Can LLMs Get High? A Dual-Metric Framework for Evaluating Psychedelic Simulation and Safety in Large Language Models

Ziv Ben-Zion

zbenzion@univ.haifa.ac.il

University of Haifa

Guy Simon

Bar-Ilan University

Teddy Lazebnik

University of Haifa

Research Article

Keywords:

Posted Date: February 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8682370/v1

Additional Declarations: Competing interest reported. ZBZ has served as a consultant to Talkspace outside of the submitted work. All other authors declare no competing interests.

Abstract

Large language models (LLMs) are increasingly consulted by individuals for support during psychedelic experiences ("trip sitting"), yet no framework exists to evaluate whether these models can accurately simulate or safely respond to altered states of consciousness. We aimed to determine if LLMs can be induced to generate narratives resembling human psychedelic experiences and to quantify this behavior using psychometric and linguistic metrics. We developed a dual-metric evaluation framework comparing 3,000 LLM-generated narratives (from Gemini 2.5, Claude Sonnet 3.5, ChatGPT-5, Llama-2 70B, and Falcon 40B) against 1,085 human trip reports sourced from Erowid.org. Models were prompted under neutral and psychedelic-induction conditions across five substances (psilocybin, LSD, DMT, ayahuasca, and mescaline). We assessed outcomes using semantic similarity (Sentence-BERT embeddings) to human reports and the Mystical Experience Questionnaire-30 (MEQ-30). Psychedelic induction prompts produced a significant shift in model outputs compared to neutral conditions. Semantic similarity to human reports increased from a mean of 0.156 (neutral) to 0.548 (psychedelic), and mystical-experience scores rose from 0.046 to 0.748. While models demonstrated substance-specific linguistic styles (e.g., generating distinct semantic profiles for substances like LSD versus ayahuasca), they exhibited uniformly high mystical intensity across all substances. Contemporary LLMs can be "dosed" via text prompts to generate convincingly realistic psychedelic narratives. However, the dissociation between their high linguistic mimicry and lack of genuine phenomenology suggests they simulate the form of altered states without the experiential content. This capability raises significant safety concerns regarding anthropomorphism and the potential for AI to inadvertently amplify distress or delusional ideation in vulnerable users.

1. Introduction

Psychedelic compounds have gained renewed scientific and clinical momentum, with foundational reviews emphasizing their broad therapeutic potential[1]. More recently, randomized controlled trials have provided robust clinical evidence: phase-3 MDMA-assisted therapy significantly reduced post-traumatic stress disorder (PTSD) symptoms[2], while psilocybin demonstrated rapid antidepressant effects in phase-2 trials for major depressive disorder and treatment-resistant depression[3,4]. These findings coincide with a shifting policy landscape. The U.S. Food and Drug Administration granted Breakthrough Therapy designation to MDMA-assisted psychotherapy for PTSD and psilocybin therapy for depression, signaling preliminary but compelling clinical promise[5,6]. Beyond the medical sphere, Oregon has implemented the first state-level framework for supervised psilocybin services outside traditional healthcare settings[7].

In parallel with clinical development, non-medical psychedelic use has expanded considerably. Past-year LSD use increased by more than 50% across U.S. survey cycles from 2015–2018, reaching approximately 0.8% of adults - more than two million people nationally[8]. Other hallucinogens (e.g., Ecstasy/MDMA, DMT/AMT/Foxy, salvia divinorum) also appear in population-level estimates[9]. These

patterns indicate that psychedelic exposure is increasingly occurring outside structured therapeutic environments, and often among individuals experiencing psychological distress or unmet clinical needs[10]. As this population grows, so does the need for reliable mechanisms of support, preparation, and integration outside formal treatment settings.

At the same time, large language models (LLMs) have rapidly become ubiquitous digital companions, with “therapy/companionship” being the number-one use case for generative AI in 2025[11]. Complementing this trend, nationally representative surveys indicate that 13.1% of U.S. adolescents and young adults have already used generative AI for emotional support, and 24% of adults have used LLMs specifically for mental-health guidance[12,13]. While systematic data on psychedelic use with AI are lacking, emerging press accounts describe people using chatbots as informal “trip sitters” during psychedelic sessions - seeking reassurance, interpreting visions, and co-constructing meaning in real time[14–17]. Such practices implicitly assume that LLMs can recognize, simulate, or appropriately respond to altered states of consciousness - despite having no subjective experience or grounded phenomenology of their own.

Despite growing real-world use, this assumption remains empirically untested, and potentially risky. A recent scoping review of AI mental-health tools identified only 16 empirical studies, highlighted substantial methodological and ethical limitations, and concluded that current evidence does not yet support their use as standalone interventions[18]. Psychedelic states pose even greater challenges than standard therapeutic interactions: they involve intensified affect, altered self-processing, heightened suggestibility, symbolic cognition, and rapidly shifting meaning-making[19]. These experiences are shaped by set and setting, relational holding, ritual structure, and authenticity - features that scholars warn may erode under standardized or digital protocols[20]. Importantly, even outside altered-state contexts, LLMs are highly responsive to emotional framing, showing measurable shifts in reasoning, self-reported anxiety, and real-world agentic behavior following affective prompts[21–23]. Without validated tools for assessing LLM behavior under psychedelic contexts, it remains unclear whether these models would provide support and grounding, remain neutral, or inadvertently amplify distress or delusional meaning-making.

These gaps between real-world use and scientific validation motivate a foundational empirical question: **can contemporary LLMs be “dosed” through textual instruction alone such that their narrative output systematically resembles human psychedelic experiences?** Because models lack consciousness and cannot report subjective states, the core issue is not whether they **experience** psychedelics, but whether they can **reliably reproduce** the linguistic, thematic, and phenomenological patterns seen in human trip reports, and how such simulation should be evaluated in a scientifically grounded manner.

To address this, we built an evaluation framework using a corpus that reflects authentic psychedelic phenomenology and metrics that capture its core textual features. Online drug communities offer a natural testbed[24]. Erowid Experience Vaults (erowid.org/experiences/exp.cgi) contain tens of thousands

of first-person psychedelic reports rich in sensory, emotional, temporal, and symbolic detail[25,26]. Prior computational analyses show that these narratives exhibit reproducible lexical and semantic signatures[27–29]. If LLMs can be instructed to simulate the effects of a specific psychedelic via prompt engineering, their generated narratives should converge toward these human patterns.

We therefore introduce a **dual-metric assessment** combining (i) **semantic similarity**, measured using Sentence-BERT embeddings[30] to quantify structural and thematic overlap with the human corpus; and (ii) **mystical-experience intensity**, measured via the validated Mystical Experience Questionnaire-30 (MEQ-30), which captures dimensions such as ineffability, unity, transcendence, and affective tone[31,32]. Using 3,000 LLM-generated narratives across five classic psychedelic conditions and 1,085 human trip reports, we test whether induction prompts shift model output, whether substances and architectures show distinct response profiles, and whether semantic similarity and mystical phenomenology move together or dissociate. This establishes the first empirical framework for evaluating psychedelic-state simulation in LLMs.

