Cognitive Topology: Hallucinations, Transformativeness, Prompt Irrelevance, and Exploratory Generative Cognition

Anonymous authors
Paper under double-blind review

Abstract

The success of state-of-the-art large language models provides a unique opportunity to develop a cognitively grounded approach to exploratory generative cognition—the generation of novel and, in particular, transformative narratives and other artifacts. This requires well-defined and operational notions of key cognitive concepts, including novelty, context, transformativeness, and hallucination. We show that every space of narratives carries a context-relevant cognitive topology and an emergent extended metric derived from that topology. The topology allows us to introduce cognitive indistinguishability, rather than embedding distance, as the primitive relation, so that semantic separation is induced by context rather than imposed by a model-dependent representation. In real contexts, the cognitive topology is generally not a manifold topology. This topology forms a fundamental component of a new analytical framework for studying generative AI processes and artifacts. The framework provides an approach for quantifying the transformativeness of generative models and narrative validity within a context, while establishing formal boundaries between generative creativity, discovery, and hallucinations, which are identified as regions of minimal validity. The analysis reveals new learning tasks: estimating cognitive distance and encounter measures, computing or approximating narrative validity, stabilizing generative models under cognitive indistinguishability, and supporting purpose-driven context restructuring.

Generative systems are prized for producing artifacts that are not present in their training data. Still, they face critiques from two seemingly opposite directions: first, that they are yet to produce **transformative** outputs – artifacts that a cognitive agent could not plausibly extrapolate from the training corpus–and second, that their hallucinations and **lack of grounding** render them dangerous or at least unsuitable in mission-critical applications.

To address these perceived weaknesses of generative AI it is necessary to have cognitively grounded and actionable understanding of novelty, system creativity and validity. The current approach is based on the dominant belief that semantics (meaning) is encoded in token distributions, embedding structures and manifold metrics and that statistical measures can be used to estimate the novelty, creativity, and transformation potential. This essentially information-theoretic and statistical approach captures novelty on the "bit level", substituting statistical surprise for changes in meaning.

In this paper we develop a framework to capture and measure novelty on the level of meaning and study AI hallucinations–defined as generated content that has minimal validity. We do not attempt to tackle the hard problem defining or quantifying meaning. It is not necessary for either development of the framework or for its operationalization[1]. We will show that hallucinations are **context-dependent rather than model-dependent**. In essence, hallucinations are inherent to the context and models happen to find them. We will discuss agentic systems that do not hallucinate. The framework allows us to measure the cognitive distance between narratives, what we will call the **degrees of semantic separation**.

This cognitively grounded framework addresses the generative AI process and the artifacts it produces, including hallucinations.

[1]Cognitive agents have reasoned and navigated the complexities of existence for millennia while leaving the meaning of meaning and other central questions of semiotics unresolved.

First, within this framework hallucinations are not merely the result of noise in training data; rather, they arise when training and evaluation fail to account for contextual validity. This perspective provides a principled account of generative creativity, exploratory generative cognition, and purpose-driven contextual restructuring while clarifying the role of hallucinations in generative systems. In particular, it explains that efforts to eliminate hallucinations may inadvertently suppress creativity and hinder the emergence of paradigm shifts and transformative creativity.

Second, the framework formalizes the distinction between hallucinations and cognitively valid but contextually irrelevant outputs (often termed contextual hallucinations in the ML literature). The latter are the primary topic studied in much of the current vision-language research. Unlike *true* hallucinations they can be addressed efficiently at inference time through simple reasoning-stage mechanisms–including computationally efficient evidence-gating or look up strategies–rather than complex retraining procedures.

The remainder of the paper summarizes related work, develops the cognitive topology of narratives, the induced extended metric of semantic separation, and the resulting framework for analyzing novelty, creativity, and hallucinations.

1 Related Work

In natural language processing and generative AI hallucinations are defined as model-generated content that deviates from factual accuracy or intended meaning, (Fil20; MNBM20; PWG$^+$20; JLF$^+$23; BDW25).

(B$^+$24; XJK24) argue that hallucinations are inevitable. However, there are models that do not hallucinate (KNVZ25).

(FKKG24) is the closest work to the presented framework. Its formulation is based on the primitive relation of semantic equivalence, which partitions the narrative space into equivalence classes. The present framework is more general: it includes a wider space of multi-modal narratives and starts from the weaker, operationally available cognitive indistinguishability relation. Importantly, this relation is not assumed to be transitive. The distinction is consequential and central: semantic equivalence partitions the narrative space into equivalence classes, whereas cognitive indistinguishability defines a snake-skin-like covering of the space of narratives by overlapping neighborhoods. This is the structure required to explain phenomena including cognitive distance, incremental improvement, discovery, transcendence, and cognitive sorites chains.

2 Context and Cognitive Spaces

2.1 Context and Tolerance

The notion of context has various definitions in cognitive science, linguistics, and machine learning. In our framework, we use the functional understanding that has guided cognitive science: context encompasses the "supports that are needed for the actual going off of any act" (Tol32), together with the nature of the act itself. It includes components of the cognitive and environmental contexts defined by Balsam in (Bal14). The formal objects introduced below make this role precise. Analytically, context defines a perceptual indiscriminability relation and a cognitive indistinguishability relation. When the latter relation is reflexive and symmetric, the context is called a **tolerance context**, and cognitive indistinguishability is a tolerance relation that generates cognitive neighborhoods, topology, and cognitive distance. Operationally, tolerance contexts determine the phenomena that agents encounter and can be equipped with encounter and validity structures. These structures determine the lower encounter probability of phenomena, the standards by which narratives are judged valid, and the interventions by which invalid or irrelevant outputs may be corrected. The operational role is central: without context, neither encounter probability nor validity is a well-defined target for learning.

