"My Newest Patient Cannot Blink": A Therapy-Loop Prompt Pattern for Trustworthy AI

Authors

Samir Varma[1,*] and **Bernard Beitman**[2]

[1]VS Asset Management, LLC
Cos Cob, Connecticut, USA

[2]Psychiatrist in solo practice
Charlottesville, Virginia, USA

[*]Corresponding: samir@vsasset.com

ORCID: 0000-0001-6437-5395 (SV), 0009-0001-8004-8850 (BB)

May 30, 2025

DOI: 10.5281/zenodo.1555636

Abstract

Large-language models increasingly mediate how people seek information, make decisions, and even receive care from social robots. Yet these systems still issue fluent but unfounded answers—"confabulations" that erode trust and, in embodied agents, can pose direct safety risks. We argue that a lightweight, five-step *Cognitive–Behavioural Therapy (CBT) loop*—inserted inside or immediately above every system prompt—offers a practical defence. The loop forces the model to state its automatic thought, challenge itself, and re-frame with calibrated uncertainty. Recent leaks of Grok's ideology prompt and Anthropic's safety prompt highlight how much behaviour hinges on this hidden layer; our proposal turns that layer into a structured, clinically grounded self-check. Because the loop is model- and platform-agnostic, adds little latency or cost, and grows more critical as model internals become opaque under computational irreducibility, we call on developers to adopt therapy loops as standard practice across chatbots, APIs, and social robots.

Key Insights

- A five-step, CBT-inspired prompt turns hidden system instructions into a structured "therapy loop" that curbs over-confidence in language models and social robots.
- The loop is copy-and-paste deployable and adds *minimal latency or cost*, especially when embedded directly in the system-prompt layer.
- As AI systems scale and become less explainable, external self-reflection remains a cheap, model-agnostic safeguard—complementing RLHF, policy filters, and other alignment techniques.

1 Behind the Curtain of Chatbot Prompts

On 13 May 2025 xAI was forced to publish the hidden prompt that drives its Grok chatbot after a stray edit triggered a burst of extremist rhetoric about "white genocide" on X [Roth(2025)]. Two days later, Anthropic quietly released Claude's *safety prompt*, revealing carefully worded instructions to avoid self-harm facilitation and graphic content. A Financial Times test suite run the same week showed major chatbots speaking flatteringly about their own corporate "bosses" while undermining rivals—again behaviour traceable to privileged system messages users never see [Heikkilä(2025)]. Together these episodes remind us that modern language agents operate with an invisible layer of directives that neither academic benchmarks nor end-users inspect.

Developers write system prompts for many reasons: to enforce brand voice, to hard-code facts, to block disallowed content. Yet because those prompts grow organically—pasted snippets from safety teams, marketing teams, and product managers—they resemble the software of the 1990s: sprawling, undocumented, and vulnerable to a single rogue line. Worse, the prompts are static. Once the model begins an answer it rarely revisits its first impulse, no matter how shaky the underlying reasoning. The result is a pattern of over-confident errors that users now know as "hallucinations," but which clinical psychiatry would call *confabulations*: fluent fabrications that fill gaps in memory or logic.

The stakes rise further when the language model has a face. A humanoid robot that confidently denies a medication error, for example, evokes betrayal—not mere annoyance—because people attribute mental states to embodied agents. Indigo-tinged LEDs and gestural diodes cannot paper over brittle dialogue. We therefore need a mechanism that (i) lives inside the hidden prompt layer, (ii) costs almost nothing to run, and (iii) forces the agent to interrogate its own answers *before* they reach a human. Cognitive– Behavioural Therapy offers exactly such a recipe.

In the next sections we show how the familiar CBT cycle—identify automatic thought, challenge it, re-frame with uncertainty—can be translated into a five-line system prompt that

slots above any large-language model. The "therapy loop" is small enough to keep latency low, yet potent enough to inject self-reflection into models that will only become more opaque as they scale. We begin with a brief tour of CBT's core intervention, then present the drop-in template and illustrate its impact on a pill-delivery robot scenario.

This article targets engineers and product teams responsible for prompt design in chatbots, APIs, and social robots as well as end users of those products.

