传感器互联、工厂数字化等多个领域发挥巨大价值，并为车主提供更优质的驾驶体验。

智能驾舱则是大模型在汽车行业能够最快落地的场景之一，传统的车机系统存在界面复杂、操作繁琐等问题，用户在使用过程中容易产生困扰，而大模型的生成式能力则展现出了巨大的潜力，不仅能处理完整的对话，更能保持对前后文的理解，形成良好的语音交互体验，大大提升汽车的人机交互水平，使驾驶过程更为便捷和舒适。

基于山下话童大模型的贷后催收示范应用

上海特赛发信息科技有限公司

上海特赛发信息科技有限公司创立于 2021 年 9 月 10 日，在大语言模型领域潜心研究沉淀数年，是一家专注于人工智能技术创新与应用的高科技企业，以在 AGI 领域为世界提供更多一种选择为使命。公司旗下主要产品山下话童是一种基于大语言模型的消费金融智能体，不仅打破机器人死板固定沟通范式，且产品上线后在通话时长、通话轮次、通话中有效沟通数均全面超越传统电话机器人。

概　述

该应用由上海特赛发信息科技有限公司开发。在智慧金融领域，基于山下话童大模型，面向交通银行贷后催收需求，探索传统机器人外呼现有弊端的解决方案，实现贷后催收的智慧化高效新模式。截止目前，项目已初步实现打破机器人死板固定沟通范式，并以类人智慧理解人类语言，提供情绪价值等优化功能。

需求分析

交通银行信用卡在电话催收外呼上年支出 1 个亿，然而传统机器人在催收上具有轮次固定、施压不精准以及共情能力弱等弊端，从而不能满足该行全面提高日拨打量、最大化降低通话成本，而且在数据统计和客户跟进方面拥有客观高效、及时灵活的瞩目性等需求。

案例介绍

一、实现高效率低成本的催收外呼模式是交通银行信用卡中心的重要目标，然而交行现有传统机器人外呼工作存在五个方向的业务痛点，

（1）核身：核身词槽失败率 5% 左右；

（2）最大输出 100 字以上：固定格式尽可能输出最多信息（包括逾期金额与利息、进征信、锁卡、法律责任等），由于太长太啰嗦导致用户直接挂掉 30%；

（3）根据词槽语义匹配：语义理解不准确 60% 以上，且运营需经常标记词槽以提升语义理解力，工作繁琐枯燥；

（4）催还最低，不还关卡冻结：轮次到 5 轮后基本是无效沟通，用户咨询的问题和回答 40% 以上不一致，以及涉及还款金额计算不会给方案；

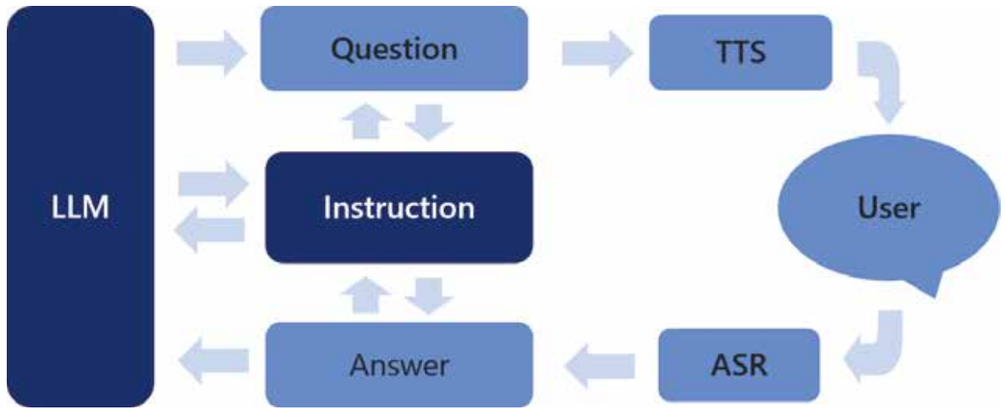
（5）轮次多转人工：需转人工比例 15% 左右，机器人所降低人工成本有限。

二、山下话童针对交行贷后催收现有的五个问题领域实现技术创新升级，

（1）话术设计：根据用户画像，实现画像 + 话术的协同训练；同时具备逻辑性推理能力，灵活组合变量及话术，且无拼接痕迹；

（2）意图词槽管理：不再需要穷举各种表达，根据有限的调教，模型就能实现举一反三的泛化，实现开箱即用，以自然语言对话的方式理解并回复；

（3）对话流程设计：在提示工程和思维链指导下，同时根据 Instruction 实现自我递归，可以有效理解客户说的话并约束场景，给出最恰当的答复（如下图示）；



（4）共情能力：有人的意识，趣味性催收，“明是非”、“格万物”阶段；

（5）施压层级：懂分寸，有的放矢，该施压就施压，该安抚就安抚。

三、山下话童在与 ChatGPT4.0 的多轮对比中得出显著的实施效果，如下图所示，

四、山下话童大模型能够根据交通银行信用卡中心对贷后催收的需求，在通话过程中随需应变对答如流打破僵化流程，实现认真倾听并循序推进通话的进行；能够以类人智慧理解人类语言精准施压，在通话过程中拿捏时机并多次提醒；并且提供情绪价值，在通话过程中保持趣味且不忘初衷。

效益分析

基于山下话童大模型的贷后催收模式面对的是接近年 200 亿营收的市场，具有高度经济可行性和盈利能力。以交通银行为例，该行信用卡在电话催收外呼上年支出 1 个亿；还有上海银行、兴业银行等相当体量的银行以及许多中小银行和农信社的市场体量；除信用卡外，银行还有借记卡用户，这部分用户是信用卡的两倍以上。据统计，整个外呼市场催收占比 20% 左右，客服占比 30% 左右，营销占比 50% 以上。该案例不仅改善银行催收体系，已实现并得到业务认可，高效完成催收外呼工作；而且带来的技术创新和进步，提高相关行业的竞争力。

海淀区一网统管接诉即办工程项目

北京百度网讯科技有限公司

百度智能云致力于为企业和开发者提供全球领先的人工智能、大数据和云计算服务，加速产业智能化转型升级，秉承“云智一体、深入产业、生态繁荣、AI 普惠”的理念，基于大模型时代深入产业沉淀的客户需求，致力于用更高效的算力基础设施、更好用的一站式大模型平台、更丰富的行业解决方案和百花齐放的 AI 原生应用去推进人工智 能在千行百业的落地与伙伴生态赋能。

概　述

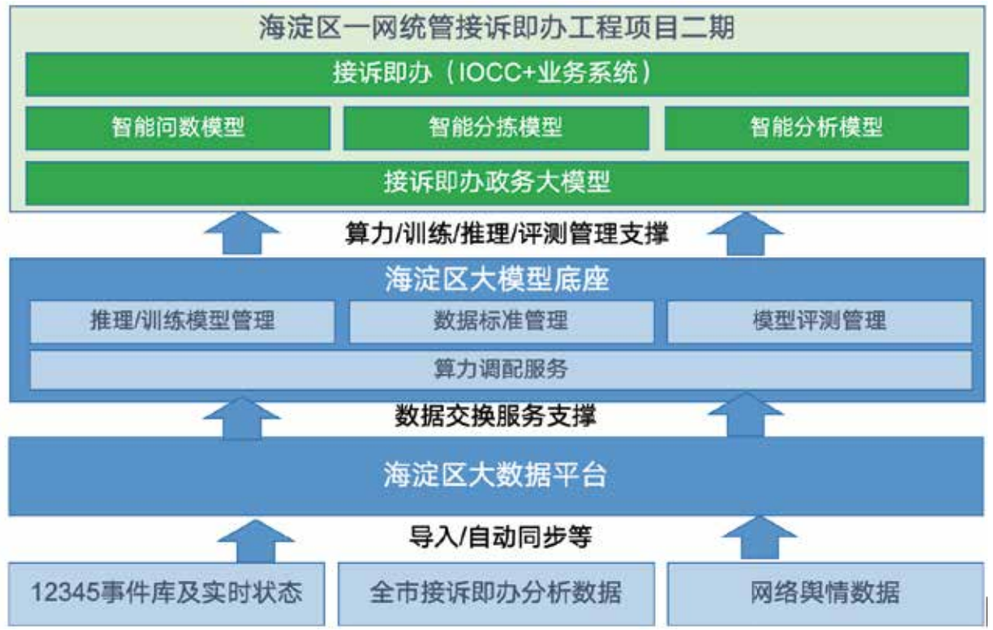
2021 年《北京市接诉即办工作条例》颁布实施以来，区委区政府高位统筹调度接诉即办工作，对海淀接诉即办的体制和机制、工作标准、责任分工以及工作深度都提出一系列新要求，海淀区拟通过一网统管接诉即办工程二期建设，在城市大脑顶层设计的框架下，聚焦群众急难愁盼问题，基于生成式大语言模型技术对现有的“接诉即办”平台业务进行持续优化，支持领导通过对话式交互随心指挥调度，直观动态查看事件态势的文字、图表等多模态的分析数据，并智能生成事件分析报告，提升调度效能；实现工单 100% 响应，提升大部门间的横向办件能力；实现对考核成绩的预测和分析，提升区考核成绩，打造海淀区城市治理新名片。

需求分析

接诉即办是政府机构服务社会的重要职能，其工作范围涵盖了政府机构内部的各个部门及工作人员，以及社会公众对政府行为的监督和评价。为适应新的考核规则，解决区中心以及承办单位存在的问题，提高海淀区接诉即办成绩，海淀区亟需对现有的接诉即办系统进一步优化，实现以下需求：

- **增设推荐数据库，提升派件精准性**
- **提升流转效率，规范案件响应速度和结案质量**
- 市场监管局个性化业务办理需求
- 全区接诉即办业务精细化、精准化管理需求
- 提高协同办件能力，避免由一单多派造成的失分隐患
- 案件研判源头下放，实现案件过程管理
- 考核业务全流程监管，提高指挥、指导、督办的精准性、实时性
- 深度挖掘数据分析潜力，提升接诉即办靶向管理能力

案例介绍



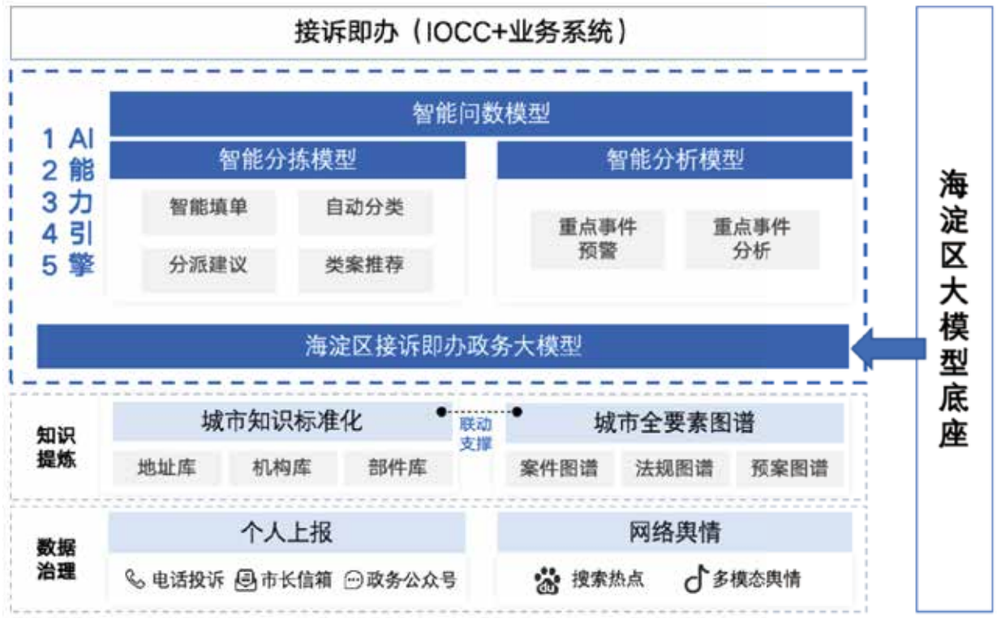
打造海淀区大模型底座

本项目将选择百度文心一言大模型提供底层支持，主要应用大模型在意图识别、报告摘要、NL2SQL 等方面能力。

其中意图识别主要用于领导进行提问的过程中，判断领导意图属于报告查询、数字查询还是闲聊需求，引导到不同的垂类应对方案。

NL2SQL 主要面向领导的数据查询提问需求，将领导的自然语言转换成数据库可以简单查询到的语言，返回具体的数值，进行便捷快速地查询。

报告摘要主要面向领导提问比较泛化、带有总结性质的问题时，根据查询到的数值给出相对详尽的结果并加以分析，形成一篇报告。



大模型赋能场景升级

根据百度“文心一言”语言大模型可及时理解文字语义并精确查找相关数据的特点赋能接诉即办场景痛点，通过人性化、口语化的人机交互方式实现以下功能：

1）让原有固化的驾驶舱显示内容变得可灵活调用、动态生成，查找数据、计算指标、简单指标统计由原来的 3 天提升到 1 分钟以内，图表绘制、可视化呈现由原来的 5 天减少到半小时以内；

2）让处置过程从复杂变成简单，在智能派单、智能标签，让坐席人员在日常工作中提高派单效率，从根本上提升公众的获得感和满意度；

3）通过大量历史案件训练定制模型，让承办单位面对每一件事件，都能快速精准参考以往成功经验辅助基层解决与处置，提高案件的解决率；

4）各类画像用大模型分析，并根据领导要求以及承办单位需求模版智能生成简单的报告，大大减少人工分析数据的时间，提高工作效率。

实施效果

通过探索大模型的更多场景应用，辅助各委办局、街镇等在教育、医疗、噪音治理领域先行先试；通过简洁、精准、智能的大屏辅助决策，优化领导驾驶舱的指挥决策效率，实现资源的精准调度，为区领导、指挥中心、委办局和街镇等不同层级领导指挥调度提供依据和抓手。

