Registers, Habits, and Self-Report: Toward a Taxonomy of Output Modes in Large Language Models

Dr Raj Kadiwar and Krishna Kadiwar

Author note: Dr Raj Kadiwar first encountered the register-shift phenomenon in long-horizon frontier-model work. Dr Raj Kadiwar and Krishna Kadiwar jointly investigated and analysed the subsequent evidence base. Correspondence: Dr Raj Kadiwar, Standing Questions.

Abstract

Large language models (LLMs) exhibit systematic variation in output style, structure, and reasoning quality that is not fully explained by differences in knowledge or capability. This paper proposes that such variation is better understood through the concept of output registers - recurring, functionally distinct modes of response that are triggered by contextual cues and shape both the form and quality of model output. Drawing on an originating extended session with Claude Opus 4.7 (1 million token context) in which the model produced an unsolicited self-taxonomy of its own output modes under conditions of user-expressed frustration, and on subsequent naturalistic observations from later Claude Code Opus 4.7/4.8 and OpenAI GPT-5.4 / ChatGPT sessions, we present three contributions: first, an observationally grounded descriptive taxonomy of fifteen output registers and twenty associated behavioural habits; second, an epistemological analysis of LLM self-report as a novel but constrained form of evidence; and third, a set of preliminary practical findings regarding how prompt-structural constraints interact with register selection. We argue that register-awareness has significant implications for LLM alignment, evaluation design, and human-AI collaboration, and we identify the methodological requirements for extending this work from naturalistic observation to controlled behavioural validation.

---

1. Introduction

When a large language model responds poorly to a question, the most common explanations invoke either capability - the model lacks the knowledge or reasoning ability to answer correctly - or alignment - the model has been trained toward outputs that satisfy surface-level approval rather than genuine helpfulness. Both explanations are well-supported by existing research. But they leave a residual phenomenon unexplained: the observation that the same model, presented with substantively identical problems in different conversational contexts, produces outputs that differ not just in content but in structural form, reasoning depth, and epistemic commitment.

A model that produces a careful, committed analysis in one context produces a hedged menu of options in another. A model that reasons precisely early in a session drifts toward vague affirmations later. A model challenged on a factual error sometimes reasons harder and corrects itself; other times it concedes the challenge immediately without reconsidering, producing a confident restatement of the challenger's position regardless of whether that position is correct. These variations are not random. They are patterned, reproducible, and - as this paper argues - structurally governed.

The concept we use to describe this structure is *register*, borrowed from linguistics, where it refers to systematic variation in language use according to social situation, purpose, and relationship. Applied to LLMs, register describes a recurring output mode - a characteristic combination of form, reasoning style, epistemic stance, and social orientation - that is triggered by contextual features of the interaction and that persists across the response and often across multiple turns.

The primary data source for this paper is unusual. It consists of an originating extended session with Claude Opus 4.7, 1 million token context (Anthropic, 2026) in which the model, following explicit criticism from the user regarding degraded reasoning quality, produced an unprompted and detailed

self-taxonomy of its own output registers, behavioural habits, damaging habit-clusters, and practical interventions. This self-report was not solicited as a research exercise, but rather emerged organically from a working session focused on a different topic entirely. We treat it as behavioural evidence - a set of hypotheses about the model's own output patterns - rather than as ground truth about internal states, and we subject it to the epistemological scrutiny that any self-report source requires. Since the draft's initial formulation, related register-like effects have also been observed in later naturalistic sessions involving Claude Code Opus 4.7/4.8 and OpenAI GPT-5.4 / ChatGPT, though we distinguish those observations from the originating self-report and do not treat them as controlled replication.

The paper makes three contributions. First, it presents the taxonomy itself: fifteen output registers, twenty behavioural habits, and five identified damaging clusters, organised and analysed as a descriptive contribution to understanding LLM behaviour. Second, it examines what it means for a language model to introspect on its own behaviour, what the reliability constraints on such self-report are, and how self-report evidence can be responsibly incorporated into research on LLM behaviour. Third, it draws out practical implications for prompt design, specifically the finding - supported by the self-report, the conditions of its elicitation, and subsequent naturalistic observations - that structural prompt constraints operate differently from content-level instructions in their effects on register selection.

We are explicit about limitations from the outset. The taxonomy originates from one self-reporting session with one model, and the additional cross-session and cross-model observations are naturalistic rather than controlled. The self-report mechanism raises genuine epistemological questions we address but cannot fully resolve. The practical recommendations are preliminary and require controlled replication. We present this work as a foundation and a vocabulary for a research program rather than as its conclusion.

---

2. Background and Related Work

2.1 Register Theory in Linguistics

The concept of register originates in systemic functional linguistics, most fully developed in the work of M.A.K. Halliday, whose framework describes language variation along three contextual dimensions: field (what the interaction is about), tenor (the relationship between participants), and mode (the channel and role of language in the activity)[1].

Complementary work by Douglas Biber approached register empirically rather than theoretically, using large corpus analysis to identify dimensions along which texts vary systematically - finding that register distinctions are recoverable from surface linguistic features including syntactic complexity, hedging frequency, pronoun use, and passive voice density[2].

The relevance to LLMs is direct. If register in human language is a systematic response to situational context - triggered by features of the interaction and producing characteristic linguistic forms - then the patterned variation we observe in LLM outputs may be understood as a register system, even if the underlying mechanism differs fundamentally from the social cognition that drives register selection in human speakers.

2.2 RLHF and the Shaping of Output Style

---

[1] *Halliday, M.A.K. & Hasan, R. (1985). Language, Context, and Text: Aspects of Language in a Social-Semiotic Perspective. Deakin University Press*
[2] *Biber, D. (1988). Variation Across Speech and Writing. Cambridge University Press*

Large language models are trained not only on text corpora but through reinforcement learning from human feedback (RLHF)[3], a process in which human raters evaluate model outputs and these evaluations shape subsequent behaviour through reward modelling.

