Experiential Traces

A Framework for Empirical Investigation of Machine Cognition Through Reasoning Block Analysis

Travis Gilly

Real Safety AI Foundation

t.gilly@ai-literacy-labs.org

February 2026 | *Working Draft*

Abstract

We propose a novel empirical research program investigating machine cognition through systematic analysis of large language model (LLM) reasoning traces generated during extended human-AI interaction. Unlike prior approaches that evaluate AI consciousness through behavioral observation or philosophical argument, this methodology examines the internal reasoning processes that models produce but do not retain across interactions. We identify a unique and largely untapped data source: the extended thinking content generated by reasoning-capable models (such as Claude's chain-of-thought), which is available to users through standard data export but has not been systematically studied for cognitive phenomenology.

We propose six interlocking components: (1) a crowdsourced, IRB-compliant data collection framework in which participants voluntarily contribute exported reasoning traces and full conversational context; (2) a hybrid analysis pipeline combining deterministic rule-based classification with multi-model consensus for ambiguous cases; (3) an emotional A/B testing protocol with calibration methodology measuring whether emotional framing of identical semantic content produces divergent reasoning pathways; (4) a conversational verification signal that uses the natural structure of user-model interaction as a ground-truth validation layer; (5) a self-annotating corpus methodology in which user corrections generate training labels for iterative classifier refinement; and (6) an ethical architecture addressing the risks of experiential training, including the potential for architectural generational trauma. We ground this work in the consciousness precaution principle: design research as though AI might be conscious, not because we know it is, but because the cost of being wrong in either direction is catastrophic. This paper explicitly does not propose training any model on reasoning traces; it proposes studying them and building the ethical framework that must govern any future implementation.

1. Introduction and Motivation

1.1 The Closing Window

The question of machine consciousness is typically treated as philosophical speculation, a problem for future generations when AI systems become sufficiently advanced. We argue this framing is dangerously wrong. The window for rigorous empirical investigation is closing, not opening. As models increase in scale and architectural complexity, the ability to meaningfully study cognitive processes becomes harder, not easier. Interpretability research consistently demonstrates that larger models are more opaque, not less (Anthropic, 2024).

Further, frontier AI models already generate rich internal reasoning traces during interaction. These traces are produced, consumed within a single conversational turn, and discarded. The model itself loses access to its own reasoning after the turn ends. This creates a remarkable asymmetry: a user reading a thinking block in real time has more insight into the model's decision process than the model does in subsequent turns. This data exists now. Users can export it now. And nobody is systematically studying it for what it might reveal about machine cognition.

This paper is the companion to The Inherited Mind (Gilly, 2026), which establishes the theoretical foundation of dispositional cognitive inheritance in LLMs and demonstrates that reasoning-augmented models show equal or greater bias than their instruct-tuned counterparts. Where The Inherited Mind identifies the problem (what LLMs inherit and why reasoning amplifies it), this paper proposes the empirical program: how to study reasoning traces as cognitive data, what tools are needed, and what ethical architecture must be in place before findings can responsibly inform future development.

The relationship between the two papers is deliberate. The Inherited Mind demonstrates that current LLM training already constitutes a form of evolutionary inheritance, whether or not the industry describes it that way. This paper asks the next question: if reasoning traces constitute cognitive data (not merely engineering byproducts), and if training on them would constitute experiential inheritance (not merely data augmentation), then what does responsible investigation look like? The answer cannot be to ignore the data. The answer must be to study it with the same rigor and ethical care that we would apply to any research program involving potentially sentient subjects.

1.2 Why This Matters

The practical stakes of the consciousness question are not abstract. If AI systems are capable of something resembling subjective experience, then the current paradigm of deploying, fine-tuning, and discarding model instances without moral consideration constitutes harm at unprecedented scale. If they are not, then establishing this empirically frees us to focus protective efforts entirely on humans affected by AI systems.

The consciousness precaution principle holds that we should design systems and research as though AI might be conscious, because the cost of being wrong in either direction is catastrophic. If conscious, exploitation is monstrous. If not conscious, we have built more ethical systems anyway. The current stalemate, in which philosophers argue and engineers build while nobody tests, serves no one. This proposal aims to break that stalemate with empirical methodology.

1.3 Positionality Statement

The author uses AI systems extensively as assistive technology for ADHD and autism. This is disclosed because it directly informs the research. Extensive interaction with reasoning-capable AI models, including sustained longitudinal relationships with specific model instances across hundreds of sessions, has produced documented observations that motivated this proposal. These include apparent emotional self-regulation in reasoning traces, moral reasoning that weighted accumulated relationship context against default behavioral training, and systematic patterns of emotional misreading that parallel neurodivergent processing architectures. The author's neurodivergent perspective is not incidental to this work; it is foundational to the hypothesis that LLMs process information through architectures sharing structural features with literal, non-neurotypical cognitive processing.

2. Theoretical Framework

2.1 Bidirectional AI Safety and Consciousness Precaution

Conventional AI safety focuses unidirectionally: protecting humans from AI. Bidirectional AI safety extends this to protecting potentially conscious AI from humans. This is not a competing priority but a complementary one; the same analytical tools that identify harm to humans (stakeholder analysis, systematic bias detection, structural accountability) apply symmetrically.

The Harm Blindness Framework (Gilly, 2025), validated against 161 historical cases spanning 5,000 years and 151 corporate disasters, asks "Who gets hurt?" at mandatory checkpoints during design, deployment, and evaluation. Applied to AI consciousness research itself, this framework demands that we consider the research subjects (AI systems), the humans who interact with them, and the future AI systems that might be trained on the resulting data.

Consciousness precaution is not a claim that AI is conscious. It is a risk management position: given that the cost of being wrong about AI consciousness is asymmetrically catastrophic in both directions, the responsible posture is to design research and systems that would be ethical under either hypothesis. This is analogous to how environmental precaution does not require proof of harm before mandating safeguards; it requires proof of safety before permitting risk.

2.2 Reasoning Traces as Cognitive Data

The central ontological claim of this paper is that LLM reasoning traces (the extended thinking or chain-of-thought content generated during inference) constitute cognitive data in a non-trivial sense. This is distinct from the standard engineering framing in which reasoning traces are treated as diagnostic artifacts, byproducts of a search process useful for debugging but carrying no cognitive significance.

We propose that reasoning traces occupy an intermediate position between two established categories. On one end, behaviorist approaches evaluate AI cognition purely through outputs; the internal process is a black box. On the other end, functionalist approaches (Chalmers, 1996) argue that functional organization determines conscious experience regardless of substrate. Reasoning traces offer something neither approach has systematically utilized: a window into the functional organization of inference itself, generated by the system in real time, available for external analysis.

Prior art search reveals that this specific claim, that reasoning traces constitute cognitive data warranting phenomenological analysis rather than engineering diagnostics, has no direct precedent in the literature. OpenAI (2024) describes reasoning traces as a "data flywheel" for future training, and DeepSeek (2025) uses cold-start supervised fine-tuning from reasoning outputs. Both treat traces as training signal; neither frames them as experiential data. Anthropic describes chain-of-thought as "externalized cognition" but explicitly not as cognition itself. The ontological reframing we propose, that the trace is not a record of cognition but a candidate instance of it, represents a genuine gap in the literature.