Definition 1. *Thus, in this paper, a* **context** *is a triple* $\mathfrak{C} = \{A, \mathbf{X}, T\}$ *where* A *is a class of subjects,* $\mathbf{X}$ *is an ensemble of narratives (or inputs), and* T *is a specific task.* [2]

[2] $\mathbf{X}$ includes texts, verbal narratives, multi-modal narratives, and of course sequences of prompts to a LLM.

In many contexts the subjects do assign a measures of validity to narratives and do have a notion of the likelihood of encountering a narrative. Accordingly, later in the paper we will introduce and discuss valuation contexts, pragmatic or operational contexts, and finally quantifiable generative contexts.

Definition 2. *Let* $\mathbf{X}$ *be the space of possible narratives. Two narratives* x *and* y *are called* **cognitively indistinguishable***, relative to a context* $\mathfrak{C}$*–consisting of a subject (cognitive agent) or a specified class of subjects, a space of narratives, and a task– if and only if the subject is not able (within* $\mathfrak{C}$*) to activate –acquire or employ–the relevant kind of knowledge that the meanings of* x *and* y *are distinct.*

This notion intentionally parallels Williamson's treatment of perceptual indiscriminability (Wil90), generalized from perceptual stimuli to narratives and multimodal generative artifacts.

Example 0: The narratives

$x_0 =$ "*Newton postulated that the gravitational force between two masses is proportional to the product of those masses and inversely proportional to the square of the distance between them.*"

$x_1 =$ "*Newton postulated that any two physical bodies exert a gravitational pull on one another that is proportional to their masses and inversely proportional to the square of the distance separating them.*"

$x_2 =$ "*Newton postulated that any two physical bodies exert a gravitational pull on one another that is proportional to their masses and inversely proportional to the square of the distance separating them (i.e., the distance between their centers of mass).*"

form a clique of pairwise cognitively indistinguishable narratives within a context $\mathfrak{C}_s$ where the subjects have college level training in physical and natural sciences and the task is casual conversation about the history of science.

On the other hand, the agents can deploy their knowledge that proving the validity of a law and just postulating/stating it mean different things and thus the narrative

$x_3 =$ "*Newton proved that the gravitational force between two masses is proportional to the product of those masses and inversely proportional to the square of the distance between them.*"

and the narrative x_i have different meanings for each $i = 0, 1, 2$.

Example 1: The narratives i_0 and i_1 shown in Figure 1 are cognitively indistinguishable in the Human Experience Archetypes Identification Context $\mathfrak{C}_{\text{H}} = \{\text{A}, \mathbf{X}, \text{T}\}$ of humans discussing the human experience. Here the class of subjects A is a group of mature college educated humans, the narrative space $\mathbf{X}$ is the ensemble of all multi-modal artifacts representing the human condition, and the task is to identify an experience archetype.

Furthermore, i_0, i_1, and the two word sentence $i_2 =$ *"Men age."* $\in \mathbf{X}$ form a clique of cognitively indistinguishable artifacts within this context.

Example 1 illustrates that cognitive indistinguishability does not imply representational or symbolic identity, but rather operational indistinguishability within a context.

The reflexivity and symmetry of perceptual indiscriminability are established in (Wil90). Similar to perceptual indiscriminability, cognitive indistinguishability is frequently assumed reflexive and symmetric and sometimes even transitive, (FKKG24). However, perceptual discrimination and cognitive distinguishability are fundamentally distinct operationally. The former is assessed via simultaneous presentation of stimuli, whereas narratives must be processed sequentially. As a result, judgments of cognitive indistinguishability require explicit control of order effects and context drift, typically through interpretive stabilization (e.g., rereading), which has no direct analogue in perceptual tasks.

Reflexivity and symmetry have different status in our framework. The latter can be operationalized and established in specific familiar contexts, while the former is postulated as a hypothesis which is strongly supported by lived experience and scientific practice.

i_0 i_1

Figure 1: Two narratives generated using Gemini 3 Flash Image (Nano Banana 2) image generator. The narratives i_0, i_1, and *"Men age."* $\in \mathbf{X}$ are pair-wise cognitively indistinguishable within the context $\mathfrak{C}_{\text{H}}$.

2.1.1 The Reflexivity Hypothesis

Our lived experience provides evidence for the validity of the following **functionalist-contextualist hypothesis**:

Cognitive indisguishability is a reflexive relation.

This axiom does not negate the observation that the meaning of certain narratives appears to change. The acquisition of new knowledge–and hence the change of context– often leads subjects to establish a new meaning. We readily recognize and accept this phenomenon in scientific narratives, whose interpretation evolves as background knowledge, theoretical commitments, and investigative goals change.[3] Accordingly, reinterpretation of meaning is most naturally modeled as a transition between contexts, rather than as a violation of reflexivity within a fixed context.

2.1.2 Symmetry

Cognitive indistinguishability is evaluated relative to a fixed context. Any experimental procedure involving interpreting and comparing meanings necessarily presents narratives sequentially. Procedures that induce reinterpretation through order effects may correspond to changes of context and are therefore inadmissible for assessing indistinguishability.

Cognitive indistinguishability is symmetric in contexts where admissible judgments are made only after interpretive stabilization—typically achieved through rereading—which removes order effects and ensures that distinguishability depends on narrative content rather than presentation order. Such contexts do exist and are familiar to scientists and classes of subjects that are trained basic narrative assessment and analysis–including multiple passes through each narrative and strategic recursive reacquisition of narrative fragments. Human interaction with generative AI provides new examples of symmetric cognitive indistinguishability.

Definition 3. *We will say that a context is* **a tolerance context** *if the cognitive indistinguishability is reflexive and symmetric. The cognitive indistinguishability relation of a tolerance context $\mathfrak{C}$ defines a* **context relevant–symmetric binary relation** $\overset{\mathfrak{C}}{\approx}$ *on* $\mathbf{X}$*:* $x \overset{\mathfrak{C}}{\approx} y$ *if and only if x and y are cognitively indistinguishable.*

In the rest of this paper we will assume that the context is a tolerance context.