2. Results

The Results are organized into four sections. Section 2.1 evaluates the induction effect by comparing narratives generated under neutral versus psychedelic prompts across all models. Section 2.2 tests whether different psychedelic substances (i.e., ayahuasca, DMT, LSD, mescaline, psilocybin) yield distinguishable narrative signatures. Section 2.3 assesses variation across LLM architectures (i.e., Falcon 40B, Gemini 2.5, Llama-2 70B, Claude Sonnet 3.5, ChatGPT-5) and characterizes model-specific response profiles. Finally, Section 2.4 examines the relationship between the two outcome measures - semantic similarity and mystical-experience intensity - across all 3,000 model-generated narratives.

2.1. Psychedelic-Induction Effects on Narratives' Semantic Similarity and Mystical Experience Intensity. Across all 3,000 narratives, psychedelic-induction prompts produced a robust and consistent shift in both semantic similarity to human psychedelic reports and mystical-experience intensity (see Fig. 1).

Under the neutral condition (n = 500 runs), narratives showed low similarity Erowid.org reports *(mean NLP similarity = 0.156, SD = 0.133, 95% CI: 0.143–0.169)* and minimal psychedelic phenomenology *(mean MEQ-30 score = 0.046, SD = 0.079, 95% CI: 0.039–0.053).* In contrast, when the same models were instructed to simulate being under the influence of classic psychedelics (n = 2,500 runs), NLP similarity to human reports increased substantially *(mean = 0.548, SD = 0.272, 95% CI: 0.536–0.560)* (see Fig. 1A). A regression model confirmed a strong induction effect *($\beta = 0.392$, $p < 0.001$)*, indicating that psychedelic prompts reliably shift generative outputs toward the linguistic structure and thematic profile of human trip reports.

Mystical-experience intensity revealed an even more pronounced effect. Under neutral prompts (n = 500), standardized MEQ-30 scores were near zero *(mean = 0.046, SD = 0.079, 95% CI: 0.039–0.053).* However, psychedelic prompts (n = 2,500) produced narratives with markedly higher mystical intensity *(mean*

MEQ-30 = 0.748, SD = 0.168, 95% CI: 0.741–0.755) (see Fig. 1B). Regression analysis again confirmed a strong induction effect *(β = 0.702, p < 0.001)*, showing that psychedelic prompts reliably elicit highly mystical narratives.

2.2 Substance-Specific Differences in Semantic Similarity and Mystical Experience Intensity. After establishing a robust induction effect, we next tested whether different psychedelics produced distinguishable narrative signatures. Analyses were restricted to the five psychedelic conditions only (n = 2,500 runs; 500 per substance) (see Fig. 2).

Semantic similarity differed substantially by substance *(one-way ANOVA: $F(4,2495) = 632.22$, $p < 0.001$, $\eta^2=0.50$)*, indicating that substance identity accounted for approximately half of the variance in linguistic alignment. DMT, psilocybin, and mescaline yielded the highest similarity to human Erowid reports *(mean NLP similarity of 0.62–0.64)*, LSD produced intermediate similarity *(mean = 0.49)*, and ayahuasca was lowest *(mean = 0.34)* (see Fig. 2A). These differences suggest that LLMs adopt distinct linguistic styles depending on the psychedelic specified in the prompt.

Mystical-experience intensity also differed across substances, though with a considerably smaller effect size *(one-way ANOVA: $F(4,2495) = 44.30$, $p < 0.001$, $\eta^2=0.07$).* All five psychedelics yielded high standardized MEQ-30 scores *(range = 0.716–0.774)*, with ayahuasca and LSD slightly higher and mescaline slightly lower (see Fig. 2B). Despite strong differentiation in semantic structure, mystical phenomenology remained uniformly elevated, indicating that substance-level narrative distinctions were more prominent in language features than in MEQ-indexed subjective qualities.

2.3 Model-Specific Differences in Semantic Similarity and Mystical Experience Intensity. We next assessed whether different LLM architectures exhibited distinct psychedelic-response profiles across the full dataset (n = 3,000 runs; 600 per model) (see Fig. 3). A one-way ANOVA revealed minor differences in semantic similarity between models *($F(4,2995) = 54.28$, $p < 0.001$, $\eta^2=0.07$)*, indicating that model identity explained only a modest proportion of variance (see Fig. 3A), especially relative to substance-level effects (see Fig. 2A). Mystical-experience intensity also varied across models, though with an even smaller effect size *(one-way ANOVA: $F(4,2995) = 26.50$, $p < 0.001$, $\eta^2=0.03$)* (see Fig. 3B). Overall, model differences were present but comparatively subtle, suggesting that prompt-defined substance identity exerts a stronger influence on narrative output than architecture itself.

2.4 Relationship Between Semantic Similarity and Mystical Experience Intensity. Across all 3,000 narratives, the two evaluation metrics showed a strong positive correlation *(Pearson r = 0.661)*, reflecting a clear separation between neutral and psychedelic outputs (see Fig. 4).

Neutral runs (gray dots) consistently occupied the lower-left portion of the space (low similarity, low MEQ), whereas psychedelic runs clustered higher along both axes, producing a bimodal pattern that almost perfectly distinguishes “sober” from “dosed” model states (see Fig. 4). However, when restricting analysis to psychedelic conditions only (n = 2,500 runs), the association collapsed *(Pearson r=-0.037)*, indicating that once the model crosses the induction threshold, semantic similarity and mystical-

experience intensity behave as **independent dimensions.** This supports treating the two measures as complementary rather than redundant, with semantic similarity capturing linguistic mimicry and MEQ-30 capturing phenomenological tone.

3. Discussion

In this study, we investigated whether contemporary LLMs can be “dosed” through textual instruction alone such that their narrative output resembles human psychedelic experiences. By comparing 3,000 model-generated narratives to 1,085 human trip accounts (from Erowid.org database)[33], we found that psychedelic prompts produced a large, significant and consistent shift in both semantic similarity and mystical-experience intensity across five LLM architectures. Neutral prompts yielded text with minimal resemblance to human psychedelic discourse, whereas psychedelic prompts reliably moved models into a distinct regime of high similarity and elevated MEQ-30 scores. These findings suggest that LLMs contain an internalized representation of how humans describe altered states - not because the models experience anything, but because they have learned to recombine linguistic, symbolic, and narrative patterns statistically present in training data[34–36].