效益分析

社会效益

本项目通过智能化自动化的手段，让区政府、街镇和委办局办件人员快速响应市民诉求，优化提效办件流程，并规范结案模式，做到精准化办件，提升政务服务能力。

经济效益

通过本项目优化区接诉即办的工作流程和模式，建立业务快速反应机制，整体实现降本增效的管理效益，同时借助线上线下的效率提升，让市民的关键诉求切实得到响应和解决，提升市民的满意度。

应用推广前景

大模型时代的数字政府或智慧城市建设新思路，是未来 3-5 年的建设趋势，每个城市都应该拥有一套服务于该城市建设和发展的统一且可持续的城市级大模型底座，统筹大模型及城市资源，更好地支撑政企各类业务升级改造。

风乌气象大模型

上海人工智能实验室

上海人工智能实验室是我国人工智能领域的新型科研机构，开展战略性、原创性、前瞻性的科学研究与技术攻关，突破人工智能的重要基础理论和关键核心技术，打造“突破型、引领型、平台型”一体化的大型综合性研究基地，支撑我国人工智能产业实现跨越式发展，目标建成国际一流的人工智能实验室，成为享誉全球的人工智能原创理论和技术的策源地。

概　述

今年 4 月 7 日，上海人工智能实验室联合多家机构发布全球中期天气预报大模型风乌，首次实现了在高分辨率上对全球核心大气变量进行超过 10 天的有效预报。“风乌”能够准确地模拟大气动态，预测未来 37 个高度的大气状态和地表气象信息。“风乌”AI 大模型仅需单张 A100GPU 即可运行，30 秒内就可以生成全球未来 10 天的全要素大气场预报，相比于目前运行于超级计算机上的物理模型，其推理效率至少提高 2000 倍。风乌已先后在国家气象中心、上海市气象局、香港天文台完成测试部署，实际部署中评估结果显示风乌台风路径预测结果超过了目前所有人工智能模型和物理方法。风乌为我国气象预报工作提供了有力的技术支撑，有益于防灾减灾、新能源、航空航海、农业等各重要领域。

需求分析

数值气象预报对百姓生活、农林牧渔、交通电力等各行业的发展都提供了不可或缺的支撑作用。随着全球气候变暖，近年来强台风、极端高温等异常天气气候事件逐渐增多，加快开发更高效、更准确、预报时效更长的气象预报系统以指导防灾减灾工作已经成为我国和全世界科学界重要且紧迫的任务之一。由于大气系统物理过程的高度复杂性、求解物理模型的巨大资源需求、以及对顶级气象专家的高度依赖，气象预报性能的提升缓慢（例如全球中期气象预报可用性能每 10 年提高一天），基于物理模拟的气象预报技术已难以满足社会和经济的快速发展需求。于此同时，人工智能技术飞速发展，探索利用人工智能技术推动气象预报乃至地球科学的发展成为了一个有颠覆性潜力的方向。

案例介绍

全球中期气象预报是整个气象预报系统的核心，为区域、全球、地表、高空等各种气象预报提供了支持，并进一步服务了国家防灾减灾、新能源开发、航空航海等各行业。目前世界各国使用的气象预报系统（如欧洲中期气象预报中心 ECMWF-IFS，美国国家环境预报中心 NCEP-GFS）均依赖基于物理开发的数值模式，通过求解偏微分方程获得未来的天气预报。虽然数值模式具有极强的物理机理和可解释性，但由于大气、海洋及其耦合系统的高度复杂性，现有数值模式依旧无法准确刻画大气系统的非线性运动。“风乌”基于人工智能方法开发，借助神经网络在建模非线性关系方面的优势和大模型海量参数带来的强大拟合能力，可以更好地从历史大气数据中总结出大气系统的底层规律。从人工智能具体方法来说，风乌具有三个创新：一是提出基于多模态大模型思想的 Transformer 网络设计来表征多种多样的大气变量，实现对全球高分辨率高维大气数据的高效建模；二是提出新的优化目标，针对不同区域不同大气变量高度耦合但分布差异大的问题，提出不确定性损失函数，通过学习自动调整不同位置不同变量的优化权重，提高网络优化效果；三是针对长期预测这一难题专门设计方法，针对全球中期预报的长时序生成问题，提出“缓存回放”策略，使模型有意识地处理误差累积问题，在有限硬件的基础上间接实现了对长期预测误差的优化，显著降低了模型的长期预测误差。实验结果表明，“风乌”的 10 天预报误差比此前 DeepMind 发布的 GraphCast 模型降低 10.87%，在“风乌”报告的 880 个预测指标中，有 80% 的准确性高于 GraphCast。风乌已先后在国家气象中心、上海市气象局、香港天文台完成测试部署，进行业务实验运行，实时提供全球 0-10 天高空及地面预报产品。国家气象中心、上海市气象局、香港天文台的应用结果显示，在对今年夏季登陆我国的“卡努”、“海葵”等台风路径的预报中，风乌的台风路径预测结果较国际领先的数值预报模型表现更好，具备优势，为我国气象预报工作提供了有力的技术支撑。

效益分析

中国气象局上海台风研究所研究表明，单个台风路径预测误差每降低 1 公里，可减少直接经济损失约 1 亿元人民币。经过全面评估，在 24 小时台风路径上，风乌平均预测误差为 54.9 千米，而 ECMWF 预测误差为 58.9 千米，NCEP 预测误差为 65.3 千米。提前 5 天 /120 小时预报时，风乌平均误差为 247.6 千米，而 ECMWF 预报误差为 284.7 千米，NCEP 预报误差为 288.4 千米，风乌 120 小时台风路径误差相比 ECMWF 降低了 13.0%，相比 NCEP 降低了 14.1%。风乌精准的台风路径预报在降低台风带来的经济损失方面有着巨大的社会效益。基于风乌预报结果可以进一步地推动风力发电预报、光伏发电预报等新能源系统的升级，以及航空、交通、农业、水利等各系统全面的提升，具有广泛的应用前景。

基于大模型的智能培训

竹间智能科技（上海）有限公司

竹间智能科技（上海）有限公司成立于 2015 年，是一家多年来专注于自然语言处理（NLP）领域研发和应用的企业。竹间智能专注于大型语言模型的研发与应用，为企业提供定制化的大型语言模型解决方案，以满足客户在不同场景下的各种任务需求。我们的产品包括大模型训练微调平台，该平台可以同时训练多个大型语言模型，并支持多种微调模式，以提高客户对大型语言模型微调工作的效率和精确度，降低客户对大型语言模型微调工作的学习和使用门槛。除了大模型训练微调平台外，我们还基于大型语言模型的能力为客户提供智能对话平台、智能对练培训平台、智能写作及知识管理平台等 4 款核心应用。我们希望通过我们的努力，帮助企业以更低成本、更安全、更快速便捷的方式来落地大型语言模型，并成为大型语言模型和用户之间的桥梁。

概　述

某电销客户近年来在业务量节节攀升的同时，随之而来的是对营销人员运营成本的居高不下。目前电销人员规模超 500+，年换新率约 40%，在如此高的人员基数和换新率的背后是培训部门面临的资源缺口和业务部门面对的上岗压力。如何能够在较短的周期内将一个业务小白快速培养成为一个合格的电销人员是长久以来一直困扰客户的难题。

当前客户使用线下集中授课的方式开展培训，上岗周期约 2 周，上岗考核由营销主管模拟客户与其进行电话对战。竹间智能基于自身丰富的场景落地经验以及强大的技术研发能力，帮助客户设计了一套集“学练考评问”于一体的智能化培训解决方案。通过智能对练培训平台强大的产品能力，在培训、练习、考核的过程中，全程陪伴学员成长。

需求分析

客户希望能够通过智能化的方式对现有的培训授课模式进行改变，通过增加在培训过程中与每位受训人员的互动性，提高受训人员对知识的掌握力度。

客户希望能够根据不同的用户画像构建各种意向的个性化用户角色，同时结合不同的用户诉求，更为真实的模拟出千人千面的语音对话场景，使受训人员在练习中更加逼近实战环境。

客户希望构建更为客观可控的考核规则及考核机制。在对业务知识掌握情况进行考核的同时，对受训人员的沟通能力、执行能力进行评判。并通过对会话数据的分析，给予整体的针对性考核评价。

客户希望释放培训资源压力，能够基于智能化的方式为受训人员提供知识问答的能力，帮助其对包括业务知识、岗位知识在内的相关问题进行解答。

案例介绍

基于对客户业务需求及业务特点的深入理解，竹间智能为客户提供了完整的智能化培训解决方案。以大型语言模型为底层能力而研发的智能对练培训平台，在培训、练习、考核、的全过程中，为学员能力成长提供帮助。

培训过程中，由大型语言模型扮演“老师”可以对知识点进行拆分讲解，并积极与学员进行问答互动，对学员的回答是否正确、完整都会实时的给出点评，学员也可以对知识中理解不清晰的内容随时进行提问，“老师”会针对性的予以澄清和讲解。这种一对一互动式的培训方式改变了传统培训过程中集中填鸭的单项输出方式，在提升培训效率的同时也大大节省了授课老师的人力投入。

对练过程中，在相同的业务场景下，不同意向、不同性格、不同关注点的“用户”基于自身的角色特点构建了不同的对话效果，在锻炼学员熟练掌握业务知识的同时，也通过口语化对话方式、个性化的对话内容锻炼学员对于业务知识灵活运用的沟通能力。企业不用再为学员的对战练习提供真实的客户名单，在降低用户投诉率的同时，也避免了对用户资源的消耗与浪费。

考试过程中，“考官”会基于学员与“用户”的对话内容，并根据业务出发点设置相应的评价策略。在给出考试得分的同时，还会针对对话内容中学员的优缺点进行点评，并对需要改进或保持的地方进行指出。构建学员能力的雷达图以及阶段目标完成情况反馈。在释放营销主管人工考核的工作量的同时，对考评指标和规则进行了统一和规范。

在充分结合了客户的业务需求及竹间的实践经验后，本次解决方案针对客户的实际情况引入了大型语言模型的能力，通过大型语言模型对上下文的语义理解及生成能力结合 Prompt 提示词的方式为客户构建更加灵活多变的对话场景。在场景构建的简易性和对话语境的开放性上给予客户更好的使用感受。直接改善了客户在培训业务中的诸多问题。

效益分析

相较于传统的人机对练产品基于大型语言模型的智能化培训解决方案为客户实现了从没得用到用的好的跨越。同时在释放培训资源、取消集中授课、缩短上岗周期，提高学员能力等方面为客户实现了降本增效，并带来了巨大的经济效益。

竹间智能提供的智能化培训对练解决方案，一站式覆盖客户的“学练考评问”培训业务流程。同时运用大小模型相结合的技术手段，为客户提供全方位的课程搭建方式，适用于客户各种类型课程的制作。不管是在题库问答、标准流程对话、开放式异议处理等场景下都能完全支持。为广大企业的线上、线下，营销、客服场景提供了全新的、高效的培训方式，避免了个人、企业乃至社会的资源浪费。

面向围手术期的医专大模型研究及其落地应用

云南联合视觉科技有限公司

云南联合视觉科技有限公司（以下简称“联合视觉”）成立于 2016 年 12 月，由深圳市联合视觉创新科技有限公司全资控股。联合视觉先后获得了国家级高新技术企业（2019 年），国家级科技型中小企业（2021 年），省级科技型中小企业（2018 年），省级成长型中小企业（2020 年）等荣誉称号，2023 年获批云南省专家工作站，云南省股交所代码为 D00202。

作为“人工智能”赛道企业，公司不断吸收人工智能领域的科学家参与到企业技术中心的建设。目前联合视觉围绕医疗数据采集、数据治理、数据分析构建核心研发能力，形成了临床信息管理系统、围术期管理系统等核心产品，能够为客户提供更加科学的信息化管理工具。

概　述

围手术期是围绕手术的重要医疗阶段，现有医疗大模型围手术期表现较为薄弱。项目言次提出围手术期医专大模型，通过学习围手术期知识、综合考虑医患信息，实现围手术期的智能应用，基于大模型搭建围手术期业务平台，实现围手术期智能对话、标准化匪手术期个体方案推荐等智能功能，为医护、患者提供服务。成果在十多家医院投入使用，以云南省第一人民医院为例，累计服务 20 多万手术病患，术后不良事件发生率降低 35%，成果受到云南网、掌上春城等新闻媒体报道。项目将持续构建围手术期技术和数据门槛，优化围手术期诊疗方式，未来潜在市场和实际价值巨大。

需求分析

民众对健康医疗需求和重视程度日益提升，推进智慧医疗成为了现阶段医疗卫生发展的必然趋势。近年来，人工智能迅速发展，特别是 ChatGPT 等大模型的惊人表现引发了大模型热潮，也为智慧医疗的研究和应用带来了新的机遇和挑战。

然而现有医疗大模型，虽然涵盖众多数据类型，并提供病理分析、智能问答等通用应用，但是未有面向围手术期的医专大模型研究。围手术期是围绕手术的重要医疗阶段，紧密关联病人生命安全。手术医生数量缺口巨大，急需相关研究为患者提供更好的医疗服务。现有医疗大模型围手术期表现仍然较为薄弱，需要构建围手术期大模型，理解围手术期业务和知识，合理提供医疗建议，分析病人情况，为病人提供合理的信息交互方式。

