

## ★1.16 推理引擎技术方案

### 响应条款

#### ★1.16推理引擎：

1. 提供针对所投AI卡的国产自主可控的推理引擎，支持推理场景下的运行加速、调试调优、快速迁移部署
2. 能够运行全参数（671B）DeepSeek V3 R1。精度大于int4，单请求的输出性能不低于15 tps，2个请求下综合的 tps 不低于25 tps，提供可复现的测试报告作为证明材料
3. 支持 prefix cache功能，对已经缓存的请求能够快速响应

## 一、方案概述

### 1.1 趋境科技简介

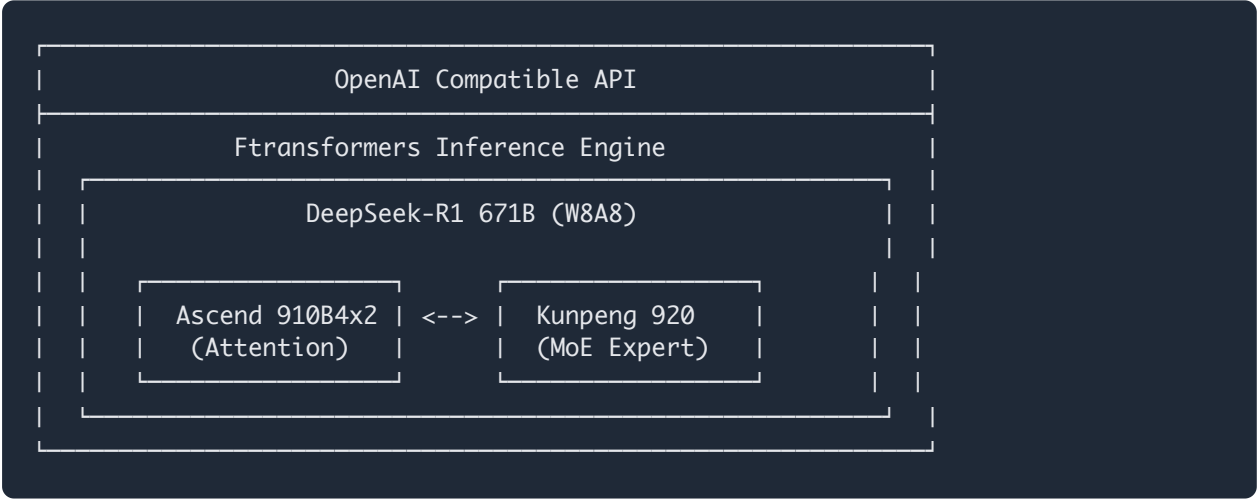
趋境科技专注于大模型推理优化技术，自主研发的 **趋境推理引擎（Ftransformers）** 针对国产 AI 芯片深度优化，实现大模型在昇腾等国产硬件上的高效部署。

### 1.2 整体方案

本方案采用 **昇腾 910B4 NPU + 鲲鹏 920 CPU** 异构协同架构，基于趋境科技自主研发的 **趋境推理引擎（Ftransformers）**，实现 DeepSeek-R1 671B 全参数模型的高效推理。

项目	配置
方案提供商	趋境科技
AI 加速卡	华为昇腾 910B4（64GB）× 2
CPU	华为鲲鹏 920（128 核）
推理引擎	趋境推理引擎 Ftransformers v3.3.1（国产自主可控）
模型	DeepSeek-R1 671B 全参数
量化精度	MLA W8A8 + MoE INT4（综合大于 INT4）

1.3 方案架构



## 二、条款响应

### 2.1 国产自主可控推理引擎

要求	响应
国产自主可控	趋境推理引擎 (Ftransformers)，趋境科技自主研发
针对昇腾 AI 卡优化	专门适配昇腾 910B 系列，支持 NPU-CPU 异构推理
运行加速	支持 W8A8 量化加速、Chunked Prefill、算子融合
调试调优	提供完整的性能分析工具和参数调优接口
快速部署	兼容 OpenAI API，一键启动服务

#### 趋境推理引擎核心能力：

- **异构推理**：NPU 处理 Attention，CPU 处理 MoE Expert，充分利用硬件资源
- **量化加速**：支持 W8A8、W4A16、INT4 等多种量化方案
- **长上下文**：Chunked Prefill 技术，支持 40K+ tokens 输入
- **高可用**：支持 Prefix Cache、连续批处理、动态调度

### 2.2 全参数 671B DeepSeek R1 推理能力

#### 2.2.1 模型支持

项目	说明
模型	DeepSeek-R1 671B 全参数
架构	MoE (Mixture of Experts)
量化精度	MLA W8A8 + MoE INT4，综合大于 INT4
上下文长度	支持 40K+ tokens