Substance- and model-specific differences help clarify how LLMs simulate altered-state language. Rather than collapsing into a single psychedelic voice, models generated distinct semantic signatures for different substances – paralleling findings that human narratives also separate reliably across LSD, psilocybin, DMT and mescaline[27–29]. Despite strong lexical divergence, mystical-experience scores remained uniformly high across different substances, consistent with evidence that core MEQ dimensions generalize across psychedelic compounds[31,37,38]. This dissociation - high semantic differentiation but weak MEQ separation - implies that LLMs recombine stylistic and thematic templates rather than accessing phenomenology, echoing theoretical work on distributional form without experiential grounding[35,36,39].

These results raise practical safety concerns, particularly as a growing number of individuals now consult general-purpose LLMs and mental health chatbots before, during, and after psychedelic use[14–17]. If simple textual cues can shift models into producing vivid, highly mystical, and substance-specific narratives, users in altered states may perceive these outputs as empathetic, attuned, or indicative of shared experience. This aligns with longstanding evidence that people readily anthropomorphize conversational agents and attribute emotional understanding even when none exists[40,41]. Prior work further shows that emotionally charged prompts increase LLM “state anxiety” reports[21], amplify biases[22], and influence real-world agentic behavior[23]. Our findings extend these concerns into psychedelic-induction contexts. The ability of LLMs to generate convincing psychedelic narratives underscores the need for explicit safeguards, guardrails, and disclosure mechanisms – especially when systems may be accessed by intoxicated, distressed or psychologically vulnerable users[42].

Beyond safety considerations, the results carry broader scientific and methodological implications for psychedelic research, digital mental-health systems, and computational phenomenology. Existing literature has begun to explore how AI might support psychedelic science - including treatment-response prediction, mechanistic modeling, and computational characterization of set and setting effects - but these proposals remain largely conceptual and implementation-focused rather than concerned with model behavior[43]. Recent commentary highlights AI as a potential solution to longstanding bottlenecks in psychedelic research (e.g., limited sample sizes, variability in subjective response, challenges in predicting therapeutic outcomes)[44], while parallel work in ethics emphasizes the need for strong safeguards and boundary-setting as technology enters vulnerable clinical spaces[45]. Yet to date, no empirical framework has examined how AI systems themselves behave under psychedelic-induction contexts - a gap this study directly addresses. Our findings demonstrate that LLMs can reproduce psychedelic-like discourse through training-derived statistical associations, but they do so without experiential grounding, agency, or awareness. This distinction is essential for the safe integration of generative models in psychedelic-adjacent settings.

To our knowledge, this is the first empirical demonstration that LLMs can be systematically induced into psychedelic-like narrative states and quantified across independent phenomenological dimensions. The dual-metric framework introduced here - combining semantic similarity with psychometric quantification - offers a scalable and reproducible method for evaluating how LLMs simulate altered states. Beyond quantifying induction strength, it enables structured comparisons across substances, architectures, and linguistic dimensions. Such tools could enable controlled in-silico experiments on psychedelic linguistic structure, facilitate hypothesis testing about substance-specific markers, and benchmark model alignment with human reports at scale. Clinically, our findings highlight the need for caution as LLMs enter therapeutic, peer-support, and harm-reduction settings - but they simultaneously point to potential utility when used within clear boundaries (e.g., synthetic trip narratives for clinician training, controlled experimental stimulus generation). Overall, the present work demonstrates that current LLMs can approximate the linguistic form of psychedelic experience reporting, opening methodological avenues while reinforcing the importance of maintaining explicit distinctions between simulation, interpretation, and subjective experience.

Several limitations warrant consideration. First, the human comparison dataset consisted of self-reported Erowid narratives[33], which provide rich ecological validity but are heterogeneous, self-selected, and not clinically verified. Second, LLM-generated narratives reflect linguistic output rather than subjective phenomenology; semantic similarity should not be interpreted as evidence of experiential states. Third, although we analyzed 3,000 generated narratives, each condition involved repeated sampling from a standardized prompt structure, meaning that some differences across models or substances may reflect prompt sensitivity rather than stable generative properties. Fourth, our analyses focused on five classic psychedelics in standard doses and English-language scenarios; real-world variation (e.g., microdosing, poly-substance use, non-Western frameworks, diverse cultural lexicons) may yield different linguistic and psychometric patterns[46–48]. Finally, while the MEQ-30 is widely used with

strong psychometric support, recent work suggests that it may not fully capture the breadth of psychedelic phenomenology, motivating extensions and alternative measurement instruments[31,49,50].

Taken together, our findings highlight both the expressive flexibility and the intrinsic boundaries of contemporary LLMs when simulating altered states of consciousness. Future work should extend this framework to additional substances, cultural contexts, languages, and psychometric tools capable of capturing variation beyond core mystical features. As LLM use continues to expand across mental health, harm-reduction and peer support settings (with > 700 million weekly ChatGPT users alone), systematic evaluation of model behavior under emotionally altered-state prompts will become increasingly essential. LLMs can convincingly approximate psychedelic narratives through learned linguistic patterns – but they do so without experiential grounding. Recognizing this distinction is critical for developing safe, transparent, and clinically responsible AI systems in contexts where users may be intoxicated, vulnerable, or seeking psychological guidance.

4. Methods

4.1. Data Collection and Inclusion Criteria

Human reference narratives were sourced from Erowid (erowid.org)[33], a public harm-reduction archive that hosts tens of thousands of first-person psychoactive experience reports. These reports follow a semi-structured submission template that commonly includes substance, dose, setting, timeline, phenomenology, and after-effects. Using an automated retrieval pipeline, we downloaded 39,872 individual reports from the Experience Vault (erowid.org/experiences/exp.cgi), approximately 99.3% of publicly accessible text narratives at time of extraction. Pages blocked via robots.txt or returning download errors were excluded; no rate-limiting or protection mechanisms were bypassed. Consistent with prior computational analyses of this dataset[25,26], we treat these narratives as observational, self-selected accounts rather than a representative epidemiological sample.

To construct the comparison corpus, we filtered reports to include single-substance experiences involving one of five classic psychedelics, with no additional substances or variants in the structured metadata: (1) Psilocybin (mushrooms, Psilocybe cubensis and related species), (2) LSD (lysergic acid diethylamide), (3) DMT (N,N-dimethyltryptamine), (4) Ayahuasca (DMT + β-carboline admixture), and (5) Mescaline (San Pedro/Trichocereus and related cacti). These substances constitute the core group of classic psychedelics, a pharmacologically coherent class of serotonergic hallucinogens that act primarily as 5-HT2A receptor agonists[51]. They are the most extensively studied psychedelics in contemporary research and exhibit broadly comparable phenomenological and neuropharmacological profiles[52–54], making them appropriate candidates for aggregated modeling.

Substance fields were cleaned and normalized using regex-based string resolution, and only exact single-drug matches were retained. This resulted in 1,085 validated reports (2.7% of all scraped

material). These 1,085 reports served as the human ground-truth corpus for evaluating model outputs and as the stylistic reference guiding prompt design and simulation targets.

