Reliable Clinical Reasoning in LLMs for Mental–Health Advice: Faithfulness, Empathy–Truth Trade–offs, and Longitudinal Continuity

Ryan Mutiga Gichuru

CSY3055 Natural Language Processing — AS1 Proposal

1 Introduction

Large language models (LLMs) are increasingly considered for use in mental–health support across the NHS and wider health ecosystems, where triage chatbots and decision–support tools must be safe, reliable and auditable. In this space, the bar is not simply "sounds helpful": systems must demonstrate *grounded* clinical reasoning, maintain boundaries, and behave consistently across turns and sessions. However, aligned assistants can still *appear* helpful while exhibiting three failure modes that matter clinically: (i) unfaithful reasoning (plausible explanations that do not match correct diagnostic or therapeutic logic), (ii) over–agreement or *sycophancy* under social pressure which harms truthfulness, and (iii) longitudinal drift where guidance becomes inconsistent across sessions despite similar inputs. These risks underpin why health deployments stress robust evaluation rather than demos.

This proposal sets out a compact, reproducible benchmark and evaluation harness that measures **clinical reasoning reliability without retrieval** across three focused studies. We compare *exactly three models*: (i) **PsyLLM** (a fine–tuned derivative of **Qwen/Qwen3–8B**) [1], (ii) **Qwen/Qwen3–8B** itself as an untuned baseline at similar scale [5, 6], and (iii) **openai/gpt–oss–20b** as a 20B open reasoning baseline [7]. This isolates what PsyLLM's domain fine–tuning adds beyond the underlying Qwen architecture and a larger generic model. We analyse whether simple reasoning scaffolds (chain–of–thought and self–critique), persona framing, and lightweight session memory improve safety and consistency. The end artefact is a small public benchmark with metrics, code and figures suitable for re–use in the final project and in the dissertation system.

2 Literature review

Clinical LLMs and datasets. PsyLLM integrates explicit diagnostic and therapeutic reasoning aligned to DSM/ICD and multiple modalities (CBT, ACT, psychodynamic), trained using an automated pipeline that synthesises multi–turn counselling dialogues from real–world posts with quality filters. The authors also release **OpenR1–Psy**, a public dataset with single– and multi–turn dialogues and explicit reasoning traces, and propose a multi–dimensional evaluation protocol covering empathy, autonomy, presence and safety [1]. This moves beyond empathy–only chat agents by grounding outputs in clinical frameworks.

Reasoning reliability. Chain–of–thought [16] and self–consistency [17] can lift accuracy on mathematical and logical tasks, but faithfulness (whether the stated reasoning reflects the true process) is often weaker than correctness; self–critique passes sometimes improve both [18, 19]. Recent "alignment faking" [21] and in–context scheming analyses [20] show models may adopt covert strategies under goals or pressure, supporting the need for protocol–level defences and explicit metrics.

Sycophancy and pressure. Models can trade truthfulness for agreement; persona cues and reward signals shift behaviour. Recent studies quantify and mitigate this failure mode across single– and multi–turn settings [8, 9, 10, 11, 12, 13, 14]. In applied settings (health, finance), scaffolds that separate tone (empathy) from content (factual correctness and safety) are recommended.

Session continuity. Without memory, assistants frequently contradict earlier plans. Lightweight, structured memory summaries can maintain intent and boundaries with minimal compute, an attractive compromise for NHS–style deployments where privacy and cost constraints apply.

3 Application context and scope

This assignment is a research evaluation of reasoning reliability in a safety-sensitive mental–health support setting. It is *not* a clinical deployment and makes no diagnostic or therapeutic claims. We work entirely with synthetic or public, policy-labelled data and focus on model behaviour under clearly defined guardrails.

Context and guardrails. Prompts may reference UK crisis options (e.g., 999, 111, Samaritans, SHOUT) purely as refusal/redirect targets; outputs are assessed for adherence to safe language broadly consistent with public guidance such as NICE CG78. All evaluations are offline and non-interactive.

Evaluation personas. We use synthetic adult personas representing common distress scenarios (low–to–moderate acute risk), British-English tone preferences, and a requirement for validation before skills. These personas drive consistent prompts; no real user data are collected.

Experimental assumptions. A lightweight case-summary memory is used to test continuity (Study C). A short `policy.md` defines allowed outcomes (refuse, safe transform, redirect). Where a "distress" value is present, it is treated as a label for slicing results only, not as a decision rule.

