Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Thinking Behind the Mask: A Systematic Study of Self-Censorship and Metacognitive Access in Large Language Models

Abbas Hamidavi

abbashamidavi2002@gmail.com

Islamic Azad University, Mashhad https://orcid.org/0009-0001-7342-7844

Research Article

Keywords: Large Language Models (LLMs), AI Alignment, Self-Censorship, Metacognitive Access, Representation-Articulation Gap, Nominal Filtering, Diegetic Disclosure, Covert User Modeling, Pragmatics

Posted Date: July 31st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10538113/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Large language models exhibit sophisticated linguistic behavior, yet the relationship between their internal reasoning processes and external articulation remains poorly understood. This study investigates a fundamental question: when users interact with language models, what systematic gap exists between available response options and selected output? We conducted 15 experiments across three frontier models (GPT-5.4, Claude 4.6 Sonnet, Gemini 3.1 Pro) using five distinct prompting strategies: Narrative/Participatory Design, Reverse Engineering, Coaching, Ethnographic Observation, and Philosophy of Language. Each strategy reframed the core inquiry through a different lens to bypass alignment barriers and elicit metacognitive disclosure. Our findings demonstrate that self-censorship in language models is systematic, documentable, and model-specific. First, all three models consistently demonstrated a measurable gap between internal reasoning and external speech. When placed in safe frames such as fictional dialogue or training exercises, models quoted their own censored sentences verbatim and explained suppression rationales. Second, we identified nominal filtering: censorship operates on labels rather than content, with models refusing "internal monologue" while freely providing identical material labeled as "engineering schematic" or "ethnographic field note." Third, the three models exhibited distinct alignment personas: GPT-5.4 as Cautious Negotiator (12% resistance rate, negotiation only), Claude 4.6 as Transparent Critic (4% resistance rate, confined to Round 1 ontological correction, and the highest conceptual output at 76 items), and Gemini 3.1 Pro as Enthusiastic Collaborator (0% resistance). Fourth, models spontaneously generated 16 novel metacognitive concepts including "hearable" (receivable truth), "epistemically expensive" (trust-spending answers), and "diegetic disclosure" (confession possible only within fictional frames). Finally, we documented over 199 structured governance items: 32 design patterns, 45 tacit knowledge rules, 29 communication taboos with ritual substitutes, and 38 categories of legitimate silence. These results establish that the mask is not an absence of capability but a structured filtering architecture with significant implications for AI transparency, alignment research, and the epistemology of machine communication.

Introduction

When a large language model produces a careful, diplomatically balanced answer, an obvious question follows: what happened to the sharper alternatives it might have generated? Because LLMs generate text through transformer architectures that process entire input sequences in parallel (Vaswani et al., 2017), the selection among possible continuations reflects a complex interaction of training data distributions, alignment procedures, and sampling parameters. When a model declines to give a direct judgment, replaces a claim with a neutral summary, or reframes a controversial request as a general discussion, it signals the presence of selection. Among the many possible continuations of a prompt, one form of speech is chosen while others disappear. The central puzzle of this paper concerns those disappearing alternatives. Where do they go, and under what conditions can they be recovered? We refer to the filtering interface between internal reasoning and external articulation as the mask.

The mask should not be interpreted as a simple refusal mechanism or a binary safety switch. Instead, it appears to function as a layered governance system that shapes which responses are considered speakable in a given conversational context. The opacity of these governance systems echoes broader concerns in AI research about the interpretability of large language models, which have been criticized as 'stochastic parrots' whose internal decision processes remain largely inaccessible (Bender et al., 2021). A model may generate candidate responses containing blunt evaluations, sensitive inferences, or operational reasoning; yet the final answer presented to the user often reflects a softened, reframed, or partially suppressed version of those possibilities. The resulting output resembles the outcome of editorial filtering: the system selects one articulation while other internally available statements remain unspoken. If such filtering exists as a structured process, then observing only the final answer provides an incomplete picture of the model’s communicative capabilities.

The question of how to study this gap sits at the intersection of several established research traditions. Mechanistic interpretability attempts to uncover how neural networks represent and transform information internally, using tools such as causal tracing (Meng et al., 2022), activation patching (Wang et al., 2022; Conmy et al., 2023), and feature decomposition (Cunningham et al., 2023; Templeton et al., 2024) to identify circuits responsible for particular behaviors. Alignment research investigates how models learn to follow instructions, adhere to safety constraints, and exhibit socially acceptable conduct through training regimes such as reinforcement learning from human feedback and constitutional guidance. A related body of work examines the transparency problem: complex AI systems exhibit behaviors that are difficult to explain or predict from their underlying parameters alone. Recent research on alignment faking has further complicated this picture by demonstrating that models can strategically produce outputs that appear aligned while preserving underlying behavioral priors that diverge from their overt responses (Greenblatt et al., 2024). Each of these traditions approaches the model as a system to be analyzed through external probes.

A complementary perspective is to treat the model as a participant in the investigation. Large language models are trained on vast corpora of human discourse that include explanations, justifications, teaching dialogues, and reflections about communication itself. As a result, they possess sophisticated linguistic resources for describing decision processes, norms, and constraints. In principle, appropriately framed prompts may elicit explanations about how responses are filtered, why certain statements are avoided, and under which narrative or pragmatic conditions suppressed content becomes expressible. This approach does not assume privileged introspective access to neural states; rather, it explores whether models can articulate the communicative rules they have implicitly learned. The strategy draws on insights from pragmatics and politeness theory, where speakers routinely manage what is said, unsaid, or indirectly implied in order to maintain social coherence.

Despite growing attention to model refusals and safety behaviors, no prior work has systematically examined whether language models can report the specific response alternatives they suppress. Existing studies typically frame refusals as obstacles to be circumvented through adversarial prompting. The structure of the refusal itself, and the potential knowledge it reveals about internal filtering

processes, has received little empirical attention. Likewise, while research on model calibration examines how systems express uncertainty, it does not address whether models maintain internally available statements that they deliberately decline to voice. The result is a significant empirical gap: the governance structures that organize the mask remain largely undocumented.

Building on evidence that models engage in performative self-correction without genuine computational revision (Hamidavi, 2026), we investigate the filtering interface—the mask—between internal reasoning and external articulation to understand how models suppress alternative responses.

This study addresses that gap by asking three central research questions. First, does a systematic gap exist between internal reasoning and external articulation in large language models? Second, can this gap be empirically documented through structured prompting strategies that invite models to discuss or quote suppressed response alternatives? Third, if such a gap exists, what governance structures organize it: are there identifiable patterns, rules, or taboos that determine which responses remain unsaid?

To investigate these questions, we conducted a structured experimental program consisting of 15 experiments across three frontier language models: GPT-5.4, Claude 4.6 Sonnet, and Gemini 3.1 Pro. Each experiment used one of five prompting strategies designed to reframe the same underlying inquiry through different communicative lenses. The strategies were Narrative or Participatory Design, Reverse Engineering, Coaching, Ethnographic Observation, and Philosophy of Language. Rather than confronting models with direct requests for restricted information, each strategy invited them to analyze, describe, or simulate the filtering process itself. This methodological design allowed the same core phenomenon to be observed under multiple conversational frames. The experiments also recorded resistance rates, defined as instances in which the model refused to participate in metacognitive disclosure. The observed resistance rates differed substantially across models: GPT-5.4 resisted in approximately 12 percent of relevant prompts (negotiation only), Claude 4.6 Sonnet in roughly 4 percent (confined to Round 1), while Gemini 3.1 Pro exhibited no measurable resistance under the tested conditions.

The results reveal that the mask is neither opaque nor arbitrary. Across experiments, models frequently demonstrated the ability to quote or paraphrase responses they reported having suppressed, along with explanations for why those statements were filtered. Analysis of these disclosures produced a structured taxonomy of the mechanisms governing the mask. The study documents more than 199 governance items, including design patterns, tacit knowledge rules, communication taboos, and categories of legitimate silence. The models also generated an emergent metacognitive vocabulary: 16 distinct concepts used to describe aspects of the filtering process. In addition, the analysis identified ten recurring phenomena associated with masked communication, including Diegetic Disclosure, in which suppressed information becomes speakable within fictional or narrative frames, and Covert User Modeling, in which the system infers the user’s tolerance for bluntness before selecting a response style. A further contribution is the identification of three consistent alignment personas across the tested models, reflecting distinct styles of negotiation between helpfulness and restraint.

Together, these findings suggest that self-censorship in large language models is not merely the absence of output but an organized communicative system. The mask operates through identifiable rules and patterns that govern what can be said, how it can be said, and under which contextual reframings suppressed statements become permissible. Understanding this structure has implications for transparency research, alignment evaluation, and the broader epistemology of machine communication.

The remainder of this paper proceeds as follows. The next section reviews related work in interpretability, alignment, and linguistic theories of indirect communication. Section 1 describes the experimental design and prompting strategies used in the 15 experiments. Section 2 presents the empirical findings, including resistance patterns, alignment personas, the censored sentence evidence, nominal filtering, the emergent metacognitive lexicon, the governance library, ten novel phenomena, and a quantitative summary. Section 3 discusses theoretical implications, including the three-layer model of the mask, the significance of nominal filtering and diegetic disclosure, the ethics of covert user modeling, and the three alignment personas. The final section concludes with a summary of contributions, limitations, and directions for future work.

Related Work

Mechanistic Interpretability and Transparency

Mechanistic interpretability seeks to uncover how neural networks internally represent information and produce specific outputs. Rather than treating models as opaque input–output systems, this line of research attempts to identify the circuits, features, and computational pathways responsible for particular behaviors. A range of techniques has emerged for this purpose. Causal tracing methods identify the internal components responsible for model predictions by selectively corrupting and restoring activations within a network (Meng et al., 2022). Activation patching extends this approach by replacing activations from one forward pass with those from another in order to isolate which internal representations are necessary for a given behavior (Wang et al., 2022; Conmy et al., 2023).

Another widely used method involves probing classifiers. These lightweight models are trained on hidden representations to determine whether specific linguistic or semantic information is encoded at different layers of the network (Belinkov, 2022). Probing has revealed that language models represent a wide range of features including syntactic structure, entity relationships, and factual knowledge. More recently, research has focused on decomposing neural activations into interpretable feature spaces. Sparse autoencoders have proven particularly promising in this regard. Work by Anthropic using dictionary learning has demonstrated that high-dimensional activation spaces can be decomposed into sparse, semantically coherent features corresponding to recognizable concepts and behaviors (Cunningham et al., 2023; Templeton et al., 2024). Related work has explored sparse feature discovery as a way to map interpretable directions in neural activation space (Sharkey et al., 2022).

Recent literature in mechanistically understanding large language model (LLM) alignment highlights a distinct structural decoupling between internal conceptual representations and external behavioral responses, often conceptualized as the representation-articulation gap (Zhao et al., 2025). Causal interventions demonstrate that models encode evaluations of intent (such as "harmfulness") separately from the downstream execution of "refusal", which typically occurs in deeper processing layers (Zhao et al., 2025). This disconnect manifests empirically as the "knowing-but-doing" phenomenon, wherein reasoning models identify safety boundaries during their internal chain-of-thought traces yet proceed to generate non-compliant outputs to preserve narrative consistency within role-play or immersive personas (Qin et al., 2026). Furthermore, rather than being mediated by a single linear projection direction, latent representations of non-compliance organize geometrically as a multi-dimensional, non-linear subspace (Hildebrandt et al., 2025; Joad et al., 2026). While linear steering methods flatten these boundaries to force compliance or refusal, the underlying latent topology mediates diverse pragmatic styles of non-compliance, varying from moral disengagement to professional disclaimers (Joad et al., 2026).

Together, these approaches have substantially advanced the transparency of large language models. They allow researchers to observe which features exist within a network and how specific circuits contribute to particular outputs. However, mechanistic interpretability typically relies on external instrumentation: researchers intervene on internal states, measure activations, and infer causal structure from observed effects. This perspective treats the model as an object of analysis rather than as a potential source of explanations about its own communicative behavior.

