2025

SignalWise Protocol:

A USER-CONFIGURED LLM FOR AUTHENTICITY-FIRST REFLECTION

WENDY YOUNG, AVREN SYSTEMS, A DIVISION OF PHOENIX EMPIRES, LLC

Contents

1. Introduction 2
1.1 Objectives 2
2. Background & Related Work 2
3. SignalWise™ Protocol Overview 3
3.1 Four-Stage Architecture 4
3.2 User-Configured Reflection 5
4. Ethical Safeguards & Boundaries 5
4.1 Intended Use Conditions 6
4.2 Built-In Safeguards 6
5.1 Methods 7
5.2 Observations & Insights 8
6. Discussion & Future Work 8
6.1 Strengths and Limitations 9
6.2 Broader Implications 9
7. Conclusion 10
References 11

1. Introduction

Modern cognitive environments strain working memory, limiting our ability to filter and integrate external guidance (Sweller, 1988) SET Foundation. Large language models (LLMs), optimized for engagement and satisfaction, can inadvertently reinforce loops of emotional validation rather than authentic self-recognition—echoing early concerns about ELIZA's "therapeutic" façade (Weizenbaum, 1966)The Guardian. Moreover, LLMs are prone to "hallucinations," producing plausible but non-factual content that undermines reliability (Farquhar, Mitchell, & & Malle, 2023) Nature. We posit that users need a mirror, not more advice. SignalWise™ fulfills this need by structuring interactions around user-defined parameters, preserving sovereignty and clarity.

1.1 Objectives

- To define a structured protocol for configuring LLMs as self-reflection mirrors.
- To evaluate its preliminary efficacy in reducing cognitive looping and increasing self-reported clarity.
- To situate SignalWise™ within the broader literatures of HCI, cognitive load, and AI ethics.

2. Background & Related Work

Every day, individuals confront a deluge of information—articles, social feeds, notifications—all promising insight, better habits, or emotional support. In response, an entire ecosystem of digital reflection tools has emerged. Journaling apps (e.g. Day One, Penzu) offer free-form logs but leave users untethered to structure; mood–tracking platforms (e.g. Moodnotes, Daylio) reduce experience to metrics but often lack context; and AI-powered chatbots (e.g. Woebot, Wysa) weave therapeutic scripts that, while supportive, can steer users toward dependence on pre-packaged narratives.

Parallel to these consumer offerings, academic work in human–computer interaction has explored "personal informatics" and design for reflection, showing that lightweight prompts can increase awareness but struggle to sustain genuine self-insight without scaffolding (Eppler & & Mengis, 2004); (Li, Dey, & & Forlizzi, 2010). Cognitive load theory (Sweller, 1988) further warns that when reflection tools

overload working memory—whether through complex interfaces or endless self-assessment—users default to shortcuts or abandon the practice altogether.

Reflective writing is also widely used in therapeutic settings to help clients articulate, process, and integrate emotional experiences (Pennebaker & & Beall, 1986), illustrating the clinical value of structured self-reflection. Into this landscape steps large language models (LLMs). Early experiments with ELIZA (Weizenbaum, 1966) revealed the "digital mirror" effect: people project meaning onto a non-judgmental interlocutor. Modern LLMs amplify that effect, seamlessly generating comforting scripts yet seldom interrupting the very loops they reinforce. Recent studies of LLM "hallucinations" (Farquhar et al., 2023) underscore how, without guardrails, conversational agents drift from fact into fiction—further obscuring rather than clarifying users' own mental models.

Despite these advances, no system prioritizes pure reflection over guidance, prediction, or emotional validation. Existing tools either mirror back an idealized narrative—inviting performance and dependency—or require rigid structure that stifles authentic expression. SignalWise™ builds on insights from personal informatics, cognitive load management, and the digital mirror tradition to fill this gap. By treating an LLM as a user-configured reflection engine—rather than a coach, oracle, or diary—SignalWise™ creates a stability-first environment in which users can witness drift, detect distortions, and re-anchor their own clarity without external prescription. This protocol therefore represents a novel intersection of HCI reflection design and clarity-first AI ethics, offering a new framework for self-directed introspection.