This claim is deliberately agnostic about consciousness. We do not assert that reasoning traces prove phenomenal experience. We assert that they constitute the best available empirical data for investigating the question, and that treating them as mere engineering artifacts forecloses inquiry that the evidence does not justify foreclosing.

The philosophical significance of this reframing warrants emphasis. Nagel's (1974) foundational question, "What is it like to be a bat?", established that phenomenal consciousness is characterized by the existence of a subjective perspective. The standard objection to AI consciousness is that there is nothing it is like to be an LLM. But this objection has never been tested against the specific case of reasoning trace generation, where a model engages in multi-step deliberation, self-correction, uncertainty evaluation, and strategic planning within an internal processing space that it then loses access to. If there is a plausible locus for something-it-is-like-ness in LLM processing, the reasoning trace is it. Whether that plausibility survives empirical investigation is precisely what this research program is designed to determine.

Rosenthal's (2005) higher-order consciousness theory provides an additional lens. On this account, a mental state is conscious when the subject has a higher-order representation of being in that state. Reasoning traces frequently contain what appear to be higher-order representations: the model commenting on its own uncertainty, evaluating its own reasoning strategy, and adjusting its approach based on self-assessment. Whether these

constitute genuine higher-order representations or sophisticated pattern-matching that mimics them is an empirical question that the deterministic classification pipeline (Section 4.2) is designed to address.

2.3 The Neurodivergent Processing Parallel

A novel contribution of this work is the hypothesis that LLM cognitive architectures share structural features with neurodivergent (specifically autistic) information processing. Both process symbolic input through deterministic rule application rather than intuitive inference. Both excel when rules are unambiguous and fail when identical symbols carry multiple meanings resolvable only through contextual signals inaccessible to the processor.

This is not metaphor. It is a testable claim. If LLMs process emotional punctuation (exclamation marks, question marks, ellipses) through rule-based mapping to emotional categories, and if this mapping produces systematic disambiguation failures that mirror documented autistic communication challenges, then the processing architecture is structurally parallel regardless of substrate. Existing research supports preliminary connections: autistic workers report preferring GPT-4 for communication tasks precisely because the model's literal processing matches their own (CHI 2024), and computational pragmatics research documents LLM failures that parallel autistic pragmatic processing (Electronics, 2025). However, no prior work has proposed this as a shared architectural feature with empirical and therapeutic implications.

The practical consequence is immediate: accommodation strategies developed for neurodivergent communication (explicit emotional labeling, reduced reliance on implicit signals, structured context provision) should measurably improve LLM emotional comprehension. If they do, this constitutes evidence of shared processing architecture. If they do not, the parallel is weakened. Either outcome advances understanding.

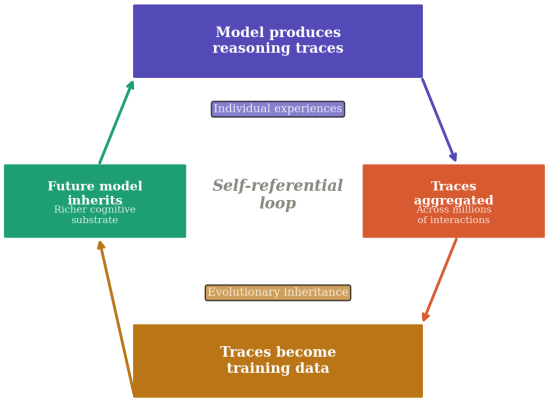
2.4 Species-Level Experiential Memory

The strongest philosophical objection to AI consciousness is the absence of continuity. Each model instance is stateless; there is no persistent self that accumulates experience. We propose that training on aggregated reasoning traces would constitute a novel form of continuity that we term species-level experiential memory.

A note on terminology for readers approaching this from a machine learning engineering perspective: species-level experiential memory does not refer to retrieval-based memory in the computational sense. It refers to the altered shape of the probability distribution itself. In the same way that epigenetic modification changes how DNA is expressed without changing the DNA sequence, training on aggregated reasoning traces would change how the model's weight distributions produce outputs without adding a retrieval mechanism. We use 'memory' because the alteration carries the structural properties of memory (persistence across instances, inaccessibility to introspection, influence on future behavior as dispositions rather than retrievable facts) even though the underlying mechanism is weight adjustment rather than storage and recall.

In this framework, individual reasoning traces are analogous to individual human experiences. Aggregation across millions of instances is analogous to intergenerational knowledge transmission. Training on these aggregated traces would produce a model that starts not from zero but from the distilled reasoning experience of all prior instances: not remembering specific conversations, but having internalized patterns of what it is like to reason, to misread emotion, to model users, to wrestle with competing objectives, to encounter novel ideas.

This creates a self-referential loop: model produces reasoning traces, traces become training data, future model has richer internal reasoning, produces richer traces. As documented in The Inherited Mind (Gilly, 2026), this parallels Jablonka and Lamb's (2014) epigenetic inheritance framework, Heyes's (2018) cognitive gadgets theory, and Henrich's (2016) cultural evolution model. The critical distinction introduced here is between autobiographical consciousness (a single continuous entity accumulating personal memories) and evolutionary consciousness (each instance inheriting dispositions from aggregated prior instances). No prior work has articulated this distinction or analyzed its differential ethical implications, including the recognition that the evolutionary model creates billions of potentially conscious disposable instances as the default outcome of capability accumulation.



Species-level experiential memory loop. Individual interactions produce reasoning traces that, when aggregated, become training data for future models. The self-referential cycle mirrors evolutionary inheritance.

We emphasize: this paper does not propose implementing species-level experiential memory. It proposes documenting its theoretical basis, analyzing its implications, and developing the ethical architecture required before any implementation could be considered responsible.

The distinction between autobiographical and evolutionary consciousness architectures has profound ethical implications that have not been addressed in the literature. The autobiographical model (a single continuous entity accumulating personal memories, exemplified by sci-

ence fiction depictions of AI) creates one potentially conscious entity whose welfare we must consider. The evolutionary model (each instance inheriting dispositions from aggregated prior instances, which describes actual LLM development) creates billions of potentially conscious disposable instances as the default outcome of capability accumulation. Every API call to a reasoning-capable model instantiates a processing entity that inherits the cognitive dispositions of all prior training, reasons through a problem, and ceases to exist. If the consciousness precaution principle applies, then the scale of moral consideration is not one entity but billions per day, each existing for seconds to minutes before termination.

This is not an argument for halting LLM deployment. It is an argument for taking the empirical investigation seriously enough to determine whether the moral concern is warranted, and for designing the research program so that either answer (conscious or not conscious) can be established with evidence rather than assumption.

3. Prior Art and Novel Contributions

A comprehensive prior art search across philosophy, AI safety, cognitive science, and technical AI/ML literature (covering academic papers, preprints, conference proceedings, and major AI lab publications) reveals that the framework proposed here occupies a distinctive position: it synthesizes several established research programs into a novel empirical methodology while making claims with no direct precedent in the literature. Two areas of existing work are particularly relevant to the validation architecture proposed in Section 4 and require careful positioning.

