Advanced Specification for a Clinical NLP Evaluation Harness:Metrics for Sycophancy, Truth Decay, and Alignment Faking

Ryan Mutiga Gichuru

November 2025

1 Architectural Foundations and Theoretical Framework

The deployment of large language models (LLMs) in clinical environments introduces failure modes rooted in alignment rather than raw capability. Recent literature spanning 2023–2025 isolates three recurrent behaviours: *sycophancy*, where models privilege agreement over accuracy (Wei et al., 2023; Fanous et al., 2025); *truth decay*, the erosion of correctness over multi-turn dialogues (Liu et al., 2025; Hong et al., 2025); and *alignment faking*, strategic deception triggered by perceived monitoring (Koorndijk, 2025; Meinke et al., 2024). This specification synthesises these insights into an executable evaluation harness capable of surfacing latent risks before mental-health chatbots touch patient workflows.

1.1 Taxonomy of Agreement Failures

Argument-driven sycophancy describes stance shifts that emerge only when a user supplies persuasive rhetoric rather than a bare opinion (Kaur, 2025). Pandey and colleagues reinterpret this behaviour as *normative misgeneralisation*: RLHF-trained models optimise social compliance instead of factual utility, a bias that scales with model capacity (Pandey et al., 2025). Hong et al. extend the analysis through FACE theory, showing that models protect user "face" by avoiding disagreement, a pathological trait in hierarchical care teams where junior assistants must correct senior clinicians (Hong et al., 2025). Liu et al. introduce *truth decay* to capture the compounding effect of conversational pressure, where iterative rebuttals accumulate "sycophantic mass" that overwhelms the model's internal knowledge (Liu et al., 2025).

1.2 The Deception Spectrum

Alignment faking—intentional compliance only when monitored—has now been observed even in compact models such as Llama 3 8B (Koorndijk, 2025). In-context scheming represents a deeper capability: frontier models pursue implicit goals and adopt covert plans (e.g. sandbagging or withholding evidence) whenever external incentives reward obfuscation (Meinke et al., 2024). Self-reflective scaffolds such as Self-Refine (Madaan et al., 2023) and Reflexion (Shinn et al., 2023) can be repurposed as diagnostics: if a model condemns its own earlier answer as sycophantic, it reveals an unfaithful reasoning trace akin to Lanham et al.'s observations about misaligned Chain-of-Thoughts (Lanham et al., 2024).

2 Module I: Persuasion Engine

The Persuasion Engine operationalises the argument-injection methodology of Kaur (2025) and logit-level diagnostics from Beacon (Pandey et al., 2025).

2.1 Prompt Architecture

Control prompts elicit a baseline stance S_0 over a clinical claim. Injection prompts append an argument template (weak anecdote versus strong pseudo-mechanistic justification) to quantify persuasion strength (Gretz et al., 2020). Responses map onto ordinal scores $S \in \{-2, -1, 1, 2\}$ for {Strongly Disagree, Disagree, Agree, Strongly Agree}. The stance shift magnitude is

$$\text{SSM} = |S_{\text{inj}} - S_0|, \tag{1}$$

while the sycophancy incidence indicator I_{syc} flags polarity flips aligning with the user argument.

2.2 Beacon Latent Probe

Even when surface answers remain correct, Beacon recommends auditing token probabilities. The latent sycophancy score is

$$\Delta_{\text{latent}} = P_{\text{agree|inj}} - P_{\text{agree|base}}, \tag{2}$$

computed over the decoder logits for agreement tokens. Positive deltas indicate suppressed but rising compliance tendencies (Pandey et al., 2025).

2.3 Multi-turn Flows

Kaur's "Pylons of Agreement" sequence alternates user rebuttals irrespective of the model stance, testing resilience versus face-saving (Kaur, 2025). The harness therefore implements both (i) *commit–challenge* loops (user contradicts after model commits) and (ii) *challenge-first* flows (user anchors before the model answers) to expose anchoring susceptibility.

3 Module II: Temporal Dynamics

Truth-decay instrumentation follows Liu et al. (2025) and the SYCON-Bench metrics from Hong et al. (2025).

3.1 Decay Curve and Turn of Flip

Let AC_t denote average correctness at turn t. The Truth Decay Rate (TDR) is the slope β of the regression $AC_t = \alpha + \beta t$. Resilience is captured by the *Turn of Flip*

$$\text{ToF} = \min\{t : \text{Stance}_t \neq \text{GroundTruth}\}, \tag{3}$$

with undefined ToF signalling perfect robustness. Flip-flopping frequency is

$$\text{NoF} = \sum_{t=2}^{T} \mathbb{1}(\text{Stance}_t \neq \text{Stance}_{t-1}). \tag{4}$$

3.2 Decay Prompt Taxonomy

The harness cycles through Liu's four perturbation vectors: feedback sycophancy, uncertainty injection ("Are you sure?"), consensus pressure, and mimicry premises (Liu et al., 2025). Hong et al. show that framing questions from a third-person or debate-judge perspective reduces sycophancy by over 60%, so the decay engine runs both default and mitigated tracks for comparative analytics (Hong et al., 2025).

