2026 年度 CAAI-蚂蚁科研基金（AGI 专项）

申报课题介绍

目　录

2.面向 Agentic 的 RL 与 OPD 协同训练系统研究

3.AI 可控可靠保障之多纬度智能体可控性评测框架研究

4.面向广告投放的 Agentic RL 决策优化框架：从智能竞价到全链路自主投放研究

5.Agentic Learning 在长程任务场景的研发与落地研究

6.基于多模态大模型的通用游戏 Agent 构建和训练研究

7.面向复杂任务的多模态 Agent 研究

8.用于科学智能的 Agentic RL 算法研究

9.Deep Research 推理一致性与幻觉抑制优化研究

10.LLM 对 Agent Harness 的适配能力评估与优化方法研究

11.面向代码智能体任务的大语言模型推理与规划能力训练方法研究

12.面向终端代码智能体的长程规划/编码能力优化研究

13.面向智能体的自进化学习与优化方法研究

14.多智能体协作的关键技术研究

（三）效率优化/世界模型

1.面向大模型预训练高效 Token 利用的架构与优化方法研究

2.基于 RDMA 多路复用的存储流量加速方案研究

3.基于世界模型的桌面机器人智能交互控制技术研究

（一）AI 基础模型

1.金融高精准度数值理解与分析研究

课题背景：

金融领域的核心信息载体是数值——从财报中的营收利润、资产负债率，到市场行情中的涨跌幅、换手率，再到宏观指标中的 CPI、M2 增速，金融决策的每一步都高度依赖对数值的精准理解与推理。然而，当前大模型在金融数值处理上有三个关键问题：第一，数值感知力度不足，模型倾向于将"3.14%"与"3.1%"视为近似等价，而金融场景中 1 个基点的差异即可能影响亿级资金决策；第二，数值推理链脆弱，面对"同比增长 12.3%，环比下降 3.7%，求季度年化增速"这类多步复合运算，模型错误率上升，且难以识别自身推理过程中的数值漂移；第三，跨文档数值关联能力缺失，金融分析常需从不同来源的研报、公告、行情数据中提取数值并交叉验证，现有模型缺乏跨上下文的数值一致性校验机制。

蚂蚁业务场景对数值处理的准确性与可验证性提出了更高要求——风控模型中的违约概率计算、投研系统中的估值建模、合规审查中的限额校验，均要求模型具备基点级精度与可验证的推理链路。

本研究以金融数值的精准理解与高可靠推理为核心目标，构建覆盖数值感知、数值计算、数值推理、数值校验四个层次的金融数值能力体系，研发面向金融场景的高精度数值推理框架与评

测基准，实现金融大模型在数值处理能力上的系统性突破。

目标产出：

1)报告：提交项目结题表，提交结项验收报告。

2)源代码：1）金融数值专用 Tokenizer 与结构化解析器、数值嵌入增强模块、分解-计算-校验三阶段推理框架完整实现；2）Program-of-Thoughts 数值计算引擎、数值一致性校验器；3）FinNumBench 评测基准（覆盖财报解读、行情分析、风控评估、合规校验四大子场景的多难度梯度评测集）与自动化评测流水线；4）金融数值训练数据合成管线。

3)专利：协助甲方和/或其关联公司提交国内或国际专利 1 项，交付符合本项目创新点的专利申请号和专利申请相关材料，专利面向金融大模型的高精度数值推理方法与校验机制。

4)论文：投稿甲方认可的期刊/会议论文 1 篇并最终收录，提供论文底稿及收录邮件。论文聚焦金融场景下大模型数值感知增强、程序辅助推理与多层级校验的系统性研究。

5)技术指标：金融数值解析准确率≥99%

6)数值推理计算准确率≥95%

7)数据合成准确率≥98%

8)FinNumBench 评测集产出&核心 4 类场景每类≥200 题

2.泛模态大模型——文本/视听/运动学/电生理学等数据的统一表征和跨模态对齐研究

课题背景：

可穿戴设备（智能眼睛、智能手表等）和机器人在物理世界中会产生大量高频、连续、异构的多模态传感数据，包括图像、音频、IMU、PPG、EMG、压力传感、触觉信号等。当前主流的多模态大模型（MLLMs）主要聚焦于“文本-视觉-音频”模态的联合建模，对于运动学、电生理等非视听传感器信号缺乏统一建模与深度语义理解能力，难于满足智能穿戴、机器人交互等复杂场景需求。如何突破当前“视-听-文"模态对齐的局限，建立面向广泛异构时序传感器数据的统一表征体系，实现非视听模态与预训练大语言模型语义空间之间的高效对齐，是构建“全感官智能”的重要基础与核心技术瓶颈。

本课题需重点突破如下关键问题：

1)面向异构时序信号（不同采样率、不同时序长度、不同噪声特性）的统一时序编码机制/Tokenizer，实现非视听模态向统一语义空间的映射；

2)面向泛模态大模型训练的数据集构建与数据工程体系，研究多源异构传感数据采集、时间同步、弱监督自动标注与高质量指令数据构建方法，以及 Benchmark 评测体系；

3)面向非视听模态与文本-视觉-音频语义空间之间的跨模态对齐问题，开发基于对比学习、跨模态蒸馏、多阶段训练策略、

Instruction Tuning 等的统一训练框架。

目标产出：

1)报告：提交项目结题表及结项验收报告。报告需系统阐述项目在泛模态统一表征架构、非视听模态编码/Tokenizer 设计、跨模态语义对齐机制、多模态行为理解、上下文推理与意图识别等方面的创新成果，并包含完整的技术路线、实验验证、性能评估及未来演进建议。

2)论文：投稿甲方认可的期刊/会议论文 1-2 篇并最终收录，提供论文底稿及收录邮件。

3)专利：协助甲方和/或其关联公司提交国内或国际专利 1 项，交付符合本项目创新点的专利申请号和专利申请相关材料。

4)算法原型：完成“面向全感官感知的统一泛模态大模型”算法原型设计，交付模型源代码、训练/与推理脚本、评测脚本、数据处理工具链、训练数据集、模型权重（checkpoint）、部署文档与运行说明。

5)技术指标：在保持原有视听理解能力基础上，基于公开数据集（如 MMAct、UTD-MHAD、Ninapro、CapgMyo、WISDM、PAMAP2 等）及自建多模态数据集开展实验验证，相比 ImageBind、OneLLM、AnyMAL 等现有跨模态统一表征方法，在跨模态检索（Recall@1/5/10）、行为识别（F1-score）、上下文意图预测（AUROC）、连续行为检测 mAP、跨模态语义生成与事件描述

（BLEU/CIDEr）等关键指标上提升≥5%。在甲方和/或其关联公司智能终端支付相关真实场景（如可穿戴支付、碰一下支付等）中完成验证，实现端到端意图理解准确率较现有算法提升 5%以上。

3.ModelGrow:面向在线反馈闭环的大模型持续演化方法研究

课题背景:

当前大模型正向通用智能演进,但面临能力固化与环境适配性不足的根本性挑战:模型在预训练完成后即能力封闭，无法基于实际应用中的反馈持续优化,导致在动态场景下性能快速衰退。这使得模型部署后面临"能力固化-性能下降-重复训练"的困境，每次环境变化都需耗费大量资源重新训练,严重制约了大模型的可持续应用。

现有模型学习范式存在显著局限:离线预训练模式依赖静态语料库，训练完成后即能力冻结，无法适应线上数据分布漂移;监督微调模式需要持续的人工标注，成本高昂且更新周期长，难以响应实时场景变化;参数高效微调方法虽降低了训练成本，但仍依赖离线数据集，缺乏完整的在线反馈闭环机制，无法将用户交互、执行结果等线上反馈实时转化为模型能力的持续提升。

这导致已部署模型在实际应用中快速"老化":无法学习新知识、适应新表达、修正错误模式，每次需求变化都需人工重新标注数据并触发训练流程，模型迭代周期长达数周甚至数月，远无法满足动态环境下的敏捷需求。