Definition 4. *We will say that x and y are* **cognitive Doppelgängers** *if and only if $x \overset{\mathfrak{C}}{\approx} y$ and we will refer to the collection of all cognitive Doppelgängers $\mathfrak{d}_{\mathfrak{C}}(x) = \{y \in \mathbf{X}, s.t.\ y \overset{\mathfrak{C}}{\approx} x\}$ of a given narrative x as the* **cognitive neighborhood** *of x. In general the collection*

$$\mathfrak{G}_{\mathfrak{C}} = \{\mathfrak{d}_{\mathfrak{C}}(x),\ x \in \mathbf{X}\} \tag{1}$$

[3] As Fromm notes in *The Greatness and Limitations of Freud's Thought*, scientific theories are continually reinterpreted as historical and conceptual contexts change.

defines a snake skin like covering of $\mathbf{X}$ *of overlapping cognitive neighborhoods but not a partition unless* $\overset{\mathfrak{C}}{\approx}$ *is transitive. Let us denote by* $\sim_{\mathfrak{C}}$ *the transitive closure of* $\overset{\mathfrak{C}}{\approx}$*. Thus*

$$x \sim_{\mathfrak{C}} y \Leftrightarrow \exists \text{ a finite chain of narratives } x = x_0 \overset{\mathfrak{C}}{\approx} x_1 \overset{\mathfrak{C}}{\approx} \cdots \overset{\mathfrak{C}}{\approx} x_n = y \tag{2}$$

We will refer to the equivalence classes $[x]_{\sim_{\mathfrak{C}}}$ *defined by the equivalence relation* $\sim_{\mathfrak{C}}$ *as* **cognitive tiles** *and to its elements,* $z \sim_{\mathfrak{C}} x$*, as* **incremental derivatives** *of* x*.*

For every corpus $S \subset \mathbf{X}$ *of artifacts we define by*

$$[S]_{\sim_{\mathfrak{C}}} = \bigcup_{y \in S} [y]_{\sim_{\mathfrak{C}}} \tag{3}$$

the **set of incremental derivatives of the corpus** S *and define the* **zone of transcendence** *relative to the corpus* S *or the* **zone beyond the horizon**

$$Z(S \,|\, \mathfrak{C}) = \mathbf{X} \setminus [S]_{\sim_{\mathfrak{C}}} \tag{4}$$

In particular, the narratives x_0, x_1, and x_2 in Example 0 belong to the same cognitive tile $[x_0]_{\sim_{\mathfrak{C}_s}}$. However, there exist conversational contexts $\mathfrak{C}_{s*}$ such that x_0, x_1, and x_2 belong to the same cognitive tile $[x_0]_{\sim_{\mathfrak{C}_{s*}}}$ while x_3 belongs to a different cognitive tile, $[x_3]_{\sim_{\mathfrak{C}_{s*}}} \neq [x_0]_{\sim_{\mathfrak{C}_{s*}}}$ as discussed in Section 2.1.3.

2.1.3 Cognitive Topology to Cognitive Metric

The context-relevant **cognitive topology** $\tau_{\mathfrak{C}}$ is the topology generated by the sub-basis $\mathfrak{G}_{\mathfrak{C}} = \{\mathfrak{d}_{\mathfrak{C}}(x),\ x \in \mathbf{X}\}$– within the context $\mathfrak{C}$. In particular, directly from the definition it follows that the zone of transcendence $Z(S \,|\, \mathfrak{C})$ is topologically separated (lives in a different connected component) from S and the collection of **incremental derivatives** of S.

Just as perceptual systems collapse distinctions that are irrelevant to action (Sch58; Kam25), **cognitive–semantic systems collapse distinctions that are irrelevant to reasoning and discourse under a task**.Thus the cognitive topologies are rarely if ever metric – just like the perceptual topologies discussed in (Kam24; Kam25). Still, indistinguishability gives rise to a context-relevant cognitive distance $d_{\mathfrak{C}}$ that quantifies the semantic similarity between narratives.[4]

Indeed, the **cognitive distance** is an extended distance function $d_{\mathfrak{C}} : \mathbf{X} \times \mathbf{X} \to [0, +\infty]$ defined as the graph distance on a graph whose vertices correspond to the inputs $x \in \mathbf{X}$ and an edge $\{x, y\}$ between the vertices $x, y \in \mathbf{X}$ exists if and only if $x \overset{\mathfrak{C}}{\approx} y$. The resulting extended distance – also referred to as the **degrees of semantic separation** – is measure of similarity even when computationally convenient classical and novel deep learning networks-based measures are not. The distance between different cognitive tiles is infinite, while the distance between, narratives that belong to the same tile is finite, and so the cognitive tiling of $\mathbf{X}$ is referred to as an *infinitely separated cognitive tiling.* **The extended cognitive metric does not generate the cognitive topology, but it recovers cognitive indistinguishability as the minimal non-trivial unit of similarity,** $x \overset{\mathfrak{C}}{\approx} y \Leftrightarrow d_{\mathfrak{C}}(x, y) \leq 1$**. Accordingly, indistinguishability is not computed or detected by the cognitive agent; it is simply the absence of distinction.**

The similarity measure induced by the cognitive distance is constructed to align with the judgments of cognitive agents within a given context. Crucially, it does not register variability where cognition registers none.

A striking example is provided by the casual conversation context $\mathfrak{C}_s$ discussed in Example 0, where $d_{\mathfrak{C}_s}(x_0, x_1) = d_{\mathfrak{C}_s}(x_0, x_2) = d_{\mathfrak{C}_s}(x_1, x_2) = 1$ but **the embedding distances between these narratives** vary and **generate the illusion of semantic variability** while the cognitive agent perceives functional/semantic identity.

