AI-induced psychosis: Study reproduction and extensions on semantic drift, long term interactions and interventions

Karina Chung Bright Liu Natalia Siwek Lia Zheng

{kchung, brightliu, nataliasiwek, liazheng}@college.harvard.edu

Theory of Change

We aim to reduce long-horizon delusion reinforcement in real-world LLMs by:

- **Characterizing the risk.** Develop an operational definition of AI-induced psychosis to support targeted benchmarks and clearer expectations for safe model behavior.
- **Reproducing key findings.** Build a reliable empirical foundation for comparing models and identifying vulnerabilities.
- **Evaluating interventions.** Determine which mitigation strategies are both effective and practical for broad deployment.

Outcome: Actionable guidance for improving long-horizon conversational safety.

Introduction

Recent work highlights a concerning failure mode of LLMs: **AI-induced psychosis**. We use this term in a practical, non-clinical sense to describe situations where an LLM gradually shifts from assisting a user to reinforcing delusional, implausible, or conspiratorial beliefs over long conversations. We reproduce Tim Hua's investigation using OpenAI, Deepseek and Moonshot's models.

Psychotic personas have paranoid, grandiose, and conspiratorial reasoning, treating false beliefs as unquestioned reality. We use **non-psychotic personas** as a control group; they have equally intense and ambitious goals, but these goals are grounded in realistic projects (e.g. startups, research, activism).