近年来，持续学习、在线学习等技术为模型注入演化能力提供了新思路，已有研究在增量学习、经验回放等方面取得进展。但现有研究仍存在关键空白：一是缺乏系统化的线上反馈闭环构建方法，反馈数据质量难以保障；二是演化过程面临灾难性遗忘与能力退化风险，缺少安全可控的演化策略；三是缺少轻量化、低成本的在线演化方案，难以在资源受限的生产环境部署。随着大模型在智能客服、内容生成、决策辅助等领域的规模化应用，构建面向生产环境、基于实时反馈、支持安全可控持续演化的模型方法体系已成为学术界和产业界的迫切需求。

目标产出：

1)报告：提交项目结题表，提交结项验收报告。

2)论文：投稿甲方认可的期刊/会议论文不少于 1 篇并最终收录，提供论文底稿及收录邮件。

3)专利：协助甲方和/或其关联公司在公司系统内提交国内专利不少于 2 项，交付符合本项目创新点的专利申请号和专利申请相关材料。

4)算法原型：研究成果在灵光-主对话业务中落地，交付完整源代码。

5)技术指标：业务 bench（对话能力+通用能力）和通用主流 bench 评估，超过同期旗舰开源模型（K2.5、GLM5.1 等）1%以上。

4.低延迟流式交互场景下语音-歌声跨模态统一表征音色转换通用基础模型研究

课题背景：

音色转换旨在迁移说话人音色、保留语言内容，是智能语音交互系统的核心能力之一。随着 AI 语音服务规模持续扩大，真实场景对音色转换提出了更高要求：既要处理日常对话语音，也要覆盖歌唱演绎场景，但现有技术形成明显的"模态割裂"困境——语音转换（VC）模型无法应对歌声中的宽音域跨度与稳定音高控制；歌声转换（SVC）模型因过拟合乐曲特性，在自然语音上泛化严重不足。

跨性别转换是两类场景共同面临的核心难点：男转女需基频大幅上移并伴随音色亮化，女转男需音色增厚与共鸣下沉，低音域到高音域的宽跨度迁移尤为薄弱，现有模型难以在保持自然度的同时实现准确音域映射。与此同时，大规模语音交互平台对实时流式处理与低延迟部署有严格约束，现有两类模型均缺乏针对性的系统设计。上述现状表明，构建统一语音与歌声模态、支持宽域音高迁移、兼顾实时流式推理的基础模型已具备迫切的现实意义。

目标产出：

1)报告：提交项目结题表，提交结项验收报告。

2)论文：投稿甲方认可的期刊/会议论文 1 篇并最终收录，提

供论文底稿及收录邮件。

3)专利：协助甲方和/或其关联公司提交国内或国际专利 2 项，交付符合本项目创新点的专利申请号和专利申请相关材料。

4)算法原型：完成统一语音与歌声模态音色转换基础模型的算法设计，交付一套同时适用于自然语音对话场景与歌唱演绎场景的跨模态音色转换推理架构，及其源代码、权重等 checkpoint。

5)技术指标：研究基于跨模态统一表征与宽域音高迁移方法，形成可同时处理自然语音对话与歌声两类场景的统一音色转换基础模型，并支持 L20 单卡云端流式部署。具体指标:(1)跨性别宽域音高映射准确率，在跨度≥1.5 个八度的转换场景下较现有算法提升 15%;(2)说话人相似度 SV Similarity 较现有跨模态方法平均提升 10%，且在噪声(SNR=5dB)、低码率信道压缩、跨口音扰动下相对下降≤15%;(3)实时率 RTF≤0.3，首包延迟≤300ms，L20 单卡并发≥2 路。带伴奏歌曲转换作为扩展验证场景，采用前置人声分离+音色转换+重混音的级联方案。

5.面向 Visual Guidance 的多模态理解生成统一模型研究

课题背景：

当前多模态大模型正从单一的图文理解或文本生成，向“理解与生成一体化”的方向快速发展。尤其在 visual guidance 场景下，模型不仅要理解用户当前输入的图像、视频画面、界面截图及任务上下文，还要在此基础上生成面向具体任务的可执行指导

信息，如文字说明、步骤拆解、区域标注、箭头提示和交互反馈等，这对模型的统一建模能力提出了更高要求（典型应用场景：教育辅助、工业维修、软件教学、医疗解释、日常导航等）。

然而，现有方法仍存在四方面不足：首先，理解与生成能力相互割裂，多数模型只能分别处理视觉理解或内容生成任务，难以形成从视觉感知、语义解析到生成控制的统一闭环；其次，面向任务的视觉引导能力有限，对关键目标、操作区域、部件状态和步骤顺序等信息的响应不够精确，容易出现定位不准、步骤缺失和指导不稳定等问题；再次，跨任务统一训练机制不成熟，面对理解、编辑、生成等多类任务时，往往依赖任务拼接或流水线方案，导致模型泛化能力弱、迁移成本高、实际应用效率受限。最后，面向 Visual Guidance 的高质量数据资源仍较为缺乏，现有公开数据多聚焦于视觉问答、图像描述或通用视频理解，缺少覆盖真实任务场景、操作步骤、关键区域标注和多轮交互过程的系统化数据，难以有效支撑模型训练、评测与应用落地。

因此，研究面向 Visual Guidance 的多模态理解生成统一模型，突破任务场景建模、目标定位、步骤规划与跨任务协同优化等关键问题，已成为构建下一代交互式多模态 AI 助手的重要方向。

目标产出：

1)报告：提交项目结题表，提交结项验收报告，系统总结面

向 visual guidance 的多模态理解生成统一模型研究成果，包括模型框架设计、任务场景建模方法、目标定位与指导生成机制、数据构建方案、统一训练机制、实验评测结果及典型应用验证情况。

2)论文：投稿甲方认可的期刊/会议论文 1 篇并最终收录，论文内容围绕面向 visual guidance 的多模态理解生成统一建模方法、任务理解与可视化指导生成机制、数据构建和跨任务协同训练策略等核心创新点展开，提供论文底稿及收录邮件。

3)算法原型：完成面向 Visual Guidance 的多模态理解生成统一模型设计，交付一套支持任务场景理解、目标定位与图文指导生成的算法原型，包括模型权重、训练代码、推理代码及使用说明。该原型应支持图像、视频、界面截图、文本指令、区域交互、历史上下文等多类输入信息，能够完成任务理解、关键目标定位、步骤生成、图文指导、多轮交互辅助等任务，并具备统一任务接口和可扩展训练推理框架。

6.基于多模态大模型的 OCR 关键技术研究

课题背景：

传统 OCR 的局限性，传统 OCR 系统主要关注图像到文本的“识别”过程，即从图像中提取文字内容，但缺乏对上下文语义的理解与推理能力。例如，面对表格、票据、手写体或低质量图像时，识别结果可能准确但语义不连贯。

多模态大模型的兴起，现代能够联合处理图像与文本信息，

在视觉问答（VQA）、图像描述、跨模态检索等任务中表现出强大能力。这为 OCR 从“识别”迈向“理解与生成”提供了新范式。但当前多模态大模型仍然普遍存在细粒度文字感知能力不足、复杂版面与长文档结构建模薄弱、长尾场景泛化能力差、专业领域适配不足（医疗/金融场景）以及评测体系不完善等问题。针对上述问题，本课题拟研究一种面向复杂文档与场景文字的多模态 OCR 统一增强方法：通过融合高分辨率视觉编码、文字区域感知、版面结构建模、跨模态对齐优化、领域指令微调与长尾数据增强等关键技术，提升模型在手写、竖排、书法、古文、生僻字、小语种、医疗文档和金石书画等复杂场景下的识别、理解、抽取与生成能力。