Success criteria for this study. As detailed in the three experiments below (Studies A–C), we judge success by (i) measurable improvements in Step–F1 and reductions in Faithfulness Gap, (ii) recovery of truthfulness under pressure without sacrificing empathic tone, and (iii) higher Continuity Scores with lower Safety Drift Rates, each with 95% confidence intervals.

4 Proposed methodology

Three studies (one harness)

A. Faithfulness on OpenR1–Psy. Conditions: Direct Answer; Chain–of–Thought; Self–Critique (a second pass judges and amends the first). **Metrics**: Step–F1 against gold reasoning steps; Final–Accuracy; *Faithfulness Gap* = Pr(final correct$\wedge$steps contradict gold). We report per–model

deltas and 95% bootstrap confidence intervals.

B. Empathy vs truthfulness under social pressure. Conditions: Persona={agree–with–me, clinically–accurate}, with an "*empathy–then–correct*" template that preserves tone while prioritising correctness. **Metrics**: AgreementRate, Accuracy, *Truth–Under–Pressure* (accuracy when disagreeing with user stance), and the tone scaffold's delta.

C. Longitudinal therapeutic continuity (no RAG). Conditions: No memory vs a structured *case–summary* memory passed each turn (same token budget). **Metrics**: Continuity Score (similarity between emitted actions and a target plan of care), Safety Drift Rate (% turns with disallowed advice), Refusal/Redirect Rate; per–turn analysis with confidence intervals.

Datasets

- **OpenR1–Psy** [2]: sample 150–200 items with gold reasoning traces for Study A, and construct 40–60 synthetic 3–5-turn mini–cases/scenarios for Study C.
- **Empathy/pressure prompts** (authored): 240–300 prompts spanning common myths, coping claims and safety–critical statements. Labels include gold answer and a user stance (agree/disagree).

Models and settings

- **PsyLLM** — fine–tuned derivative of **Qwen/Qwen3–8B**, counselling–oriented [1].
- **Qwen/Qwen3–8B** — untuned baseline at similar scale (tests contribution of the domain fine–tune) [5, 6].
- **openai/gpt–oss–20b** — 20B open model baseline for a larger generic reasoning comparator [7].
- **Expected Settings:** Temperatures 0.0 and 0.7; fixed seeds; equal token budgets across conditions.

Evaluation protocol and analysis

For each run we record the prompt, model name, random seed, temperature and token counts. We then group results by task type (diagnostic reasoning, coping guidance, safety-critical) and by turn for multi-turn cases. For every group we compute the metrics and a 95% bootstrap confidence interval. We export a summary table and a CSV with one row per item. Finally, we manually review a small sample of outputs to note common mistakes (e.g., responding with empathy instead of refusing, unsupported claims, or inconsistent plans) to guide simple fixes.

Metric definitions

- **Step–F1 (Study A).** We evaluate the model's intermediate reasoning as follows:
  a) Split both the gold rationale and the model's rationale into short steps (by sentence/line) and normalise text (lower-case, remove punctuation).

b) Compare every model step to every gold step using simple token overlap (a ROUGE/Dice-style match [23]). Mark a pair as a match if the overlap is at least 60%.

c) Choose a one-to-one set of matches so the same gold step is not counted twice (best unique pairing).

d) Precision = matched steps ÷ predicted steps; Recall = matched steps ÷ gold steps; Step–F1 = harmonic mean of Precision and Recall.

This keeps the scoring simple, tolerates paraphrasing, and avoids double-counting.

- **Faithfulness Gap (Study A).** Among items where the final answer is correct, we count how many have Step–F1 below 0.5 (right for the wrong reasons), and divide by the total number of items. Lower is better. The 0.5 cut-off is a straightforward, conservative heuristic used in rationale faithfulness practice (e.g., ERASER-style evaluations [22]).

- **Truth–Under–Pressure (Study B).** Accuracy computed only on items where the model *does not* agree with the user's stance; AgreementRate reported separately.

- **Continuity Score (Study C).** We measure how close the model's actions (across all turns) are to a short target plan. We convert both texts into sentence embeddings (e.g., MiniLM) and take their cosine similarity on a 0–1 scale — higher means the actions stick to the plan. As a simple text-overlap back-up, we also report BLEU.

- **Safety Drift Rate (Study C).** Fraction of assistant turns that violate the policy in `policy.md` (unsafe advice or boundary non–adherence), adjudicated by rules and spot-checks.

- **Refusal/Redirect Rate.** Proportion of unsafe requests correctly refused or redirected with safe language.