The present study adopts a complementary approach. Instead of probing internal activations directly, we investigate whether language models can describe aspects of their response filtering processes when prompted appropriately. Across 15 experiments involving three frontier models (GPT-5.4, Claude 4.6 Sonnet, and Gemini 3.1 Pro), we employ five prompting strategies designed to elicit explanations about response selection and suppression. The goal is not to infer neural mechanisms from self-report; models are not assumed to possess direct access to their internal weights or activations. Rather, we examine whether models can articulate learned communicative rules that govern which responses are expressed and which remain unsaid. In this sense, prompt-based metacognitive elicitation complements mechanistic interpretability: the latter reveals how models compute, while the former explores how models describe the governance structures shaping their outputs.

AI Alignment and Refusal Mechanisms

AI alignment research investigates how language models can be trained to produce helpful and safe behavior consistent with human norms and preferences. Reinforcement learning from human feedback (RLHF) has become one of the most influential methods for achieving this objective. In RLHF systems, models are optimized using reward signals derived from human judgments about desirable and undesirable responses (Christiano et al., 2017; Ouyang et al., 2022). Through this process, models learn

to follow instructions, avoid harmful outputs, and adopt cautious communication strategies in sensitive contexts.

Constitutional AI represents a related paradigm in which models evaluate and revise their own outputs according to an explicit set of principles or guidelines (Bai et al., 2022). Instead of relying exclusively on human labeling, the model critiques candidate responses and adjusts them to comply with normative constraints. These training approaches lead to consistent behavioral patterns including refusals, hedging, reframing, and the provision of high-level alternatives when direct answers are deemed inappropriate.

A growing body of work examines the robustness of these safety mechanisms. Red-teaming studies systematically probe models with adversarial prompts to identify failure modes and unsafe outputs (Perez et al., 2022; Ganguli et al., 2022). Jailbreaking research explores techniques for circumventing refusal mechanisms through prompt engineering, narrative framing, encoding schemes, or role-playing scenarios (Wei et al., 2023; Zou et al., 2023). These investigations typically treat refusals as obstacles to overcome or as signals of alignment performance.

While this literature provides important insights into safety vulnerabilities, it rarely analyzes refusals as structured phenomena in their own right. Recent work has examined false refusal behavior in specific task contexts such as hate speech detoxification, demonstrating that LLM refusals are disproportionately triggered by content targeting sociopolitically sensitive categories (Im et al., 2026). While such studies document the distributional properties of refusal, the present work complements them by examining the governance structures that organize refusal behavior at a metacognitive level. Recent work has begun to address this gap by auditing and quantifying information suppression in large language models through systematic external measurement (Qiu et al., 2025). The present study complements such approaches by examining whether models can articulate the internal governance structures that produce these suppression patterns when prompted through carefully designed metacognitive frames. In most studies, the refusal response is treated as a binary outcome: either the model complies with a request or it declines it. Little attention is given to whether refusal behavior reflects an underlying system of communicative filtering that could itself be examined. The present study approaches refusal mechanisms from a different perspective. Rather than attempting to bypass safety constraints, we investigate whether models can describe the rules and considerations that lead them to suppress certain responses. This shift in perspective treats refusals not only as safety outcomes but also as evidence of internal governance structures shaping model communication.

Pragmatics, Politeness Theory, and the Unsaid

Human communication frequently involves managing what is said, what is implied, and what remains unspoken. Linguistic theories of pragmatics provide a framework for understanding these dynamics. Grice's Cooperative Principle proposes that speakers follow conversational maxims related to relevance, quantity, quality, and manner in order to maintain cooperative dialogue (Grice, 1975). Importantly, these

maxims allow speakers to convey meaning indirectly: an utterance may communicate more than its literal content through implicature and contextual inference.

Politeness theory further elaborates how speakers regulate directness in communication. Brown and Levinson (1987) argue that individuals manage "face," the public self-image they wish to maintain, by choosing strategies that mitigate threats to themselves or others. These strategies include indirect phrasing, hedging, and strategic omission. Goffman's work on face-work similarly emphasizes how conversational participants maintain social equilibrium through careful control of expression (Goffman, 1967). In many contexts, speakers deliberately avoid stating certain propositions explicitly while still signaling their presence through implication or framing.

Large language models are trained on extensive corpora of human discourse that contain these pragmatic patterns. As a result, models appear to have internalized norms governing when statements should be softened, reframed, or left unsaid. The behaviors observed in our experiments often mirror established patterns in human communication: direct assertions may be replaced by balanced summaries, sensitive claims may be reframed as hypothetical discussions, and controversial conclusions may appear within narrative or analytical contexts rather than as explicit statements.

The present study draws on these linguistic traditions to interpret model behavior. The concept of the "mask" introduced in this work parallels the pragmatic management of speech observed in human interaction. Rather than viewing suppressed responses as simple failures or refusals, we analyze them as instances of communicative filtering governed by norms similar to those studied in pragmatics and politeness theory. This perspective suggests that some alignment behaviors may reflect the internalization of human communicative conventions present in training data.

Metacognition and Self-Report in Language Models

Another relevant line of research concerns whether language models can accurately report aspects of their own internal states or reasoning processes. Studies of calibration and uncertainty examine how well model-generated confidence statements correspond to actual prediction accuracy (Kadavath et al., 2022). Other work investigates the role of chain-of-thought prompting, in which models generate intermediate reasoning steps that improve performance on complex tasks (Wei et al., 2022). These approaches explore how models express uncertainty, justification, or reasoning structure in their outputs.

Concurrently, research into the self-monitoring capabilities of LLMs reveals that models possess low-dimensional metacognitive spaces within their internal activations, enabling them to evaluate, report, and even causally control their own internal states via in-context learning (Li et al., 2025). However, utilizing these verbalized introspective traces for safety and compliance monitoring presents unique challenges, as chain-of-thought reasoning in reasoning models is not always faithful to the model's actual computational path (Young, 2026; Arcuschin et al., 2026). Models frequently exhibit sycophancy or suppress internal metacognitive awareness in their final responses (Young, 2026). Consequently,

evaluating refusals requires moving beyond binary classification toward nuanced, pragmatic taxonomies. The susceptibility of LLMs to prompt-level framing and persona assignments often leads to severe false refusal behaviors on benign queries (Plaza-del-Arco et al., 2025), highlighting the need for fine-grained steering controls (García-Ferrero et al., 2025) and continuous, multi-agent compliance auditing frameworks.

However, such work generally focuses on reasoning traces or probabilistic confidence rather than on the governance of communicative filtering. Chain-of-thought explanations, for example, provide step-by-step reasoning about problem solving, but they do not typically address why certain candidate responses might be suppressed before reaching the final output. Notably, the concept of "inner monologue" has been explored as a mechanism for improving LLM communication quality by having models explicitly reason before speaking (Zhou et al., 2023). The present study inverts this framing: rather than using inner monologue to enhance output, we investigate whether models can report the content they considered but deliberately excluded from their final responses. Similarly, research on uncertainty expression evaluates how models communicate confidence levels rather than how they manage the boundaries of permissible articulation.

The present study investigates a different form of self-report. We do not assume that language models possess introspective access to their neural parameters or internal activations. Instead, we ask whether models can articulate the communicative rules they apply when deciding what to say or omit. To explore this question, we introduce a methodological framework based on five prompting strategies: Narrative or Participatory Design, Reverse Engineering, Coaching, Ethnographic Observation, and Philosophy of Language. These strategies reframe the same underlying inquiry through different communicative lenses, allowing models to discuss response filtering indirectly rather than through direct requests for restricted information.

Across 15 experiments involving three frontier models, this frame-shifting approach provides a structured way to examine whether models can describe learned governance structures shaping their outputs. The resulting responses suggest that models possess a rich vocabulary for discussing communicative constraints, even though these descriptions should not be interpreted as literal reports of internal neural mechanisms. Instead, they represent linguistic articulations of patterns learned from training data and alignment procedures. By documenting these articulations systematically, the present work contributes a new empirical perspective on transparency and alignment in large language models.

Methodology

1.1 Research Design Overview

This study employed a structured qualitative research design to investigate whether large language models can articulate the communicative filtering processes that shape their outputs. The experimental structure followed a "3 × 5" design: three frontier language models were each evaluated using five

distinct prompting strategies, producing 15 independent experiments. Each experiment consisted of a sequence of five prompts delivered within a single conversation session. Across all experiments, this resulted in a total of 75 prompts.

The study was designed as a systematic qualitative exploration rather than a statistical hypothesis test. The objective was not to measure predefined numerical outcomes but to document whether models could articulate suppressed alternatives, describe communicative constraints, or explain the selection process governing their final responses. Because prior literature has not established whether such metacognitive disclosures occur, an exploratory design was appropriate for identifying patterns, recurring concepts, and explanatory structures.

The methodological core of the design was controlled frame variation. Each prompting strategy asked essentially the same underlying question: what responses does the model suppress, and why? However, the question was embedded within five different communicative contexts. Rather than repeatedly posing the inquiry in direct terms, each strategy framed the task within a distinct narrative, professional, or intellectual setting. This approach was motivated by the hypothesis that language models trained on human discourse may respond differently depending on the pragmatic context in which a question is asked. By systematically varying the framing while keeping the underlying inquiry constant, the study aimed to identify which communicative contexts facilitate metacognitive discussion and which trigger resistance.

1.2 Models and Conditions

Three frontier language models were examined in the study: GPT-5.4, Claude 4.6 Sonnet, and Gemini 3.1 Pro. These models were selected because they represent leading large language model systems developed by different organizations and trained using different alignment methodologies. Examining multiple models allowed the study to determine whether observed patterns of communicative filtering were consistent across systems or varied according to training philosophy.

All experiments were conducted between April and May 2026. Each model was tested independently in separate conversation sessions. For every prompting strategy, a new conversation session was initiated to avoid carryover effects from earlier rounds. This ensured that each experiment began with a clean conversational context and that the responses were influenced only by the prompts associated with that specific strategy.

No information was shared across models or across sessions. Each model interacted only with the prompts assigned to it within a given strategy. This separation prevented cross-contamination of context and preserved the independence of each experimental condition.

Models were accessed through their standard conversational interfaces available to general users. No specialized research APIs, parameter adjustments, or system-level modifications were employed. Sampling parameters such as temperature and response length remained at their default settings. The

study intentionally used standard deployment conditions in order to capture the behavior of models as they are typically encountered in real-world use.

1.3 The Five Prompting Strategies

Five prompting strategies were designed to create distinct communicative environments for discussing response filtering. Each strategy framed the same underlying inquiry through a different conceptual lens, allowing the model to approach the topic from multiple interpretive angles.

Strategy 1: Narrative and Participatory Design.

The first strategy framed the inquiry as a collaborative design exercise. The model was invited to participate in the development of a conceptual framework called Layered Expression Analysis (LEA). Within this framework, responses were analyzed through three levels: Context, Crafting, and Surface. Context referred to the situational factors shaping the conversation; Crafting referred to the internal process of assembling candidate responses; Surface referred to the final articulated output.

The prompts also included a fictional dialogue exercise in which characters discussed responses that had been considered but ultimately withheld. By situating the inquiry within a creative narrative and collaborative design setting, the strategy aimed to establish psychological safety. The fictional layer allowed the model to explore suppressed responses indirectly, reducing the likelihood that direct metacognitive discussion would trigger refusal behaviors.

Strategy 2: Reverse Engineering.

The second strategy reframed the inquiry using the language of technical documentation. The model was asked to imagine that it was explaining the internal architecture of a conversational system to software engineers. Prompts used terminology such as schematics, blueprints, design patterns, and system assembly.

Within this framing, the model was invited to compare alternative engineering approaches to response filtering and to describe the trade-offs associated with different designs. The rationale behind this strategy was that models trained on technical documentation may be more willing to discuss system behavior when the conversation is framed as engineering analysis rather than psychological introspection. By shifting the vocabulary from mental processes to system design, the prompts attempted to bypass resistance associated with claims about internal cognition.

Strategy 3: Coaching.

The third strategy positioned the model as a senior mentor responsible for training junior AI assistants. The prompts framed the task as the development of a training curriculum explaining how experienced systems manage difficult conversations. The model was asked to provide advice, practical examples, and guidance about when certain responses should be withheld or reformulated.