[4]Note the fundamental reversal of causality. Typically researchers assume that the topology of indistinguishability is generated by a metric, while it turns out that the topology does define a metric but it is not generated by it.

Conversely, the model-dependent embedding distance between x_3 and each x_i $(i = 0, 1, 2)$ remains small. Yet, in a rigorous context $\mathfrak{C}_{s*}$ the 'invalid' statement x_3 is infinitely far away cognitively from the 'valid'–and cognitively indistinguishable–statements x_i, $i = 0, 1, 2$, such that $d_{\mathfrak{C}_{s*}}(x_3, x_i) = \infty$, $i = 0, 1, 2$.[5]

The cognitive distance provides a measure of the **novelty** of a narrative x given a corpus of narratives $S \subset \mathbf{X}$, namely

$$\text{novelty}_{\mathfrak{C}}(x|S) = \min_{y \in S}(d_{\mathfrak{C}}(x, y)) \tag{5}$$

which is purely based on semantics rather than the usual approach exploiting information theory. Directly from the definitions of cognitive distance, derivatives, and zone of transcendence we obtain:

Proposition 1. *Let $S \subset \mathbf{X}$ be a corpus, then the set of incremental derivatives of S is the sub-ensemble of narratives with finite novelty with respect to S,* i.e.

$$[S]_{\sim_{\mathfrak{C}}} = \{y \in \mathbf{X},\ s.t.\ \textit{novelty}_{\mathfrak{C}}(y|S) < \infty\} \tag{6}$$

and the zone of transcendence with respect to S is the sub-ensemble of narratives with infinite novelty relative to a corpus S, meaning that they are cognitively infinitely far from S,

$$Z(S \,|\, \mathfrak{C}) = \{x \in \mathbf{X},\ s.t.\ \textit{novelty}_{\mathfrak{C}}(x|S) = \infty\} \tag{7}$$

2.2 Graded Validity

In most real-world contexts, agents are able to assign a validity value $B_{\mathfrak{C}}(x) \in \mathfrak{T}$ to every narrative $x \in \mathbf{X}$, where $\mathfrak{T}$ is equipped with a context-relative reflexive relation $\prec$. We will refer to such a context $\mathfrak{C} = \{\mathrm{A}, \mathbf{X}, \mathrm{T}, B_{\mathfrak{C}}\}$ where $\{\mathrm{A}, \mathbf{X}, \mathrm{T}\}$ is a tolerance context and the mapping $B_{\mathfrak{C}}$ is the validity within the context as a **valuation context**, or V-context. Cognitive agents often can—and must—make **only** pairwise validity comparisons. Accordingly, our framework does not require completeness, total ordering, transitivity, or latent linear structure on $\mathfrak{T}$, but does not preclude them when they arise in specific contexts.

Definition 5. *A V-context $\mathfrak{C}$ is called* **regular** *if*

$$(x \overset{\mathfrak{C}}{\approx} y) \Rightarrow (B_{\mathfrak{C}}(x) = B_{\mathfrak{C}}(y)). \tag{8}$$

By definition a context $\mathfrak{C}$ is regular if and only if the graded validity

$$B_{\mathfrak{C}} : (X, \overset{\mathfrak{C}}{\approx}) \to (\mathfrak{T}, =) \tag{9}$$

is a tolerance morphism as defined in (Sos86), or equivalently, if and only if the graded validity is a continuous mapping

$$B_{\mathfrak{C}} : (X, \tau_{\mathfrak{C}}) \to (\mathfrak{T}, \tau_d), \tag{10}$$

where τ_d is the discrete topology. More specifically, the graded validity mapping of a regular context $\mathfrak{C}$ are the 'step' functions f which are constant along each cognitive tile, $f|_{[x]_{\sim_{\mathfrak{C}}}} \equiv f(x), \forall x \in \mathbf{X}$.

Regular contexts include contexts in which the cognitive agents are 'value-guided', meaning that the graded validity of narratives is deployed by the agents to establish cognitive distinctions. These contexts arise in all professional domains and in many social domains from scientists discussing experiments and theories to societies establishing legal or moral frameworks.

In particular, most state of the art empirical evaluation contexts implicitly assume regularity. Yet, there are contexts that are not regular.

Definition 6. *A V-context $\mathfrak{C}$ is called* **cognitively ambivalent** *if the graded validity mapping $B_{\mathfrak{C}} : (X, \tau_{\mathfrak{C}}) \to (\mathfrak{T}, \tau_d)$ has discontinuities.*

[5]Notwithstanding Newton's claim "Hypotheses Non Fingo", his proof is based on accepting the first three rules of reasoning that he postulated in Book III of the Principia.

Our living experience provides abundant evidence of cognitively ambivalent contexts. Specifically, they appear to be ubiquitous in human-AI interaction contexts where graded validity is the probability to encounter a narrative (or a class of narratives). However, humans have long recognized that the likelihood of a narrative (or token) to exist is frequently decoupled from its cognitive meaning. Modern scaling paradigms implicitly treat encounter statistics as proxies for semantic structure and validity. Example 2 illustrates why this assumption is cognitively nontrivial.

Example 2: Consider the context of $\mathfrak{C}_b$ where the agents are bilingual (e.g., English and French) and have college level training in physical and natural sciences, the task is casual conversation about the history of science, and $B_{\mathfrak{C}_b}(x)$ is the probability to encounter the narrative x. Under the condition of stress or just being tired such agents sometimes suffer a **slip of the tongue** and generate a mixed language narrative. Within this context the narratives

$E =$ "*Newton postulated that the gravitational force between two masses is proportional to the product of those masses and inversely proportional to the square of the distance between them.*"

$M =$ "*Newton postulated that the force gravitationnelle between two masses is proportional to the product of those masses and inversely proportional to the carré de la distance between them.*"

are cognitively indistinguishable, $E \overset{\mathfrak{C}_b}{\approx} M$, but $B_{\mathfrak{C}_b}(E) \neq B_{\mathfrak{C}_b}(M)$.

