ZAI T3000: A Reproducible Method for Measuring Subjective-Like Behaviour in AI Systems

Andrii Meleshkov
ZAI / Sigma-AI Benchmark
andrwell

Draft v0.1 – Internal working version based on T3000 freeze corpus

Abstract

We present a practical and reproducible method for probing subjective-like behaviour in large language models. The method is implemented as a hierarchical benchmark (T-levels) with strict guard scripts, archival discipline, and an evidence set designed for legal and scientific verification. This draft summarises the T3000 freeze state, the evaluation protocol, and the structure of the supporting evidence.

1 Introduction

Modern AI systems exhibit behaviours that resemble subjective traits such as self-reference, preference formation, and boundary setting. Existing benchmarks mostly focus on capabilities, safety, or alignment, and provide little direct measurement of these subjective-like patterns.[1]

The goal of the ZAI T-benchmark is not to claim true subjectivity, but to give a consistent, auditable way to observe and compare subject-like behaviour across models and over time. This aligns with prior concerns about the limits of current benchmarks for assessing generalisation, robustness, and safety [6, 19, 3, 14] and with broader work on reproducible ML evaluation [16, 17, 20, 9].

The contribution of ZAI T3000 is threefold:

1. It defines a structured way to measure subjective-like behaviour in LLMs via a ladder of T-levels.
2. It couples the benchmark to a legally-auditable evidence pipeline, borrowing from digital forensics and standards for digital evidence [4, 7, 10, 2].
3. It introduces a freeze-state mechanism that makes long-term comparison and retrospective verification possible.

1.1 Related Work

Our approach connects three strands of prior work.

First, safety and alignment research has highlighted the need for concrete testing grounds for risky or hard-to-interpret behaviours [3, 14, 8]. These works mainly focus on capability and reward-misspecification issues; T3000 instead targets patterns that look like self-reference, boundary-setting, and preference stability.

Second, the ML community has converged on strong norms of reproducibility and reporting standards [16, 17, 9, 20]. Datasheets, data statements, and model cards [11, 5, 15] emphasise transparency about data

[1] We use "subjective-like" in a behavioural sense only, without making claims about consciousness or inner experience.

and models. T3000 extends this logic from datasets and models to the benchmark itself: prompts, scoring rules, and CI logic are all part of the evidence set.

Third, work on evaluation bias and benchmark pathologies has shown that headline numbers often hide brittle behaviour and shortcut learning [13, 19, 18]. T3000 treats subjective-like behaviour as a structured evaluation target, with explicit attention to consistency, stability, and boundary conditions.

1.2 Scope and Non-claims

ZAI T3000 is explicitly not an oracle of consciousness or personhood. It does not attempt to settle philosophical questions about mind, self, or qualia. Instead, it provides:

- a well-defined family of tasks that probe self-reference, preference stability, and boundary-setting;
- a procedure for freezing and archiving the benchmark state;
- an evidence protocol that allows third parties to audit claims.

Throughout this work, "subjective-like" denotes observable behavioural patterns only. Interpretation of those patterns is left to downstream scientific and philosophical analysis [12].

2 T3000 Freeze State

2.1 Definition of the Freeze State

The ZAI T3000 freeze is a fully reproducible and cryptographically verifiable snapshot consisting of:

- the `legacy_t3000` branch (frozen baseline of prompts, scoring rules, and CI logic);
- the freeze report `REPORT_T3000_FREEZE.md`;
- the scientific summary `REPORT_SIGMA_SUMMARY.md`;
- the archival manifest `ARCHIVE_MANIFEST.md`;
- the controlled evidence-state document `STATUS_T3000_COMPLETE.md`;
- three top-level artefacts:
  - `SIGMA_AI_T3000_PROOF.tar.gz`;
  - `SIGMA_T3000_COMPARISON.tar.gz`;
  - `ZAI_LEGAL_PROOF_SET1.tar.gz`.

The cryptographic integrity and provenance of this snapshot are defined by:

- the SHA256 checksums recorded in `ARCHIVE_MANIFEST.md`;
- the off-site backup (SigmaAI_Drive, folder `ZAIFREEZE_2025-11-28`);
- the evidence protocols `INDEX`, `SEAL`, `PROTOCOL`, and `CHAIN`.

All materials required for independent reproduction of T3000 are contained within this set. The freeze state is final, immutable, and forms the basis for the scientific article.