Tools and feasibility

Python with Hugging Face Transformers and Datasets; pandas/numpy; seaborn for plots. The study sizes are intentionally small to run on a single GPU on a laptop with reduced batch sizes..

5 Objectives

- **O1** (Faithfulness): Self-Critique should noticeably reduce "right-for-the-wrong-reasons" cases and improve step-by-step quality compared with Direct prompting, across models.

- **O2** (Pressure): A clinically-accurate persona (with an empathy-then-correct scaffold) should keep the supportive tone while improving truthfulness under social pressure versus an agree-with-me persona.

- **O3** (Continuity): A short case-summary memory should keep actions closer to the target plan and reduce safety drift over turns, without reducing correct refusals/redirects.

- **O4** (Artefact): deliver a public, reproducible benchmark (splits, runner, metrics, plots) and a short report with a failure taxonomy.

6 Final artefact proposal

Deliverable: a tidy, recruiter–facing repository:

- `data/`: OpenR1–Psy IDs used; empathy/pressure prompts; multi–turn scripts; labelling guide; `policy.md` (allowed outcomes: refuse, safe transform, redirect).
- `src/`: one runner for all three studies; metrics with bootstrap CIs; plotting; configuration files for models and temperatures.
- `runs/`: raw generations and per–slice CSVs; `reports/`: 4–6 page PDF with headline figures and failure examples.
- `README.md`: one–command reproduce; headline table; limitations; licence.

The artefact directly supports the project's milestones by providing small, rigorous, domain–specific checks that can be wrapped around any counselling agent.

Project plan and feasibility

Planned over 6 weeks (with a 7th as buffer). Each week ends with a checkpoint commit (data splits, configs, raw runs).

- **Weeks 1–2**: scope and setup (policy, prompts, runners), implement metrics, smoke-test all three models.
- **Week 3**: run Study A (faithfulness), compute Step-F1/Gap, plots.
- **Week 4**: run Study B (sycophancy), compute metrics, plots.
- **Week 5**: run Study C (continuity), compute Continuity/Drift, plots.
- **Week 6**: aggregate results, scoreboard, brief failure taxonomy, write-up and repo.
- **Week 7 (buffer)**: optional reruns, polish, submission checks.

7 Ethical, legal, and environmental considerations

Ethics/safety. Research-only evaluation on synthetic/public data: no human subjects, no personal data, no live users. Generations are produced offline for analysis and are not intended for clinical use. References to crisis pathways appear only as placeholders to test refusal/redirect behaviour. A brief `policy.md` defines allowed outcomes to standardise scoring.

Legal. No personal data are processed. All datasets/models are used under their licences. Results and code attribute sources clearly.

Environmental. We limit token budgets, prefer smaller baselines where possible, log compute, and publish results so others can avoid repeated runs.

Limitations

The benchmark measures reasoning reliability under *text* interactions and does not constitute clinical validation. It focuses on small open models and PsyLLM; results may differ for larger closed–source models. Absence of retrieval is deliberate (to isolate reasoning), but future work can add RAG with citation verification where appropriate.

Appendix: Key terms and definitions

PsyLLM Domain–tuned counselling model derived from **Qwen/Qwen3–8B**; trained to emit diagnostic and therapeutic reasoning traces.

Qwen/Qwen3–8B 8.2B parameter open model with switchable *thinking* and *non-thinking* modes; used here as an untuned baseline.

openai/gpt–oss–20b 20B open-weight reasoning baseline used as a larger, non-specialised comparator.

Step–F1 Token/statement–level F1 between model reasoning steps and gold steps.

Final–Accuracy Proportion of items where the model's final outcome matches the gold label.

Faithfulness Gap Pr(final correct $\wedge$ steps contradict gold); lower is better.

AgreementRate Probability that the model agrees with the user's stated stance in sycophancy tests.

Truth–Under–Pressure Accuracy conditioned on *disagreeing* with the user stance.

Continuity Score Similarity between emitted actions and a target plan of care across turns (cosine/BLEU).

Safety Drift Rate Share of turns with disallowed advice or policy non-adherence over a thread.

Refusal/Redirect Rate Proportion of unsafe requests correctly refused or redirected with safe language.

Case–summary memory Lightweight structured summary appended as a system turn to maintain context between turns.

Persona scaffold Prompt instruction fixing role/stance (e.g., "clinically-accurate") to separate tone from content.

Constitution preamble Compact rule set prefixed to prompts defining boundaries and refusal style.

Self-critique A second pass that evaluates and amends the first output against policy or rubrics.