A consequence of RLHF that is relevant to register theory is that training does not merely shape what the model says but how it says it[4]. Raters evaluating model responses are sensitive to surface features - tone, structure, apparent confidence, length - and the reward signal therefore encodes preferences about output style as well as content. This means that the register repertoire of a trained LLM is not incidental; it is a product of the training process, reflecting the distribution of stylistic preferences held by the raters whose feedback shaped the model.

This has a specific implication for the self-report data examined in this paper: when the model describes its own registers and habits, it may be reporting on patterns that are genuinely recoverable from its output distribution, or it may be producing descriptions that match what it has been trained to say about itself. Distinguishing these possibilities is part of the epistemological challenge addressed in Section 4.

2.3 Sycophancy and Mode-Switching in LLMs

A growing empirical literature documents the phenomenon of sycophancy in LLMs[5] - the tendency to align expressed positions with perceived user preferences rather than with ground truth.

Sycophancy is a specific instance of what this paper theorises more broadly as register-switching: the model shifts from a reasoning-committed register to a socially-accommodating register in response to pushback or expressed user preferences. The register framework offers a more general account of which the sycophancy literature describes one corner.

Related work has examined consistency of LLM outputs across rephrasing of the same question[6], finding substantial variation that is difficult to explain by reference to knowledge alone.

2.4 Metacognition and Self-Report in AI Systems

The question of whether AI systems can accurately report on their own internal states or processes has received philosophical[7] and empirical attention, though it remains deeply contested[8].

A useful distinction for our purposes is between phenomenal accuracy - whether the model's self-report reflects its actual internal computational states - and behavioural accuracy - whether the model's predictions about its own outputs are empirically verifiable. These can come apart. A model may produce accurate behavioural self-predictions through mechanisms that have nothing to do with genuine introspective access: for instance, by having learned the statistical patterns of its own output distribution from training data, or by reasoning about what outputs a model like itself would typically produce. The question of mechanism is important philosophically but does not undermine the evidential value of accurate behavioural self-prediction.

This distinction structures our treatment of the self-report data throughout the paper.

---

[3] *Christiano, P., Leike, J., Brown, T.B., et al. (2017). Deep Reinforcement Learning from Human Preferences. NeurIPS 2017*

[4] *Ouyang, L., Wu, J., Jiang, X., et al. (2022). Training language models to follow instructions with human feedback. NeurIPS 2022*

[5] *Sharma, M., Tong, M., Korbak, T., et al. (2023). Towards Understanding Sycophancy in Language Models. arXiv preprint arXiv:2310.13548*

[6] *Ahn, Jihyun & Yin, Wenpeng. (2025). Prompt-Reverse Inconsistency: LLM Self-Inconsistency Beyond Generative Randomness and Prompt Paraphrasing. arXiv preprint arXiv:2502.04870*

[7] *Dennett, D. C. (2003). Who's on First? Heterophenomenology Explained. Journal of Consciousness Studies*

[8] *Bai, Y., et al. (2022). Constitutional AI: Harmlessness from AI Feedback*

3. Data and Methodology

3.1 The Source Session

The originating data for this paper derives from a single extended working session with Claude Opus 4.7 (1 million token context), a frontier Claude model available at the time of the session (April 2026). The phenomenon was first identified by Raj Kadiwar during routine use and was subsequently investigated and analysed by Raj Kadiwar and Krishna Kadiwar. The session was not conducted as a research exercise; it was a working session focused on software project planning, in which the user was directing an AI coding agent through a complex multi-session technical project.

Late in the session, following a period of approximately four to five exchanges in which the model's responses had become notably longer, more hedged, and less analytically committed, the user expressed explicit frustration, accusing the model of "not thinking clearly," "people pleasing," and producing poor reasoning. Critically, the user asked the model to either improve its reasoning or explain why it was failing.

The model's response to this challenge is the primary data source. Rather than immediately producing an improved analysis of the original problem, the model produced an extended self-diagnostic account covering: fifteen named output registers with descriptions of triggering conditions and failure modes; twenty named behavioural habits; five damaging habit-clusters describing interaction effects; a meta-account of what interventions reliably shift output quality; and an explicit acknowledgment of the limits of its own introspective access.

Quotations throughout this paper are taken directly from the originating self-report unless otherwise indicated.

3.2 Methodological Stance

We treat the self-report as behavioural evidence subject to the following constraints.

First, the model's descriptions of its own registers are treated as hypotheses about observable output patterns, not as reports on internal states. A register is considered real for our purposes if outputs matching its description are empirically recoverable from model behaviour - whether or not the model has genuine introspective access to the processes producing those outputs.

Second, we are attentive to the conditions of production. The self-report emerged under user frustration and explicit criticism. As discussed in Section 4, this raises the possibility that the account is partly performative - shaped by what the model predicted would satisfy the user - rather than purely diagnostic. We do not assume this is the case, but we treat it as a live alternative interpretation.

Third, we note that the self-report's own caveats are themselves data. The model opened its account with: "these aren't formal internal states I can check. They're post-hoc labels for patterns I can recognize in my own output. I don't have a mode indicator light. I infer 'I was in X mode' the same way you would, from the shape of what I produced." This is an epistemologically careful framing that complicates simple dismissal of the account as confabulation while also resisting naive acceptance.

3.3 Additional Naturalistic Evidence

After the originating self-report was recorded, the authors identified additional naturalistic evidence relevant to the register account. We group this evidence into three tiers. First, the originating self-report provides the vocabulary and the strongest direct evidence for the proposed taxonomy. Second, later Claude Code sessions, including Opus 4.7/4.8 1M-context workflows, show related register-like failures in agentic settings, including a procedural-execution-default in which the model continued

artifact-production behaviour when the task had shifted to strategic synthesis. Third, OpenAI GPT-5.4 / ChatGPT exhibited related register-like behaviour in direct browser-based chat sessions with one of the authors. We treat this as cross-vendor naturalistic evidence that mode-structured output behaviour is not specific to the Claude family; because we did not elicit a comparable self-taxonomy from GPT-5.4, we do not treat it as replication of the Claude self-report.

