平台引擎

# 现状

1. 团队成员不了解项目组的整体目标和阶段目标。
2. 缺少开源系统使用、问题排查、运维管理等相关经验。
3. 亟需与领导和组员沟通、对齐整体平台引擎的**目标和行动路线**。

进展：

1. 单机Havenask环境就绪。
2. 阅读仓库LLM

# 成员分工

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **成员** | **技术分工** | **关联开源组件** | **员工/学生** | **定位** |
| 贺广福 | 实时RAG推理引擎和检索算子国产适配 | havenask | 员工 | 核心研发 |
| 徐天宇 | 推理引擎国产GPU适配和优化 | RTP-LLM | 员工 | 核心研发 |
| 姬晓光 | 图引擎国产GPU适配和优化 | havenask | 员工 | 核心研发 |
| 徐志斌 | 实时RAG推理引擎的流式处理 | havenask | 员工 | 生态建设 |
| 金龙 | Havenask 管控系统 | havenask | 员工 | 生态建设 |
| 张皓晶 | 平台引擎能力对接到agent平台 | havenask | 员工 | 生态建设 |
| 王云鹤 | 运维监控管理 | havenask | 员工 | 运维管理 |
| 马振洲 | RAG长链路推理方向 | havenask | 学生 | 算法研究 |
| 姜铄 | 智能体实时记忆系统 | havenask | 学生 | 算法研究 |

# 推进方案

## 阶段目标（3个月） -> 用起来

1. 各引擎独立**部署**、基础**功能可用**
2. 掌握**核心 API**
3. 了解和构建基础**监控与日志**体系
4. 掌握**核心功能使用**与**调试**能力
5. 与阿里团队建立协作机制
6. 支持上层业务应用，尝试少量业务场景

### 关键任务 – 开源组件上手【1-2周】

目标：让每个小组能独立部署所负责引擎，跑通一个最小示例

关键任务：

* 搭建开发/测试环境（**Docker**/K8s）
* 完成官方 Quick Start 示例（如：插入文档并搜索、导入图数据并查询、vLLM 启动模型并推理、LLM 多模态大模型启动推理）
* 输出《快速上手指南》（含部署命令、配置说明、常见报错）

研讨：

1. 现有工作与平台引擎工作的关系，哪些可以集成？哪些可以组合使用？哪些可以替换？
2. 上手过程中的痛点问题 -》 体验优化的参考 + 系统优化的目标

### 关键任务 - 场景验证 【8周】

1. 场景设计【1周】：
   1. 检索引擎：1000w新闻，关键词+向量+过滤
   2. 图引擎：千万级顶点和百万级边， 多度关系+过滤
   3. 推理引擎：高负载下推理效果；并行推理；MOE模型推理；多模态模型推理；
2. 功能、性能测试【5周】
3. 输出《场景验证报告》【1周】

**研讨**：现有业务适配情况和问题交流。【1周】

1. 相比其他开源/自研组件的优劣势。
2. 痛点问题。
3. 难点问题。

## 阶段目标（6个月）- 开源共建

[平台引擎 - 百度脑图](https://naotu.baidu.com/file/27714bc702f117ee30000f5ebb54f52e)

### 关键任务 – 数据引擎适配国产芯片【贺广福、姬晓光】

主要是适配【昇腾/海光】

1. **向量检索算子**国产GPU加速：设计基于国产GPU芯片的向量检索算子适配优化，大幅提升国产平台向量检索吞吐

2. **图采集算子**国产GPU加速：采用异步流水线预取策略，将海光DCU/昇腾NPU显存访问延迟大幅降低

### 关键任务 – 适应国产芯片的模型推理计算【徐天宇】

#### 动态编译技术

基于MLIR框架开发芯片专用IR（中间表示），将高层计算图（如PyTorch模型）自动转换为海光DCU/昇腾NPU原生指令集。

#### 通信层加速

1. 梯度稀疏化协议：针对昇腾Link带宽限制（仅200GB/s），开发 TopK梯度压缩算法（压缩率80%），减少跨卡通信量；

2. 拓扑感知集合通信：在强化学习模块中，根据物理拓扑动态优化AllReduce路径，提升训练加速比。

#### 模型层兼容

1. 混合精度训练：引入自动精度切换机制，在国产芯片平台实现 Loss Scaling梯度补偿，避免低精度训练发散，提升模型收敛速度

2. 异构GPU集群推理：在多种GPU构成的集群上实现大模型并行推理。【后续可扩展到训练？】

### 关键任务 - 增量图学习【姬晓光】

#### 动态子图划分

将超大规模知识图谱（如10亿边）按智能体访问热点切割为动态子图，通过LRU缓存机制实现实时高效更新

#### 图推理算子

适配动态子图的算子

#### 智能体实时记忆系统【姜烁】

基于图和索引引擎，构建智能体实时记忆系统。

### 关键任务 - 实时RAG推理引擎【贺广福】

#### 多模态索引融合

联合向量检索 + 知识图谱，构建统一检索API，支持自然语言查询到图谱节点的端到端映射；

#### （多模态）长链路推理【马振洲】

通过扩展推理链来增强模型的复杂问题解决能力。

模拟人类的思维,将复杂问题分解为多个中间步骤，并逐步推导出最终答案。

> 长思维链（Long CoT）与短思维链（Short CoT）代表了两种截然不同的推理范式。短思维链以浅层、线性的推理方式快速得出结论，逻辑路径短，探索性低，适用于结构清晰、解答明确的问题。而长思维链则强调深度推理、广泛探索和可行性反思，允许模型在更复杂的逻辑网络中展开深入分析，发现隐藏关系，并优化推理路径

#### 流式处理架构【徐志斌】

基于Apache Flink开发 事件驱动更新管道，使行为记忆写入延迟大幅降低

## 阶段目标（3个月） - 体验优化 -> 好用

### 关键任务 – 开发者友好的多语言SDK

### 关键任务 – 文档完善

### 关键任务 – 标杆Examples

### 关键任务 – 社区互动