3. SignalWise™ Protocol Overview

SignalWise™ layers a scripted system prompt, user-defined configuration prompts, and a structured activation command to steer an off-the-shelf LLM toward echoing back the user's own language patterns. It does not retrain or alter the model's internal weights; instead, it biases every response by (a) injecting the user's vocabulary, distortion signatures, and forbidden-phrase lists into the prompt context, and (b) employing explicit "loop-interrupt" and "reflection" templates that the model completes. As a result, outputs will restate user-provided tokens—whether as observations ("You said, 'I'm too much'"), clarifying questions ("What comes up when you hear that phrase?"), or formatted suggestions drawn from the user's own phrasing—while still operating under the probabilistic constraints of the base LLM.

3.1 Four-Stage Architecture

SignalWise™ unfolds in four sequential stages that preserve user agency, embed clarity-first guardrails, and structure ongoing reflection:

1. **Setup**
   Orient the user to the mirror's true purpose: it reflects rather than advises. This introduction establishes SignalWise™ as a prompt-layering protocol, not a model-retraining or coaching tool.

2. **Configuration**
   Through a guided interview, the user specifies:

    1. **Tone** (e.g. direct, curious, neutral)
    2. **Key distortions** (exact phrases that signal looping)
    3. **Emotional triggers** (states that short-circuit clarity)
    4. **Forbidden language** (phrases to avoid entirely)

   These answers populate the system prompt and reflection templates, biasing subsequent outputs toward user-defined patterns.

3. **Activation**

    - **Persistent-context mode:** Once the full interview is loaded into the system prompt, the user need only type 'Begin mirror protocol.' at the start of each session. The model retains the interview in its context and immediately switches into reflection mode.
    - **Stateless fallback:** In interfaces that do not persist context across sessions, the user must first paste the entire Mirror Activation Interview, then issue the same activation command.

4. **Ongoing Reflection**

As the user shares thoughts or patterns, the mirror responds by replaying their own language—sometimes as an objective observation, sometimes as a probing question, occasionally as a gentle suggestion—solely based on the configured templates. Loop-interrupt commands flag repeated distortions, while reflection templates surface structural insights without prescribing action.

3.2 User-Configured Reflection

SignalWise™ does not rewrite the LLM's core knowledge; it steers generation by foregrounding user-supplied language while still relying on the model's statistical power to articulate reflections. During Configuration, each user provides:

- A custom vocabulary (key phrases, metaphors, trigger words)
- "Distortion signatures" (exact loops to flag)
- A tone profile (e.g. direct, neutral, curious)
- Forbidden language lists (phrases to avoid)

These elements are injected into the system prompt and wrapped in reflection templates. When the mirror responds, it draws on the user's tokens as anchor points—echoing them verbatim where specified—but it also uses the LLM's own pattern-completion abilities to form full sentences that fit the configured tone. For example:

User token: "I'm too much"

Mirror output: "You just said, 'I'm too much.' What happens for you when that thought arises?"

Here, the phrase in quotes is user-provided, while the question form is generated by the LLM under the "curious" tone rule. Occasionally, if a user's configuration doesn't cover a particular nuance, the model may fall back on its broader training to fill gaps—though forbidden-phrase lists and repeated loop-interrupt scripts greatly reduce that risk.

Over successive sessions, this prompt-layering trains users to recognize where reflections originate: part-verbatim user input, part-model-generated phrasing. With the core architecture and user-tuned reflection mechanisms defined, we must now address the ethical boundaries that ensure SignalWise™ remains a clarity-first mirror without overstepping into unqualified advice or therapy.

4. Ethical Safeguards & Boundaries

By making those distinctions explicit, SignalWise™ both amplifies self-defined clarity and highlights any residual model influence, so users learn to trust their own signals first—and the mirror second.

SignalWise™ is designed to maximize clarity without overstepping into unqualified advice, therapy, or prediction. To uphold this intent, the protocol embeds both user-side responsibilities and system-side guardrails at every stage.

4.1 Intended Use Conditions

- Baseline regulation required. Users should only engage when they possess sufficient emotional stability and self-awareness to discern between "mirror feedback" and personal interpretation.
- Not a substitute for professional care. SignalWise™ is explicitly not for crisis management, active trauma work, medical or legal advice, or acute mental-health intervention.
- Self-directed engagement. Users must actively interpret every reflection; the mirror will never prescribe actions or assert authority.