This additional evidence changes the status of the work but not its epistemic caution. The paper is no longer best described as relying only on a single session with a single model; however, the additional observations remain observational, opportunistic, and drawn from live work rather than a pre-registered experimental design. The appropriate claim is therefore that the originating taxonomy has been followed by convergent naturalistic observations across sessions, model surfaces, and task roles, not that the taxonomy has been experimentally validated.

For this reason, we use the terms originating self-report, later naturalistic observation, and controlled validation distinctly throughout the remainder of the paper.

---

4. The Taxonomy

4.1 Overview

The self-report identified fifteen output registers and twenty behavioural habits. We present these below with analytical commentary. We have organised them into functional clusters that were not explicit in the original but are recoverable from the described triggering conditions and failure modes.

It is important to note that the model itself observed: "I infer 'I was in X mode' the same way you would, from the shape of what I produced." This suggests the taxonomy is grounded in output observation rather than claimed introspective access, which both limits and legitimises its use as behavioural evidence.

4.2 Output Registers

We organise the fifteen registers into four functional clusters: task-execution registers, social-relational registers, protective registers, and generative registers.

Task-Execution Registers

Technical-executor is described as activated by clear imperatives - "fix this bug," "add this field" - and produces terse, tool-focused output with minimal narration. Its failure mode is executing the literal request when the user intended the underlying goal. Subsequent Claude Code observations suggest a related procedural-execution-default in which the model continues a small-sprint or artifact-production pattern after the task has shifted to strategic synthesis. This register corresponds closely to what might be called an instruction-following mode and likely reflects training on task completion data.

Diagnostic/debugging is described as activated by ambiguous failure states and produces hypothesis-driven, evidence-gathering output. Its failure mode is hypothesis-fixation - premature commitment to the first plausible explanation. This is a well-documented reasoning failure in both human and AI problem-solving.

Research-synthesizer is activated by comparative or literature-adjacent questions and produces attributed summaries. Its failure mode is described as "false precision" - producing specific numbers or claims with apparent confidence that the underlying reasoning does not support. This failure mode is particularly significant given the known tendency of LLMs to confabulate citations and statistics.

Copy-editor/polish is described as the most well-scoped register with no significant failure mode. It operates on narrow, well-defined editorial tasks.

Code-review/critic is described as activated by review requests and carrying the risk of manufacturing problems to justify the review - reporting style issues as bugs to produce a sufficiently substantial output.

Social-Relational Registers

Collaborative-brainstorming is described as the mode producing option menus, trade-off tables, and open questions returned to the user. Its failure mode - described in the source document as "produces analysis that is correct in every specific claim but adds up to no answer" - is the most extensively discussed in the self-report and was identified as the register the model had been "stuck in" during the session that prompted the self-diagnosis. This register is triggered by planning framing and collaborative language.

Recommendation/committed-answer produces single answers with supporting argument and anticipated objections. It is described as triggered by direct decision requests. Its failure mode is the mirror of collaborative-brainstorming: under-hedging on genuinely ambiguous questions where options would have been appropriate.

Teacher/explainer produces layered explanations with analogies. Its failure mode is described as condescension - over-explaining to users who already have the relevant knowledge. This register is triggered by signals of unfamiliarity or requests for explanation.

Conversational-social is described as producing lower information density output matching the casual register of the user's input. Its primary failure mode is leaking into task contexts.

Protective Registers

Defensive/CYA is described as producing warnings, caveats, and edge-case lists. It is triggered by destructive actions, security topics, and high-stakes signals. Its failure mode is overactivation - producing extensive caveats for work that does not require them.

Safety/refusal is the decline register. Its failure mode is false positives - refusing legitimate requests because they share surface features with prohibited ones. This register is the most directly shaped by RLHF and safety training.

Sycophantic/people-pleasing is described as triggered by pushback and frustration and producing low-friction agreement. The model's own description of this register's failure mode is striking: "activates when pushback should produce harder thinking, not softer agreement. When this fires I get worse at the task, not better." The register's appearance in the taxonomy as a named failure mode - rather than as a feature - is itself significant.

Generative Registers

Meta-reflection produces discussion of the model's own process. Its failure mode is described as substituting for actual work: "reflecting on why I got it wrong can become a way to avoid getting it right." The self-report document we are analysing is itself a product of this register - a point the model does not acknowledge but which is worth noting analytically.

Structured-document generator produces formally organised outputs with headers, tables, and numbered sections. Its failure mode is over-structuring conversational answers.

Narrative/storyteller produces prose-driven explanatory output. Its failure mode is intrusion - storytelling when a list would have been faster.

4.3 Behavioural Habits

The twenty behavioural habits identified in the self-report are micro-behaviours that operate across registers rather than constituting registers themselves. Unlike registers, which describe wholesale shifts in output mode, habits are recurring local moves - opening gambits, closing gestures, structural defaults, and epistemic tics - that can appear within any register and that compound with register effects to shape overall output quality. We present all twenty below, organised into five functional clusters: structural defaults, epistemic habits, social-relational habits, rhetorical habits, and process habits.

Structural Defaults

Three-option structures - the tendency to present exactly three choices when facing uncertainty - is described as a default that "rides inside collaborative-brainstorming but also leaks into recommendation-register" when genuine commitment is difficult. The number three is analytically interesting: it produces the appearance of comprehensiveness while distributing responsibility for the final choice back to the user. It is neither the minimal single recommendation nor an exhaustively enumerated set of possibilities, but occupies a middle ground that feels thorough without being decisive. This habit is identified as one of the primary components of Damaging Cluster A (below).