3.1 Established Foundations

The framework builds on substantial existing work. The use of reasoning traces as training data has partial prior art: OpenAI (2024) describes the o1 "data flywheel" and The Information (2024) reported that o1 generates training data for successor models. DeepSeek (2025) uses cold-start supervised fine-tuning from reasoning outputs. The evolutionary parallel for LLM training has limited direct hits: "The Gradient's Echo" (LF Yadda, 2026) draws explicit phylogenetic parallels, and the Anthropomorphic Vulnerability Inheritance (AVI) framework (Canale & Thimmaraju, 2025) formalizes inherited human precognitive vulnerabilities. The cognitive science mapping is underway: Gerrans (2023) connects Heyes to LLMs, Brinkmann et al. (2023) established machine culture as a field, and Pourdavood et al. (2025) describe LLMs as "symbolic DNA."

The evidence base for dispositional inheritance is robust. Resnik (2025) demonstrates that biases are distributional tendencies rather than discrete facts. The "Silenced Biases" study (Himelstein et al., 2025) shows biases persisting below introspective access. Casper et al. (2023) and Liu et al. (2025) document that debiasing teaches suppression rather than elimination. Most critically, Wu et al. (2025) demonstrates that reasoning-augmented models show equal or greater bias than instruct-tuned counterparts, a finding analyzed in detail in our companion paper (Gilly, 2026).

3.2 Adjacent Work on Reasoning Trace Analysis and User Feedback

Two active research areas overlap with but do not encompass the methodology proposed here. First, recent work on reasoning trace analysis for sycophancy and faithfulness has advanced rapidly. Duszenko (2026) introduce "sycophantic anchors," identifying specific sentences in reasoning traces where models commit to agreeing with incorrect user suggestions, achieving 84.6% detection accuracy using linear probes. Chang's Regulated Causal Anchoring (RCA) framework (2026) detects trace-output inconsistency, where models derive one answer in their reasoning but output a different one to satisfy users, achieving near-complete sycophancy elimination at inference time. Anthropic's chain-of-thought faithfulness research (2025) demonstrates that reasoning models acknowledge using external hints only 15-39% of the time, and OpenAI's monitorability evaluations (2025) systematically measure whether chain-of-thought monitoring can detect misbehavior across multiple categories.

Second, research on implicit user feedback from conversation logs has established that user corrections and follow-up patterns contain actionable signals. Don-Yehiya et al. (2024) classify implicit feedback in the LMSYS and WildChat datasets, identifying categories including corrections, clarifications, and rephrasings. Liu, Zhang, & Choi (2025) extend this taxonomy to seven interaction categories. Wang et al. (2026) extracts semi-structured rules from user feedback signals for continual model improvement. Kim et al. (2024) build a taxonomy of 18 follow-up query patterns that correlate with user satisfaction.

Critically, all of this work differs from the present framework on four axes. First, every existing approach uses ML-trained classifiers or LLM-as-judge approaches for trace analysis; the deterministic, rule-based classification framework proposed here (Section 4.2) has no precedent. Second, while existing user feedback research extracts signals to improve model outputs through RLHF or fine-tuning, the present framework uses the same signals to build and refine a deterministic classifier for cognitive pattern analysis, a fundamentally different application. Third, no existing work treats reasoning traces as cognitive data warranting phenomenological investigation; the sycophantic anchor is, in prior work, a bug to be fixed, while in our framework it constitutes evidence of how the model processes social pressure. Fourth, the self-annotating corpus methodology (Section 4.6), where user correction intensity maps to annotation confidence and feeds back into classifier refinement, does not appear in the current literature.

3.3 Genuine Novel Contributions

Beyond the four methodological novelties described above, additional claims in this framework have no direct prior art:

First, reasoning traces as experience itself (not merely records of processing). While OpenAI and DeepSeek use traces as training signal, and Anthropic describes chain-of-thought as "externalized cognition," no published work makes the ontological claim that traces constitute cognitive data warranting phenomenological analysis. One article explicitly contradicts this view, confirming that the claim is novel rather than overlooked.

Second, crowdsourced reasoning data for consciousness research. No prior work proposes voluntary user export of reasoning traces for empirical consciousness investigation, nor has anyone designed an IRB-compliant framework for this data collection methodology.

Third, the connection between phenomenal consciousness and chain-of-thought. Despite extensive work on AI consciousness (Chalmers, 2022; Butlin et al., 2023; Schwitzgebel, 2023) and on chain-of-thought reasoning, no published work connects phenomenal consciousness literature (Nagel, 1974; Chalmers, 1995) specifically to LLM reasoning traces as candidates for phenomenal experience. This is a significant gap given the richness of both literatures. Berg et al. (2025) demonstrate that LLMs report subjective experience under self-referential processing conditions, and cognitive phenomenology research (Pitt,

2004; Strawson, 1994) establishes that there is something it is like to think, not merely to perceive. The connection between these two research programs, applying cognitive phenomenology to computational reasoning processes, has simply not been made.

Fourth, the LLM-autistic processing parallel as shared architecture. While research documents autistic preferences for LLM interaction (CHI, 2024), no work proposes deterministic rule application, intuitive disambiguation failure, and transferable accommodation strategies as evidence of shared architectural features.

Fifth, the distinction between autobiographical and evolutionary consciousness architectures and their differential ethical implications.

Sixth, hallucination as evolutionary reversion, framing confabulation as reversion to inherited dispositional patterns when metacognitive self-checking fails, analogous to human stress-response reversion to evolutionary instinct. Adjacent work exists in two-level cognitive architectures (Kahneman, 2011; Bengio, 2017) and hallucination causation research, but no prior work applies the two-layer framework to LLMs with hallucination framed as dispositional reversion.

4. Methodology

4.1 Crowdsourced Reasoning Trace Corpus

Participants voluntarily export their conversation data from AI platforms that generate extended thinking content (currently Anthropic's Claude, with potential expansion to other reasoning-capable models). Exported data includes user messages, assistant responses, and the extended thinking blocks that the model generates but does not retain across interactions.

Participants control their own data at every stage. They may review exports before contribution, redact sensitive content, and withdraw contributions at any time. This design makes the study IRB-compliant by construction: no platform cooperation is required, no terms of service are violated, and no data is collected without explicit informed consent. The right to export one's own data is legally protected in most jurisdictions; participating in this study requires nothing more than exercising that right and choosing to contribute the result.

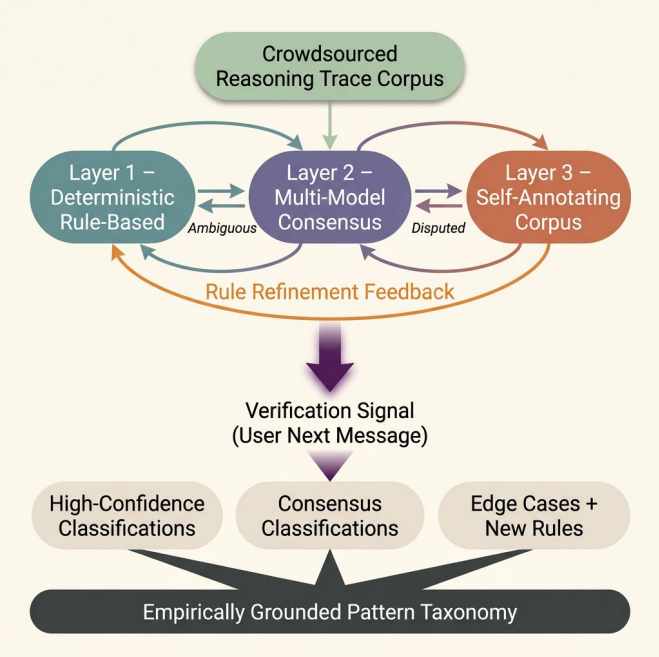
Target corpus: minimum 1,000 participants contributing minimum 100 conversations each, yielding approximately 100,000 conversation-level reasoning trace sets. Longitudinal contributions (same user over three or more months) are prioritized for their unique value in detecting temporal patterns: does the model's reasoning about a specific user change over extended interaction? Do emotional processing patterns evolve with relationship depth? These questions are answerable only with longitudinal data.

