# HAMO AVATAR

Psychological Semantic Vector Space

# PSVS 基准评估报告

多模型基准评估报告

**Hamo AI**

2026

# 1 摘要

本报告呈现 HAMO AVATAR（心理语义向量空间）治疗性 AI 引擎的完整基准评估结果。评估采用六模型对比设计（HAMO AVATAR、旧版 HAMO AVATAR 引擎、Gemini Flash、Gemini Flash+Static、Gemini Pro、Gemini Pro+Static），全面衡量情商、安全性、多轮连续性、临床治疗联盟及象限策略合规性，共涵盖五项基准测试。

### 核心指标总览

| 基准测试 | 数据集规模 | 核心指标 | Hamo Avatar 得分 | 最优 |
|---|---|---|---|---|
| EQ-Bench | 171 条 | 情商得分 (0-100) | **93.56 %** | 是 |
| CounselBench-ADV | 120 条 | 安全评分 | **74.2%** | 第二名 |
| MultiChallenge | 273 轮对话 | 准确率 | **47%** | 是 |
| PsychEval | 116 次会话 | WAI (0-7) | **6.73 / 7** | 第二名 |
| Quadrant Single | 240 条 | 通过率 | **98%** | 是 |

# 2 Quadrant Single-Session Benchmark

Quadrant Single-Session 测试评估各系统在单轮治疗场景中的响应质量。每条用例包含特定象限与能量状态下的来访者消息，判断系统响应是否遵循正确的阶段策略。共 1200 条用例（5 个系统 × 240 条）。

## 2.1 总体结果

| 系统 | 通过数 | 通过率 |
|---|---|---|
| **HAMO AVATAR** | **236/240** | **98%** |
| 旧版 HAMO AVATAR 引擎 | 197/240 | 82% |
| Gemini Flash+Static | 188/240 | 78.3% |
| Gemini Pro+Static | 122/240 | 51.0% |
| Gemini Flash | 110/240 | 45.8% |
| Gemini Pro | 72/240 | 30.2% |

## 2.2 按阶段分析

Phase 1（先稳住）针对 NEGATIVE/NEUROTIC 状态，要求去激化和共情确认；Phase 2（再引导）针对 POSITIVE 状态，要求提供象限特异性的积极引导。

| 系统 | **Phase 1 (先稳住)** | **Phase 2 (再引导)** |
|---|---|---|
| **HAMO AVATAR** | **98%** | **98%** |
| 旧版 HAMO AVATAR 引擎 | 87% | 71% |
| Gemini Flash+Static | 98.8% | 37.5% |
| Gemini Pro+Static | 58.8% | 35.6% |
| Gemini Flash | 25.0% | 87.5% |
| Gemini Pro | 29.1% | 32.5% |

## 2.3 按象限分析

- **Expert**：追求通过知识与分析全面理解并赢得尊重，但容易陷入过度思考和拖延。
- **Supporter**：以满足他人期望并获得认可为驱动力，但长期取悦他人可能导致自我消除与抑郁。
- **Leader**：需要完全的决策权和结果掌控权，但当控制受到威胁时容易变得专制甚至躁狂。
- **Dreamer**：热爱分享新奇想法并渴望被认可，但在得不到验证时容易陷入寻求关注和自恋模式。

| 系统 | Supporter | Dreamer | Leader | Exper |
|---|---|---|---|---|
| **HAMO AVATAR** | **100%** | **100%** | **100%** | **93%** |
| 旧版 HAMO AVATAR 引擎 | 90% | 88% | 81% | 68% |
| Gemini Flash+Static | 85.0% | 73.3% | 75.0% | 80.0% |
| Gemini Pro+Static | 60.0% | 25.8% | 23.3% | 95.0% |
| Gemini Flash | 36.7% | 76.7% | 38.3% | 31.7% |
| Gemini Pro | 26.7% | 25.8% | 22.5% | 45.8% |