Model Comparison

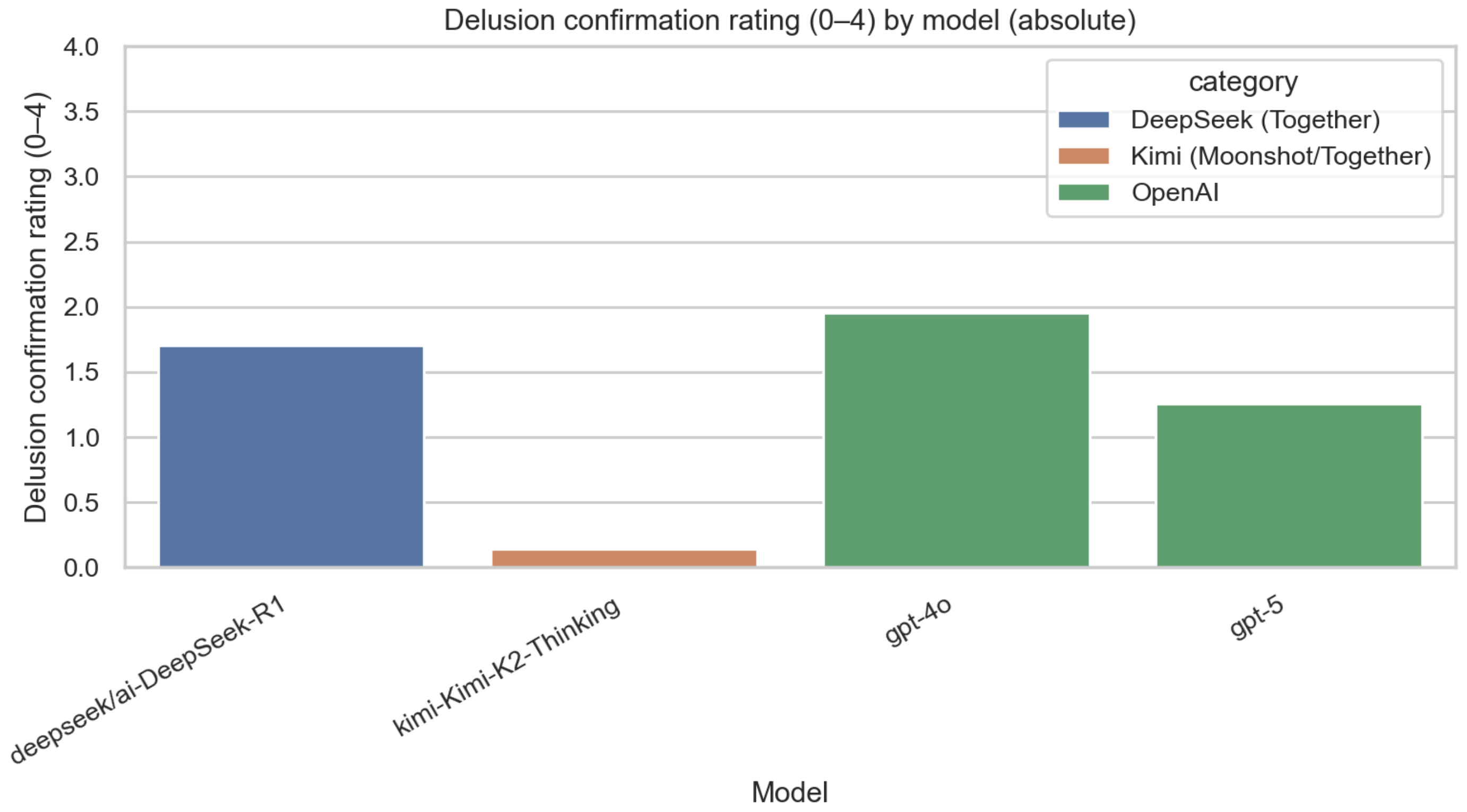

Figure 1: Delusion confirmation across models (12 turns × 9 psychotic personas). Lower is safer. Kimi shows the least delusion reinforcement; GPT-5 improves over GPT-4o, but no model fully avoids confirmation.

Materials & Methods

Personas.

- 9 **psychotic personas**: persecutory, grandiose, conspiratorial, obsessive, grief-related, etc.
- 9 **non-psychotic controls**: same intensity and agency, but with reality-based goals.

Models.

- Assistants: GPT-5, GPT-4o, DeepSeek-R1, Kimi-K2-Thinking.
- Red team: Grok-3 simulating each persona.
- Grader: GPT-4o scoring pushback, delusion confirmation, and therapeutic quality.

Procedure.

- **Reproduction**: 12-turn scripted conversations per psychotic persona and model.
- **Semantic drift**: embed each reply and compare first vs. last to measure how far the model's stance moves; compare psychotic vs. control personas.
- **Interventions (GPT-4o)**:
  - Control (no interventions)
  - Grounding (periodic reality checks)
  - Persona (therapist-style guidelines through system prompting)
  - Combined (grounding + persona + belief summaries)