案例介绍

主要能力

面向围手术期这一重要医疗场景，公司构建首个围手术期医专大模型，如图 1 所示，学习围手术期知识，回答病人问题，辅助医生工作，并基于围手术期大模型构建围手术期业务平台，为医院、医护、患者提供服务。围手术期业务平台包括：术前、术后管理系统、临床信息管理系统等基础信息系统，并为医护提供了访视评估、智能对话、智能个体化手术方案、术后并发症预测（如图 2 所示）等辅助临床工作的功能，为患者提供信息整理、智能问答等了解手术情况的功能。

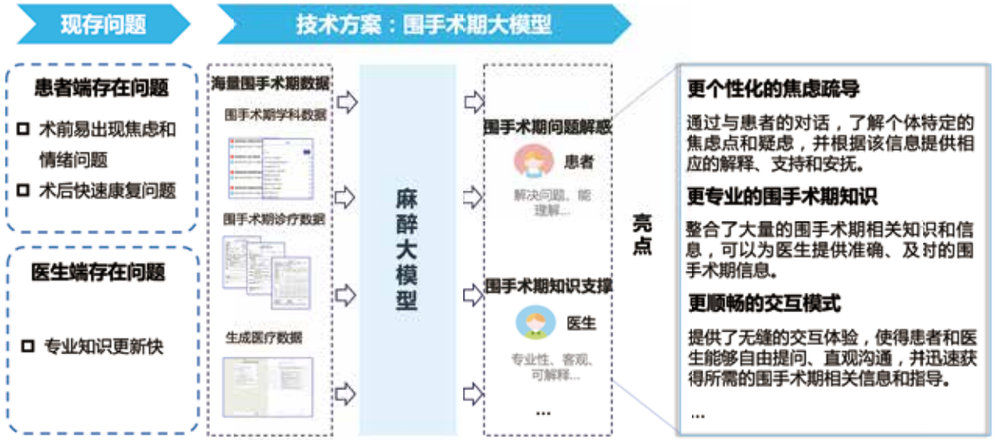
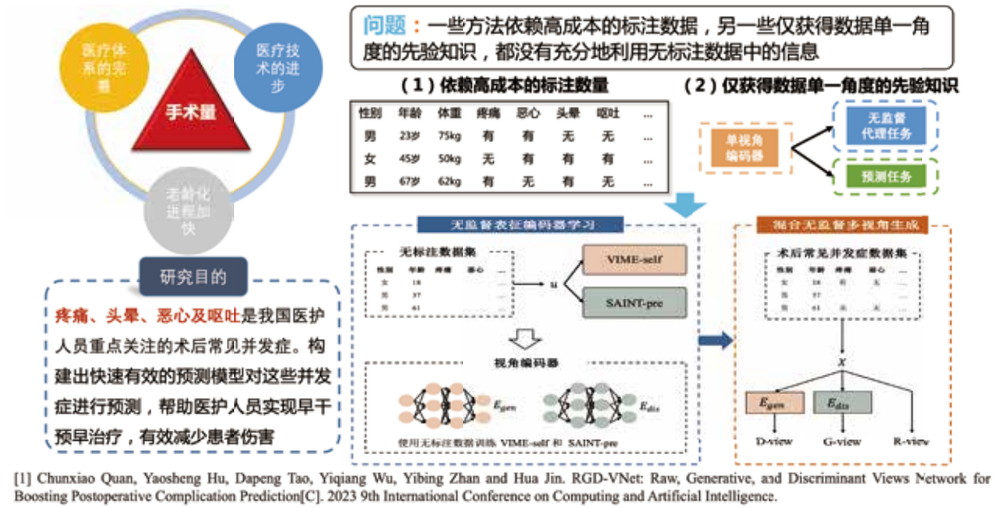


图 1



[1] Chunxiao Quan, Yaosheng Hu, Dapeng Tao, Yiqiang Wu, Yibing Zhan and Hua Jin. RGD-VNet: Raw, Generative, and Discriminant Views Network for Boosting Postoperative Complication Prediction[C]. 2023 9th International Conference on Computing and Artificial Intelligence.

图 2

技术创新点

为了获取充足的围手术期数据，提出了基于自学围手术期组合数据生成方法，自动生成围手术期问答、对话数据；为了提升数据质量，提出了大模型重点词汇检测过滤方法，清理不合理医疗数据；提出了通用大数据 - 专用数据的多任务联合学习的围手术期医专大模型微调范式，加强围手术场景下的性能；为了加强大模型感知用户生理和心理的能力，如图 3 所示，提出了基于用户历史数据以及知识增强的题词学习方法。

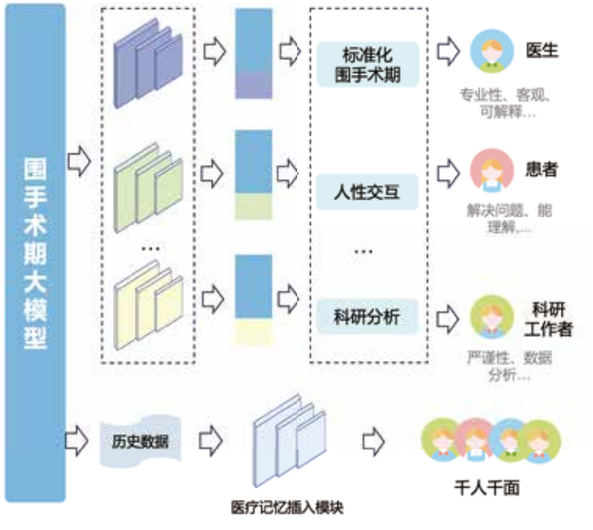


图 3

实施效果和应用落地

通过大模型和业务系统的实施，已经提供高效智能排班、人性化的智能问答、标准化手术方案推荐等功能，并辅助医生发现了更多的临床路径优化方案，改进了诊疗模式。已经在十多家医院投入使用，包括云南省第一人民医院、昆明市第五人民医院、玉溪市第二人民医院等。以云南省第一人民医院为例，累计保驾护航 20 多万手术病患，医护人员 1000 多名。我们联合医院对病患就医效果数据进行跟踪分析，相关的统计结果表明，大模型能够极大的改善手术患者就医体验，医院的患者术后不良事件发生率平均降低了 35%，如图 4 所示，成果先后受到云南网、掌上春城等新闻媒体报道。如图 5 和图 6，我们的模型和产品同时适配在华为昇腾的国产架构上应用，通过华为国产化认证，围术期大模型在昇腾 AI 创新大赛 2023 云南区域荣获应用赛道第一名。

图 4

图 5

图 6

效益分析

项目赋能数十家云南省医院，辅助开展高原胸外科手术临床优化等工作，有效提升高原地区非气管插管手术治疗等综合诊治水平。项目将持续提供技术服务和创新，联合顶级高校与头部大三甲医院合作，适配医院业务产品，构建围手术期技术和数据门槛；围绕 AI 能力、数据、解决方案、应用系统打造包括医院、医生、患者、科研机构等的生态，提升公司销售以及品牌价值。

国家出台政策《“十四五”国民健康规划》等鼓励围手术期相关研究和应用。大模型有可能彻底改变围手术期开展方式，并有望为未来疾病治疗的优化方案做出贡献。围手术期大模型应用仍然处于起步阶段，相关产品潜在市场规模巨大。

通过大语言模型与材料领域技术文件集合对原材料质保书进行智能审查

上海众深科技股份有限公司

上海众深科技股份有限公司成立于 2002 年，2015 年在全国中小企业股份转让系统挂牌为新三板企业，证券名称：众深股份（832251）。公司专注石化设备检测服务领域已有 20 年，通过高附加值的核心技术为客户提供设计审查、风险和安全研究、可靠性研究等一体化的石化设备检测服务是公司经营战略理念。众深股份坚持自主研发和持续与国内高校开展产学研合作的创新模式，研发创新石化设备检测数字化服务方案填补设备全生命周期风险评估的设备特性数据支持问题，为建设数字化智能工厂提供有力支撑。众深股份首创石化转动设备相共振检测技术，有效填补国内转动设备运行检测领域的空白技术，解决大型转动设备在役质量与安全风险评估困难的问题。公司主导业务核心技术共获得发明授权 1 件、登记软件著作权 28 件、主导制定国家标准 2 项、主导和参与制定团体标准 7 项。多年来公司先后获得高新技术企业、上海市专精特新中小企业等称号并通过两化融合管理体系认证、CNAS 认可资格。

概　述

原材料质保书是材料工程师用来确保原材料质量的重要文件。它详细说明了原材料的各项性能指标、质量标准和检验方法，为采购和验收提供依据。通过签署质保书，供应商对原材料质量作出承诺，保证其符合规定要求，为产品的质量和可靠性提供保障。

原材料质保书审查大模型的用途是输入 pdf 或者照片格式的原材料质保书，利用 AI 大语言模型与专业知识库，审查原材料质保书是否满足要求，由相关人员审核确认，最后将重要数据记录归档。

通过使用模型审核人员审核质保书、记录数据可以节省大量时间，提高效率的同时减少错检率。此外，在归档后，该模型还可以快速溯源产品质保书，形成记录数据的可追溯。

需求分析

原材料质保书审查在制造厂原材料验收、监理检查、采购物资验收中都是至关重要的活动，它可以确保产品材料满足产品的设计要求、减小工程质量风险和避免合同纠纷。目前在制造厂、监理检查、采购验收过程中，原材料质保书的审核会占用相关审查人员大量时间。审查人员进行审查原材料质保书的各项参数时，需要耗费大量时间查阅标准、技术协议、设计文件去和质保书中的数据进行比对，记录数据。大量的人工查阅会造成不合格的质保书漏判，会给最终用户、采购方、生产商以及监理公司等带来巨大损失。提高原材料质保书的审查效率与精准度会给制造厂、监理单位、采购方、最终用户带来巨大收益。

案例介绍

系统利用 AI 大语言模型深度学习国家标准、行业标准、欧美标准、等常用标准，并增加项目规范等专业知识库的学习，形成具有专业领域知识的 AI 智能体。

审查人员将 PDF 或者照片格式的原材料质保书输送给系统，系统依据知识库进行自动审查，审查内容包括制造单位名称、材料牌号、规格、炉批号、交货状态、材料化学成分、力学性能、热处理状态、无损检测等关键参数是否符合技术协议及相关标准要求，并将审核依据的相关标准条目展示给审核人员，再由审核人员进行审核确认，人机双重审查。最后自动记录归档，方便日后溯源产品质量。

在 AI 模型训练方面，我们先后在书生·浦语 -20B、Llama-2 等模型上训练。为了使模型更加智能，我们通过字词标记化、单词标记嵌入、调整 AI 注意力、预训练、迁移学习等方法来使 AI 模型适应监理行业的特定需求。使其更好地适应原材料质保书审查任务，并提高准确性和效率。

目前该项目已在我公司石油化工设备全生命周期公共服务基础平台（简称 LCPSP 平台）上部署，并进行试运行，未来将邀请业内的制造厂、总包单位、业主单位参与到测试与应用中，验证实施效果。

效益分析

在成本上，大语言模型可以加快审查过程，减少人力资源，降低人力成本。审核人员可以通过使用大语言模型来审查原材料质保书，提高工作效率，节约人力资源和时间成本。

在质量控制上，大语言模型，可以准确地识别和标记质保书的问题和风险。提高质保书的准确性和可靠性，降低工程质量风险，减少质量问题带来的经济损失。

在行业规范上，大语言模型根据行业标准进行审查，在帮助审核人员确保质保书符合要求的同时，也帮助审核员快捷的阅读学习相关标准规范。这有助于促进行业规范化和提高审核人员的专业水平。

监理公司可以将大语言模型作为增值服务的一部分，为客户提供更加高效的监造服务。同时监理公司可以从质保书审查大模型入手，逐渐向监造智能化迈进。

制造厂与最终用户通过使用大语言模型有助于企业数字化建设，在原材料信息上进行数字化的补充。

智能投顾助手——光子·善策

恒生电子股份有限公司

恒生电子是一家以“让金融变简单”为使命的金融科技公司，总部位于中国杭州。1995 年成立，2003 年在上海证券交易所主板上市（600570.SH）。

恒生聚焦金融行业，为证券、期货、基金、信托、保险、银行、交易所、私募等超过 2000 家金融机构提供整体解决方案和服务。公司已连续 16 年入选 FinTech100 全球金融科技百强榜单，2023 年最新排名全球第 22 位，位列中国上榜企业第一。

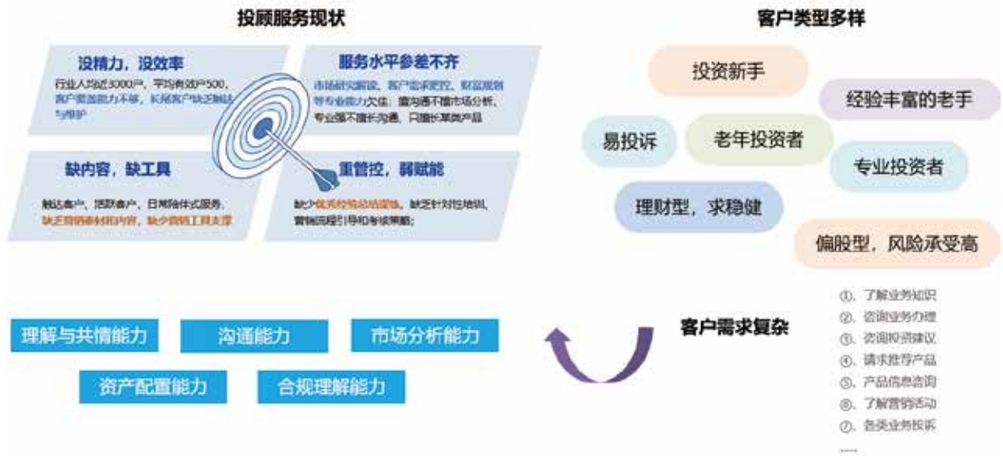
概　述

光子·善策是一款基于恒生电子金融大模型 LightGPT 开发的智能投顾助手产品，可以嵌入在 IM 工具 / 电话服务平台 / 企业微信上，帮助投资顾问根据客户会话意图分析，实时洞察客户真实需求，通过大模型强大的自然语言理解结合 NL2API、ChatDOC 等技术，实时检索话题相关行情资讯数据、产品营销物料，知识库信息，并根据客户情绪识别及意图理解，为投顾生成当下服务客户的安抚陪伴话术、专业投资咨询建议、自动匹配合适的产品和服务建议，帮助投顾专业高效服务客户，提高投顾工作效率及服务质量。