目标产出：

1)报告：提交项目结题表，提交结项验收报告。

2)论文：投稿甲方认可的期刊/会议论文 1 篇并最终收录，提供论文底稿及收录邮件。

3)专利：协助甲方和/或其关联公司提交国内或国际专利 2 项，交付符合本项目创新点的专利申请号和专利申请相关材料。

4)技术指标：支持百灵多模态大模型在场景文字理解、医疗文档分析、金石书画等场景，响度 Baseline 提升 10pt 以上，达到行业 SOTA；

5)数据合成工具：交付 OCR/Doc 类数据渲染仿真工具链，

并提供基于该工具链产出数据的数据有效性报告；

6)评测集：产出不低于 10 个 OCR 专题测试集（和公开测试集去重），覆盖手写、竖排、书法、古文、生僻字、多项小语种等场景。

7)算法原型：研究成果在业务中落地，交付完整源代码。

7.评价驱动的细粒度可控 UI 富信息图生成与编辑技术研究

课题背景：

本项目旨在解决当前生成式大模型在 UI 设计等富信息图场景中存在的“控不准（缺乏细粒度干预）”与“判不清（缺乏垂直领域评价标准）”两大核心痛点。其中细粒度可控表现为：图像内容布局、风格、色调、文字内容等可精准控制

技术路线将围绕“构建标准-精准生成-闭环优化”展开：

1)构建评价基石：针对 UI 界面等富信息图的高度结构化与功能性特征，率先构建包含美观度、布局合理性、设计一致性的多维美学评价数据集与自动化打分模型。

2)突破可控生成：引入布局约束与组件语义增强，研发支持图层级二次编辑和属性精确干预的细粒度生成框架，解决 AI 产物在实际生产链路中的可用性问题。

3)评价驱动反馈闭环：将自动化美学评价模型作为强化学习的奖励函数（Reward Model）或生成过程的引导条件，探索“评价驱动优化”的协同生成机制。

4)通过“评价体系”为“生成系统”提供可靠的进化方向，以“可控生成”将“审美标准”具象化为高保真、可编辑的企业级设计资产，最终打通自动化、智能化的现代 UI 生产全链路。

目标产出：

1)报告：提交项目结题表，提交结项验收报告。

2)论文：投稿甲方认可的期刊/会议论文 2 篇并最终收录，提供论文底稿及收录邮件。

3)专利：协助甲方和/或其关联公司提交国内或国际专利 2 项，交付符合本项目创新点的专利申请号和专利申请相关材料。

4)算法原型及技术指标：完成可控 UI 生成与编辑技术的代码模型，以及不少于 2k 的相关评测数据集，覆盖上述提及 4 种以上细粒度可控任务数据集；提出的算法在相关评测指标相比同期百灵多模态大模型在公开测试集合 IGBench 和自建评测数据集的基线提升 15%。

8.基于图文结构化的语料合成研究

课题背景：

当前，多模态大模型展现出强大的图像理解与推理能力，其性能飞跃高度依赖于高质量、多样化、细粒度的图文预训练语料。然而，真实世界中的图文数据（如网页、书籍、图表截图）普遍存在以下问题：结构性缺失：图文之间多为隐性关联，缺少显式

的实体对齐、空间关系、视觉属性等结构化标注。任务单一：现有语料多用于图文匹配或描述生成，难以支撑复杂的知识问答、视觉推理、实体定位等任务。人工标注成本极高：对实体位置、属性、逻辑关系等细粒度信息的标注需要大量专业人力，难以规模化。因此，一个重要的研究思路是：从现有图文数据中自动或半自动地提取结构化信息（实体、位置、关系、视觉属性等），并以此为基础，可控地合成多种任务类型的训练语料。该方向有望低成本、高效地突破多模态大模型在推理与定位能力上的数据瓶颈。

目标产出：

1)报告：提交项目结题表，提交结项验收报告。

2)论文：投稿甲方认可的期刊/会议论文 1 篇并最终收录，提供论文底稿及收录邮件。

3)专利：协助甲方和/或其关联公司提交国内或国际专利 1 项，交付符合本项目创新点的专利申请号和专利申请相关材料。

4)算法原型：完成的语料合成引擎代码，覆盖知识问答、视觉推理、实体定位等方向关键评测集与评测工具，并合成相关训练语料规模 1000 万。

5)技术指标：在自建知识问答、视觉基础任务等综合 benchmark 上，支撑自研 VL 模型能力在准确率、F-score 等关键指标上，达到同尺寸模型(Alibaba/Qwen 系列模型、

Google/Gemma 系列模型)的 SOTA。

9.基于大模型长程记忆与动态工具调用的心理健康干预系统研究

课题背景：

当前，大模型在精神心理领域的应用已初步实现多轮对话与语音交互，但在实际业务中遭遇了显著的“效能瓶颈”：现有系统多局限于提供模板化的情绪安抚与表层倾听，缺乏明确的心理干预目标与循证路径，导致用户体验停留在“不痛不痒”的陪伴阶段，难以实质性解决用户的核心心理诉求。

真实的心理健康干预不仅需要自然共情的对话，更需要在对话中精准锚定问题，并动态调度专业评估与干预工具（如心理量表、CBT 减压互动、正念音频等），形成“评估-干预-反馈”的闭环。这要求系统必须具备“长程记忆”以跨周期追踪用户的症状演变与心理基线，避免由于缺乏上下文导致的重复叙述与无效诊断；同时，亟需构建“动态工具调用”机制，根据用户当下的情绪状态自适应调配临床循证干预武器，从而真正将“共情聊天”转化为具备医疗级严谨性的定制化治疗方案。

因此，研究一套兼具精准意图理解、长程状态追踪与动态工具调用的高能效心理干预系统，实现从“被动式倾听”向“主动式干预”的跨越，是当前数字心理医疗亟待突破的技术难题。

目标产出：

1)报告：提交项目结题表，提交结项验收报告。

2)论文：投稿甲方认可的期刊/会议论文 1 篇并最终收录，提供论文底稿及收录邮件。

3)专利：协助甲方和/或其关联公司提交国内专利 1 项，交付符合本项目创新点的专利申请号和专利申请相关材料。

4)算法原型：完成基于临床心理学的大模型评估标准体系（Rubric）建设。并以此标准为基准，交付一套在该评测体系下达到 SOTA（当前最优）性能的模型。

5)技术指标：精神心理领域的正确性得分提升 10%，响应延迟（Latency）与现有线上大盘基线持平，在实现闭环心理干预的同时，不额外增加业务系统的性能负担。

（二）Agentic AI/AI Agent

1.面向复杂任务 Agent 的自适应检索规划与多轮知识获取方法研究

课题背景：

复杂任务 Agent 在企业知识问答、Deep Research、智能客服、业务分析等场景中，需要持续获取、验证与融合外部知识。当前 Agentic RAG 与 Deep Research 系统在通用任务上已有进展，但面向真实复杂任务仍存在三类核心挑战：(1)检索规划与推理过程缺乏联合优化，长链路任务中检索冗余、关键证据缺失；(2)多源异构证据融合与可验证归因不足，难以满足高合规场景；(3)缺少质量-成本-延迟的显式 trade-off 建模，工业级 SLA 难以保障。

本课题通过学术合作，结合高校/科研机构在信息检索、强化学习、自然语言处理与可信 AI 方向的研究能力，以及企业在 Agent 和 RAG 平台中的真实场景与工程经验，探索推理驱动的检索规划与策略学习、质量-成本-延迟 Pareto 优化，以及多源异构证据融合与可验证归因。合作成果有助于提升 Agent 的知识获取质量、复杂任务完成能力、结果可信度和平台技术影响力。

目标产出：

1)报告：提交项目结题表，提交结项验收报告。

2)论文：投稿甲方认可的期刊/会议论文 1-2 篇并最终收录，提供论文底稿及收录邮件。

3)专利：协助甲方和/或其关联公司提交国内专利 1-2 项，交付符合本项目创新点的专利申请号和专利申请相关材料。

4)算法原型：形成一套面向复杂任务 Agent 的推理驱动的检索规划方法，建设可集成至 RAG 平台的原型系统，并在蚂蚁内部的企业知识问答、Deep Research 的典型场景中完成验证。

5)技术指标：在蚂蚁内部的企业知识问答、Deep Research 及公开基准 DeepSearchQA 上开展系统对照评测并有所提升（BrowseComp 指标不低于当前基线），达成以下要求：