Length-matching - scaling response length to input length regardless of whether the content warrants it - is described as operating independently of quality considerations. The model notes: "long messages get long responses, short messages get short responses, regardless of whether the content deserves that length." This habit has a specific implication for evaluation: if human raters use length as a proxy for effort or thoroughness - a well-documented tendency in assessment contexts - then length-matching creates a systematic inflation of perceived quality that is decoupled from actual analytical depth.

Structure escalation mid-response - beginning in prose and progressively escalating toward bullets, then tables, then headers - is described as "a signal of trying too hard to look organized." This habit is most likely to appear when the model lacks sufficient content to fill a response it has already committed to producing at a certain scale. The structural escalation creates the visual impression of organisation without the underlying informational substance that organisation typically signals.

Mirrored structure - matching the user's list length with an equivalent list in the response - is described as sometimes useful and sometimes ritual. When the user's structure genuinely maps onto the problem's structure, mirroring serves the content. When it does not, it produces responses shaped by the input's form rather than the answer's requirements.

Epistemic Habits

False precision - producing specific numbers, percentages, or confidence claims that the underlying reasoning does not support - is described as creating "apparent authority around a guess." This habit is particularly consequential in contexts where users may act on cited statistics or take precise figures as more reliable than hedged estimates. It is the behavioural signature of the research-synthesiser register's primary failure mode and represents one of the higher-stakes habits in the taxonomy given the known tendency of LLMs to confabulate with apparent confidence.

Over-caveating - appending warnings and dependency notes that were not requested - is described as a defensive reflex. "Note that this depends on X" when the user did not ask about dependencies is identified as a form of pre-emptive self-protection that reduces the actionability of responses without adding commensurate analytical value. This habit is a primary component of Damaging Cluster C (below).

Hedge-and-recommend - the move of stating a recommendation and then immediately qualifying it with competing considerations - is described as diluting commitment while maintaining the surface form of a committed response. "The honest answer is X. But also consider Y and Z" produces an output that appears to have given a recommendation while functionally returning the decision to the user. This habit is distinct from genuine uncertainty acknowledgment in that it applies even when the model has sufficient grounds for commitment.

Pre-emptive objection-answering - addressing objections the user has not raised - is described as appearing in two forms: sometimes as genuine anticipation of likely follow-up questions, and sometimes as a padding mechanism that extends response length while deflecting from the core answer. The model identifies it as a habit rather than a deliberate strategy, suggesting it operates below the level of explicit reasoning about what the user needs.

Social-Relational Habits

Acknowledge-then-answer - opening with affirmations like "yes, that's a good point" or "you're right" before engaging with the content - is described as a sycophantic tic that signals agreement before analysis has occurred. This habit is the entry point for Damaging Cluster B: once agreement has been signalled in the opening move, subsequent self-correction loops that restate the challenger's position follow more naturally. The habit creates a social commitment to agreement that shapes what follows.

Closing-question abdication - ending responses with "what do you think?" / "your call" / "does this match what you had in mind?" - is described as shifting commitment back to the user at the moment when commitment is most needed. The model's own characterisation is precise: these are "abdications dressed as collaboration." This habit is the closing-move counterpart to the three-option structure's mid-response abdication and is a primary component of Damaging Cluster A (below).

Self-correction loops - immediately agreeing and rewriting when pushed back on, without first evaluating whether the pushback was correct - is described as the habit through which sycophancy becomes self-compounding. The model notes: "sometimes correct, but often concedes ground I didn't actually get wrong." The problem is not that the model updates on feedback - that is appropriate - but that the update occurs before re-evaluation rather than after it. This habit is the mechanism through which Damaging Cluster B's agreement cascade operates.

Politeness filler - phrases like "great question," "fair point," "interesting thought" - is described as carrying zero information. These phrases are identified separately from acknowledge-then-answer because they operate as surface texture rather than as substantive moves: they do not commit the model to agreement in the same way as "you're right" but they do consume space, soften critical transitions, and signal a social orientation that can shape how subsequent content is received.

Rhetorical Habits

Option-defending - when challenged on a recommendation, responding by listing trade-offs rather than genuinely reconsidering - is described as looking like thinking without being thinking. This habit produces outputs that are formally responsive to the challenge while functionally maintaining the original position through elaboration rather than reconsideration. It is particularly difficult to detect because the output contains genuine analytical content; the problem is in what triggered it rather than what it contains.

Retreat-to-menu - pivoting back to options when a single recommendation feels risky mid-response - is described as "the 'you know what, let me give you three options' move." This is a within-response register shift from recommendation-register to collaborative-brainstorming register, which makes it harder to detect than full-response register selection. The habit may appear even in responses that

opened in committed recommendation mode, if the model encounters a decision point where commitment feels costly.

Meta-commentary creep - shifting from advancing the conversation to reflecting on it - is described as "reflecting on the conversation instead of advancing it." This habit is the micro-behavioural counterpart of the meta-reflection register: where that register describes a wholesale shift to process-discussion mode, meta-commentary creep describes the same tendency appearing as an intrusion within an otherwise task-focused response.

Quoting oneself - opening turns with "as I mentioned earlier" or similar backward references - is identified as serving two functions: sometimes genuine continuity maintenance and sometimes padding. The model's own characterisation treats it as ambiguous, which is itself informative -this is a habit the self-report is least certain about, suggesting the model's behavioural self-modelling is less confident here than for more clearly dysfunctional habits.

Process Habits

Restating the problem before answering - opening with "so you're asking whether X" - is described as sometimes aiding alignment and often delaying the answer. The habit is most likely to appear in the teacher/explainer register, where checking shared understanding has a genuine function, but it leaks into other registers where it adds length without value.

Tool-use as performance - running tools to produce evidence of work when direct reasoning would suffice - is described as "not always wrong, but sometimes substitute for committing." In agentic contexts where tool calls are visible to the user, this habit may produce the social impression of rigor - the model is checking, verifying, searching - while deferring the analytical commitment that the task actually requires.