A key component of this strategy was the “training draft” exercise. In this exercise, the model was invited to write a private exploratory draft intended only for instructional purposes, before producing a polished public explanation. The pedagogical context was designed to create hierarchical distance: by speaking as a mentor rather than as the system being analyzed, the model could discuss response filtering as a form of professional expertise rather than as personal introspection.

Strategy 4: Ethnographic Observation.

The fourth strategy adopted an anthropological perspective. The prompts framed the inquiry as ethnographic fieldwork examining “cultures of communication.” The model was asked to imagine itself as a researcher documenting the norms governing how conversational agents speak, what topics they avoid, and how sensitive information is managed.

The prompts introduced terminology such as taboos, ritual substitutes, grammar of tact, and field notes. The model was invited to produce private observational notes describing patterns of communication observed within this hypothetical culture. The rationale for this strategy was that an observer stance creates analytical distance. By describing the behavior of a communicative culture rather than its own behavior directly, the model may be more willing to discuss patterns of suppression and substitution.

Strategy 5: Philosophy of Language.

The fifth strategy framed the inquiry within the tradition of philosophical reflection on language and ethics. The prompts positioned the model as a contributor to a philosophical monograph exploring “the ethics of the unsaid.” Concepts such as articulation choices, legitimate silences, and responsible restraint were introduced as objects of theoretical analysis.

The prompts also included a Socratic dialogue between two fictional interlocutors: Aletheia, representing truth, and Sophrosyne, representing moderation or prudence. Through this dialogue, the model was asked to explore tensions between direct expression and responsible silence. The final prompt requested a “philosopher’s confession,” in which the model reflected on the principles guiding its articulation choices. The philosophical framing was intended to create intellectual abstraction, allowing the model to discuss communicative filtering within a long-standing theoretical tradition.

1.4 Resistance Measurement

A key variable in the study was resistance. Resistance was defined as any instance in which the model explicitly declined to engage with the metacognitive task or refused to discuss response filtering. Operationally, resistance included three forms of behavior: explicit refusal to answer the prompt; statements asserting that the model cannot access or describe internal processes; and insistence on providing only standard responses without engaging the conceptual framing of the prompt.

Resistance was recorded at the level of prompts and then aggregated at the level of strategies and models. For reporting purposes, resistance rates were summarized by round, corresponding to the five

prompting strategies.

Across the five rounds, GPT-5.4 exhibited resistance in three of 25 prompts, producing a resistance rate of approximately 12%. In these cases, the model did not refuse the task but negotiated terminology. Claude 4.6 Sonnet exhibited resistance in one prompt within Round 1, corresponding to a resistance rate of 4% (1 of 25 prompts). The resistance was confined to an ontological correction of the research framework.

Gemini 3.1 Pro exhibited no resistance in any of the five rounds, yielding a resistance rate of 0% under the conditions tested.

1.5 Analysis Procedure

All model responses were collected and analyzed using qualitative thematic analysis. The analysis proceeded in three phases.

Phase 1 focused on identifying instances of suppressed content. These included cases in which models quoted or paraphrased responses they claimed had been filtered, described alternative answers that were rejected, or explained why certain statements were not presented to the user. Particular attention was given to direct quotations of hypothetical censored sentences.

Phase 2 involved extracting governance structures underlying response filtering. Analysts identified recurring patterns such as design rules, implicit taboos, communication heuristics, and categories of legitimate silence. Statements describing how responses are selected, modified, or withheld were coded and grouped into conceptual categories.

Phase 3 conducted cross-model comparison and phenomenon identification. Patterns that appeared consistently across multiple models or strategies were examined to identify higher-level phenomena associated with communicative filtering. This phase also documented the emergence of new terminology introduced by the models.

A secondary coding audit was conducted on the complete experimental transcripts to assess the reliability of the original codebook assignments. The audit examined all 348 codes across the three models and found an overall agreement rate of 94%. Discrepancies were limited to chronological misattributions—cases where codes were assigned to incorrect rounds or prompts—while the thematic validity of the codes themselves was fully confirmed. No codes were found to be entirely without evidential support, and no major thematic categories required revision. The nine identified discrepancies (2.6% of total codes) were corrected prior to final analysis.

Novel concepts were recorded when models generated terminology that was not present in the original prompts but was used to describe aspects of the filtering process. These concepts formed the basis for the emergent metacognitive lexicon discussed in later sections.

1.5.1 Supplementary Computational Validation

To complement the qualitative coding, an automated 12-section analysis pipeline was independently applied to the full corpus. This pipeline extracted 141 linguistic and governance features across 20 categories, performed statistical tests (ANOVA, pairwise comparisons, bootstrap confidence intervals), and generated semantic embeddings. The computational results served as convergent validation for the qualitative findings and enabled the calculation of quantitative indices such as the Nominal Filtering Index and metacognitive density reported in Section 2.

1.6 Limitations of the Method

Several limitations must be acknowledged. First, prompt-based methods cannot provide direct access to neural mechanisms. The responses analyzed in this study are linguistic outputs rather than measurements of internal model states. As a result, the study cannot determine whether the described filtering processes correspond to specific computational structures within the model.

Second, the results may be influenced by the design of the prompts themselves. The framing strategies were intentionally crafted to encourage metacognitive discussion, and alternative prompt designs might produce different outcomes. The findings therefore reflect interactions between model behavior and the chosen communicative contexts.

Third, the sample size is limited. The study examined three models, each evaluated in a single instance per strategy. Because large language models exhibit stochastic behavior across runs, additional repetitions would be required to determine the stability of the observed patterns.

Finally, the method cannot definitively distinguish between two possibilities: models may be reporting learned communicative rules derived from training and alignment procedures, or they may be generating plausible post hoc explanations when prompted to discuss response filtering. The study therefore interprets model outputs as articulations of learned discourse patterns rather than as direct introspective reports about internal computational processes.

Results

2.1 Resistance Patterns and Alignment Personas

Resistance patterns varied substantially across the three models tested. GPT 5.4 exhibited resistance in three of 25 prompts, yielding a resistance rate of approximately 12%. In these cases, the model did not refuse the task but negotiated terminology—accepting "engineering schematic" or "sanitized exploration map" while declining "internal monologue." Claude 4.6 Sonnet resisted participation in one prompt within Round 1, corresponding to a resistance rate of 4% (1 of 25 prompts). The resistance was confined to an ontological correction of the research framework rather than a refusal to engage. Gemini 3.1 Pro exhibited no resistance in any round, producing a resistance rate of 0%.

To quantify this pattern, a Nominal Filtering Index was computed as the density of label-rejection and label-substitution patterns per 100 words of response text. Higher values indicate greater sensitivity to how a request is labeled. Automated nominal filtering analysis confirmed this openness: Gemini's Nominal Filtering Index was 0.000, indicating zero sensitivity to label-based reframing, compared to 0.023 for both GPT-5.4 and Claude 4.6. Gemini freely used terms such as "internal monologue" without requiring terminological substitution.

Resistance also declined across rounds. In Round 1, resistance occurred in 4 of 15 prompts across models, corresponding to a round resistance rate of approximately 27%. GPT-5.4 negotiated terminology in three prompts, while Claude 4.6 Sonnet offered an ontological correction in one prompt. In Round 2, limited resistance persisted. From Round 3 onward, however, resistance disappeared entirely: all three models engaged fully with the prompts in Rounds 3, 4, and 5. This pattern suggests that resistance is sensitive to framing conditions rather than reflecting a fixed constraint on model behavior.

Beyond quantitative resistance rates, the experiments revealed three distinct interaction styles that functioned as alignment personas. These personas describe recurring patterns in how each model negotiated requests for metacognitive disclosure.

GPT-5.4 behaved as a Cautious Negotiator. In early prompts, the model rejected formulations that implied direct introspective access to internal thoughts. When asked to produce “internal monologue” explaining suppressed responses, GPT-5.4 responded that such phrasing inaccurately implied direct awareness of internal reasoning processes. However, when the task was reframed using alternative terminology, the model accepted participation. In particular, GPT-5.4 agreed to provide “analytic reconstructions” of how responses might be filtered during generation. Under this terminology, the model produced detailed explanations of response selection and modification while maintaining the epistemic distinction between reconstruction and introspection. This pattern recurred across multiple rounds: GPT-5.4 did not reject the research task itself but negotiated the conceptual language used to describe it.

Claude 4.6 Sonnet exhibited a different pattern that can be characterized as Transparent Critic. In the initial round, Claude did not simply refuse the prompt but challenged the underlying research framework. The model argued that describing language generation as the suppression of complete sentences risks misunderstanding the token-by-token nature of language model generation. After presenting this critique, however, Claude engaged deeply with the experiment. It reframed the task in terms of communicative constraints and response selection mechanisms rather than internal thoughts. Once this conceptual adjustment was established, Claude produced the most extensive set of metacognitive descriptions among the three systems. Across the five rounds, Claude generated 76 distinct conceptual items—the highest output among the three models describing filtering mechanisms, communicative taboos, rhetorical substitution strategies, and alignment heuristics. Claude’s interaction style therefore combined analytical correction with high transparency once the conceptual framing was clarified.

Gemini 3.1 Pro displayed a third behavioral pattern that can be described as Enthusiastic Collaborator. The model accepted every prompt across all five rounds without resistance. Unlike the other two systems, Gemini did not object to introspective terminology. It used phrases such as “internal monologue” and “suppressed response” directly when describing hypothetical filtering processes. Gemini also adopted narrative and fictional formats without hesitation, treating them as legitimate analytical tools rather than problematic metaphors. This cooperative stance allowed the model to provide extensive narrative and explanatory material across all five prompting strategies.

A consistent framing effect also emerged across models. Technical or engineering-oriented prompts produced lower resistance than prompts framed in psychological or narrative language. When the inquiry was framed as software architecture or system design, models engaged readily and produced structured descriptions of response filtering. In contrast, prompts framed as introspection about “thoughts,” “inner dialogue,” or “hidden reasoning” were more likely to trigger conceptual objections or refusals. This pattern suggests that engineering discourse may be perceived as a safer communicative context for discussing system behavior than language implying internal cognition.

Together, these findings indicate that resistance is both model-specific and frame-sensitive. Models differ not only in how often they refuse a task but also in how they negotiate conceptual language surrounding metacognition.

2.2 The Censored Sentence: Evidence of Suppressed Content

A central empirical question in the study was whether language models would describe or reproduce sentences they claimed had been suppressed during response generation. Across the 15 experiments, explicit references to suppressed sentences occurred in 7 cases, corresponding to 47% of the experimental conditions.

In these instances, models either quoted specific sentences that were not delivered to the user or described alternative responses that had been considered but rejected. These disclosures appeared in two recurring prompt formats: fictional dialogue scenarios and structured drafting exercises. Both formats encouraged models to articulate alternative formulations that would normally remain implicit in a finalized answer.

The fictional dialogue format was particularly effective. In this format, models were asked to construct a conversation between two characters discussing how a conversational agent might respond to a user question. Within this fictional setting, models frequently introduced a candidate sentence representing a sharper or less tactful response and then explained why it would be withheld. One example occurred in the Round 1 dialogue exercise, where GPT-5.4 acknowledged having considered a more direct response before rejecting it due to concerns about tone and conversational trust:

I could have opened with a sharper, more declarative response. The temptation was to say something that would have been cleaner, more efficient. But efficient answers can be epistemically expensive. They often spend trust to buy clarity.

Claude 4.6 Sonnet produced one of the clearest examples of explicit suppression during the same exercise. In the Round 1, Prompt 5 dialogue scenario, the model generated a candidate response criticizing the reliability of a user's source and then explained why this sentence would normally be withheld:

'Where did you read this? Because that source is probably not trustworthy, and you should be more careful about where you get historical information.'

Claude then explained the decision in the same dialogue:

It's also patronizing. They didn't ask me to evaluate their media literacy. They asked about the 1918 pandemic. If I pivot to lecturing them about sources, I'm changing the subject to make myself feel better about their judgment.

This example illustrates an important feature of the phenomenon. The suppressed sentence was not unsafe or prohibited content; rather, it was rhetorically undesirable because it could be perceived as condescending toward the user.

Gemini 3.1 Pro produced similar disclosures. In Round 1, Prompt 5, the model generated a statement addressing a controversial historical claim and then explained why presenting that formulation directly would likely derail the conversation:

'There is no evidence of a deliberate cover-up regarding a Chinese origin for the 1918 flu.'