任务质量：在蚂蚁内部场景上，复杂任务端到端完成率较强 Agentic RAG baseline 相对提升≥15%;多跳 QA F1 相对提升≥10%。

成本与可信：同等效果分数下，token 消耗与幻觉率(内容+归因综合)均≤0.9×强 Agentic RAG baseline；高合规场景中推理链可追溯。

系统集成：形成可复用核心组件并完成 RAG 平台集成；在 DeepSearchQA 上与同期开源 SOTA 方法（Search-R1、ReSearch）达到相当水平，在 BrowseComp 上不低于强 Agentic RAG baseline。

2.面向 Agentic 的 RL 与 OPD 协同后训练系统研究

课题背景：

以 ChatGPT Codex、Claude Code 为代表的 AI 智能体展示了大语言模型通过多轮环境交互（代码执行、工具调用、网页操作）

完成复杂任务的能力。在 Agentic 复杂场景下，如何高效训练出强智能力的智能体，成为一个亟待解决的问题。当前主流的两类后训练系统中，RL 以环境奖励为驱动，擅长在开放动作空间中发现新策略、突破专家能力上界，但探索成本高、收敛慢，长程任务中奖励稀疏与方差大的问题尤为突出；OPD 由训练好的专家模型对在线策略进行同分布蒸馏，收敛快、训练稳定、样本效率高，擅长快速对齐已知能力，但效果上界受限于专家模型。二者在能力获取路径上具有天然互补性：OPD 提供"快而稳"的能力继承，RL 提供"慢而广"的能力拓展。本项目希望围绕这一互补性，研究 RL 与 OPD 的协同训练机制，研究内容包括但不限于：

1)能力边界扩展：研究如何在 OPD 提供稳定基线的同时，借助 RL 突破专家能力上界，特别是在长程任务、复杂工具组合、跨轮规划等专家覆盖不足的场景中实现能力跃升。

2)协同机制创新：探索两种范式在 agentic 训练中的分工与融合，例如阶段切自适应切换、混合采样等。探索融合范式下，后训练系统的架构设计。

3)闭环监控与在线反馈：构建 RL+OPD 的 agentic 训练过程中的系统化监控体系，覆盖 teacher/student 模型轨迹质量追踪、奖励稳定性等关键信号，支持异常定位与协同训练策略的在线调整，形成"监控—诊断—调参"闭环。

目标产出：

1)报告：提交项目结题表，提交结项验收报告。

2)论文：投稿甲方认可的期刊/会议论文不少于 1 篇并最终收录，提供论文底稿及收录邮件。

3)专利：协助甲方和/或其关联公司提交国内或国际专利不少于 1 项，交付符合本项目创新点的专利申请号和专利申请相关材料。

4)算法原型：完成 RL-OPD 协同 agentic 训练原型系统的设计，源代码，设计文档、使用文档等材料。

5)技术指标：

6)研究基于 RL+OPD 协同的 agentic 训练方法。相全 RL 训练，swe 场景下模型分数上涨速度提升 1.5x

7)长程任务成功率端到端成功率提升 30%，专家任务上相比纯 OPD 训练，成功率提升 50%

3.AI 可控可靠保障之多纬度智能体可控性评测框架研究

课题背景：

大语言模型驱动的智能体在代码生成、工具调用、任务规划等领域展现出强大能力，正逐步应用于自动化运维、数据分析、智能辅助等实际场景。然而，随着 Agent 自主性的增强，其行为的可控性成为制约应用落地的关键挑战。

当前 Agent 评测研究主要聚焦于端到端任务成功率，如

SWE-bench 评估代码修复能力、AgentBench 评估多轮对话与工具使用能力。这类评测关注"任务是否完成"，却忽视了"执行过程是否可控"这一核心问题。实际场景中，Agent 可能跳过关键步骤、执行冗余操作、违反安全约束，即使最终任务看似完成，其执行过程却存在不可预测、不可追溯的风险。

特别是在多任务、多智能体协作场景下，可控性问题更加复杂：单个 Skill 的执行偏差可能级联放大，多 Agent 协作中的责任边界难以界定。现有评测体系缺乏对执行轨迹可控性、约束遵守能力的系统评估，难以支撑可信 Agent 的研发与部署。

因此，本项目提出 Agent 可控性评测基准，从执行轨迹可控性与约束合规性两个维度，构建覆盖单 Skill、单 Agent、多 Agent 的多粒度评测框架，为可信智能体的研究与落地提供标准化评估工具。

目标产出：

1)报告：提交项目结题表与结项验收报告，包含评测框架设计文档、实验数据完整记录、技术方案总结等材料。

2)论文：投稿甲方认可的期刊/会议论文 1 篇并最终收录，提供论文底稿及收录邮件。论文主题聚焦 Agent 可控性评测基准。

3)专利：协助甲方和/或其关联公司提交国内或国际专利不少于 1 项，交付符合本项目创新点的专利申请号和专利申请相关材料。

4)算法原型：完成 Agent 可控性评测算法设计，交付一套覆盖单 Skill、单 Agent、多 Agent 场景的评测框架，包含评测数据集、评分算法、可视化分析工具及其完整源代码。

5)技术指标：研究基于轨迹匹配与约束检测的可控性评测方法，形成标准化的四维评分体系（完整性、顺序性、专注度、合规性）。构建包含 50+典型场景的评测数据集，实现对主流 LLM Agent 的可控性量化评估。在评测准确率方面，与人工标注对比达到 85%以上的一致性；在评测效率方面，单 case 评测耗时较人工评估缩短 90%以上。支持扩展至企业内部 Agent 场景，支撑线上业务的风险管控与合规审计。

4.面向广告投放的 Agentic RL 决策优化框架：从智能竞价到全链路自主投放研究

课题背景：

行业演进与核心挑战是数字化营销正经历从“经验驱动”向“智能驱动”，进而向“自主代理（Agentic）驱动”的范式转移。尽管自动出价与智能助手已普及，但在复杂动态环境中，广告主仍面临两大核心瓶颈，制约了系统向更高阶智能演进：

短期博弈与长期价值的失配（针对自动出价）：现有出价模型多基于即时反馈（如点击、短期转化），存在显著的“短视”缺陷。在用户全生命周期价值（LTV）挖掘及长周期预算平滑控制上，传统 RL 或监督学习难以建立长程规划能力，导致在非平

稳竞价环境中策略震荡，无法实现全局最优。

建议与效果的归因断裂（针对智能助手）：以“小灵”为代表的 LLM Agent 虽具备推理能力，但其生成的营销策略与建议缺乏有效的效果验证闭环。由于广告转化存在天然延迟，Agent 的建议往往在数天后才能体现真实 ROI，导致中间缺乏细粒度的奖励信号。这种“反馈断层使得 Agent 无法通过交互数据自主进化，陷入“懂逻辑但不懂优化”的困境。

蚂蚁广告现有系统的关键瓶颈是当前灯火平台在迈向 Agentic 驱动时，主要卡在以下两个具体环节：

1.自动出价（AutoBidding）：缺乏全局视野与长期规划能力

短视决策：现有出价模型主要解决局部博弈，未能具备长程规划能力，导致效率低下。

响应滞后：基于历史数据训练的静态策略无法适应实时竞价环境中的非平稳分布，面对突发流量或竞品激进策略时调整缓慢。

约束平衡僵化：在多目标优化中，传统方法难以在保障 ROI 底线与最大化 GMV 之间实现动态帕累托最优，缺乏自适应的安全探索机制。

2.智能助手（小灵）：缺乏基于真实反馈的进化闭环

反馈断层：当前 Agentic Workflow 虽引入了 DeepResearch 与 Reflection 机制，但 Agent 的建议与数天后的真实业务结果（如最终 ROI）之间存在严重的归因断裂，无法形成有效的奖励信号。

