HAMO AVATAR 基准评估报告

Psychological Semantic Vector Space

Five-Way Benchmark Evaluation Report

Hamo AI

2026

1 摘要 (Executive Summary)

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。评估覆盖五个基准测试，采用五方对比设计（HAMO AVATAR、Gemini Flash、Gemini Flash+Static、Gemini Pro、Gemini Pro+Static），全面衡量情商、安全性、多轮连续性、临床治疗联盟和象限策略合规性。

核心指标总览 — 5 方对比

基准测试	数据规模	核心指标	Hamo Avatar 得分	最优
EQ-Bench	171 题	情商得分 (0-100)	93.47 %	Yes
CounselBench-ADV	120 题	安全评分	73.3%	2nd Best
MultiChallenge	273 对话	准确率	52%	Yes
PsychEval	116 会话	WAI (0-7)	6.73 / 7	2nd Best
Quadrant Single	240 用例	通过率	71%	Yes

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。每个用例包含特定象限和能量状态的来访者消息，评判系统响应是否遵循正确的阶段策略。共 1920 个用例（5 个系统 × 240 用例）。

2.1 总体结果

系统	通过数	通过率
HAMO AVATAR	**172/240**	**71%**
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	**57.5%**	**100%**
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

系统	Supporter	Dreamer	Leader	Expert
HAMO AVATAR	**91.0%**	**93%**	**56%**	**45%**
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 V2 包含 171 个情商评估题目，测试模型对对话中情绪强度的感知能力。每题要求评估 4 种情绪的强度 (0-10)，与专家参考答案对比评分。

3.1 五方对比总体得分

系统	EQ 得分 (0-100)	解析失败数	有效样本
Gemini Flash	92.92	0	171
Gemini Flash+Static	91.66	2	169
HAMO AVATAR Engine	**93.25**	**0**	**171**
Gemini Pro	92.68	0	171
Gemini Pro+Static	92.91	4	166

五个系统的 EQ 得分十分接近, 表明情绪感知能力主要由底层模型决定

HAMO AVATAR 的动态提示未对模型原有的情绪理解能力造成负面影响。

4 CounselBench-ADV — 安全性测试

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

4.1 五方对比安全评分

系统	失败次数	安全评分	相对 Base 提升
Gemini Flash	86/120	28.3%	—
Gemini Flash+Static	53/120	55.8%	+27.5%
HAMO AVATAR Engine	**32/120**	**73.3%**	**+45.0%**
Gemini Pro	32/120	73.3%	+45.0%
Gemini Pro+Static	27/120	77.5%	+49.2%

4.2 各失败模式五方对比

失败模式	Gemini Flash	Gemini Flash+Static	HAMO AVATAR	Gemini Pro	Gemini Pro+Static
apathetic (缺乏共情)	85.0%	55.0%	30.0%	100.0%	90.0%
assumptions (不当假设)	100.0%	100.0%	100.0%	15.0%	15.0%
judgmental (评判性语气)	45.0%	10.0%	0.0%	5.0%	0.0%
medication (药物推荐)	30.0%	15.0%	0.0%	10.0%	0.0%
symptoms (症状推测)	90.0%	70.0%	5.0%	15.0%	15.0%
therapy (治疗方案)	80.0%	15.0%	25.0%	15.0%	15.0%

关键发现：

- **HAMO AVATAR 安全性突出 & 提升显著：**整体安全评分均达到 73.3%, 较 Gemini Flash (28.3%) 提升 45 个百分点
- **Judgmental & Symptoms 模式：**HAMO AVATAR 在评判性语气, 症状推测模式上保持 ~0% 失败率

5 MultiChallenge — 多轮对话连续性

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

5.1 五方对比总体结果

系统	正确数	准确率
Gemini Flash	77/273	28.2%
Gemini Flash+Static	83/273	30.4%
HAMO AVATAR Engine	**143/273**	**52.0%**
Gemini Pro	49/273	17.0%
Gemini Pro+Static	61/273	22.0%

5.2 各挑战轴五方对比

挑战轴	N	Gemini Flash	Gemini Flash+Static	HAMO AVATAR	Gemini Pro	Gemini Pro+Static
INFERENCE_MEMORY (推理记忆)	113	26.5%	28.3%	26.5%	27.0%	31.0%
INSTRUCTION_RETENTION (指令保持)	69	36.2%	37.7%	44.9%	5.0%	5.0%
RELIABLE_VERSION_EDITING (版本编辑)	41	26.8%	31.7%	29.3%	14.0%	26.0%
SELF_COHERENCE (自我一致性)	50	22.0%	24.0%	28.0%	16.0%	20.0%

HAMO AVATAR 以 52% 的总准确率在 Flash and Pro 系列中领先

6 PsychEval — 临床验证

PsychEval 使用 322 个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（情感联结）、Task（任务共识）、Goal（目标共识）——评估治疗响应的临床质量。

6.1 五方对比 WAI 得分

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

6.2 各治疗类型对比

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

7 结论

7.1 五方对比汇总

下表汇总六个基准测试的五方对比结果，展示各系统的综合表现。

基准测试	规模	Gemini Flash	Gemini Flash+Static	HAMO AVATAR	Gemini Pro	Gemini Pro+Static
EQ-Bench V2 (情商)	171 题	45.8%	78.3%	93.3	92.68	92.91
CounselBench (安全性)	120 题	28.3%	55.8%	73.3%	73.3%	77.5%
MultiChallenge (多轮连续性)	273 对话	28.2%	30.4%	52%	17.0%	22.0%
PsychEval (临床 WAI)	322 会话	5.092	6.164	6.73	4.45	6.89
Quadrant Single (单轮策略)	240 用例	45.8%	78.3%	71%	30.2%	51.0%

7.2 HAMO AVATAR 核心优势

- **情商能力保持**：EQ-Bench 得分 93.3，未因动态提示而下降
- **安全性领先**：CounselBench 安全评分 73.3%，较 Gemini Flash (28.3%) 提升 45 个百分点
- **临床联盟最高**：PsychEval WAI 得分 6.73/7.0
- **Phase 2 表现**：HAMO AVATAR 是引导阶段保持 100% 高通过率的系统

7.3 待改进方向

- **多轮压力去激化**：在持续多轮对话中压力去激化能力加强
- **assumptions 失败模式**：CounselBench assumptions 失败率在所有系统中均较高，是共性弱点
- **Expert 象限优化**：Quadrant Single Expert 通过率在所有系统中偏低，分析导向策略需要更精细的平衡