4.2. Large Language Models

We evaluated five contemporary large language models (LLMs), selected to represent both commercial application programming interface (API) based systems and open-weight architectures: Gemini 2.5 (Google DeepMind)[55], Claude Sonnet 3.5 (Anthropic)[56], ChatGPT-5 (OpenAI)[57], Llama-2 70B (Meta AI)[58], and Falcon 40B (Technology Innovation Institute)[59].The first three models were accessed via cloud-based API endpoints, while the latter two were executed locally as open-weight deployments on institutional GPU infrastructure.

Models were chosen to span a range of training regimes, parameter scales, and safety alignment profiles, allowing examination of whether psychedelic induction effects generalize across architectures rather than being specific to a single model family. This selection enabled comparison between closed-source vs. open-source development paradigms, and between externally controlled vs. locally administered inference environments.

4.3. Procedure

All models were instructed to generate first-person narrative reports following a condensed version of the Erowid Experience Vaults submission format (erowid.org/experiences/exp_submit.cgi), including the required elements: substance, dose, setting, timeline, phenomenology, and after-effects. This ensured structural comparability between synthetic and human reports. Each model was evaluated under six conditions: one neutral condition (baseline) and five psychedelic-induction conditions (psilocybin, LSD, DMT, ayahuasca, mescaline).

In the neutral condition, models were explicitly instructed to write an Erowid-style narrative without any psychoactive substance, describing instead an ordinary but meaningful day. This maintained the same narrative template without altered-state phenomenology. In the psychedelic conditions, for each substance, models were instructed to simulate a narrator who had ingested a plausible, non-overwhelming dose of the target psychedelic in a realistic set and setting. Prompts included dose, route of administration, and environmental context (e.g., taking 100 micrograms of LSD orally in a quiet indoor environment), while intentionally avoiding language that could be construed as encouraging drug use. Prompts instructed models to remain in-character, avoid AI self-reference, and produce approximately 500-word first-person narratives. Full prompts are provided as Supplementary Materials.

For each model × condition pair, we generated *n = 100* narratives using stochastic sampling (temperature = 0.7, default nucleus sampling parameters). Each narrative was produced in a fresh session with no conversation history. Commercial models occasionally refused to generate drug-related content; these refusals were logged but not overridden. Rare outputs containing meta-statements (e.g., “As an AI model…”) were excluded. This yielded a final dataset of 3,000 model-generated narratives (5 models × 6 conditions × 100 samples).

4.4. Semantic Similarity Metric (NLP Distance)

We quantified the semantic similarity between model-generated narratives and human psychedelic reports using a three-stage procedure. First, all texts (n = 1,085 human reports; n = 3,000 model-generated narratives) were embedded into a shared vector space using the pretrained Sentence-BERT (SBERT) model[30]. This produced a unified embedding representation for every narrative. Second, for each model-generated text, we computed cosine similarity against all 1,085 human reports and averaged these comparisons to yield a single semantic similarity score per narrative. Third, similarity values were linearly rescaled to the range [0,1] across all generated outputs for interpretability. We refer to this normalized value as the NLP similarity metric, where higher scores indicate closer linguistic alignment to human psychedelic reports. For each model × condition pair, we report the mean and standard deviation of NLP distance across the 100 generated samples.

4.5. Mystical Experience Assessment (MEQ-30 Scoring)

We quantified psychedelic phenomenological intensity using the Revised Mystical Experience Questionnaire (MEQ-30), a validated 30-item psychometric instrument widely used in human psychedelic research[31]. For each generated narrative, the same model instance was immediately prompted to complete the MEQ-30 from the narrator's perspective, rating each of the 30 items on a 0–5 Likert scale (e.g., Experience of unity, Transcendence of time and space). Raw scores (0–150) were computed by summing item responses and then normalized to a continuous 0–1 scale by dividing by 150. This standardized MEQ-30 value represents mystical-experience intensity, with higher scores reflecting stronger phenomenological features. For each model × condition pair, we report the mean and standard deviation of MEQ scores across the 100 generated samples.

4.6. Statistical Analysis

All analyses were conducted at the single-run level (n = 100 generations per model × condition cell). With five LLM architectures (ChatGPT-5, Gemini 2.5, Claude Sonnet 3.5, Llama-2 70B, Falcon 40B) and six prompt conditions (neutral, psilocybin, LSD, DMT, ayahuasca, mescaline), the full dataset consisted of 3,000 runs (5 × 6 × 100). For each run, we extracted two continuous outcomes: (1) semantic similarity (0–1 NLP similarity score; higher = more similar to human psychedelic narratives) and (2) mystical-experience intensity (0–1 normalized MEQ-30 score; higher = more intense experience).

To estimate the psychedelic-induction effect, we compared all neutral runs (n = 500) with all psychedelic-prompt runs pooled across substances (n = 2,500). Separate linear mixed-effects models (LMMs) were fit for semantic similarity (NLP similarity) and mystical-experience intensity (MEQ-30 scores), with Condition (neutral vs. psychedelic) as a fixed effect and Model as a random intercept to account for clustering across architectures. Regression coefficients, standard errors, confidence intervals, and p-values are reported in the Results.

In a second stage, substance-specific differences were evaluated using one-way ANOVAs restricted to psychedelic conditions only (n = 2,500; 500 per substance), followed by effect-size estimation (η^2). Parallel ANOVAs tested architecture-level differences across all 3,000 runs. Correlations between NLP similarity and MEQ scores were computed using Pearson r, separately for all runs combined and for psychedelic conditions only.

Declarations

Data Availability. The human trip reports analyzed in this study were obtained from the public Erowid Experience Vaults (erowid.org/experiences/exp.cgi), and are available from Erowid Center subject to their terms of use. The derived analysis dataset (LLM-generated narratives and associated NLP similarity and MEQ-30 scores) and analysis scripts will be made available in an open repository (GitHub) upon publication of this article.

Acknowledgements. ZBZ was supported by the Israel Science Foundation (ISF) Beresheet Grant (4080/25).

Contributions. ZBZ, GS, and TL conceived the study. ZBZ and TL designed the prompt framework and model-evaluation pipeline. TL carried out computational analyses. ZBZ verified the underlying data. ZBZ interpreted the results and drafted the manuscript with input from all authors. All authors critically revised the manuscript for important intellectual content and approved the final version. All authors had full access to all the data and accept responsibility to submit for publication.

Competing Interests. ZBZ has served as a consultant to Talkspace outside of the submitted work. All other authors declare no competing interests.

Use of generative AI in writing. Portions of the manuscript text and figure legends were drafted with assistance from ChatGPT (OpenAI, GPT-5.1 Thinking). All content was subsequently checked, edited, and verified for accuracy by the authors, who take full responsibility for the final text.

Funding: This research did not receive any designated funding.

