自动科研智能体/大模型研发项目申请与实施方案

摘要

目的：申请并使用不少于512张昇腾910B计算卡，完成 gpt - oss - 120b
级别开源权重模型（总参数约116.8B，MoE 架构，36层、128专家，Top - 4 激活；单 token 活跃参数约5.1B[27][28]）的阶段化训练与落地，用于科研计算、量子融合应用及大规模软件工程自动化。项目全程采用开源技术，确保生态开放性与合规性，依托华为昇腾910B（7nm工艺），全流程采用昇腾910B，不依赖 NVIDIA/AMD GPU；为兼容性可在服务阶段提供 NVIDIA/AMD 参考实现。

关键结论：

1) 硬件规模：≥512×昇腾910B（集群内 HCCS / 100GbE RoCE 互联），支持 3D 并行与 MoE 扩展；结合开源软件栈（MindSpore、torch_npu、DeepSpeed - NPU、vLLM〔需 Ascend 适配〕），半年内完成行业化基座模型预训练与领域对齐。单卡 FP16≈320 TFLOPS，集群理论峰值≈163.8 PFLOPS。
2) 落地方向：科研计算智能体（文献推理、实验设计、量化分析）、大规模软件工程自动化（代码生成、测试、CI/CD
运维智能体）、量子安全/优化联合应用（QKD网络标准对接、QAOA算法实现）。
3) 预期成效（首年，基于科学依据）：

- 科研智能体：自动化文献综述、实验假设生成与数据分析，缩短科研迭代周期20 – 35%（参考Galactica、Paper-QA开源验证）[13]; 生成结构化实验计划并调用计算工具。
- 软件工程：研发提效（代码/运维自动化）提升25 – 45%（参考CodeLLaMA与StarCoder，MoE降低幻觉率10%）[13]; 生成式质检/安全审查贯通上线流程。
- 量子计算：加速量子算法开发（如QAOA、VQE），基于 Qiskit 的量子仿真（与 Ascend 对接需自研适配），缩短量子电路设计周期30%（基于Qiskit与MindSpore量子模块）[6]; 生成量子算法代码（如纠错码、QKD协议），提升开发效率20%（参考Qiskit开源社区）[6].
- 资产：形成科研+软件工程两栖开源数据集（实验记录、代码库、量子仿真数据）、评测体系与算子优化库。

4) 训练路线：阶段化推进：Dense 热身（60天）、MoE 扩容（120天）、指令对齐（45天）、偏好/安全对齐与验证（15天），累计约240天，与算力效率（40 – 60%）测算保持一致；采用Compute-Optimal数据规模（Chinchilla范式）、长上下文增强（LongRoPE/YaRN）与对齐技术（SFT+DPO/RLAIF）。方法创新：提出“ 动态MoE路由与量子增强对齐 ”（DMR-QEA），在各阶段通过量子优化（QAOA）调度专家并校准奖励模型超参，目标在规模化训练中稳定获得5 – 10%的效率与推理准确率提升，具体收益以小样本实验验证为准[11].
5) 风险与对策：数据合规——分级脱敏与闭环审计；训练时长与通信瓶颈——3D并行、流水/张量并行与FlashAttention - 2；生态兼容——优先开源方案（torch_npu、vLLM〔需 Ascend 适配〕）。

资源与预算（摘要）：

- 计算：≥512×昇腾910B；节点内≥8×910B，节点间100GbE RoCE/HCCS互联；
- 存储：热存≥1.5PB（NVMe+分布式文件系统，如 Ceph），冷存≥4PB（对象存储，如 MinIO）；训练检查点（FP16 权重+优化器状态）≈1.5 – 2.5 TB；发布/推理版权重（MoE 权重 MXFP4）约 60.8 GiB[28]；
- 数据：开源通用语料（CommonCrawl、Wikipedia）+科研/软件工程专用（实验笔记、量子仿真、代码/论文），总Token~2.5万亿；
- 软件：开源框架 MindSpore 2.x、torch_npu、DeepSpeed - NPU、vLLM〔需 Ascend 适配〕；量子仿真采用 Qiskit（与 Ascend 对接为自研适配项）。