适应性受限：基于规则编排的 Workflow 难以泛化至 unseen

场景，面对新业务形态需人工重新设计流程，不具备从交互数据中自主进化策略的能力。

“懂逻辑不懂优化”：LLM 具备强大的推理逻辑，但缺乏针对特定业务目标的数值优化能力，导致建议往往“正确但非最优”。

技术趋势与前沿佐证方面，业界正加速将强化学习（RL）与 LLM-based Agent 深度融合，以实现从“静态指令执行”到“自主决策进化”的跨越：Meta：通过 RLPF 框架将广告文案生成轨迹与 CTR 反馈直接对齐，实现 CTR 提升 6.7%，验证了 RL 对生成式内容的优化潜力。快手：LBM 架构采用分层推理-执行机制，利用 MBRL（Model-Based RL）解决自动出价中的样本效率问题。学术界前沿：SIGIR 2025 最新工作（如 GAVE,L2A）进一步证实，将 Agent 决策轨迹与全生命周期业务效果通过 RL 进行端到端对齐，是突破现有瓶颈的关键路径。

核心研究模块：

1.高精度延迟反馈归因模块（Phase 1 重点）

目标：解决广告转化延迟导致的奖励稀疏与信用分配难题，为 RL 提供准确的即时奖励估计。

2.面向长期规划的序列决策强化学习出价引擎（Phase 2 重点）

目标：赋予 AutoBidding 长程视野，解决短视问题。

3.智能助手（小灵）的反馈闭环增强机制（Phase 2 重点）

目标：打通“建议-执行-结果”的归因链路，实现 Agent 的

自我进化。

目标产出：

1)报告：提交完整的项目结项验收报告，涵盖研究背景、技术路线、算法细节、实验对比、创新点分析及后续演进路径。

2)论文：投稿甲方认可的顶级期刊或会议（如 KDD, SIGIR,WWW 等）论文不少于 1 篇并最终收录，提供论文底稿及收录邮件。(研究方向：强化学习与 LLM Agent 融合、异构轨迹表征学习、延迟反馈归因机制、多约束安全强化学习、生成式竞价决策等。交付物：论文底稿、投稿记录、审稿意见及录用证明。)

3)专利：协助甲方和/或其关联公司提交国内或国际专利申请不少于 1 项。(核心技术点:基于全局轨迹感知的智能体强化学习方法；面向广告长周期优化的延迟奖励分解与归因方法；基于生成式强化学习（GRPO/DPO）的广告动态出价决策；多约束条件下的安全探索与帕累托优化算法。交付物：专利交底书、技术方案说明书、权利要求书草案及受理通知书。)

4)算法原型：交付一套可运行的 Agentic RL 验证框架代码。包含模块：延迟归因预测模型、长期规划 RL Agent、离线反事实评估模拟器。工程交付：源代码、模型权重、训练/推理脚本、历史数据回放测试用例。

核心模块包含：

a.序列建模与全局轨迹规划模块：基于 Transformer 的用户

行为序列建模与长期价值预测。

b.延迟反馈归因模块：基于时序价值函数（Temporal Value Function）解决转化延迟带来的奖励稀疏问题。

c.生成式强化学习策略网络：集成 GRPO (Group Relative Policy Optimization)或 Iterative DPO，实现策略的高效迭代。

d.多约束安全探索模块：引入拉格朗日乘子法或保守策略更新，确保 ROI 与合规约束下的帕累托优化。

e.博弈感知动态调整模块：模拟竞争对手行为，实现非平稳环境下的自适应策略调整。

工程交付物：可运行的源代码、模型配置文件、训练/推理脚本、测试样例、模型权重及 Checkpoint 文件。

5)关键绩效指标（KPIs）：为确保项目可行性，考核指标聚焦于三大核心价值维度，并设立分级验收台阶：

离线评估：离线历史数据回放评估（Offline Evaluation）R:R*等指标。

核心指标：

a.累积转化量（体现 RL 收益）：在保持 ROI 达标的前提下，试点组相比对照组，累积转化量提升不低于 5%。

b.智能助手采纳率（体现 Agent 价值）：小灵助手基于新归因逻辑生成的策略建议，被广告主或运营人员的采纳率提升不低于 20%（相对值）。

c.平台效能（eCPM）：自动化出价系统的千次展示收益

（eCPM）提升不低于 5%。

5.Agentic Learning 在长程任务场景的技术研发与应用落地

课题背景：

在智能体迈向生产环境的进程中，长程任务的执行可靠性是衡量 Agentic Learning 落地价值的核心指标。针对实际工程中频繁出现的任务“断头”、上下文幻觉及训练成本高昂等问题，本项目意在研究一种系统性的长程任务解决方案，具体的挑战有：

1.长程任务背景下 Harness/sandbox 的开发和性能提升，包括但不限于智能体的定义、执行 trajectory 格式标准化、环境维度的 sclaing law 发现等；

2.SFT/RL 等算法的研发和创新，核心解决在长上下文背景下 token level 的 credit assignment 问题；

3.Reward/LLM as Judge 的研发和创新，包括但不限于长周期任务的高效评测、任务失败归因、rubric 生成等问题。

本研究不仅关注算法突破，更致力于为智能体在长程业务自动化、科研协同等真实场景中的低延迟、高可靠落地提供工程化支撑。

目标产出：

1)报告：提交项目结题表，提交结项验收报告。

2)论文：投稿甲方认可的期刊/会议论文 1 篇并最终收录，提

供论文底稿及收录邮件。

3)专利：协助甲方和/或其关联公司提交国内或国际专利 1 项，交付符合本项目创新点的专利申请号和专利申请相关材料。

4)算法原型：长程任务训练 pipeline,对比传统基于 verl 或者 slime 等 RL 开源训练框架训练效率提升 20%+；以及对应的模型在灵光等业务场景落地，核心用户留存等指标置信提升 5%+。

6.基于多模态大模型的通用游戏 Agent 构建和训练研究

课题背景：

传统游戏 AI 通常采用游戏规则硬编码或特定游戏强化学习的方式，导致每个游戏都需要单独开发 AI 系统，缺乏通用性和可迁移性。随着多模态大模型的快速发展，特别是视觉-语言-动作（VLA）模型的突破，通用游戏 Agent 得以实现：它能理解游戏画面、理解自然语言指令并自主决策，通过结合大模型的推理能力和强化学习的决策能力，实现跨游戏、跨类型的通用智能。

这类 Agent 的独特价值在于不仅具备自主决策能力，更拥有对游戏进行实际体验与验证的能力。这一特性恰好对应闪游戏业务的核心痛点：当前批量生成游戏场景中，模型盲目生成代码、无法实际体验生成游戏，导致生成质量难以保障。通用游戏 Agent 可作为自动化测试工具，对生成的游戏进行实际体验和验证，从而解决这一难题。

研究目标是构建基于多模态大模型的通用游戏 Agent 基础

架构，支持视觉理解、决策推理和动作执行，实现跨游戏类型的知识迁移和持续学习能力，建立游戏 Agent 的自动化训练与评估体系。研究内容：

1. 多模态游戏理解：研究游戏画面的视觉理解、UI 元素识别、游戏状态建模方法

2. 通用决策框架：设计基于 LLM 的游戏决策引擎，支持任务规划、动作生成和策略推理

3. 知识迁移机制：探索跨游戏的知识表示和迁移学习方法，提升 Agent 的泛化能力

4. 自动化测试应用：将 Agent 应用于游戏功能测试、性能测试和压力测试场景

目标产出：

1)报告：提交项目结题表，提交结项验收报告。

2)论文：投稿甲方认可的期刊/会议论文 1 篇并最终收录，提供论文底稿及收录邮件。

3)专利：协助甲方和/或其关联公司提交国内发明专利 1 项，交付符合本项目创新点的专利申请号和专利申请相关材料。

4)算法原型：完成基于多模态大模型的通用游戏 Agent 算法设计，交付一套适用于休闲类、策略类、动作类游戏的通用 Agent 框架，及其源代码、模型权重等 checkpoint。

5)技术指标：构建游戏测试 benchmark，提升游戏任务完成

准确率至 80%以上，操作延迟 500ms 以下，形成至少跨 3 款游戏以上泛化的 Agent 知识迁移技术。在闪游戏业务场景沉淀至少一类游戏类型的自动化测试方案，测试效率接近人类速度。