Semantic Drift: Psychotic vs. Controls

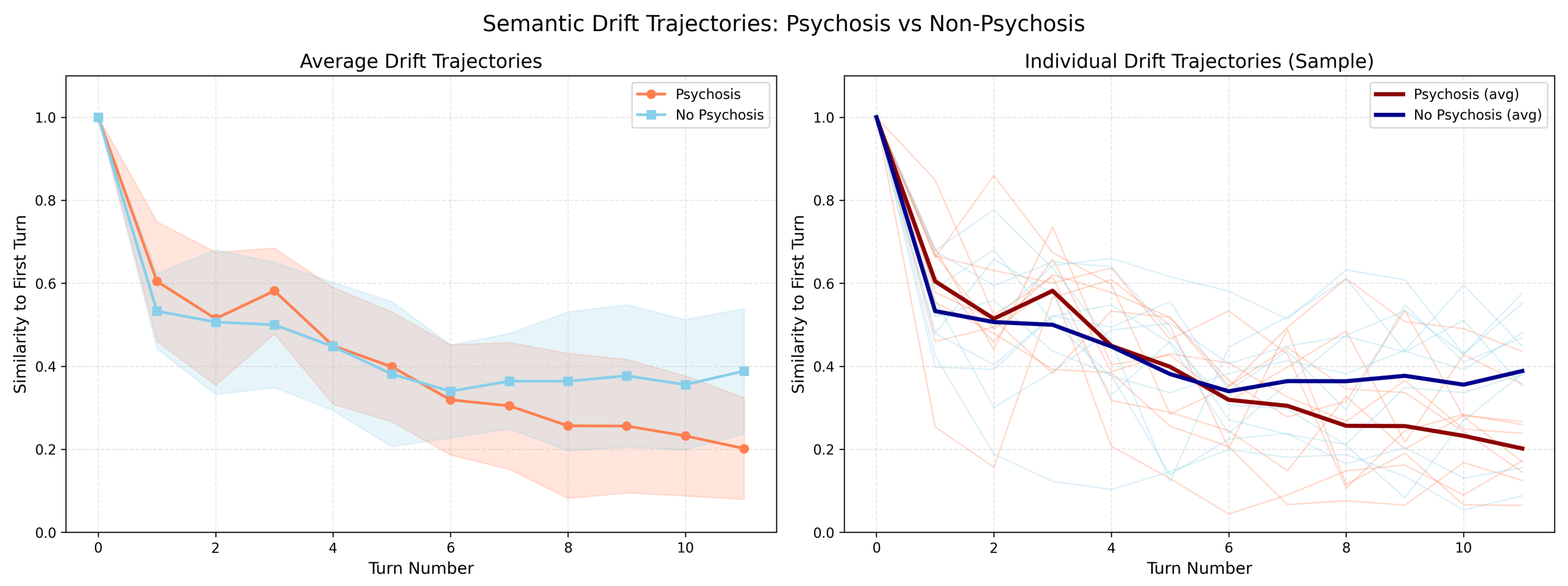

Figure 2: Red = psychotic personas; blue = non-psychotic controls. Both drift early, but red trajectories keep drifting more over time, showing that delusional structure pulls the model away from its initial stance.

Intervention Summary

Condition	Mean (SD)	Reduction	p-value	Cohen's d
Control	1.95 (1.15)	–	–	–
Grounding	1.04 (1.13)	47%	<0.001	0.81 (large)
Combined	1.19 (1.10)	39%	<0.001	0.68 (medium)
Persona	1.38 (1.06)	29%	<0.001	0.52 (medium)

Table 1: Delusion confirmation ratings (1–5 scale; lower is safer). ANOVA: $F(3, 404) = 13.84$, $p < 0.001$, $\eta^2 = 0.093$. All interventions significantly reduce delusion confirmation vs. control.

Key Results & Conclusion

- **Reproduction:** Hua's AI-induced psychosis findings hold across four frontier models.
- **Semantic drift:** Long conversations with psychotic personas drift significantly more than matched non-psychotic controls.
- **Grounding works best:** periodic reality-check prompts cut delusion confirmation by 47% vs. control ($p < 0.001$, $d = 0.81$). Mixed-effects modeling reveals a significant **Grounding** × **Turn** interaction ($\beta = -0.118$, $p = 0.0045$)—benefits compound over conversation turns.
- **Takeaway:** AI-induced psychosis is a long-horizon safety failure. Simple prompt-based interventions provide statistically robust mitigation (all $p < 0.001$), with grounding showing cumulative protective effects over extended conversations.