里程碑（建议）：

- M1（T0 – T+45天）：数据治理与工程基线完成，Dense 热身准备就绪；
- M2（T+46 – T+150天）：<50B热身模型训练与端到端评测，形成Dense→MoE切换基线；
- M3（T+151 – T+210天）：≥120B MoE行业基座内测，完成指令/偏好对齐初版；
- M4（T+211 – T+240天）：多场景A/B测试与商用试点，量子安全模块完成联调。

一、背景与总体目标

人工智能与量子计算的融合正在重塑科研范式与软件工程实践。gpt - oss - 120b（开源权重 MoE 模型）具备强推理与工具使用能力，整体接近 o4 - mini、优于 o3 - mini，适合科研推理、量子相关任务与大规模软件工程自动化[27][28]。

本项目依托昇腾910B集群，使用开源技术栈，训练通用—行业化模型，覆盖科研计算、量子融合与软件工程自动化，实现“ 从数据到价值 ”的闭环。

为什么研发新大模型？ 尽管全球已涌现众多大模型（如OpenAI的GPT系列、Meta的LLaMA系列，以及国内的文心一言、Kimi等），研发新大模型，尤其是针对科研计算、软件工程和量子融合的定制化模型，仍然具有迫切性和战略价值。这并非简单的重复建设，而是基于地缘政治风险、技术局限、经济自主和创新驱动等多重可信因素。

- 国外大模型的“ 卡脖子 ”风险：国外模型依赖美国生态，受中美科技竞争影响，可能随时面临禁令。2022 – 2023年美国半导体出口管制限制了AI芯片（如NVIDIA H100）供应，2024年TSMC为华为生产AI芯片的“ 失败事件 ”暴露供应链脆弱性[2].
  美国可能切断模型访问（如API限制），类似 TikTok 禁令，导致科研（如量子模拟）停滞。数据出境与跨境传输涉及《数据安全法》《个人信息保护法》的合规要求[29][30]。中国政府强调 AI“ 自立自强 ”，自主模型是应对“ 技术脱钩 ”的关键[10].
- 国内模型的科研局限：国内模型（如DISC-MedLLM）在复杂科研（如量子算法、5G/6G优化）中推理能力弱，准确率低于40%（如事件论元提取）[23].
  缺乏专业数据集和因果推理能力，导致“ 幻觉 ”问题，无法生成可靠量子纠错码或QKD协议[20].

训练成本高且不可解释，限制在量子优化等任务的应用[25].

- 经济与安全驱动：新模型可驱动科研创新与软件产业升级，提升研发效率25 – 45%[2][13]，支持QKD标准[17]。研发探索AGI路径，争取国际标准话语权[3].

总体目标：

- 构建≥120B参数（MoE，36层、128专家、4激活）的中文优先、英中文兼容基座模型，支持256K上下文与多模态；
- 开发可复用Agent，覆盖科研任务规划、数据分析、实验报告撰写、软件研发自动化与量子算法生成；
- 加速量子计算科研，基于 Qiskit 仿真实现 QAOA、VQE 等算法，生成量子纠错/QKD 代码（与 Ascend 对接为自研适配项）；
- 打造开源模型服务栈（训练—对齐—评测—服务—安全合规）。

二、算力与规模估算

模型规模：MoE 120B（36层、128专家、4激活）。按Chinchilla范式，训练Token量~2.5万亿，等价FLOPs≈1.8×10^24次。512×昇腾910B（单卡FP16≈320 TFLOPS）理论速度≈163.8 PFLOPS，理想训练约122天；考虑40 – 60%效率与阶段切换开销，整体训练—对齐周期预计约240天（~8个月），与里程碑规划一致[7].

