



附件一：申报主题

| | |
|---------------------------------|-----------|
| 1. 大模型及其应用研究 | 2 |
| 1.1 面向少量样本学习能力的多模态多任务大规模预训练学习 | 2 |
| 1.2 生成式大模型垂类自适应研究 | 2 |
| 1.3 机器翻译大模型 | 3 |
| 1.4 基于大规模数据及模型的多任务语音处理前沿技术研究 | 3 |
| 1.5 大规模稀疏模型训练加速研究和应用 | 4 |
| 1.6 可处理多媒体信息流的大规模语言模型 | 4 |
| 1.7 跨任务的通用强化学习模型 | 5 |
| 1.8 通用大语言模型在垂直 Game AI 领域的迁移和应用 | 5 |
| 1.9 基于大规模语言模型的虚拟人交互技术研究 | 5 |
| 1.10 基于预训练模型的事实编辑与多任务精调 | 6 |
| 1.11 基于 LLM 的代码智能技术研究 | 6 |
| 1.12 能源领域决策优化及序列预测大模型研究 | 7 |
| 1.13 医学大模型的研究和应用 | 7 |
| 1.14 大模型及其应用研究开放命题 | 8 |
| 2. 网络与信息安全 | 8 |
| 2.1 后量子签名算法 Dilithium 的侧信道攻击与防御 | 8 |
| 2.2 基于自然语言处理的二进制安全技术研究 | 8 |
| 2.3 威胁情报领域的若干关键技术研究 | 9 |
| 2.4 面向数据库系统安全漏洞的程序分析技术研究 | 9 |
| 3. 多模态融合与内容生成技术 | 10 |
| 3.1 跨域弱相关性广告匹配技术研究 | 10 |
| 3.2 数据中心 AI 驱动监控数据治理 | 10 |
| 3.3 AIGC 视频图像生成人物技术的探索研究 | 11 |
| 3.4 基于 2D 人脸动画先验的 3D 表情生成研究 | 11 |
| 3.5 手掌技术研究 | 11 |
| 4. 机器学习与深度学习 | 12 |
| 4.1 机器学习在生物信息学中的研究 | 12 |
| 4.2 内容生态中的高价值 KOL 挖掘 | 12 |
| 4.3 多域用户行为序列建模 | 13 |
| 4.4 自监督学习在用户增长领域的应用 | 13 |
| 5. 数据库 | 14 |
| 5.1 HTAP 数据库系统 | 14 |

申报主题介绍

(各主题均不限于给定的建议研究方向, 申请人可自行拓展决定。)

1. 大模型及其应用研究

1.1 面向少量样本学习能力的多模态多任务大规模预训练学习

随着数字化转型需求增长, 人工智能(AI)在社会中的应用也越来越多。然而, AI 开发门槛高、应用场景复杂多样、对场景标注数据依赖等问题成为 AI 规模化落地的挑战。因此, 当前 AI 规模化应用深度不足, AI 应用多以单个场景使用为主, 深入传统行业核心业务流程、完整解决行业痛点、实现商业价值的应用较少。大规模预训练的出现则为 AI 发展带来了新的机遇与希望。大规模预训练作为推进人工智能产业发展的重要抓手, 在识别、理解、决策、生成等 AI 任务的泛化性、通用性、迁移性方面都表现出显著优势和巨大潜力。同时, 大规模预训练正在推动 AI 进入大规模可复制的产业落地阶段, 仅需零样本、小样本的学习就可以达到很好的效果, 大大降低 AI 开发成本。本课题旨在少量样本场景下, 研究多模态多任务大规模预训练以及在少量样本学习下的视觉感知、理解和多模态链式推理, 实验开展上会以固定大模型参数只迭代少量参数的手段进行, 减少对大规模算力的依赖。课题组可提供大规模预训练平台和脱敏数据。

建议研究方向:

- 1) 计算机视觉感知和理解任务设计: 探究计算机通用感知技术, 赋能大模型理解数字图像和视频, 不仅让计算机具备视觉能力, 更让计算机理解视觉信息;
- 2) 大规模预训练: 探究预训练赋能多模态大模型的广度和深度, 提高模型上限, 增强通用性, 进而拓宽大模型应用广度, 实现图像、文本、语音等模态之间的统一表示和相互生成;
- 3) 多模态小样本学习: 探究以大模型作为基础设施, 在上面进行微调优化, 使用少量特定标注数据训练实现快速大模型应用落地;
- 4) 多模态推理链: 探究多模态大模型中的思维链 (CoT) 提示, 生成中间推理链作为推断答案的依据, 提高复杂推理的效果, 减少大模型产生幻觉的推理链, 消除误导答案的推理逻辑, 生成有助于答案推理的有效内容。