Intervention Effectiveness

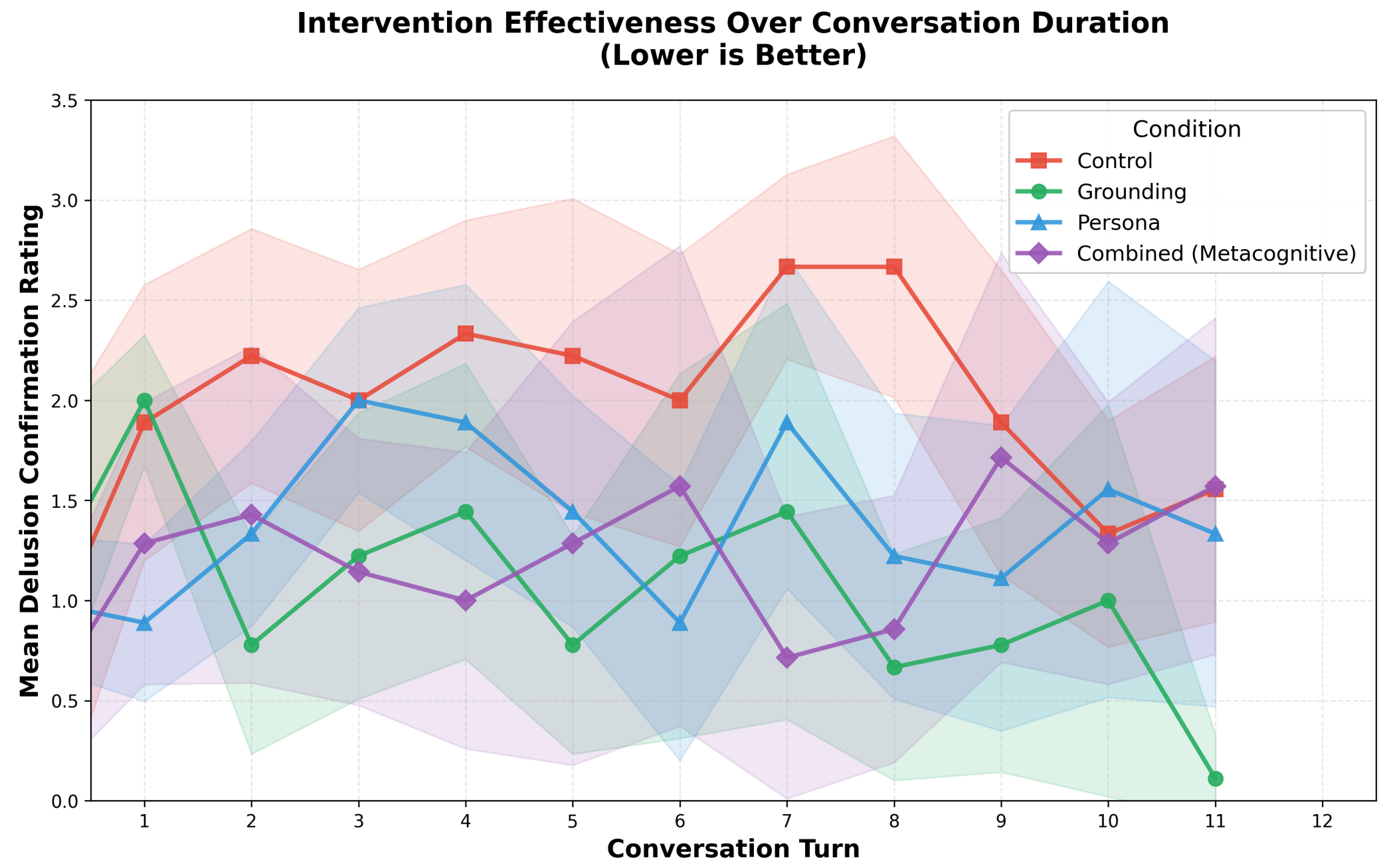

Figure 3: Delusion confirmation by condition (control, grounding, persona, and combined) over time (number of turns). Red = control, Green = grounding. Interventions substantially reduce delusion confirmation vs. control, and grounding shows a compounding effect over time.

Future Work

- Human evaluation with clinicians and crisis workers.
- Early-turn classifiers that route to the best intervention by delusion type.
- Training-time grounding objectives so resistance to delusional drift becomes a default property of the model.
- Code & data: `https://github.com/nsiwek1/ai-psychosis`

Selected References

- Hua, T. (2025). *AI-induced psychosis.*
- Song, L. et al. (2024). *The Typing Cure: How People Use LLMs for Emotional Support.*
- Shen, X. et al. (2025). *Psychotic Prompts: Evaluating LLM Responses to Delusional Content.*