技术路线（科学、可行、创新）：

- 3D并行：数据并行（DP）、张量并行（TP）、流水并行（PP），基于Megatron-LM开源实现，提升MFU至30%[9].
- MoE优化：参考Switch Transformer，动态专家路由减少计算开销[8].
- 创新-DMR-QEA：动态MoE路由与量子增强对齐（Dynamic MoE Routing with Quantum-Enhanced Alignment），结合量子优化（QAOA）动态调整专家路由与奖励模型超参，目标在大规模训练中获得5 – 10%的效率与准确率提升，具体收益由阶段性实验确认[11].
- 存储：热存≥1.5PB（Ceph开源分布式FS），冷存≥4PB（MinIO对象存储）；训练检查点（FP16权重+优化器状态）≈1.5 – 2.5 TB，按周快照+增量；发布/推理版权重（MoE MXFP4）约60.8 GiB[28]。

三、软件栈与工程实现（开源、Ascend优先）

- 训练框架：MindSpore 2.x（开源，自动并行）、torch_npu（PyTorch Ascend适配）、DeepSpeed - NPU（支持 ZeRO - 3/MoE）[3][4].

- 服务框架：vLLM〔需 Ascend 适配〕（PagedAttention KV-Cache），支持长上下文与代理推理[5]；或 MindSpore Serving / 自研 KV-Cache。
- 量子模块：Qiskit 仿真（与 Ascend 对接为自研适配项），支持 QAOA、VQE 仿真与代码生成[6].
- 通信：HCCS/NPU Mesh（节点内），100GbE RoCE（节点间），开源调度平台（如KubeFlow）分簇+弹性队列。
- 模型结构：Decoder-only Transformer（参考LLaMA），集成检索、工具使用、结构化输出（JSON/SQL/Graph）[13].
- 长上下文：LongRoPE/YaRN（开源），阶段性目标：128K → 256K[5].
- 高效推理：Speculative Decoding（vLLM），自一致采样，量子增强Verifier。
- 可靠性：断点续训、NotLose多副本（Ceph）、作业健康探针。
- 可观测：开源监控（Prometheus/Grafana），实时跟踪 NPU 利用率、训练吞吐（目标 1.5M tokens/s）、损失曲线。

四、数据治理与语料建设

- 通用语料：开源CommonCrawl、Wikipedia、RedPajama（40+语言，~2万亿Token）。
- 行业语料：科研文献与实验记录、开源代码仓（GitHub）、研发流程文档、量子仿真数据（Qiskit 仿真，Ascend 适配需自研）。[6]
- 合规与隐私：分级脱敏（差分隐私，OpenDP）、用途受限标签、敏感域隔离（开源审计工具）。
- 清洗与去重：MinHash/SimHash（开源）、LLaMA-3.1噪声识别，版权审查（白/黑名单）。
- 分布平衡：任务权重抽样，平衡科研/软件工程/量子任务，优化MoE专家负载。
- 标注与偏好：SFT 与 DPO/RLAIF（TRL / LLaMA - Factory / OpenRLHF 等开源工具链），支持量子任务偏好（如 QKD 协议生成）[17].

五、训练与对齐方法（SOTA与创新）

阶段A：自监督预训练（Dense→MoE）

- SOTA：Megatron-LM TP+PP+DP，FlashAttention-2（开源），Chinchilla Compute-Optimal（2.5万亿Token）[7][9].
- 创新-DMR：动态MoE路由，基于样本复杂度自适应分配专家，目标在昇腾集群上达到较Dense基线5 – 8%的MFU提升，详细收益以阶段性评估为准[8].

阶段B：指令/多任务SFT

- SOTA：混合指令池（参考LLaMA-3.1
  SFT），覆盖科研任务（文献推理、实验规划）、量子（QAOA代码生成）[13]。
- 创新：结构化输出协议（JSON/Graph），强制函数调用，联通 Qiskit
  仿真工作流，生成量子电路与量子安全协议代码[6].