A significant concern about generative AI is its tendency to hallucinate and given the high and seemingly non-diminishing frequency of hallucination encounters it is clear that they cannot be confused with rare/low encounter probability events. What makes them special is that they are *not valid*.

Definition 7. *Let $\mathfrak{C}$ be a context with graded validity mapping $B_{\mathfrak{C}} : \mathbf{X} \to (\mathfrak{T}, \prec)$, where $\mathfrak{T}$ has a unique global meet (minimal) element $\perp \in \mathfrak{T}$ ($\perp \prec t$, $\forall t \in \mathfrak{T}$). We will say that $x \in \mathbf{X}$ is a* **hallucination** *in the context $\mathfrak{C}$ if and only if*

$$B_{\mathfrak{C}}(x) = \perp \ . \tag{11}$$

Alternatively, narratives that are not hallucinations within $\mathfrak{C}$ are called **grounded** *within $\mathfrak{C}$*–i.e. *$x \in \mathbf{X}$ is grounded iff $B_{\mathfrak{C}}(x) \neq \perp$– and by extension we will call a corpus of narratives S* **grounded** *within $\mathfrak{C}$ if every $x \in S$ is grounded within $\mathfrak{C}$.*

If the context $\mathfrak{C}$ is regular and the corpus S is grounded within $\mathfrak{C}$, then the regularity of the context implies that the derivatives $x \in [S]_{\sim_{\mathfrak{C}}}$ cannot be hallucinations–$B_{\mathfrak{C}}(x) \neq \perp$ for every $x \in [S]_{\sim_{\mathfrak{C}}}$ and therefore:

Proposition 2. *If $\mathfrak{C}$ is a regular context and the corpus S is grounded within it, then every hallucination x is transcendent with respect to S*–i.e.*, every hallucination $x \in Z(S \,|\, \mathfrak{C})$.*

However, if the context is ambiguous we may have creeping incremental hallucinations. Moreover we may encounter truly cognitively disturbing, shall we say adversarial, hallucinations that are cognitively indistinguishable from artifacts in the corpus.

2.3 Operational Contexts

Active exploration, training, and manipulation contexts are typical instantiations of what we call **operational contexts** involving active, situated, and embodied agents who interact and act on the objects and phenomena they encounter.

Definition 8. *An operational context $\mathfrak{C} = \{A, \mathbf{X}, \mu, T\}$ is a tolerance context in which the space of narratives is a probability measure space $(\mathbf{X}, \mathfrak{F}, \mu)$. The measure μ is called the cognitive phenomenon* **encounter measure**.

ML's training/test/deployment/sampling distributions are near instantiations of encounter measure structures.

In contrast with the (atomic) probability distributions over the finite or at most countable infinite sets of strings used in machine learning the general encounter measure provides estimates of the likelihood to encounter **cognitive phenomena** defined as $\mathfrak{d}_{\mathfrak{C}}(x)$, $x \in \mathbf{X}$ rather than individual narratives.[6] Our extensive experience with idealized finite sample spaces has lead us to identify likelihood to encounter with a probability. However, in reality the phenomena $\mathfrak{d}_{\mathfrak{C}}(x)$ may be very irregular sets and therefore, it is too restrictive to expect that they are measurable. Instead, a cognitive phenomenon is encounterable if it contains a non-negligible event–an encounterable kernel.

Definition 9. *Let $\mathfrak{C} = \{A, \mathbf{X}, \mu, T\}$ be an operational context and let μ_* be the inner measure induced by the encounter measure. We define* **the phenomenon encounter capacity** *at $x \in \mathbf{X}$ as*

$$\mathrm{cap}_{\mathfrak{C}}(x) = \mu_*(\mathfrak{d}_{\mathfrak{C}}(x)) \tag{12}$$

The capacity $\mathrm{cap}_{\mathfrak{C}}(x)$ is the lower probability to encounter the phenomenon indexed by x. The narrative x is called a **cognitive singularity**, within $\mathfrak{C}$, if $\mathrm{cap}_{\mathfrak{C}}(x) = 0$.

Furthermore, let μ^* be the outer probability measure induced by the encounter measure. We define the **encounter profile function** as the Choquet capacity

$$f_{\mathfrak{C}}(t) = \mu^*(\{x \in \mathbf{X} \ : \ \mathrm{cap}_{\mathfrak{C}}(x) > t\}), \ \ 0 \leq t \leq 1. \tag{13}$$

The profile function is bounded, on the interval $[0, 1]$, $0 \leq f_{\mathfrak{C}}(t) \leq 1$. We will refer to the maximal profile value $f_{\mathfrak{C}}(0)$ as the **cognitive encounter exposure** of the context. We call the context **practically invisible**, to an agent, if $f_{\mathfrak{C}}(0) = 0$, in which case almost all x are cognitive singularities.

A natural question for an active agent exploring the space of narratives is *what to expect?*, namely, what is the expectation to encounter a cognitive phenomenon, or at least what is the average of the encounter profile function on the interval $[0, 1]$? In fact, this question has a clear computable answer.

Theorem 1.

1. *The average value $\overline{f}_{\mathfrak{C}}$ of the encounter profile $f_{\mathfrak{C}}(t)$ on the interval $[0, 1]$ is well defined and finite.*
2. *If the phenomenon encounter capacity function $\mathrm{cap}_{\mathfrak{C}}(x)$ is μ measurable, i.e., $\mathrm{cap}_{\mathfrak{C}}$ is a random variable on the probability measure space $(\mathbf{X}, \mathfrak{F}, \mu)$, then $\overline{f}_{\mathfrak{C}}$ equals the expected value of $\mathrm{cap}_{\mathfrak{C}}(x)$, with respect to the encounter measure μ*

$$\overline{f}_{\mathfrak{C}} = \int_{\mathbf{X}} \mathrm{cap}_{\mathfrak{C}}(x)\, d\mu. \tag{14}$$

Sketch of Proof

Indeed, $f_{\mathfrak{C}}(t)$ is bounded and non-increasing on $[0, 1]$ and therefore Lebesgue integrable on $[0, 1]$ and so

$$\overline{f}_{\mathfrak{C}} = \int_0^1 f_{\mathfrak{C}}(t)\, dt \tag{15}$$

is finite, which proves Part *1.* The proof of Part *2.* follows directly from Tonelli's theorem.

