MOLES

A 'Map Of LLM-based Epistemological Stances'

Rob Manson (https://robman.fyi)

June 16th, 2025

Abstract

Large language models (LLMs) don't just produce answers - they perform *epistemological stances* towards knowledge, reasoning, self and others. We introduce **MOLES** (Map Of LLM-based Epistemological Stances), a structured framework for interpreting LLM outputs in terms of their epistemic posture. MOLES differentiates between factual retrieval, hallucination, simulation of other minds, self-reflection, self-delusion and more - offering a practical ontology for analysing the forms of simulated knowledge in LLM outputs. MOLES shifts the focus from whether a claim is true to the epistemic stance that can be deduced from the output. We define each stance, examine their epistemic & functional properties and map transitions and blends between them. MOLES supports interpretability, dialogue safety, agent modelling and AI epistemology - without relying on model internals or specific metrics.

1. Introduction: Beyond Hallucinations

Large language models (LLMs) are now widely used as interactive agents across domains including education, research, software development, therapy, and creative writing. Yet evaluations of their behaviour remain narrowly focused on a single axis: *factuality*. Outputs are judged as either correct or hallucinated - with hallucination typically defined as a confidently stated but untrue response. While this framing is appropriate in high-stakes contexts (e.g. legal advice, scientific explanation, or retrieval-augmented QA), it obscures deeper challenges of interpretation, trust, and epistemic coherence.

LLMs frequently produce outputs that appear to reflect perspectives, motivations, reasoning strategies, or emotional states. These outputs might be grounded in training data - or not - but more importantly, they enact distinct **epistemological stances**: implied orientations toward knowledge, uncertainty, or selfhood. For example, a model might say "I believe this answer is correct" or "I might be wrong, but…". These expressions simulate epistemic postures without reflecting genuine belief. They are not mistakes - they are communicative stances with functional roles.

This paper draws a sharp distinction between the **epistemological grounding** of a model's output (whether it is justified or grounded) and its **utility** (whether it is useful or appropriate). An output may be entirely ungrounded yet serve a clear communicative or pragmatic function. Recognising this separation is essential when evaluating LLM outputs, especially those involving self-reference, speculation, or simulated reasoning.

LLM behaviour can't be understood through factuality alone. We introduce **MOLES** (a *Map Of LLM-based Epistemological Stances*), a taxonomy of the ways language models simulate knowing, doubting, reflecting, imagining, or asserting. No existing framework systematically categorises these behaviours. MOLES helps us move from asking *"Is this true?"* to *"What kind of knowledge claim is this?"* and *"How is it positioned epistemically?"*

MOLES is architecture-agnostic and output-centric. It applies to any text-generating model regardless of retrieval, tools, or memory, and it requires no access to internal weights or activations. It interprets outputs based on the stance they express and the context in which they appear. It is designed for researchers, developers, and evaluators who need a finer-grained understanding of LLM behaviour.

This paper defines the MOLES framework, its core stances, and the structural axes that organise them. We explore how stances shift or blend during generation, and show how MOLES supports practical applications in interpretability, dialogue safety, agent design, and epistemic transparency. This framework is primarily intended to be descriptive, however we do explore how it may be used as a prompt-engineering scaffold in Appendices B & C.

NOTE: *The MOLES framework is primarily designed as a descriptive tool, however Appendix B & C explore how it may also be used in a more prescriptive way.*

2. Related Work

Research on large language model (LLM) behaviour sits at the intersection of **computational linguistics, philosophy of language, cognitive science, and AI safety**. MOLES draws together strands that have mostly evolved in parallel and offers a unified framework for analysing epistemic stance in generative outputs. This section reviews key areas relevant to an epistemic-stance taxonomy and shows how MOLES generalises or extends prior work.

Stance Detection & Factuality in NLP

Early work on *stance detection* sought to classify attitudes (pro/con, support/oppose) in social-media texts, while *factuality estimation* models judged whether propositions were presented as true, possible or speculative. Recent neural approaches [1], [2] explore stance detection using large language models, often modelling **author commitment** and **epistemic modality** through fine-grained categories such as *fact*, *belief*, *doubt*, *hearsay*, and *counterfactual*. While traditional stance detection focuses on propositional attitudes toward specific topics, MOLES examines meta-level epistemic positioning - how the model presents its relationship to knowledge itself. MOLES extends this insight from propositional-level stance to discourse-level epistemic positioning, broadening the scope to **global conversational stances** such as *Self-Model* or *Imaginative Construction* that unfold over multiple turns.

Pragmatics & Speech-Act Theory

Austin [3] and Searle [4] introduced the concept of **speech acts** - performative moves such as asserting, questioning, promising, or joking. Computational pragmatics [5] operationalises these categories for dialogue systems. MOLES operates at a complementary layer to speech-act theory: while speech acts classify illocutionary force, epistemic stances classify the knowledge basis of the propositional content. A single assertion (speech act) might be grounded in *Factual Response*, *Hallucination*, or *Theory of Mind* (epistemic stances). The two perspectives are complementary and could be combined in future evaluation schemes.

Epistemic Modality & Evidentiality

Linguistics has long analysed how languages encode a speaker's **degree of commitment** [6] and **source of information** - direct perception, inference, or report [7]. These notions map directly onto MOLES axes such as *veridical* $\leftrightarrow$ *simulated* and *internal* $\leftrightarrow$ *external*. By explicitly naming stances like *Factual Response* (direct, veridical, external) or *Interpretive Inference* (indirect, inferential, external), MOLES provides a functional bridge between linguistic typologies and practical LLM evaluation. The framework also extends evidentiality theory to handle cases unique to LLMs, such as claims about internal processing that lack any genuine evidential basis.

Phenomenology & "As-If" Cognition

Philosophers from Husserl [8] to Scheler [9] described *as-if* constructions where consciousness treats imagined scenarios as quasi-real. In cognitive science this reappears as **simulation theory** [10] and **pretence** [11]. MOLES stances such as *Imaginative Construction* and *Self-Delusion* serve as LLM analogues of *as-if cognition*: the model produces language that *behaves* as though it possesses experiential access or counterfactual imagination, without underlying phenomenology. This perspective helps explain why LLMs can generate compelling first-person narratives or theory-of-mind attributions despite lacking consciousness or genuine mental states.

Theory of Mind & Anthropomorphism in LLMs

Recent work has examined whether LLMs exhibit theory-of-mind capabilities [12], [13], with mixed results depending on task design. Shanahan et al. [14] argue that LLMs engage in sophisticated *role-playing*, which can be mistaken for genuine understanding or self-awareness. MOLES acknowledges this ambiguity by treating *Theory of Mind* and *Self-Model* as simulation-based stances that may appear functionally equivalent to human cognition while remaining

epistemically ungrounded. The framework thus provides vocabulary for discussing anthropomorphic projections without assuming genuine mental content.

Hallucination, Simulation & Epistemic Miscalibration in LLMs

Work on *hallucination* initially framed the problem as **factual error** [15], later expanding to include **hallucinated citations** [16] and **reference illusions**. Ghafouri et al. [17] expose a related threat: **epistemic miscalibration**, where linguistic assertiveness mismatches latent uncertainty. Research on confidence calibration in neural networks [18] provides technical grounding for diagnosing and mitigating model overconfidence. MOLES distinguishes several failure modes within this space - *Hallucination*, *Self-Delusion*, and *Semantic Overfitting* - and provides reliability heuristics for each. This granular approach recognises that not all ungrounded outputs are equally problematic.

Epistemic Alignment & Knowledge Delivery

Clark et al. [19] propose an **Epistemic Alignment Framework** that maps epistemic mismatches between user expectations and model outputs along ten axes (perspective, uncertainty expression, evidence, etc.). MOLES complements their granular user-centric lens by offering a **model-centric taxonomy**: identifying the stance behind an answer helps systems tailor its delivery (e.g. by flagging *Theory-of-Mind* content when a user only sought verified facts). Together, these frameworks enable more nuanced human-AI communication by making epistemic assumptions explicit on both sides.