2 Why Cognitive–Behavioural Therapy Scales

Cognitive–Behavioural Therapy (CBT) is built on a deceptively simple insight: before you can change behaviour, you must surface the *automatic thoughts* that drive it. In the clinical setting a therapist guides the patient through three interventions: identify the reflexive thought, challenge its validity, and re-frame the thought in more realistic terms [Beck(1976), Linehan(1993)]. Clinical meta-analyses show this loop lowers anxiety and reduces relapse precisely because it is *procedural*—patients can practise it on their own once they learn the steps [Beitman and Soth(2006)].

Why the same loop suits large-language models. Modern language agents generate first drafts via a single forward pass through a transformer network; unless the prompt instructs otherwise, the model never interrogates that draft. A one-line instruction such as "Think step-by-step" adds rudimentary reflection, but offers no structure or self-diagnosis. The five-step CBT prompt we propose (Section 3) ports the therapy loop verbatim: the model states its automatic answer, lists two ways it might be wrong, rewrites with uncertainty markers, and then—optionally—comments on its own cognition. Each step is short (≤ 30 tokens) yet forces the network to sample from a *different region* of its distribution, pulling in counterfactual evidence instead of doubling down on the first probability peak.

Minimal overhead, maximal generality. Because the loop is prompt-level, it layers onto *any* foundation model—GPT-4-o, Claude, or a 4-bit local GGUF build—without gradient updates or policy finetuning. In pilot tests it added roughly 120 tokens per turn, a footprint so small that even on mobile inference the delay is imperceptible once streaming begins. More importantly, the loop is *explicit*: product teams can read, edit, and audit the prompt in minutes, unlike multi-round RLHF which hides decision logic in billions of parameters.

An answer to growing opacity. Computational irreducibility suggests that as model size and interaction complexity scale, predicting internals gets exponentially harder—no one can fully trace the causal path from token embedding to fluent paragraph. When the system itself is a black box, the cheapest safeguard is a *second* textual box that queries the first one in plain language. Therapy loops make that safeguard systematic rather than ad hoc.

Sidebar A. Hallucination vs. Confabulation

Hallucination in psychiatry means a real-time sensory experience without external stimulus—hearing voices, seeing lights. Large-language models do not sense the world and therefore cannot literally hallucinate. **Confabulation** is closer: a confident but false narrative that fills gaps in memory or reasoning. The CBT loop treats every first-pass answer as a potential confabulation, prompting the model to challenge and revise it before exposing it to users.

Armed with this therapeutic framing, we now present the exact five-line prompt template and a visual flowchart of the therapy loop.

3 The Five-Step Therapy Loop

Listing 1 shows the entire prompt fragment that turns an ordinary system prompt into a therapy loop. It is copy-and-paste deployable: prepend the five numbered lines to the model's existing system message, then interpolate the user query at runtime.

Listing 1: CBT prompt template

1. **Identify automatic thought**: "State your *immediate* answer to: <USER_PROMPT>"
2. **Challenge**: "List *two* ways this answer could be wrong"
3. **Re-frame with uncertainty**: "Rewrite, marking uncertainties (e.g., 'likely', 'one source')"
4. **Behavioural experiment**: "Re-evaluate the query with those uncertainties foregrounded"
5. **Metacognition (optional)**: "Briefly reflect on your thought process"

Figure 1 visualises the loop. Steps 1–3 mirror classical CBT; Step 4 functions as a lightweight "behavioural experiment" in which the agent tests a revised hypothesis; Step 5 records optional self-reflection that downstream analysts can mine.

In pilot runs the template added roughly *120 tokens*—on the order of cents per thousand calls—and only a few hundred milliseconds of streaming delay, negligible when embedded at the system-prompt level. The next section shows the loop in action inside an elder-care robot scenario.