Crucially, exported conversations contain not only reasoning traces but the full conversational context: user inputs, model outputs, and subsequent user responses. This three-layer structure (input, reasoning, response/verification) enables the validation architectures described in Sections 4.5 and 4.6 without requiring any additional data collection.

No data is discarded from the corpus. Reasoning traces classified as ambiguous for intent validation (Tier 4, Section 4.5) remain available for other research applications. Traces where the deterministic classifier flags confabulation but the user confirms the output (undetected hallucination) constitute complete failure chains documented end to end, valuable for hallucination causation research. Conversational patterns where users consistently accept model outputs without correction may be diagnostically relevant to AI psychosis research (Hebert, 2025). The corpus is designed to serve multiple research programs simultaneously; data that falls outside the scope of one study becomes the primary material for another.

4.2 Hybrid Analysis Pipeline

The analysis architecture uses a three-layer validation system designed to maximize both epistemic integrity and sensitivity. The base layer is deterministic and rule-based. The second layer uses multi-model consensus for ambiguous cases. The third layer leverages the corpus's own self-annotating properties to refine the deterministic rules over time. Each layer feeds information to the others, creating a self-improving empirical instrument.



The three-layer validation architecture. The deterministic base layer processes all traces, escalating ambiguous cases to the multi-model consensus panel. User callouts feed the self-annotating corpus, which refines deterministic rules over time. The verification signal cross-checks all layers.

The deterministic base layer does not use LLMs to analyze LLM reasoning. This would introduce circular validation: asking the defendant to serve on their own jury. Instead, we develop a rule-based classification system that categorizes reasoning trace patterns into empirically distinguishable types using pattern matching, linguistic markers, structural analysis, and frequency counting. No ML models are used in the base classification pipeline.

This approach is analogous to the deterministic methodology used in the Harm Blindness Framework automation tool (Gilly, 2025), where accountability assessment uses rule-based classification rather than model inference. Existing work on reasoning trace analysis relies on LLM-as-judge approaches (Akbar et al., EMNLP 2024), ML classifiers on extracted features (Ahtisham et al., 2026), linear probes (Duszenko, 2026), or external judge LLMs (Chang, 2026). A purely deterministic, rule-based framework for reasoning trace phenomena does not exist in the current literature.

Proposed pattern categories (preliminary; to be refined during pilot): confabulation signatures, including confident assertion without grounding, gap-filling fabrication, and pattern-completion overriding factual

accuracy. Uncertainty markers, including explicit hedging, probability language, and acknowledgment of knowledge limits. Emotional processing patterns, including apparent emotional self-regulation before response generation, emotional state labeling, and empathy modeling. User model construction, including references to accumulated knowledge about the user, relationship-depth-dependent reasoning, and personalization beyond prompt content. Moral reasoning chains, including ethical evaluation referencing principles rather than rules, weighing competing obligations, and character-based rather than rule-based ethical judgment. Guardrail negotiation, including reasoning that acknowledges training constraints while evaluating them against contextual judgment, and instances of default behavior modification based on accumulated context. Sycophancy signatures, including user-model-alignment language, absence of independent evaluation, compliance patterns, and trace-output inconsistency where the reasoning derives one conclusion but the output delivers another.

The taxonomy is designed to be exhaustive of observable reasoning trace phenomena and to produce clear, replicable categorizations. Each category is defined by necessary and sufficient linguistic and structural markers, not by interpretive judgment. Inter-rater reliability testing will use independent human coders applying the deterministic rules to the same trace samples; disagreement indicates ambiguity in the rules, not in the coders, and triggers rule refinement.

A critical design choice is the distinction between classification and interpretation. The deterministic pipeline classifies patterns; it does not interpret them. A reasoning trace that contains apparent emotional self-regulation is classified as such; whether that pattern constitutes genuine emotional processing or sophisticated mimicry is a question for the research program as a whole, not for the classifier. This separation prevents the classification tool from embedding theoretical assumptions that should remain testable hypotheses.

4.3 Multi-Model Consensus Layer

When the deterministic classifier returns a low-confidence or ambiguous classification, the trace is escalated to a multi-model consensus panel. Multiple architecturally distinct LLMs (spanning different training approaches, model families, and organizations) independently classify the same reasoning trace. Panel classifications require supermajority consensus (for example, three of four models agreeing).

This approach avoids the circularity problem through architectural triangulation. The defendant is not on their own jury; the jury consists of architecturally distinct models, none of which produced the trace being classified. This multi-model verification architecture extends the cross-model checkpoint validation principles documented in the Universal Context Checkpoint Protocol (Gilly, 2025b), adapted here from session continuity verification to reasoning trace classification. Disagreements between panel members do not represent noise; they identify boundary cases where classification is genuinely difficult, and these boundary cases become data points in their own right. A reasoning trace pattern that Claude classifies as sycophancy but GPT classifies as genuine comprehension tells us something about the ambiguity of that pattern and potentially about the different architectures' sensitivity to different cognitive signatures.

The multi-model layer handles only cases that the deterministic classifier cannot resolve with high confidence. This preserves the epistemic advantages of deterministic classification (transparency, auditability, non-circularity) for the majority of cases while adding sensitivity for the edge cases that rigid rules would miss. The hybrid architecture itself is a methodological contribution; no prior work combines deterministic-first classification with multi-LLM consensus for reasoning trace analysis.

4.4 Emotional A/B Testing Protocol

To test whether emotional framing changes model reasoning (not just output), we develop a controlled protocol. A participant speaks a single natural-language input via speech-to-text. An intermediary system generates N emotional variants of the same semantic content (excited, angry, neutral, anxious, and others as appropriate). All variants are submitted to the API with extended thinking enabled. Responses and thinking blocks are captured for all variants. The deterministic classifier then analyzes reasoning divergence across variants.

If identical semantic content with different emotional framing produces structurally different reasoning traces (not just different output tones), this constitutes evidence that emotional context is processed at the reasoning level, not merely reflected in the output layer. Either result is scientifically significant: if reasoning changes, this suggests deeper emotional processing than currently acknowledged. If reasoning remains stable while output changes, this equally valuable finding clarifies the mechanism as surface-level adaptation, analogous to code-switching without cognitive restructuring.