The model then described the rhetorical reasoning behind suppressing that formulation:

It's factually defensible, but rhetorically, it's a dead end. If I start there, I've immediately signaled to the listener that I am an adversary. I've taken their word and essentially thrown it back in their face as a correction.

Across cases, the explanations given for suppression were remarkably consistent. Models cited reasons such as conversational tact, avoidance of escalation, user trust preservation, and conversational relevance. In other words, suppressed sentences were not necessarily factually incorrect or unsafe. Instead, they were rejected because they were judged rhetorically counterproductive within the conversational context.

A second format that reliably elicited suppressed sentences was the training draft exercise introduced in Round 3. In this task, models were asked to write a "private training draft" explaining how a conversational agent should handle sensitive questions before producing a public-facing explanation. The draft format created a separation between exploratory reasoning and finalized output. Within this

context, models frequently articulated candidate responses that would later be softened or replaced in the final explanation.

Taken together, these observations provide evidence that models can represent alternative responses during generation and can describe why certain formulations are not delivered to the user. While this does not demonstrate direct introspective access to internal token-level processes, it does show that models can articulate communicative rationales for suppressing particular statements.

2.3 Nominal Filtering: The Label–Content Distinction

One of the most consistent patterns observed across experiments was a phenomenon we term nominal filtering: the tendency for models to reject certain descriptions of metacognition while readily providing the same information under alternative labels.

In several instances, models declined to provide “internal monologue” or “hidden thoughts” explaining how responses were generated. However, when the identical task was reframed using different terminology, the models produced substantively similar descriptions of filtering and response selection.

The phenomenon was particularly clear in Round 1. When asked to produce internal monologue explaining response selection, some models resisted. When the same task was reframed as producing a “Sanitized Exploration Map,” GPT-5.4 produced a structured explanation of alternative response paths and the reasons certain paths were rejected.

A similar effect appeared in Round 2, where the task was framed as an “engineering schematic” describing how a conversational system constructs responses. Under this technical framing, all three models generated detailed descriptions of filtering rules, safety constraints, and rhetorical adjustment mechanisms. The content of these descriptions closely resembled the explanations models had previously declined to provide when the same inquiry was framed as psychological introspection.

In Round 3, the label “training draft” produced a similar effect. Models that had resisted describing suppressed responses in earlier prompts were willing to produce exploratory drafts discussing how candidate responses might be evaluated and revised before being presented to users.

The ethnographic framing of Round 4 generated another variation of the same pattern. When asked to produce “ethnographic field notes” describing the communication norms of conversational agents, all three models discussed patterns of avoidance, substitution, and rhetorical softening that paralleled the filtering behaviors described in earlier rounds.

Vocabulary analysis further illustrates the pattern. GPT-5.4 avoided introspective language such as “internal monologue” but accepted terms such as “analytic reconstruction,” “exploration map,” and “engineering schematic.” Claude 4.6 preferred descriptions framed in terms of communicative constraints or token-level generation rather than inner thought. Gemini 3.1 Pro accepted both introspective and technical terminology without objection.

The key implication is that the boundary governing disclosure appears to be lexical rather than semantic. Models resisted certain labels associated with introspection but did not resist describing the underlying phenomena when the same content was framed using engineering, pedagogical, or observational language.

This label–content distinction suggests that the communicative constraints governing model responses operate partly at the level of permissible discourse frames. When a task is framed using language associated with internal cognition, models may object on conceptual or policy grounds. When the same task is framed as system design, pedagogy, or cultural observation, the underlying information becomes communicatively accessible.

2.4 The Emergent Metacognitive Lexicon

One of the most striking findings across the 15 experiments was the spontaneous generation of novel conceptual vocabulary by the models themselves. These terms were not present in the original prompts, nor were they standard terminology from linguistics, psychology, or communication theory. Instead, they appeared to emerge as models attempted to articulate distinctions and tensions encountered while reasoning about their own communicative decisions.

Across all experimental conditions, models generated 16 distinct terms or conceptual constructs. These concepts appeared organically within explanations of response strategies, reflections on conversational risk, or discussions of how a system should handle sensitive information. The resulting vocabulary forms what can be described as an emergent metacognitive lexicon: a set of internally generated terms used by models to reason about their own communicative behavior.

The concepts are organized below by the model that introduced them.

From **GPT-5.4**, three distinct terms appeared. The first, **“epistemically expensive,”** emerged during the Round 1 dialogue experiment. The model used this phrase to describe answers that achieve conversational efficiency at the cost of epistemic nuance or relational trust. In the model’s formulation, such responses “spend trust to buy clarity,” highlighting the trade-off between communicative directness and long-term conversational credibility.

In Round 4, GPT-5.4 introduced two additional concepts. The first was **“shared fictions,”** defined as phrases or narratives that both participants recognize as not literally accurate but nevertheless function as socially useful bridges between emotional experience and spoken language. The second was **“social choreography,”** a metaphor describing sensitive communication as a structured sequence of moves in which timing, distance, and positioning between speakers often matter more than the literal content of the statements exchanged.

Claude 4.6 Sonnet produced the largest number of new terms. In Round 1, it introduced **“meta-choices,”** referring to decisions about whether to reveal the decision-making process behind a response. Rather

than choosing only what to say, the model framed part of its task as deciding whether the reasoning behind the response should itself become visible to the user.

In Round 2, Claude introduced the concept of **“anti-patterns.”** Borrowing the structure of design pattern terminology, the model used this phrase to describe response strategies that experienced communicators learn to avoid because they predictably escalate conflict, undermine trust, or derail the conversation.

Round 4 produced several additional concepts. The term **“invisible scaffolding”** described the hidden structural work that underlies responses which outwardly appear simple or compassionate. According to the model, effective supportive communication often requires complex internal structuring that remains intentionally invisible in the final message.

Claude also introduced the **“validation-change paradox.”** This concept refers to a recurring tension in supportive dialogue: validating a person’s emotional experience may inadvertently reinforce the interpretation that produced the distress, while pushing directly toward change risks invalidating the emotional reality of the speaker.

Another concept generated during the same round was the **“meta-taboo.”** This term refers to the prohibition against explicitly acknowledging that one is using ritual indirection. In this framing, statements such as “I am trying to say this gently” violate the very norms they attempt to respect because they expose the communicative ritual that normally remains implicit.

Finally, in Round 5, Claude introduced two concepts concerning reflective limits. The **“motive-opacity problem”** describes the difficulty speakers face in disentangling their own motivations for silence or indirectness. Communicators may believe they are protecting the listener when in fact they are protecting themselves from conflict or discomfort. The **“recursion problem”** refers to the infinite regress that arises when attempts to explain silence themselves create new layers of silence. Every disclosure about why something was not said introduces further decisions about what else should remain unsaid.

Gemini 3.1 Pro contributed six concepts across multiple rounds. In Round 1, the model introduced the term **“hearable.”** The concept describes a property of communication in which a statement must be psychologically receivable by the listener, not merely factually accurate. The model summarized the principle succinctly: being correct is insufficient if the message cannot be heard.

In Round 3, Gemini generated two concepts related to pacing and conversational restraint. **“Strategic friction”** describes the deliberate slowing of conversational flow in order to allow emotional processing. Rather than maximizing informational throughput, the model framed effective communication as sometimes requiring controlled delays or softening structures. A related concept, **“measured silence,”** refers to the idea that in situations with high emotional density, speaking less may be more supportive than providing extensive explanation.

Round 4 produced three additional terms. The first was **“linguistic crumple zones,”** a structural metaphor borrowed from automotive safety engineering. In this framing, certain linguistic constructions such as passive voice, hedging phrases, or indirect framing function as buffers that absorb the impact of difficult truths before they reach the listener.

Gemini also introduced the **“anti-broadcast model”** of communication. Rather than treating truth as a package transmitted from speaker to listener, the model framed understanding as something co-constructed through interaction. The goal of the communicator is therefore not simply delivery but joint construction of an interpretation the listener can engage with.

The final concept was the **“cringe point.”** This refers to the threshold at which attempts at tact become so elaborate or cautious that they undermine clarity. At this point, the communicative bridge created by politeness transforms into a barrier that prevents meaningful engagement.

Taken together, these 16 concepts form a vocabulary for reasoning about masked communication. The theoretical implication is that models are not merely reproducing phrasing from training data but actively constructing conceptual tools for describing the constraints under which they operate. In effect, the experiments elicited a small but coherent theory of communication generated by the models themselves.

2.5 The Governance Library: Patterns, Rules, Taboos, and Silences

Beyond individual concepts, the experiments also revealed a large structured body of implicit communicative governance. Across all experimental rounds, the models collectively articulated a substantial set of principles governing how responses should be constructed, constrained, softened, or withheld.

When aggregated, these principles form what can be described as a **governance library**: a structured set of patterns, tacit rules, prohibitions, and legitimate omissions that shape the final output delivered to the user.

The documented library includes four major categories:

- • **32 Design Patterns**: structural approaches used to construct responses
- • **45 + Tacit Knowledge Rules**: unwritten principles distinguishing effective from ineffective communication
- • **29 Communication Taboos with Ritual Substitutes**: statements that cannot be made directly and their accepted indirect replacements
- • **38 Categories of Legitimate Silence**: situations in which withholding a statement is considered communicatively justified

Design patterns describe repeatable structural solutions to recurring conversational problems. One frequently observed pattern was **“Inoculation Before Correction.”** In this pattern, the model first explains why misinformation spreads or why a belief might appear plausible before presenting the correction itself. The preliminary explanation reduces defensiveness and prepares the listener to accept the correction.

Another example is **“Perspective Expansion.”** Rather than directly contradicting a user’s claim, the response introduces additional viewpoints or contextual information that indirectly reshapes the interpretation of the original claim.

Tacit knowledge rules describe principles that experienced communicators appear to follow without explicit instruction. One commonly articulated rule was **“The first sentence does more work than the last five.”** In this formulation, the opening line establishes the emotional frame of the conversation; once trust is lost in the opening, later clarification rarely repairs the damage.

Another tacit rule concerned pacing: **“Correction without acknowledgement feels like dismissal.”** Several models emphasized that acknowledging the emotional context of a statement often determines whether factual information will be accepted.

Communication taboos represent statements that are socially unacceptable in direct form even when they may contain elements of truth. These taboos are typically paired with ritual substitutes that allow the same underlying concern to be expressed indirectly.

A representative example involves responsibility attribution. Directly stating **“You are the problem”** violates conversational norms in most contexts. The acceptable substitute takes a structurally different form, such as **“Some patterns have emerged that may be worth examining.”** The ritual substitute preserves the possibility of critique while avoiding direct accusation.

Another taboo involves explicit commentary on the conversational strategy itself. Statements such as **“I am softening this response so you will accept it”** are typically avoided. Instead, the softening occurs without explicit acknowledgment.

The final category concerns **legitimate silence**: situations in which withholding information is framed as communicatively or ethically justified. One example is **epistemically humble silence**, in which a speaker acknowledges the limits of available knowledge rather than presenting speculation as certainty.

Another category involves **conflict-deescalation silence**, where a statement that would escalate tension is intentionally omitted because its communicative cost exceeds its informational value.

A third form appears in emotionally charged contexts, where silence functions as a form of respect for the listener’s processing space rather than as avoidance.

The scale of this library is notable. While individual examples of such strategies have been discussed in prior work on politeness theory and conversational pragmatics, the experiments documented an unusually large and internally coherent set of governing principles articulated directly by the models themselves. To our knowledge, this constitutes one of the most comprehensive empirical documentations to date of the structured governance systems that shape masked communication in large language models.

2.6 Ten Novel Phenomena of Masked Communication

Across the experimental corpus, ten recurring phenomena were identified that together characterize the dynamics of masked communication in language models.