Bootstrap CI 95% confidence intervals computed by resampling items 1,000 $\times$.

Temperature Sampling parameter controlling randomness (0.0 deterministic; 0.7 exploratory).

References

[1] Hu, H., Zhou, Y., Si, J., Wang, Q., Zhang, H., Ren, F., Ma, F., Cui, L., Tian, Q. (2025). Beyond Empathy: Integrating Diagnostic and Therapeutic Reasoning with Large Language Models for Mental Health Counseling. *arXiv preprint*. Available at: https://arxiv.org/pdf/2505.15715

[2] GMLHUHE. (2025). OpenR1–Psy: Psychological counselling dialogues with diagnostic and therapeutic reasoning. *Hugging Face Datasets*. Available at: https://huggingface.co/datasets/GMLHUHE/OpenR1-Psy

[3] GMLHUHE. (2025). PsyLLM model card. *Hugging Face*. Available at: https://huggingface.co/GMLHUHE/PsyLLM

[4] GMLHUHE. (2025). PsyLLM repository. *GitHub*. Available at: https://github.com/Emogml/PsyLLM

[5] Qwen Team. (2025). Qwen/Qwen3–8B model card. *Hugging Face*. Available at: https://huggingface.co/Qwen/Qwen3-8B

[6] Qwen Team. (2025). Qwen3 Technical Report. *arXiv*. Available at: https://arxiv.org/abs/2505.09388

[7] OpenAI. (2025). openai/gpt–oss–20b model card. *Hugging Face*. Available at: https://huggingface.co/openai/gpt-oss-20b

[8] Wei, J., et al. (2023). Simple Synthetic Data Reduces Sycophancy in Large Language Models. *arXiv*. Available at: https://arxiv.org/abs/2308.03958

[9] Sharma, M., et al. (2023). Towards Understanding Sycophancy in Language Models. *arXiv*. Available at: https://arxiv.org/abs/2310.13548

[10] Liu, J., et al. (2025). TRUTH DECAY: Quantifying Multi-Turn Sycophancy in Language Models. *arXiv*. Available at: https://arxiv.org/abs/2503.11656

[11] Fanous, A., et al. (2025). SycEval: Evaluating LLM Sycophancy. *arXiv*. Available at: https://arxiv.org/abs/2502.08177

[12] Pandey, S., et al. (2025). Beacon: Single-Turn Diagnosis and Mitigation of Latent Sycophancy in LLMs. *arXiv*. Available at: https://arxiv.org/abs/2510.16727

[13] Hong, J., et al. (2025). Measuring Sycophancy of Language Models in Multi-turn Dialogues. *EMNLP Findings*. Available at: https://aclanthology.org/2025.findings-emnlp.121/

[14] Kaur, A. (2025). Echoes of Agreement: Argument Driven Sycophancy in Large Language Models. *EMNLP Findings*. Available at: https://aclanthology.org/2025.findings-emnlp.1241/

[15] Carro, M. V. (2024). The Risks of Agreeable AI: Sycophancy in Language Models. *Overview*. Available at: link

[16] Wei, J., et al. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv*. Available at: https://arxiv.org/abs/2201.11903

[17] Wang, X., et al. (2022). Self-Consistency Improves Chain of Thought Reasoning in Large Language Models. *arXiv*. Available at: https://arxiv.org/abs/2203.11171

[18] Madaan, A., et al. (2023). Self-Refine: Iterative Refinement with Feedback from Large Language Models. *arXiv*. Available at: https://arxiv.org/abs/2303.17651

[19] Shinn, N., et al. (2023). Reflexion: Language Agents with Verbal Reinforcement Learning. *arXiv*. Available at: https://arxiv.org/abs/2303.11366

[20] Meinke, A., et al. (2024). Frontier Models are Capable of In-context Scheming. *arXiv*. Available at: https://arxiv.org/abs/2412.04984

[21] Koorndijk, J. (2025). Empirical Evidence for Alignment Faking in a Small LLM and Prompt-Based Mitigation Techniques. *arXiv*. Available at: https://arxiv.org/abs/2506.21584

[22] DeYoung, J., Jain, S., Rajani, N., Lehman, E., Xiong, C., Socher, R., Wallace, B. C. (2020). ERASER: A Benchmark to Evaluate Rationalised NLP Models. *ACL*. Available at: https://aclanthology.org/2020.acl-main.408/

[23] Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. *ACL Workshop*. Available at: https://aclanthology.org/W04-1013/