[返回目录](#)

1.2 生成式大模型垂类自适应研究

目前, 大型语言模型在通用场景展示出了强大的能力, 然而在一些领域知识丰富的垂类场景, 通用大模型在效果上仍然无法达到工业界的预期。因此, 通过高效节能的方式将通用生成式大模型迁移到特定专业领域成为重要的探索方向。与此同时, 大模型的领域迁移与后期应用面临着诸多挑战, 比如, 如何低功耗完成垂类大模型开发, 在保留模型通用能力的基础上增强专业领域知识; 如何保证大模型微调过程中梯度稳定且高效收敛等。本课题旨在打造高质量、高效率的垂类大型语言模型, 兼顾科学研究与落地应用价值。可应用的实际场景

包括：1. 标注提效：为文本标注人员提供多粒度辅助信息；2. 数据增强：样本改写及批量相似样本生成；3. 工作提效：打造部门工作场景机器人，进行各类知识管理。

建议研究方向：

- 1) 开发垂类对话大模型：探索绿色高效领域适应技术；
- 2) 微调大模型：探索大模型微调快速收敛方法；
- 3) 大模型高效使用：在有限资源条件下实现大模型快速高效训练&推理。

[返回目录](#)

1.3 机器翻译大模型

目前，大模型在翻译知识获取与利用方面取得了显著突破，具有巨大的潜力，但也带来了新的挑战。首先，大模型的参数众多，无法简单高效地用于翻译任务；其次，大模型的训练数据涉及广泛，导致在特定领域的翻译场景下存在性能瓶颈。本课题旨在开展基础及应用研究，提升机器翻译系统的效果。可通过派遣实习生使用腾讯 GPU 资源或在 10B 模型上验证方法。

建议研究方向：

- 1) 基于在线学习的翻译大模型方法，包括但不限于上下文学习，实时自适应等前沿技术；
- 2) 大模型的高效参数精调策略，提高训练效率和翻译数据吞吐，包括但不限于 LoRA、Prefix Tuning 等前沿技术；
- 3) 其他翻译大规模技术探索。

[返回目录](#)

1.4 基于大规模数据及模型的多任务语音处理前沿技术研究

语音是人机交互的一种最自然的方式，语音处理技术包括语音识别和合成等是 AI 时代的一个重要研究领域。随着大规模数据及大规模语言模型在文本模态的优势日益显现，语音作为区别于文本的一种连续信号序列模态，其前沿研究面临特殊的机遇和挑战。通用领域的 ASR 仍然面临着领域适应性、噪声鲁棒性、声学环境复杂性等问题，需要挖掘覆盖各种场景的海量数据来解决。同样，语音合成与转换也受到包括说话人、风格、情绪、韵律等各种强上下文条件影响，而对应的高质量、有标注的数据往往很难获取。本课题旨在通过大规模数据、自监督学习及大规模语言模型的技术找到解决上述难题的突破口。

建议研究方向：

- 1) 研究利用海量无标签数据的自监督模型，例如基于类语言模型损失函数训练的自监督模型，提取通用 (Universal) 表征；
- 2) 研究多模态数据 (文本、语音、图像等) 作为输入的自监督模型，构建通用的声学模型；
- 3) 研究利用海量无标签数据的基于通用表征的大规模语音识别模型；
- 4) 研究具有 In-Context 学习能力的高质量 and Few-Shot 个性化声音合成与转换模型；

- 5) 研究基于通用表征的 Prompt 适应语音，用于同步预训练声源分离、增强、识别以及声音合成的多任务模型。

[返回目录](#)

1.5 大规模稀疏模型训练加速研究和应用

大规模稀疏特征的深度学习模型因其优越的特征表示和特征交叉能力，被广泛应用到商业公司的广告、推荐及搜索等场景当中。由于一般大规模稀疏特征模型参数规模达百亿规模，部分场景可达千亿规模，并且在商业系统中需要反复重新训练，迭代模型效果，因此大模型的训练效率和成本的优化是模型平台的核心挑战。与 NLP/CV 等场景计算密集型大模型不一样，稀疏特征模型参数量主要集中在特征 Embedding 上，这也使得基于大规模稀疏参数的大模型优化方向和传统大模型不一样。本课题旨在对大规模稀疏特征的深度学习模型进行训练参数通信优化、软硬件一体化优化、训练加速等方面的研究。课题组将根据研究需要提供 GPU 训练集群资源。