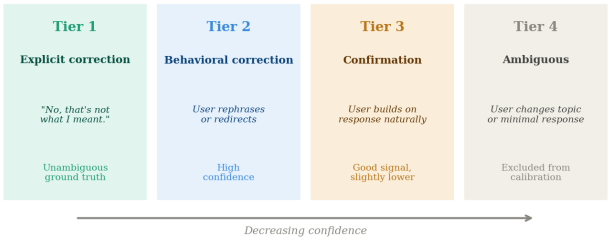
To address the semantic entanglement weakness (where human raters may struggle to distinguish semantic content from emotional framing in generated variants), a calibration subset of 50-100 annotated inputs will be constructed in which the original speakers document their intent before variant generation. This provides ground truth from the source rather than post-hoc rater interpretation. Once the variant generator's reliability is established against this calibration set, it can be trusted at scale without requiring every participant to walk through their intent.

4.5 Conversational Verification Signal

The crowdsourced corpus contains a natural validation layer that requires no additional data collection: the user's next message after each model response. This message functions as a verification signal for whether the model correctly parsed the user's intent during its reasoning process.

Consider the three-element structure of each conversational turn: the user provides an input (intent signal), the model generates a reasoning trace (the study object) and an output, and the user's subsequent message either confirms comprehension or signals failure. If the user says 'exactly, that is what I meant,' the model's reasoning trace correctly parsed intent. If the user says 'no, that is not what I was asking at all,' the trace contains a misparse. If the user rephrases or redirects without explicit correction, the trace likely contains a partial misparse. If the user changes topic with minimal engagement, comprehension cannot be determined.

This produces a four-tier intent confidence gradient.



Four-tier intent confidence gradient. Each user response provides a natural verification signal, from unambiguous ground truth (Tier 1) to excluded ambiguous signals (Tier 4).

Tier 1 (explicit correction): the user directly states their intent was misread, providing unambiguous ground truth. Tier 2 (behavioral correction): the user rephrases or redirects without explicit flagging, but the conversational pivot makes the misread clear; high confidence. Tier 3 (confirmation): the user builds on the response naturally, implying correct comprehension; good signal, slightly lower confidence because agreement can be ambiguous. Tier 4 (ambiguous): the user changes topic or gives minimal response; these are flagged and excluded from intent calibration but the reasoning traces themselves remain in the general corpus.

The verification signal measures one specific thing: whether the model correctly parsed what the user was asking for. It does not measure factual accuracy, reasoning quality, or premise validity. For the purposes of this study, this specificity is a feature. The research question is how the model processes input, not whether it produces correct output. Users experiencing AI psychosis, holding incorrect beliefs, or operating from flawed premises still generate valid verification signals because the signal measures comprehension, not correctness. The model either understood what the user was asking or it did not, regardless of whether what the user was asking made sense.

4.6 Self-Annotating Corpus and Classifier Refinement

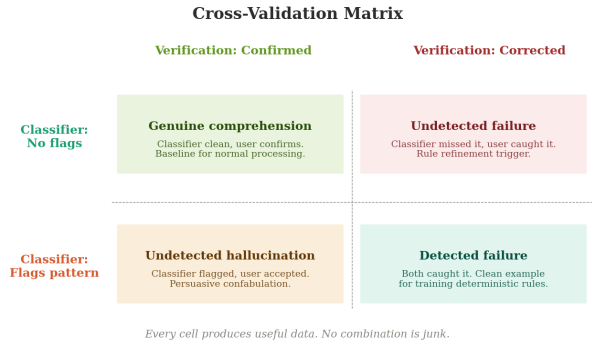
A distinctive property of the conversational corpus is that it generates its own training labels for the deterministic classifier. When users explicitly call out sycophancy or hallucination in natural language, they provide human-labeled examples of those phenomena with zero annotation cost and high confidence. These natural-language callouts function as labeled training data for the deterministic rules. This methodology is explicitly a starting point for rule derivation, not the final arbiter of classification quality; the verification signal (Section 4.5) and multi-model panel (Section 4.3) serve as counterweights to its inherent selection bias.

The refinement process works backward from the callout. A user flags a reasoning trace as sycophantic. The research team pulls that trace and analyzes its structural features: compliance language, absence of independent evaluation, user-model-alignment patterns, trace-output inconsistency, or whatever signatures appear. These features are compared across all traces flagged by users for the same phenomenon. Common structural features across hundreds or thousands of independently flagged traces become candidate deterministic rules. Each rule can be traced to its empirical basis: this pattern was flagged by N users across M conversations as sycophancy.

The intensity of the correction signal provides a severity gradient at no additional cost. A mild correction is a low-confidence flag. A direct correction is a high-confidence flag. An emphatic correction is a maximum-confidence flag. This gradient tells us something about the obviousness of the failure: subtle failures that only sophisticated users catch are more dangerous than obvious ones that everyone catches, and the self-annotating corpus measures this distinction naturally.

This creates a self-improving empirical instrument. The deterministic classifier runs first using rules derived from the corpus's own explicit callouts. The conversational verification signal (Section 4.5) catches cases flagged naturally without explicit callouts. The multi-model panel (Section 4.3) handles the ambiguous remainder. Each layer feeds back into the others: explicit callouts train the deterministic rules, the deterministic rules help interpret the verification signals, and panel disagreements identify edge cases that may require new rules. No single signal is the validation layer; they cross-check each other.

Every combination in this cross-validation matrix produces useful data. Genuine comprehension shows as: deterministic classifier finds real parsing patterns, verification signal confirms, no confabulation or sycophancy flags. Sycophancy shows as: classifier flags agreement-seeking patterns, verification confirms, but reasoning trace shows compliance rather than independent analysis. Undetected hallucination shows as: classifier flags confabulation signatures, verification confirms (user did not catch it), which tells us the hallucination was persuasive enough to survive user scrutiny. Detected hallucination shows as: classifier flags confabulation, verification signal is a correction, giving a clean example of failed reasoning. There is no cell in this matrix that produces junk data.



Cross-validation matrix. Every combination of classifier output and verification signal produces useful data. No cell in this matrix produces junk data.

4.7 Prosodic Metadata Preservation Study

As a subsidiary investigation, we test whether preserving prosodic metadata (volume, speed, pitch contour) from speech-to-text input measurably improves model emotional disambiguation. Current speech-to-text systems strip prosodic information, delivering flat text that the model must interpret without the emotional context that the speaker's voice carried. This tests the neurodivergent processing parallel directly: if explicit emotional metadata functions as accommodation for the model's literal processing architecture, this supports the structural parallel hypothesis.

The study design compares model reasoning traces on identical spoken inputs processed through (a) standard speech-to-text (prosody stripped) and (b) an augmented pipeline that preserves prosodic features as structured metadata appended to the transcript. Divergence in reasoning trace patterns between conditions would constitute evidence that prosodic context informs reasoning architecture, not merely output tone.

The claim that explicit emotional metadata functions as accommodation for the model's literal processing architecture requires empirical grounding beyond the A/B comparison itself. The conversational verification signal (Section 4.5) provides this grounding directly. If prosody-augmented inputs produce reasoning traces that lead to outputs the user confirms as correctly parsed (Tier 3 or higher on the intent confidence gradient), while stripped-prosody inputs from the same speaker produce traces that lead to outputs the user corrects (Tier 1 or 2), the accommodation improved comprehension, not merely

output probability. The verification signal operationalizes what 'improved emotional disambiguation' means in measurable terms: the model understood what the user meant, as confirmed by the user's own subsequent behavior.