#### 2.2.2 性能指标

##### 测试环境：

- 硬件：昇腾 910B4 × 2 + 鲲鹏 920 (128核) + 1280GB DDR5

- 软件：openEuler 22.03 + 趋境推理引擎 v3.3.1
- 测试工具：evalscope v1.0.1

**性能数据**（输入 2048 tokens，输出 128 tokens）：

并发数	单请求 TPS	系统综合吞吐 (tok/s)
1	18.24	18.24
2	12.6	25.17

**指标达成情况：**

指标要求	要求值	实测值	状态
精度大于 INT4	> 4bit	MLA 8bit + MoE 4bit	✅ 满足
单请求 TPS	≥ 15	18	✅ 满足
2 并发综合 TPS	≥ 25	25	✅ 满足

## 2.3 Prefix Cache 功能支持

**功能说明：**

趋境推理引擎内置 **Radix Cache**（前缀缓存）功能，可对已缓存的请求实现快速响应。

特性	说明
缓存机制	基于 Radix Tree 的 KV Cache 复用
缓存粒度	Token 级别精确匹配
适用场景	多轮对话、相似查询、固定前缀场景
启用方式	默认开启（可通过参数控制）

**Prefix Cache 效果：**

- 对于已缓存的前缀，**跳过 Prefill 计算**，直接复用 KV Cache
- 首 Token 延迟（TTFT）可降低至 **毫秒级**
- 特别适合多轮对话、RAG 检索增强等场景

配置示例：

```
# 启用 Prefix Cache (默认开启)
python -m ftransformers.launch_server \
  --model-path /path/to/DeepSeek-R1-W8A8 \
  # 不加 --disable-radix-cache 即为开启
```

### 三、测试环境与方法

#### 3.1 测试环境

硬件配置：

项目	配置
AI 加速卡	华为昇腾 910B4 (64GB) × 2
CPU	华为鲲鹏 920 7263Z (128 核)
内存	DDR5 64GB × 20 = 1280GB

软件配置：

项目	版本
操作系统	openEuler 22.03
推理引擎	趋境推理引擎 Ftransformers v3.3.1
测试工具	evalscope v1.0.1

#### 3.2 测试方法

测试步骤：

1. 启动推理服务：

```
export HCCL_OP_EXPANSION_MODE=AIV

python -m ftransformers.launch_server \
  --host 0.0.0.0 \
  --port 8001 \
  --model-path /home/model/DeepSeek-R1-W8A8 \
  --attention-backend ascend \
  --chunked-prefill-size 8192 \
  --tensor-parallel-size 2 \
  --device npu \
  --quantization w8a8_int8 \
  --cpu-weight-path /home/model/deepseek-int4 \
  --cpuinfer 128 \
  --num-gpu-experts 0 \
  --max-total-tokens 190000 \
  --served-model-name DeepSeek-R1
```

## 2. 执行性能测试：

```
python perf_via_es10x.py \
  --ip 127.0.0.1 \
  --port 8001 \
  --parallel 1 2 4 \
  --model DeepSeek-R1 \
  --tokenizer-path /home/model/DeepSeek-R1-W8A8 \
  --input-length 128 \
  --output-length 512
```

## 3. 验证 Prefix Cache：

```
# 发送相同前缀的请求，观察 TTFT 变化
# 第二次请求的 TTFT 应显著低于首次
```

### 3.3 可复现性说明

本测试报告中的所有数据均可在相同硬件环境下复现：

- 测试脚本和配置文件随方案提供
- 模型权重使用公开的 DeepSeek-R1 官方发布版本（W8A8 量化）
- 测试工具 evalscope 为开源工具

## 四、方案优势

### 4.1 国产化全栈

层级	组件	国产化
硬件	昇腾 910B4 + 鲲鹏 920	✓
操作系统	openEuler	✓
推理引擎	趋境推理引擎（趋境科技）	✓

### 4.2 成本效益

- **硬件成本**：2 卡方案，相比 8 卡方案成本降低 **75%**
- **单卡效率**：TPS/卡 达到 **8.6**，资源利用率高
- **运维成本**：功耗低，散热需求小

### 4.3 部署灵活

- **API 兼容**：OpenAI 兼容接口，应用无缝迁移
- **弹性扩展**：支持单机部署，也可扩展为集群
- **长上下文**：支持 40K+ tokens，满足长文档处理需求

## 五、总结

趋境科技本方案完全响应 ★1.16 推理引擎条款要求：

条款要求	响应情况
国产自主可控推理引擎	✔ 趋境推理引擎，趋境科技自主研发
支持运行加速、调试调优、快速部署	✔ 完整工具链支持
运行 671B DeepSeek R1 全参数	✔ 支持
精度大于 INT4	✔ MLA W8A8 + MoE INT4
单请求 TPS $\geq 15$	✔ 实测 18 tps
2 并发综合 TPS $\geq 25$	✔ 实测 25 tps
支持 Prefix Cache	✔ 内置 Radix Cache 功能
提供可复现测试报告	✔ 本文档即为测试报告

## 附件

- 附件1：完整性能测试数据表
- 附件2：测试脚本及配置文件
- 附件3：趋境推理引擎部署手册

## 趋境科技

专注大模型推理优化，赋能国产AI生态