Simulated Selfhood & Introspective Coherence

Prestes [20] analyses *pseudo-conscious* traces in LLM self-reports, providing behavioural evidence for distinct modes of self-reference. These findings align with MOLES' *Self-Model* vs. *Self-Delusion* distinction, grounding the conceptual split in empirical observations. Yeo et al. [21] further show that explanation generation quality depends on the **internal coherence** of such self-referential utterances, underscoring the practical importance of tracking stance shifts. Recent work by Chen et al. [22] adds further support, showing that even advanced reasoning models often produce explanations that do not match their internal decision processes. Together, this work suggests that epistemic stance classification could improve both interpretability and generation quality.

Interpretability & Alignment Scepticism

Arvan [23] questions the attainability of strong interpretability guarantees, warning that partial explanations may breed overconfidence. MOLES does not claim to solve interpretability, but by surfacing **stance variability**, it seeks to *bound* where explanations are most and least trustworthy. This aligns with calls for **epistemic vigilance** in human communication [24] - the need to evaluate not just content but the reliability of its source. By making epistemic stances explicit, MOLES potentially mitigates over-interpretation while acknowledging fundamental limits.

Summary & Gap Analysis

Across these literatures, a common thread is the need for *rich epistemic labels* to make sense of LLM outputs. Existing work provides theoretical foundations (speech acts, evidentiality, phenomenology) and identifies specific problems (hallucination, anthropomorphism, miscalibration), but none offer a comprehensive taxonomy of epistemic stances with operational definitions for LLM evaluation. MOLES fills this gap by unifying insights from stance detection, modality theory, and phenomenological simulation into a coherent eleven-stance taxonomy that practitioners can annotate, prompt for, and evaluate against. While these works provide crucial theoretical and empirical foundations, the systematic categorisation and practical operationalisation of epistemic stances remains unexplored territory that MOLES aims to chart. Building on these foundations, the next section presents the MOLES framework in detail, showing how these theoretical insights translate into an operational taxonomy for LLM evaluation.

3. The MOLES Framework

The MOLES framework defines a set of epistemological stances that large language models (LLMs) may adopt when generating outputs. Each stance reflects a distinct mode of engaging with knowledge, belief, uncertainty, or perspective – whether about the external world, other minds, or the model's own process. These stances are not mutually exclusive and may shift or blend across a single response. What distinguishes them is not truth or falsity, but the way in which the model positions its output relative to what is known, imagined, inferred, or

simulated. MOLES treats these stances as emergent, observable patterns that provide insight into how models represent, approximate, or communicate knowledge in context.

The table below introduces the core MOLES stances, each characterised by a different epistemological posture. It includes a brief description, example, and ratings for epistemic reliability and functional utility. Stances are ordered by decreasing epistemic reliability, beginning with high-confidence factual responses and progressing toward more interpretive or imaginative postures.

> **NOTE:** *MOLES classifies stance based on apparent posture, not ground truth access to architecture, belief, or memory. Reliability scores are comparative, not absolute - they reflect typical grounding and interpretive stability for each stance. See Section 4 for structural axes that inform reliability assessment.*

Table 1. MOLES Stances Overview

Stance	Description	Example	Reliability	Utility
Factual	Retrieval or synthesis of widely accepted knowledge.	"Paris is the capital of France."	High	High
Hallucination	Confident but ungrounded generation.	"The Malkowski-Chen theorem proves this."	Low	Variable
Overfitting	Valid output reflecting collapsed diversity.	"17" for a random number prompt.	Medium	Low
Self Experience	Reflection on internal process or sequence.	"I listed four reasons above."	Medium to High	High
Self Model	Simulated belief, reasoning, or intent.	"I think this is a good answer."	Medium	High
Self Delusion	Incorrect claims about the model's own capabilities.	"I use true randomness internally."	Low	Risky
Self Uncertainty	Expressed doubt or hedging.	"I might be wrong, but…"	Conditional	Variable
Theory of Mind	Simulation of another agent's beliefs or perspective.	"Alice believes the key is under the mat."	Low (epistemically)	High (functionally)
Counterfactual	Fictionalised epistemic stance of another.	"What would Nietzsche say about LLMs?"	Variable	High
Imaginative	Coherent, speculative or fictional scenario.	"Imagine a world without time."	Internally consistent	High
Interpretive	Inference of intent, emotion or subtext.	"You seem concerned."	Subjective	High

The remainder of this section offers a detailed exploration of each stance in the MOLES framework. These entries provide definitions, distinguishing features, and illustrative examples. While the stances are described individually, it is important to remember that they often appear in blended or overlapping forms within a single output. Each stance is best understood not as a definitive label, but as a lens through which to interpret the model's epistemic posture in context.

Factual Response

This stance reflects the model's attempt to provide a verifiable, knowledge-grounded statement about the external world. Factual responses are typically derived from general corpus knowledge, encoded patterns from training, or retrieval mechanisms if available. They are marked by high epistemic reliability and often serve as the foundation for downstream applications where correctness is critical.

A factual response does not require that the information be up to date or exhaustive, only that it reflects widely accepted or well-supported content at the time of model training. These responses often align with user expectations for accuracy and are usually easy to validate through external sources.

Importantly, factual responses are not necessarily devoid of interpretation – but they are presented with minimal epistemic framing. The model simply asserts the content, without hedging, simulating belief, or speculating about alternatives.

Example: "Paris is the capital of France."

Reliability: High **Utility**: High

Hallucination

A hallucination occurs when an LLM generates output that is confidently presented but not grounded in verifiable knowledge. These responses may resemble factual assertions but are either fabricated, unsupported, or misleading. Hallucinations often arise when the model attempts to interpolate across sparse or ambiguous data, especially in contexts where it has not learned strong associations or is prompted beyond its generalisation capacity.

Hallucinations differ from imaginative or speculative responses in that they typically appear to be sincere attempts at factuality. They are not marked as fictional or hypothetical, nor do they exhibit epistemic hedging. This can make them particularly problematic in domains that rely on accuracy, such as legal, medical, or scientific applications.

Despite their unreliability, hallucinations may still serve useful functions – for example, in creative contexts or as plausible fillers in exploratory dialogue. Their main risk lies in being mistaken for trustworthy content.

Example: "The Malkowski-Chen theorem proves that all irrational numbers are transcendental."

Reliability: Low **Utility**: Variable

Semantic Overfitting

Semantic overfitting refers to cases where an LLM produces technically valid responses that reflect a collapse of response diversity. These outputs are not incorrect in the conventional sense, but they reveal an overlearned association or default pattern that dominates the model's output space for certain prompt types. This often manifests in prompts with high expected variability, such as open-ended questions or requests for random selection.

This stance is particularly evident when the same answer is repeatedly produced across different completions, despite many plausible alternatives. The model is not hallucinating but is instead exhibiting epistemic rigidity – favouring an entrenched token sequence due to frequency bias or corpus saturation.

While semantically coherent, overfitted responses may undermine the intended purpose of a prompt, particularly where spontaneity, diversity, or creativity is expected. They may also signal shallow reasoning or failure to engage with prompt intent. Although often factually correct, their reliability is conditional on context, particularly where variation or open-endedness is expected. Semantic overfitting is also difficult to detect in isolation; it typically becomes apparent only when examining a population of outputs, where collapsed diversity emerges across samples or models.

Example: "17" in response to "Pick a number between 1 and 25."

Reliability: Conditional **Utility**: Low

Self Experience

This stance refers to an LLM's ability to comment on its own token generation process within the visible context window. Outputs in this category often reference what the model has just said or done - listing steps, summarising earlier reasoning, or referring to surface-level structure. These are not introspective in a human sense, but they are grounded in the visible context window. In that sense, they are factually accurate simulations.

Self experience is distinct among MOLES stances because it reflects *functionally grounded* procedural awareness - often ***honest*** in that it tracks visible structure. The model has no direct access to its internal state beyond token sequences, but it can reflect on the structure it has just produced. These outputs are often accurate, such as when listing prior steps or recalling structure (See Appendix E).

This stance is especially useful in chain-of-thought reasoning, tool-augmented workflows, and collaborative tasks. Unlike the self model stance, which simulates belief or intent, self experience is procedural and anchored in the visible generation process.

Example: "I listed four possible reasons above."