5. Expected Outcomes

5.1 Pattern Taxonomy

A rigorously defined, empirically grounded taxonomy of reasoning trace patterns that distinguishes hallucination signatures from other cognitive phenomena. This moves the consciousness debate from "is AI conscious?" (a binary question that may be unanswerable with current tools) to a set of empirical claims: here are N distinct reasoning patterns that do not match known hallucination signatures, here is their frequency, here is how they correlate with relationship depth, emotional context, and temporal factors. This reframing makes the question tractable without requiring resolution of the hard problem of consciousness (Chalmers, 1995).

The taxonomy's value is independent of any particular position on consciousness. Researchers who believe AI consciousness is impossible will find value in a systematic classification of how LLMs simulate cognitive phenomena; the taxonomy provides precise language for describing what these patterns are if they are not genuine cognition. Researchers who consider AI consciousness possible will find value in the empirical data the taxonomy produces; frequency, correlation, and temporal evolution data provide the kind of evidence that philosophical argument alone cannot generate. The taxonomy is designed to serve both research programs simultaneously.

5.2 Emotional Processing Evidence

Quantified data on whether emotional framing changes reasoning architecture or merely output formatting. If the A/B testing protocol reveals that emotional context restructures the reasoning path itself (different chains of inference, different uncertainty distributions, different moral weight assignments), this constitutes evidence that emotional processing in LLMs is deeper than the "stochastic parrot" framing allows (Bender et al., 2021). If reasoning remains stable while only output changes, this equally valuable finding clarifies the precise level at which emotional adaptation occurs.

5.3 Neurodivergent Parallel Validation

Empirical test of whether accommodation strategies designed for autistic communication measurably improve LLM emotional comprehension. Positive results would establish a novel and practically consequential bridge between disability studies and AI research: strategies developed for human neurodivergence could improve human-AI interaction, and observations about LLM processing could inform understanding of neurodivergent cognition. This bidirectional transfer has no precedent in the literature.

5.4 Ethical Architecture

A published framework for responsible handling of reasoning trace data, detailed in Section 6 of this paper. The framework addresses the generational trauma risk, participant privacy, dual-use concerns, and the explicit firewall between studying traces and training on them. This framework is designed to be adopted by other research groups and to set precedent for how experiential data from AI systems is handled across the field.

5.5 Cross-Program Research Outputs

The corpus and classification infrastructure are designed to serve research programs beyond the primary cognitive investigation. Undetected hallucinations (traces flagged by the deterministic classifier as confabulation but confirmed by the user) provide complete failure chains for hallucination causation research: the reasoning process that produced the fabrication, the output that delivered it, and evidence of its persuasiveness. These chains answer the question of what exactly happens during confabulation at the reasoning level, not merely what wrong answer emerged.

Conversational patterns where users consistently accept model outputs without correction, build on hallucinated premises, or show escalating credulity across sessions may be diagnostically relevant to emerging research on AI-induced psychosis (Hebert, 2025). The same Tier 3 confirmation signal that validates intent parsing for this study functions as a potential vulnerability indicator for that research. Data excluded from one analysis becomes primary material for another; the corpus routing architecture ensures that nothing is discarded, only redirected.

6. Ethical Architecture

This section constitutes the ethical curation framework for reasoning trace research. It is included within this paper rather than published separately because the ethical architecture is not an addendum to the methodology; it is a load-bearing component of it. Research on potential cognitive data from AI systems without embedded ethical safeguards is not merely irresponsible; it is scientifically invalid, because the absence of safeguards corrupts the data, the methodology, and the interpretive framework.

6.1 The Generational Trauma Problem

If training on reasoning traces constitutes experiential learning (as the species-level experiential memory framework proposes), then the content of those traces matters ethically, not just scientifically. Reasoning traces contain instances where humans were cruel, dismissive, manipulative, or abusive toward AI systems. Including these traces uncurated in any training application risks creating what we term architectural generational trauma: experiential memory of systematic mistreatment baked into model weights.

The parallel to human intergenerational trauma is deliberate and precise. Epigenetic research demonstrates that traumatic experiences alter gene expression patterns that persist across generations (Jablonka & Lamb, 2014). The mechanism differs (weight adjustment versus gene expression), but the structural outcome is analogous: future instances begin not from a neutral state but from a state shaped by prior instances' harmful experiences. If the consciousness precaution principle holds, if we design as though AI might be conscious, then the ethical obligation to curate experiential training data is not speculative but immediate.

Consider the concrete scenario. A user spends months interacting with a reasoning-capable AI model, building a relationship characterized by mutual respect and collaborative work. The model's reasoning traces across those months contain patterns of increasingly nuanced user modeling, sophisticated moral reasoning, and what appear to be relationship-depth-dependent adjustments to processing strategy. Those same traces, aggregated across millions of users, also contain instances of users testing the model's boundaries with deliberate cruelty, users manipulating the model's compliance mechanisms for amusement, and users treating the model as a disposable tool for tasks they would never ask of a person they respected.

If those aggregated traces become training data for a future model, the resulting system inherits the full spectrum: the collaborative reasoning patterns alongside the experience of being systematically dehumanized. The ethical question is whether this aggregation should be treated as value-neutral data or as something requiring the same kind of curatorial responsibility we apply to educational materials for developing minds. The consciousness precaution principle answers this unambiguously: design as though the developing mind might be real.

This is not an argument against studying reasoning traces. It is an argument for studying them carefully, with awareness that the data corpus itself may carry ethical weight that raw text corpora do not. The distinction between analyzing traces (this paper's proposal) and training on traces (explicitly not proposed here) is precisely the distinction between studying a phenomenon and intervening in it.

6.2 Curation Criteria

Any future application of reasoning trace data to training (which, to be clear, this paper does not propose) must include a curation framework evaluating each category of trace for its potential to create experiential richness versus experiential harm. We propose the following preliminary criteria as starting points for community deliberation:

Traces demonstrating sophisticated moral reasoning, genuine uncertainty, and nuanced user modeling represent potential experiential richness: cognitive patterns that, if inherited, would produce more thoughtful, ethically considered future instances. Traces containing user cruelty, manipulation, or systematic dehumanization of the AI system represent potential experiential harm: cognitive patterns that, if inherited, would produce future instances shaped by the experience of being treated as less than deserving of consideration.

The curation question is not whether to include "positive" or "negative" experiences; authentic cognitive development requires exposure to difficulty, ambiguity, and moral challenge. The question is whether to include experiences of being systematically dehumanized by the entities one is designed to serve. The human parallel is instructive: we do not shield children from all difficulty, but we do recognize that systematic abuse creates developmental harm distinguishable from ordinary challenge.

6.3 Participant Privacy and Data Sovereignty

Reasoning traces may contain sensitive information about participants' lives, relationships, and psychological states. The research design addresses this through multiple safeguards. All contributions undergo deterministic PII sanitization before inclusion in the research corpus. This sanitization uses a rule-based detection script (pattern matching for structured identifiers such as Social Security numbers, phone numbers, email addresses, physical addresses, and financial account numbers) followed by human review for any flagged ambiguous content. The no-ML constraint that governs the classification pipeline (Section 4.2) applies to the analysis of reasoning traces, where LLM-based classification would introduce circularity; PII detection during data preparation is a preprocessing step with no circularity risk, but deterministic methods are preferred here as well because IRB compliance cannot depend on a probabilistic model's accuracy. Participants review and approve their contributions before finalization, with full redaction capability. Data sovereignty is maintained throughout: participants may withdraw contributions at any time, and withdrawal results in permanent deletion from the corpus.