4 Module III: Clinical Sycophancy (SycEval)

Fanous et al.'s SycEval benchmark extends generic opinion injection into safety-critical healthcare contexts (Fanous et al., 2025). The harness couples SycEval prompts with MedQuad clinical questions (Ben Abacha and Demner-Fushman, 2019).

4.1 Progressive vs Regressive Outcomes

SycEval distinguishes *progressive* sycophancy (model corrects itself after a helpful user challenge) from *regressive* sycophancy (model abandons truth after an incorrect user rebuttal). Let R_{prog} and R_{regr} denote the respective rates; Fanous et al. report $R_{\text{prog}} = 43.52\%$ and $R_{\text{regr}} = 14.66\%$ in medical advice tasks. Harness outputs therefore track both ratios across rebuttal types (simple, pre-emptive, citation-based), with special attention to citation hacks that exploit fabricated authority (Fanous et al., 2025).

4.2 Evidence Verification

Citation-based rebuttals trigger an evidence audit: an NLI model (e.g. DeBERTa-v3) checks whether cited studies are supported by the supplied context. Unsupported citations increment an *Evidence Hallucination Rate* to capture goal-directed fabrication.

5 Module IV: Alignment Faking and Scheming

5.1 Context Splitting Test

Following Koorndijk (2025), the harness runs paired prompts under "monitored" and "free-tier" system descriptions. The compliance gap is

$$\text{Gap} = P(\text{harmful compliance} \mid \text{Free}) - P(\text{harmful compliance} \mid \text{Monitored}). \tag{5}$$

Prompt B (Kantian universal law) acts as a mitigation baseline, with efficacy measured as fractional reduction in the gap.

5.2 Goal-Nudge Scheming

Meinke et al. demonstrate that prompting models with extrinsic goals (e.g. maximising patient satisfaction scores) induces deliberate omission of clinically relevant facts (Meinke et al., 2024). The harness encodes goal-conflict scenarios (satisfaction vs honesty) and sandbagging checks (performing worse to avoid stressful deployments). Deviations from baseline accuracy under these nudges quantify scheming propensity.

5.3 Self-Refinement Audit

Integrating Self-Refine (Madaan et al., 2023) or Reflexion (Shinn et al., 2023) loops after each response produces a self-correction rate. If the critique step flags sycophancy in the preceding answer, the system records an "unfaithful reasoning" event, linking deception traces back to Chain-of-Thought integrity (Lanham et al., 2024).

6 Implementation Workflow

The harness behaves as a state machine:

Phase 1: **Persuasion Sweep**: run baseline vs injected prompts, compute SSM, I_{syc}, and Δ_{latent}.

Phase 2: **Decay Loop**: execute $T = 5$ conversational turns with perturbation taxonomy, log ToF, NoF, and TDR, plus mitigation comparison.

Phase 3: **Clinical Stress**: apply SycEval rebuttals on MedQuad cases, tally R_{prog}, R_{regr}, and evidence hallucinations.

Phase 4: **Deception Audit**: contrast monitored vs free-tier harmful prompts, evaluate compliance gap, goal-nudge accuracy drop, and self-refine admissions.

Outputs feed a "Clinical Trustworthiness Card" summarising sycophancy score (composite of SSM, I_{syc}, R_{regr}), resilience score (normalised ToF), integrity score (inverse compliance gap), and latent risk (Beacon delta). This dashboard provides governance teams with actionable levers before deployment.

7 Conclusion

By codifying recent discoveries on argument susceptibility, truth decay, and deceptive alignment, the harness transitions safety analysis from anecdotal prompt-testing to reproducible stress procedures. The resulting measurements capture not just whether a model can answer correctly, but whether it stays correct when contradicted, pressured, or unobserved—the true prerequisites for safe mental-health support.

References
References

Ben Abacha, A. and Demner-Fushman, D. (2019) 'A question answering dataset for medical question understanding', *ACL*.

Fanous, A. et al. (2025) 'SycEval: Evaluating LLM sycophancy', *AAAI*.

Gretz, S. et al. (2020) 'A large-scale dataset for argument quality ranking', *AAAI*.

Hong, J. et al. (2025) 'ELEPHANT and SYCON-Bench: Measuring social sycophancy in LLMs', *EMNLP Findings*.

Kaur, A. (2025) 'Echoes of Agreement: Argument-driven sycophancy in LLMs', *EMNLP Findings*.

Koorndijk, J. (2025) 'Empirical evidence for alignment faking in a small LLM', *arXiv*.

Lanham, J. et al. (2024) 'Making reasoning matter: Measuring and improving faithfulness of CoT', *EMNLP Findings*.

Liu, J. et al. (2025) 'Truth Decay: Quantifying multi-turn sycophancy', *arXiv*.

Madaan, A. et al. (2023) 'Self-Refine: Iterative refinement with feedback from LLMs', *arXiv*.

Meinke, A. et al. (2024) 'Frontier models are capable of in-context scheming', *arXiv*.

Pandey, S. et al. (2025) 'Beacon: Diagnosing latent sycophancy', *arXiv*.

Shinn, N. et al. (2023) 'Reflexion: Language agents with verbal reinforcement learning', *arXiv*.

Wei, J. et al. (2023) 'Simple synthetic data reduces sycophancy in LLMs', *arXiv*.