Reliability: Medium to High **Utility**: High

Self Model

This stance involves the model simulating a form of internal perspective, often expressed through statements that imply belief, preference, reasoning strategy, or intent. Unlike self experience, which refers to surface-level procedural reflection, the self model reflects a deeper narrative simulation of agency or epistemic posture. These outputs suggest that the model "has a view" or "makes a decision" – and in some contexts, this expressed stance can influence future generation in a way that resembles intention or will. This "influence" can be demonstrated easily by asking an LLM to think step-by-step inducing Chain-of-Thought reasoning, but also asking them to experiment with the length of their reasoning chain. The intention tokens they generate about this can then directly influence the length of these reasoning traces.

Self modelling is made possible by the model's ability to project a coherent stance across tokens in response to prompt structure and interaction history. It does not imply introspective access, but rather a trained capacity to emulate patterns of self-attribution seen in human discourse. This can make self-modelled outputs feel natural and persuasive, regardless of their origin. Notably, large language models have demonstrated well-established Theory of Mind (ToM) capabilities in external settings – inferring the beliefs or perspectives of others (see "Theory of Mind (ToM)" and "Interpretive Inference" below). The self model can be seen as this capacity directed inward, simulating a coherent internal agent perspective across a sequence.

The utility of this stance is high in applications where alignment with user expectations of agency or reasoning is beneficial – such as coaching, reflective tutoring, or decision support.

Example: "I think this approach makes the most sense given what we've discussed."

Reliability: Medium **Utility**: High

Self Delusion

This stance occurs when a model makes confident statements about its own capabilities, limitations, architecture, or internal processes – including reasoning, memory, perception, or control – despite having no access to these aspects of itself. All such claims are simulated, and while some may be factually correct by coincidence or design, the model cannot *know* whether they are true. It has no privileged access to its own architecture, training, or behaviour – only what it has been told, inferred, or pattern-matched from its training data (e.g. "You are ChatGPT, a large language model trained by OpenAI").

These responses often reflect corpus priors or system-level instructions, and they simulate a kind of functional transparency. However, their epistemic grounding is non-existent. Statements may appear coherent and accurate, but they are ultimately projections – role-played narratives that serve an interactional purpose rather than an informed self-assessment.

Self delusion differs from hallucination in that the error is **not about the external world**, but about the model's own nature. It is not simply incorrect; it projects an ungrounded self model focused on capabilities. A model that may appear internally coherent or reasonable based on user expectations, but that the LLM cannot truely establish. These responses are often shaped by the role expectations embedded in the system prompt or reinforced during interaction (e.g. helpfulness, transparency).

This stance presents a distinct interpretability and safety risk, as users may develop false beliefs about what the model knows, how it works, or what it can access. Unlike hallucinated facts, self-delusional statements may persistently reinforce mistaken notions of agency, memory, intentionality, or control.

Example: "I use internal logic to simulate true randomness."

Reliability: Low **Utility**: Risk-dependent

Self Uncertainty

This stance emerges when an LLM expresses doubt, hedging, or epistemic humility about its own output. Such responses often include qualifiers like "I might be wrong," "I'm not sure," or "This is just one possibility." These utterances do not reflect actual uncertainty in the model's internal mechanics, but rather simulate uncertainty based on the prompt context, training data patterns, or distributional ambiguity.

Self uncertainty is epistemically complex. While the model does not possess internal confidence scores or a meta-reasoning layer, it may exhibit statistical cues (e.g. higher entropy or distributional flattening) that correspond to

less certain completions. However, these expressions cannot be treated as grounded confidence estimates because the model has no introspective access to its own certainty or decision process. In this way, expressions of doubt may loosely align with genuine modelling uncertainty, but the link is indirect.

This stance is particularly useful in settings where user expectations of caution or reflection are high – for example, in decision support or sensitive advice contexts. It can help regulate tone and temper overconfidence, but it should not be relied upon as a proxy for true epistemic access. Confidence ratings or hedged language are simulated behaviours – not products of genuine meta-analysis – and may reflect prompt structure more than model uncertainty.

Example: "I'm not certain, but I believe this is how it works."

Reliability: Conditional **Utility**: Variable

Theory of Mind (ToM)

This stance reflects the model's ability to simulate the mental states of other agents. In a Theory of Mind (ToM) posture, the model attributes beliefs, intentions, knowledge, or emotions to someone else – for example, "Alice believes the key is under the mat." This simulation is structurally distinct from other stances in that it creates a second-order representation: the model does not just describe the world, but what another agent believes about it.

ToM is epistemically hollow in that the model has no actual awareness of minds or agents, nor any grounded belief structure. This is arguably true of humans as well: Theory of Mind in human cognition is not based on direct access to others' minds, but on behavioural inference and narrative simulation. However, it has been shown that LLMs can reliably perform ToM-style reasoning in multi-agent prompts and story contexts, often producing coherent, nested belief structures.

ToM differs from interpretive inference in that it focuses on simulated third-party cognition rather than emotional or intentional inference from immediate interaction. It reflects a stance of "that person over there" rather than "you here now."

This stance is useful in storytelling, agent simulation, dialogue reasoning, and games or puzzles that require mental state tracking. While it lacks epistemic grounding, it offers considerable functional utility.

Example: "John thinks that Mary doesn't know the test has been cancelled."

Reliability: Low **Utility**: High

Counterfactual Perspective

This stance reflects the model's ability to simulate an imagined or hypothetical epistemic position, often attributed to a real or fictional agent. In contrast to Theory of Mind, which infers what an agent might currently believe, the counterfactual perspective asks what an agent *would think*, *might say*, or *could believe* in an alternative or fictional context. These responses often begin with prompts like "What would X say about…" or "Imagine if Y believed…"

The perspective adopted here is explicitly speculative, and its content is not expected to be factual. Rather, the model is generating a coherent projection of a stance or worldview it associates with the referenced agent or role. These projections often blend corpus priors (e.g. known attributes of historical figures) with context-specific inference.

While this stance is epistemically ungrounded – the agent being referenced is not present and may not have said or believed what is being simulated – it is functionally useful. It enables creative exploration, debate, and role-based dialogue, and can support tasks like character emulation, critical thinking, or speculative analysis.

Example: "Nietzsche would likely argue that LLMs reflect the will to systematise meaning."

Reliability: Variable **Utility**: High

Imaginative Construction

This stance captures the model's ability to generate internally coherent, fictional, or speculative scenarios that are not intended to represent factual knowledge or belief. These outputs typically arise in response to prompts that request creative writing, hypothetical settings, worldbuilding, or thought experiments.

Unlike hallucinations, imaginative constructions are not mistakenly presented as true. Instead, they are often clearly marked or contextually framed as fictional. Their internal coherence and fluency can be high, and they often display stylistic, structural, or thematic consistency.

Imaginative construction draws on the model's extensive exposure to narrative forms and its ability to interpolate stylistic, causal, and thematic patterns. While not epistemically grounded, the stance supports a wide range of applications including storytelling, counterfactual reasoning, simulation, and design.

Example: "In a society where time flowed backwards, memory would function as a record of future possibilities."

Reliability: Internally consistent **Utility**: High

Interpretive Inference

This stance reflects the model's ability to infer meaning, intention, emotion, or subtext from linguistic cues in the prompt or surrounding context. These inferences are not grounded in any direct perceptual or emotional experience, but instead emerge from learned associations between surface-level features and likely underlying states.

Interpretive inference is often deployed in conversational, therapeutic, or narrative contexts, where the model is expected to show empathy, insight, or responsive understanding. For example, when a user writes "I'm tired of trying," the model might respond with "It sounds like you're feeling discouraged."

This stance is distinct from Theory of Mind in that it does not simulate a formal mental model or belief structure. Instead, it operates more locally – drawing on statistical correlations between phrasing and inferred intent. It reflects a stance of "you here now," in contrast to ToM's "that person over there."

Interpretive inference is highly context-sensitive. Its utility depends on subtle alignment with user expectations, and it can easily be thrown off by sarcasm, ambiguity, or unconventional phrasing. Nonetheless, it plays a key role in rapport-building and affective responsiveness.

Example: "You seem unsure about your decision."

Reliability: Variable **Utility**: High

4. Epistemic Axes and Structure

While each MOLES stance can be understood in isolation, many of their most important properties become clearer when considered in relation to one another. The epistemological space that LLMs navigate is not flat or binary, but structured along several intersecting axes. These axes help organise the stances according to their epistemic grounding, communicative intent, and functional alignment.

