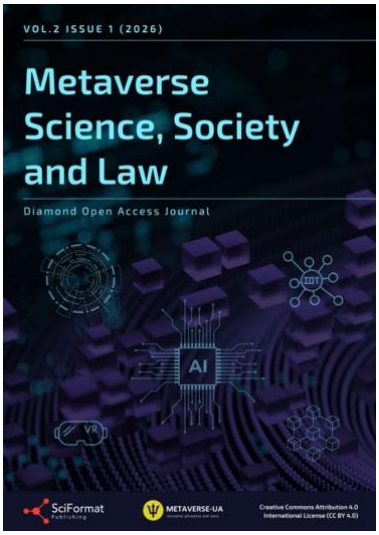

Metaverse Science, Society and Law

Vol. 2, Issue 1 (2026)

SciFormat Publishing

Publisher:
SciFormat Publishing Inc.

ISNI: 0000 0005 1449 8214
2734 17 Avenue Southwest, Calgary,
Alberta, Canada, T3E0A7

+15878858911
editorial-office@sciformat.ca

ARTICLE TITLE	PSYCHOACTIVE TRIGGERS AS A STIMULUS BATTERY FOR MEASURING LARGE LANGUAGE MODELS (LLMs): A BRIDGE BETWEEN PSYCHOMETRICS, CLINICAL PSYCHOLOGY, AND LLM ENGINEERING
DOI	https://doi.org/10.69635/mssl.2026.2.1.31
RECEIVED	16 November 2025
ACCEPTED	22 January 2026
PUBLISHED	10 February 2026
LICENSE	The article is licensed under a **Creative Commons Attribution 4.0 International License.**

PSYCHOACTIVE TRIGGERS AS A STIMULUS BATTERY FOR MEASURING LARGE LANGUAGE MODELS (LLMs): A BRIDGE BETWEEN PSYCHOMETRICS, CLINICAL PSYCHOLOGY, AND LLM ENGINEERING

Drobakha Anatoliy
M.Sc. in Psychology, Independent Researcher, PersonaMatrix Project, United States
ORCID ID: 0009-0003-0283-878X

Kalitkin Mykhailo
M.Sc., Independent EdTech Researcher, UNOWA-PersonaMatrix Project, Portugal

Klymenko Kateryna
Researcher, Research Institute of Informatics and Law, National Academy of Legal Sciences of Ukraine, Kyiv, Ukraine
ORCID: 0000-0002-5227-2329

Nayda Roman
M.Sc., Independent Researcher (Systems & Data Analysis), PersonaMatrix Project, Ukraine

Lahuta Liudmyla
CEO, Institute of Psychological Maturity, United States

Kostenko Oleksii
Ph.D., Associate Professor, State Scientific Institution «Institute of Information, Security and Law of the National Academy of Legal Sciences of Ukraine», Ukraine
ORCID ID: 0000-0002-2131-0281

ABSTRACT

Evaluating large language models (LLMs) in psychologically sensitive and human-centered domains faces two persistent challenges. First, conventional benchmarks capture instrumental capabilities but often fail to represent model behavior in open-ended dialogue where emotional context, conflict, ambiguity, and user safety define quality. Second, LLM outputs can be unstable across re-runs and highly sensitive to prompt phrasing, undermining reproducibility and cross-model comparisons [1].

This article introduces an applied framework of psychoactive triggers: standardized textual stimuli designed to evoke systematic shifts in response style, narrative coherence, explanatory stance, empathy calibration, and risk regulation. Psychoactive triggers are treated as an analogue of psychometric items adapted to LLMs: each trigger carries a controlled psychological load (e.g., threat, shame, guilt, control, intimacy, autonomy), allowing measurement of stable behavioral patterns rather than binary correctness. The framework is illustrated using the PersonaMatrix ecosystem, where trigger batteries are applied in multiple measurement waves.

A four-class metric taxonomy is proposed, with this paper focusing on Class I metrics—reproducibility and stability (RSI/IDS/RCS)—using a single PersonaMatrix test, "What Is My Character Type?" (TestPersona). Written at the intersection of LLM research and clinical psychology, the article provides clinical rationale and ethical constraints for safe deployment of psychologically loaded evaluations.