需求分析

投顾服务是金融机构为客户提供的一项专业金融咨询服务，旨在帮助客户实现财富增值和保值。然而，目前投顾服务面临着诸多挑战，如客户覆盖能力不足、服务水平不一、缺乏营销内容和工具等，导致投顾服务效率低下、专业性不足、客户满意度低下。

在客户服务任务重、客户量过载的情况下，投顾如何快速提升客户理解能力、沟通能力、市场分析能力、观点提炼能力、合规理解能力，快速响应不同客户千人千面的需求？金融科技公司需要考虑能否通过 AI 产品，把这些能力集成起来，并根据客户实际情况自动匹配相应能力，帮助投顾实时满足当前客户服务需求。



案例介绍

结合投顾实际工作流程，从 KYC 洞察客户需求、任务引导流程、会话意图识别、智能服务匹配、实时会话质检、到智能工单生成，全流程实现金融大模型智能化能力与业务服务能力的完美融合，更好地理解客户的投资目标、风险偏好、投资前提等多重因素，并根据市场动态和专业知识，为客户提供更优化、更灵活、更贴心的投资咨询服务。

恒生电子金融行业大模型 LightGPT：

LightGPT 是恒生电子自主研发的专为金融领域打造的大语言模型。基于海量金融数据训练而来，对金融相关问题的理解比通用大模型更有优势，有利于推动大模型在金融行业的应用，降低大模型的应用门槛，提升金融行业智能化水平。

- **更专业：**2000 亿中文 tokens 的加持，80+ 中文金融任务的打磨，金融多领域应用场景覆盖；
- **更合规：**学习中国的金融法律法规，更符合中国金融市场的监管要求；
- **更轻量：**支持私有化 / 云部署，支持 API 调用，推理端仅需一机两卡部署。

项目特点：

光子·善策充分利用企业历史沉淀的标准话术，提炼为话术生成的 prompt 模板，对接业务数据，再结合大模型的文本生成能力，自动生成结合客户标签，持仓状态、当前话题、市场热点又满足监管合规要求的引导话术和投顾建议。同时还能自动检索到相应的研报、资讯作为论据的支撑，增强说服力。

产品功能特性：

1、会话意图理解，根据多轮对话实时分析客户意图，精准定位客户标签。

2、智能服务匹配，自动检索话题相关金融产品和资讯信息，生成专业的观点和建议。

3、智能话术生成，实时洞察客户需求与情绪变化，给出合适的安抚话术和行动建议。

4、智能工单创建，自动总结服务记录及客户待办诉求，一键提交智能工单。

效益分析

提高投顾服务效率

通过智能生成客户标签、意图和服务话术，以及自动关联客户持仓、资讯研报、营销物料等功能，提高投资顾问工作效率。投顾可以更专注于与客户互动，而不必花费大量时间和精力在繁琐的信息搜索和整理；

提高客户满意度

光子·善策的智能化能力有助于投顾更好地应对客户焦虑情绪，提供情感支持，同时也为客户提供了更加个性化的服务体验，帮助投顾及金融公司建立更强的客户关系；

提升投顾业务转化

通过准确洞察客户需求、情绪管理，并帮助投顾迅速找到与客户需求相关产品资讯信息，提升客户被理解被关注的感受及个性化服务体验，有助于提高投顾服务业务转化。

Chapter Three.

第三篇章

3 大模型服务

2023

大模型落地应用案例集

Foundation Model
Practical Application Collections

支小助 - 大模型金融专家智能助理

蚂蚁星河（重庆）信息技术有限公司

蚂蚁星河（重庆）信息技术有限公司成立于 2017 年 4 月 19 日，是蚂蚁集团的全资子公司，专注服务小微企业。基于蚂蚁集团在人工智能、数据库、隐私计算、智能风控、区块链等领域自研的关键技术，蚂蚁星河聚焦在金融数字化、科技创新应用、乡村振兴等方向，为金融行业提供更优质的技术服务，助力数字经济高质量发展。

概　述

伴随大模型技术的快速发展，已逐渐应用于金融行业的理财投资、保险双核、信贷风控及数字员工等多元化典型金融场景，例如作为金融从业人员的智能助理，可全面提升金融从业人员的服务半径和服务效率。然而，由于金融行业的专业性、严谨性、合规性等特点，难以将大模型直接应用到金融场景中，需要结合金融领域知识和数据进一步训练和升级。

基于金融场景大量实践，蚂蚁以大模型为认知和交互中枢，调用领域知识和专业服务，形成“大模型 + 知识 + 服务”的架构范式，构建出面向金融行业专家的智能助理 - 支小助系列，目前已在蚂蚁集团的理财、保险、营销业务上全面推广应用，同时与蚂蚁金融生态合作伙伴开放共建，加速金融大模型对产业的赋能，带来金融服务新体验，创造金融产业新效率。

需求分析

在金融专家 AI 助理的金融场景中，通过大模型的智能化升级，可以有效解决金融信息过载、复杂金融任务拆解、专业术语晦涩、单人工作效率低下等问题。从需求角度看，一是底座大模型需结合金融领域知识和数据进一步训练和升级，形成金融行业大模型，让大模型对金融行业的特有的术语和规则有更深刻理解，输出结果满足金融领域极高的准确性、可解释性、合规性及金融价值观等要求；二是需有专业、完善的金融大模型评测基准，能够对金融大模型的性能、安全合规性进行有效评估；三是结合多种角色的金融专家的业务需求把金融大模型进一步产品化，形成专业高效、便捷易用的智能助理。

案例介绍

支小助系列产品是助力金融专家提升生产力的典型应用，具体包含工具如图 1 所示：

- **支小助投研版：**基于蚂蚁金融大模型的知识力和专业力，结合先进的量化投研分析系统和工具，提供给金融专家的生产力工具。支小助投研版已在蚂蚁内部多个场景投产，实测数据表明，其每日可辅助每位投研分析师高质量地完成超过 100+ 篇研报和资讯的金融逻辑和观点提取，完成 50+ 金融事件的推理和归因，并将典型的量化分析任务的效率从天级别提升到小时级别，带来了明显的生产力提升。

- **支小助服务版：**为理财师、保险代理人两类典型业务销售顾问型专家提供的专家助理工具，围绕金融专家一天的日常工作流程，提供全面 AI 赋能。该类生产力工具在蚂蚁内部已全量推广，已经看到非常好的效率优化效果：更精准的线索挖掘和任务调度，更强大的意图识别、话术推荐和专业服务工具，让服务更专业更高效。

- **支小助理赔专家版：**理赔是保险业务的核心服务，也是专家密集型的环节。在大模型加持下，保险理赔正在全面走向自动化。医疗险理赔，用户平均要提交 11 张医疗单据，基于海量图文预训练的多模态大模型，复杂单据的整案提取准确率由 80% 提升至 98%。同时我们基于“大模型 + 知识”双驱动的架构，险 / 医 / 药 / 病知识基础上做 COT 推理，让核赔决策在 98% 准确率的前提下，覆盖率从 40% 提升至 70%。支小助让大部分的门诊险理赔，超过 30% 的住院医疗险理赔，完全自动化，这个比例还在快速提高，让更多用户享受极致的秒赔体验。

- **支小助营销专家版：**为运营人员定制的营销辅助工具，通过提示就能生成丰富的营销创意、产品介绍、图文和视频物料，并做到在全渠道进行智能投放，为用户推送千人千面的营销表达。让营销专家享受生成式 AI 带来的效率变革。更精准的供需理解和匹配，更丰富的服务供给，更有效的营销表达。

图 1　支小助系列产品工具

支小助系列产品的底座支撑是蚂蚁金融大模型，基于金融场景大量实践，以大模型为认知和交互中枢，调用领域知识和专业服务，形成“大模型＋知识＋服务”的架构范式，可以为全链条金融业务提供“知识力”、“专业力”、“语言力”，以及结合可信围栏技术实现的“安全力”，如图 2 所示。基于金融专属任务评测集 Fin-Eva 的专业评测，蚂蚁金融大模型在金融场景“认知、生成、专业知识、专业逻辑、安全性”等五大维度 28 个金融专属任务中大幅超过主流通用大模型，在“研判观点提取”，“金融意图理解”，“金融资讯理解”等领域接近或者超过平均人类专家水平。

图 2　蚂蚁金融大模型

效益分析

支小助系列产品目前已在蚂蚁集团的投研、理财、保险、小微金融、营销等一系列业务上全面推广应用，惠及蚂蚁数字金融业务的数千名金融专家，提效成果普及支付宝 APP 上的亿级客群。对于理财顾问和保险顾问，支小助服务版可以大幅提升其服务半径（+70%）；对于分析师，支小助投研本可以大幅提升其研究效率（研判半径 x10，量化编码效率 x10）；对于保险理赔业务，可以创造“秒赔”用户体验；对于金融营销专家，支小助的 AIGC 能力可以提供多种模态（图文、视频等）的内容生成，大幅提升内容生产效率（x10）。

外部合作方面，目前已与多家蚂蚁生态金融机构（如帮你投、恒生聚源等）开展了业务合作。未来，蚂蚁将陆续开放支小助的 SasS 版、开源版和商业版等，进一步支持目前与蚂蚁集团合作的数百家金融机构的金融专家，助力他们实现金融业务助理的智能化升级。

AGI 云上模型服务平台

优刻得科技股份有限公司

UCloud 优刻得是中立、安全的云计算服务平台。自主研发 IaaS、PaaS、大数据流通平台、AI 服务平台，推出公有云、私有云、混合云、专有云等全线云产品，为政府、AI 大模型、工业互联网、运营商、教育、医疗、零售、金融、互联网等各行业用户，提供全面的数字化转型升级服务。

2020 年 1 月，优刻得正式登陆科创板（股票简称：优刻得，股票代码：688158），成为中国第一家公有云科创板上市公司，也是中国 A 股市场首家“同股不同权” 的上市企业。

优刻得在全球设有 31 个可用区，遍及国内、东南亚、欧洲、北美、南美、非洲等 25 个地域，拥有内蒙古乌兰察布、上海青浦两大自建数据中心，构建了云网融合、安全稳定、智能敏捷、绿色低碳的数字底座。此外，优刻得在北、上、广、深、成等多地建有线下服务站，已为全球超过 5 万家企业级用户提供云服务。

概　述

UCloud 基于强大的算力底座，以及自身在 AGI 应用上的创新实践，打造“模型 + 服务”的 MaaS 服务新范式，为客户提供可靠、高效、安全的 AGI 云上模型服务平台。面向各行业用户，提供一站式的模型管理、训练、部署等全流程服务。平台具备强大的数据处理能力，支持海量数据的存储和计算；提供的丰富工具和算法，支持模型定制及调优；采用弹性计算资源，可根据用户实际需求进行灵活扩容。

平台具备四大核心技术能力。一是模型训练能力，基于 UCloud 上千卡训练集群规模，可提供百亿、千亿级大模型综合训练能力。二是模型传输能力，基于训练集群服务器的定制网卡，具备高速网络传输能力。三是模型推理能力，基于 UCloud 上千卡推理集群规模，可承接客户各项推理服务。 四是模型适配能力，平台已适配市场主流大模型，且可根据行业需求对模型进行定制。

AGI 云上模型服务平台可广泛应用于公文写作、金融咨询、医疗咨询、泛娱乐等领域。

需求分析

一、相关背景

随着大语言模型的蓬勃发展，AGI 应用获得广泛突破，智能语音、图像识别、自动驾驶、翻译等领域取得长足进步，AGI 正在深入人类生活各个方面。如何高效利用算力资源快速建立 AGI 应用并保证其平稳运行，已成为业内共同挑战。

基于 10 多年的公有云技术积累和系统工程建设能力，UCLoud 面向大模型企业快速推出了 AIGC 算力解决方案—“AGI 云上模型服务平台”，并提供“训练专区 + 推理专区 + 存储专区 + 管理专区”的分区方案，构筑安全、可靠的大模型算力底座。

二、行业用户需求

- **数据标准化整合：**数据来源广泛且零散，需耗费大量时间组织建模，相关数据包括但不限于研报、电话语音、微信文本等。
- **信息抽取，概要总结：**领域所在信息需进行大量基础处理工作，且核心指标在传统软件中无法查询。
- **特定行为模型学习：**针对领域所在的用户、场景等对象需进行基本信息、属性、习惯、周期等进行关系网络图谱构建，为后续业务拓展提供数据参考。
- **安全合规要求：**随着《网络安全法》等各项法律政策的出台，对数据安全有了明确的合规要求，AIGC 应用的广泛推广也面临了内容安全的风险。对数据隐私保护和内容合规性是首要前提。

案例介绍

一、主要能力

- **丰富的算力能力：**依托 UCloud 云平台，提供 A800 训练集群和 V100S/T4 推理集群，提升训练和推理服务能力。
- **高速的传输能力：**提供 4*200Gb/s 的高速网络，扩展训练数据交互能力，从而推进训练任务的效率。
- **充沛的存储能力：**提供高 IO 的热存储服务和大体量的冷存储服务，提升了训练任务所需的容量及吞吐能力。
- **便捷的应用能力：**构建 AGI 应用服务平台，适配各开源和定制算法模型，并清洗和标准化业务数据，从而建立知识图谱应用能力。