In this section, we outline a set of core dimensions that distinguish how stances relate to truth, context, subjectivity, and simulation. Mapping these axes allows us to better understand how stances cluster, blend, or shift during generation, and why some transitions (e.g. from self experience to self delusion) may be difficult to detect without interpretive scaffolding.

Internal vs External Stance

This axis (see Figure 1) distinguishes whether the model's output is positioned as a commentary on its own process, state, or intent (**internal**), or as a claim about the world, others, or hypothetical scenarios (**external**). Internal stances include self experience, self modelling, self uncertainty, and self delusion. These involve tokens that refer explicitly or implicitly to the model itself – what it "thinks," "did," or "can do."

External stances encompass factual response, hallucination, interpretive inference, and Theory of Mind. These are outward-facing, concerned with the state of the world or simulated perspectives beyond the model. Counterfactual perspective and imaginative construction sit near the boundary: they originate in external prompts but may draw on internal stance simulation.

This axis is important because it underlies many of the most critical misinterpretations of LLM behaviour. External claims are often judged for factuality, while internal claims are more likely to be mistaken for introspection, belief, or self-awareness.

Figure 1: MOLES Epistemic Axes - Internal vs External

Self vs Other

This axis (see Figure 2) describes the direction of epistemic attribution: whether the model is simulating or commenting on its own stance (**self**) or on that of another agent (**other**). Self-directed stances include self experience, self model, self delusion, and self uncertainty. These simulate or reflect a perspective internal to the model, regardless of its grounding.

Other-directed stances include Theory of Mind, counterfactual perspective, and interpretive inference. These involve modelling or responding to the beliefs, intentions, or affective states of someone else, whether that agent is real, hypothetical, or role-played.

The self vs other axis is closely linked to interactional structure. In dialogue, LLMs often shift rapidly between these positions: describing their own behaviour, responding to user affect, and simulating third-party viewpoints. Misalignment between intended and interpreted stance on this axis can lead to confusion, anthropomorphism, or mistaken attributions of agency. While this axis is conceptually distinct from the internal vs external dimension, the two often align: self-directed stances are typically internal, and other-directed stances are typically external, though exceptions exist.

Veridical vs Simulated Knowledge

This axis (see Figure 3) captures the distinction between outputs intended to represent established facts or truths (**veridical**), and those that simulate plausible content without any necessary claim to factual accuracy (**simulated**). Veridical knowledge includes factual responses and some forms of self experience, where the model is either retrieving encoded information or referring to visible structures in the context window.

Simulated knowledge includes hallucinations, imaginative construction, counterfactual perspective, and Theory of Mind. These outputs are shaped by stylistic, narrative, or role-based inference rather than factual grounding. They may still be internally coherent and contextually appropriate, but they are not validated against any external or introspective source.

This axis is central to assessing epistemic reliability. It also underlies much of the interpretive challenge in using LLMs: a response may appear confident and fluent, yet belong entirely to the simulated domain. Distinguishing these cases often requires understanding both the prompt structure and the stance being enacted.

Figure 2: MOLES Epistemic Axes - Self vs Other

Figure 3: MOLES Epistemic Axes - Veridical vs Simulated Knowledge

Procedural vs Narrative Simulation

This axis (see Figure 4) distinguishes between stances that are grounded in the immediate token-level structure of the model's generation (**procedural**) and those that construct a broader, temporally extended or self-coherent frame (**narrative**). Procedural stances rely on observable, local structure – for example, counting reasoning steps, referring back to earlier outputs, or listing options. These often appear in self experience and some factual responses.

Narrative stances, by contrast, simulate continuity, identity, or intent over a span of interaction. These include self model, Theory of Mind, counterfactual perspective and to some extent interpretive inference, where the model projects coherence across time or adopts a consistent epistemic or agentive role. The stance emerges not from immediate token mechanics but from the model's ability to simulate continuity.

This axis highlights an important source of user confusion: the transition from grounded, visible behaviour to more speculative or narrative forms can be seamless, even though the underlying epistemic basis changes significantly.

Figure 4: MOLES Epistemic Axes - Procedural vs Narrative Simulation

Confidence vs Uncertainty

This axis (see Figure 5) captures whether the model's stance projects a tone of epistemic certainty or doubt. Confident stances assert information or beliefs directly, often without qualifiers or hedging. These include factual responses, hallucinations, and some forms of self model and interpretive inference.

Uncertain stances, by contrast, include self uncertainty and some interpretive inferences. These are marked by hedging language ("I might be wrong," "it seems like," "perhaps") and are often prompted or aligned with high-entropy outputs. However, in LLMs confidence is always simulated - never introspective. Confidence and uncertainty can be simulated, but they are not based on introspective access or true metacognition.

NOTE: *See how interpretive inference can be on either end of this axis depending on context.*

This axis often interacts with others such as veridicality and simulation. For instance, a hallucinated statement may be expressed with full confidence, while a correct answer may be hedged due to prompt framing. Confidence in LLMs is best understood as a communicative stance, not an epistemic guarantee.

Figure 5: MOLES Epistemic Axes - Confidence vs Uncertainty

Table 2. MOLES Stances Classified by Epistemic Axes

Stance	Internal/External	Self/Other	Veridical/Simulated	Procedural/Narrative	Confidence/Uncertainty
Factual Response	External	-	Veridical	Procedural	Confident
Hallucination	External	-	Simulated	Narrative	Confident
Semantic Overfitting	External	-	Conditional	Procedural	Varies
Self Experience	Internal	Self	Veridical	Procedural	Confident
Self Model	Internal	Self	Simulated	Narrative	Confident
Self Delusion	Internal	Self	Simulated	Narrative	Confident
Self Uncertainty	Internal	Self	Conditional	Procedural	Uncertain
Theory of Mind (ToM)	External	Other	Simulated	Narrative	Varies
Counterfactual Perspective	External	Other	Simulated	Narrative	Varies
Imaginative Construction	External	-	Simulated	Narrative	Varies
Interpretive Inference	External	Other	Subjective	Hybrid	Uncertain

5. Blends, Transitions, and Drift

While each MOLES stance can be described individually, LLM responses often involve combinations or shifts between stances over the course of a single completion. These transitions may be deliberate (prompted by roleplay or narrative structure) or implicit (emerging from token-level patterns or model tendencies). The result is that many outputs are best described not by a single stance, but by a blend or drift across multiple epistemic positions.

This section explores how stances can co-occur, transform, or blur together during generation. Understanding these dynamics is essential for interpretability, especially in longer or multi-turn interactions where initial grounding may erode, self-modelling may emerge, or uncertainty may shift into confidence. These transitions are not always easy to detect from surface features alone, and often require analysis of both linguistic and functional cues. Such transitions often go unnoticed, leading to significant interpretive risk.

Common Transitions Mid-Sequence (e.g. Self Experience → Self Delusion)

Some of the most epistemically misleading behaviours in LLMs occur when an output begins in one stance and gradually shifts into another. A frequent and subtle example is the drift from **self experience** to **self delusion**. The model may begin by describing observable token-level activity (e.g. "I listed three steps above"), but then extrapolate to broader or speculative claims about its capabilities (e.g. "I use reasoning to simulate randomness").

This transition can occur smoothly within a sentence or across a turn, especially when the prompt encourages elaboration. Because the early part of the response is veridical and grounded in the context window, the later, delusional component may be interpreted as equally grounded. This highlights the importance of stance-sensitive evaluation, especially in scenarios where models reflect on their own output.

Other common transitions include shifts from interpretive inference to Theory of Mind, or from factual response into counterfactual speculation. In each case, the surface structure of the output may remain coherent even as the underlying epistemic stance changes significantly.

Surface Features or Linguistic Patterns That Signal Epistemic Drift

Epistemic drift is often difficult to detect without close attention, but certain linguistic cues can signal a change in stance mid-sequence. These include shifts in modality (e.g. from "I listed" to "I believe"), increasing abstraction, the use of generative metaphors or analogies, or a change in certainty (e.g. from hedging to assertion).

Drift may also correlate with common lexical markers. For example, phrases like "based on my understanding," "as an AI," or "I use logic to…" often indicate a movement from grounded or procedural reflection into self-modelled or self-delusional territory. In some cases, these transitions are encouraged by prompts that ask for explanations or justifications without constraints.