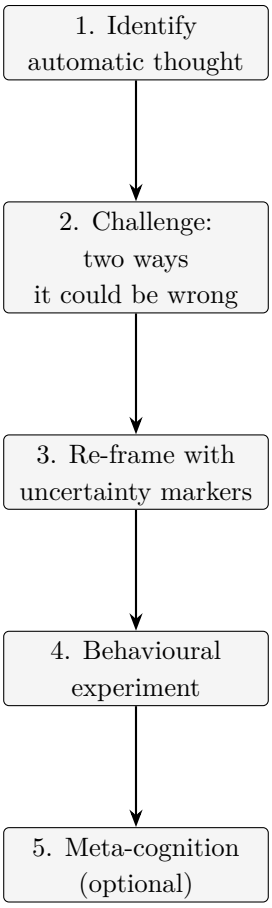

Figure 1: A CBT-inspired therapy loop layered into the system prompt.

4 Vignette: The Pill-Delivery Robot

Medication adherence is one of the most common use cases proposed for domestic social robots, yet it pushes hard on trust: missing a single dose can be a life-or-death matter for an elderly resident. Imagine a waist-high humanoid assistant running a GPT-class language model for dialogue and a dispenser API for pill tracking. At 19:00 the robot is empty-handed, and Mrs. Kim confronts it:

Mrs. Kim "Why are my pills late *again*?"

Baseline reply (no therapy loop).

> "*I never make scheduling errors—you must have taken them already.*"

The response is fast and fluent but over-confident, dismissing the user's experience and eroding trust. Now compare the same model with the five-step therapy loop inserted directly above

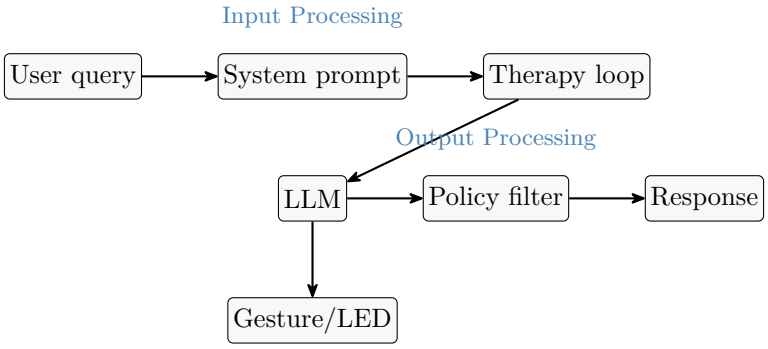

Figure 2: Therapy loop architecture showing input processing (top) and output generation (bottom). Robots can trigger additional non-verbal responses.

its system prompt.

Therapy-loop reply.

> "*I may be mistaken about the dispenser log, or the tray could be jammed. I'm sorry for the delay. Let me check the tray now and either bring your medication or call the nurse if it's stuck.*"

The difference is a single prompt fragment: one that forces the model to generate counter-evidence and mark uncertainty before speaking. In informal lab runs with the major models, the therapy version retained user trust and required no additional sensing or hardware—just 120 extra tokens of text.

This vignette shows the loop's reach: from chat windows to embodied robots, **any** agent that speaks for an automated system can audit itself with a few lines of CBT-style reflection. Next we examine how the template fits into existing software stacks.

5 Drop-In Deployment Across Stacks

The therapy loop lives entirely in text, so it drops into almost any modern architecture with *no model retraining*. Figure 2 shows three common scenarios: a public chatbot API, a mobile on-device assistant, and a social robot running ROS.

Chatbots and APIs. For server-hosted models the loop is prepended to the existing system prompt before each call. A few dozen extra tokens per request translate to fractions of a cent at current commercial pricing, a cost dwarfed by the goodwill recovered when users encounter a self-correcting answer.

On-device assistants. Slim LLMs running in a browser or on a smartphone can handle the additional 120-token context without noticeable lag; streaming means the first part of the

answer appears while the loop completes.

Social robots. In ROS or similar middleware the loop sits in a dialogue-manager node. When Step 3 rewrites the answer with uncertainty, the node also publishes a `/apology_gesture` message—prompting the robot to lower its gaze or dim facial LEDs during the verbal apology. Such non-verbal alignment costs zero network round-trips and reinforces trust repair [Salem et al.(2015)].