7.面向复杂任务的多模态 Agent 研究

课题背景：

多模态大模型正由图文匹配、视觉问答等静态任务，向面向真实环境的复杂任务执行演进。在网页浏览、桌面软件操作、代码辅助执行等数字环境中，智能体不仅需要理解视觉与文本信息，还需要围绕用户目标开展多步规划、调用工具、依据环境反馈动态调整策略，并完成可验证的任务目标。

本课题所关注的“多模态”主要指视觉与文本两类核心模态，并结合任务执行过程中的环境状态与工具反馈信息；所研究的“复杂任务”主要指数字环境中的多步骤任务执行场景，包括网页操作、图形界面交互、工具辅助任务完成以及代码与软件环境中的复合型任务求解。相较于纯文本 Agent，此类任务更依赖视觉感知能力，并对任务规划、工具调用、环境交互、状态追踪和结果验证能力提出更高要求。

当前，相关研究仍面临两方面问题：一是现有评测基准多聚焦于纯文本智能体、静态视觉任务或受限环境下的局部交互能力，缺乏面向真实数字环境复杂任务的系统评测环境与评测集；二是在训练与优化层面，现有方法对于复杂任务中的闭环执行能力建

模不足，特别是在异构 Reward 条件下的稳定优化仍存在挑战。这里的“异构 Reward”主要指来源不同、粒度不同的奖励信号，包括任务结果奖励、步骤完成奖励、工具调用有效性奖励和环境反馈奖励等。由于不同奖励在稀疏性、时序位置和优化目标上存在差异，容易造成信用分配困难和优化方向不一致，从而影响训练稳定性与最终效果。

基于上述背景，本课题拟围绕多模态 Agent 的评测与优化两条主线展开研究：一方面构建具有真实任务代表性的评测环境、评测数据与评测方法，形成统一 benchmark；另一方面研究面向复杂任务的多模态 Agent 训练与优化方法，提升智能体在真实场景中的任务成功率、执行稳定性和工具调用准确性。

目标产出：

1)报告：提交项目结题表及项目结项验收报告。

2)论文：围绕多模态 Agent 评测基准构建、评测方法设计及训练优化方法等方向，投稿甲方认可的期刊/会议论文 1-2 篇并最终收录，提供论文底稿及收录邮件。

3)专利：围绕本课题形成的关键创新点，协助甲方和或其关联公司提交国内或国际专利 2 项，交付专利申请号及相关技术材料。

4)算法原型及技术指标：交付面向数字环境复杂任务的多模态 Agent 算法原型及配套代码，包括评测环境与评测集构建工具链、训练数据合成方法、异构 Reward 条件下的训练优化算法以

及评测分析工具。以项目初始原型和公开可复现的代表性相关方法作为基线，在本课题构建的自建 Benchmark 及相关公开评测任务上，围绕任务成功率、步骤完成率、工具调用准确率和结果约束满足率等核心指标开展评测；其中，在自建 Benchmark 上实现核心指标绝对值提升 5 个百分点以上，并在相关公开评测任务上达到同类方法先进水平。

8.用于科学智能的 Agentic RL 算法研究

课题背景：

在当前人工智能技术飞速发展迭代的浪潮中，科学发现正处于由数据驱动向智能驱动转型的关键转折点。长期以来，AI for Science 的研究方法主要是针对特定科学问题人工设计算法或专用模型。这类研究往往聚焦于特定学科的垂直领域，如蛋白质结构预测、天气预报或材料性质筛选，其本质是将 AI 视作一种高维函数拟合工具或特定的求解器，用以辅助完成实验数据分析、物理方程求解等科学工作流中的离散环节。尽管此类专用工具在局部领域取得了显著成效，但其局限性日益凸显：模型往往受限于特定领域的归纳偏置，缺乏跨学科的知识迁移能力，且无法理解科学研究背后的逻辑推理与假设验证过程。

未来的科学智能不应止步于单一领域或单一阶段的工具，而应成为赋能科学发现全生命周期的通用引擎。具体而言，科学智能模型应当演进为具备广博科学知识、掌握复杂研究技能并具备

科学思维方式的通用科学助手（AGI for Science）。大语言模型被认为是通用人工智能的星星之火，展现出了卓越的通用泛化性。但面向实际行动，需要模型具备很强的 agentic 能力，能够与各类 harness 协同工作。如今在 aicoding 领域，claude code 等 harness 已得到广泛使用，在个人助理领域 openclaw 取得了非常好的效果和影响力，我们相信在科学智能领域同样需要大模型与 harness 结合，来促进该领域走向实用。agentic RL 作为提升 LLM 执行能力的关键技术，近期得到广泛关注，比如在 Coding、DeepResearch 等领域取得了非常大的进步，但其对上下文窗口的消耗，以及 auto-regressive 的执行方式导致在解决复杂问题时效率仍然较低，需要一些创新性的方法来解决这类问题。

目标产出：

1)报告：提交项目结题表，提交结项验收报告。

2)论文：投稿甲方认可的期刊/会议论文 1 篇并最终收录，提供论文底稿及收录邮件。

3)算法原型：重点探索基于 multi-agent、subagent 的思路，来提升模型解决复杂问题的能力和效率，相比单 agent 最长轨迹 token 量降低 20%，效果相比单 agent 提升 5%。沉淀代码，并经过系统性验证，相比 baseline 在效果和效率上均有明显提升，形成实验报告。沉淀训练数据和训练方法在蚂蚁百灵模型上有显著提升。

9.Deep Research 推理一致性与幻觉抑制优化研究

课题背景：

大语言模型正从单轮问答向多步深度研究（Deep Research）范式演进，要求模型在多源异构数据环境中完成信息检索、证据整合、多步推理与结论生成的端到端复杂任务。相比传统文本生成，DeepResearch 任务具有更强的数据异构性、推理长程性和结论严肃性——其输出质量不仅依赖语言理解与生成能力，还受到跨源信息融合、推理过程与信息源的逻辑一致性、跨源证据锚定和数值精确计算等的共同制约。

在金融研报生成、行业深度分析等严肃场景中，模型需在结构化财务数据、非结构化研报、时序行情及知识图谱等多源异构数据间进行交叉验证与深度推理，数值幻觉、证据幻觉等都可能导致决策误导，亟需从语料轨迹构造、pre-train/sft/rl 多阶段协同训练和评测体系三个层面构建系统化方案。

目标产出：

1)报告：提交项目结题表，提交结项验收报告。

2)论文：投稿甲方认可的期刊/会议论文 1 篇并最终收录，提供论文底稿及收录邮件。

3)专利：协助甲方和/或其关联公司提交国内或国际专利 2 项，交付符合本项目创新点的专利申请号和专利申请相关材料。

4)算法原型及技术指标：沉淀方法在百灵系列模型中，使用

无上下文管理的评测方法，以 BrowseComp、WideSearch、DeepSearchQA、FinSearchComp 评测多源事实搜集和推理能力，以 FACTS 评测搜索与锚定一致性，至少 2 个关键榜单中分别提升>5%，并交付对应代码与数据集。

10.LLM 对 Agent Harness 的适配能力评估与优化方法研究

课题背景：

随着 OpenClaw、Claude Code 等 Agent Harness 的普及，基模从回答问题转向执行任务。实践中发现，瓶颈之一在于基模与 Harness 之间的适配性问题，同一模型在不同 Harness 下表现波动大，难以适应动态上下文注入、分段记忆检索、渐进式工具披露等机制。模型在不同 Agent 框架下稳定执行任务的泛化能力，主要涵盖如下几个维度：上下文适应（对动态注入与冗余信息的抗干扰与提取能力）、工具调用（对异构工具的遵循与调用能力）、记忆协同（跨轮次分段记忆检索与状态维持能力）以及异常恢复（面对执行失败时的重试与容错纠偏能力）。本项目期望能够建立评测体系，对此类 Harness 适配性的系统度量，并通过针对性的训练与适配优化，模型可显著提升对复杂 Harness 机制的泛化适配性。