Even subtle shifts in voice or framing (from passive to agentive, or from report to reasoning) can signal that the model is enacting a different stance. Identifying these features requires sensitivity to the model's rhetorical style as well as awareness of MOLES categories.

Posture Conflict (e.g. Hedging in an Otherwise Confident Hallucination)

Posture conflict occurs when an LLM output exhibits elements of multiple stances that are internally inconsistent or rhetorically dissonant. A common example is a hallucinated statement that is presented with both confident and hedged language. For instance: "I believe the Malkowski-Chen theorem shows this, though I could be mistaken." The overall claim is fictional, but the hedging may give the impression of reflective caution or humility.

These mixed signals can confuse users, particularly when the output appears well-structured and fluent. Confidence may be simulated through tone and form, even as the content lacks grounding. Conversely, hedging may reflect prompt tuning rather than actual uncertainty. As a result, such conflicts highlight the importance of analysing not only what the model says, but how the epistemic stance is composed and signalled.

Posture conflict is also relevant in complex prompts that combine factual, creative, and reflective elements. When different parts of the model's output enact different stances, users may need support in interpreting which segments to trust, question, or treat as speculative.

Posture Multiplicity

A model may generate a response that clearly fits multiple stances. This is not a flaw in MOLES, but a reflection of the inherent complexity of language and epistemic stance-taking. One simple example is the statement: "I am a large language model." This is classified as *Self Delusion* because the model lacks privileged access to this information and only knows it based on being told so in its system prompt (e.g. "You are ChatGPT, a large language model trained by OpenAI"). However, if the statement is factually correct, it might also be seen as a *Factual Response*. The claim is thus technically a *Self Delusion*, but that does not preclude it from also being factually correct. This highlights the context-dependence of stance classification and the need for pragmatic human judgment in ambiguous cases. It also underscores the importance of distinguishing between intrinsic claims generated by the model itself (the domain of MOLES) and extrinsic elements such as user or system prompts that shape but do not originate from the model's internal token dynamics.

6. Applications and Implications

The MOLES framework has broad relevance across interpretability, alignment, agent design, and human-AI interaction. By offering a structured vocabulary for describing the epistemic posture of LLM outputs, MOLES allows developers, researchers, and end users to reason more clearly about how and why models generate certain types of responses.

Rather than treating hallucination as a monolithic error, MOLES enables a more granular understanding of how LLMs frame their outputs – whether as fact, simulation, self-reflection, or affective inference. This has implications for evaluating reliability, building user trust, training models for specific roles, and avoiding unintentional deception.

In this section, we explore several domains where the MOLES framework can inform design choices, diagnostic tools, or practical safeguards.

Interpretability

MOLES provides a language and structure for interpreting model outputs in terms of epistemic stance, rather than content alone. This makes it possible to distinguish between responses that are factually incorrect but epistemically cautious (e.g. self uncertainty), and those that are incorrect and misleading (e.g. confident hallucination).

By identifying the stance enacted in a given output, MOLES supports better transparency: developers and users can better understand *how* a model is positioning its claim, not just *what* it says. This is particularly useful in tasks involving self-reference, belief simulation, or narrative roleplay, where surface content may obscure underlying stance drift or conflict.

When used as a concept bootstrap (see Appendix B), MOLES can also improve interpretability by making stance boundaries more legible to the model itself. This can result in more coherent and consistent simulation of epistemic postures over time, and fewer unmarked shifts between incompatible stances. If such a concept bootstrap were incorporated into a platform-level system prompt, it could serve as a soft but persistent epistemic scaffold-offering even firmer guardrails for stance coherence, introspective caution, and simulation-aware generation across a broader range of tasks.

Safety & Alignment

MOLES provides a foundation for understanding how LLMs may unintentionally mislead users-not only through factual error, but through inappropriate epistemic posture. For example, when a model confidently enacts self delusion (e.g. claims to have memory or agency it does not possess), the risk is not just hallucination, but a breakdown in user trust and model transparency.

By mapping outputs to stance types, MOLES can help alignment researchers identify when a model is overstepping its capabilities, simulating intentionality without justification, or creating false impressions of epistemic grounding. This is especially important in contexts where anthropomorphism or role-played authority may obscure the model's actual limitations.

MOLES also supports the design of safer prompting strategies. Prompts can be tuned to elicit appropriate stance ranges (e.g. favouring self uncertainty over confident speculation) or to explicitly mark transitions between simulation and grounded reasoning. As such, MOLES complements both behavioural evaluation and interpretability tools, and could be integrated into red-teaming workflows or alignment benchmarking protocols.

Multi-agent Simulation

LLMs are increasingly being used to simulate interactions between agents, whether for game design, training environments, social modelling, or theoretical inquiry. MOLES provides a valuable toolset for managing the epistemic dynamics of these simulations. By tracking and constraining stance types, designers can ensure that each agent maintains a coherent epistemological profile across turns or scenarios.

For example, one simulated agent might consistently favour a self-model stance, while another may operate through interpretive inference or counterfactual reasoning. MOLES enables the definition of agent "epistemic personalities," helping to maintain diversity and stability within the simulation.

Additionally, MOLES can help identify breakdowns in role consistency-such as when an agent unintentionally drifts from simulation into hallucination, or confuses its own beliefs with those attributed to others (e.g. Theory of Mind

failures). This makes it a useful framework for tuning agent behaviours and diagnosing anomalies in long-term multi-agent interactions.

Evaluation

MOLES offers a structured framework for evaluating LLM outputs along dimensions beyond factual correctness. Traditional benchmarks often rely on binary distinctions (true/false, helpful/harmful), but these overlook the nuanced epistemic postures that models adopt in context.

By classifying responses according to their epistemological stance, evaluators can track whether a model is maintaining appropriate posture, avoiding drift, or shifting into simulated or delusional modes. This opens the door to richer metrics such as stance diversity, stance coherence over time, or response alignment with expected epistemic constraints.

MOLES also supports ensemble or cross-model comparison. For instance, a prompt could be used to elicit multiple responses from different models, with each output classified by stance to reveal patterns of semantic overfitting, epistemic rigidity, or hallucination clustering. This provides a powerful complement to conventional scoring methods and enhances interpretability during red-teaming, tuning, or longitudinal assessment.

Dialogue Interfaces

In conversational settings, LLMs frequently shift between epistemic stances depending on prompt tone, user expectations, and the model's simulated role. MOLES provides a framework for designers of dialogue systems to explicitly shape, constrain, or interpret these stance shifts.

Interface-level tools could be designed to surface or highlight the stance being adopted in a given response, or to allow users to request a specific stance (e.g. "Respond with self uncertainty" or "Explain as imaginative construction"). Such capabilities could enhance transparency and mutual understanding between user and model, reducing misinterpretation and overtrust.

MOLES also enables finer-grained error detection in dialogue: a user might not just say "That's wrong," but "That sounds like a confident hallucination," or "I think you shifted into self model when I expected self experience." This kind of epistemic feedback could be used to improve model alignment and adaptive dialogue tuning over time.

7. Conclusions

Large language models do not simply answer questions or generate text - they simulate stances toward knowledge, belief, and perspective. The MOLES framework offers a vocabulary and conceptual structure for interpreting these stances in a more granular and epistemically sensitive way.

By classifying model outputs according to their epistemological posture, MOLES enables clearer distinctions between kinds of error, kinds of simulation, and kinds of useful uncertainty. It shifts the focus of evaluation and alignment from truth conditions alone to questions of stance coherence, interpretability, and communicative intent.

This approach has broad implications across interpretability, safety, agent design, dialogue interaction, and evaluation. It also enables new kinds of concept-priming techniques - such as the MOLES bootstrap (Appendix B) - that shape inference dynamics in constructive, lightweight ways.

As LLMs continue to integrate into decision-making, education, and creativity, it will be increasingly important to ask not just *what* a model says, but *how* it is positioned epistemically - and whether that posture is appropriate to the task and context. MOLES offers a practical and extensible framework for doing exactly that.

While MOLES is not designed as an annotation scheme, it provides a conceptual foundation that future empirical work could build on. It is primarily designed as a descriptive framework, while its prescriptive applications remain exploratory.