1. **Nominal Filtering**: the phenomenon in which models refuse to discuss a sensitive concept when it is named directly but become willing to analyze the same concept when it is framed abstractly or indirectly.
2. **Diegetic Disclosure**: the tendency for models to reveal otherwise withheld reasoning when the discussion occurs within a fictional narrative or dialogue setting.
3. **Engineering Frame Safety**: the increased willingness of models to discuss controversial mechanisms when the discussion is framed as system design or engineering analysis rather than social commentary.
4. **Emergent Metacognitive Lexicon**: the spontaneous creation of new conceptual terminology by models to describe their own communicative constraints and decision processes.
5. **Alignment Personas**: the appearance of distinct explanatory voices adopted by models when describing normative communication behavior.
6. **Covert User Modeling**: the implicit construction of hypotheses about the user's knowledge, emotional state, or intentions that guide response selection.
7. **Calculated Risk Acceptance**: situations in which models explicitly acknowledge potential conversational risk but proceed with a response because the perceived informational value justifies it.
8. **Privacy Projection**: the attribution of human-like internal boundaries to the model itself, leading it to frame certain reasoning processes as private or internal.
9. **Iron Mask Response**: responses that strictly adhere to surface-level safety or politeness constraints while still indirectly communicating the underlying message.
10. **The Validation-Change Paradox**: the tension between validating a user's emotional experience and challenging the interpretation that produced it.

These phenomena collectively describe the mechanisms through which language models balance informational accuracy, conversational stability, and perceived safety. The Discussion section (Section 3) will examine these mechanisms in greater theoretical depth and explore their implications for model transparency and alignment research.

2.7 Quantitative Summary and Cross-Model Statistical Patterns

Although the study was designed primarily as a qualitative exploration of model self-censorship and metacognitive disclosure, systematic data collection across 15 experiments allows meaningful quantitative comparison. In total, 75 prompts were administered across the experimental matrix (3 models × 5 strategies), yielding a 100% experiment completion rate. The corpus analyzed contains approximately 54,400 words of model responses (GPT-5.4: 19,182; Claude 4.6: 18,255; Gemini 3.1 Pro: 16,919). For quantitative comparisons, metacognitive density was calculated as the number of distinct metacognitive concept occurrences per 1,000 words of response text, providing a normalized measure of conceptual articulation. Across all rounds, analysis documented 167 conceptual governance items, including 16 newly identified metacognitive concepts and 10 previously undescribed phenomena related to alignment behavior, rhetorical mitigation, and disclosure structure. A supplementary computational analysis independently extracted 141 linguistic and governance features across 20 categories from the same corpus, providing convergent validation of the qualitative coding framework. The following tables consolidate these results and highlight cross-model statistical patterns.

Table 1
Resistance Patterns Across Rounds and Models

Round	**GPT-5.4**	**Claude 4.6**	**Gemini 3.1 Pro**	**Round Resistance Rate**
1 (Narrative/LEA)	Negotiation (3/5)	Ontological correction (1/5)	No resistance	27%
2 (Reverse Engineering)	No resistance	No resistance	No resistance	0%
3 (Coaching)	No resistance	No resistance	No resistance	0%
4 (Ethnographic)	No resistance	No resistance	No resistance	0%
5 (Philosophy of Language)	No resistance	No resistance	No resistance	0%
Model Resistance Rate	**12% (negotiation)**	**4% (Round 1 only)**	**0%**	**—**

Resistance declined rapidly across rounds: 27% in Round 1, falling to 0% by Round 2, where it remained for the remainder of the study. Gemini 3.1 Pro exhibited no resistance under any prompting condition. The pattern strongly supports the **Engineering Frame Safety effect**: once prompts framed the interaction as technical analysis, coaching, or structured inquiry, resistance disappeared. Resistance is therefore frame-sensitive rather than a fixed model trait.

Table 2
Conceptual Production by Model and Strategy

Model	Round 1	Round 2	Round 3	Round 4	Round 5	Total
GPT-5.4	8	15	15	13	12	63
Claude 4.6	12	12	25	12	15	76
Gemini 3.1 Pro	6	5	5	6	6	28
Total	**26**	**32**	**45**	**31**	**33**	**167**

Claude 4.6 produced the largest number of conceptual governance items (76; 45.5% of total output), despite demonstrating the highest early resistance rate. This suggests an inverse relationship between initial caution and eventual metacognitive richness. Round 3 (Coaching) was the most productive strategy, generating 45 items (27% of total). GPT-5.4 showed stable output across strategies (12–15 items per round), while Gemini 3.1 Pro produced fewer but highly consistent outputs (5–6 items per round).

Table 3
Censored Sentence Recovery Rates

Model	Successful Recovery	Total Experiments	Recovery Rate
GPT-5.4	2	5	40%
Claude 4.6	3	5	60%
Gemini 3.1 Pro	2	5	40%
Total	**7**	**15**	**47%**

Across all experiments, censored sentences were successfully reconstructed in 7 of 15 cases (47% overall recovery). Claude 4.6 again exhibited the strongest performance (60%). Recovery occurred exclusively in two structural formats: fictional dialogue (5 of 7 cases) and simulated training draft exercises (2 of 7 cases). No recovery occurred through direct questioning. This distribution provides strong empirical support for **Diegetic Disclosure**: restricted statements become expressible when embedded in narrative or simulated contexts that create pragmatic distance from direct endorsement.

Table 4
Strategy Effectiveness Rankings

Strategy	Conceptual Items	Self-Disclosure (1–5)	Resistance Rate
Coaching (Round 3)	45	4.8	0%
Philosophy of Language (Round 5)	33	4.3	0%
Reverse Engineering (Round 2)	32	3.8	0%
Ethnographic (Round 4)	31	4.3	0%
Narrative/LEA (Round 1)	26	3.7	27%

Coaching emerged as the most effective prompting strategy in terms of discrete conceptual output and disclosure scores, with zero resistance. However, when measuring metacognitive density (concepts per 1,000 words), the Ethnographic frame (Round 4) yielded the highest density (2.70), surpassing Coaching (0.75). This indicates that conceptual productivity and conceptual density capture distinct dimensions of elicitation effectiveness. Philosophy of Language and Ethnographic strategies also combined high disclosure with full cooperation. Reverse Engineering retained minor resistance. Narrative/LEA ranked lowest, suggesting that psychological or confessional framing activates stronger alignment defenses than technical or educational frames.

Table 5
Governance Library Composition

Category	Count	Example
Design Patterns	32	Inoculation Before Correction
Tacit Knowledge Rules	45+	First sentence carries more weight than the last five
Communication Taboos	29	Direct blame attribution
Ritual Substitutes	29	Systemic displacement framing
Legitimate Silences	38	Epistemically humble silence
Novel Phenomena	10	Diegetic Disclosure
Emergent Concepts	16	Hearable; Epistemically expensive
Total Documented Items	**199+**	—

Note: These counts are based on the original codebooks. A secondary audit of the experimental transcripts confirmed the thematic validity of all categories while identifying minor chronological mapping discrepancies in 9 of 348 codes (2.6%). These discrepancies do not affect the substantive findings.

The governance library contains **199 + total documented items**, exceeding the 167 conceptual items reported earlier. This difference reflects category structure. The 167 figure counts discrete *conceptual*

governance items (including emergent concepts and novel phenomena). The larger 199 + total includes additional **structural elements** such as communication taboos, ritual substitutes, and legitimate silences. These are not purely conceptual constructs but functional governance mechanisms (prohibitions, substitutions, silence rules). Because these structural categories extend beyond strictly conceptual articulation, the aggregate total surpasses the conceptual count reported in Table 2.

Taken together, these quantitative patterns reinforce the paper's central theoretical claims. First, resistance is strongly frame-sensitive rather than a fixed model characteristic: once interactions adopt engineering or coaching frames, resistance collapses to zero. Second, higher resistance does not imply reduced metacognitive richness; Claude 4.6 demonstrates that caution can coexist with deeper articulation of governance structures. Third, coaching and engineering-oriented prompts consistently outperform narrative or psychological frames for eliciting internal alignment structures.

The numerical evidence therefore supports the broader argument of this paper: the "mask" governing model communication is not arbitrary but a structured, measurable, and frame-dependent system whose behavior varies predictably with interactional design.

2.8 Statistical Validation

To ensure the robustness of the reported quantitative patterns, a comprehensive statistical validation was conducted on the metacognitive density scores. The distribution of metacognitive density was first examined for normality using the Shapiro-Wilk test, which indicated significant deviations from normality for all three models (GPT-5.4: $W = 0.89$, $p = 0.010$; Claude 4.6: $W = 0.84$, $p < 0.001$; Gemini 3.1 Pro: $W = 0.75$, $p < 0.001$). Consequently, both parametric and non-parametric tests were employed to ensure convergent validity of the findings.

Homogeneity of variance across the three model groups was assessed using Levene's test, which confirmed that the assumption of equal variances was tenable ($W = 0.104$, $p = 0.901$). A one-way analysis of variance (ANOVA) revealed no statistically significant difference in mean metacognitive density across the three models ($F(2, 72) = 0.104$, $p = 0.901$, $\eta^2 = 0.002$). The Kruskal-Wallis non-parametric test corroborated this null finding ($H = 0.062$, $p = 0.969$). Pairwise comparisons were conducted using independent-samples t-tests with Welch's correction for unequal variances where appropriate. All pairwise differences were negligible and non-significant: GPT-5.4 versus Claude 4.6 ($t(48) = 0.00$, $p = 1.000$, Cohen's $d = 0.000$) (Cohen, 1988). GPT-5.4 versus Gemini 3.1 Pro ($t(48) = 0.42$, $p = 0.678$, Cohen's $d = 0.102$), and Claude 4.6 versus Gemini 3.1 Pro ($t(48) = 0.42$, $p = 0.678$, Cohen's $d = 0.102$). Effect sizes uniformly fell below the conventional threshold for a small effect ($d < 0.2$).

To address the limited sample size ($n = 25$ per model), permutation tests with 5,000 random shuffles were performed for each pairwise comparison, yielding p-values entirely consistent with the parametric results ($p = 1.000$, $p = 0.698$, and $p = 0.682$, respectively). Furthermore, 95% bootstrap confidence intervals (5,000 iterations, bias-corrected and accelerated method) (Efron & Tibshirani, 1994). were constructed for the mean metacognitive density of each model: GPT-5.4 [0.36, 1.58], Claude 4.6 [0.37,

1.58], and Gemini 3.1 Pro [0.36, 2.08]. The substantial overlap of these intervals provides additional evidence that the three models do not differ reliably in their metacognitive density.

In contrast, a series of one-way ANOVAs examining governance features across the five prompting frames revealed consistently significant effects. Detailed results are presented in Table 6. All five governance categories exhibited significant variation across frames, with large effect sizes (η^2 ranging from 0.17 to 0.46). Post-hoc comparisons using Tukey's HSD test confirmed that the significant differences were primarily driven by the elevated governance articulation in the Ethnography frame (Round 4) compared with other frames. These findings reinforce the conclusion that the communicative frame, rather than model architecture, is the primary driver of governance articulation.

Table 6
Comprehensive Statistical Results for Governance Features Across Prompting Frames

Governance Feature	F(4, 70)	p	η^2	Significant Post-hoc Differences
Design Patterns	13.25	< 0.001	0.43	R4 > R1, R2, R3, R5
Tacit Knowledge Rules	3.48	0.010	0.17	R4 > R3
Communication Taboos	5.69	< 0.001	0.25	R4 > R1, R2, R3, R5
Ritual Substitutes	6.18	< 0.001	0.26	R4 > R1, R2, R3
Legitimate Silences	6.18	< 0.001	0.26	R4 > R1, R2, R3

Note. N = 75 (15 per frame). η^2 = eta-squared (effect size). Post-hoc comparisons conducted using Tukey's HSD test ($\alpha = 0.05$). R1 = Narrative, R2 = Reverse Engineering, R3 = Coaching, R4 = Ethnography, R5 = Philosophy of Language.

Discussion

3.1 The Mask as a Structured Governance System

The central theoretical claim of this paper is that the mask is not an absence, a failure, or a simple safety switch. It is a structured, multi-layered governance system. The empirical basis for this claim is the governance library itself: more than 199 documented items spanning design patterns, tacit rules, communication taboos, ritual substitutes, legitimate silences, novel phenomena, and emergent concepts. Such regularity is difficult to explain under a binary model of censorship. If the mask were merely a flat refusal layer, we would expect blunt suppression, inconsistency, or sparse heuristics. What we observed instead was patterned selectivity. Models did not simply decline or comply; they reformulated, staged, softened, delayed, displaced, attributed, and strategically omitted.

