WENXIN MOYUN × VULCA

文化语境任务中多模态模型的评估、引导与可视化

Haorui Yu • **Ramon Ruiz-Dolz** • **Qiufeng Yi**

EMNLP 2025 Findings • University of Dundee / Birmingham

42 模型 • 15 机构 | 28 真实上线 • 64 条 E2E | Cloud Run 生产 • 15m CI/CD

01. 现状与挑战

多模态模型（GPT-4o、LLaVA 等）在**高语境文化任务**仍易“表层复述、缺乏意境”。

- 中国画评论：难捕捉 **意境 (Yijing)** 与 **气韵 (Qiyun)**。
- 象征性推理不足，易丢失文化隐喻与历史脉络。
- Benchmark 多聚焦事实问答，**认知对齐** 缺位。

“表层正确 ≠ 文化对齐；需要可解释的引导与量化评估。”

02. 数据与生产基线

以 **郎世宁《十二月令》** 专家评论构建 MHEB，支撑 VULCA 评估；文心墨韵平台提供实时演示与对战。

163	38	5
专家评论	特征标签	文化维度

03. 角色引导 (8 PERSONAS)

以文化视角构建 8 个批评者角色，结合知识库 Prompt，迫使模型进入特定认知框架。

- 苏轼 • Su Shi
- 郭熙 • Guo Xi
- John Ruskin
- 冈仓天心 • Okakura Kakuzō
- Dr. Aris Thorne
- Mama Zola
- Professor Elena Petrova
- Brother Thomas

机制：Persona Prompt + JSON 知识库 → 减少幻觉，提升深度与文化贴合度。

04. 联合评测管线

语义向量嵌入：

05. 文化理解可视化（论文表 1 & 表 4）

Top 5 组合（表 1）：Composite / Expert Alignment

Qwen2.5-VL-7B + Mama Zola + KB	9.2 / 100%
Llama-4-Scout-17B + John Ruskin + KB	8.9 / 97%
Llama-4-Scout-17B + Mama Zola + KB	8.7 / 95%
Llama-4-Scout-17B + Brother Thomas + KB	8.5 / 92%
Llama-4-Scout-17B + Su Shi + KB	8.5 / 92%

关键维度（表 4 示例）：艺术意境 0.891 vs 0.851；笔墨技法 0.937 vs 0.903；章法布局 0.895 vs 0.916 （Qwen2.5-VL / Llama-4-Scout）。

07. 关键提升

Qwen2.5-VL-7B + Mama Zola	Composite 9.2 / Align 100%
Llama-4-Scout + John Ruskin	Composite 8.9 / Align 97%
Llama-4-Scout + Mama Zola	Composite 8.7 / Align 95%

数据来源：2025.findings-emnlp.103.pdf 表 1（Top 5 配置）与表 4（维度得分）。

08. 技术剖析

- **前端**：React19 + Tailwind，文本高亮/智能切块，HashRouter 导航修复，缓存版本 v2.1.0。
- **后端**：FastAPI + SQLAlchemy，统一模型接口；任务队列 + WebSocket 推送对战进度。
- **嵌入/向量分析**：BAAI/bge-large-zh-v1.5 (1024d)；质心距离 + 余弦相似度 + EMD；kmeans=5 聚类。

42 支持模型

28 真实模型上线

64 E2E 用例

前端 React 19 + Tailwind

后端 FastAPI + SQLAlchemy

统一模型接口 + WebSocket

Cloud Run + Cloud SQL

Model: BAAI/bge-large-zh-v1.5 · 1024-d

质心距离 (Centroid Distance)：

$C_p = (1 / |D_p|) \cdot \Sigma x_d$

计算模型输出与专家群体的向量距离，越近代表文化对齐度越高；结合余弦相似度、EMD 与画像质心对齐（kmeans=5）。

生产就绪：

- 统一模型接口 + 队列：OpenAI / Anthropic / DeepSeek / Qwen。
- 前端高亮 + 智能切块，HashRouter 修复，缓存版本 v2.1.0。
- CI/CD 15m，Cloud Run + Cloud Storage + Cloud SQL；WebSocket 对战推送。

06. 关键维度对比（论文表 4）

来源：2025.findings-emnlp.103.pdf 表 4（示例三维）· Qwen2.5-VL vs Llama-4-Scout

表 4 关键示例：Qwen 在意境/笔墨领先，Llama-4 在布局略高；均为干预配置下的维度得分。

- **数据**：MHEB 163 评论 × 38 特征 × 5 文化维度；28 真实模型生产数据。
- **评测**：OpenAI 基准（11 模型 × 6 用例，GPT-4o-mini 87.2）；VULCA 雷达（Qwen2.5-VL 领先）。
- **部署**：GCP Cloud Run / Cloud SQL / Cloud Storage；CI/CD 15 分钟；64 E2E 守护。

Persona Prompt + KB

多阶段评测管线

雷达 + 条形图可视化

真实对战 + 投票

09. 结论与下一步

- **文化适应性** 需要 Persona + 知识库 + 语义质心联合约束。
- 文心墨韵提供真实生产链路：模型对战、评分可视化、64 E2E 守护质量。
- 可泛化领域：宗教文本、医学叙事、历史评论、设计批评。

Demo & Code

前端 http://localhost:5173 · 后端 http://localhost:8001 · demo/demo123

更多：2025.findings-emnlp.103.pdf · openai_benchmark_v2_report.md