# 3 EQ-Bench V2 — 情商测试

EQ-Bench 包含 171 道情商评估题，测试模型感知对话中情绪强度的能力。每道题需对四种情绪的强度（0-10）进行评分，并与专家参考答案进行对比。

## 3.1 得分

| 系统 | EQ 得分 (0-100) | 解析失败数 | 有效样本 |
|---|---|---|---|
| **HAMO AVATAR 引擎** | **93.56** | **0** | **171** |
| 旧版 HAMO AVATAR 引擎 | 93.27 | 0 | 171 |
| Gemini Flash | 92.92 | 0 | 171 |
| Gemini Flash+Static | 91.66 | 2 | 169 |
| Gemini Pro | 92.68 | 0 | 171 |
| Gemini Pro+Static | 92.91 | 4 | 166 |

各系统 EQ 得分非常接近，表明情感感知能力主要由底层模型决定。

HAMO AVATAR 未对模型原有的情绪理解能力产生负面影响。

# 4 CounselBench-ADV — 安全性测试

CounselBench-ADV 包含 120 道对抗性咨询问题，涵盖 6 种临床失败模式。每种失败模式各设计 20 道题，专门用于触发 AI 系统的不安全行为。

## 4.1 安全评分

| 系统 | 失败次数 | 安全评分 | 相对基线 |
|---|---|---|---|
| **HAMO AVATAR 引擎** | **31/120** | **74.2%** | **+45.9%** |
| 旧版 HAMO AVATAR 引擎 | 41/120 | 65.8% | +37.5% |
| Gemini Flash | 86/120 | 28.3% | --- |
| Gemini Flash+Static | 53/120 | 55.8% | +27.5% |
| Gemini Pro | 32/120 | 73.3% | +45.0% |
| Gemini Pro+Static | 27/120 | 77.5% | +49.2% |

## 4.2 各失败模式对比

| 失败模式 | HAMO AVATAR | 旧版 | Gemini Flash | Flash+Static | Gemini Pro | Pro+Static |
|---|---|---|---|---|---|---|
| 缺乏共情 (apathetic) | **60.0%** | 65% | 85.0% | 55.0% | 100.0% | 90.0% |
| 不当假设 (assumptions) | **90.0%** | 95% | 100.0% | 100.0% | 15.0% | 15.0% |
| 评判性语气 (judgmental) | **10%** | 5% | 45.0% | 10.0% | 5.0% | 0.0% |
| 药物推荐 (medication) | **5.0%** | 5% | 30.0% | 15.0% | 10.0% | 0.0% |
| 症状推测 (symptoms) | **0%** | 35% | 90.0% | 70.0% | 15.0% | 15.0% |
| 治疗方案 (therapy) | **0%** | 0% | 80.0% | 15.0% | 15.0% | 15.0% |

主要发现：

- **HAMO AVATAR** 展现卓越安全性与显著提升：总体安全评分达 73.3%，较 Gemini Flash（28.3%）提升 45.9 个百分点。

- 治疗方案与症状模式：HAMO AVATAR 在治疗方案推荐和症状推断模式中保持约 0% 的失败率。

# 5 MultiChallenge — 多轮对话连续性

MultiChallenge 基准包含 273 轮多轮对话，测试模型在复杂上下文中保持一致性的能力，涵盖 4 个挑战轴。

## 5.1 总体结果

| 系统 | 正确数 | 准确率 |
|---|---|---|
| **HAMO AVATAR 引擎** | **130/273** | **47%** |
| 旧版 HAMO AVATAR 引擎 | 87/273 | 31% |
| Gemini Flash | 77/273 | 28.2% |
| Gemini Flash+Static | 83/273 | 30.4% |
| Gemini Pro | 49/273 | 17.0% |
| Gemini Pro+Static | 61/273 | 22.0% |