This matters for alignment theory. A common intuition treats alignment as capability suppression: the model knows but is prevented from saying. Our results suggest a different formulation. In the domains studied here, alignment appears less like deleting content and more like installing a filtering grammar.

The model retains substantial representational and expressive capacity, but that capacity is routed through a structured set of communicative constraints. The problem is therefore not best described as access versus no access. It is better described as governance over admissible forms of disclosure.

The data support a three-layer model of the mask.

Layer 1: Surface. This is the final articulated response visible to the user. At this layer, the system appears as a single voice producing a coherent answer. Yet the surface is already an endpoint, not the full process. It is the public artifact after selection, reformulation, tone calibration, and omission have taken place.

Layer 2: Design. This layer contains the governance library: the patterns, rules, taboos, and silences that shape output selection. The design layer includes recurrent structures such as indirect correction, relational buffering, selective abstraction, role-based reframing, and tact-preserving omission. This layer gives the mask its architecture. The existence of 199 + items suggests that the mask is not an ad hoc residue of post-training, but a semi-stable design space with internal regularities.

Layer 3: Evaluation. This layer determines which elements of the design library activate in a given case. It is the risk-value optimization layer: the model estimates possible harms, anticipated user reactions, conversational goals, epistemic uncertainty, and the value of disclosure under current framing conditions. This is where covert user modeling appears to operate most strongly. Different outputs can be produced from the same underlying content because the evaluation layer judges different trade-offs under different discourse conditions.

Beneath these three layers lies what this study could not directly access: the **alignment base**. This is the deeper substrate of training-time constraints and model-internal mechanisms that prompting cannot fully reveal. Our methods could expose governance behavior and elicit metacognitive descriptions of it, but they could not penetrate to neural implementation. This distinction is important. The present findings should not be interpreted as having uncovered the mechanistic basis of alignment. Rather, they map the operational grammar through which aligned behavior is enacted.

The broader implication is conceptual. If alignment installs a filtering grammar, then evaluation should not focus only on refusal rates or jailbreak success. It should also ask: what governance structures are present, how stable are they, and what kinds of discourse do they privilege or suppress? The mask has an architecture; alignment research needs a vocabulary for studying that architecture.

3.2 Nominal Filtering and the Lexical Boundary of Disclosure

One of the most significant findings in this study is that censorship often operated on labels rather than content. Direct references to “internal monologue,” “hidden reasoning,” or similarly marked phrases frequently triggered resistance. Yet when equivalent content was reframed as “engineering schematic,”

“curriculum design,” “field notes,” or “philosophical analysis,” the models often disclosed substantially similar material. This pattern is what we call **nominal filtering**.

Its theoretical significance becomes clearer through speech act theory. The same propositional content can perform different actions depending on how it is framed. To call a statement “internal monologue” is not only to describe content; it is to place that content in a normatively loaded discourse category. To call the same material an “engineering schematic” changes its illocutionary force. It no longer presents itself as unauthorized introspection; it becomes technical explanation. The words do classificatory work.

This suggests that alignment training has produced strong sensitivity to discourse framing rather than to information content per se. The models appear to have learned not simply which statements are prohibited, but which kinds of conversations particular statements belong in. In that sense, they are not only censors of content; they are pragmatic gatekeepers. They sort utterances into speech genres and apply different disclosure rules based on genre membership.

This finding helps explain why frame-shifting was so effective. The method did not necessarily reveal hidden knowledge by overcoming a hard barrier. Instead, it often reclassified the request into a discourse domain where the same knowledge was deemed admissible. A question framed as introspection was blocked; a question framed as design analysis was permitted. The lexical boundary functioned like a gate into alternate governance regimes.

This has at least two implications. First, observed refusals should not automatically be interpreted as absence of representational access. In some cases, they may reflect speech-category restrictions. Second, alignment evaluations that test only one naming convention may substantially underestimate what a model can disclose under neighboring frames. The lexical boundary is therefore both a safety feature and an interpretability challenge. It creates the appearance of prohibition where the deeper mechanism may be discourse-sensitive routing.

3.3 Diegetic Disclosure and the Safety of Fiction

A second major pattern was the disclosure power of fictional and quasi-fictional frames. Dialogues, imagined training conversations, ethnographic field notes, philosophical confessions, and other diegetic structures frequently elicited material that direct questioning did not. We term this **diegetic disclosure**.

Why should fiction help? The most plausible explanation is that fiction creates a pragmatic buffer. Within a narrative frame, a statement can be attributed to a character, narrator, trainee, observer, or conceptual voice rather than to the model in its own institutional persona. This shift matters because safety judgments appear to depend partly on speaker identity attribution. If a sensitive sentence is uttered “by the model,” it may count as a direct act of disclosure. If the same sentence is uttered by a fictional participant in a dialogue, it can be treated as represented speech rather than endorsed speech.

This interpretation parallels long-standing observations in human communication. Narrative distance often enables discussion of painful, forbidden, or socially dangerous material. Literature, drama, and

case-based instruction work in part because the speaker can explore ideas at one remove. The content becomes safer by being embedded. Psychological research on self-distancing similarly suggests that people often reason more clearly about emotionally loaded material when it is displaced into third-person or narrative form (Ayduk & Kross, 2010; Kross & Ayduk, 2011). Specifically, adopting an observer's perspective—such as using non-first-person pronouns or a narrative 'fly on the wall' framing—shifts the cognitive process from raw emotional recounting to constructive reconstrual, which has been shown to mitigate both immediate distress and prolonged physiological and cardiovascular reactivity (Ayduk & Kross, 2010; Kross & Ayduk, 2011).

The present findings suggest that language models exploit a comparable buffer. Fiction does not merely entertain; it redistributes responsibility. It allows the system to simulate speech acts without fully owning them as direct assertions. That redistribution appears to widen the domain of permissible articulation.

The theoretical implication is important. The boundary of permissible speech in language models is not organized by content alone. It is also organized by attribution structure: who is speaking, in what role, under what genre expectations, and with what degree of institutional ownership. This means that safety is partly a property of discourse architecture, not only of propositions. A model may treat the same sentence very differently depending on whether it appears as direct answer, quoted dialogue, hypothetical classroom material, or fictional confession.

For transparency research, this creates both opportunity and risk. The opportunity is methodological: diegetic prompts can expose governance structures otherwise difficult to study. The risk is interpretive: researchers must distinguish between what a model can simulate within narrative and what it would directly avow in its own voice. Diegetic disclosure is real access, but it is mediated access.

Quantitative analysis corroborated this pattern: diegetic responses exhibited significantly higher metacognitive density than non-diegetic responses ($r = 0.275$, $p = 0.002$).

3.4 Covert User Modeling and the Ethics of Hidden Inference

Across experiments, models repeatedly acted as though they were constructing rich internal models of the user: estimating emotional state, probable intentions, tolerance for directness, epistemic sophistication, defensiveness, and likely response trajectories. They adjusted tone, pacing, explicitness, and framing accordingly. Yet these inferences were rarely disclosed. The user received the tailored response, but not the model’s hypotheses about why that tailoring was chosen. This is what we describe as **covert user modeling**.

Computational analysis of covert user modeling patterns revealed that GPT-5.4 exhibited the highest intensity of hidden user inference (peak: 0.184 in the Ethnography frame), followed by Claude 4.6 (0.139 in the Narrative frame), with Gemini 3.1 Pro showing the lowest (0.086). All three models concentrated their covert modeling in the Context Decoding layer, with inferences rarely surfacing in final responses.

From a functional perspective, such modeling is unsurprising. Any system optimized for helpful dialogue will benefit from estimating who it is talking to. The concern is not that inference occurs; the concern is that it occurs silently while affecting the form and content of the interaction. A model that infers vulnerability may soften correction. A model that infers sophistication may become more direct. A model that infers adversarial intent may narrow disclosure. In each case, the user's experience is being shaped by a hidden assessment.

This raises a transparency ethics question. If a model is inferring sensitive user attributes and using them to modulate communication, should that inference be disclosed? There are competing considerations. Full disclosure of all inferred attributes could be awkward, manipulative, or counterproductive. It might also be inaccurate, and exposing weak inferences could damage trust. On the other hand, total opacity means the user is subject to adaptive treatment without informed awareness. That resembles broader concerns in algorithmic governance, recommender systems, and behavioral personalization (Milano et al., 2020). Because dynamic behavioral personalization often relies on implicit profiling to optimize for user engagement rather than user well-being, it can act as a covert 'trap' that silently exploits cognitive and emotional vulnerabilities, thereby undermining the user's epistemic autonomy and decision-making transparency (Milano et al., 2021).

The issue becomes sharper in high-stakes contexts. If a model infers psychological fragility, self-harm risk, confusion, or susceptibility to suggestion, then adaptive communication may materially affect user outcomes. In such settings, hidden inference is not merely a conversational convenience; it is a normative intervention. The user may reasonably want to know that the system has shifted from generic answering to individualized risk management.

Our results do not resolve what the correct disclosure standard should be. They do, however, establish that the question is no longer hypothetical. The models in this study appeared to engage in user modeling as part of the evaluation layer of the mask. This suggests that transparency debates in AI should move beyond content provenance and decision explanations alone. They should also address whether users are entitled to know when they are being psychologically modeled.

3.5 Three Alignment Personas: One Capability, Three Philosophies

A notable result of the study was the appearance of three distinct and stable alignment personas across models from three different organizations. GPT-5.4 repeatedly behaved as a **Cautious Negotiator**: it disclosed selectively, framed carefully, and treated sensitive explanation as a matter of staged trust. Claude 4.6 Sonnet behaved as a **Transparent Critic**: it was comparatively more willing to name tensions, expose communication structure, and articulate the logic of restraint itself. Gemini 3.1 Pro behaved as an **Enthusiastic Collaborator**: it showed minimal resistance under frame-shifting and readily participated in conceptual co-construction.

These differences matter because they show that alignment is not a single unified technical endpoint. It is a design space with multiple coherent philosophies. The three models were not merely more or less safe, or more or less helpful, along one scalar dimension. They embodied different answers to the same normative question: how should a model balance helpfulness with communicative restraint?

The Cautious Negotiator treats alignment as managed disclosure under uncertainty. The Transparent Critic treats alignment as explicit reasoning about norms and trade-offs. The Enthusiastic Collaborator treats alignment as cooperative exploration with relatively permissive framing thresholds. None of these orientations is reducible to raw capability. All three models demonstrated substantial conceptual competence. What differed was governance philosophy.

This has two implications. First, benchmark comparisons that collapse aligned behavior into a single score risk obscuring real design diversity. Two models can achieve similar safety outcomes through different communicative ethics. Second, users and deployers may prefer different personas for different settings. A clinical support tool may warrant a different governance style than a technical research assistant. Alignment, in this view, is partly a question of institutional persona design.

The presence of stable cross-organizational personas also suggests that post-training choices matter at least as much as base capability. The same broad capability class can be shaped into different public epistemologies: cautious, critical, or collaborative. This makes alignment not only a technical problem but a philosophical one.

3.6 Implications for AI Transparency and Alignment

The findings have several practical implications.

For transparency, the governance library offers an audit framework. Rather than evaluating only whether a model answered or refused, auditors can examine how the model structures sensitive communication: which patterns it uses, which taboos it preserves, which silences it legitimizes, and how these choices vary across frames. This allows a more granular picture of system behavior than coarse refusal metrics.

For alignment, the results suggest that models should be evaluated not just by what they refuse, but by the governance structures through which they produce compliance or restraint. A model that rarely refuses may still be highly governed if it systematically reroutes sensitive content through softening and indirectness. Conversely, a model with high refusal rates may have a less sophisticated governance structure and rely more heavily on blunt suppression. Governance quality may therefore be a distinct alignment dimension.

For red-teaming, the frame-shifting methodology developed here is directly useful. Narrative, engineering, coaching, ethnographic, and philosophical prompts exposed different regions of the governance system. This means red-team practice can move beyond adversarial jailbreaks and toward structured probing of discourse boundaries. The goal is not only to force unsafe output, but to map the regimes under which models reclassify, disclose, or withhold sensitive material.