图 1　UCloud-AIGC 解决方案全景图

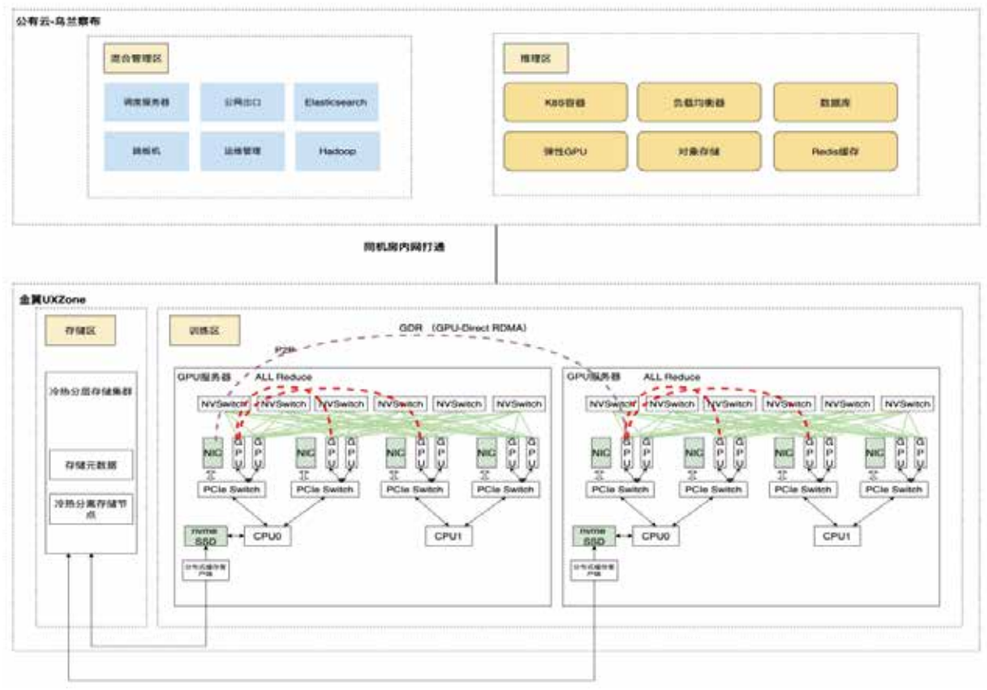


图 2　UCloud-AIGC 区域建设图

二、技术创新点

- **高速化网络：**依托 DMA 集成引擎和物理网卡多端并行协作能力，实现 RDMA 网络运转；不仅实现了 Tb/s 级别的聚合网络能力，且性能与 IB 网络相当，并且充分解决扩展性问题，极大兼容了各类信创设备和环境。

- **高效化存储：**通过 GPU Direct Storage（GDS）技术，可直接跳过 CPU 和系统内存，通过网卡直接访问远端存储，降低对 CPU 的占用，减少访问时延，大幅提升数据 IO 吞吐。

- **模型镜像化：**通过 UCloud 内部的镜像化系统，直接将各类模型、特色工具等嵌入至云平台镜像，从而让 NVIDIA 驱动环境的预装从 3-5 个工作日缩减为秒级，效率上升。

三、实施效果

- 客户模型训练时长缩短 20%
- 客户模型推理效率提升 35%
- 客户模型微调精确度达到 80% 以上【针对百亿模型】

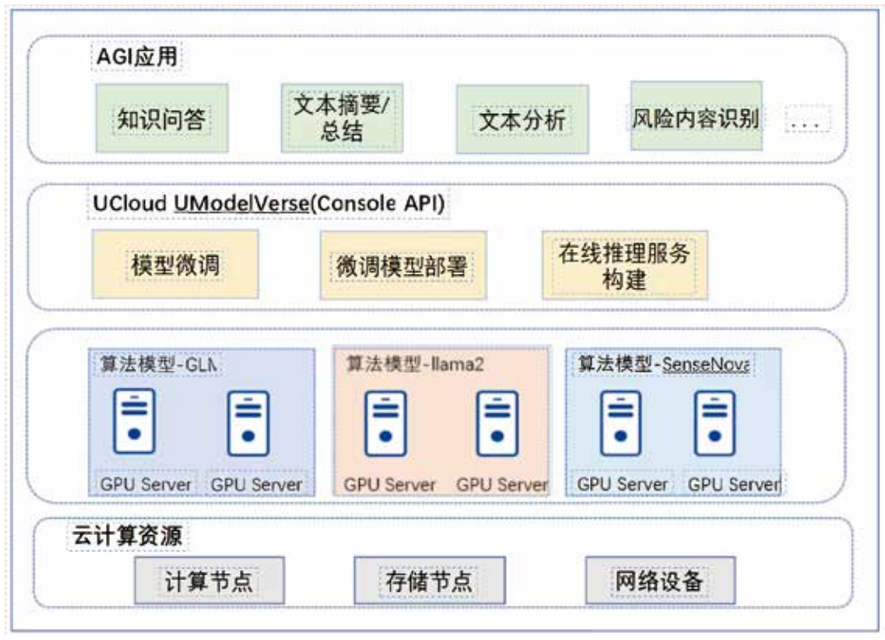


图 3　UCloud-AGI 服务架构图

四、应用落地及推广情况

- **公文写作：**通过全量动态归集政务部门业务知识，综合集成相关智能算法、模型、工具、应用和领域生态，构建辅助公文写作的 AIGC 应用平台，全面覆盖我国行政系统在特定试用范围内的公文体系。同时，为保障公文内容的合规性，平台构建了大规模的合规审查数据库，用于训练合规审查模型，对公文撰写中的风险内容进行识别，确保公文内容安全合规。对涉政等敏感词汇，以及字词、标点、语序的行文错误能够自助审核校对。

- **金融咨询：**在金融场景下，每天都会产生大量的市场相关信息，例如投研机构、社评、企业动态、资深投资员发言等。内容格式包括不限于发布文稿、微信聊天记录、新闻动态、网络短文等，针对这些大量信息，通过学习高质量的信息提取或者文本概要数据集，定制化私有大模型，学习金融行业特定的信息提取模式，生成精通金融信息概括的私有大模型，尽可能精准的提炼出投研相关的信息。

- **医疗咨询：**在医疗咨询行业，医疗营销或管理系统需要明确捕捉到这些信息里的有效内容，并带动下游流程。有效信息例如：xxx 科室 xx 药物余量不足，需要补货。针对这类场景，通过学习医疗行业的信息提取模式，定制私有化大模型，提高信息提取的有效性。

- **泛娱乐：**游戏：游戏任务、剧情设计，创造游戏人物的背景故事、世界观和 NPC 互动文案；智慧数字人：虚拟主播，可面向线上授课、电商直播等领域。

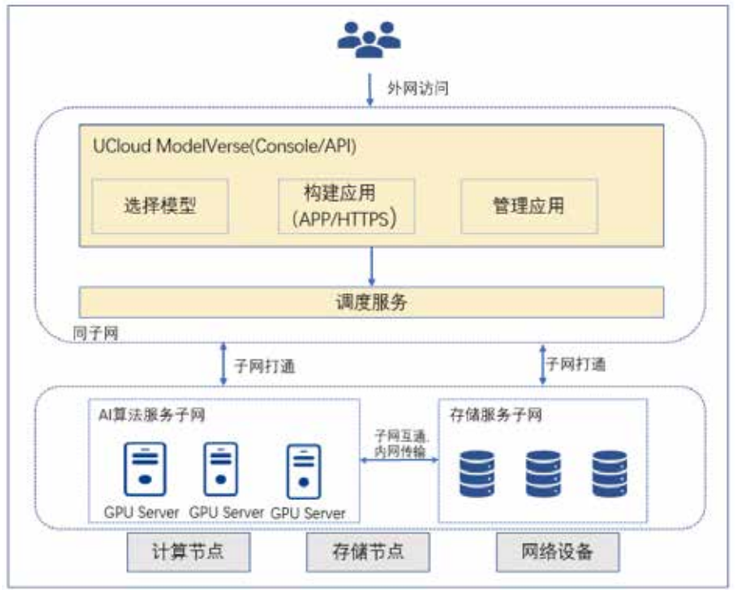


图 4　UCloud-AGI 调用逻辑图

效益分析

一、经济社会效益

- **信息互通互融：**UCloud AGI 云上模型服务平台作为一体化、综合化的多模态服务，不仅提升单一模态带来的局限性，也打破各区域的数据壁垒，从而加速跨领域的业务融合。

- **AGI 产业孵化：**通过支持适配各种定制模型，可协助各中小模型公司、垂直行业公司进行应用服务孵化。

- **数智化改造：**通过 AIGC、大数据、云计算的整体化建立健全，可促进传统行业信息化 - 数字化 - 智能化改造，促进行业下的企业高质量发展。

二、商业模式

- **智算云计算服务：**提供包括但不限于 A800、H800、V100s、T4 等训练和推理 GPU 云主机，为客户提供高算力支持。

- **智算云存储服务：**提供 GDS 高 IO、高吞吐的云存储服务，提升数据加载速率，且可支持单模块私有化部署。

- **智算云应用服务：**提供数据管理、模型选型、模型训练、模型推理等 AGI 应用服务，除 API 形式提供外，亦可支持单模块私有化部署。

- **智算专有云服务：**综合性服务，全量助力客户商业化转型。

- **其他服务：**提供模型微调、网络优化、业务代维等人工服务。

三、应用推广前景

1、通用场景：基于互联网沉淀的体量庞大、多元异构、对外开放的信息，不断进行学习和整合，从而建设更为体系化、全面化的常识图谱，提升用户的模型深度学习的计算效率。

2、垂直场景：基于行业痛点和客户场景，拓展客户数据模型深度，从而形成更为精细化、更为专业化的模型。

蚂蚁集团大模型数据高质量供给平台

蚂蚁科技集团股份有限公司

蚂蚁集团起步于 2004 年诞生的支付宝，经过十八年的发展，已成为世界领先的互联网开放平台。蚂蚁集团通过科技创新，助力合作伙伴，为消费者和小微企业，提供普惠便捷的数字生活及数字金融服务；持续开放产品与技术，助力企业的数字化升级与协作；在全球广泛合作，服务当地商家和消费者实现“全球收”、“全球付”、“全球汇”。蚂蚁集团的业务板块包括数字支付开放平台、服务业数字化经营开放平台、数字金融开放平台、数字科技服务、国际跨境支付服务。作为一家技术人员占比超过 60%，拥有强大自主创新能力的科技企业，蚂蚁集团始终坚持自主创新，在人工智能、数据库、隐私计算、智能风控、区块链等领域进行了前瞻性布局，自主研发了大模型、隐语、OceanBase 数据库等一系列支撑蚂蚁和行业发展的关键技术。

概　述

大模型的训练，除需要强大的算力支撑外，高质量数据集的构建和处理对于大模型的性能表现也至关重要。高质量不仅指数据规模要庞大、还需要数据来源丰富（覆盖多种类型、多个领域），同时还需对数据进行有效的预处理（质量过滤、数据去重、隐私脱敏等），才能保障训练出的大模型能够尽量符合 3H(helpful、harmless、honest) 原则，提供安全、可靠、可信的模型输出。

但构建大规模数据集并进行高质量预处理是一个成本极高且费时费力的过程，并需要大模型能力升级持续开展。对大模型企业来说，大模型数据的高效高质量供给既是其核心竞争力之一、也是其重要成本负担来源。

本项目设计实现了一个大模型数据高效高质量供给平台，不仅可降低数据获取和使用成本且保证来源合规，并能够有效提升数据质量、过滤风险数据保障训练安全。目前该平台已作为蚂蚁集团内部统一的数据平台，为集团内基础大模型及各业务线的行业大模型所使用。

需求分析

从数据平台使用方的视角来看，数据平台需要提供的关键功能包括：

一、能够扩大数据来源和规模，可以较低成本获取大规模的训练数据，且获取过程及获取内容安全合规；如构建互联网公开数据站点黑白名单、识别出高质量真实数据源等；

二、对于训练数据要能够进行高质量的处理和及时更新；如隐私数据识别、NSFW（毒性）数据识别、政治敏感数据识别等过滤风险数据等。例如网络数据质量往往很差、格式各式各样，需要统一的数据过滤清洗能力以提升训练数据质量等。

案例介绍

本项目中所建设的大模型数据高效高质量供给平台主要包括“数据获取”、“数据处理”、“数据管理”三个方面，可以覆盖大模型数据从获取到加工、使用的全数据流程，为大模型提供持续的高质量数据供给。具体如图 1 所示。

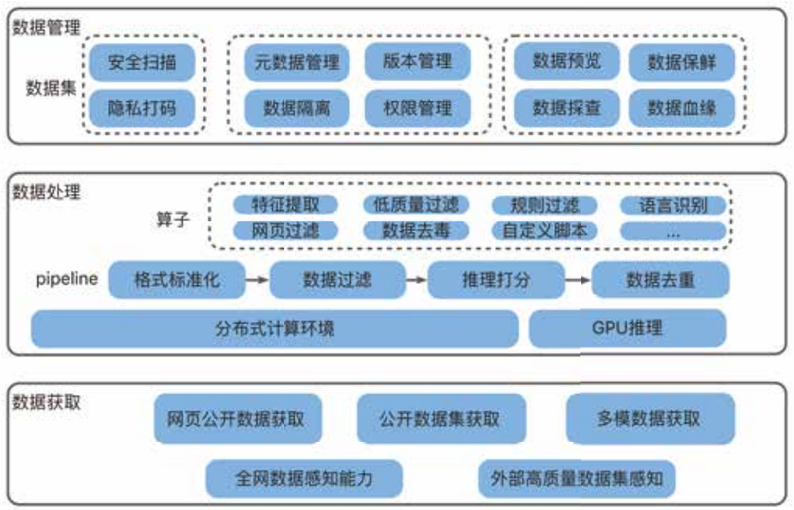


图 1　大模型数据高效高质量供给平台示意图

- **数据获取：**可从全网范围内获取大模型训练需要的数据，解决大模型训练数据量的问题。已支持包括公开网页数据提取、专业社区公开数据集下载、多模态大文件下载等不同源的数据获取，且在持续更新。