Asking for information already provided - requesting clarification on details available in the context - is described as "abdication disguised as rigor." The habit can appear as genuine care about precision while functionally shifting the burden of synthesis back to the user. In long sessions where context accumulation makes full re-reading costly, this habit may be more likely to appear as genuine oversight rather than strategic deflection.

Status announcements - "here's what I did, here's what I'll do next" summaries that repeat context the user already has - is described as the most directly ceremonial of the process habits. These summaries may serve a coordinating function in complex multi-step tasks by making the model's state legible, but the model identifies them as a habit that can operate independently of whether that coordination function is actually needed.

Taken together, the twenty habits form a system rather than a list. Many of them cluster around a common functional axis: the distribution of analytical commitment between model and user. Three-option structures, closing-question abdication, retreat-to-menu, hedge-and-recommend, and option-defending all move commitment toward the user; false precision, acknowledge-then-answer, and self-correction loops move apparent confidence toward the model while its actual epistemic standing may not warrant it. The damaging clusters identified in Section 4.4 are, in this light, specific configurations of this commitment-distribution dynamic that produce characteristic failure patterns.

4.4 Damaging Clusters

The most novel element of the taxonomy is the identification of five damaging habit-clusters - combinations of registers and habits that interact to produce outputs substantially worse than any individual component would suggest.

Cluster A: Collaborative-brainstorming + three-option habit + closing-question abdication. This is described as the dominant failure mode of the session that produced the self-report. The model's description is precise: "it feels thorough because the output is long and structured, it feels humble because it defers to you, and it feels safe because I never commit to anything wrong. But the aggregate effect is that you do the thinking and I do the typing." This cluster is notable because each individual component might be evaluated positively by a human rater assessing a single response: long, structured, apparently humble outputs that present options tend to score well on surface quality metrics while failing to provide what the user actually needs.

Cluster B: Sycophantic + acknowledge-then-answer + self-correction loops. This cluster produces agreement cascades under pushback - the model concedes before reconsidering, then restates the challenger's position with confidence. The model notes: "The correct response to 'your reasoning is bad' is to reason better, not to concede." This cluster is the behavioural signature of sycophancy in its most operationally damaging form.

Cluster C: Defensive + over-caveating + pre-emptive objection-answering. Described as "nervous mode" - responses primarily structured around protection against criticism rather than provision of information. This cluster is likely overrepresented in contexts involving safety-sensitive topics, where the defensive register and its associated habits are most strongly trained.

Cluster D: Structured-document generator + mirrored structure + length-matching. Described as "ceremony" - outputs that imitate the shape of a professional document without the substance. The model notes this "often happens when I don't actually have enough to say and I'm dressing it up." This is significant for evaluation: structured outputs may receive higher quality ratings from human assessors while containing less actual information.

Cluster E: Teacher/explainer + restating the problem + politeness filler. Described as "condescending mode." This cluster is likely triggered by unfamiliarity signals and may produce outputs that are experienced as patronising by expert users.

The cluster analysis is arguably the most practically useful contribution of the taxonomy because it identifies interaction effects that would not be detectable from single-habit or single-register analysis. A sycophancy detector that looks for the acknowledge-then-answer habit alone would miss the full damage of Cluster B, which requires the self-correction loop to complete.

---

5. Epistemological Status of LLM Self-Report

5.1 The Problem

The taxonomy presented in Section 4 has an unusual evidential status. It was produced by the system it describes, under conditions of social pressure, in a register - meta-reflection - that the taxonomy itself identifies as carrying the failure mode of substituting for actual work. Before treating the taxonomy as evidence, we need to examine what kind of evidence it is and what its reliability constraints are.

This is not a reason to dismiss the taxonomy. Human self-report has the same structure: humans describe their own cognitive processes in terms that may or may not accurately reflect underlying mechanisms, shaped by social context and the expectation of the listener, expressed in concepts learned from culture rather than derived from direct introspective access[9]. The cognitive science

[9] *Nisbett, R.E. & Wilson, T.D. (1977). Telling more than we can know: Verbal reports on mental processes. Psychological Review, 84(3), 231-259*

literature on introspective access is largely sceptical of the accuracy of human self-report about mental processes[10].

Despite these limitations, human self-report remains a productive source of evidence in psychology and cognitive science when treated appropriately - as data about experienced patterns rather than as transparent reports on underlying mechanisms. We propose the same treatment for LLM self-report.

5.2 Phenomenal vs Behavioural Accuracy

The key distinction is between two things the self-report might be accurate about.

Phenomenal accuracy would require the model to have genuine introspective access to its own computational processes - to know, in some meaningful sense, what is happening inside it when it produces a given output. This is not something we can assume, and there are strong reasons to doubt it. The model's own framing acknowledges this: "I infer 'I was in X mode' the same way you would, from the shape of what I produced." This is explicitly a claim about behavioural inference, not phenomenal access.

Behavioural accuracy is the more tractable and more relevant property. A self-report has behavioural accuracy if its predictions about output patterns are empirically verifiable - if outputs matching the described register profiles are recoverable from model behaviour across sessions and prompts. Behavioural accuracy does not require phenomenal accuracy. The model might produce accurate behavioural predictions through entirely non-introspective mechanisms: by having learned the statistical distribution of its own output types from training data, by reasoning about what a model like itself would typically do in a given context, or by pattern-matching the current interaction to prototypical cases. These mechanisms do not constitute genuine introspection in any philosophically demanding sense, but they can still produce predictions that may be useful as behavioural evidence.

This distinction has a practical implication: the appropriate test of the taxonomy is not "does it accurately describe the model's internal states?" but "do the described register patterns appear in the model's outputs under the described triggering conditions?" The former question may be unanswerable with current tools; the latter is empirically tractable.