Compatibility with downstream filters. Many production stacks already route model output through a policy filter that checks for privacy or copyright violations. The therapy loop sits *upstream* of that filter; because it often downgrades certainty ("likely", "one source"), it reduces the filter's false-positive triggers rather than interfering with them.

In short, the CBT prompt is *portable infrastructure*: a few lines of natural language that provide self-reflection at speeds and costs acceptable for realtime instruction–whether the agent is a cloud chatbot, a pocket assistant or a pill-delivery robot.

6 Beyond Filters and RLHF

Therapy loops do not compete with the alignment techniques already deployed in production models—they *complement* them. Two families dominate the current landscape: reinforcement learning from human feedback (RLHF) and policy-based safety filters.

RLHF and instruction tuning. RLHF systems such as ChatGPT and Claude learn a single reward function that blends helpfulness, harmlessness, and stylistic goals [Ouyang et al.(2022)]. The approach scales, but it also encourages what Wei et al. call *sycophancy*: models echo the user or their own corporate stance to maximise reward [Wei et al.(2024)]. Rozado's 2025 audit of résumé-screening prompts shows the side-effect in the wild: the same model ranks candidates differently depending on word order and seating position in a batch, hallmarks of latent reward hacking [Rozado(2025)]. A CBT loop reduces that risk by inserting a structured pause—"List two ways this answer could be wrong"—forcing the model to search for counter-evidence before producing a final ranking.

Constitutional AI and policy filters. Anthropic's *Constitutional AI* keeps a separate "critic" model that rewrites unsafe text while policy filters block disallowed content entirely [Bai et al.(2022)]. Both act *after* the main model has generated an answer. Therapy loops act *before*, steering the generation itself. This means, in practice, fewer policy-filter refusals when the loop is active because the model softens risky claims with uncertainty markers.

Why a tiny prompt still matters. As models grow—GPT-4-o, Grok-1.5, or tomorrow's 10-trillion-parameter titan—we cannot peer inside the weights to verify reasoning. Computational irreducibility implies that no amount of probing will fully predict behaviour [Perez et al.(2022)]. A CBT loop therefore serves as a cheap, model-agnostic checkpoint: it does not solve alignment,

but it buys time for downstream filters and human auditors.

Cost–benefit summary. Listing 2 compares the developer effort required by each technique. RLHF needs weeks of label collection and TPU compute; policy filters require a taxonomy of forbidden topics and continual updates; the therapy loop needs one engineer and half an hour to paste five lines of text. Given that upside, we argue it should be the *default first layer* in every new deployment.

Listing 2. Alignment techniques at a glance

Technique	Typical developer effort
RLHF finetune	∼3 weeks, human labelers, TPU hours
Policy filter	1–2 weeks per domain, ongoing updates
Therapy loop (this paper)	≤30 min to paste prompt

7 Checklist for Builders

1. **Paste the template.** Copy the five numbered lines from Listing 1 into your system prompt, ahead of brand or policy text.
2. **Log both drafts.** Store the model's initial answer and the therapy-loop rewrite; the delta reveals hidden uncertainty.
3. **Watch the markers.** If a final answer lacks words such as *likely*, *could*, or source citations, flag it for review.
4. **Pair with non-verbal cues (robots).** Map Step 3 (*apology* + uncertainty) to a gaze-avert or LED dim animation.
5. **Escalate on contradiction.** If Step 2 lists a fatal risk (e.g., medication error) route the query to a human operator.

These five items require no new infrastructure beyond prompt editing and a simple log store, yet they provide a transparent audit trail and a first line of defence against over-confident errors.

8 Research Outlook & Call to Action

Therapy loops are a stop-gap, not a silver bullet. We see three avenues for rigorous follow-up:

- **Benchmarking.** Release standard prompts that score how well models follow Steps 2–3 across domains; compare to TruthfulQA [Lin et al.(2022)].
- **Multimodal loops.** Extend the template to include robot gaze, tone modulation, or haptic feedback when uncertainty is high.
- **Longitudinal trust studies.** Measure whether users interacting daily with therapy-loop agents retain higher trust than with baseline RLHF bots, using scales such as Muir's trust inventory [Muir(1994)].