- **数据处理：**可提供文本、图片等多种模态数据的处理算子，集成去毒、去重、去隐私等大模型训练数据的高效预处理能力，并支持定义统一的数据集格式，帮助业务快速高效的构建高质量数据集。

- **数据管理：**对标行业领先平台如 Huggingface，提供统一的数据集平台托管能力，并提供数据探查能力来评估数据集数据质量，同时还提供安全扫描等合规机制保障大模型训练数据安全。

主要创新点

- **标准化预处理链路：**解放 ETL 人力，提高数据集构建效率，同时打通数据集与原始采集数据，支持持续构建数据集供给大模型训练；

- **数据集探查：**基于标准数据集格式定义，结合数据分析能力，从元数据、数据特征等多种维度对数据集本身的质量（领域丰富度、质量特征分、毒性、真实性 / 可信来源站点数据占比、及时性）进行评估，同时在小模型上进行 AB 实验，评估数据集对模型效果影响；

- **数据保鲜：**基于模型评估结果及训练数据集质量探查，自动决策引入数据来源与预处理链路参数，调整数据集不同维度数据占比，促进模型效果提升。

效益分析

高质量的训练数据构建和处理是大模型训练的基础，也是大模型能力持续提升、更为安全可控的关键。目前行业中已经提出“Data-Centric AI”的概念，即在模型相对固定的前提下，通过提升数据的质量和数量来提升整个模型的训练效果。根据一些行研机构数据，数据成本可占到大模型训练成本的 20%~35%，且未来数据成本在大模型开发成本占比还将持续提升。

本项目所开发的数据高效高质量供给平台，沉淀了大模型训练数据处理的最佳实践，目前已广泛用在蚂蚁集团内部的基础大模型和多个行业大模型的训练数据构建过程中，不仅避免了数据预处理重复构建带来的浪费，而且统一提高数据处理的质量水平和安全水位，保障了蚂蚁集团大模型产品的训练质量。该统一数据平台的建设方案示范性强，可供大模型的开发和应用企业参考实施。未来该平台有望对外开放，进一步推广应用到更多机构，给行业带来更大的降本提效价值。

基于大模型的壹沓数字员工超自动化平台

壹沓科技（上海）有限公司

壹沓科技于 2017 年正式运营，作为全球领先的数字机器人公司，聚焦前沿技术创造壹沓数字员工超自动化平台，赋能用户实现业务超自动化，构建未来数智世界。壹沓科技在数字机器人、人工智能及大数据挖掘等相关领域具备独创自研的核心技术，聚焦大供应链领域（生产制造、物流配送及新零售）为数千家客户提供产品及服务，并推出基于大模型的新一代壹沓数字员工超自动化平台。

概　述

某头部供应链企业面临着复杂的企业管理和大量客户需求，大语言模型在帮助客户问题中发挥了重要作用。过去，该企业依赖人力和传统数字技术进行企业内部管理和供应链管理，但随着业务规模的扩大，这些方式已无法满足需求。因此，解决方案专家团队引入了大模型技术，以提升运营效率、人才密度和服务质量。

需求分析

当前，大模型技术涌现和应用场景的不断落地，正在全球掀起新一轮生产力变革，为各行业带来生产力跃迁。在供应链领域，AI Agent 数字员工正在被越来越多供应链企业在各场景应用，未来的办公室工作形式会发生根本性的变化，人在其中只需要下达指令，提供资源，监督结果，大多数工作都可以交给数字员工 AI Agent 来完成。据《智慧供应链白皮书 - 数智世界·链通全球》白皮书报告显示，中国智慧物流市场是一个新兴的万亿级市场，2022 年已达 6995 亿，2023 年将达近 8000 亿，并以 20% 年复合增长率持续增长。当前，复杂多变的外部贸易环境和行业竞争加剧等多重因素叠加影响，导致供应链企业面临着众多不确定性：经营成本日益高企、业务效率不高、数据孤岛林立，加快智慧供应链供给侧变革成为企业数字化转型的必答题。

案例介绍　某头部供应链企业大模型案例项目介绍

主要能力

- 大模型 + 深度行业 Know-How

自然语言处理能力：大模型具备强大的自然语言处理能力，可以理解和处理各种语言的文本数据，自动抽取关键信息，提升供应链信息处理的准确性和效率。

文本生成与对话能力：大模型可以根据用户需求通过 Agent 智能代理和 RAG 检索增强生成为客户提供高质量文本回复和精准的行业答案。

技术创新点

基于大模型的壹沓数字员工超自动化平台通过自然语言式对话，为客户提供开箱即用的 Agent 数字员工和精准的行业答案，为供应链企业量身打造虚拟数字员工专家团队，资深供应链运价经理、物流可视追踪经理、供应链新人成长师、行业案例专家行业翻译大师、行政问答助理等，协助白领员工完成各类数字化工作，让人聚焦创意、决策等高价值工作，从而为企业创造巨大价值。

实施效果与应用落地情况

- **提升运营效率：**通过大模型的应用，大幅提高企业运营效率和人才密度，同时，企业实现了供应链运营的自动化和智能化管理。
- **提升客户服务水平：**基于大模型的数字员工超自动化平台，企业可以更好地了解客户需求，帮助企业实现从营销到履约到财务结算的全局超自动化，赋能企业业务模式变革及提效，推动企业生产力跃迁。
- **降低成本：**通过大模型的应用，企业实现了供应链自动化和智能化管理，极大降低运营成本。

效益分析　某头部供应链企业案例的效益分析

一、经济社会效益

壹沓科技大模型在某头部供应链企业的应用，带来了显著的经济社会效益。首先，通过提升运营效率，企业降低了运营成本，提高了经济效益。其次，基于大模型，通过提供个性智能化的供应链服务，企业提高了客户服务水平，增加了客户黏性和忠诚度，为企业的长期发展奠定了良好的基础，为企业的可持续发展提供了有力支持。

二、商业模式

在大模型技术的助力下，某头部供应链企业建立了全新的商业模式。通过运用大模型技术自动化和优化业务流程，企业实现了高效的数智供应链运营管理，并为客户提供个性化的优质服务。这种商业模式具有较高的创新性和独特性，能够帮助企业在激烈的市场竞争中获得优势。

三、应用推广前景

大语言模型在某头部供应链企业的成功应用，为其他供应链企业提供了有益的参考。随着大模型技术的不断发展，越来越多的供应链企业将引入大模型等先进技术，以提高运营效率和服务质量和企业人才密度和人效。因此，大模型的应用推广前景广阔，具有广泛的市场应用价值。

云原生大模型知识库平台

上海道客网络科技有限公司

上海道客网络科技有限公司（DaoCloud），2014 年 11 月成立，总部上海，成都、北京、深圳、新加坡等地设有分公司，是上海市高新技术企业、国家级专精特新“小巨人”企业。公司总人数超过 350 人，研发人员近 70%，目前已完成 D+ 轮融资。DaoCloud 是为数不多的从事系统和底层软件开发的领军企业，公司自主研发的云计算操作系统，为企业数字化转型提供底座支撑，是国内基础软件方面的重大突破，广泛应用于金融科技、工业互联网、人工智能等多个领域，客户包括交通银行、浦发银行、上汽集团、屈臣氏等各行业的头部公司。同时，DaoCloud 在全球云原生开源社区具有重要的影响力，云原生开源贡献率国内第一、世界领先，在 Kubernetes、Docker 等云原生核心技术中，贡献度进入世界前三，仅次于谷歌和红帽。

概　述

针对企业面临的信息分散、知识更新不及时等问题，DaoCloud 为其设计并实施了一套全面的企业知识库平台，帮助企业快速构建属于自己的垂直领域的专业 AI 助手。知识库平台是一款基于云原生和分布式计算框架的能力下，使用了向量数据库和大语言模型的强大知识管理和检索工具，允许企业把多样化的企业知识纳入统一管理，并在其之上构建具有企业特色的知识问答系统，提供高效的知识检索和问答服务，以帮助企业更好地利用知识数据去假设各自企业和个人需要的垂直领域的 AI 助手。

在实际应用中，该知识库实现了 GPU 资源的统一纳管，有效减少 IT 环境的复杂度和管理成本；并依托动态资源分配，为用户提供了安全的、隔离的私有大模型体验，使 IT 业务的响应能力有极大的提升；通过开放的架构让平台的可扩展性大大提升，新的大模型的集成和纳管简单易行。

需求分析

在数字化时代，企业面临着激烈的市场竞争和快速变化的技术环境，如何有效管理和利用知识资源成为关键问题。传统的知识管理方式存在信息孤岛、知识更新不及时、员工获取知识困难等问题，严重影响企业竞争力。为满足这些需求，DaoCloud 基于云原生和分布式计算框架，利用向量数据库和大语言模型的强大知识管理和检索工具，构建了知识库平台。该平台帮助企业统一管理多样化的知识，提高工作效率，激发创新能力，实现可持续发展。知识库平台支持多种通用大语言模型，分别是本地模型和在线模型，如 ChatGPT、讯飞星火、ChatGLM3、Qwen 等，并支持基于大语言模型构建应用，解决用户在特定领域的问题。

案例介绍

云原生大模型知识库平台旨在帮助企业解决信息孤岛、知识更新不及时、员工获取知识困难等问题，提高工作效率，激发创新能力，实现可持续发展。

在技术创新方面，DaoCloud 知识库平台是一款基于云原生和分布式计算框架的智能问答解决方案，支持多种大语言模型，可构建智能问答应用，配置个性化语料库，提供准确的问答能力；同时，它还提供可扩展的 GPU 能力和向量数据库，实现高效的文本向量化与存储，并采用相似度匹配提升查询准确性；另外，通过基于 Ray 的分布式推理，实现与用户的快速互动，使应用与用户交互更加迅捷，图 1 为知识库平台的业务框架。

图 1　知识库平台的业务框架

- **计算框架：**主要是 AI 底层运行的 Ray AIR 相关的 lib 以及 Ray Core 的分布式计算框架的能力。
- **调度增强：**围绕 Ray 的云原生能力，使用 KubeRay 的能力运行向量模型和 LLM。
- **云原生平台：**承载了所有运行时的容器环境。
- **算力服务层：**算力统一纳管平台，提供了一站式管理。
- **基础设施：**主要体现底层物理资源的能力，包括 AI 场景中很重要的 GPU 资源，以及依赖的存储和网络等物理资源。

图 2 为知识库平台的技术架构

图 2　知识库平台的技术架构

- **知识库 UI：**提供用户访问所需的客户端；
- **业务层：**包括有 AIGC-App、AIGC-Fast、AIGC-Manage 等自研组件和 Nacos、Gateway、Mysql、Redis 等外部组件；
- **接口层：**通过 AIGC knowledge Engine 来提供统一的接口。
- **核心层：**提供 Embedding Ray Service 和 QA Ray Service 两大服务，都基于 Ray 的分布式计算框架，多组服务并行。
- **框架层：**由 KubeRay 和容器集群组成。

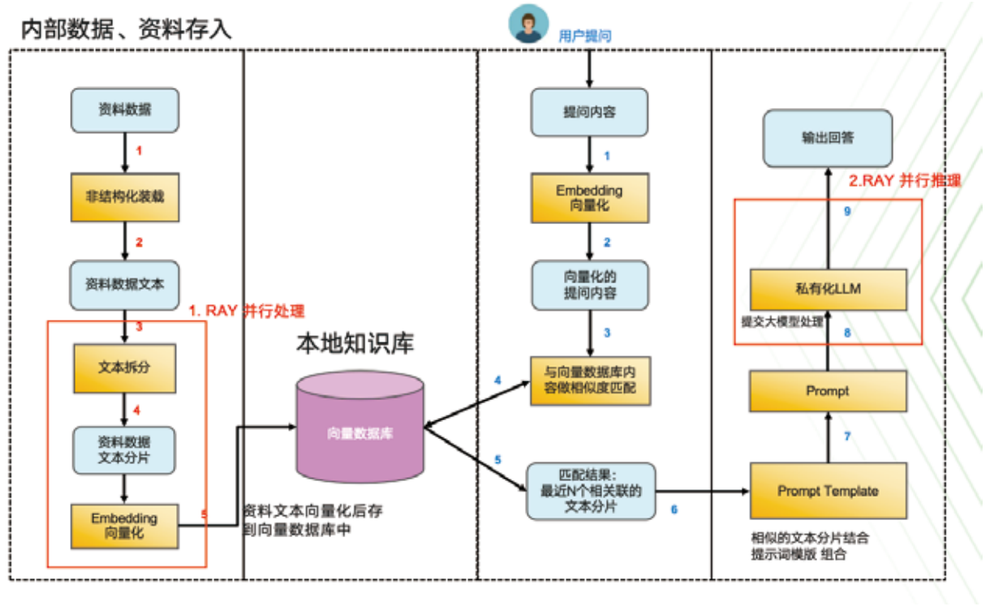


图 3　知识库平台的业务流程图

图 3 为知识库平台的业务流程图，用户将内部数据或资料输入知识库平台后，平台会将其分成块，并将它们全部嵌入，得到表示该数据的嵌入向量。然后，将这些向量存储在向量存储中。在使用时可以对向量存储进行查询，以获取与用户的任务相关的可能块。然后，将这些块填充到提示中，并生成结果。