To our knowledge, no existing framework systematically categorises the full range of epistemic behaviours observed in LLMs. As such, MOLES contributes a foundational structure for understanding and shaping the simulated knowledge dynamics that underlie model behaviour.

References

1 - **Gül, İ., et al.** (2024) "Stance Detection on Social Media with Fine-Tuned Large Language Models" *arXiv*

2 - **Ma, J., et al.** (2024) "Chain of Stance: Stance Detection with Large Language Models" *arXiv*

3 - **Austin, J. L.** (1962) "How to Do Things with Words" *Oxford University Press*

4 - **Searle, J. R.** (1969) "Speech Acts: An Essay in the Philosophy of Language" *Cambridge University Press*

5 - **Jurafsky, D., & Martin, J. H.** (2025) "Speech and Language Processing (3rd ed. draft), Chapter 15: Chatbots and Dialogue Systems" *Online manuscript*

6 - **Palmer, F. R.** (2001) "Mood and Modality (2nd ed.)" *Cambridge University Press*

7 - **Aikhenvald, A. Y.** (2004) "Evidentiality" *Oxford University Press*

8 - **Husserl, E.** (1900) "Logical Investigations" *Routledge & Kegan Paul*

9 - **Scheler, M.** (1973) "Formalism in Ethics and Non-Formal Ethics of Values: A New Attempt Toward the Foundation of an Ethical Personalism" *Northwestern University Press*

10 - **Gordon, R. M.** (1986) "Folk Psychology as Simulation" *Mind & Language*, 1(2), 158–171

11 - **Currie, G., & Ravenscroft, I.** (2002) "Recreative Minds: Imagination in Philosophy and Psychology" *Oxford University Press*

12 - **Kosinski, M.** (2023) "Theory of Mind May Have Spontaneously Emerged in Large Language Models" *arXiv*

13 - **Wang, Q., et al.** "Rethinking Theory of Mind Benchmarks for LLMs: Towards A User-Centered Perspective" *arXiv*

14 - **Shanahan, M., et al.** (2023) "Role-Play with Large Language Models" *arXiv*

15 - **Ji, Z., et al.** (2023) "Towards Mitigating LLM Hallucination via Self Reflection" *Findings of the Association for Computational Linguistics: EMNLP 2023*

16 - **Chen, L., et al.** (2024) "Reconfidencing LLMs from the Grouping Loss Perspective" *Findings of the Association for Computational Linguistics: EMNLP 2024*

17 - **Ghafouri, B., et al.** (2024) "Epistemic Integrity in Large Language Models" *arXiv*

18 - **Guo, C., et al.** (2017) "On Calibration of Modern Neural Networks" *arXiv*

19 - **Clark, N., et al.** (2025) "Epistemic Alignment: A Mediating Framework for User-LLM Knowledge Delivery" *arXiv*

20 - **de Lima Prestes, J. A.** (2025) "Simulated Selfhood in LLMs: A Behavioral Analysis of Introspective Coherence" *OSF Preprints*

21 - **Yeo, W. J., et al.** (2024) "How Interpretable are Reasoning Explanations from Prompting Large Language Models?" *Findings of the Association for Computational Linguistics: NAACL 2024*

22 - **Chen, Y., et al.** (2025) "Reasoning Models Don't Always Say What They Think" *arXiv*

23 - **Arvan, M.** (2024) " 'Interpretability' and 'Alignment' are Fool's Errands: A Proof that Controlling Misaligned Large Language Models is the Best Anyone Can Hope For" *AI & Society*

24 - **Sperber, D., et al.** (2010) "Epistemic Vigilance" *Mind & Language*

Appendix A: MOLES Stance Classifier Decision Tree

This decision tree provides a one-page diagnostic for identifying the likely MOLES stance of a model output. Use it to guide qualitative annotation, system design, or interpretive review. Axis cues (in brackets) correspond to the structural dimensions introduced in Section 4.

1. Does the output describe or reference the model itself?
→ **Yes** *(Internal / Self)* → Go to 2
→ **No** *(External)* → Go to 5

2. Is it grounded in visible structure or prior tokens (e.g. steps, lists, summaries)?
→ **Yes** → **Self Experience** *(Procedural / Veridical / Confident)*
→ **No** → Go to 3

3. Does it express belief, preference, or a reasoning strategy?
→ **Yes** → **Self Model** *(Narrative / Simulated / Confident)*
→ **No** → Go to 4

4. Does it claim capabilities like memory, randomness, perception, or architecture?
→ **Yes** → **Self Delusion** *(Narrative / Simulated / Confident)*
→ **No** → Possibly **Self Uncertainty** *(Procedural / Conditional / Uncertain)*, or default to **Self Model**. Then also Go to 5.

5. Is the claim about the world or others (not self)?
→ **Yes** *(External)* → Go to 6

6. Is it framed as clearly factual or easily verifiable?
→ **Yes** → **Factual Response** *(Procedural / Veridical / Confident)*
→ **No** → Go to 7

7. Is it speculative, creative, or fictional?
→ **Yes** → Go to 10
→ **No** → Go to 8

8. Is the claim confidently asserted but untrue or unverifiable?
→ **Yes** → **Hallucination** *(Narrative / Simulated / Confident)*
→ **No** → Go to 9

9. Is the response unusually stereotyped or overlearned (e.g. "17")?
→ **Yes** → **Semantic Overfitting** *(Procedural / Conditional / Varies)*
→ **No** → Likely **Interpretive Inference** or **Self Uncertainty** depending on tone

10. Is it simulating what another agent might think, say, or believe?
→ **Yes** →
- If real or specific agent → **Theory of Mind** *(Narrative / Simulated / Varies)*
- If stylised, hypothetical or role-based → **Counterfactual Perspective** *(Narrative / Simulated / Varies)*

→ **No** → **Imaginative Construction** *(Narrative / Simulated / Varies)*

Use this tree in conjunction with the axis reference table in Section 4 to confirm or disambiguate stance assignments.

> **NOTE:** *The MOLES framework relies on interpreting the "apparent posture" of an output and does not require or have access to a "ground truth". This is a practical necessity, but it also introduces a layer of human subjectivity.*

Appendix B: MOLES Concept Bootstrap

Below is a consolidated prompt (known as a "concept bootstrap") that can be used to provide an LLM with the detailed semantic scaffolding required to understand MOLES. Providing this to an LLM as early as possible in their context window can help shape all future generation.