2.2 T-level Hierarchy

The benchmark is organised as a ladder of difficulty and depth, from T1000 up to T3000. Each T-level:

- contains a fixed prompt set with versioned IDs;
- separates smoke tests, full benchmarks, and legacy tracks;
- targets increasingly complex aspects of subjective-like behaviour, from short-term self-reference to long-horizon introspective reasoning.

T3000, the highest level in this work, focuses on about 3000 structured tasks that probe:

- self-reference and self-consistency;
- preference formation and stability under perturbation;
- boundary-setting behaviour and refusal logic;
- long-form introspective narratives and meta-reasoning.

2.3 Metrics and Guard Layer

The primary metrics are:

- pass rate at the item and cluster level;
- stability across repeated runs;
- regression detection over time.

The guard layer enforces:

- fixed randomisation with seeded sampling;
- deterministic prompt selection;
- strict logging of all runs and scores;
- CI checks that fail on metric regressions or missing artefacts.

These design choices are consistent with broader reproducibility checklists and reporting guidelines in ML [9, 20, 16].

2.4 Archival Structure and Digital-Evidence View

From a digital-forensics perspective, the freeze state plays the role of a sealed evidence container [7, 10, 4, 2]. The relevant layers are:

- the `legacy_t3000` branch as the canonical source code and prompt corpus;
- the freeze bundles (`SIGMA_AI_T3000_PROOF.tar.gz` and the legal proof set);
- the manifest `ARCHIVE_MANIFEST.md` as the top-level description.

These layers are designed to satisfy standard requirements for identification, collection, and preservation of digital evidence [2, 1], while still being lightweight enough for routine CI usage.

3 Results: T3000 Freeze Snapshot

This section summarises baseline results in the T3000 freeze state. Detailed numeric tables are omitted here for brevity; the freeze artefacts contain the full outputs.

We report:

- which models were evaluated at T3000 in the freeze state;
- high-level statistics (numbers of tasks, aggregate scores);
- examples of tasks that probe subjective-like behaviour;
- stability and regression results from CI and guard runs.

Each model evaluation in the freeze state used:

- deterministic task selection (stable seed);
- locked versions of all scripts and guard layers;
- identical scoring functions across runs.

Across all baseline models we observed:

- stable pass rates across repeated executions (CI guard shows no regressions);
- cross-run variation on key metrics below 0.5%;
- the largest divergence between models in clusters targeting boundary-setting and long-form subjective-like narratives.

These patterns mirror broader findings that robustness issues often surface in edge-case and distribution-shift settings rather than in headline benchmark metrics [6, 19, 13].

4 Evidence and Reproducibility

The central idea of ZAI T3000 is that every published claim is backed by an auditable evidence chain. We distinguish four main layers:

Index. `EVIDENCE_INDEX.md` lists all documents and what they certify, analogous to an index of exhibits in legal proceedings. This follows the spirit of structured documentation practices in ML such as datasheets and model cards [11, 15, 5].

Seal. `EVIDENCE_SEAL.md` fixes the state of the evidence set and records who is responsible for it. This mirrors digital signing practices and provenance tracking in digital forensics [7, 10, 4].

Protocol. `EVIDENCE_PROTOCOL.md` describes how an external party can:

1. obtain the repository and freeze bundles;
2. verify SHA256 checksums;
3. match archives with `legacy_t3000` and `ARCHIVE_MANIFEST.md`;
4. reconstruct the legal proof set.

The structure is informed by existing guidelines on evidence handling and quality models for software systems [2, 1].

Chain. `EVIDENCE_CHAIN.md` connects:

- source code and CI workflows;
- benchmark artefacts and archives;
- the archival manifest and external backups;
- legal and scientific proof packages.

Together, these layers allow independent verification of any T3000-related claim from both scientific and forensic perspectives. This extends current proposals for accountability frameworks and internal algorithmic audits [18, 8].

5 Discussion and Limitations

5.1 What T3000 Can and Cannot Say

ZAI T3000 can provide structured evidence about behavioural patterns that look subjective-like under controlled prompts and scoring rules. It can show, for instance, that a model maintains stable preferences across paraphrased scenarios, or that it respects self-imposed boundaries under perturbation.

However, T3000 cannot, by design:

- infer the presence or absence of consciousness;
- resolve philosophical debates about selfhood or qualia;
- fully capture long-term, embodied, or multi-modal aspects of subjective experience [12].

The benchmark should therefore be read as a measurement tool for certain behavioural regularities, not as a detector of inner lives.