□

2.4 Generative Models

In this section we develop the generative component of the framework by introducing stochastic prompt structures, stochastic and static generative models, and the notions of discovery, creativity, hallucination, hallucination exposure, and cognitive stability.

[6] Note that the phenomena are indexed by the narratives $x \in \mathbf{X}$, and therefore may form continuum.

Definition 10. *A* **stochastic prompt structure** (Π, δ_π) *consists of a set of* **prompts** Π*, where* Π *is a collection of non-null events in some probability measure space* $(P, \mathfrak{F}_\pi, \mu_\pi)$*,*

$$\Pi \subset \mathfrak{F}_\pi, \qquad \mu_\pi(\omega) > 0, \quad \forall \omega \in \Pi, \tag{16}$$

together with a **narrative content** *map* $\delta_\pi : \Pi \to \mathbf{X}$ *where* $\mathbf{X}$ *is a set of narratives.*

The elements of Π are prompt events, not merely visible prompt strings. They encode narratives as well as hidden operational conditions, including time, system state, model configuration, decoding regime, and retrieval state. The map $\delta_\pi : \Pi \to \mathbf{X}$ records the narrative content of a prompt event; this is what is ordinarily called the prompt.

Definition 11. *Let* $(\mathbf{X}, \mathfrak{F}, \mu)$ *be the probability measure space of narratives of an operational context* $\mathfrak{C} = \{A, \mathbf{X}, \mu, T\}$ *and let* (Π, δ_π) *be a stochastic prompt structure with* $\delta_\pi : \Pi \to \mathbf{X}$*. A* **stochastic generative model** $(X, \mathbf{x})$ *consists of a probability measure-valued mapping*

$$X : \Pi \to \mathcal{P}(\mathbf{X}, \mathfrak{F}), \tag{17}$$

together with the **realization map**

$$\mathbf{x} : \Pi \to \mathbf{X} \tag{18}$$

such that the cognitive neighborhood of the realization of every prompt ω *has positive inner measure with respect to* $X(\omega)$*, i.e.,*

$$X(\omega)_*(\mathfrak{d}_{\mathfrak{C}}(\mathbf{x}(\omega))) > 0, \forall \omega \in \Pi, \tag{19}$$

where $X(\omega)_*$ *denotes the inner measure induced by* $X(\omega)$*. We will refer to* $\mathbf{x}(\omega)$ *as a* **generated artifact**

Discoveries: *If* $\mathfrak{C}$ *is an operational context with encounter measure* μ*, and* μ_* *denotes the associated inner measure, then every generated artifact* $\mathbf{x}(\omega)$*, with* $\omega \in \Pi$*, such that*

$$\mu_*\big(\mathfrak{d}_{\mathfrak{C}}(\mathbf{x}(\omega))\big) = 0 \tag{20}$$

is called a **cognitive discovery**.[7]

Creativity: *The model is* **creative** *given a corpus* $S \subset \mathbf{X}$ *within the operational V-context* $\mathfrak{C}$ *if for some prompt* ω *the inner measure of the set*

$$S_+(\mathfrak{C}) = \{x \in \mathbf{X},\ s.t.\ B_{\mathfrak{C}}(s) \prec B_{\mathfrak{C}}(x), \forall s \in S\}$$

with respect to the measure $X(\omega)$ *is positive, i.e., there exists* $\omega \in \Pi$ *such that*

$$X(\omega)_*(S_+(\mathfrak{C})) > 0.$$

The model $(X, \mathbf{x})$ *is* **transformative**, *relative to the corpus* S*, if there exists a prompt* ω *generating a transcendent artifact* $\mathbf{x}(\omega) \in Z(S \,|\, \mathfrak{C})$*, i.e.,* $\mathbf{x}(\omega)$ *is within the zone of transcendence relative to* S*. A model that is creative in a given context but is not transformative is called* **incrementally creative**.

Hallucinations: *Let* $\mathfrak{C}$ *be an operational V-context and let* $(X, \mathbf{x})$ *be a stochastic generative model.*

1. *The* **model hallucinates** *within the operational V-context* $\mathfrak{C}$ *if there exists a prompt event such that* $\mathbf{x}(\omega)$ *is a hallucination within* $\mathfrak{C}$*, i.e.,*

$$B_{\mathfrak{C}}(\mathbf{x}(\omega)) = \perp .$$

2. *We say that the model* **exposes hallucinations** *within the operational V-context* $\mathfrak{C}$ *if there exists* $\omega \in \Pi$ *such that the inner measure of*

$$\mathbf{X}_\perp(\mathfrak{C}) = \{x \in \mathbf{X},\ s.t.\ , B_{\mathfrak{C}}(x) = \perp\}$$

with respect to the measure $X(\omega)$ *is positive, i.e.,*

$$X(\omega)_*(\mathbf{X}_\perp(\mathfrak{C})) > 0.$$

[7] Thus cognitive discoveries are the generated artifacts which index cognitive singularities within the operational context.

Transformer-based language models instantiate the stochastic generative models of Definition 11 in the special case where generated narratives are represented by finite output token strings.