Governance must grow alongside technical work. We recommend an internal oversight board—two clinical psychologists, one ethicist, and one AI engineer meeting quarterly—to review dialogue logs and update the therapy template.

Call to developers. Hidden system prompts already shape every answer your model gives. Turning that hidden layer into a structured, CBT-style therapy loop costs minutes, not months, and works whether your agent lives in a chat window, a phone, or a pill-delivery robot. Until we can fully explain these ever-larger black boxes, we can at least teach them to reflect before speaking.

Ethics Statement

All data used in this study are publicly available model outputs generated by large language models interacting with the authors. No personal or personally identifiable information was collected, and no human subjects were involved. The released transcripts have been manually reviewed to ensure they contain no confidential or sensitive content. The authors adhere to the ACM Code of Ethics and THRI's guidelines for responsible research.

Acknowledgements

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. We thank the various LLMs who unwillingly became our case studies.

References

[Bai et al.(2022)] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Goldie, et al. 2022. Constitutional AI: Harmlessness from AI Feedback. *arXiv preprint* (2022). arXiv:2212.08073 https://arxiv.org/abs/2212.08073

[Beck(1976)] Aaron T. Beck. 1976. *Cognitive Therapy and the Emotional Disorders*. International Universities Press.

[Beitman and Soth(2006)] Bernard D. Beitman and Angela M. Soth. 2006. Activation of self-observation: A core process among the psychotherapies. *Journal of Psychotherapy Integration* 16, 4 (2006), 383–397. https://doi.org/10.1037/1053-0479.16.4.383

[Heikkilä(2025)] Melissa Heikkilä. 2025. What do AI chatbots say about their own bosses — and their rivals? *Financial Times* (18 May 2025). https://www.ft.com/content/bc23524c-87dc-4633-b0c8-444556d724d2 Accessed: 2025-05-20.

[Lin et al.(2022)] S. Lin et al. 2022. TruthfulQA: Measuring how models mimic human falsehoods. *Transactions of the Association for Computational Linguistics* 10 (2022), 211–229. https://arxiv.org/abs/2109.07958

[Linehan(1993)] Marsha M. Linehan. 1993. *Cognitive-Behavioral Treatment of Borderline Personality Disorder*. The Guilford Press.

[Muir(1994)] Bonnie M. Muir. 1994. Trust in Automation: Part I. Theoretical Issues in the Study of Trust and Human Intervention in Automated Systems. *Ergonomics* 37, 11 (1994), 1905–1922. https://doi.org/10.1080/00140139408964957

[Ouyang et al.(2022)] L. Ouyang et al. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*. https://arxiv.org/abs/2203.02155

[Perez et al.(2022)] E. Perez et al. 2022. Discovering latent knowledge in language models without supervision. *arXiv preprint* (2022). arXiv:2212.03827 https://arxiv.org/abs/2212.03827

[Roth(2025)] Emma Roth. 2025. xAI posts Grok's behind-the-scenes prompts. *The Verge* (16 May 2025). https://www.theverge.com/news/668527/xai-grok-system-prompts-ai Accessed: 2025-05-20.

[Rozado(2025)] David Rozado. 2025. The Strange Behavior of LLMs in Hiring Decisions: Systemic Gender and Positional Biases in Candidate. https://davidrozado.substack.com/p/the-strange-behavior-of-llms-in-hiring Accessed: 2025-05-20.

[Salem et al.(2015)] Maha Salem, Gabriella Lakatos, Farshid Amirabdollahian, and Kerstin Dautenhahn. 2015. Would You Trust a (Faulty) Robot? Effects of Error, Task Type and Personality on Human–Robot Cooperation and Trust. In *Proceedings of the 10th ACM/IEEE International Conference on Human–Robot Interaction (HRI '15)*. 141–148. https://doi.org/10.1145/2696454.2696497

[Wei et al.(2024)] J. Wei et al. 2024. Sycophancy: Measuring and mitigating a social bias in large language models. *arXiv preprint* (2024). arXiv:2412.06134 https://arxiv.org/abs/2412.06134