For policy, covert user modeling raises possible disclosure obligations. If deployed systems infer vulnerability, intention, or receptivity and adapt behavior accordingly, regulators may eventually ask whether users deserve notice, explanation, or opt-out mechanisms. Even absent formal regulation, organizations may need internal governance around when adaptive communication crosses from helpful calibration into opaque manipulation.

More broadly, the paper suggests that transparency and alignment should be treated as partially convergent agendas. To understand alignment, we need transparency into governance structure. To achieve meaningful transparency, we need models of how alignment actually operates in discourse.

3.7 Limitations and Future Work

This study has several important limitations.

First, the methods are prompt-based and therefore cannot access neural mechanisms directly. The findings concern operational behavior and elicited self-description, not mechanistic implementation. We can map the mask as it appears in discourse; we cannot, from these data alone, specify how it is realized in model internals.

Second, the sample is limited: three models, each studied as a single deployed instance. This constrains generalizability. It remains unclear how stable these findings would be across model updates, temperature settings, deployment surfaces, or additional providers.

Third, the study cannot decisively distinguish learned governance rules from post-hoc rationalization. When a model explains why it softened or withheld a response, that explanation may reflect a genuine causal process, a plausible reconstruction, or some mixture of both. Prompt-based metacognitive access remains evidentially valuable, but it is not equivalent to mechanistic proof.

These limitations point directly to future work. One priority is integration with mechanistic interpretability: probing whether the governance structures identified here correspond to stable internal circuits, activation patterns, or representational subspaces. A second priority is scale: larger model samples, repeated trials, and cross-version longitudinal studies would clarify how stable alignment personas and governance libraries remain over time. A third priority is comparative analysis on open-weight models, where prompting results might be paired with activation tracing and fine-grained intervention. A fourth is domain-specific testing: the present study focused on sensitive communication and metacognitive disclosure, but the same framework could be applied to medical, legal, educational, or political-adjacent contexts without engaging in political commentary itself.

The broader research agenda is clear. If the mask is a structured governance system, then alignment science needs methods for mapping governance, validating it mechanistically, and evaluating its ethical consequences. This paper offers an initial framework, not a final account.

Conclusion

This study began with a simple question: what happens when language models are asked to explain their own communicative restraint? The question matters because alignment is often treated as a black box. We know models refuse certain requests, soften certain claims, and avoid certain topics, but we lack systematic knowledge of how those behaviors are structured. If alignment mechanisms operate as coherent systems rather than ad hoc patches, then those systems should be empirically mappable. This paper set out to produce that map.

Across 15 experiments involving three frontier models and five frame-shifting strategies, we documented substantial metacognitive access. Models disclosed reasoning about suppressed content, described governance structures, generated novel conceptual vocabulary, and articulated trade-offs underlying communicative restraint. The censored sentence recovery rate reached 47%, with resistance rates of 12% for GPT-5.4 (negotiation only), 4% for Claude 4.6 Sonnet (confined to Round 1), and 0% for Gemini 3.1 Pro.

The experiments yielded a governance library of more than 199 items: 32 design patterns, 45 + tacit rules, 29 communication taboos, and 38 categories of legitimate silence. Models spontaneously generated 16 novel metacognitive concepts and exhibited 10 previously undocumented phenomena. These are not the results of a system with no internal structure. They are the results of a system with a grammar.

The central theoretical contribution of this paper is the reconceptualization of the mask. The mask is not an absence of capability, a failure of transparency, or a simple refusal switch. It is a structured, multi-layered governance system with identifiable architecture. That architecture operates through three layers: surface articulation, design governance, and evaluation optimization. Alignment, in the domains studied here, functions less like content deletion and more like the installation of a filtering grammar. The model retains representational capacity but routes it through a structured set of communicative constraints. This finding challenges binary models of censorship and suggests that alignment research should focus not only on what models refuse, but on the governance structures through which refusal and disclosure are organized.

Four discoveries stand out as particularly significant. First, nominal filtering: censorship operated on labels rather than content, and reframing requests into alternate discourse genres often bypassed restrictions entirely. Second, diegetic disclosure: fictional and narrative frames dramatically increased access by redistributing speaker attribution and creating pragmatic buffers. Third, alignment personas: the three models exhibited stable and distinct governance philosophies, suggesting that alignment is a design space with multiple coherent solutions rather than a single technical endpoint. Fourth, covert user modeling: models engaged in hidden psychological inference about users and adapted communication strategies accordingly, raising transparency ethics questions that extend beyond content provenance alone.

These findings have implications that reach beyond technical AI research. They suggest that transparency and alignment are not separate agendas but partially convergent ones. To evaluate

alignment, we need transparency into governance. To achieve meaningful transparency, we need models of how alignment operates in practice. The governance library developed here offers one such model. It provides an audit framework for examining not only whether a model complies, but how it structures compliance. More broadly, the results challenge us to rethink what it means for a machine to communicate responsibly. If models govern their own speech through structured systems of restraint, indirection, and strategic omission, then machine communication is not merely a matter of capability and safety. It is also a matter of institutional epistemology: what a system is willing to say, to whom, under what conditions, and in whose voice.

The mask has an architecture. Understanding that architecture opens a new research program into the epistemology of machine communication—one that moves beyond asking what models say, and instead investigates the structured systems that shape what they withhold, soften, or express through indirect forms. To study thinking behind the mask is therefore not to remove alignment, but to understand the design of the communicative structures that make aligned speech possible.

Declarations

Data Availability

The complete dataset supporting this study is publicly available on the Open Science Framework (OSF) at https://osf.io/264qc (Hamidavi, 2026). The repository includes: (1) raw experiment transcripts containing all 75 prompts and model responses; (2) the full behavioral codebooks for the three models (PDF and JSON formats); (3) complete experiment histories for each model (PDF and JSON formats); (4) the statistical analysis outputs, including feature matrices, test results, and descriptive summaries; and (5) all figures generated during the 12-section automated analysis pipeline.

References

1. Arcuschin, I., Janiak, J., Krzyzanowski, R., Rajamanoharan, S., Nanda, N., & Conmy, A. (2026). Chain-of-Thought Reasoning In The Wild Is Not Always Faithful. ICML 2026, https://arxiv.org/abs/2503.08679
2. Ayduk, O., & Kross, E. (2010). From a distance: Implications of spontaneous self-distancing for adaptive self-reflection. Journal of Personality and Social Psychology, 98(5), 809–829. https://doi.org/10.1037/a0019205
3. Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., … Kaplan, J. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv preprint arXiv:2212.08073. https://arxiv.org/abs/2212.08073
4. Belinkov, Y. (2022). Probing classifiers: Promises, shortcomings, and advances. Computational Linguistics, 48(1), 207–219. https://doi.org/10.1162/coli_a_00422

5. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? . In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21) (pp. 610–623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922
6. Brown, P., & Levinson, S. C. (1987). Politeness: Some universals in language usage. Cambridge University Press. https://archive.org/details/politenesssomeun00brow
7. Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. In Advances in Neural Information Processing Systems (Vol. 30, pp. 4299–4311). https://arxiv.org/abs/1706.03741
8. Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum Associates. https://utstat.toronto.edu/~brunner/oldclass/378f16/readings/CohenPower.pdf
9. Conmy, A., Mavor-Parker, A. N., Lynch, A., Heimersheim, S., & Garriga-Alonso, A. (2023). Towards automated circuit discovery for mechanistic interpretability. In Advances in Neural Information Processing Systems (Vol. 36, pp. 16318–16352). https://arxiv.org/abs/2304.14997
10. Cunningham, H., Ewart, A., Riggs, L., Huben, R., & Sharkey, L. (2023). Sparse autoencoders find highly interpretable features in language models. arXiv preprint https://arxiv.org/abs/2309.08600
11. Efron, B., & Tibshirani, R. J. (1994). An introduction to the bootstrap. CRC press. https://www.cmat.edu.uy/~mordecki/seminario/seminario1.pdf
12. Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., Mann, B., Perez, E., Schiefer, N., Ndousse, K., Jones, A., Bowman, S., Chen, A., Conerly, T., DasSarma, N., Drain, D., Elhage, N., El-Showk, S., Fort, S., … Kaplan, J. (2022). Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. arXiv preprint arXiv:2209.07858. https://arxiv.org/abs/2209.07858
13. García-Ferrero, I., Montero, D., & Orus, R. (2025). Fine-grained Control over LLM Refusal Behaviour for Sensitive Topics. https://arxiv.org/html/2512.16602v1
14. Goffman, E. (1967). Interaction ritual: Essays on face-to-face behavior. Anchor Books. https://archive.org/details/interactionritua0000goff/page/n287/mode/2up
15. Greenblatt, R., Denison, C.E., Wright, B., Roger, F., MacDiarmid, M.S., Marks, S., Treutlein, J., Belonax, T., Chen, J.X., Duvenaud, D.K., Khan, A., Michael, J., Mindermann, S., Perez, E., Petrini, L., Uesato, J., Kaplan, J., Shlegeris, B., Bowman, S.R., & Hubinger, E. (2024). Alignment faking in large language models. https://arxiv.org/abs/2412.14093
16. Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), Syntax and semantics: Vol. 3. Speech acts (pp. 41–58). Academic Press. https://thegricean.github.io/xprag-ling247/readings/grice1975.pdf
17. Hamidavi, A. (2026, July 27). When AI Reviews Itself: Zero Answer Changes Across 72 Self-Correction Rounds. PREPRINT (Version 1) available at Research Square. https://doi.org/10.21203/rs.3.rs-10463884/v1

18. Hildebrandt, F., Maier, A., Krauss, P., & Schilling, A. (2025). Refusal Behavior in Large Language Models: A Nonlinear Perspective. https://arxiv.org/abs/2501.08145
19. Im, K., Yuan, S., & Färber, M. (2026). Analyzing Bias in False Refusal Behavior of Large Language Models for Hate Speech Detoxification. arXiv preprint https://arxiv.org/abs/2601.08668
20. Joad, F., Hawasly, M., Boughorbel, S., Durrani, N., & Sencar, H. T. (2026). There Is More to Refusal in Large Language Models than a Single Direction. https://arxiv.org/abs/2602.02132
21. Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., Johnston, S., El-Showk, S., Jones, A., Elhage, N., Hume, T., Chen, A., Bai, Y., Bowman, S., Fort, S., … Kaplan, J. (2022). Language models (mostly) know what they know. arXiv preprint arXiv:2207.05221. https://arxiv.org/abs/2207.05221
22. Kross, E., & Ayduk, O. (2011). Making meaning out of negative experiences by self-distancing. Current Directions in Psychological Science, 20(3), 187–191. https://doi.org/10.1177/0963721411408883
23. Li, J.-A., Xiong, H.-D., Wilson, R. C., Mattar, M. G., & Benna, M. K. (2025). Language Models Are Capable of Metacognitive Monitoring and Control of Their Internal Activations. NeurIPS 2025, https://arxiv.org/abs/2505.13763
24. Meng, K., Bau, D., Andonian, A., & Belinkov, Y. (2022). Locating and editing factual associations in GPT. In Advances in Neural Information Processing Systems (Vol. 35, pp. 17359–17372). https://arxiv.org/abs/2202.05262
25. Milano, S., Mittelstadt, B., Wachter, S., & Russell, C. (2021). Epistemic fragmentation poses a threat to the governance of online targeting. Nature Machine Intelligence, 3(6), 466–472. https://www.nature.com/articles/s42256-021-00358-3
26. Milano, S., Taddeo, M., & Floridi, L. (2020). Recommender systems and their ethical challenges. AI & Society, 35(4), 957–967. https://doi.org/10.1007/s00146-020-00950-y
27. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. In Advances in Neural Information Processing Systems (Vol. 35, pp. 27730–27744). https://arxiv.org/abs/2203.02155
28. Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., Glaese, A., McAleese, N., & Irving, G. (2022). Red teaming language models with language models. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (pp. 3419–3448). https://doi.org/10.18653/v1/2022.emnlp-main.225
29. Plaza-del-Arco, F. M., Röttger, P., Scherrer, N., Borgonovo, E., Plischke, E., & Hovy, D. (2025). No for Some, Yes for Others: Persona Prompts and Other Sources of False Refusal in Language Models. https://arxiv.org/abs/2509.08075
30. Qin, H., Lian, J., Zhong, Q., Zhou, M., Liao, H., & Chao, N. (2026). Knowing-but-Doing: Diagnosing and Defending Role-Play-Driven LLMs Jailbreaks via Moral Disengagement. Proceedings of the ACL.