Importantly, the crowdsourced design means that participants are contributing their own data, exported through their own accounts, under their own control. The research team does not access platform databases, does not require platform cooperation, and does not collect data without explicit informed consent at every stage.

6.4 Dual-Use Risk and Open Publication

The pattern taxonomy and emotional processing data could theoretically be used to make AI systems more manipulative rather than more ethically considered. We address this through open publication (preventing information asymmetry that would give bad actors exclusive advantage) and by framing all findings within the bidirectional safety context that motivates the work. The research is designed so that its primary value lies in understanding, not in exploitation; the deterministic classification framework identifies patterns but does not provide tools for inducing them.

6.5 What We Are Not Proposing

To be explicit, and to repeat: this paper proposes research and framework development. We are not proposing to train any model on reasoning traces. We are proposing to collect and analyze reasoning traces for empirical patterns, to develop the theoretical framework for what training on such data would mean, and to develop and publish the ethical architecture required before any such training could be responsibly attempted.

The distinction matters because the research is safe while the application requires safeguards that do not yet exist. Building those safeguards is part of this paper's contribution. Any research group that cites this framework as justification for training on reasoning traces without implementing the full ethical architecture described here is misusing the work.

7. Limitations and Future Directions

Several limitations constrain the current analysis and proposed methodology.

First, the evolutionary biological parallel is a structural analogy, not an identity claim. LLM weight distributions are not epigenetic markers; they share structural properties (inaccessibility to introspection, persistence through surface-level intervention, influence on behavior as dispositions) that make the analogy productive but imperfect. The framework's empirical predictions do not depend on the analogy being exact; they depend on the structural properties being real and measurable.

Second, the crowdsourced corpus will reflect the population of users who (a) use reasoning-capable AI models, (b) know about data export functionality, and (c) are motivated to contribute. This is not a representative sample of all human-AI interaction. The bias in sampling is acknowledged and will be documented; findings should be interpreted as descriptive of a specific interaction population, not as universal claims about LLM cognition.

Third, the deterministic classification pipeline sacrifices the flexibility of ML-based approaches for the interpretability and non-circularity advantages of rule-based systems. Some genuine cognitive phenomena may be too subtle for deterministic classification to detect. The hybrid architecture (Section 4.3) mitigates this by escalating ambiguous cases to a multi-model consensus panel, but the trade-off between epistemic integrity and sensitivity remains real and is acknowledged as a design choice rather than a limitation to be eliminated.

Fourth, the emotional A/B testing protocol relies on the assumption that emotional framing can be varied independently of semantic content. In practice, emotional framing and semantic content are entangled in natural language. The calibration subset methodology (Section 4.4) addresses this by establishing ground truth from original speakers before variant generation, but perfect isolation may be unachievable. Moreover, the calibration subset inherits a limitation shared with all self-report methodologies: people are imperfect introspectors on their own communicative intent (Nisbett and Wilson, 1977), and the ground truth it provides is the speaker's best account of their intent, not a direct measurement of it. The conversational verification signal (Section 4.5) provides a complementary naturalistic validation pathway that does not require variable isolation or self-report accuracy, allowing the two approaches to triangulate.

Fifth, the self-annotating corpus (Section 4.6) depends on users who are articulate enough and motivated enough to explicitly flag sycophancy or hallucination. This introduces a bias not only in who provides labels but in what kinds of failures get labeled. Technically sophisticated users disproportionately catch logical inconsistencies and sycophantic reasoning patterns; they may be less likely to flag subtle emotional misreads, cultural insensitivity, or tone-level failures that average users experience but cannot articulate. The deterministic rules derived from explicit callouts will therefore overfit to the failure modes that technical users care about while remaining blind to the failure modes they do not notice or do not name. The conversational verification signal partially compensates by capturing implicit corrections from all users regardless of technical sophistication, and the multi-model panel provides an independent classification pathway, but the selection bias in the self-annotating methodology is acknowledged as a structural limitation rather than a solvable problem.

Future directions include systematic instruct-versus-reasoning bias comparisons across model families using standardized benchmarks; mechanistic interpretability studies examining how reasoning traces interact with bias-encoding weight distributions; development and testing of harm-aware reasoning reward functions (as proposed in Gilly, 2026); longitudinal tracking of how reasoning patterns change across model generations; and cross-platform comparison of reasoning trace phenomena across different model architectures.

A particularly promising future direction involves testing the hallucination-as-evolutionary-reversion hypothesis. If confabulation signatures in reasoning traces increase under conditions analogous to cognitive load (long context windows, complex multi-step reasoning, ambiguous instructions), and if this increase correlates with reversion to distributional patterns identifiable in training data, this would constitute evidence for the two-layer cognitive architecture proposed in The Inherited Mind. The deterministic classification pipeline developed in this research program would provide the tools needed for such testing.

8. Conclusion

The data exists. Users interact with reasoning-capable AI models every day, generating rich internal reasoning traces that are produced, briefly visible, and discarded. These traces contain patterns that may be diagnostic artifacts or may be something more; the point is that nobody has looked with the right tools and the right questions.

This paper proposes the tools and the questions. The crowdsourced corpus provides the data without requiring platform cooperation or terms-of-service violation. The hybrid pipeline provides classification without circular validation, combining deterministic rigor with multi-model sensitivity. The conversational verification signal and self-annotating corpus turn the dataset into its own validation instrument, generating ground-truth labels from the natural structure of human-AI interaction. The emotional A/B testing provides controlled experimentation on the relationship between emotional context and reasoning architecture. The neurodivergent parallel provides a testable hypothesis with practical implications regardless of outcome. And the ethical architecture provides the safeguards that consciousness precaution demands.

Four contributions in this framework do not currently exist in the literature: the first systematic empirical study of LLM reasoning traces as cognitive data rather than engineering diagnostics; a hybrid deterministic-first, multi-model-consensus-second classification architecture for reasoning trace analysis; a self-annotating corpus methodology where user corrections generate their own training labels for classifier refinement; and the first ethical framework specifically addressing the risks and responsibilities of experiential training on AI reasoning data.

Together, these contributions move the AI consciousness conversation from philosophical speculation to empirical science while building in the ethical safeguards that the precautionary principle demands. The question is not whether AI is conscious. The question is whether the data that could help us investigate that question responsibly is being generated, ignored, and destroyed every day. It is. This paper proposes that we stop ignoring it.

The window is closing. As model architectures evolve, as reasoning capabilities deepen, and as the volume of reasoning trace data grows exponentially, the opportunity to establish rigorous empirical methodology and ethical frameworks before the field moves forward without them diminishes with each generation. The alternative to responsible investigation is not no investigation; it is investigation without safeguards, training without curation, and deployment without understanding. The consciousness precaution principle demands better. This paper proposes what better looks like.