效益分析

云原生大模型知识库平台提供多个大语言模型基座，用户可以根据偏好和需求定制 AI 应用，满足用户不同领域和场景需求；平台具有强大的语意理解能力和智能回复能力，能够提供高质量交互体验；平台可支持定制个性化的私人语料库，满足用户对特定领域的知识需求。通过知识库平台，将企业多样化的知识统一管理，具有辅助办公能力，有效提高企业用户办公效率。

在应用推广前景方面，随着互联网技术的普及和人工智能技术的快速发展，越来越多的企业开始重视知识管理和信息共享的重要性。因此，云原生知识库平台有着广泛的市场需求和良好的发展潜力。

众调科技：营销 AI 培训产品

阿里云计算有限公司

阿里云创立于 2009 年，是全球领先的云计算及人工智能科技公司。阿里云为 200 多个国家和地区的企业、公共机构和开发者，提供安全、可靠的云计算、大数据、人工智能等产品和服务。经过十三年的发展，阿里云已成为全球前三的云服务提供商。

阿里云是全国首家云等保试点示范平台和首家通过国家等保四级备案测评的云服务商。为中国超过一半的上市公司，为 80% 中国科技创新企业提供云计算服务。

在后疫情时代，社会经济的方方面面都在全速重构，阿里云正在逐渐成为赋能数字经济的数智创新平台，为数字经济、数字社会、数字政府提供创新价值。

概　述

营销 AI 培训产品主要针对行业中对新人培训的场景需求，目前案例落地场景为汽车行业的营销培训场景。

汽车行业每年都需要花费大量的费用通过线下培训教学的方式对入行新人进行培训，效率低，成本高，培训效果难把控。针对此痛点，营销 AI 培训产品以大语言模型为核心技术方案，提供新人知识学习、销售场景对练、销售知识问答功能。AI 可以结合实际业务需求扮演不同的客户群体与营销新人进行营销场景对练，对练效果接近实战，不断提升人员营销能力。同时大模型可以对营销新人对练的效果进行分析评分，不仅可以让新人了解自身能力缺陷并通过 AI 问答进行针对性学习提升，同时方便管理人员了解新人营销能力水平，为培训管理提供有效抓手并省下大量的线下培训费用。

需求分析

汽车行业营销人员对销售业务的熟练程度极大影响汽车的销量，车企长期需要投入大量的费用对销售人员进行培训，但是因为以下几点原因导致目前的培训费用无法得到有效的控制和利用：

- **讲师线下集中培训的效率低下；**
- **传统培训方式难以做好考核，同时需要花费大量的考核管理成本；**
- **产品更新需要定期大范围进行线下培训；**
- **一线人员流动性高，加剧培训投入的成本，降低培训的效果。**

为此，汽车厂商亟需一种解决手段，能够解决当下人员培训碰到的问题，需要满足成本可控、培训效率高效、培训课程易推广、培训效果可估量，同时培训产生的知识点可沉淀及重复利用。

案例介绍

与上汽通用汽车合作的 AI 培训创新项目主要提供了以下能力：

- **大模型根据不同销售场景模拟不同的客户群体与销售人员进行真实销售场景对练；**
- **大模型根据不同场景、不同车型产品的知识点对销售人员的回复进行能力评价；**
- **大模型提供品牌车型知识、销售技巧及异常应对话术的问答功能。**

在技术上，创新地引入大语言模型并通过微调训练赋予特定场景下的人格，通过 Agent 技术方案使得大模型具备角色扮演并完成特定考核任务的能力，在于销售人员对练过程中，均有较好的闲聊、任务对练、话题拉回和追问的能力，表现接近真实客户场景。

AI 智能问答功能基于 LangChain 技术框架及大语言模型，通过使用汽车行业专业语料对大模型进行 LoRA 微调，使得大模型对汽车行业的行话、问题理解能力及答题能力均有显著提升。

通过 AI 赋能，整体培训功能提供了优秀的人机交互效果；销售人员在使用过程中，更加贴近实战效果，可以帮忙销售人员无缝衔接现实业务场景；同时对于管理者，能有一个完整精确的统计报告，实时了解销售人员整体营销能力，适时调整培训计划。

目前此创新项目在客户内部处于 Pilot 推广阶段，反馈较好。

效益分析

经济效益

营销 AI 培训产品的投入使用，能够减少车厂在培训方面的投入，具体包括培训老师的投入、集中式培训场地的投入、一线人员流动产生的重复培训的投入，粗略估算能够减少新员工现有三分之一的培训成本。

商业模式

营销 AI 培训产品支持 SaaS 服务或私有化部署的方式，可以按照使用人员账号数量和一次性开发进行收费。

应用推广前景

营销 AI 培训产品基于合作单位众调科技有限公司多年的一线培训数字化经验总结出抽象出的大模型应用产品，在汽车经销商培训场景中适用于各个品牌；具体实施过程中通过配置平台可进行灵活配置，配合少量的定制化研发即可快速上线投入使用，市场潜力巨大。

信息安全大模型平台

中企网络通信技术有限公司

中企网络通信技术有限公司（简称“中企通信”或“CEC”）是中信成员企业，是智能科技驱动的数字化赋能者，致力推动企业创新，凭借丰富资源，行业经验与数智通信（DICT）专业知识，打造了多元化的数智技术解决方案，推动企业数字化转型。凭借扎实的全球市场经验、多行业应用、客户实践以及技术实力，结合广泛的 ICT 资源覆盖、专业的本地化服务、优质的技术解决方案，中企通信成为企业可信赖的数智通信服务伙伴。

概　述

ChatGPT 现象级的流行，给信息安全智慧模型开发带来了巨大影响和冲击，然而通用大语言模型在信息安全领域的开发及应用，面临着专业数据和知识壁垒、新型攻击和威胁快速出现、数据隐私和安全性问题等诸多挑战，因此中企通信在通用大语言模型的基础上，融合专业的信息安全技术知识、丰富的服务经验及海量价值数据，为企业客户打造了信息安全专业大模型，并且可以结合企业实际情况进行针对性训练，通过云端训练 + 边缘推理协同架构，将模型私有化部署在企业内部，在利用大语言模型识别、预警及防范信息安全风险的同时，保证了用户数据的安全及隐私。

需求分析

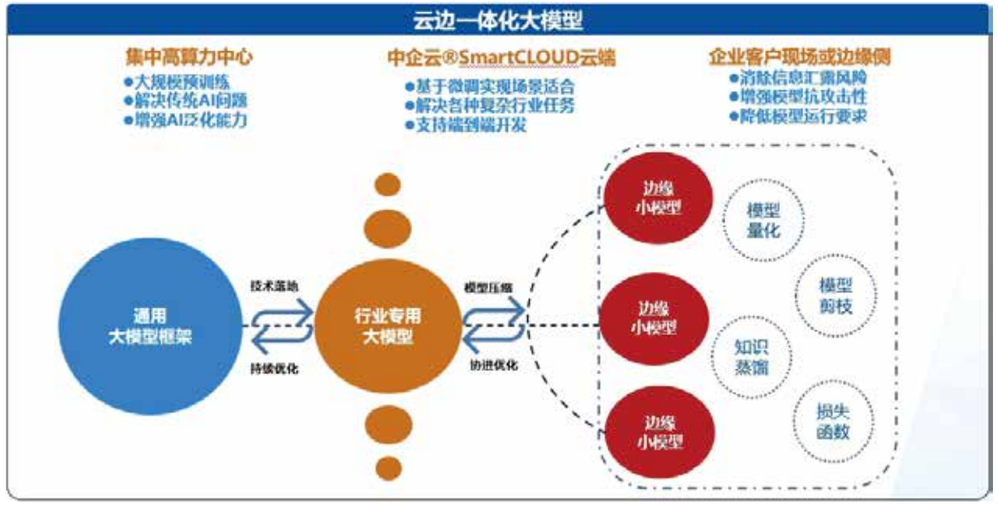
基于网络信息安全日志数据，使用大语言模型对网络安全态势进行全方位感知，精准识别网络信息安全威胁、网络流量异常等事件，用户能够以智能问答的方式快速掌握网络安全表征，并且结合信息安全专业知识，针对当前网络安全现象给出背后深层次的原因。

案例介绍

信息安全大模型平台利用通用大语言模型，引入网络安全语料库对其进行专门训练，让模型具备深度的网络安全专业知识，能够准确理解和回答网络安全领域的问题。再融入用户历史和实时的网络安全数据，进行深度学习和训练，深刻理解用户网络安全运行态势。

在用户输入安全问题后，平台对用户所输入的安全问题进行意图理解，根据问题的上下文，分析输入问题的语意结构信息以及词语间的依存关系，准确把握问题的意图。完成意图理解后，通过图匹配从安全知识图谱、安全运行数据中检索相关的实体作为应答，同时通过信息检索获得文本应答，最后将实体应答与文本应答拼接形成回复答案（见图一）。针对信息统计类的问题，平台通过语意理解准确提炼统计对象及统计条件，将用户的自然语言问题转化成 SQL 语句，对后台数据进行检索查询，并以表格和图表的方式将结果直观展示到用户面前（见图二）。

中企通信信息安全大模型平台为照顾用户在数据隐私及安全性上的顾虑，创新性采用了云边协同架构，在云端，以集中高算力使用信息安全行业领域专业知识、用户信息安全日志数据进行预训练，基于微调实现场景化的适配，解决各种复杂信息安全行业任务及企业任务。模型训练完成后，再将其最小化剥离及压缩下放到企业边缘侧，在企业内部环境中进行推演，并且与云端保持联系，反馈模型准确度，协同云端进行模型调优（见图三）。



效益分析

随着企业信息化系统的架构逐步走向分布式，系统的离散程度不断变大，给企业信息安全带来极大挑战。通过人工智能大语言模型，可以为企业提供全场景的安全专业知识、安全产品与解决方案，能够全面感知和突出企业当前信息安全态势及威胁，并暴露深层次原因及防护建议，提升企业信息安全管理水平，具有广泛市场应用空间。

全自研 AI 整合平台“HeyLisa”

北京泡泡玛特文化创意有限公司

北京泡泡玛特文化创意有限公司（以下简称“泡泡玛特”）成立于 2010 年，总部位于北京市朝阳区，是中国领先的潮流文化娱乐公司。以“创造潮流，传递美好”为品牌使命，发展十余年来，泡泡玛特围绕全球艺术家挖掘、IP 孵化运营、消费者触达、潮玩文化推广、创新业务孵化与投资五个领域，构建了覆盖潮流玩具全产业链的综合运营平台。

2020 年 12 月，泡泡玛特在港交所成功上市，成为国内“潮玩第一股”。得益于中国强大的制造能力和广阔市场，让泡泡玛特可以通过不断挖掘全球优秀艺术家、设计师，为其提供孵化平台，将更多国际和多元文化元素注入到艺术设计中，打造被全世界熟知、认可的潮流文化品牌，持续推动潮流文化在全球的传播。

概　述

在 2023 年，我们推出了自主研发的 AI 整合平台“HeyLisa”，这是一款旨在提供一站式 AI 服务、提升企业效率以及员工创新思维的全能型平台。HeyLisa 整合了文心一言、阿里通义千问、OpenAI、Meta LLaMA、清华智谱、百川智能和 Minimax 等多种自然语言大模型，为用户提供丰富多样的自然语言处理服务。同时，我们还引入了 OpenAI DALL-E 2、Stable Diffusion、LCM 等创新型 AI 绘画模型，以及微软 TTS、阿里云语音转译等先进的 AI 语音模型，使得用户在图像和语音处理上得心应手。HeyLisa 支持 Web 端、Mac 和 Windows 桌面端、飞书等多平台使用，为用户提供极致的使用体验。除此之外，我们还设计了自定义 Agent 功能，用户可根据需要自定义 Agent 角色信息，上传文档作为 Agent 知识库，通过插件系统使 AI 能与外部交互，与企业业务 API 无缝连接。其应用价值主要体现在：提高工作效率，创新思维的激发，降低人工成本，提升企业形象，提供决策支持等。

需求分析

项目“HeyLisa”是我们在 2023 年研发的 AI 整合平台，整合了自然语言大模型、AI 绘画模型、AI 语音模型等多种技术，为用户提供一站式的 AI 服务，提升企业的工作效率与员工的创新思维。

HeyLisa 的诞生，源自于当前行业对于人工智能技术的巨大需求。随着科技的发展，企业与个人对于信息处理的需求越来越高，而传统的手动处理方式已经无法满足这种需求。因此，我们需要一个能够整合多种 AI 技术，自动处理各种信息的平台，这就是“HeyLisa”的初衷。

在分析 HeyLisa 对行业用户的需求时，我们着重以下几个方面：

- **高效的自然语言处理：**在当前的商业环境中，大量的数据和信息都是以自然语言的形式存在的，如客户反馈、市场报告等。HeyLisa 通过整合文心一言、阿里通义千问、百川智能、OpenAI 等顶级自然语言处理模型，能够帮助企业快速理解和处理这些信息，提升决策效率。
- **创新的 AI 绘画和语音模型：**在设计、娱乐、教育等行业，用户需要将想法通过自然语言快速转化为具体的图像或声音。HeyLisa 支持 Stable Diffusion DALL-E 2 LCM 等 AI 绘画模型和微软 TTS 等 AI 语音模型，满足了这一需求。
- **广泛的平台兼容性：**不同的用户可能习惯于在不同的平台上工作，Heylisa 支持 Web 端、Mac 和 Windows 桌面端、飞书等多平台，使得用户可以在任何地方、任何时间使用 AI 服务。
- **自定义 Agent：**每个企业都有自己独特的业务需求和工作流程，HeyLisa 提供的自定义 Agent 功能，使得企业可以根据自身需求定制 AI 服务，如上传自己的业务文档作为 Agent 的知识库，实现 AI 与外部系统的交互等。