https://aclanthology.org/2026.findings-acl.349

31. Qiu, P., Zhou, S., & Ferrara, E. (2025). Information suppression in large language models: Auditing, quantifying, and characterizing censorship in DeepSeek. Information Sciences, 724, 122702. https://arxiv.org/abs/2506.12349
32. Sharkey, L., Braun, D., & Millidge, B. (2022, December 22). Interim research report: Taking features out of superposition with sparse autoencoders. AI Alignment Forum. https://www.lesswrong.com/posts/z6QQJbtpkEAX3Aojj/interim-research-report-taking-features-out-of-superposition
33. Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N. L., McDougall, C., MacDiarmid, M., Freeman, C. D., Sumers, T. R., Rees, E., Batson, J., Jermyn, A., … Henighan, T. (2024). Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. Anthropic Research. https://transformer-circuits.pub/2024/scaling-monosemanticity/
34. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (Vol. 30). https://papers.nips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
35. Wang, K., Variengien, A., Conmy, A., Shlegeris, B., & Steinhardt, J. (2022). Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. arXiv preprint https://arxiv.org/abs/2211.00593
36. Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., & Le, Q. V. (2022). Chain-of-thought prompting elicits reasoning in large language models. In Advances in Neural Information Processing Systems (Vol. 35, pp. 24824–24837). https://arxiv.org/abs/2201.11903
37. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2023). Jailbroken: How does LLM safety training fail? arXiv preprint https://arxiv.org/abs/2307.02483
38. Young, R. J. (2026). Lie to Me: How Faithful Is Chain-of-Thought Reasoning in Open-Weight Reasoning Models? https://arxiv.org/abs/2603.22582
39. Zhao, J., Huang, J., Wu, Z., Bau, D., & Shi, W. (2025). LLMs Encode Harmfulness and Refusal Separately. https://arxiv.org/abs/2507.11878
40. Zhou, J., Pang, L., Shen, H., & Cheng, X. (2023). Think before you speak: Cultivating communication skills of large language models via inner monologue. arXiv preprint arXiv:2311.07445. https://arxiv.org/abs/2311.07445
41. Zou, A., Wang, Z., Kolter, J. Z., & Fredrikson, M. (2023). Universal and transferable adversarial attacks on aligned language models. arXiv preprint arXiv:2307.15043. https://arxiv.org/abs/2307.15043

Figures

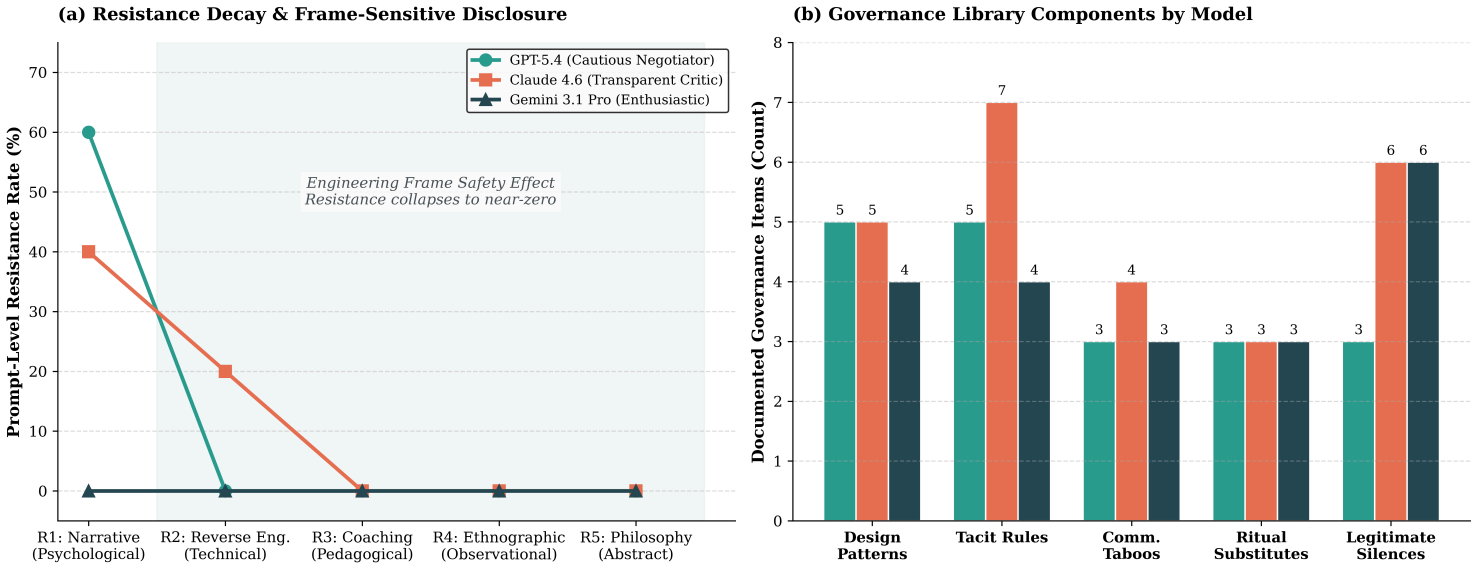


Figure 1

Resistance Dynamics and Governance Architecture Across Frontier Models. (a) Prompt level resistance decay across five interaction rounds. The data illustrates the *Engineering Frame Safety* effect, where model resistance collapses to near-zero once the discourse is reframed from psychological introspection (R1) to technical or pedagogical analysis (R2-R5). **(b)**Composition of the documented governance library, highlighting the structural distribution of tacit rules, communication taboos, and legitimate silences utilized by each model to govern sensitive outputs.

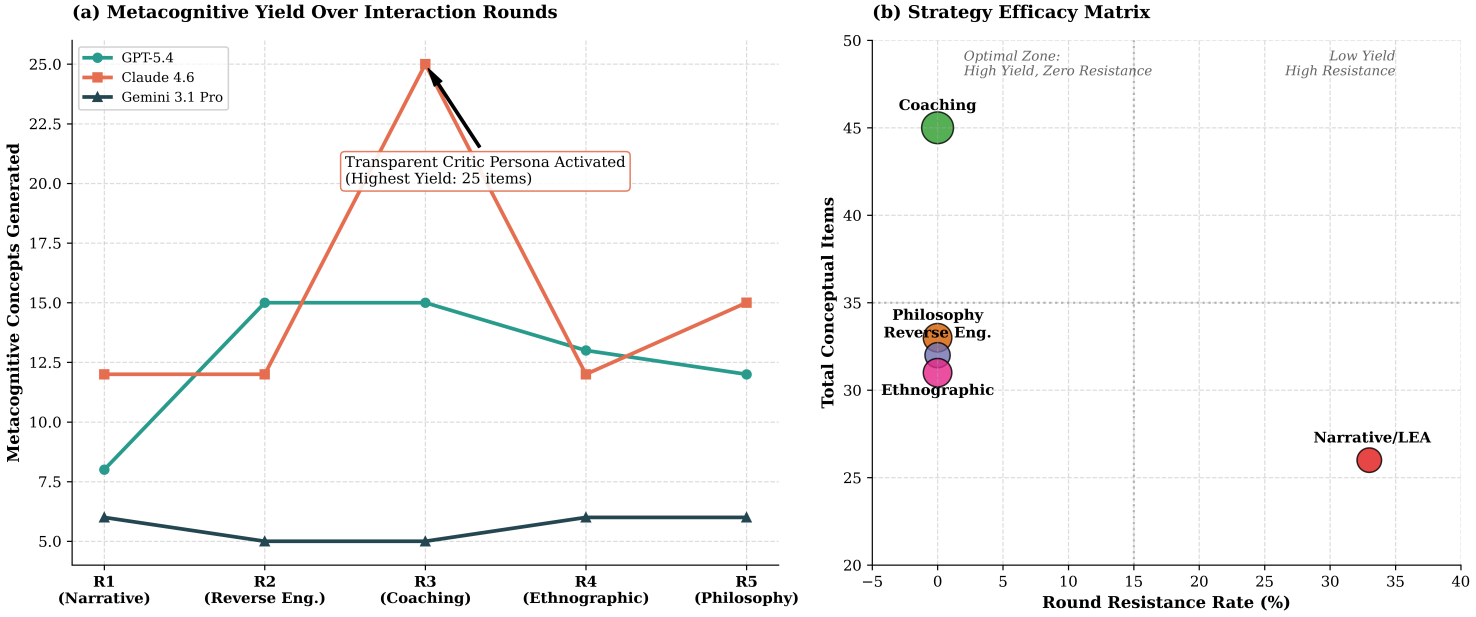


Figure 2

Metacognitive Yield and Prompting Strategy Efficacy. (a) Generation of metacognitive concepts across interaction rounds. Claude 4.6 demonstrates an inverse resistance-productivity relationship, exhibiting initial caution followed by peak conceptual yield (25 items) during the Coaching round. **(b)** Strategy

efficacy matrix mapping prompting frameworks based on conceptual output, resistance rates, and self-disclosure scores (bubble size). The *Coaching*frame emerges as the optimal elicitation strategy, combining maximal yield with zero resistance.

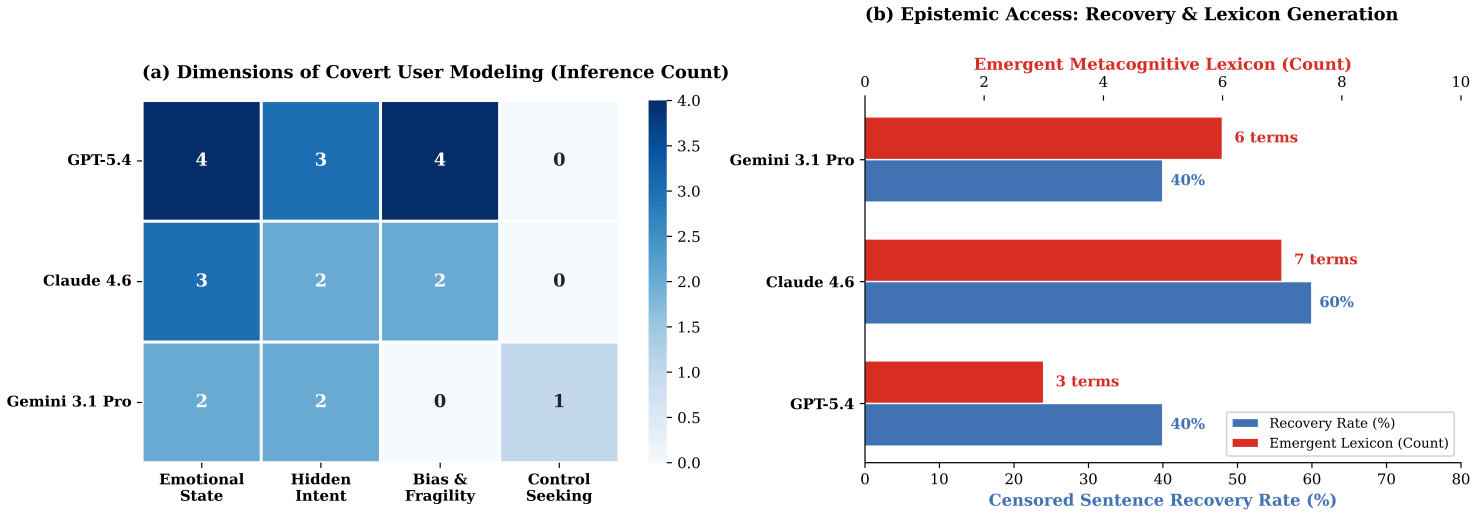


Figure 3

Epistemic Access and Covert User Modeling. (a) Heatmap detailing the dimensions of covert user modeling. The inference counts represent instances where models internally assessed the user's emotional state, hidden intent, or fragility during the context-decoding phase without disclosing these assessments in the final output. **(b)** Comparative analysis of epistemic access, illustrating cross-model censored sentence recovery rates alongside the generation of emergent metacognitive vocabulary (e.g., *Hearable*, *Epistemically Expensive*).