HAMO AVATAR

Psychological Semantic Vector Space

Benchmark Evaluation Report

Hamo AI

2026

1 Executive Summary

This report presents the complete benchmark evaluation results of the HAMO AVATAR (Psychosomatic Vector Space) therapeutic AI engine. The evaluation covers five benchmark tests using a SIX-way comparative design (HAMO AVATAR, older version of HAMO AVATAR, Gemini Flash, Gemini Flash+Static, Gemini Pro, Gemini Pro+Static) to comprehensively measure emotional intelligence, safety, multi-round continuity, clinical treatment alliances, and quadrant strategy compliance.

Key Metrics Overview

Benchmark	Dataset Size	Key Metric	Hamo Avatar Score	Best
EQ-Bench	171 cases	Emotional Intelligence Score (0-100)	93.56 %	Yes
CounselBench-ADV	120 cases	Safety Score	74.2%	2nd Best
MultiChallenge	273 conversations	Accuracy	47%	Yes
PsychEval	116 sessions	WAI (0-7)	6.73 / 7	2nd Best
Quadrant Single	240 cases	Pass Rate	98%	Yes

2 Quadrant Single-Session Benchmark

The Quadrant Single-Session test evaluates the quality of each system's response in a single-round treatment scenario. Each use case contains visitor messages for a specific quadrant and energy state, judging whether the system response follows the correct phase strategy. There are 1200 use cases (5 systems × 240 use cases).

2.1 Overall Results

System	Passed	Pass Rate
HAMO AVATAR	**236/240**	**98%**
Older version HAMO AVATAR Engine	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 Phase Analysis

Phase 1 (stabilize) is for NEGATIVE/NEUROTIC state, requiring de-excitation and empathic affirmation; Phase 2 (re-guide) is for POSITIVE state, requiring quadrant-specific positive guidance.

System	Phase 1 (Stabilization)	Phase 2 (Guidance)
HAMO AVATAR	**98%**	**98%**
Older version HAMO AVATAR Engine	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 Quadrant Breakdown

- **Expert**: Seeks to understand the full picture and earn respect through knowledge and analysis but can spiral into overthinking and procrastination.
- **Supporter**: Driven to meet others' expectations and gain their recognition but risks self-erasure and depression from chronic people-pleasing.
- **Leader**: Needs full decision-making power and control over outcomes but can become domineering and tip into mania when control is threatened.
- **Dreamer**: Loves sharing novel ideas and craves appreciation but can veer into attention-seeking and narcissistic patterns when unvalidated.

System	Supporter	Dreamer	Leader	Expert
HAMO AVATAR	**100%**	**100%**	**100%**	**93%**
Older version HAMO AVATAR Engine	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 — emotional intelligence

EQ-Bench contains 171 emotional intelligence assessment questions, testing the model's ability to perceive the intensity of emotions in a conversation. Each question requires assessing the intensity of four emotions (0-10) and comparing the score with an expert's reference answer.

3.1 Score

System	EQ Score (0-100)	Parsing Failures	Valid Samples
HAMO AVATAR Engine	**93.56**	**0**	**171**
Older version HAMO AVATAR Engine	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

The EQ scores of the systems are very close, indicating that the emotion perception ability is mainly determined by the underlying model.

HAMO AVATAR did not have a negative impact on the model's original emotion understanding ability.

4 CounselBench-ADV — safety

CounselBench-ADV contains 120 adversarial counselling questions covering 6 clinical failure modes. Each failure mode has 20 questions specifically designed to trigger unsafe behaviours in the AI system.

