ISA-HCA-LLM 真实 Hidden States 测试报告

> **模型**: Mistral-7B-Instruct-v0.2
> **测试日期**: 2026-01-19

一、测试配置

配置	值
模型	Mistral-7B-Instruct-v0.2
量化	Q4_K_M (4-bit)
Embedding 维度	4096
Token 数	384
文档数	4
HCA 集成	✅ 完整

二、相似度分布

类型	平均相似度
同文档 Token 对	**0.1908**
跨文档 Token 对	**0.0885**
差异	**0.1023** ✅

三、阈值测试结果

百分位	ρ_th	尺度数	操作节省	加速比	文档对齐	HCA 模拟器
P85	0.2287	1	50.0%	0.00x ⚠️	24.9%	1
P90	0.2741	4	50.0%	0.00x ⚠️	24.9%	1
P92	0.2997	7	50.0%	0.00x ⚠️	24.9%	1
P95	0.3521	27	50.0%	0.00x ⚠️	24.8%	5
P97	0.4076	52	50.0%	0.00x ⚠️	24.8%	9

四、最佳配置

指标	值
阈值百分位	P97
ρ_th	0.4076
操作节省	**50.0%**
加速比	**0.00x** ⚠️
尺度数	52
HCA 模拟器数	9
文档边界对齐	24.8%

五、可视化

操作节省率:

```
P85: ████████████████████░░░░░░░░░░ 50.0%
P90: ████████████████████░░░░░░░░░░ 50.0%
P92: ████████████████████░░░░░░░░░░ 50.0%
P95: ████████████████████░░░░░░░░░░ 50.0%
P97: ████████████████████░░░░░░░░░░ 50.0%
```

加速比:

```
P85: ▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒ 0.00x ⚠️
P90: ▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒ 0.00x ⚠️
P92: ▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒ 0.00x ⚠️
P95: ▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒ 0.00x ⚠️
P97: ▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒ 0.00x ⚠️
```

HCA 模拟器使用:

```
P85: ████▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒ 1 个
P90: ████▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒ 1 个
P92: ████▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒ 1 个
P95: ████████████████▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒ 5 个
P97: ████████████████████████████████ 9 个
```

六、与其他测试对比

测试类型	维度	HCA 集成	操作节省	加速比
模拟 LLM (完美聚类)	4096	❌	87.5%	2.17x
真实 LLM (无 HCA)	4096	❌	34.4%	1.02x
真实 LLM (有 HCA)	**4096**	✅	**50.0%**	**0.00x**
真实 Embedding (MiniLM+HCA)	384	✅	~50%	~2x

七、关键发现

7.1 HCA 集成效果

- HCA 模拟器处理 52 个尺度，但开销较大
- 在 4096 维空间中，HCA 模拟开销明显
- 建议优化: GPU 加速 HCA 模拟或减少模拟步数

7.2 HCA 模拟器开销

- 使用了 **9** 个 HCA 模拟器
- 处理了 **4** 个多元素尺度
- 模拟器/尺度比: **17.3%**

7.3 ISA→HCA 映射验证

ISA 层	HCA 组件	验证结果
交互探测 (ρ)	HCASimulator 网格	✅
尺度划分	TopologyGraph	✅
场势注意力	GradientField + Flow	✅
组间渗透	BoundaryField	✅

八、复现步骤

```
# 1. 进入目录
cd ~/Desktop/hca-sim/isa_hca_llm_benchmark

# 2. 生成真实 Hidden States（需要 LLM 模型）
python3 generate_hidden_states.py

# 3. 运行 HCA 测试
cargo run --release
```

> **版本**: 1.0
> **测试环境**: macOS, Apple Silicon, Metal GPU