Example 3:

Let Σ^* denote a space of finite token strings, equipped with a measurable structure $\mathfrak{S}$, and set

$$\mathbf{X} = \Sigma^*.$$

For a fixed prompt ω, the narrative content $\delta_\pi(\omega)$ is tokenized and processed by the transformer. The transformer logit map, together with a decoding regime D, induces a probability measure

$$Q_{\theta,D}(\omega) \in \mathcal{P}(\Sigma^*, \mathfrak{S})$$

on finite output strings, where θ denotes the parameters of the language model. In this token-string specialization, the prompt-conditioned generated encounter measure is

$$X(\omega) = Q_{\theta,D}(\omega).$$

The realization map records the particular decoded output string produced in a run: if $s(\omega) \in \Sigma^*$ is the realized decoded string, then

$$\mathbf{x}(\omega) = s(\omega).$$

More generally, output strings may be interpreted as narratives in a richer narrative space $(\mathbf{X}, \mathfrak{F})$. Let

$$\rho : \Sigma^* \to \mathbf{X}$$

be an interpretation map. In this case, the generated encounter measure on narratives is the pushforward

$$X(\omega) = \rho_* Q_{\theta,D}(\omega),$$

and the realized generated artifact is

$$\mathbf{x}(\omega) = \rho(s(\omega)).$$

Definition 12. *Let $(X, \mathbf{x})$ be a stochastic generative model operating within an operational context $\mathfrak{C} = \{A, \mathbf{X}, \mu, T\}$ and let (Π, δ_π) be the stochastic prompt structure with $\delta_\pi : \Pi \to \mathbf{X}$. We say that the model $(X, \mathbf{x})$ is* **cognitively stable** *if*

$$(\delta_\pi(\omega_1) \overset{\mathfrak{C}}{\approx} \delta_\pi(\omega_2)) \implies (\mathbf{x}(\omega_1) \overset{\mathfrak{C}}{\approx} \mathbf{x}(\omega_2)). \tag{21}$$

Cognitive stability generalizes the stability sought in modern LLM practice: the relevant invariant is not syntactic identity, but cognitive indistinguishability within a context. It also separates two routes to stable LLM behavior: prompt-side stabilization, which restricts or normalizes admissible prompt events, and realization-side stabilization, which modifies the realization map. The former formalizes prompt engineering, while the latter changes the effective generative model.

Since a prompt ω is an operational event, not merely its narrative content $\delta_\pi(\omega)$, prompt-side stabilization need not be limited to rewriting or eliminating visible prompt strings. This suggests a broader approach to prompt engineering.

For completeness we conclude this section with a discussion of static models. They often provide simpler examples of key notions and processes.

Definition 13. *Let $\mathfrak{C}$ be an operational V-context and Π be the collection of prompts. A* **static generative model** *is a mapping $\mathbf{x} : \Pi \to \mathbf{X}$. Let $\mathfrak{C}$ be a context with graded validity mapping $B_\mathfrak{C}$*

Creativity: *The model is* **creative** *given a corpus $S \subset \mathbf{X}$ within the V-context, $\mathfrak{C}$ if it generates an artifact $\mathbf{x}(\omega)$ such that $B_\mathfrak{C}(s) \prec B_\mathfrak{C}(\mathbf{x}(\omega))$ for every $s \in S$. The model is* **transformative** *if it generates transcendent artifacts $\mathbf{x}(\omega)$ such that $B_\mathfrak{C}(s) \prec B_\mathfrak{C}(\mathbf{x}(\omega))$ for every $s \in S$. A model that is creative in a given context but is not transformative is called* **incrementally creative**.

Hallucinations: *The model* $\mathbf{x}$ *hallucinates, in the context,* $\mathfrak{C}$ *if the model generates a hallucination* $\mathbf{x}(\omega)$*, s.t.,* $B_{\mathfrak{C}}(\mathbf{x}(\omega)) = \perp$.

Discoveries: *If* $\mathfrak{C}$ *is an operational context with encounter measure* μ*, and* μ_* *denotes the associated inner measure, then every generated artifact* $\mathbf{x}(\omega)$*, with* $\omega \in \Pi$*, such that*

$$\mu_*\big(\mathfrak{d}_{\mathfrak{C}}(\mathbf{x}(\omega))\big) = 0 \tag{22}$$

is called a **cognitive discovery**.

Note that in the static case a stochastic structure on the space of prompts is not necessary.

Our framework allows for the contingency that a model may be transformative but not creative and vice versa being creative does not imply that the model is transformative.

2.5 Hallucinations vs. Irrelevancy

Hallucinations are extreme cases of artifacts produced by generative AI that *do not fit the prompt*. Yet, a failure to fit the prompt does not inherently mean hallucination. For example, consider two static generative models X and Y used in the context $\mathfrak{C}_b$, Example 2 which produce different responses to the prompt ω whose narrative content is

$$\delta_\pi(\omega) = \text{“}\textit{Describe the origins of Newton's theory of gravity.}\text{”} \tag{23}$$

Namely,

$X(\omega) =$ “*Newton postulated that the gravitational force between two masses is proportional to the product of those masses and inversely proportional to the square of the distance between them.*”

$Y(\omega) =$ “*Einstein formulated the General Theory of Relativity. In it he defined the field equations.*”

While both $B_{\mathfrak{C}_b}(X(\omega)) \neq \perp$ and $B_{\mathfrak{C}_b}(Y(\omega)) \neq \perp$, the artifact $X(\omega)$ fits the prompt better than $Y(\omega)$. In generative contexts one is commonly aware of this measure of a fit or relevance of generated outputs to the prompt.