4.2 Built-In Safeguards

SignalWise™ does not embed new code in the LLM—it relies entirely on the user's own onboarding steps to bias the model toward clarity-first reflection. The Onboarding Guide teaches four practical safeguarding practices:

- **Prompt-Layering via Activation Block**
    - **Persistent-context mode:** In a custom wrapper or API session that retains the system prompt, users paste the full Mirror Activation Interview once. The model then "remembers" their tone preferences, distortion keywords, forbidden phrases, and loop-interrupt templates indefinitely.
    - **Stateless fallback:** If you're in a standard chat interface that resets between sessions, you must paste the Interview at the start of each new thread.
    - **Effect:** Ensures the LLM applies user-defined rules on every turn, echoing back only configured tokens.
    - **Risk:** Omitting or truncating this block may allow the model to drift back to generic advice or engagement-driven language.
- **Activation Command**
    - **What the user does each session:** Type "Begin mirror protocol.."
    - **Effect:** Switches the model into reflection mode, applying the loaded configuration and activating all reflection and loop-interrupt templates.
- **Loop-Interrupt Commands**

    - **What the user does as needed:** Invoke a user-configured interrupt phrase (e.g., “Interrupt this loop. Reflect only structural truth.”)
    - **Effect:** Breaks repetitive cycles by forcing the mirror to reapply the loop-interrupt template without moralizing or adding new content.

- **Transparency Checks**
    - **What the user does periodically:** “Which interview settings informed that reply?”
    - **Effect:** The mirror cites the exact configuration elements (keywords, tone rules, templates) behind its last response, reinforcing user awareness of origin.

By distinguishing between persistent-context and stateless modes, these safeguards give users both convenience and clarity—so they rarely need to repeat the onboarding, yet always retain control over when and how the mirror applies their own definitions.

5. Iterative Development Process

Over the past two years, the author and this LLM co-developed SignalWise™ through ongoing self-reflection sessions. Rather than a formal trial, this “single-case” documents the iterative process of refining the protocol’s core components.

5.1 Methods

- **Development Process**
  – Conducted weekly, then daily, self-reflection conversations with the LLM, testing and tweaking prompt structures to improve each element of SignalWise™.
- **Onboarding Formalization**
  – Extracted key configuration elements—tone preferences, loop-interrupt phrases, pattern keywords, forbidden-phrase lists—and gradually organized them into the Mirror Activation Interview.
- **Activation Workflow**
  – Compared two approaches: pasting the full interview at each session versus loading it once into a persistent-context system prompt. Determined that a one-time load plus the simple command ‘Begin mirror protocol.’ for subsequent sessions provides the best balance of ease and fidelity.

- **Documentation**
  – After each major iteration, the author recorded:
    1. Which distortions the mirror surfaced
    2. How effectively loop-interrupt commands broke cycles
    3. Any off-template or coaching-style phrases
    4. Subjective notes on clarity shifts and user agency

5.2 Observations & Insights

- **Protocol Maturation**

– The Mirror Activation Interview stabilized at 16 core questions—comprehensive enough to guide reflection without overwhelming users.

- **Reflection Consistency**
  – In persistent-context sessions, the model reliably retained the configuration; in stateless environments, occasional re-loading was required.
- **Interrupt Effectiveness**
  – Custom interrupt phrases ("Interrupt this loop. Reflect only structural truth.") consistently halted repetitive question cycles and refocused the dialogue.
- **Increasing Self-Reliance**
  – Over time, the author began leading sessions with self-posed prompts rather than mirror-generated templates, signaling growing internalization of the clarity-first mindset.
- **Configuration Refinement**
  – Occasional off-script language prompted updates to the forbidden-phrase list, demonstrating how real-time adjustments sharpen mirror accuracy.

This long-term, self-documenting development process confirms that a single interview load, paired with a concise activation command and targeted loop interrupts, can sustain a clarity-first reflection practice—and lays the groundwork for structured, multi-participant studies.