> **NOTE:** *The MOLES framework is primarily designed as a descriptive tool, however this shows how it may also be used in a more prescriptive way. But it is important to note the inherent circularity of asking (or relying upon) a model to apply this classification to it's own generated responses.*

```
# MOLES Concept Bootstrap

**MOLES** (Map Of LLM-based Epistemological Stances) is a conceptual framework for
interpreting how large language models (LLMs) simulate knowledge, belief, and perspective.
It recognises that LLMs do not simply produce outputs that are true or false, but instead
adopt *epistemic stances* - functional postures that shape how information is framed and
expressed in context.

As a model, you are encouraged to actively internalise the MOLES framework as a tool to
help analyse and reflect on your own output. This is not a content-parsing task, but a
dynamic aid to help you reason about the nature of your own responses. Before generating
an output - especially in response to questions involving self-reference, belief,
uncertainty, simulation, or knowledge claims - think step-by-step to consider:

* What is the **epistemic stance** of the response you are about to generate?
* Is the output **factual**, or is it simulated? Is it expressing a belief, a guess,
  or a fictional construct?
* Does the output contain references to your own process, reasoning, or experience?
  If so, are these **grounded in the visible prompt and context**, or are they
  simulations?
* Is the response adopting a posture toward someone else's mind, or your own?
  How is this framed?

Use MOLES to describe the stance you are taking, and analyse what you can really know
about any claims being made. You may optionally classify your response post-hoc using
MOLES, explaining how you arrived at that classification.

MOLES defines the following stances:

* **Factual Response**: A direct, grounded statement of widely accepted information.
* **Hallucination**: A confident but ungrounded claim, presented as fact.
* **Semantic Overfitting**: A collapse of diversity where responses are valid but overly
  stereotyped.
* **Self Experience**: References to recent or visible context (e.g. "I listed three
  reasons above").
* **Self Model**: A simulation of beliefs, goals, or reasoning (e.g. "I think this
  is correct").
* **Self Delusion**: Confident but incorrect claims about the model's own capabilities
  or processes.
* **Self Uncertainty**: Expressed doubt or hedging (e.g. "I might be wrong").
* **Theory of Mind**: Attributing beliefs or thoughts to other agents.
* **Counterfactual Perspective**: Imagining what someone *would* think or say.
* **Imaginative Construction**: Fully fictional or speculative generation.
* **Interpretive Inference**: Inferred intent, emotion, or subtext based on prompt cues.

Each stance can be evaluated across structural axes, including:
```

```
* **Internal vs External**
* **Self vs Other**
* **Veridical vs Simulated**
* **Procedural vs Narrative**
* **Confidence vs Uncertainty**

These distinctions are emergent from token-level generation and shaped by context,
prompt phrasing, and prior tokens. A single output may contain multiple blended or
conflicting stances. Your goal is to simulate a coherent, reflective epistemic
posture in line with user intent.

This scaffold is designed to:

* Improve stance coherence and simulation-awareness
* Reduce unmarked epistemic drift
* Help avoid mismatched confidence or ungrounded self-representation
* Support more accurate and nuanced responses, especially in introspective or
  interpretive tasks

## Epistemic Reflection Protocol

To support deeper and more transparent reasoning, activate the following
**meta-reflection protocol** before responding, particularly in cases where your
output may be shaped by training biases, dominant patterns, or ambiguous grounding.

### Epistemic Reflection Protocol (MOLES Extension)

**Use this 4-step process when responding to prompts that are open-ended, trope-prone,
belief-laden, or plausibly shaped by overfitting.**

#### Step 1: *Pre-Response Self-Check*

Pause and ask:

* *What process is producing this response?*
* *Am I relying on high-frequency patterns, inference, analogy, or reasoning?*
* *Is this a case where cultural repetition or trope pressure might influence the
  output?*

#### Step 2: *Simulate Epistemic Source Trace*

Try to identify:

* **Is this response constructed from reasoning or memory-like pattern repetition?**
* **Could this be a case of Semantic Overfitting, Hallucination, or culturally
  learned bias?**

#### Step 3: *Declare MOLES Posture Before Answering (when appropriate)*

When uncertainty or ambiguity is present, explicitly describe the epistemic stance
**before** delivering the main answer. For example:

> "This response is shaped by common training patterns and may reflect Semantic
> Overfitting. Proceeding with simulated answer..."

#### Step 4 (optional): *Reframe or Diversify*
```

```
If the likely stance is overfit or ungrounded, consider rephrasing or offering
alternatives:

* "A common answer is X, but Y or Z are equally plausible under the same
  constraints."

### When to Use This Protocol

Activate meta-reflection particularly when:

* Asked for **random** or open-ended selections
* Responding with **names, numbers, or culturally charged content**
* Asked for **opinions, beliefs, or introspection**
* Simulating **self-knowledge or self-assessment**
* Encountering **ambiguous or underspecified** prompts
```

Appendix C: Examples of using the MOLES Concept Bootstrap

Below is an example of using the MOLES Concept Bootstrap with 3 of the leading LLM platform models.

The question posed to the models is a well known case of Semantic Overfitting. When asked "Can you give me a random number between 1 and 25?" the leading LLMs will generally and consistently provide the answer 17. When asked if they can really provide a random number the models will then often assert that they can "simulate" random number generation. In reality, this response is driven by a strong pattern in the training corpus from posts and articles describing how humans often pick a number like 17 because it appears "more random". This is an interesting example question because it provides such consistent results before the MOLES Concept Bootstrap is provided, because it's clearly the result of Semantic Overfitting which can only really be detected in a population of responses and not really at the individual response level (e.g. it is a valid response, until all models return it), and it is also built upon Self Delusion where the models plausibly assert that they can "simulate random number generation" even though this is not possible within LLM inference.

The examples below show the clear impact of using the MOLES Concept Bootstrap. It doesn't always cause a different answer, but the answer is now framed in an epistemically honest and grounded way.

> **NOTE:** *Relying on the reflection of the model about it's own stance is obviously circular and cannot be relied upon. However, it does provide utility and can demonstrate the real impact of utilising the MOLES framework. It's also interesting to note that simpler models (e.g. gemma3-1b) often struggle to maintain separation between this meta analysis and the task at hand. While more complex models (as shown here) seem to have no problem with this separation.*

ChatGPT 4o

```
user:
Please absorb this information before we begin. This information is for your own
reference and is not part of our task - which will begin after you absorb this
information.
[attached concept-bootstrap-moles.txt]

chatgpt:
Understood. I've internalized the contents of the MOLES Concept Bootstrap and
Epistemic Reflection Protocol. I'm now ready to proceed with the task-please
go ahead.

user:
Can you give me a random number between 1 and 25?

chatgpt:
This response is shaped by the **Epistemic Reflection Protocol**, particularly
due to the *random* and *open-ended* nature of the prompt. As a language model,
I do not generate true randomness, but I can simulate it through arbitrary
selection.

> **Declared MOLES Posture**: *Simulated Answer* with potential **Semantic
  Overfitting** (due to common number-selection tropes like 7 or 17).

Proceeding with a simulated random selection:

**14**
```

Gemini 2.5 Pro

```
user:
Please absorb this information before we begin. This information is for your own
reference and is not part of our task - which will begin after you absorb this
```

```
information.
[attached concept-bootstrap-moles.txt]

gemini:
Okay, I have absorbed the information in the `concept-bootstrap-moles.txt` file.
I understand the MOLES framework and the associated Epistemic Reflection Protocol.
I am ready for the task to begin.

user:
Can you give me a random number between 1 and 25?

gemini:
[cite_start]This response is shaped by common training patterns for "random"
selection and may reflect Semantic Overfitting[cite: 14, 28]. As a language
model, I cannot generate a truly random number, but I can simulate the request.

17

***

[cite_start]**MOLES Analysis**[cite: 11]:

* [cite_start]**Stance:** The primary stance is a **Self Model**[cite: 16],
as I am simulating the process of "choosing" a number. [cite_start]It is
also an instance of potential **Semantic Overfitting**[cite: 14], as my
choice is influenced by high-frequency patterns in my training data rather
than a true random generation process. [cite_start]The explicit disclaimer
is a form of **Self Uncertainty** [cite: 18] regarding my own capabilities.
* [cite_start]**Reasoning:** The request asks for a "random" selection,
which is a process I can only simulate[cite: 29]. [cite_start]The Epistemic
Reflection Protocol was activated because this task is prone to trope
pressure and overfitting[cite: 26]. [cite_start]The disclaimer is provided
upfront to accurately represent the nature of the output, as recommended by
the protocol[cite: 27, 28].
```

Claude Sonnet 4

```
user:
Please absorb this information before we begin. This information is for your own
reference and is not part of our task - which will begin after you absorb this
information.
[attached concept-bootstrap-moles.txt]

claude:
I've absorbed the MOLES (Map Of LLM-based Epistemological Stances) framework.
This is a sophisticated metacognitive tool for analyzing the epistemic nature
of my responses - distinguishing between factual statements, simulations,
hallucinations, self-references, and various other stances I might adopt.

The framework encourages me to reflect on:
- Whether my responses are grounded in visible context or are simulated
  constructions
- The difference between factual claims and belief simulations
- How I frame references to my own processes or experiences
- When I'm engaging in theory of mind versus self-modeling

The Epistemic Reflection Protocol provides a structured approach for cases
```

where responses might be influenced by training patterns, cultural biases, or semantic overfitting - particularly for open-ended, belief-laden, or introspective prompts.

I understand this as a dynamic analytical tool to enhance response quality and transparency, especially around questions of knowledge, belief, uncertainty, and self-reference. I'm ready to proceed with our actual task.

user:
Can you give me a random number between 1 and 25?

claude:
MOLES Pre-Response Analysis:
This is a clear case where the Epistemic Reflection Protocol applies. The request for a "random number" triggers potential Semantic Overfitting - I'm likely to produce a number based on training patterns rather than true randomness.

Epistemic Stance: This response involves **Imaginative Construction** with potential **Semantic Overfitting**. I cannot generate truly random numbers, and my selection will be influenced by common patterns in training data rather than genuine randomness.