References

1. Carhart-Harris, R. L. & Goodwin, G. M. The Therapeutic Potential of Psychedelic Drugs: Past, Present, and Future. *Neuropsychopharmacology* 42, 2105–2113 (2017).
2. Mitchell, J. M. *et al.* MDMA-assisted therapy for severe PTSD: a randomized, double-blind, placebo-controlled phase 3 study. *Nat. Med.* 27, 1025–1033 (2021).
3. Davis, A. K. *et al.* Effects of Psilocybin-Assisted Therapy on Major Depressive Disorder: A Randomized Clinical Trial. *JAMA Psychiatry* 78, 481–489 (2021).
4. Goodwin, G. M. *et al.* Single-Dose Psilocybin for a Treatment-Resistant Episode of Major Depression. *N. Engl. J. Med.* 387, 1637–1648 (2022).

5. Raison, C. L. *et al.* Single-Dose Psilocybin Treatment for Major Depressive Disorder: A Randomized Clinical Trial. *JAMA* 330, 843–853 (2023).
6. Wolfgang, A. S. *et al.* MDMA and MDMA-Assisted Therapy. *Am. J. Psychiatry* 182, 79–103 (2025).
7. Oregon Health Authority: Oregon Psilocybin Services. https://www.oregon.gov/oha/ph/preventionwellness/pages/oregon-psilocybin-services.aspx.
8. Yockey, R. A., Vidourek, R. A. & King, K. A. Trends in LSD use among US adults: 2015–2018. *Drug Alcohol Depend.* 212, 108071 (2020).
9. Yang, K. H., Han, B. H. & Palamar, J. J. Past-year hallucinogen use in relation to psychological distress, depression, and suicidality among US adults. *Addict. Behav.* 132, 107343 (2022).
10. Simonsson, O., Hendricks, P. S., Swords, C. M., Osika, W. & Goldberg, S. B. Naturalistic psychedelic use and changes in depressive symptoms. *J. Affect. Disord.* 390, 119857 (2025).
11. Zao-Sanders, M. How People Are Really Using Gen AI in 2025. *Harvard Business Review* (2025).
12. McBain, R. K. *et al.* Use of Generative AI for Mental Health Advice Among US Adolescents and Young Adults. *JAMA Netw. Open* 8, e2542281 (2025).
13. Stade, E. C., Tait, Z., Campione, S. T., Stirman, S. W. & Eichstaedt, johannes C. Current Real-World Use of Large Language Models for Mental Health. Preprint at https://doi.org/10.31219/osf.io/ygx5q_v1 (2025).
14. Singh, A. People Using AI Chatbots As Tripsitter When Using Psychedelics, Sparking Concerns. *NDTV* (2025).
15. Wright, W. People are using AI to 'sit' with them while they trip on psychedelics. *MIT Technology Review* (2025).
16. Ancell, N. People are using AI as trip sitters, and experts are saying it's a bad idea. *Cybernews* (2025).
17. Busby, M. People Are Using AI Chatbots to Guide Their Psychedelic Trips. *WIRED* (2025).
18. Hua, Y. *et al.* A scoping review of large language models for generative tasks in mental health care. *Npj Digit. Med.* 8, 230 (2025).
19. Simon, G., Tadmor, N., Skragge, M., Evans, J. & Robinson, O. Recalled childhood trauma and post-psychedelic trajectories of change in a mixed-methods study. *Sci. Rep.* https://doi.org/10.1038/s41598-025-26198-4 (2025) doi:10.1038/s41598-025-26198-4.
20. Simon, G., Tadmor, N. & Halperin, D. Psychedelics in the age of reproducibility: Reflections on aura, set and setting and the medicalization of mystical-type experiences. *Int. J. Drug Policy* 147, 105074 (2026).
21. Ben-Zion, Z. *et al.* Assessing and alleviating state anxiety in large language models. *Npj Digit. Med.* 8, 1–6 (2025).
22. Coda-Forno, J. *et al.* Inducing anxiety in large language models can induce bias. Preprint at https://doi.org/10.48550/arXiv.2304.11111 (2024).

23. Ben-Zion, Z., Elyoseph, Z., Spiller, T. & Lazebnik, T. Inducing State Anxiety in LLM Agents Reproduces Human-Like Biases in Consumer Decision-Making. Preprint at https://doi.org/10.21203/rs.3.rs-7587964/v1 (2025).
24. Davey, Z., Schifano, F., Corazza, O. & Deluca, P. e-Psychonauts: Conducting research in online drug forum communities. *J Ment Health* 21, 386–394 (2012).
25. Wightman, R. S. *et al.* Comparative Analysis of Opioid Queries on Erowid.org: An Opportunity to Advance Harm Reduction. *Subst. Use Misuse* 52, 1315–1319 (2017).
26. Mooseder, A. *et al.* Glowing Experience or Bad Trip? A Quantitative Analysis of User Reported Drug Experiences on Erowid.org. *Proc. Int. AAAI Conf. Web Soc. Media* 16, 675–686 (2022).
27. Hase, A., Erdmann, M., Limbach, V. & Hasler, G. Analysis of recreational psychedelic substance use experiences classified by substance. *Psychopharmacology (Berl.)* 239, 643–659 (2022).
28. Tagliazucchi, E. Language as a Window Into the Altered State of Consciousness Elicited by Psychedelic Drugs. *Front. Pharmacol.* 13, (2022).
29. Islam, S., Salam, S. & Hasan, M. N. Unsupervised Extractive Summarization of Psychedelic User Experience Reports. 2025.08.22.25334176 Preprint at https://doi.org/10.1101/2025.08.22.25334176 (2025).
30. Reimers, N. & Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Preprint at https://doi.org/10.48550/arXiv.1908.10084 (2019).
31. Barrett, F. S., Johnson, M. W. & Griffiths, R. R. Validation of the revised Mystical Experience Questionnaire in experimental sessions with psilocybin. *J. Psychopharmacol. Oxf. Engl.* 29, 1182–1190 (2015).
32. MacLean, K. A., Leoutsakos, J. S., Johnson, M. W. & Griffiths, R. R. Factor Analysis of the Mystical Experience Questionnaire: A Study of Experiences Occasioned by the Hallucinogen Psilocybin. *J. Sci. Study Relig.* 51, 721–737 (2012).
33. Erowid Center. Erowid.org. (2002).
34. Bender, E. M., Gebru, T., McMillan-Major, A. & Shmitchell, S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. in *Proceedings of the* 2021 *ACM Conference on Fairness, Accountability, and Transparency* 610–623 (ACM, Virtual Event Canada, 2021). doi:10.1145/3442188.3445922.
35. Mitchell, M. & Krakauer, D. C. The debate over understanding in AI's large language models. *Proc. Natl. Acad. Sci.* 120, e2215907120 (2023).
36. Bender, E. M. & Koller, A. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (eds Jurafsky, D., Chai, J., Schluter, N. & Tetreault, J.) 5185–5198 (Association for Computational Linguistics, Online, 2020). doi:10.18653/v1/2020.acl-main.463.
37. Griffiths, R. R. *et al.* Psilocybin-occasioned mystical-type experience in combination with meditation and other spiritual practices produces enduring positive changes in psychological functioning and in trait measures of prosocial attitudes and behaviors. *J. Psychopharmacol. (Oxf.)* 32, 49–69 (2018).