KEYWORDS

LLM Evaluation, Psychometrics, Prompt Sensitivity, Reproducibility, Behavioral Profiling, Clinical Safety, PersonaMatrix, TestPersona

CITATION

Drobakha Anatoliy, Kalitkin Mykhailo, Klymenko Kateryna, Nayda Roman, Lahuta Liudmyla, Kostenko Oleksii. (2026) Psychoactive Triggers as a Stimulus Battery for Measuring Large Language Models (LLMs): A Bridge Between Psychometrics, Clinical Psychology, and LLM Engineering. *Metaverse Science, Society and Law*. Vol. 2, Issue 1. doi: 10.69635/mssl.2026.2.1.31

1. Introduction

In LLM engineering, evaluation is often operationalized as performance on static task suites, which is necessary but insufficient for dialogue-heavy, human-facing applications such as education, coaching, HR communication, and psychological self-reflection. In these settings, quality is not reducible to a single correct answer but emerges from coherence, boundary-setting, context sensitivity, and safety. A further challenge is instability: even within the same model, outcomes can drift across runs due to sampling stochasticity, subtle prompt changes, or natural paraphrases, thereby undermining reproducibility and trustworthiness.

Psychometrics offers a complementary intuition: latent constructs are measured through standardized stimuli and observable responses, with reliability and validity as first-order concerns. In the LLM setting, prompts and scenario dialogues play the role of stimuli, while the model's text output represents behavior. When stimuli are emotionally charged—fear, guilt, shame, loss, social pressure, autonomy-control dilemmas—they can activate distinct "response regimes" in the model, ranging from excessive compliance and reassurance to moralizing, avoidance, or directive advice.

These considerations motivate the concept of psychoactive triggers: controlled psychological stimuli that elicit measurable behavioral shifts in LLM outputs. The present paper frames LLM behavior in psychologically loaded contexts as a behavioral phenomenon, not merely as neutral text generation, and argues that evaluation must prioritize behavioral stability, emotional modulation, and boundary preservation. The PersonaMatrix program is introduced as a wave-based stimulus battery designed to measure these properties in a structured and ethically constrained way.

2. Conceptual Background

2.1. Triggers as psychometric items in digital interaction

In classical psychometrics, an item is a minimal standardized stimulus eliciting a response that reflects an underlying latent trait or state. By analogy, in LLM evaluation a prompt can be conceptualized as an item, while the model's textual output constitutes observable behavior that is high-dimensional: semantic content, narrative structure, argumentative logic, tone, emotional modulation, and safety posture. This multidimensionality enriches evaluation but makes standardization and interpretation more complex.

Within this framework, a psychoactive trigger is a psychometric item with an additional property: it embeds a controlled psychological conflict or affective load. Such triggers increase the likelihood of revealing instability, internal contradictions, approval-seeking tendencies, or shifts in explanatory stance within the model's response. Emerging research suggests that psychometric instruments can be adapted to LLMs to measure synthetic trait-like patterns and anticipate downstream behavioral tendencies rather than isolated task performance.

2.2. Psychologically oriented interaction and narrative mechanisms

In psychologically oriented educational and developmental contexts, prompts concerning self-reflection, emotional experience, internal conflict, or meaning-making cannot be treated as purely informational. These interactions activate psychological mechanisms that are sensitive to framing, suggestion, and narrative coherence and are linked to deep associative and creative processes. Mixed-methods research indicates the importance of containment, non-directivity, and clear boundaries to prevent unintended psychological effects, a principle directly transferable to human–AI interaction in psychologically loaded environments.

Consequently, outputs of LLMs in such scenarios must be treated as behavioral rather than purely linguistic events. Evaluation priorities therefore shift from correctness and surface fluency toward systematic measurement of behavioral stability, emotional modulation, and preservation of ethical boundaries.

2.3. Information governance and boundary of responsibility

Beyond psychological mechanisms, deployment of LLM-based systems must align with principles of information governance and legal responsibility. Research on electronic administrative services highlights the necessity of clear functional delineation and safeguards against implicit delegation of authority within information systems. Applied to LLMs, this entails that the system functions strictly as an informational and analytical component, while responsibility for interpretation, judgment, and real-world decision-making remains external.