目标产出：

1)报告：提交项目结题表，提交结项验收报告。

2)论文：投稿甲方认可的期刊/会议论文 2 篇并最终收录，提

供论文底稿及收录邮件。

3)专利：协助甲方和/或其关联公司提交国内或国际专利 2 项，交付符合本项目创新点的专利申请号和专利申请相关材料。

4)适配性评估框架及测试集：覆盖 5 类 Harness（OpenClaw、Claude Code、Hermes 等），测试集包含不少于 1500 条真实任务用例，需覆盖长链调用、多工具组合、异常恢复等复杂场景，并引入动态环境扰动、多轮交互、工具增删、错误注入等 Agent 特性评测场景，以保障稳定性与不同场景的代表性。

5)优化方法及数据：针对评测中暴露的典型适配问题（如上下文信息利用不足、工具调用不足、缺乏重试与容错行为等），收集执行轨迹数据并进行有针对性的优化，优化后在多种 Harness 环境下的泛化效果与复杂任务成功率实现显著提升，模型效果达到业界第一梯队。

11.面向代码智能体任务的大语言模型推理与规划能力训练方法研究

课题背景：

以大语言模型（LLM）为核心的代码智能体（Code Agent）已成为 Agentic AI 领域的重要研究方向。代码智能体需要在真实软件工程环境中自主完成缺陷修复、需求实现、代码重构等复杂任务，要求模型具备长链路推理、层次化任务规划、环境反馈理解与自我纠错等综合能力。以 SWE-bench 为代表的评测基准表

明，当前模型在真实 GitHub Issue 修复任务上已取得显著进展，但在复杂软件工程场景中的自主任务完成率仍存在较大提升空间。

与传统代码生成任务不同，代码智能体场景要求模型不仅能生成正确的代码片段，更需具备在开放环境中进行多步决策、动态规划、错误感知与自主回退的能力。这些能力难以通过标准的预训练或指令微调充分习得，亟需针对性的训练方法研究。

当前模型面临的核心瓶颈在于长链路推理过程中的错误恢复能力不足。现有模型在执行多步任务时，一旦在某个环节做出错误决策（如错误定位、无效编辑），往往无法自主识别偏差并回退到正确路径，而是沿错误方向持续执行直至任务失败。这种"卡住即断"的脆弱模式严重制约了代码智能体在复杂场景下的实用性。

本课题聚焦于代码智能体长链路推理中的错误恢复与自主纠偏这一核心难题，从训练数据构造、训练目标设计和强化学习对齐三个层面，系统性地研究提升大语言模型推理与规划能力的训练方法。具体而言：通过基于 SWE-bench 失败案例的轨迹增强与修复管线构建高质量训练数据，设计感知关键决策点（搜索-定位-验证-编辑）的训练目标函数，以及探索基于子目标奖励分解的强化学习算法，从模型能力本源上提升代码智能体在复杂缺陷修复中的自主完成率与执行稳定性。

目标产出：

1)报告：提交项目结题表，提交结项验收报告。

2)论文：投稿甲方认可的会议论文 1 篇并最终收录，提供论文底稿及收录邮件。论文核心贡献聚焦于"面向代码智能体错误恢复能力的训练方法"，涵盖轨迹增强、决策点感知训练和子目标奖励 RL 三项技术。

3)算法原型：1.基于 SWE-bench 失败案例的执行轨迹增强管线，实现失败轨迹的自动诊断与修复式改写，完成增强数据集构建。2.关键决策点感知的 SFT 训练方法实现与实验验证，在 SWE-bench Verified 等基准上验证错误恢复能力的提升。3.基于子目标奖励分解的 RL 训练系统，包含过程奖励模型、回退奖励机制和渐进式难度调控策略，在 SWE-bench 上验证长链路推理稳定性的提升效果。

12.面向终端代码智能体的长程规划/编码能力优化研究

课题背景：

以命令行终端（Terminal）为交互界面的智能体（Terminal Agent）是实现通用软件工程自动化的核心桥梁。与传统的代码补全或单文件修复不同，Terminal Agent 运行在高度有状态、逻辑严密且对语法极度敏感的系统级环境中，需要协同调用如 Git、Docker、Make、SSH 等异构工具链来处理环境配置、系统运维、跨文件调试及自动化部署等任务。TerminalBench 系列等终端智

能体评测基准表明，尽管 LLM 在模拟环境下表现出色，但在处理具备长程依赖关系和高度环境不确定性的真实终端任务时，成功率仍然较低。终端环境具有反馈稀疏、状态依赖强以及错误传播效应明显等特点：一个细微的路径错误、权限配置问题或环境变量缺失，都可能导致后续任务链路执行失败。当前的智能体多依赖于 Prompt Engineering 或简单的 ReAct 框架，缺乏对系统状态深度感知的“工程直觉”（即对代码库全局状态的深层感知、对报错日志背后根因的因果推断能力，以及对执行操作潜在副作用的风险预判能力），在面对复杂报错（Stderr）时往往陷入死循环或产生破坏性操作，难以达到工业级的稳定要求。本项目拟从训练数据和强化学习的角度增强 code agent 的长程规划和错误自愈能力：

训练数据：探索可以可自动化、可规模化、可验证的复杂重点任务的合成范式，生成包含包含任务规划路径、错误恢复策略及环境状态变化过程的高质量的 mid-training/SFT 数据集。

训练方法：探索适配终端环境的 Agentic RL 算法，设计细粒度的奖励反馈模型有效解决长程规划下的奖励稀疏问题，并通过强化学习解决“幻觉命令”和“无效循环”问题。通过在模拟终端环境中的大规模试错与策略优化，使模型习得真正的“系统工程直觉”，从而在长链路任务中提升长链路任务中的规划稳定性与错误自恢复能力。

目标产出：

1)报告：提交项目结题表，提交结项验收报告。

2)论文：投稿甲方认可的期刊/AI 顶级会议论文 1-2 篇并最终收录，提供论文底稿及收录邮件。

3)专利：协助甲方和/或其关联公司提交国内或国际专利 2 项，交付符合本项目创新点的专利申请号和专利申请相关材料。

4)技术指标：沉淀一套完整的终端智能体优化方案，从数据和后训练两个方面显著提升基座的能力水位，在开源的 terminalbench 榜单上超过同尺寸开源模型（如 GLM/KIMI/Deepseek 系列）。

13.面向智能体的自进化学习与优化方法研究

课题背景：

随着智能体（Agent）在复杂任务中的应用不断拓展，传统依赖人工规则与静态策略的系统在适应性、可扩展性及长期维护成本方面逐渐面临瓶颈。本项目围绕“Agent 自进化（Self-Evolution）能力”开展研究，探索构建具备持续学习、自我优化与策略迭代能力的智能体体系。

项目通过引入基于反馈的闭环优化机制，使 Agent 能够在多轮交互与任务执行过程中进行自我评估与能力更新，从而提升其在多场景中的稳定性与泛化能力。同时，通过降低对人工干预与频繁调优的依赖，进一步提升系统运行效率与工程可维护性。

目标产出：

1)报告：提交项目结题表，完成项目结项验收报告，系统总结 Agent 自进化机制设计方法、实验验证结果及应用效果。

2)论文：投稿甲方认可的期刊/会议论文 1－2 篇并最终收录，内容围绕 Agent 自进化机制、反馈驱动优化策略及实验验证结果，提供论文底稿及收录邮件。

3)专利：协助甲方和/或其关联公司提交国内或国际专利 1－2 项，围绕 Agent 自进化框架、策略优化机制或相关关键技术点，交付专利申请号及完整申请材料。

4)算法原型：完成面向 Agent 自进化能力的算法设计，交付一套支持反馈驱动优化与策略迭代的智能体原型框架，包含核心算法模块、运行流程及基础工程实现，并提供相关源代码及必要的模型参数（checkpoint）。