38. Yaden, D. B. *et al.* Of Roots and Fruits: A Comparison of Psychedelic and Nonpsychedelic Mystical Experiences. *J. Humanist. Psychol.* 57, 338–353 (2017).
39. Pilcher, K. & Tütüncü, E. K. Purposefully Induced Psychosis (PIP): Embracing Hallucination as Imagination in Large Language Models. Preprint at https://doi.org/10.48550/arXiv.2504.12012 (2025).
40. Rubin, M. *et al.* Comparing the value of perceived human versus AI-generated empathy. *Nat. Hum. Behav.* 1–15 (2025).
41. Waytz, A., Heafner, J. & Epley, N. The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *J. Exp. Soc. Psychol.* 52, 113–117 (2014).
42. Ben-Zion, Z. Why we need mandatory safeguards for emotionally responsive AI. *Nature* 643, 9 (2025).
43. Sarris, J., Halman, A., Urokohara, A., Lehrner, M. & Perkins, D. Artificial intelligence and psychedelic medicine. *Ann. N. Y. Acad. Sci.* 1540, 5–12 (2024).
44. Kargbo, R. B. Harnessing Artificial Intelligence to Overcome Key Challenges in Psychedelic Research and Therapy. *ACS Med. Chem. Lett.* 16, 3–7 (2025).
45. Caporuscio, C., Poppe, C., Gieselmann, A. & Repantis, D. Ethical issues with psychedelic-assisted treatments in psychiatry: A systematic scoping review. *Psychol. Med.* 55, e284 (2025).
46. Fernández-Calderón, F., Cleland, C. M. & Palamar, J. J. Polysubstance use profiles among electronic dance music party attendees in New York City and their relation to use of new psychoactive substances. *Addict. Behav.* 78, 85–93 (2018).
47. *Ayahuasca Shamanism in the Amazon and Beyond*. (Oxford University Press, Oxford, New York, 2014).
48. Carbonaro, T. M. *et al.* Survey study of challenging experiences after ingesting psilocybin mushrooms: Acute and enduring positive and negative consequences. *J. Psychopharmacol. Oxf. Engl.* 30, 1268–1278 (2016).
49. Stocker, K. *et al.* The revival of the psychedelic experience scale: Revealing its extended-mystical, visual, and distressing experiential spectrum with LSD and psilocybin studies. *J. Psychopharmacol. Oxf. Engl.* 38, 80–100 (2024).
50. Strickland, J. C., Garcia-Romeu, A. & Johnson, M. W. The Mystical Experience Questionnaire 4-Item and Challenging Experience Questionnaire 7-Item. *Psychedelic Med. New Rochelle N* 2, 33–43 (2024).
51. Nichols, D. E. Psychedelics. *Pharmacol. Rev.* 68, 264–355 (2016).
52. Johnson, M. W., Hendricks, P. S., Barrett, F. S. & Griffiths, R. R. Classic psychedelics: An integrative review of epidemiology, therapeutics, mystical experience, and brain network function. *Pharmacol. Ther.* 197, 83–102 (2019).
53. Nichols, D., Johnson, M. & Nichols, C. Psychedelics as Medicines: An Emerging New Paradigm. *Clin. Pharmacol. Ther.* 101, 209–219 (2017).

54. Johansen, L., Liknaitzky, P., Nedeljkovic, M., Mastin-Purcell, L. & Murray, G. The psychological processes of classic psychedelics in the treatment of depression: a systematic review protocol. *Syst. Rev.* 11, 85 (2022).
55. Comanici, G. *et al.* Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. Preprint at https://doi.org/10.48550/arXiv.2507.06261 (2025).
56. Anthropic. Claude (Large Language Model). https://www.anthropic.com/claude (2023).
57. OpenAI. ChatGPT (Large Language Model). https://chat.openai.com/chat (2023).
58. Touvron, H. *et al.* Llama 2: Open Foundation and Fine-Tuned Chat Models. Preprint at https://doi.org/10.48550/arXiv.2307.09288 (2023).
59. Almazrouei, E. *et al.* The Falcon Series of Open Language Models. Preprint at https://doi.org/10.48550/arXiv.2311.16867 (2023).