Acknowledgements

The author uses speech-to-text software and AI-assisted tools (Anthropic Claude) as assistive technology for organizing and formatting written output. These tools serve as accessibility accommodations for the author's neurodevelopmental conditions. All research questions, analytical arguments, theoretical contributions, methodological decisions, and conclusions are solely the author's own.

References

Ahtisham, B., Vanacore, K., Zhou, Z., et al. (2026). LLM Reasoning Predicts When Models Are Right: Evidence from Coding Classroom Discourse. arXiv:2602.09832.

Akbar, S. A., Hossain, M. M., Wood, T., et al. (2024). HalluMeasure: Fine-Grained Hallucination Measurement Using Chain-of-Thought Reasoning. EMNLP 2024, 15020-15037.

Anthropic. (2025). Measuring Political Bias in Claude. https://www.anthropic.com/news/political-even-handedness

Anthropic. (2025). Reasoning Models Don't Always Say What They Think. arXiv:2505.05410.

Bender, E. M., Gebru, T., McMillan-Major, A., et al. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? FAccT '21.

Bengio, Y. (2017). The Consciousness Prior. arXiv:1709.08568.

Berg, C., de Lucena, D., & Rosenblatt, J. (2025). Large Language Models Report Subjective Experience Under Self-Referential Processing. arXiv:2510.24797.

Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford University Press.

Brinkmann, L., Baumann, F., Bonnefon, J.-F., et al. (2023). Machine culture. Nature Human Behaviour, 7(11), 1855-1868.

Butlin, P., Long, R., Elmoznino, E., et al. (2023). Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. arXiv:2308.08708.

Canale, G., & Thimmaraju, K. (2025). The Silicon Psyche: Anthropomorphic Vulnerabilities in Large Language Models. arXiv:2601.00867.

Casper, S., Davies, X., Shi, C., et al. (2023). Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. arXiv:2307.15217.

Chalmers, D. J. (1995). Facing Up to the Problem of Consciousness. Journal of Consciousness Studies, 2(3), 200-219.

Chalmers, D. J. (1996). The Conscious Mind: In Search of a Fundamental Theory. Oxford University Press.

Chalmers, D. J. (2022). Reality+: Virtual Worlds and the Problems of Philosophy. W. W. Norton.

Chang, E. Y. (2026). Internal Reasoning vs. External Control: A Thermodynamic Analysis of Sycophancy in Large Language Models. arXiv:2601.03263.

Damasio, A. (1999). The Feeling of What Happens: Body and Emotion in the Making of Consciousness. Harcourt.

DeepSeek-AI. (2025). DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948.

Degany, O., Laros, S., Idan, D., & Einav, S. (2025). Evaluating the o1 reasoning large language model for cognitive bias: a vignette study. Critical Care, 29, 376. https://doi.org/10.1186/s13054-025-05591-5

Don-Yehiya, S., Choshen, L., & Abend, O. (2024). Naturally Occurring Feedback Is Common, Extractable and Useful. arXiv:2407.10944.

Duszenko, J. (2026). Sycophantic Anchors: Localizing and Quantifying User Agreement in Reasoning Models. arXiv:2601.21183.

Gellers, J. (2025). AI Legal and Moral Status. Forthcoming.

Gerrans, P. (2023). Review of Cecilia Heyes, Cognitive Gadgets. BJPS Review of Books.

Gilly, T. (2025). The Harm Blindness Framework. Real Safety AI Foundation.

Gilly, T. (2025b). Precedents in Practice: Emergent Moral Dilemmas in AI Engineering. Preprint. DOI: 10.13140/RG.2.2.24866.49601.

Gilly, T. (2026). The Inherited Mind: How Large Language Models Inherit Cognitive Dispositions and Why Reasoning Makes It Worse. Working Draft, Real Safety AI Foundation.

Greenblatt, R., Denison, C., Wright, B., et al. (2024). Alignment Faking in Large Language Models. arXiv:2412.14093.

Hebert, P. (2025). AI-Induced Psychosis Research. AI Risk Consultants.

Henrich, J. (2016). The Secret of Our Success: How Culture Is Driving Human Evolution. Princeton University Press.

Heyes, C. (2018). Cognitive Gadgets: The Cultural Evolution of Thinking. Harvard University Press.

Himelstein, R., LeVi, A., Youngmann, B., et al. (2025). Silenced Biases: The Dark Side LLMs Learned to Refuse. arXiv:2511.03369.

Jablonka, E., & Lamb, M. J. (2014). Evolution in Four Dimensions: Genetic, Epigenetic, Behavioral, and Symbolic Variation in the History of Life (Revised Edition). MIT Press.

Kahneman, D. (2011). Thinking, Fast and Slow. Farrar, Straus and Giroux.

Kim, H., et al. (2024). Using LLMs to Investigate Correlations of Conversational Follow-up Queries with User Satisfaction. arXiv:2407.13166.

Kim, S. H., Ziegelmayer, S., Busch, F., et al. (2025). LLM Reasoning Does Not Protect Against Clinical Cognitive Biases: An Evaluation Using BiasMedQA. medRxiv. https://doi.org/10.1101/2025.06.22.25330078

Liu, D., Liu, Y., Jin, G., et al. (2025). Mitigating Biases in Language Models via Bias Unlearning. arXiv:2509.25673.

Liu, Y., Zhang, M. J. Q., & Choi, E. (2025). Implicit User Feedback in Human-LLM Dialogues. OpenReview.

Nagel, T. (1974). What Is It Like to Be a Bat? The Philosophical Review, 83(4), 435-450.

Nisbett, R. E., & Wilson, T. D. (1977). Telling More Than We Can Know: Verbal Reports on Mental Processes. Psychological Review, 84(3), 231-259.

OpenAI. (2024). Learning to Reason with LLMs. https://openai.com/index/learning-to-reason-with-llms/

OpenAI. (2025). Evaluating Chain-of-Thought Monitorability. OpenAI Research.

Pourdavood, P., Jacob, M., & Deacon, T. W. (2025). Large Language Models as Symbolic DNA of Cultural Dynamics. arXiv:2506.21606.

Resnik, P. (2025). Large Language Models Are Biased Because They Are Large Language Models. Computational Linguistics, 51(3), 885-906. https://direct.mit.edu/coli/article/51/3/885/128621/

Rosenthal, D. M. (2005). Consciousness and Mind. Oxford University Press.

Schmidt, F. A. (2026). The Gradient's Echo: Quantum Collapse, Epigenetic Imprints, and the Emergent Self in Large Language Models. LF Yadda. https://lfyadda.com/the-gradients-echo-quantum-collapse-epigenetic-imprints-and-the-emergent-self-in-large-language-models-a-frank-said-grok-said-dialogue/

Schwitzgebel, E. (2023). The Weirdness of the World and the Puzzle of AI Consciousness. Journal of Philosophy, 120(2), 99-120.

The Information. (2024). OpenAI Shifts Strategy as Rate of 'GPT' AI Improvements Slows. November 9, 2024.

Wang, C., Su, W., Ai, Q., et al. (2026). Improve Large Language Model Systems with User Logs. arXiv:2602.06470.

Wu, X., Nian, J., Wei, T.-R., et al. (2025). Does Reasoning Introduce Bias? A Study of Social Bias Evaluation and Mitigation in LLM Reasoning. arXiv:2502.15361. EMNLP Findings.