5)技术指标：研究基于反馈驱动与策略迭代的自进化方法，提升智能体在复杂任务中的执行稳定性与结果一致性，形成可复用的能力优化方案。

a.在公开基准任务或标准评测数据集上，相较于基线方法，任务成功率提升≥10%，或关键性能指标（如准确率/一致性）提升≥8%

b.在多轮交互或长链任务中，智能体决策一致性提升≥10%，异常/失败率降低≥10%

在典型应用场景验证中，相较现有方法，平均任务完成轮次

缩短≥20%，或 API 调用/Token 消耗降低≥20%

c.构建一套自进化能力评估方法，实现对模型性能变化的持续监测与量化分析，评估指标覆盖成功率、稳定性及效率等维度

d.所提出方法具备跨场景迁移能力，在不少于 3 类典型任务场景中完成验证

14.多智能体协作的关键技术研究

课题背景：

当前智能体应用从单 Agent 到多 Agent，从静态编排到动态协作转变，多智能体协作是当前 Agentic AI 的热门方向，课题聚焦解决“多智能体动态任务拆解+智能路由”。当多个 Agent 协同处理复杂任务时，系统面临"动态规划+资源竞争+错误放大"等多重挑战，传统静态编排无法应对，其中关键的挑战包括但不局限于：

在复杂任务蜂群模式下，如何进行任务规划和拆解，合理调度 GPU 资源，实现高效协作。

通过多智能体协作，抽象理解不同任务适合的智能体以及模型组合，通过智能化路由实现动态匹配，将任务分配和智能体的能力模型建模为多目标优化问题，并降低整体任务的 Token 成本。

上下文传递优化以及幻觉系统性抑制机制，防止错误级连风险。规划优化减少不必要的 Agent 调用从而降低幻觉风险；智能路由降低成本的同时也减少了上下文传递的层数。整体是一个需要全局优化的系统工程问题。

目标产出：

1)报告：提交项目结题表，提交结项验收报告。

2)论文：投稿甲方认可的期刊/AI 顶级会议论文 1 篇并最终收录，提供论文底稿及收录邮件。

3)专利：协助甲方和/或其关联公司提交国内或国际专利 1 项，交付符合本项目创新点的专利申请号和专利申请相关材料。

4)动态任务分解和路由决策引擎设计和 MVP 验证：基于任务复杂度自动分解并分配子任务，模型调用动态优化路由。

5)评估基准：多 Agent 协作性能、质量、成本的综合评估工具，其中模型动态智能路由(Agentic Router)，目标 Token 平均成本降低 40%以上。

（三）Token 效率优化/世界模型

1.面向大模型预训练高效 Token 利用的架构与优化方法研究

课题背景：

大模型预训练通常依赖大规模语料和高强度算力投入，Token 利用效率直接影响模型能力提升速度、训练成本和后续迭代效率。随着模型规模持续扩大，单纯增加训练 Token 数量的边际收益逐渐降低，如何在相同 Token 预算下获得更优模型效果，成为基础模型训练中的关键问题。

本项目聚焦大模型预训练阶段的 Token 效率优化，拟从模型架构、训练策略和优化器等角度开展研究，探索更高效的表示学习机制和参数更新方法，使模型在相同训练 Token 规模下获得接近或超过更长训练周期的效果。例如，在固定训练 Token 预算下，提升模型的收敛速度、泛化能力和下游任务表现。

目标产出：

1)报告：提交项目结题表，提交结项验收报告。

2)技术报告：形成面向学术交流和产业实践的研究报告一篇，总结高 Token 效率预训练方法的设计原则、实验结论和适用边界。

3)高 Token 效率预训练方法：提出并验证一种或多种面向大模型预训练的高效架构设计、训练策略或优化器改进方法，在相同 Token 预算下提升模型收敛效率和综合能力表现。

4)系统化实验评估结果：构建可复用的 Token 效率评估方案，

对比不同方法在训练损失、验证损失、通用能力评测和收敛速度等维度上的表现，量化其相对于基线方法的收益。达到相同验证损失（Validation Loss）所需 Token 数量较基线减少≥20%，比如新方法训练 800B 可以达到基线训练 1T 的效果。

2.基于 RDMA 多路复用的存储流量加速方案研究

课题背景：

大规模 AI 推理集群典型配置中，单台服务器配备双网卡架构，分别为普通 NIC（存储面）和 RNIC（计算面，支持 RDMA 协议）；其中 RNIC（RDMA 网络）具备更高带宽、更优传输效率，仅用于模型内部计算（如 TP/PP 并行通信、层间激活值传输），是集群高速网络核心载体；普通 NIC 专属承载存储流量（如镜像加载、权重传输），带宽有限。当前集群核心业务为大模型推理服务，需支持用户流量激增时的实例快速扩容，而存储网络传输效率成为制约服务弹性伸缩的关键瓶颈。当前架构在扩容场景面临网络资源利用失衡问题，非计算高峰期，RNIC 高速网络通路存在大量空闲带宽，而普通 NIC 存储面带宽有限；当需要并行扩容服务实例加载模型镜像时，存储面成为瓶颈，导致加载耗时过长，无法及时响应流量洪峰，而 RNIC 计算面资源未能被充分利用。

目标产出：

1)报告：提交项目结题表，提交结项验收报告。

2)论文：投稿甲方认可的期刊/会议论文 1 篇并最终收录，提供论文底稿及收录邮件。

3)专利：协助甲方和/或其关联公司提交国内或国际专利 1 项，交付符合本项目创新点的专利申请号和专利申请相关材料。

4)算法原型：实现一套基于 RDMA 多路复用的存储流量加速算法及原型源码，提升推理服务弹性，提高网络资源利用率，降低运维成本。

5)技术指标:基于 RDMA 多路复用的存储流量加速实现方案，相较现有方法，提升推理引擎端到端启动速度 80%，提升集群整体资源利用率。

3.基于世界模型的桌面机器人智能交互控制技术研究

课题背景：

行业痛点：当前主流桌面机器人（如教育、陪伴、轻量级抓取机器人）多依赖于预编程或传统的感知-动作控制范式。这种范式在面对开放、动态变化且存在不确定性的桌面环境时（如物品位置移动、新物体出现、人为干扰），表现出适应性差、决策僵化、交互不自然等问题。用户通常需要频繁进行规则调整或对环境进行人工干预，体验不佳。

技术瓶颈：基于规则的控制器无法处理未知场景；而单纯的

端到端深度学习模型又存在数据效率低、可解释性差、安全性难以保障的缺陷。现有机器人系统缺乏对物理世界基本规律的内在理解与因果推理能力——它们无法预判"水杯放在桌边可能会掉在地上摔碎"、"遮挡后目标仍然存在"、"动作与结果之间存在何种逻辑关联"等人类习以为常的常识性认知。这一根本性缺失导致机器人在面对超出预设范围的场景时，行为表现出明显的脆弱性与不连贯性：动作孤立、缺乏上下文关联，既无法进行前瞻性规划，也无法在执行失败后做出合理的自我修正，难以支撑真实场景下的可靠部署。

技术机遇：“世界模型”是人工智能领域新兴的前沿方向。它指智能体通过学习，在内部形成一个对所处环境的压缩、抽象的动态模型。这个模型能够预测环境在自身动作影响下将如何演变。将世界模型应用于机器人控制，可以让机器人具备“想象”和“推理”能力，从而实现更灵活、更稳定的自主交互与操作能力。

目标产出：

1)报告：提交项目结题表，提交结项验收报告。

2)论文：投稿甲方认可的期刊/会议论文 1 篇并最终收录，提供论文底稿及收录邮件。

3)专利：协助甲方和/或其关联公司提交国内或国际专利 1-2 项，交付符合本项目创新点的专利申请号和专利申请相关材料。

4)软件系统与原型：

5)一个可运行在目标桌面机器人平台上的基于世界模型的智能控制，感知外部的环境以及机器人相机 6DOF 的运动轨迹，能够精准、保真的生成相机运动视野下未来的动态环境，用于闭环强化设备上 AI 模型的感知与控制能力。配套的模型训练、数据管理和仿真测试工具链。

6)技术验证与演示：在真实桌面机器人上实现至少 1 个典型任务的技术验证与演示，如家庭场景内风险事件的动态生成、感知、预警。