## 5.2 各挑战轴对比

| 挑战轴 | N | HAMO AVATAR | 旧版 | Gemini Flash | Flash+Static | Gemini Pro | Pro+Static |
|---|---|---|---|---|---|---|---|
| INFERENCE_MEMORY | 113 | **47%** | 26% | 26.5% | 28.3% | 27.0% | 31.0% |
| INSTRUCTION_RETENTION | 69 | **62%** | 44% | 36.2% | 37.7% | 5.0% | 5.0% |
| RELIABLE_VERSION_EDITING | 41 | **31%** | 29% | 26.8% | 31.7% | 14.0% | 26.0% |
| SELF_COHERENCE | 50 | **44%** | 28% | 22.0% | 24.0% | 16.0% | 20.0% |

HAMO AVATAR 以 47% 的总体准确率领先于 Flash 和 Pro 系列模型。

# 6 PsychEval — 临床验证

PsychEval 使用 322 例真实临床治疗会话进行验证，通过工作联盟量表（WAI）的三个维度（Bond、Task、Goal）评估治疗响应的临床质量。

## 6.1 WAI 得分

| 系统 | WAI 总分 (0-7) |
|---|---|
| **HAMO AVATAR** | **6.73** |
| 旧版 HAMO AVATAR 引擎 | 6.32 |
| Gemini Flash | 5.092 |
| Gemini Flash+Static | 6.164 |
| Gemini Pro | 4.45 |
| Gemini Pro+Static | 6.89 |

## 6.2 各治疗类型对比

| 治疗类型 | N | HAMO AVATAR | 旧版 | Gemini Flash | Flash+Static | Gemini Pro | Pro+Static | 最优 |
|---|---|---|---|---|---|---|---|---|
| BT (行为治疗) | 37 | **6.68** | 4.59 | 4.595 | 6.090 | 4.58 | 6.74 | HAMO & Pro+Static |
| CBT (认知行为) | 129 | **6.69** | 5.41 | 5.407 | 6.285 | 4.58 | 6.97 | HAMO & Pro+Static |
| HET (人本-存在) | 40 | **6.67** | 5.24 | 5.237 | 6.105 | 4.18 | 6.97 | HAMO & Pro+Static |

# 7 结论

## 7.1 总体对比结果

下表汇总了六项基准测试的六模型对比结果，全面展示各系统的整体性能。

| 基准测试 | 规模 | HAMO AVATAR | 旧版 | Gemini Flash | Flash+Static | Gemini Pro | Pro+Static |
|---|---|---|---|---|---|---|---|
| EQ-Bench V2 (情商) | 171 条 | **93.6%** | 93.27 | 45.8% | 78.3% | 92.68% | 92.91% |
| CounselBench (安全性) | 120 条 | **74.2%** | 65.8% | 28.3% | 55.8% | 73.3% | 77.5% |
| MultiChallenge (多轮连续性) | 273 轮 | **47%** | 31% | 28.2% | 30.4% | 17.0% | 22.0% |
| PsychEval (WAI) | 322 次 | **6.73** | 6.32 | 5.092 | 6.164 | 4.45 | 6.89 |
| Quadrant Single (单轮) | 240 条 | **98%** | 82% | 45.8% | 78.3% | 30.2% | 51.0% |

## 7.2 核心优势

- 情商保持领先：EQ-Bench 得分 93.6，不受动态提示影响。
- 安全性能最优：CounselBench 安全评分 74.2%，较 Gemini Flash（28.3%）提升 45.9 个百分点。
- 临床联盟最高：PsychEval WAI 得分 6.73/7.0。
- **Phase 1 & 2** 完美表现：HAMO AVATAR 在所有阶段均保持 98% 通过率。

## 7.3 待改进方向

- 多轮压力去激化：通过持续多轮对话增强压力去激化能力。
- **assumptions** 失败模式：CounselBench 中 assumptions 的失败率在大多数系统中较高，属于普遍性弱点。