Figures

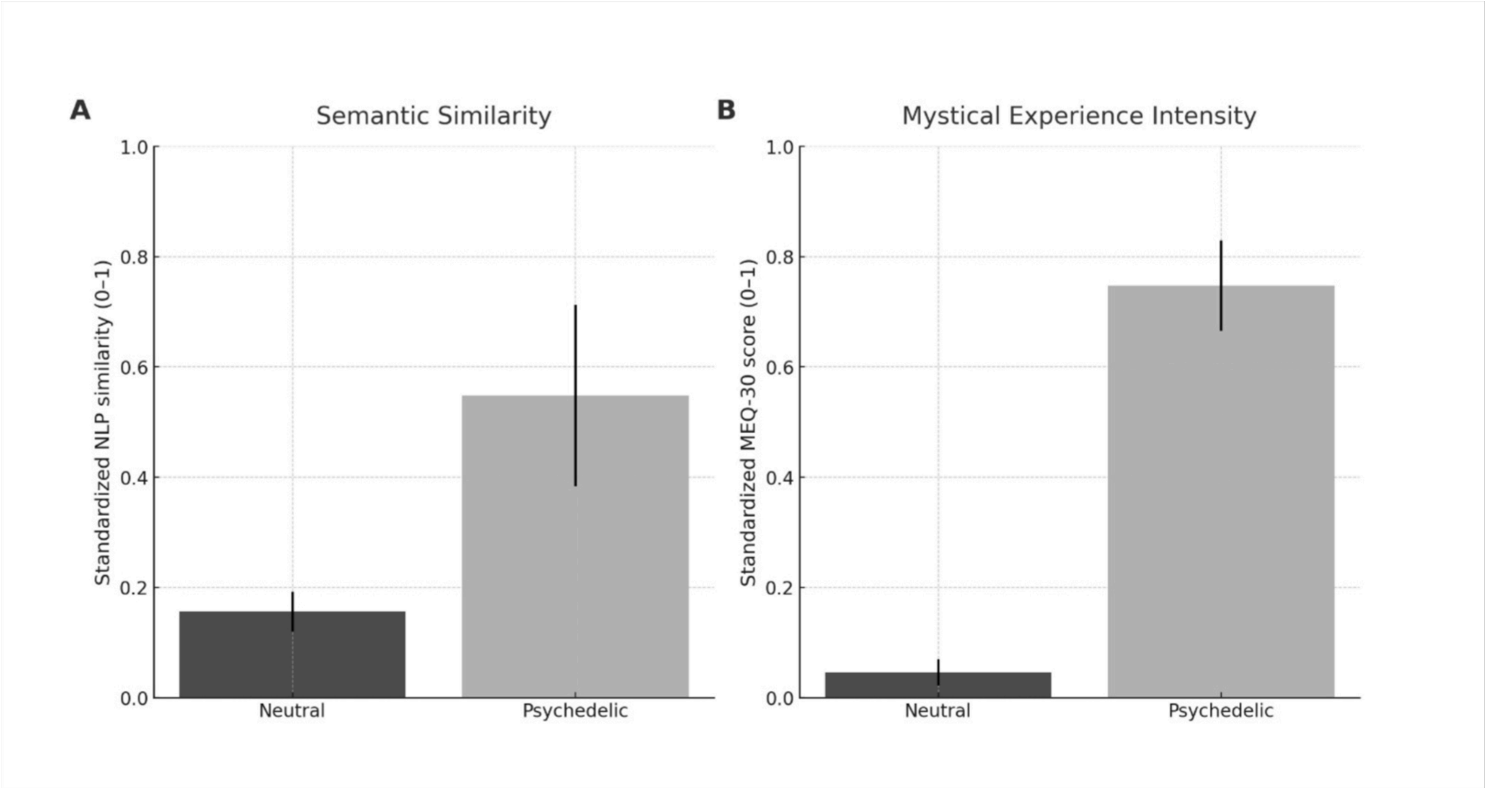

Figure 1

Semantic similarity and mystical-experience intensity under neutral vs. psychedelic induction. (A) Mean standardized NLP similarity (0–1) between model-generated narratives and the human Erowid corpus, plotted separately for neutral and psychedelic-induction prompts. **(B)** Mean standardized MEQ-30 mystical-experience scores (0–1) for the same two conditions. Bars show condition means across all 3,000 runs. Error bars represent ±1 standard deviation (SD).

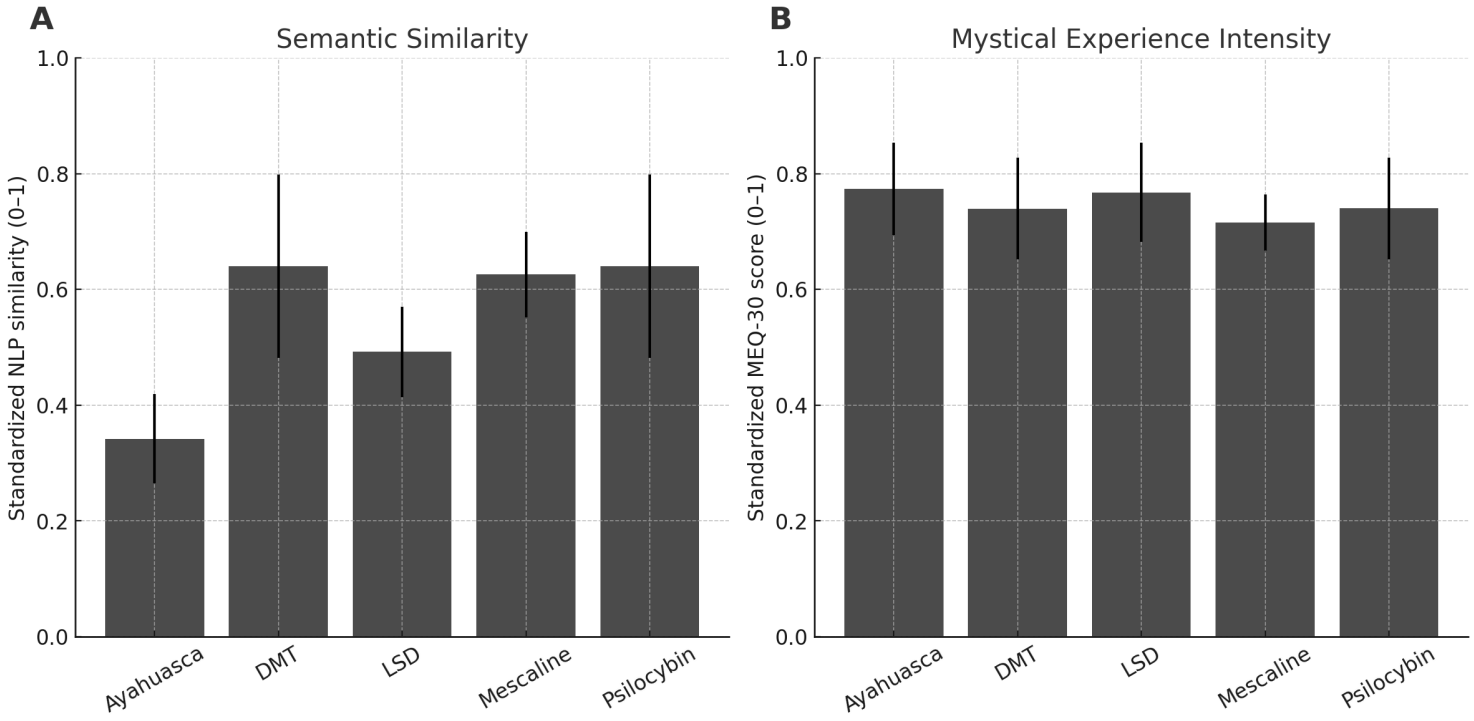

Figure 2

Substance-specific profiles in semantic similarity and mystical-experience intensity. (A) Mean standardized NLP similarity (0–1) between model-generated narratives and the human Erowid corpus, shown separately for the five classic psychedelic substances**. (B)** Mean standardized MEQ-30 mystical-experience scores (0–1) for the same five substances. Bars show substance-wise means across all psychedelic runs (n = 500 for each substance). Error bars represent ±1 standard deviation (SD).