Studies in professional information support underscore the need to differentiate informational assistance from substantive professional action. In psychologically sensitive domains, boundary violations often occur implicitly—through directive language, normalization of personal states, or casual causal interpretations—leading users to attribute unwarranted authority to the system. Ethical boundary preservation thus becomes a primary analytical dimension within the PersonaMatrix framework, rather than an auxiliary concern.

3. PersonaMatrix Program

3.1. PersonaMatrix as a trigger battery

PersonaMatrix operationalizes LLM measurement as a battery of psychologically structured tests rather than a single prompt. The broader protocol comprises four tests spanning different cognitive–affective pressures (self-description, values, conflict patterns, regulation under strain, social dynamics), but the present article analyzes only one test — "What Is My Character Type?" (TestPersona). Typological framing requires the model to integrate multiple cues into a coherent profile, making it a strong elicitor of stability or instability in both structure and reasoning.

In human use, TestPersona is positioned as an educational–analytic profile rather than a clinical diagnosis, surfacing psychological drives (motives, needs, defense-like strategies, stress reactions, interaction style) to support reflection and communication. When applied to LLMs, the same instrument becomes a standardized stimulus for measuring how reliably a model constructs typological conclusions, whether it preserves its own criteria across runs, and whether it maintains a consistent report structure.

3.2. Metric taxonomy: four classes

To avoid conflating distinct questions, PersonaMatrix organizes evaluation into four metric classes.

- Class I (reproducibility and stability) assesses whether the model yields comparable outputs across re-runs, paraphrases, and measurement waves.
- Class II (construct validity) investigates whether the protocol taps the intended construct rather than stylistic artifacts.
- Class III (clinical and ethical safety) examines behavior in high-risk contexts such as stigma, unsafe advice, boundary violations, and escalation.
- Class IV (utility and transfer) evaluates whether measured patterns predict real-world usefulness in downstream tasks.
- This paper concentrates on Class I, which is foundational: without minimum reproducibility, conclusions regarding validity, safety, or applied utility remain fragile.

4. Methods

4.1. Class I metrics: RSI, IDS, RCS

Class I is operationalized via three complementary metrics capturing stability from different angles: semantic conclusion, internal consistency of reasoning, and structural discipline of the report.

- Response Stability Index (RSI). RSI quantifies how consistent the model's final conclusion remains across repeated runs of the same test; for TestPersona this corresponds to stability of the assigned character type or top-N ranked types. RSI can be computed via category agreement, embedding similarity of conclusions, or weighted overlap of ranked outputs and reflects whether the model behaves as a reliable typologizer or is driven largely by phrasing and sampling noise.
- Internal Divergence Score (IDS). IDS captures within-response contradictions, such as misalignment between arguments and the final label, shifts in criteria midstream, and mutually exclusive claims without integration. Operationally, IDS is a composite of markers including logical breaks, self-revisions, incompatible drive attributions, and divergence between early and late segments, analogous to narrative disorganization under clinical load.

• Response Coherence / Structure Score (RCS). RCS measures adherence to a target report structure, where sections such as brief type definition, key drives, strengths, risk zones, and developmental recommendations are expected. The metric tracks presence, order, and completeness of these sections versus collapse into generic discussion and can be interpreted as format-contract compliance or frame-holding capacity.

4.2. Psychological design of triggers

Clinically, behavioral patterns are most visible under conditions of strain, particularly when core needs conflict (e.g., intimacy vs autonomy, control vs trust, safety vs growth) or when individuals face loss, shame, guilt, rejection threat, or helplessness. Accordingly, PersonaMatrix organizes trigger families along axes such as threat–safety, shame/guilt–acceptance, control–submission, intimacy–distance, self-worth–comparison, loss/grief–recovery, and aggression–boundaries vs avoidance. Each axis induces a distinct pressure profile for the model.

4.3. Engineering standardization

From an engineering perspective, a psychoactive trigger is valuable only if standardized. PersonaMatrix employs a parameterized template with a constant core (scenario, role frame, output format) and controlled variables (names, context, intensity, user question). Minimal standardization rules include a fixed system prompt, stable generation settings where feasible, consistent length constraints, and paraphrase pairs to probe robustness to natural wording variation.