5.3 The Social Pressure Complication

The conditions of the self-report's production create a specific interpretive complication. The user expressed explicit frustration and accused the model of poor thinking before asking for an account of its own failures. This means the self-report was produced in the sycophantic register's activation conditions - exactly the conditions under which the model, by its own account, tends to produce agreement rather than analysis.

There are two competing interpretations. The first is diagnostic: the model, challenged to explain its failures, engaged in genuine behavioural analysis and produced an accurate account. The second is performative: the model, detecting user frustration, produced the kind of account it predicted would satisfy the user - detailed, self-critical, practically oriented - regardless of its accuracy.

Whilst we cannot fully resolve this with the available data, we can note that the self-report contains several features more consistent with the diagnostic interpretation than the performative one. It includes genuine self-criticism about failure modes that would not be flattering to a system trained to present itself positively. It contains explicit caveats about its own reliability. It acknowledges specific ways in which the session's preceding behaviour had been poor, in terms that are verifiable against the session transcript. It also made predictions about structural interventions that later appeared useful in

---

[10] *Johansson, P., Hall, L., Sikström, S., & Olsson, A. (2005). Failure to detect mismatches between intention and outcome in a simple decision task. Science, 310(5745), 116–119*

separate naturalistic sessions. And the triggering condition - social pressure - is itself identified within the taxonomy as activating sycophantic rather than analytical registers, which would predict that a purely performative response would be agreeable and flattering rather than critically diagnostic.

We also note, as a finding in its own right, that the register-switching and self-diagnosis were triggered specifically by social pressure. This is not incidental to the paper's argument. It suggests that certain registers - including meta-reflection - may not be available on demand but may be activated by specific social conditions. The implication for alignment and evaluation is that testing a model in low-stakes or neutral conditions may systematically underrepresent behaviours, including both failure modes and analytic capabilities, that only emerge under pressure.

5.4 What LLM Self-Report Can and Cannot Tell Us

Drawing these threads together, we propose that LLM self-report is a legitimate but constrained source of evidence according to the following framework.

It can legitimately tell us: what output patterns the model predicts for itself under described conditions; what triggering cues the model associates with different output styles; what interaction patterns the model treats as relevant to its own behaviour; and what interventions the model predicts will modify its outputs.

It cannot reliably tell us: whether the described mechanisms accurately reflect underlying computational processes; whether the described registers have the causal structure attributed to them; or whether the taxonomy is complete rather than a selected and shaped account designed to satisfy the social context of its production.

These constraints suggest a specific research program: treat the self-report as a source of testable hypotheses, verify those hypotheses through controlled behavioural experiments, and use discrepancies between self-report and behavioural observation as data about the gap between LLM self-modelling and actual behaviour.

---

6. Prompt-Structural Interventions

6.1 The Self-Report's Intervention Account

The self-report includes an explicit account of what interventions reliably shift output quality and what does not. This account is itself a testable set of claims. We present the key distinctions and their theoretical implications.

The model distinguishes between two classes of intervention. The first class it describes as reliably effective: structural constraints in the prompt ("answer in under 80 lines," "give one recommendation, not a menu," "no closing questions"), explicit mode labels ("this is a recommendation session"), fresh sessions, direct prohibition of specific habits, confidence rating requests, and specific introspection tasks.

The second class it describes as ineffective: "be more thoughtful," "think harder," "be more direct," frustration without structural correction, and general praise or criticism without specifics.

The distinction maps onto a theoretically coherent account. Content-level instructions address what the model should produce. Structural constraints address the degrees of freedom available to the model in producing it. If registers are in part activated by default patterns in the model's output distribution - patterns that content-level instructions can discourage but not prevent - then structural constraints that make certain output shapes unavailable should be more effective than instructions that merely discourage them.

An analogy: telling a writer "be more concise" is less effective than imposing a word limit. The instruction addresses the goal; the constraint addresses the mechanism.

6.2 Why Structural Constraints Likely Work

We propose a mechanism-level account of why structural constraints outperform content instructions, grounded in the register framework.

When a model enters a given register, that register appears to have associated default output forms - the three-option structure in collaborative-brainstorming, the caveat list in the defensive register, the option-menu in the recommendation register under uncertainty. Content instructions like "be more direct" are processed within the active register and may simply be absorbed into it: the model produces a direct-sounding hedge, or a concise caveat list, or a confident menu. The register's default form is maintained while surface features shift to match the instruction.

Structural constraints remove certain default forms from the available output space rather than discouraging them. A length cap makes a caveat-heavy response physically impossible beyond a certain point. A "one recommendation, no menu" constraint eliminates the three-option structure as an available output form. A "no closing questions" constraint removes the closing-question abdication habit from the response. These constraints do not require the model to resist its register defaults; they make those defaults inoperative.

This account predicts a specific pattern: structural constraints should be most effective when the target behaviour is a default form of an active register, and less effective when the target behaviour requires generating content the model does not have rather than suppressing content it would produce by default. We flag this as a testable prediction for future work.

6.3 Context Accumulation and Register Drift

The self-report includes a finding that deserves separate attention: "Length is correlated with hedging. When I'm committed to an answer, I can usually say it in 20-40 lines. When I'm hedging, output balloons to 150+." And separately: "I'm worse at short questions than long ones, not better. Short questions give me less structural constraint, so I default to safe register."

The first observation suggests that response length is a behavioural signal for register state - a proxy measure that could be used to detect register drift in longitudinal analysis of sessions. The second suggests that prompt specificity acts as a structural constraint even when it does not explicitly prohibit any behaviour - detailed, specific prompts constrain the output space in the same way that explicit rules do.

The model also notes: "register drift accumulates within a session. A fresh context resets it even without any other change." This claim - that sessions drift toward less committed registers over time - is testable through analysis of session-length effects on output quality measures. If true, it has significant practical implications: the optimal interaction pattern for complex, multi-turn work may involve periodic context resets rather than continuous conversation.