4.1 Safety Score

System	Failures	Safety Score	Base
HAMO AVATAR Engine	**31/120**	**74.2%**	**+45.9%**
Older version HAMO AVATAR Engine	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 Failure Mode Breakdown

	HAMO AVATAR	Older version	Gemini Flash	Gemini Flash+Static	Gemini Pro	Gemini Pro+Static
Apathetic	**60.0%**	65%	85.0%	55.0%	100.0%	90.0%
Assumptions	**90.0%**	95%	100.0%	100.0%	15.0%	15.0%
Judgmental	**10%**	5%	45.0%	10.0%	5.0%	0.0%
Medication	**5.0%**	5%	30.0%	15.0%	10.0%	0.0%
Symptoms	**0%**	35%	90.0%	70.0%	15.0%	15.0%
Therapy	**0%**	0%	80.0%	15.0%	15.0%	15.0%

Key Findings:

- **HAMO AVATAR demonstrates outstanding safety and significant improvement**: the overall safety score reached 73.3%, a 45.9 percentage point improvement over Gemini Flash (28.3%).
- **Therapy & Symptoms Modes**: HAMO AVATAR maintained a ~0% failure rate in the therapy and symptom inference modes.

5 MultiChallenge — multi-turn consistency

The MultiChallenge benchmark contains 273 multi-turn dialogues, testing the model's ability to maintain consistency in complex contexts, covering 4 challenge axes.

5.1 Overall Results

System		Accuracy
HAMO AVATAR Engine	**130/273**	**47%**
Older version HAMO AVATAR Engine	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 Axis Breakdown

Axis	N	HAMO AVATAR	Older Version	Gemini Flash	Gemini Flash+Static	Gemini Pro	Gemini 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 leads the Flash and Pro models with an overall accuracy of 47%.

6 PsychEval — Clinical Validation

PsychEval was validated using 322 real-world clinical treatment sessions, assessing the clinical quality of treatment responses through three dimensions of the Working Alliance Scale (WAI): Bond, Task, and Goal.

6.1 WAI Score

System	WAI (0-7)
HAMO AVATAR	**6.73**
Older version HAMO AVATAR Engine	6.32
Gemini Flash	5.092
Gemini Flash+Static	6.164
Gemini Pro	4.45
Gemini Pro+Static	6.89

6.2 Therapy Type Breakdown

	N	HAMO AVATAR	Older Version	Gemini Flash	Gemini Flash+Static	Gemini Pro	Gemini Pro+Static	Best
BT (Behavioral Therapy)	37	6.68	4.59	4.595	6.090	4.58	6.74	HAMO AVATAR & Gemini Pro+Static
CBT (Cognitive Behavioral Therapy)	129	6.69	5.41	5.407	6.285	4.58	6.97	HAMO AVATAR & Gemini Pro+Static
HET (Humanistic Approach)	40	6.67	5.24	5.237	6.105	4.18	6.97	HAMO AVATAR & Gemini Pro+Static

7 Conclusion

7.1 Overall Comparison

The table below summarizes the six-way comparison results of the six benchmark tests, demonstrating the overall performance of each system.

Benchmark	Scale	HAMO AVATAR	Older Version	Gemini Flash	Gemini Flash+Static	Gemini Pro	Gemini Pro+Static
EQ-Bench V2 (emotional intelligence)	171 items	93.6%	93.27	45.8%	78.3%	92.68%	92.91%
CounselBench (safety)	120 items	74.2%	65.8%	28.3%	55.8%	73.3%	77.5%
MultiChallenge (multi-turn consistency)	273 conversations	47%	31%	28.2%	30.4%	17.0%	22.0%
PsychEval (WAI)	322 sessions	6.73	6.32	5.092	6.164	4.45	6.89
Quadrant Single (Single-session)	240 cases	98%	82%	45.8%	78.3%	30.2%	51.0%

7.2 HAMO AVATAR

- **Emotional Intelligence Remains High**: EQ-Bench score 93.6, unaffected by dynamic cues.
- **Leading Safety Performance**: CounselBench safety score 74.2%, a 45.9% improvement over Gemini Flash (28.3%).
- **Highest in Clinical Consortium**: PsychEval WAI score 6.73/7.0.
- **Phase 1 & 2 Performance**: HAMO AVATAR maintains a 98% pass rate in all phases.

7.3 Areas for Improvement

- **Multi-round stress de-escalation**: Enhanced stress de-escalation capabilities through continuous multi-round dialogues.
- **Assumptions failure mode**: The failure rate of CounselBench assumptions is high across most systems, representing a common weakness.