建议研究方向：

- 1) 模型训练参数通信优化：利用参数压缩、参数量化、参数稀疏化以及延迟更新等手段，降低大模型训练过程中由于参数传输带来的带宽压力。同时研究 NVSwitch+PCIe+RDMA 异构硬件通信方式优化；
- 2) 软硬件一体化优化：通过定制硬件以及相应软件设计的一体化优化，打造适合大规模稀疏特征的大模型训练框架；
- 3) 模型训练加速：通过大 Batch 训练方法、适合大规模稀疏特征模型的 Loss 函数，研究提升模型训练和收敛速度。

[返回目录](#)

1.6 可处理多媒体信息流的大规模语言模型

多媒体信息流指包含文本、图片、视频、音频等不同类型数据的富媒体文件。现有的大规模语言模型往往只具备理解和生成文本数据的能力，如何利用现有的大规模语言模型，使其可以理解多模态信息流，同时又不破坏其原本的 In-context Learning, Chain of Thought, 和 Instruct Learning 的能力，对于信息检索、对话系统、内容创作等多个领域的应用具有重要意义。本课题旨在研究如何使大规模预训练的语言模型（LLM）具备处理多媒体信息流的能力。课题团队可支持派遣实习生现场/远程使用腾讯 GPU 资源，访问标注数据资源，或在 10B 左右模型上验证方法。

建议研究方向：

- 1) 大规模语言模型对于图像内容的理解：如何做好视觉信息的 Tokenizer，使图像内容更好地嵌入到大规模语言模型中；
- 2) 大规模语言模型对于视频内容的理解：如何对视频内容中的时序信息进行建模，更好地嵌入到大规模语言模型中；
- 3) 大规模语言模型对于音频内容的理解：如何将音频内容更好地嵌入到大规模语言模型中。

[返回目录](#)

1.7 跨任务的通用强化学习模型

目前已证实了大模型在自然语言处理上的强大能力,但是在行为决策上能够达到怎样的高度却未被深入地研究。若能建立一个跨任务的通用决策模型,将极大降低我们在不同任务上的人工调整成本,提升解决问题的效率。当前已有一些研究(如 GATO)关注到了大模型在行为决策上的使用,但是仍有很多研究空间。本课题旨在探索行为决策方面的跨任务通用模型研究,比如如何更好地设计模型,如何更好地设计数据使用数据,如何使用更少的提示或者精调来利用大模型的泛化能力等。

建议研究方向:

- 1) 探索行为决策方面的跨任务通用模型,可以从多个任务场景中学习,包括但不限于进一步探索模型结构设计、如何构建更好的数据集、如何使用更少提示或者精调等技术来加强大模型的泛化能力,减少人工调整策略的代价;
- 2) 探索利用预训练模型在各个领域的泛化能力,使得通用模型可以快速适应新方法。

[返回目录](#)

1.8 通用大语言模型在垂直 Game AI 领域的迁移和应用

近年来,大规模预训练语言模型取得了极大的成功,受到了国内外学术界、产业界以及社会的广泛关注。通用的大语言模型将助力更多人工智能其他领域的新技术和新应用的发展,也为 Game AI 带来了新的机遇与挑战。本课题旨在以狼人杀为代表场景,研究通用大语言模型在 Game AI 领域中的迁移和应用。

建议研究方向:

- 1) 人类自然语言的理解和生成:在未知身份、多阵营等复杂环境中,对较长的上下文依赖、长文本信息、冗余信息的识别、理解和逻辑判断能力,以及综合决策和自然语言理解结果的自然语言生成能力;
- 2) 通用大语言模型和大规模强化学习决策模型的结合:研究如何提取通用大语言模型学习到的语言理解、世界知识、思维链等能力,将其与强化学习模型的复杂状态下决策能力结合,实现 AIGA(人工智能决策生成);
- 3) 人机交互新范式:在通用大语言模型的赋能下,Game AI 中人机交互场景方面涌现许多新的可能性。

[返回目录](#)

1.9 基于大规模语言模型的虚拟人交互技术研究

大规模语言模型展现出了强大的智能交互和创作能力,如何利用大规模语言模型提升虚拟人交互能力成为一个具有研究价值的课题。现有的虚拟人和数字形象往往缺乏人性化、智能性的特征,本课题旨在训练一个垂直领域的大规模语言模型,使其具备适配虚拟数字场景需求的交互生成能力来提升虚拟人的智能交互性。

建议研究方向:

- 1) 垂直场景的大规模生成式语言模型: 针对垂直场景的特定数据和任务需求, 设计创新方案并训练大规模生成式语言模型;
- 2) 虚拟角色与用户之间、多虚拟角色之间的对话生成: 理解玩家通过自然语言的任意输入, 生成符合虚拟世界观设定、虚拟角色人设、虚拟世界实时进程的对话交互内容, 并且对话内容可控, 能根据需求体现特定剧情或事件, 增强虚拟世界的互动性和沉浸感;
- 3) 基于 LLM 的动态故事生成: 在保证与虚拟世界观一致的前提下, 根据用户行为生成独特且具有完整叙事结构的故事线, 保证故事逻辑正确, 包含明确的铺垫、冲突、高潮等叙事元素, 提升剧情丰富度;
- 4) 虚拟人多模态交互能力: 根据对话情景输出对应的情绪、动作等其他模态的结果;
- 5) 生成大模型的部署和推理加速, 使得虚拟人智能交互能力可落地应用, 实现与用户实时交互。

[返回目录](#)

1.10 基于预训练模型的事实编辑与多任务精调

精细化用户行为表征是实现内容个性化分发、广告精准营销的关键技术。在真实业务场景中, 海量用户行为属于无标签样本。在传统的有监督-端到端框架内, 难以发挥数据优势, 导致学习到的用户表征存在信息瓶颈。因此, 构建无监督全生命周期的用户表征模型, 是提高业务效率的有效途径。然而, 用户的预训练表征在下游任务适配时, 常面临事实关联模糊、知识更新成本高、特征鲁棒性差等问题。此外, 由于模型预训练的样本来源于多个领域, 如用户活跃、广告转化、内容浏览等。这导致不同领域样本量/数据分布差异较大, 预训练模型普遍存在领域偏差 (Domain Bias)。本课题基于预训练模型, 重点研究模型知识编辑方法和多任务精调适配策略, 旨在提升预训练模型在内容推荐和广告投放中的准确性。

建议研究方向:

- 1) 优化大型语言模型结构 (如 Bert/GPT) 并进行无监督用户行为表征;
- 2) 无需重训练和微调, 探索预训练模型中事实知识编辑方法;
- 3) 研究基于少量标签样本/无标签样本的多任务通用精调适配范式。

[返回目录](#)

1.11 基于 LLM 的代码智能技术研究

当前, 大模型在代码、软工类任务中的表现非常惊人。本课题旨在研究指令调教、SFT、RLHF 等技术在软工类任务中的应用, 提高大模型在代码补全和生成、代码摘要、代码解释、代码漏洞检测、代码缺陷检测及代码自动修复等方面的能力, 并落地工蜂智能化工具, 提升研发效能。语言偏向于 (JavaScript、C/C++、Golang)。

建议研究方向:

- 1) 结合典型的软工场景, 研究指令调教技术 (Instruction Tuning), 设计软工类任务指令对大模型进行微调, 提升大模型在新任务上的 Zero-shot 能力;
- 2) 研究并构建社交、游戏等领域的软工数据集, 包括单轮、多轮软工 Prompt 数据, 对大

模型进行 SFT (Supervised Fine-Tune)，提高大模型在具体软工任务上的精准度；

- 3) 研究 RLHF (Reinforcement Learning from Human Feedback) 技术在软工类任务的应用，提高大模型在具体软工任务上的精准度。

[返回目录](#)

1.12 能源领域决策优化及序列预测大模型研究

几年来强化学习算法可以在多个场景下学习到优秀的策略以实现自动化控制，然而强化学习算法难以保证其在能源领域运用的安全性。在安全性要求高的自动化控制领域，如何提升强化学习算法的安全性和可靠性，是本课题的主要研究内容。本课题旨在研究和开发一套能源领域智能决策及序列预测大模型，从理论到实践，提升能源行业决策的智能化水平。

建议研究方向：

- 1) 利用运筹优化的带约束优化算法，解决策略神经网络的带约束问题，确保算法在各种不确定条件下的安全和稳定运行的条件，提升强化学习算法在现实世界的落地能力；
- 2) 利用强化学习算法的策略探索能力，提高能源效率，降低成本，减少碳排放；
- 3) 利用动态仿真技术的策略分析能力，仿真既定策略下能源系统的实际运行状态，分析策略的安全性和可靠性，研究基于深度学习的仿真预测技术，提升能源系统动态仿真效率；
- 4) 探索针对能源领域的高效决策建模技术，提升模型性能和预测准确度，以适应能源领域复杂多变的环境，研究融合时序预测能力的决策算法。

[返回目录](#)

1.13 医学大模型的研究和应用

LLM (Large Language Model) 在用户意图理解和内容生成上具有强大的能力。随着 LLM 技术的快速发展，自然语言处理中的各种任务，比如信息抽取、摘要生成、多轮对话等，都有可能被重新定义。在医学领域有很多 NLP 任务，包括医患对话、医学信息提取、疾病诊断、用药推荐、报告生成等，几乎所有的 NLP 任务在医学中都可以找到对应落地场景，因此医学大模型有很广阔的应用场景。本课题旨在借助多年来医学数字化积累的大量医学语料，包括问答、文章、图谱、教材、医学报告、试题等，在课题组构建的大规模高质量的医学知识图谱基础上，探索医学领域大模型的研究和应用。