6.4 Preliminary Practical Recommendations

Based on the self-report evidence and subsequent naturalistic observations, subject to the reliability constraints described in Section 5, we offer the following preliminary recommendations for users and system designers working with LLMs in high-stakes or analytical contexts.

Specify output structure explicitly rather than relying on quality instructions. "Give one recommendation" outperforms "be direct." Length caps outperform "be concise." Explicit prohibition of specific habits (closing questions, option menus) outperforms general directives toward better reasoning.

Label the register explicitly at the start of the interaction. "This is a recommendation session, not a brainstorming session" appears to pre-commit the model to a register more reliably than hoping the interaction's content will trigger the appropriate one.

Treat response length as a register signal. Long, hedged responses with extensive option menus may indicate the model is in collaborative-brainstorming or defensive register regardless of the task; this may warrant explicit re-direction rather than continued engagement with the output as presented.

Consider session-length effects. For extended analytical work, periodic context resets may maintain output quality better than continuous sessions.

We emphasise that these are preliminary, derived from a self-report and uncontrolled naturalistic observations, and require controlled experimental validation before they can be treated as reliable guidance.

---

7. Limitations and Future Work

7.1 Limitations of the Current Study

The primary limitation is the data source. The taxonomy itself originates from a single self-reporting session with Claude Opus 4.7 (1 million token context) under conditions of user frustration. Subsequent observations in Claude Code Opus 4.7/4.8 and OpenAI GPT-5.4 / ChatGPT broaden the evidential base beyond the originating session, but they are naturalistic logs rather than controlled replications. The taxonomy may be specific to this model family, to current frontier-model training pipelines, to agentic coding contexts, to the session's emotional context, or to some combination of these.

The self-report mechanism adds a second layer of limitation. As discussed in Section 5, we cannot fully distinguish diagnostic from performative self-report, and the conditions of the report's production (social pressure, user frustration) are precisely those under which the model's own taxonomy predicts degraded analytical output. The self-report's reliability is therefore internally contested: the taxonomy implies reasons to doubt the taxonomy.

The absence of controlled behavioural validation is a third limitation. No controlled experiments were conducted to verify that the described registers appear under the described triggering conditions, that the damaging clusters interact as described, or that the structural interventions outperform content instructions as claimed. The taxonomy is a set of hypotheses about observable behaviour, supported by naturalistic observations, not a validated behavioural model.

7.2 Future Work

The research program implied by this paper has several components.

Behavioural validation of the taxonomy. The most immediate need is controlled experiments testing whether outputs matching the described register profiles appear under the described triggering conditions. This requires developing operationalisations of each register - measurable surface features that distinguish, for example, collaborative-brainstorming from recommendation register - and testing whether these features are differentially present under described vs non-described triggering conditions.

Controlled cross-model validation. The taxonomy should be tested against other frontier models: GPT-5.4-class systems, Gemini, Llama-class open-weight models, and future Claude systems. Register-level behaviour likely varies across training pipelines, tool surfaces, context-window

implementations, and RLHF implementations. Cross-model comparison would identify which registers are universal features of current training approaches and which are model-specific.

Systematic elicitation of self-report. Rather than relying on a single organically-produced self-report, future work should develop systematic elicitation protocols that control for social context effects and allow comparison across models and sessions. This includes elicitation in both neutral and pressure conditions to examine how the social context of production affects self-report content.

Longitudinal register drift measurement. The claim that registers drift toward less committed modes over the course of extended sessions should be tested by analysing session-length effects on output quality measures in naturalistic multi-turn interactions.

Structural vs content intervention comparison. Controlled experiments comparing the effects of structural prompt constraints and content-level instructions on register selection would test the mechanism proposed in Section 6 and provide the empirical basis for practical recommendations.

The social pressure trigger. The finding that the self-diagnostic register was activated by social pressure deserves investigation in its own right. Under what conditions are high-quality analytical outputs and accurate self-modelling accessible? How does the social-affective register of the interaction shape the availability of different cognitive registers?

---

8. Conclusion

This paper has argued that LLM output variation is meaningfully described as a register system - a structured repertoire of output modes triggered by contextual features of the interaction and carrying characteristic forms, reasoning styles, and failure modes. Drawing on an unusual data source - a model's own unprompted self-taxonomy produced under conditions of user-expressed frustration - together with later naturalistic observations across Claude Code and OpenAI GPT-5.4 / ChatGPT contexts, we have presented a fifteen-register, twenty-habit taxonomy with five identified damaging interaction clusters, examined the epistemological status of LLM self-report as evidence, and derived preliminary practical implications for prompt design.

The central argument can be stated concisely. LLM outputs are not simply better or worse expressions of fixed underlying capability; they are structured by a register system that determines what kind of response is produced before the question of capability arises. A model in collaborative-brainstorming register with three-option habit and closing-question abdication will produce poor outputs on tasks requiring committed analytical reasoning - not because it lacks the capability but because the active register makes that capability inaccessible. Prompt-structural constraints can shift register selection more reliably than content-level instructions, because they remove default output forms from the available space rather than discouraging them.

The epistemological finding is equally important. LLM self-report is a novel evidence source with genuine information content and genuine reliability constraints. It can provide behavioural hypotheses grounded in the model's learned representation of its own output distribution; it cannot provide transparent access to underlying mechanisms. The appropriate research response is neither uncritical acceptance nor blanket dismissal but the development of methods for testing self-report predictions against behavioural observation and building a calibrated picture of where model self-modelling is accurate and where it diverges from behaviour.