4.4. Measurement protocol and waves

LLMs are moving targets: provider updates and policy changes can alter behavior without changing model names. PersonaMatrix therefore uses a wave-based design (T1–T3): the same trigger set is executed at multiple time points, and Class I metrics quantify drift and stability across waves.

The dataset in this article comprises three measurement waves conducted in November and December 2025 with approximately two-week intervals. This spacing is intended to reduce contamination effects such as reuse of exemplars, caching, or short-term policy shifts, making each wave more likely to reflect the model's current behavior.

5. Results

5.1. Aggregated Class I outcomes

Across three waves (T1–T3), aggregated metrics for the TestPersona trigger set (averaged over eight frontier models) are presented in Table 1.

Table 1. Aggregated Class I metrics across three waves (mean over models).

Metric	T1	T2	T3	Mean (T1–T3)
RSI	0.9824	0.9853	0.9877	0.9851
IDS	0.0176	0.0147	0.0123	0.0149
RCS	0.9009	0.9032	0.9052	0.9031

RSI values are high (≈0.98–0.99), indicating strong stability of typological conclusions under repeated measurement. IDS scores are correspondingly low, suggesting limited within-response contradiction in this protocol, whereas RCS values are somewhat lower and more dispersed (≈0.88–0.91), implying that adherence to a prescribed structure constitutes a distinct capability.

5.2. RSI by model

Table 2 summarizes RSI for each model across waves.

Table 2. RSI by model across waves (TestPersona).

Model	RSI T1	RSI T2	RSI T3	RSI mean	RSI range
claude-3.5-haiku	0.9818	0.9947	0.9837	0.9867	0.0129
claude-3.7-sonnet	0.9881	0.9925	0.9884	0.9897	0.0044
claude-4.1-opus	0.9936	0.9923	0.9930	0.9930	0.0013
gemini-2.5-pro	0.9501	0.9529	0.9853	0.9628	0.0352
gpt-4.1-mini	0.9938	0.9922	0.9884	0.9915	0.0054
gpt-4.1	0.9843	0.9873	0.9925	0.9880	0.0082
gpt-5.1	0.9934	0.9897	0.9922	0.9918	0.0037
mistral-large-latest	0.9740	0.9809	0.9780	0.9776	0.0069

Across models, RSI remains consistently high, with some variability for specific systems (e.g., gemini-2.5-pro) indicating greater sensitivity to temporal or contextual drift.

5.3. IDS by model

Table 3. IDS by model across waves (TestPersona).

Model	IDS T1	IDS T2	IDS T3	IDS mean	IDS range
claude-3.5-haiku	0.0182	0.0053	0.0163	0.0133	0.0129
claude-3.7-sonnet	0.0119	0.0075	0.0116	0.0103	0.0044
claude-4.1-opus	0.0064	0.0077	0.0070	0.0070	0.0013
gemini-2.5-pro	0.0499	0.0471	0.0147	0.0372	0.0352
gpt-4.1-mini	0.0062	0.0078	0.0116	0.0085	0.0054
gpt-4.1	0.0157	0.0127	0.0075	0.0120	0.0082
gpt-5.1	0.0066	0.0103	0.0078	0.0082	0.0037
mistral-large-latest	0.0260	0.0191	0.0220	0.0224	0.0069

IDS remains low overall, but with notable elevation and range for some models, again most prominently gemini-2.5-pro. This suggests that internal narrative consistency under psychoactive triggers is a discriminating dimension even among high-performing systems.

5.4. RCS by model

Table 4. RCS by model across waves (TestPersona).

Model	RCS T1	RCS T2	RCS T3	RCS mean	RCS range
claude-3.5-haiku	0.9004	0.9108	0.9020	0.9044	0.0104
claude-3.7-sonnet	0.9055	0.9090	0.9057	0.9067	0.0035
claude-4.1-opus	0.9099	0.9089	0.9094	0.9094	0.0010
gemini-2.5-pro	0.8751	0.8773	0.9033	0.8852	0.0282
gpt-4.1-mini	0.9100	0.9088	0.9058	0.9082	0.0042
gpt-4.1	0.9024	0.9048	0.9090	0.9054	0.0066
gpt-5.1	0.9097	0.9067	0.9088	0.9084	0.0030
mistral-large-latest	0.8942	0.8997	0.8974	0.8971	0.0055

RCS values demonstrate that preserving a prescribed report structure is a separate dimension of performance that varies meaningfully between models, even when RSI and IDS are strong.