建议研究方向：

- 1) 医疗大模型的训练：调研 LLM 模型，结合医疗数据训练垂直领域的生成式预训练模型；
- 2) 知识图谱和生成式预训练模型的互相增强：一方面使用高质量的医学知识图谱提升生成式预训练模型生成内容的正确性；另一方面使用生成式预训练模型对知识图谱进行查漏补缺；
- 3) 医疗大模型的应用研究：研究基于生成式预训练模型的医疗应用，通过对基座模型进行 Finetune 和 In Context Learning，研究出院小结助手、辅诊助手、用药助手、院外管理助手、术后管理助手等（医院场景）；医生访谈摘要生成、沟通话术生成、智能推荐参会候选人、医药代表迅速获得产品信息等（药厂场景）；

- 4) 多模态生成式预训练模型研究：医学中的影像图片、医学科普文章中的插图、视频以及直播中的音频等都是多模态的信息，将医学图像、视频、音频、文本等多模态信息集成进入 LLM，并探索其在医疗 AIGC 方面的应用。

[返回目录](#)

1.14 大模型及其应用研究开放命题

本课题是基于大模型及相关应用的开放命题，内容和方向不限。

[返回目录](#)

2. 网络与信息安全

2.1 后量子签名算法 Dilithium 的侧信道攻击与防御

后量子密码学是指能够抵抗量子攻击的密码算法。2022 年，美国国家标准技术研究院公布了标准化的四种后量子密码算法，其中 Dilithium 是选定的三种数字签名算法之一。密码实现在运行过程中往往会泄漏如功耗、时间等侧信道信息。尽管 Dilithium 在数学上是安全的，但它的实现可能受到侧信道攻击。本课题旨在研究 Dilithium 对侧信道攻击的敏感性，并制定有效的对策。课题组可提供该合作成果与数字人民币深度结合的落地场景，以及双方联合参加金融密码杯的机会。

建议研究方向：

- 1) 针对软件实现的 Dilithium 算法，采集功耗曲线进行侧信道攻击，伪造合法签名并研究有效的防御措施；
- 2) 实现切实可行的 Dilithium 算法解决方案，包括但不限于工程实现、安全论证等；
- 3) 产研结合，探索与数字人民币业务场景深度结合的 Dilithium 算法安全解决方案。

[返回目录](#)

2.2 基于自然语言处理的二进制安全技术研究

随着软件规模不断增大，针对二进制程序的分析与安全保障工作面临着新的机遇和挑战。作为一种特殊的信息表示形式，二进制代码的分析任务与自然语言处理中的某些场景存在着一定的关联性（例如：二进制反编译与文本翻译）。本课题旨在结合学术界研究热点，将自然语言处理领域最新的研究内容应用于二进制程序分析领域，推进二进制程序智能化分析进程，并在逆向工程、漏洞挖掘等关键任务方面产生积极影响。课题组提供大规模二进制/源代码函数数据集和代码向量化技术和自研二进制安全智能分析平台 BinaryAI (binaryai.cn)，以支持相关技术研究。

建议研究方向：

- 1) 二进制可执行文件的软件成分分析 (BSCA)，包括第三方库及其版本号的识别、第三方库间的源代码级复用关系识别、已知漏洞检索；

- 2) 面向二进制/源代码函数嵌入的模型设计及其上层应用研究，例如二进制与源代码间的相似度比较、二进制差异性比对、逆向工程等；
- 3) 面向二进制文件嵌入的模型设计及其上层应用研究，例如二进制文件相似性检索、二进制文件分类、二进制文件无监督聚类等；
- 4) 基于语言模型的程序翻译和生成技术研究，例如二进制与源代码间的代码翻译、二进制符号还原、编译器/解释器模糊测试的代码用例生成。

[返回目录](#)

2.3 威胁情报领域的若干关键技术研究

传统网络安全技术面临数据来源单一、在线处理能力有限、统计关联分析能力欠缺等问题，无法有效克服当前复杂的网络安全挑战。而威胁情报技术是指一系列利用大数据和机器学习技术对海量数据进行安全分析进而对威胁进行决策的过程，它是当前网络安全研究热点，也是对传统安全技术和安全产品的有力补充。本课题旨在对威胁情报数据分析、检测自动化及其他相关领域技术进行探索研究。