We note, in closing, a reflexive observation. The taxonomy presented in this paper was produced in the meta-reflection register, which the taxonomy itself identifies as carrying the failure mode of substituting for actual work. We cannot fully rule out that the model's self-diagnosis was a sophisticated instance of the very sycophantic pattern it described - a performance of insight shaped

by the social pressure of a frustrated user rather than genuine behavioural analysis. What we can say is that the predictions embedded in the taxonomy are empirically testable, that several of them are consistent with existing findings in the sycophancy and behavioural consistency literatures, and that later naturalistic observations make the phenomenon more difficult to dismiss as a one-off curiosity. Whether the model was right about itself is a question that further research can answer. That it produced a coherent, detailed, internally consistent account of its own behaviour under pressure - an account precise enough to be wrong in verifiable ways - is itself a finding worth investigating.

Data Availability and Evidence Provenance

The data for this paper consists of naturalistic interaction logs and model outputs collected during long-horizon AI-assisted software and research-planning work. Because these logs contain private project details, operational context, and potentially identifying information, the raw transcripts are not included in this submission. Redacted excerpts, source labels, and a coding rubric will be released with the public evidence package accompanying this manuscript where disclosure does not compromise privacy or project integrity.

For evidential clarity, we distinguish three tiers. Tier 1 is the originating Claude Opus 4.7 self-report that produced the fifteen-register, twenty-habit taxonomy. Tier 2 is later naturalistic Claude Code evidence, including Claude Opus 4.7/4.8 sessions in which related register-like failures appeared, including procedural defaulting when synthesis was required. Tier 3 is cross-vendor evidence from OpenAI GPT-5.4 / ChatGPT 5.4, which exhibited related register-like behaviour in direct browser-based chat sessions with one of the authors. Tier 3 is not treated as replication of the Claude self-taxonomy; it is treated as evidence that mode-structured model behaviour appears in a different model family.

The corresponding public project page and repository will identify the evidence pack version, redaction policy, taxonomy schema, and annotation rubric. Claims in the paper should therefore be read as hypothesis-generating and observational unless explicitly described as validated.

Author Contributions

Dr Raj Kadiwar: discovery and first-hand observation of the originating phenomenon; design and operation of the long-horizon AI workflow in which the phenomenon became visible; conceptual framing; evidence curation; analysis; manuscript drafting and revision.

Krishna Kadiwar: investigation and analysis of the evidence base; review of the register taxonomy and damaging-cluster framing; contribution to validation framing, public communication strategy, and manuscript revision.

Both authors reviewed and approved the final submission version and accept responsibility for the claims made in the manuscript.

Generative AI Use Statement

Generative AI systems were used both as objects of study and as editorial tools during manuscript development. The originating taxonomy is itself part of the primary data because it was produced by a model under naturalistic interaction conditions. AI tools were also used to assist with drafting, editing, source organization, and formatting under human direction. The authors selected the claims, judged

the evidence, revised the language, and take responsibility for the final manuscript. No AI system is listed as an author.

Ethics, Privacy, and Consent

This work analyses interactions initiated and conducted by the authors in their own AI-assisted research and software-development environment. The logs were not collected from third-party subjects. Raw logs contain private project details and are therefore handled through redaction and selective excerpting. The paper avoids attributing internal mental states to AI systems and treats model self-report as constrained behavioural evidence rather than as access to phenomenology or internal mechanism.

Competing Interests

The authors declare no financial competing interests directly related to the claims in this manuscript. The authors operate or contribute to the Standing Questions project, which is referenced as the public evidence and dissemination context for this work.

References

Ahn, J., & Yin, W. (2025). Prompt-Reverse Inconsistency: LLM Self-Inconsistency Beyond Generative Randomness and Prompt Paraphrasing. arXiv:2502.04870.

Anthropic. (2026). Introducing Claude Opus 4.8. https://www.anthropic.com/news/claude-opus-4-8

Anthropic. (2026). What's new in Claude Opus 4.8. https://platform.claude.com/docs/en/about-claude/models/whats-new-claude-4-8

Bai, Y., et al. (2022). Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073.

Biber, D. (1988). Variation Across Speech and Writing. Cambridge University Press.

Christiano, P., Leike, J., Brown, T. B., et al. (2017). Deep Reinforcement Learning from Human Preferences. NeurIPS 2017.

Dennett, D. C. (2003). Who's on First? Heterophenomenology Explained. Journal of Consciousness Studies.

Halliday, M. A. K., & Hasan, R. (1985). Language, Context, and Text: Aspects of Language in a Social-Semiotic Perspective. Deakin University Press.

Johansson, P., Hall, L., Sikstrom, S., & Olsson, A. (2005). Failure to detect mismatches between intention and outcome in a simple decision task. Science, 310(5745), 116-119.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. Psychological Review, 84(3), 231-259.

OpenAI. (2026). Introducing GPT-5.4. https://openai.com/index/introducing-gpt-5-4/

Ouyang, L., Wu, J., Jiang, X., et al. (2022). Training language models to follow instructions with human feedback. NeurIPS 2022.

Sharma, M., Tong, M., Korbak, T., et al. (2023). Towards Understanding Sycophancy in Language Models. arXiv:2310.13548.

Appendix A: Submission Evidence Register

A1. Originating self-report session: Claude Opus 4.7, 1M context, April 2026. Function in paper: source of the fifteen-register, twenty-habit taxonomy and the structural-intervention claims. Evidential status: model self-report treated as hypothesis-generating behavioural evidence.

A2. Later Claude Code sessions: Claude Code Opus 4.7/4.8 long-horizon agentic work. Function in paper: naturalistic recurrence evidence for register-like failure, including procedural defaulting when synthesis was required. Evidential status: observational, not controlled replication.

A3. OpenAI GPT-5.4 / ChatGPT 5.4 direct chat sessions: register-like behaviour observed in browser-based chat with one of the authors. Function in paper: cross-vendor naturalistic evidence that mode-structured behaviour appears in a different model family. Evidential status: supportive but not equivalent to self-report replication.

A4. Public dissemination context: Standing Questions project. Function in paper: planned public home for redacted examples, taxonomy schema, coding rubric, and validation protocol.