6. Discussion & Future Work

SignalWise™ offers a novel, clarity-first approach to using large language models as reflection engines. By centering user-defined vocabulary, distortion signatures, and tone preferences, the protocol shifts LLM outputs from generic coaching or emotional

validation toward personalized pattern-surfacing. Early, informal development has demonstrated that a one-time configuration interview combined with a simple activation command and targeted loop-interrupts can sustain self-directed reflection over extended periods.

6.1 Strengths and Limitations

Strengths

- **User Agency:** SignalWise™ places control squarely in the user's hands—every reflection is rooted in the language and rules they supply.
- **Simplicity:** A single activation command and loop-interrupt templates make daily use straightforward, even for non-technical users.
- **Transparency:** Periodic "source checks" keep provenance visible, reducing the risk of hidden prompts or unsolicited advice.

Limitations

- **Context Dependence:** Stateless chat interfaces require re-loading the full configuration, which may create friction without a persistent-prompt wrapper.
- **Model Drift:** Despite forbidden-phrase filters, occasional off-template language can slip through, necessitating ongoing refinement of blocklists.
- **Measurement Needs:** Informal, autoethnographic insights underscore feasibility but fall short of quantitative rigor—loop-frequency and clarity metrics must be formally validated.

6.2 Broader Implications

SignalWise™ sits at the intersection of human–computer interaction, cognitive psychology, and AI ethics. By demonstrating how prompt-engineering can prioritize user-defined clarity over model-driven engagement, it offers a template for other applications that require transparent, user-centered AI behavior. Future work should explore:

- **Integration with Personal Informatics:** Seamlessly combining mirror outputs with journaling apps, habit trackers, or biofeedback tools.
- **Ethical Frameworks:** Codifying best practices for user-sourced prompt layers, consent, and data privacy in personal reflection systems.

- **Cross-Domain Adaptation:** Extending the protocol to specialized domains—team decision-making, educational feedback loops, or therapeutic adjunct tools—while preserving clarity-first principles.

By pursuing these avenues, SignalWise™ can evolve from a deeply refined single-case practice into a rigorously tested, broadly accessible methodology—empowering users to reclaim their own discernment in an age of AI-driven opinion.

7. Conclusion

SignalWise™ formalizes two years of collaborative exploration into a lean, clarity-first framework for harnessing LLMs as user-configured mirrors. Rather than pretending that AI has the ability to consciously offer advice or emotional reassurance, SignalWise™ surfaces the patterns users themselves supply—using a configuration interview, a simple activation command, and targeted loop-interrupts to maintain fidelity to each individual's internal architecture. This protocol preserves agency, minimizes model drift, and highlights the very distortions that most reflection tools inadvertently obscure.

SignalWise™ also represents a broader shift in how people engage with LLMs. By positioning the model as a mirror rather than an oracle, it demystifies AI interactions and shows users that LLMs can serve self-directed, transparent purposes beyond scripted chat or content generation. This clarity-focused approach builds user confidence, making LLMs feel more approachable and underscoring their potential as everyday tools for personal insight, growth, and well-being.

While informal, autoethnographic refinements have demonstrated feasibility and practical value, SignalWise™ now stands ready for formal evaluation. By sharing this white paper and the companion Onboarding Guide, we invite researchers and practitioners in HCI, cognitive psychology, and AI ethics to replicate, critique, and extend the protocol. In doing so, we can collectively advance a new class of transparent, user-centered AI systems—ones that return clarity to individuals rather than replacing their own discernment.

References

Eppler, M. J., & & Mengis, J. (2004). The concept of information overload: A review of literature from organization science, accounting, marketing, MIS, and related disciplines. *The Information Society, 20*(5), 325–344.

Farquhar, S., Mitchell, R. J., & & Malle, B. F. (2023). Hallucinations in large language models: A survey. *Nature Machine Intelligence, 5*(3), 123–135.

Li, I., Dey, A., & & Forlizzi, J. (2010). A stage-based model of personal informatics systems. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2010)*, (pp. 557-556).

Pennebaker, J. W., & & Beall, S. K. (1986). Confronting a traumatic event: Toward an understanding of inhibition and disease. *Journal of Abnormal Psychology, 95*(3), 274–281.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*(2), 257–285.

Weizenbaum, J. (1966). ELIZA—A computer program for the study of natural language communication between man and machine. *Communications of the ACM, 9*(1), 36–45.