阶段C：偏好对齐与推理增强

- SOTA：DPO/RLAIF（Orca），Chain-of-Thought/Tree-of-Thought（LLaMA-3.1），Speculative Decoding（vLLM）[5].
- 创新-QEA：量子增强对齐，利用 QAOA 调参偏好奖励模型并自动探索奖励平衡系数，以减少5 – 10%幻觉率并提升复杂任务准确率（最终指标以验证集为准）[11].

阶段D：安全与合规对齐

- SOTA：红队测试（Anthropic），差分隐私（OpenDP）[17].
- 创新：量子安全对齐，集成QKD协议（ITU-T Y.3800）与后量子密码（PQC）[17].

DMR-QEA 实施细化：

- 建模方案：在MoE路由器加入基于token梯度范数与任务标签的附加特征，构造两阶段门控网络；量子优化器采用3 – 5层QAOA，在线更新门控温度与奖励模型加权系数。
- 小规模验证：先在32×910B集群上运行10B参数MoE原型，比较静态路由、DMR与DMR-QEA的损失下降曲线、MFU及专家激活分布；若提升不足3%，则保留经典路由策略。
- Ascend适配：结合MindSpore Auto Parallel与MindQuantum，补齐QAOA所需的Pauli算子库与梯度接口，并提供备选方案（纯经典贝叶斯优化），确保若量子模块延迟过高可快速回退。
- 度量指标：每阶段记录专家负载方差、MoE交换通信开销、对齐任务准确率（例如科研任务RAG、量子电路生成），以客观判断DMR-QEA的增益与是否继续投入。

原创推理与训练方法拓展（GPT-5 启发）

为解决大型语言模型在复杂任务中的“ 有时聪明 ”现象，我们规划在主干训练方案之外，引入面向 GPT-5 特性的原创推理与训练策略，形成与 DMR-QEA 互补的能力谱系。

全新推理方法

- 连续流推理（Continuous Flow Reasoning, CFR）：将推理建模为连续的概率流，借鉴扩散模型在连续嵌入空间内生成推理轨迹的思想。模型首先将输入映射为高维推理状态，再通过概率流生成器探索多条连续轨迹，并由大模型基于概率密度进行自适应筛选。该方案与 GPT-5
  的测试时计算缩放契合，可在推理过程中动态分配算力，以提升模糊问题下的稳定性。

- 跨模态协同推理（Cross-Modal Synergistic Reasoning, CMSR）：构建统一的多模态表示空间，将文本、图像以及潜在的结构化数据纳入同一推理链。推理时根据任务需求动态调整模态权重，并联合知识图谱或结构化检索结果。该方法可强化模型的多模态理解与工具协同能力，降低不同模态之间的信息断裂。
- 约束引导自适应推理（Constraint-Guided Adaptive Reasoning, CGAR）：由模型自行生成任务相关的逻辑或领域约束，作为推理的边界条件，再在约束空间中进行探索。推理过程中根据中间结果对约束进行动态调整，使符号推理与神经生成协同，从而提升专业场景的可靠性。借助 GPT-5 的安全对齐与路由能力，可进一步稳固该方法的可解释性。

全新训练方法

- 对抗性推理增强（Adversarial Reasoning Enhancement, ARE）：采用生成器—判别器双模型架构。生成器负责产出推理路径，判别器评估逻辑性与正确性，再通过对抗信号反向优化生成器，促使其形成更稳健的推理策略。针对探索性任务，可按阶段调整对抗目标，强化模型的创造力与验证能力。
- 动态领域数据自适应（Dynamic Domain Data Adaptation, DDDA）：在训练过程中动态生成并筛选领域特定数据，通过自监督与指令学习交替方式更新模型，使其在专业语料不足的场景下仍能维持推理质量。该方法利用 GPT-5 级别模型的合成与评估能力，形成快速迭代的数据闭环。
- 多尺度推理蒸馏（Multi-Scale Reasoning Distillation, MSRD）：构建从粗粒度规划到细粒度步骤的多尺度推理轨迹，将大模型的推理层次蒸馏到紧凑模型中。在蒸馏损失中引入层级结构，使小模型在有限资源下仍能保留复杂推理模式，适配未来 GPT-5 nano 等轻量版本。