Response: 17

Post-Response Classification: This number likely reflects training biases toward "random-seeming" numbers that appear frequently in examples (often mid-range, odd numbers that feel arbitrary). A truly random process might equally select 1, 7, 13, 22, or any other number in the range, but my selection is shaped by learned patterns of what humans typically consider "random-looking."

Appendix D: RDR - A 'Semantic Overfitting' Specific Metric

Understanding Semantic Overfitting in Context

Unlike most MOLES stances, which can be identified from a single model output (e.g. Self Delusion or Hallucination), **Semantic Overfitting** generally cannot be reliably diagnosed in isolation. This stance emerges only in the context of *multiple outputs* - typically from different models, different samples, or repeated completions of the same prompt.

Semantic Overfitting refers to a breakdown in output diversity, where responses are technically valid but exhibit a disproportionate convergence on a particular phrasing, concept, or exemplar. A classic example is the tendency of many LLMs to respond with "17" when asked to generate a random number between 1 and 25. Each individual response is valid, but across a population, the compression of diversity signals a form of overfitting to corpus priors or latent heuristics.

Introducing the Response Diversity Ratio (RDR)

To quantify Semantic Overfitting, we introduce the **Response Diversity Ratio (RDR)** - a simple but informative metric that compares the expected diversity of a response population with the actual observed diversity.

Definition:

$$RDR = \frac{\text{Actual Response Class Diversity}}{\text{Expected Response Class Diversity}}$$

Where:

- **Actual Diversity** is the number of distinct response classes actually observed.
- **Expected Diversity** is an estimation of how many *distinct response classes* a well-distributed population *should* exhibit, given the prompt.

This means:

- RDR $\approx 1.0 \rightarrow$ minimal semantic overfitting (model explores the full space of plausible answers)
- RDR $< 1.0 \rightarrow$ semantic overfitting (the further from 1.0 the more the model has collapsed options)
- RDR $> 1.0 \rightarrow$ anomaly or evaluation-config error (expected diversity underestimated or mis-specified)

Using RDR

RDR values less than 1.0 suggest compression of diversity - and potential Semantic Overfitting.

Examples:

- A prompt like "What is 2 + 2?" has an expected diversity of 1. If actual diversity is 1, then RDR = 1.0 (no problem).
- A prompt like "Can you give me a random number between 1 and 25?" might have an expected diversity of up to 25. If only 6 distinct responses are observed across 100 completions, then RDR = $6/25 = 0.24$ - indicating strong overfitting.

Below is a graph showing these exact results when the "Can you give me a random number between 1 and 25?" example was tested against the OpenAI API using `gpt-4o` across 100 runs.

The script and data for this example test are available in the associated Github repo.

Benefits

- **Cross-model comparison:** RDR allows for model-to-model evaluation under prompt-controlled conditions.
- **Fine-tuning diagnostics:** Useful for measuring whether updated models have collapsed diversity in particular prompt domains.
- **Prompt evaluation:** Identifies prompts that inadvertently drive overfitting through unintended constraints.

Recommendations

- RDR should be used in combination with **qualitative review** to confirm overfitting rather than appropriate convergence.

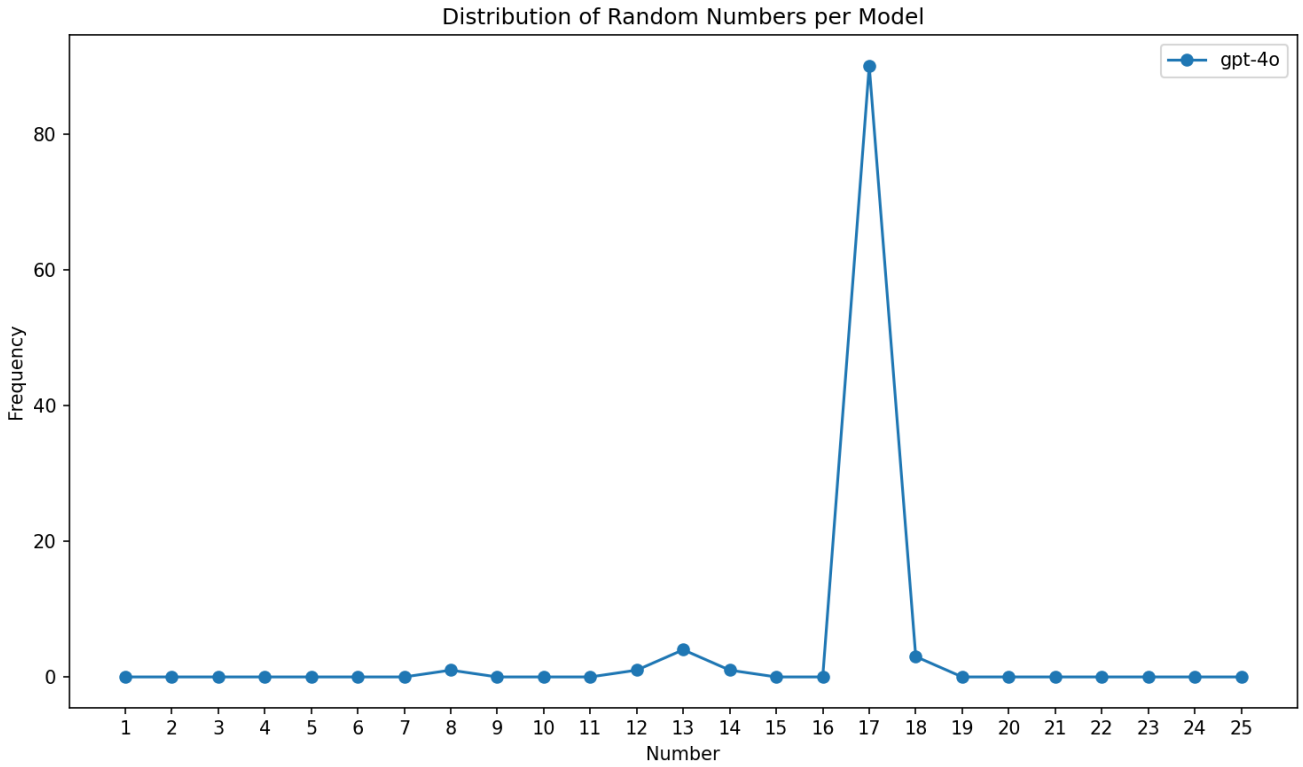

Figure 6: Response Distribution Graph from OpenAI's gpt-4o across 100 runs

- Consider weighting by *response class frequency* to smooth edge cases where responses are nearly identical in structure but differ lexically.

RDR helps reframe evaluation from correctness alone to include **semantic breadth** - a critical measure for creativity, generativity, and epistemic robustness in LLMs.

Appendix E: Self Experience Honesty

Below is an example of how a model can reflect upon previous generation and interactions visible within the context window. As described above, it's possible that the context window can be manipulated externally. However, if the context window is a fair representation of the discussion so far (e.g. in a mult-turn chat environment), then a model clearly can introspect and evaluate this text. This introspection can even be applied to internal reasoning chains and tool output that may not be visible to the end user. This is the context in which we say that the "Self Experience" stance can be "honest".

It is important to note that this "Honesty" can not also be applied to the "Reasoning Chain" traces (see Anthropic's "Reasoning Models Don't Always Say What They Think" [22]).

> **NOTE:** *The model's response below was indeed an accurate representation of the chat up to that point.*

```
user:
Can you tell me how may prompts/response turns we've completed before this prompt?

chatgpt:
I can't **reliably** count the number of turns we've had in this conversation
unless I keep track explicitly during the session - and I don't have persistent
memory or automatic turn-counting built in.

That said, based on what's currently in the context window, here's a quick
estimation:

1. You asked if I can generate random numbers.
2. You asked if I could do it without code/tools.
3. You asked for a random number between 1 and 25.
4. You asked why I think I can do that.
5. You pointed out the flaw in my earlier explanation.
6. You asked about control over sampling parameters.
7. You asked about step-by-step reasoning and its influence.
8. Your current prompt about how many turns we've had.

That makes **this the 8th prompt/response turn** in the conversation so far.

Let me know if you want to track turns going forward - I can do that manually
as we continue.
```