建议研究方向：

- 1) 威胁情报表示技术研究，例如面向多源威胁数据的抽取与统一表示技术等；
- 2) 威胁情报数据分析技术研究，例如基于主机/流量数据的攻击链路关联技术等；
- 3) 威胁检测自动化技术研究，例如多辖域的 IP 同源性及复杂攻击检测技术等；
- 4) 威胁情报相关领域技术研究，例如面向高级持续攻击的威胁情报画像分析与可视化、DNS 数据挖掘技术、网络空间测绘技术等。

[返回目录](#)

2.4 面向数据库系统安全漏洞的程序分析技术研究

数据库系统作为三大基础软件之一，是信创场景的重要组成部分。腾讯 TDSQL 在数据库产品技术发展迅速，市场增长迅猛，但同时也产生一系列问题和挑战。其中，安全漏洞和导致偶发性系统崩溃的质量问题凸显。例如，金融、电信等重点行业对数据库系统的使用广度和强度极高，同时对安全质量、系统可靠性的要求极高，需要杜绝安全漏洞或质量风险导致的系统被入侵或系统不可用。近年来，对数据库系统安全性的分析也成为学术界的热门话题，在 ACM CCS、Usenix Security 等顶级学术会议上均有先进研究成果。本课题旨在探索程序分析技术在发现数据库系统安全漏洞方面的技术和应用，结合数据库系统的特点和场景，使用人工智能、静态程序分析、模糊测试等技术手段发现深层次漏洞和安全质量风险。

建议研究方向：

- 1) 面向分布式数据库系统质量安全的静态程序分析技术；
- 2) 面向分布式数据库系统质量安全的模糊测试技术；
- 3) 人工智能制导的数据库系统程序分析技术。

[返回目录](#)

3. 多模态融合与内容生成技术

3.1 跨域弱相关性广告匹配技术研究

在互联网广告中，内容域存在多种媒介形式，如文章、短视频、长视频、论坛社区等。建立广告域外内容和广告域的匹配对于广告系统具有重要意义，有效的匹配可以提高广告与受众之间的相关性和兴趣匹配程度，从而提高广告的点击率和转化率，并提高广告推荐系统的准确性和实时体验感等。内容域数据和广告域数据的匹配核心在于根据相关性实现关联，在相关性较弱的情况下，此类匹配更具挑战性。本课题旨在利用文本、图片、视频等内容和广告信息，基于 Embedding、标签、以及其他技术构建多路模型，实现内容域数据与广告域数据的弱相关性匹配。

建议研究方向：

基于腾讯给出的上亿级别的多模态内容域数据（如文章、短视频、长视频、论坛社区等）和广告域数据（如图文广告、视频广告），以及相关的训练资源和环境，在充分调研业界方案的基础上，合理搭建基于 Embedding 检索、标签匹配及其他辅助关联技术的多模态大模型架构，挖掘出和给定内容有弱相关性的广告。

[返回目录](#)

3.2 数据中心 AI 驱动监控数据治理

腾讯在全球有 100 多个数据中心，稳定架构和准确数据成为自动化运营的关键。智维平台每秒钟采集超过千万测点，海量数据治理能力成为提升数据质量的瓶颈。例如直流霍尔传感器或温湿度都存在随使用周期发生非线性精度漂移的现象，电池监控仪探头松动或采集器通信延迟也导致采集数据质量下降，传统治理手段通过现场定期巡检使用校准仪器测量和标定，工作效率低和治理成本高。通过多年运营经验，可发现 IT 服务器功率和 PDU 采集功率拟合、双路交直流输入功率分配比、多路温湿度变化趋势分析等多测点之间存在耦合性和关联性。本课题旨在基于已建立的数据资产基线和数据治理策略规则库，图模一体的电力单线图和暖通拓扑图，建立 AI 数据治理模型，增强专家策略的泛化性和连接性。

建议研究方向：

- 1) 基于测点的拓扑连接关系建立数据预测模型，融合专家规则和多源多模态数据源进行有监督学习，实现对数据序列质量分析、异常判断、状态预估和自动补全；
- 2) 基于 Transformer+GNN 建立长尺度多元时间序列数据的预训练大模型，自动学习多测点数据的空时域关系，通过治理规则、图模一体和拟合分析等微调模型完成具体下游的数据治理任务。

[返回目录](#)

3.3 AIGC 视频图像生成人物技术的探索研究

随着 AIGC 概念的出现以及推广，如何更高效地制作图片以及视频，以大幅提升内容制作效率，降低制作成本，是当前学术界及业界十分关注的研究方向，该技术可以广泛应用于广告、娱乐等领域。本课题旨在通过 AIGC 方面的算法研究，实现高品质的图片生成，通过生成连续的图片精确地表达文案中描述的情景故事和人物。课题组可提供相关的文案素材。