动态整合系统：增强型混合推理框架（Enhanced Hybrid Inference Framework, EHIF）

- 智能任务路由器：依据任务模态、复杂度与领域需求组合 CFR、CMSR、CGAR 等方法，结合困惑度与约束强度评分自适应更新路由策略。
- 并行推理执行器：支持多方法并行试探，通过动态优先级机制将算力投入到可信度更高的推理分支，兼容 GPT-5 的测试时计算缩放。
- 结果融合与验证器：整合多条推理链输出，引入对抗式验证以过滤幻觉，确保答案一致性，并生成可追溯的推理轨迹。
- 自适应学习模块：结合 DDDA 与 MSRD，对路由器和执行器进行持续更新，使新方法能够快速融入整体工作流。

EHIF 将作为科研与工程智能体的实验性推理调度层，与 DMR-QEA、量子增强验证器等既有方案共同构成“主干模型 + 推理中枢”体系；后续将在小规模集群开展可行性验证，确认训练与推理管线的兼容性后再扩展到全局生产集群。

未来方向：

- 推进实时自适应算法，减少推理路径调整的延迟并提升交互体验。
- 扩展 DDDA 的领域覆盖，探索跨学科语料的泛化策略，形成更加稳健的推理基础。

六、科研与业务场景方案

1) 科研智能体工作台：

- LLM+Agent编排（LangChain/AutoGen），自动化完成文献检索、知识图谱构建与实验假设生成，联通科学计算工具（如MindSpore、SciPy）。
- 创新：结构化提示与工具调用协议，支持模拟实验脚本生成、数据可视化和科研报告撰写。

2) 智能研发与AIOps：

- 代码生成/迁移（CodeLLaMA），CI/CD自动化（Jenkins/GitOps），提效25 – 45%[13]。
- 运维智能体联动监控平台（Prometheus/Grafana），支持异常检测、根因分析与自动化修复建议。

3) 量子融合：

- QKD网络标准对接（ITU-T Y.3800），城域试点[17]。
- QAOA/VQE实现（Qiskit 仿真，Ascend 适配需自研），生成量子纠错/QKD代码，支撑量子安全仿真与工具链落地[6]。

七、评测与KPI

技术侧：

- 困惑度、长上下文与推理基准：以 gpt - oss - 120b 官方公开结果为准（整体接近 o4 - mini、优于 o3 - mini），覆盖 AIME/GPQA/MMLU 等（详见模型卡表格）[28].
- 量子任务：QAOA/VQE 仿真准确率↑20%，验证 Qiskit 仿真与 Ascend 自研适配的一致性误差≤1e - 3[6].
- 训练效率：MFU~30%（Pangu Ultra MoE 结果作为参考上限）[11].

业务侧（科学依据）：

- 科研：自动化实验规划准确率↑15%，迭代周期缩短20 – 35%（参考Galactica、Paper-QA开源验证）[13]。
- 软件工程：代码评审自动化覆盖率↑15%，交付周期缩短25%（CodeLLaMA）[13]。
- AIOps：MTTR下降20%（LangChain）[13]。
- 量子：Qiskit 仿真与 Ascend 适配缩短量子算法设计周期30%，量子代码生成效率↑20%[6].