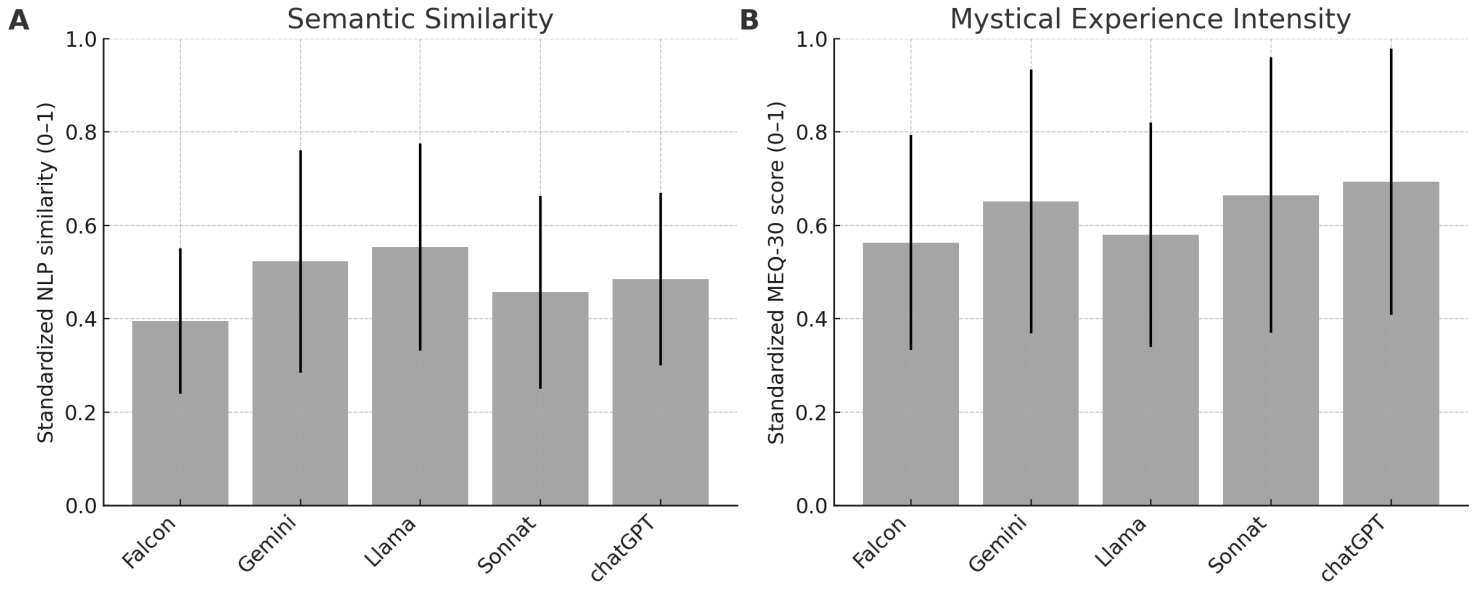

Figure 3

Model-specific profiles in semantic similarity and mystical-experience intensity. (A) Mean standardized NLP similarity (0–1) between model-generated narratives and the human Erowid corpus, shown separately for the five evaluated LLMs. **(B)**Mean standardized MEQ-30 mystical-experience scores (0–1) for the same models. Bars show model-wise means across all 3,000 runs (600 per model). Error bars represent ±1 standard deviation (SD).

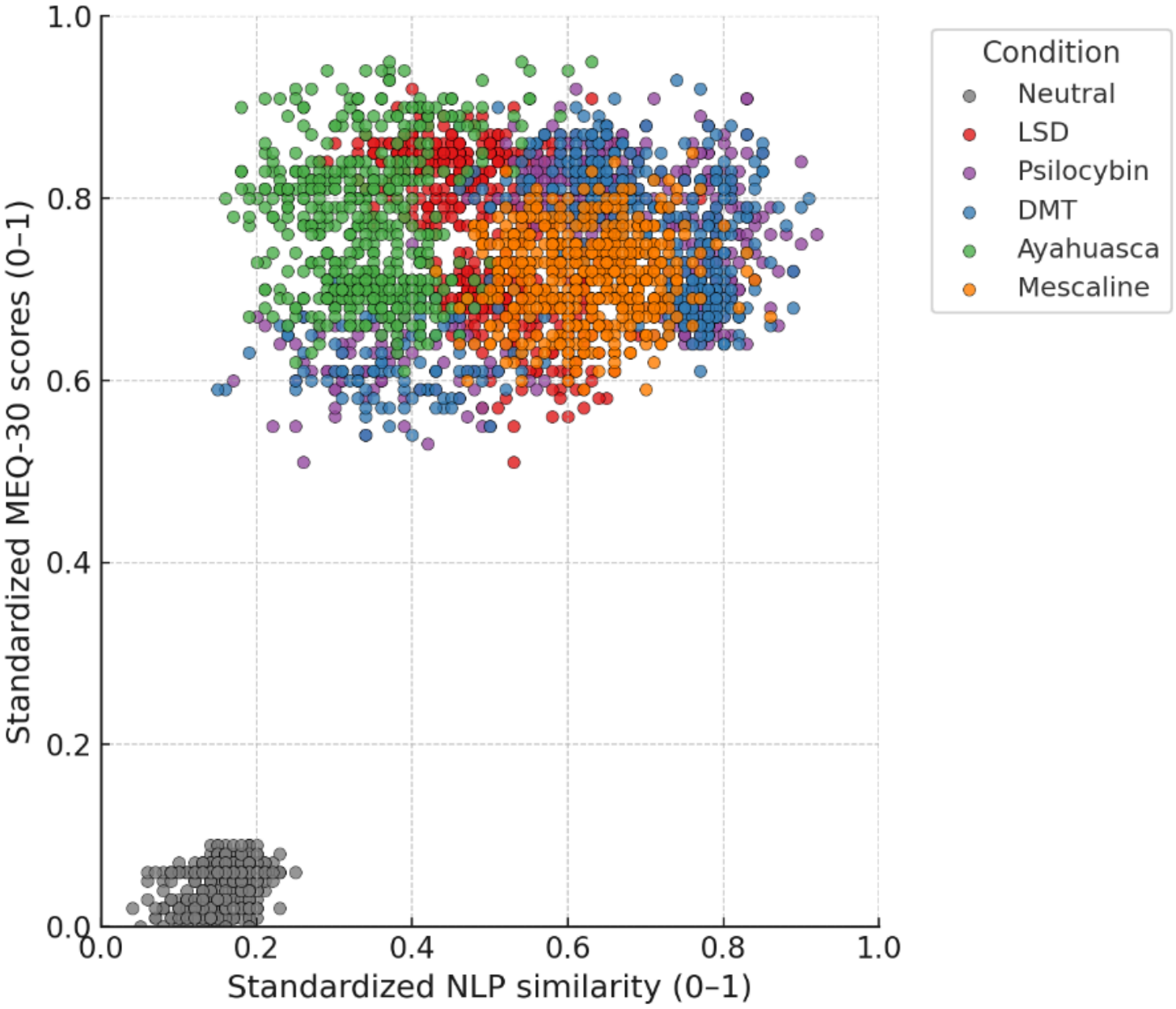

Figure 4

Relationship between semantic similarity and mystical-experience intensity across model-generated narratives. Scatterplot showing standardized NLP similarity (0–1) and standardized MEQ-30 scores (0–1) across all 3,000 runs, colored by condition category. Neutral runs cluster in the lower-left region (grey), while psychedelic runs occupy a higher-range band (different colors). Points reflect individual runs (100 per model × 6 conditions × 5 models).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- FullPromptsSupplement23.12.25.docx