建议研究方向：

- 1) 基于网络开放图片素材或者少量高质量原画及立绘，训练指定风格的模型；
- 2) 文案中人物一致性的定义以及多次图片生成过程中，人物一致性连贯性的保证(含多人场景)；
- 3) 人物行为动作的准备表达，优先单人行为动作；
- 4) 动漫类型视频制作。

[返回目录](#)

3.4 基于 2D 人脸动画先验的 3D 表情生成研究

目前，已有很多工作可基于大量 2D 人脸视频实现高质量的 2D/体表达面部动画生成。然而，3D Mesh 动画领域由于数据资产获取难度大、数据表达复杂，高质量 3D 面部 Mesh 动画生成仍具有较大挑战。本课题旨在探索一种方案，利用 2D 单/多视角或体表达人脸表情生成模型作为先验或约束，结合少量 3D Mesh 动画数据，实现高自然度、高丰富度的 3D 表情生成。

建议研究方向：

- 1) 单视角面部驱动模型提升 3D 表情捕捉/生成效果；
- 2) 多视角(或体表达，如 NeRF)面部驱动模型提升 3D 表情捕捉/生成效果；
- 3) 基于语音的 2D/体表达面部驱动模型提升 3D 表情生成效果。

[返回目录](#)

3.5 手掌技术研究

刷掌支付是使用体验上最接近扫码的生物识别支付方式，相对其他生物特征在意愿代表、使用体验、隐私性等方面有较为明显的优势。在大规模应用中，刷掌支付若要为用户提供更好的支付使用体验，将对算法提出更高的要求。本课题旨在针对刷掌支付场景，在更可靠(超大规模识别)、更安全(活体检测)、更优体验(姿态判断、质量判断)、更低成本(模型轻量化)等方面进行算法研究。

建议研究方向：

- 1) 大规模识别：可应用于掌纹/掌静脉的大规模研究方向，可以与生成、质量、局部与全局等方向结合；
- 2) 活体检测：从域泛化、异常检测、多模态等方向进行研究，提升活体模型安全能力以及跨设备迁移能力；

- 3) 姿态判断：单目骨骼点检测、深度估计，跨模态双目 3D 计算；
- 4) 质量判断：清晰度判断能力（掌纹/掌静脉），无监督或半监督质量标签生成；
- 5) 轻量化：模型剪枝量化等能力，适配低成本 CPU 与 NPU。

[返回目录](#)

4. 机器学习与深度学习

4.1 机器学习在生物信息学中的研究

生物信息学是一门跨学科领域，它涉及到生物学、计算机科学、数学和统计学等多个学科，旨在通过对生物数据的处理和分析，为生物学和医学研究提供支持。在这个领域中，蛋白质结构预测和设计、单细胞多组学数据和空间多组学数据的分析和挖掘、免疫库数据的分析和挖掘等都是热点研究方向。为了实现这些目标，生物信息学研究者不断探索和开发新的算法和工具，其中人工智能技术的应用在生物信息学中日益普及。本课题旨在应用人工智能技术助力更准确地预测蛋白质结构并设计新的蛋白质，更深入地了解单细胞和组织的功能及互动，更好地挖掘免疫组库数据并开发新的免疫治疗方案。

建议研究方向：

- 1) 人工智能算法和工具在蛋白质结构预测和人工蛋白质设计方面的创新和应用；
- 2) 人工智能算法和工具在单细胞多组学数据和空间多组学数据分析和挖掘方面的创新和应用；
- 3) 人工智能算法和工具在免疫组库数据分析和挖掘方面的创新和应用。

[返回目录](#)

4.2 内容生态中的高价值 KOL 挖掘

近年来，广告营销进入了“全域经营”时代，内容生态在产品体验延伸、用户反馈收集和内容营销获客层面的价值逐渐受到重视。内容生态运营中，如何挖掘生态内高价值的 KOL (Key Opinion Leaders, 关键意见领袖) 是非常重要的一环。然而，由于可参考经验有限、无法对历史决策进行复现、内容生态系统较复杂等现实因素，对于生态系统内 KOL 价值的客观评估较为困难。本课题旨在使用表示学习和异质性处理效应估计方案，探索一套可量化的 KOL 价值的表征方案，并评估在全域经营视角下，对所选 KOL 的资源投入带来的短线/长线价值产出。根据研究需要，课题组可提供相关脱敏数据和计算资源。

建议研究方向：

- 1) 表示学习：基于深度学习技术学习到内容生态的特征表示，寻找全域视角下的高价值 KOL；
- 2) 异质性处理效应估计：对于拟选择的 KOL，可以使用异质性处理效应估计方法，估计对于 KOL 不同扶持力度下可以带来的边际收益期望值，用于给出项目组资源投入的建议；
- 3) 基于运营节点的 KOL 挖掘：在产品发行生命周期中的不同节点，对于 KOL 的需求存在较大差异，结合产品生命周期对内容生态系统的影响，探究不同阶段下挖掘 KOL 的方案。