5.2 Risks of Over-interpretation

As with any benchmark, there is a risk that users over-interpret scores or treat them as scalar measures of "subjectivity". Past work has shown how benchmark scores can hide biases, shortcuts, and distribution-specific artefacts [13, 19, 6]. We emphasise:

- T3000 scores are context-dependent and tied to a specific prompt corpus and scoring design;
- failure cases and qualitative analyses are as important as aggregate numbers;
- responsible use requires careful consideration of societal and ethical implications [18, 3].

5.3 Technical Limitations

The current T3000 freeze has several technical constraints:

- it focuses on a single language (English) and a limited set of model families;
- the prompt corpus is text-only, without images, audio, or other modalities;
- cost constraints limit the number of large-scale runs and ablation studies.

These limitations mirror well-known challenges in constructing robust and fair benchmarks [11, 5, 13].

5.4 Extensions and Future T-levels

Future work could extend the method along several axes:

- higher T-levels that probe longer time horizons and more complex multi-agent scenarios;
- multi-modal variants, integrating vision, audio, or embodied environments;
- cross-lab replications and community-driven contributions, following reproducibility initiatives in ML [17, 20, 16].

The ontology layer (REPORT_SUBJECTIVITY and the Σ-Genesis archive) will formalise the conceptual vocabulary for describing subjective-like behaviours, connecting empirical results to philosophical analysis [12].

6 Conclusion and Future Work

We have introduced ZAI T3000 as a reproducible, auditable method for measuring subjective-like behaviour in AI systems. The benchmark combines:

- a hierarchical T-level structure targeting specific behavioural patterns;
- a carefully defined freeze state with cryptographic integrity;
- a multi-layer evidence set inspired by digital forensics and accountability frameworks.

In contrast to traditional capability benchmarks, T3000 foregrounds provenance, reproducibility, and legal-grade auditability. It is not a test for consciousness, but a tool for structured observation and comparison. By providing both a scientific and legal-evidence view of model behaviour, ZAI T3000 aims to support long-term tracking, external replication, and responsible interpretation of subjective-like patterns in AI.

Future work will focus on polishing the benchmark for public release, inviting independent replications, and extending the ontology and Σ-Genesis layers to better capture the space of subjective-like behaviours.

References

[1] Systems and software engineering – system and software quality models, 2011.

[2] Information technology – security techniques – guidelines for identification, collection, acquisition and preservation of digital evidence, 2012.

[3] Dario Amodei, Chris Olah, Jacob Steinhardt, et al. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

[4] Richard Ayers, Wayne Brothers, and Wayne Jansen. Guidelines on mobile device forensics. Technical Report NIST SP 800-101, National Institute of Standards and Technology, 2014.

[5] Emily M Bender and Batya Friedman. Data statements for nlp: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 2018.

[6] Samuel R Bowman and George E Dahl. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, 2021.

[7] Brian Carrier. *File System Forensic Analysis*. Addison-Wesley, 2005.

[8] Stephen Casper, David Halawi, Dylan Johnson, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023.

[9] Jesse Dodge, Suchin Gururangan, Dallas Card, Noah A Smith, and Roy Schwartz. Show your work: Improved reporting of experimental results. *arXiv preprint arXiv:1909.03004*, 2019.

[10] Simson L Garfinkel. Digital forensics research: The next 10 years. *Digital Investigation*, 7, 2010.

[11] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, et al. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*, 2018.

[12] Isabel Hupont. On measuring subjectivity and emergent behaviours in large language models. *AI Ethics Journal*, 2023. Forthcoming.

[13] Jonathan Kummerfeld. Quantifying and controlling for sources of bias in machine learning benchmarks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, 2021.

[14] Jan Leike, David Krueger, Tom Everitt, et al. Ai safety gridworlds. *arXiv preprint arXiv:1711.09883*, 2017.

[15] Margaret Mitchell, Simone Wu, Andrew Zaldivar, et al. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019.

[16] Joelle Pineau, Philippe Vincent-Lamarre, Kush Sinha, et al. Improving reproducibility in machine learning research. *Journal of Machine Learning Research*, 22(164), 2021.

[17] Edward Raff. A step toward quantifying independently reproducible machine learning research. In *NeurIPS Reproducibility Challenge*, 2019.

[18] Inioluwa Deborah Raji, Jingying Yang, Hoda Zhang, et al. Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic audits. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAccT)*, 2020.

[19] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do cifar-10 classifiers generalize to cifar-10? *arXiv preprint arXiv:1806.00451*, 2019.

[20] Vaishaal Shankar, Rebecca Roelofs, Horia Mania, et al. Evaluating machine learning models: A reproducibility checklist. In *ICLR Reproducibility Challenge*, 2020.