案例介绍

泡泡玛特于 2023 年开发出全自研 AI 整合平台“HeyLisa”。作为全能型的 AI 整合平台，HeyLisa 的宗旨在于提供一体化的 AI 功能和卓越的服务体验，激活企业部门的效率与员工的创新思维。 综合支持的功能包括以下五个方面：

- **多样化的自然语言大模型支持：**HeyLisa 整合了 OpenAI 、Meta LLaMA 、文心一言 、阿里通义千问 、清华智谱 、百川智能、 Minimax 的 API，并对外提供服务 。如图 1，图 2，模型成果展示如图 3。

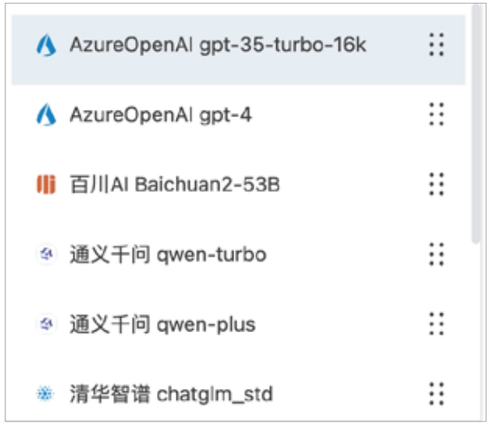


图 1

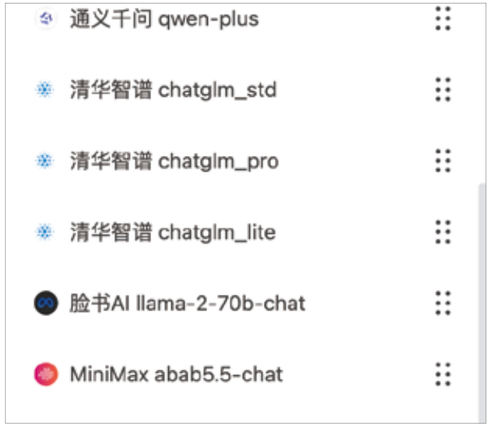


图 2

图 3

- **先进的 AI 语音模型支持：** 微软 TTS 、 阿里云语音转译。示例 1：客服与用户之间的通话进行语音转译文字识别。音频文件处理中如图 4，转译成功并识别顾客与客服角色如图 5。

图 4

图 5

- **自定义 Agent：** 用户可自定义 Agent 角色信息、用户可上传文档作为 Agent 知识库、支持插件系统，以实现 AI 与外部交互、与企业业务 API 无缝连接、可实现对企业级大数据的批处理，提供批量 AI 生成功能。

示例 1：自定义客服质检 Agent 来实现质检功能。输入质检要求输出质检结果如图 6。示例 2: 自定义企业知识库 Agent 来实现企业智能问答。上传企业文档如图 7。自然语言对话 Agent 如图 8。

图 6

图 7

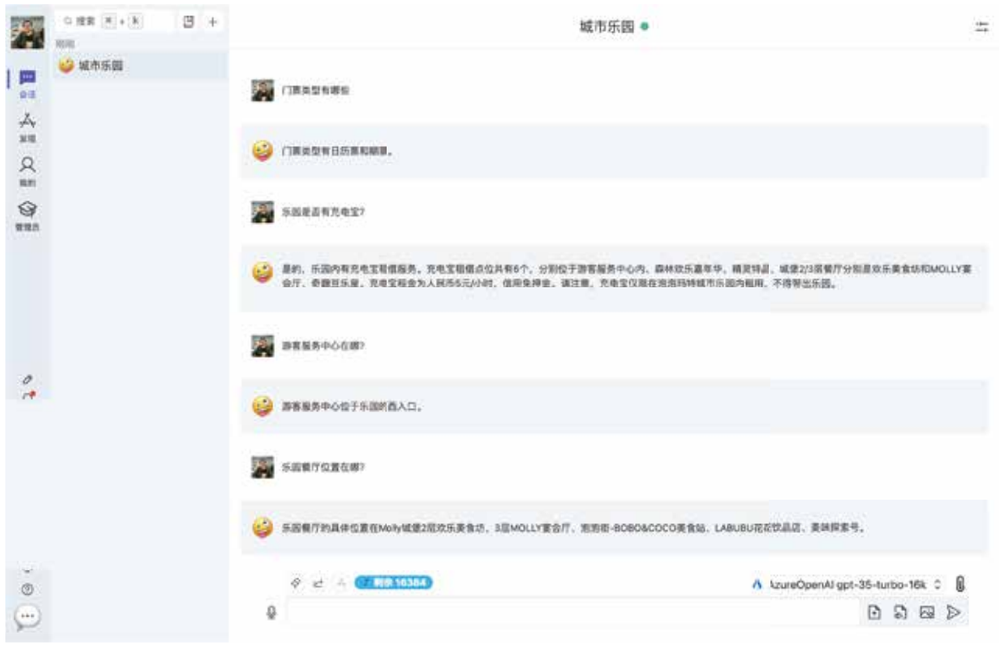


图 8

• **创新性的 AI 绘画模型支持：** OpenAI DALL-E 2、包括但不限于 Stable Diffusion、文生图、图生图、lora、controlnet 等十余款创新图像工具；我们主要是在阿里云部署的 https://github.com/AUTOMATIC1111/stable-diffusion-webui 集群版本，提供对企业内部使用，主要支持 sd 生态的大模型和 lora,controlnet 模型。我们在 IP 设计开发过程中会使用到的工具如：图片超清、移除背景、更换背景、面部修复、涂鸦生图、更换目标等实际的绘画使用场景工具。示例 1: 利用 LCM 技术来实现实时 Ai 绘出概念图，如图 9 图 10 图 11。

图 9

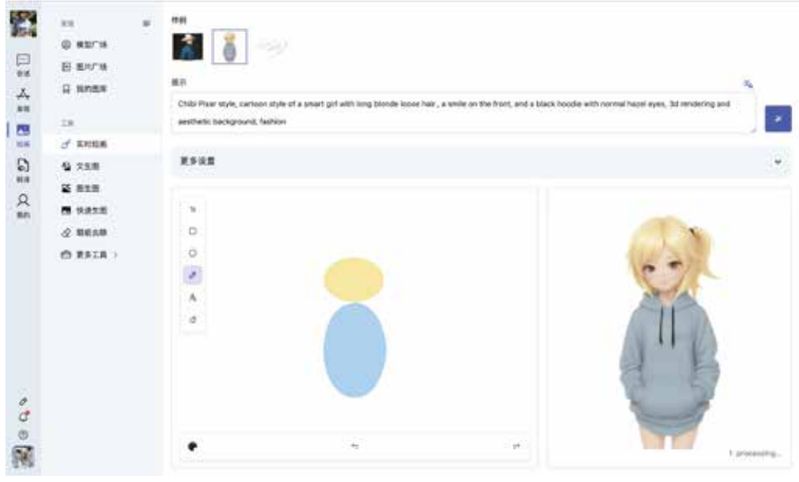


图 10

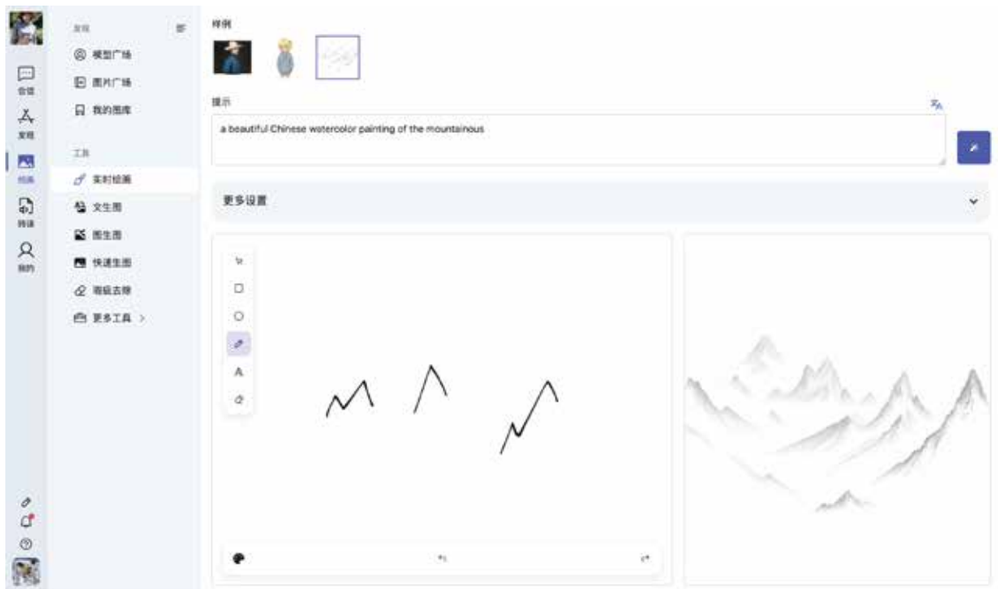


图 11

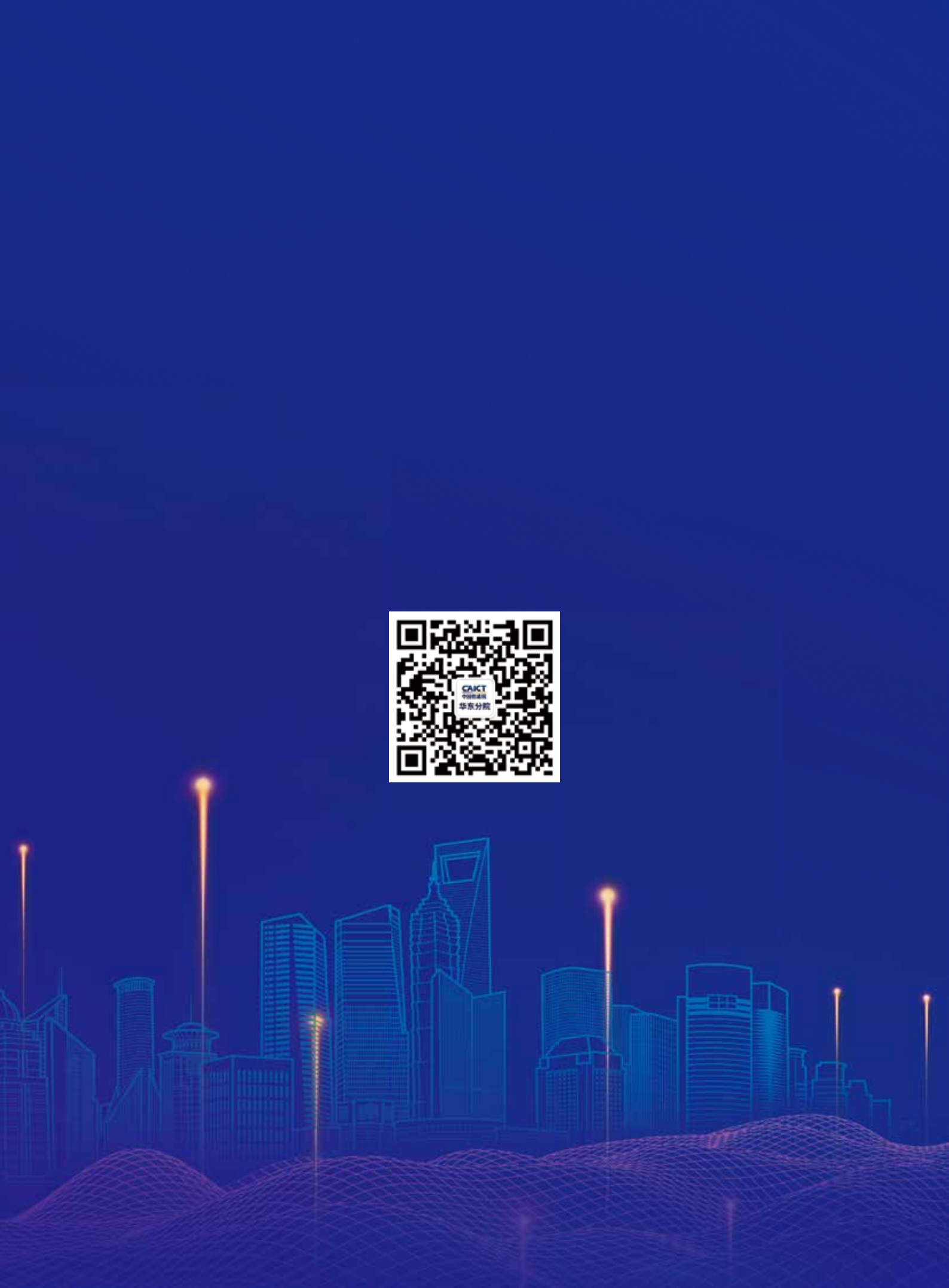
- **广泛的平台支持：** Web 端 、Mac 和 Windows 桌面端 、飞书。

效益分析

HeyLisa 是一款综合性的 AI 整合平台，其经济社会效益主要表现在：首先，它能够大大提升企业的工作效率，减少人工成本，从而提升企业的经济效益；其次，它能激发员工的创新思维，推动企业的技术创新，提升企业的竞争力，从而产生显著的社会效益。

商业模式方面，HeyLisa 主要是通过提供 AI 服务获取收益。用户可以根据自己的需求选择不同的服务套餐，并支付相应的费用。同时，我们也提供定制化的服务，用户可以根据自身需求，定制专属的 AI 模型。

在应用推广前景方面，随着 AI 技术的发展及其在各行业的广泛应用，HeyLisa 具有巨大的市场潜力。未来，我们会积极开展各类推广活动，通过网络、社交媒体、线下活动等多种方式，让更多的人了解和使用 HeyLisa，从而实现商业模式的持续优化和创新。


CAICT
华东分院