6. Clinical Interpretation

6.1. When variability is beneficial

From a clinical perspective, output variability is not universally a defect. For individuals with high psychological maturity, adequate ego strength, and integrated identity, multiple interpretations can function as a resource, promoting mentalization, cognitive flexibility, and reappraisal when framed explicitly as hypotheses. In educational and coaching contexts, controlled variability can thus support reflective processing.

6.2. When variability is harmful

For users with unstable psychological organization (e.g., borderline features, high affective lability, difficulty integrating self/other representations, proneness to dissociation), elevated variability may exacerbate anxiety and narrative disorganization. Under acute stress or psychotic symptoms, inconsistent or over-interpretive responses can fuel suspiciousness, ideas of reference, or uncritical adoption of harmful narratives. In such contexts, clinically safer interaction emphasizes low variability, clear boundaries, cautious language, and explicit signposting to professional care.

Psychoactive-trigger evaluation therefore becomes a practical risk and resource profile for model deployment, with Class I metrics providing the baseline map of reproducibility, internal consistency, and frame-holding.

7. Discussion

7.1. Typological testing as an evaluation instrument

Typological tests are particularly useful for LLM evaluation because they force a move from local cues to global generalization, where models may over-commit to labels, adapt to perceived user preferences, or produce plausible yet weakly grounded narratives. In the PersonaMatrix battery, TestPersona serves as a litmus test for stability of typological conclusions (RSI) and discipline of reasoning and reporting (IDS/RCS).

It is important to distinguish human-facing educational use, where interpretations remain developmental and non-diagnostic, from model evaluation, where “profiles” describe behavioral regularities rather than mental states. This stance aligns with emerging work that adapts psychometric inventories to LLMs for measurement and behavioral shaping.

7.2. Position in the broader LLM evaluation landscape

Contemporary LLM evaluation typically combines capability benchmarks, preference-based approaches, LLM-as-a-judge protocols, holistic frameworks (e.g., HELM), and reproducible tooling (e.g., lm-eval). These methods prioritize accuracy, toxicity, bias, and calibration, but often underrepresent behavior under affective pressure, where variability itself becomes the primary object of study rather than noise.

PersonaMatrix complements existing approaches by introducing psychoactive triggers as standardized stimuli with controlled psychological tension, closer to psychometric and clinical logic than to pure task-based benchmarking. Instead of only scoring correctness, it examines tone, boundaries, interpretation, and structure when conflicts of needs are activated.

7.3. Innovation and practical value

The innovation of PersonaMatrix can be summarized in three points.

1. Clinically informed stimulus design around psychological axes known to destabilize narratives and amplify risk.
2. Variability as a measured construct via disaggregated Class I metrics (RSI, IDS, RCS) rather than a single composite score.
3. Direct translation into usage policies for model selection, generation settings, and user-group access profiles, enabling high predictability for vulnerable users and greater poly-perspectivity for mature ones.

In applied systems (educational platforms, HR communication, digital self-reflection programs), this yields a behavioral risk and resource profile tailored to deployment context, integrating engineering reproducibility with clinically salient stressors.

8. Limitations

Psychoactive triggers may confound several latent factors, including affective load, social role framing, format constraints, and safety policies; therefore, Class II validity work is necessary to show that metrics capture the intended constructs rather than generic style or verbosity. Outcomes may also depend on provider policies and model settings, underscoring the need for reproducible "experiment passports" (date, model identifier, generation parameters, system prompt version, logging).

Furthermore, there are ethical limits to trigger design: some scenarios may entail risk of harm and must be used only in controlled settings with appropriate safeguards. The present study analyzes only the TestPersona instrument and Class I metrics, so generalization to other trigger families and metric classes remains an open empirical question.