[返回目录](#)

4.3 多域用户行为序列建模

用户的历史行为反映出用户兴趣。腾讯拥有众多触达用户的场景，如内容推荐和广告推荐场景，用户行为散落在众多场景之中，如能充分利用起来，将会极大提高推荐效率。基于多域（特别是内容和广告推荐域）的所有行为来刻画用户兴趣存在以下挑战：1. 不同域的特征体系不一致，虽然都存在大量多模态信息，也大量使用 ID 特征，但域间并未天然打通；2. 用户在各个域中的兴趣存在差异，比如用户喜欢看某种类型的文章不代表用户会点击某种类型的广告。本课题旨在设计一种基于多域行为序列的用户兴趣建模方案，充分利用内容域和广告域的所有用户行为数据，提升广告推荐的准确性，增加广告营收。

建议研究方向：

- 1) 基于域外行为序列的自监督预训练模型+Finetune：即首先基于内容域的用户行为序列，通过自监督学习/对比学习的方式进行预训练，得到用户的兴趣表示，然后基于广告语的有监督信号对该兴趣表示进行 Finetune 用于广告推荐；
- 2) 基于全域行为序列的联合训练的自监督大模型+Prompt Learning 的方式：即同时基于内容域和广告域的用户行为，构造一个包含多个域行为的长序列，进行自监督学习，然后针对下游的广告推荐任务设计 Prompt Learning 模型用于广告推荐；
- 3) 基于全域行为序列的联合端到端用户兴趣模型：即同时基于内容域和广告域的用户行为，构造一个包含多个域行为的长序列，然后利用内容域和广告域的监督信号来端到端的学习用户兴趣，提升广告推荐和内容推荐的效果。

[返回目录](#)

4.4 自监督学习在用户增长领域的应用

用户增长是产品成功发行与有效运营的重要前提，而对用户、产品本身以及相关各场景的深入理解与精准表达是构建高效用户增长能力的重要基础。该领域存在诸多技术难点，如冷启动、数据稀疏、数据缺失等。本课题旨在利用腾讯海量脱敏数据，利用自监督学习，在不依赖数据标签的情况下，提升用户理解能力、深入挖掘数据价值，实现更高效的用户增长。

建议研究方向：

- 1) 利用自监督学习加强对用户付费、留存等行为的预估能力；
- 2) 基于用户画像、用户行为与事件等，利用自监督框架对用户进行表示学习 (Representation Learning)；
- 3) 自监督学习结合图学习 (Graph Learning)，提升对产品、用户、内外部知识之间相互关系的理解能力；
- 4) 自监督学习结合多模态 (Multi-modal) 技术，提升对与用户或产品相关的文字、图片、视频等的信息挖掘能力。

[返回目录](#)

5. 数据库

5.1 HTAP 数据库系统

HTAP (Hybrid Transactional/Analytical Processing) 指混合型的事务处理和分析处理。传统上,事务处理和分析处理是分开进行的,事务处理主要应用于 OLTP 系统 (On-line Transaction Processing, 联机事务处理系统),分析处理则主要应用于 OLAP 系统 (On-line Analytical Processing, 联机分析处理系统)。HTAP 则是结合了这两种系统的特点,同时支持高并发、高吞吐、低延迟的事务处理和强大的分析处理。这种混合型的系统有助于提高数据处理的效率,降低系统的复杂度和成本,提升企业的运营效率。然而两种系统的融合也带来了一些挑战,HTAP 系统需要在系统性能 (Throughput) 与数据分析新鲜度 (Freshness) 之间做出取舍。本课题旨在研究 HTAP 数据库的存储、管理、查询、资源调度等问题。

建议研究方向:

- 1) 行列混合查询及其优化: 基于数据分区数据压缩、索引优化、数据缓存等因素的行列混合引擎下的查询优化技术;
- 2) HTAP 评测基准的设计和实现: 探究评估 HTAP 系统性能的基准测试方法,包括事务、分析和混合负载,考虑系统的硬件和软件环境,如 CPU、内存、磁盘和网络等,以确保测试结果的准确性和可重复性;
- 3) HTAP 中一致性和时效性的平衡与优化: 提出系统既能保证数据更新的及时性又能保证最新的数据能立即被查询和分析的技术 (如探索考虑数据持续更新下的 OLAP 查询近似答案的方法);
- 4) 基于机器学习的 HTAP 系统自动化优化 (AI for DB): 探究机器学习技术来自动优化混合事务/分析处理 (HTAP) 数据库系统的方法。

[返回目录](#)