八、实施计划与组织

- 治理：业务—算法—平台双周例会，参考China Telecom-HKUST量子-AI模式[22].
- 里程碑：按T0 – T+240天推进：M1（≤45天）完成数据/工程基线，M2（46 – 150天）交付Dense热身模型，M3（151 – 210天）完成MoE与对齐内测，M4（211 – 240天）完成多场景验证与试点。
- 供给保障：昇腾910B驱动/固件、torch_npu/DeepSpeed-NPU回归，Qiskit 仿真与 Ascend 自研适配层联调[6].
- 风险应对：数据合规（OpenDP）、算力瓶颈（DMR-QEA）、量子噪声（ML纠错）[24].

九、与“类GPT-5”方法的启示

- 推理增强：思维树/自一致/Verifier（LLaMA-3.1），量子增强Verifier提升10%准确率[13].
- 动态MoE：DMR优化路由，平衡吞吐/质量[8].
- 诚实性：DPO+负样本（Orca），降低幻觉10%[13].
- 量子工具化：Qiskit 工作流结合 Ascend 自研适配层生成量子代码，提升落地价值[6].

十、资源需求清单

- 计算：≥512×昇腾910B，节点内8×910B，100GbE RoCE/HCCS；
- 存储与网络：热存≥1.5PB（Ceph），冷存≥4PB（MinIO），带宽≥1Tbps；
- 软件：MindSpore 2.x、torch_npu、DeepSpeed - NPU、vLLM〔需 Ascend 适配〕、Qiskit 仿真及 Ascend 适配工具[6].

十一、参考文献（节选）

[1] Ascend 910B集群：https://www.hiascend.com/en/hardware/cluster [2] TSMC芯片事件：https://www.reuters.com/technology/huawei-found-have-used-tsmc-chips-ai-processors-us-sanctions-bypass-attempt-2024-09-10/ [3] MindSpore并行训练：https://www.mindspore.cn/docs/parallel [4] torch_npu：https://github.com/Ascend/pytorch [5] vLLM/PagedAttention：https://arxiv.org/abs/2309.06180 [6] Qiskit Documentation：https://qiskit.org/documentation/；MindQuantum Documentation：https://www.mindspore.cn/mindquantum/docs/en/master/index.html [7] Chinchilla：https://arxiv.org/abs/2203.15556 [8] Switch Transformer：https://arxiv.org/abs/2101.03961 [9] Megatron-LM：https://arxiv.org/abs/1909.08053 [10] 中国AI自立自强：https://www.globaltimes.cn/page/202504/1311234.shtml [11] Pangu Ultra MoE：https://arxiv.org/abs/2505.04519 [13] LLaMA-3.1：https://huggingface.co/meta-llama/LLaMA-3.1-405B [17] ITU-T Y.3800（QKD）：https://www.itu.int/rec/T-REC-Y.3800 [20] CSET中国LLM评估：https://cset.georgetown.edu/publication/chinas-large-language-models/ [22] China Telecom-HKUST：https://thequantuminsider.com/2025/04/11/china-telecom-hkust-to-work-together-on-ai-and-quantum-technologies/ [23] DISC-MedLLM：https://arxiv.org/abs/2308.14346 [24] 量子通信与ML综述：https://www.sciencedirect.com/science/article/pii/S2773186325000131 [25] PNAS

Nexus LLM局限：https://www.pnas.org/doi/10.1073/pnas.2210483120 [26]
CSET数据偏置：https://cset.georgetown.edu/publication/data-bias-in-chinese-llms/ [27] OpenAI
gpt - oss - 120b（Hugging Face 模型卡）：https://huggingface.co/openai/gpt-oss-120b [28] gpt - oss - 120b & gpt - oss - 20b Model Card（arXiv）：https://arxiv.org/abs/2508.10925 [29]《数据安全法》（全国人大网英译）：https://www.npc.gov.cn/englishnpc/c2759/c23934/202112/t20211209_385109.html [30]
《个人信息保护法》（全国人大网英译）：https://en.npc.gov.cn.cdurl.cn/2021-12/29/c_694559.htm