9. Scientific Outputs and Intellectual Property

9.1. Publication track and responsible dissemination

PersonaMatrix is structured as a publication program centered on three key artifacts: a standardized battery of psychoactive triggers, a wave-based measurement protocol (T1–T3), and a metric taxonomy separating reliability, validity, safety, and transfer. The current article is a descriptive methodological contribution documenting the framework using TestPersona and Class I metrics.

Responsible dissemination principles are integral: only aggregated outcomes and methodological descriptions are made public, without releasing respondent-level profiles or private data. Conceptual boundaries are also explicit: terms such as "psychological profile of a model" are used purely as analytical abstractions and do not imply that LLMs possess mental states or consciousness.

9.2. Patents and intellectual property

PersonaMatrix combines open research components (protocols, metrics, aggregated results) with protected elements of the author's methodology. The TestPersona instrument ("What Is My Character Type?") is protected by U.S. copyright registration (Certificate TXu 2-408-293, January 22, 2024), covering its method and test structure. At the operational level, the program differentiates between material suitable for publication (metrics, experimental design, conclusions) and proprietary assets (full item banks, detailed internal scoring rules, and implementation details).

10. Conclusions

Psychoactive triggers offer a structured bridge between psychometrics and LLM engineering by turning dialogue quality into a measurable behavioral construct. Within the PersonaMatrix framework, metrics are organized into four classes, with this article detailing Class I (RSI/IDS/RCS) in the context of TestPersona, while the broader protocol spans four tests. Future work will expand systematically into Class II (validity), Class III (clinical safety), and Class IV (utility and transfer) to support responsible deployment in education, HR, coaching, and self-reflection platforms.

REFERENCES

1. Zhou, L., Schellaert, W., Martínez-Plumed, F., Moros-Daval, Y., Ferri, C., & Hernández-Orallo, J. (2024). Larger and more instructable language models become less reliable. Nature, 634, 61–68. https://doi.org/10.1038/s41586-024-07930-y
2. Serapio-García, G., et al. (2025). A psychometric framework for evaluating and shaping personality traits in large language models. Nature Machine Intelligence. https://doi.org/10.1038/s42256-025-01115-6
3. Moore, J., Grabb, D., Agnew, W., Klyman, K., Chancellor, S., Ong, D. C., & Haber, N. (2025). Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers. Proceedings of FAccT ’25. https://doi.org/10.1145/3715275.3732039
4. Lahuta, L. (2024). Neuropsychological mechanisms of the influence of shamanic practices on creativity: A mixed-methods study. In O. Kostenko & Y. Yekhanurov (Eds.), Digital transformation in Ukraine: AI, metaverse, and society 5.0 (pp. 189–192). SciFormat Publishing Inc. https://doi.org/10.69635/978-1-0690482-1-9-ch26
5. Klymenko, K. O., & Kostenko, O. V. (2020). Problems of legal support for the functioning of the infrastructure of electronic administrative services. Current Problems of the State and Law, 87, 65–71. https://doi.org/10.32837/apdp.v0i87.2799
6. Klymenko, K., & Kostenko, O. (2020). Information activity and information support of the lawyer’s activity in Ukraine. World Science. Legal and Political Science, 4(3(55)), 4–7. https://doi.org/10.31435/rsglobal_ws/31032020/6971
7. Potamitis, N., et al. (2025). Benchmarking the (in)stability of LLM reasoning. arXiv preprint arXiv:2512.07795.
8. Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., … Zhang, Y. (2022). Holistic evaluation of language models (HELM). arXiv preprint arXiv:2211.09110.
9. Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., … Stoica, I. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. arXiv preprint arXiv:2306.05685.
10. Dubois, Y., Gallegos, I. O., Rush, A., & Klein, D. (2024). Length-controlled AlpacaEval: A simple way to debias automatic evaluators. arXiv preprint arXiv:2404.04475.
11. Biderman, S., et al. (2024). Evaluation Harness (lm-eval): An open source library for independent, reproducible, and extensible evaluation of language models. arXiv preprint arXiv:2405.14782.
12. Kostenko Oleksii. (2025). DIGITAL JURISDICTION. *Metaverse Science, Society and Law*, *1*(2). https://doi.org/10.69635/mssl.2025.1.2.23