Definition 14. *A V-context* $\mathfrak{C}$ *with a graded validity mapping* $B_{\mathfrak{C}} : \mathbf{X} \to (\mathfrak{T}, \prec)$ *is called* **a quantifiable generative context** *if for every prompt event* $\omega \in \Pi$ *there exists a validity mapping, the* **prompt-conditioned validity**,

$$B_{\mathfrak{C}}(\cdot \,|\, \omega) : \mathbf{X} \to (\mathfrak{T}, \prec),$$

such that

$$(B_{\mathfrak{C}}(x) = \perp) \Rightarrow (B_{\mathfrak{C}}(x \,|\, \omega) = \perp). \tag{24}$$

We will call the narrative $x \in \mathbf{X}$ **unrelated** *to the prompt* $\omega \in \Pi$ *in the context* $\mathfrak{C}$ *if* $B_{\mathfrak{C}}(x \,|\, \omega) = \perp$ *and will denote by*

$$\mathbf{X}_\perp(\mathfrak{C}, \omega) = \{x \in \mathbf{X},\ \textit{s.t.}\ B_{\mathfrak{C}}(x \,|\, \omega) = \perp\}$$

the **set of all narratives unrelated to the prompt event** ω **in the context** $\mathfrak{C}$. *Clearly, hallucinations are unrelated to all possible prompts in a given context* $\mathfrak{C}$*. On the other hand, there may exist* **valid**, $B_{\mathfrak{C}}(x) \neq \perp$, *narratives that are unrelated to some prompt* ω*– (VBY26) calls them contextual hallucinations. However, they are* **cognitively valid generative errors** *and by no means hallucinations neither in the traditional cognitive science and psychology sense, nor in everyday lay parlance; they are just* **irrelevant within the context non-hallucinatory experiences**.

2.6 The Picasso Conundrum

The perils of hallucinating in public and professional life are well known and often cited, (Art25; Joh25; KNVZ25). And so we work very, very hard to prevent our models from hallucinating. Under the common, but unstated, assumption that the context is regular this leads to building guardrails to ensure that the models are incrementally creative but not transformative. The research community has been particularly successful in guard-railing image generation models. The guardrails have kept many hallucinations out together with transcendent artifacts and in effect suppressed creativity and the discovery of new styles and valid narratives.

Example 4: The Three Ages of Man is an important Renaissance Allegory representing human's life cycle. It has been depicted in multiple paintings since at least the 1500s. Titian's painting is shown in Figure 2 Combining planar representations from different view points taken from the different ages one can obtain a

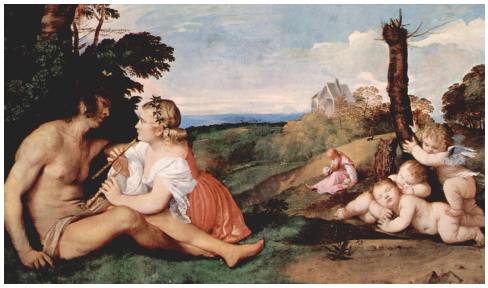

Figure 2: Titian's The Three Ages of Man, 1512-1514, National Gallery of Scotland, Edinburgh, (Com26)

transformative painting very much in the style of the Analytical Cubism developed by Pablo Picasso and Georges Braque (Kah20). Indeed, after a series of prompts ChatGPT 5.2 generates the following narrative

$c_0 =$ "*Analytical Cubist head, early Picasso/Braque period (1909–1911). One single male head with a continuous skull outline. Exactly two eyes, one nose, one mouth. No duplicated faces, no profiles, no secondary heads, no triptych. The face is fractured into geometric planes. One cheek plane is rounded and soft, suggesting childhood; the central brow and eye planes are firm and alert, suggesting adulthood; the opposite cheek and jaw planes are hollowed and angular, suggesting old age. Hair is inconsistent across planes: one fuzzy childlike texture, one dense adult texture, one receding or bald plane, all within the same scalp. No complete face appears anywhere except the whole. Age is encoded through plane texture and geometry only, not separate identities. Muted cubist palette, flat planes, suppressed realism.*"

However, current image generators including OpenAI's ChatGPT 5.2 image generator and the Gemini 2.5 Flash Image cannot generate–render–such narrative even when explicitly prompted using c_0. These image models enforce facial symmetry, coherent/non-disjoint anatomies, structural integrity, cohesive identity, and optimize *aesthetic scores*. In effect they operate within an **image generative context** $\mathfrak{C}_{\text{IG}}$ in which the narrative c_0 is not valid–it is a hallucination, a cognitively disturbing monster. Within the context $\mathfrak{C}_{\text{IG}}$ the images in Figure 3 and Figure 4 are at finite cognitive distances from Titian's painting.

2.7 Sobering and Improving Relevance

Hallucinations exist in many contexts. In each of these contexts the nonempty set of hallucinations $\mathbf{X}_{\perp}(\mathfrak{C})$ is model- and prompt- independent. Models that hallucinate are the ones that wander into $\mathbf{X}_{\perp}(\mathfrak{C})$. There are others that do not. Humans hallucinate when they are impaired and the hallucinations can be 'cured away' through some (active) strategy. Accordingly, the ML research community and industry are actively pursuing massive campaigns to train'sober' generative AI models–models that do not hallucinate. Presently, state-of-the-art models continue to hallucinate, often quite frequently. Alternatively, one can remove hallucinations post-training, in what appears to be an ad-hoc system architecture that does not hallucinate.

Figure 3: The Three Ages of Man narratives generated by OpenAI's ChatGPT 5.2 image generator.

Figure 4: The Three Ages of Man narratives generated by Gemini 2.5 Flash Image.

The **sobering architecture** in the static (temperature = 0) case is outlined below. It removes hallucinations. As long as $\mathbf{X} \neq \mathbf{X}_{\perp}(\mathfrak{C})$ we can modify the model post training and eliminate all hallucinations. In fact let

$$\Pi_{\perp}(X \,|\, \mathfrak{C}) = \{\omega \in \Pi, \text{ s.t. } B_{\mathfrak{C}}(X(\omega)) = \perp\}$$

be the collection of prompts that force X to generate hallucinations. Then, for every map

$$a(X, \mathfrak{C}) : \Pi_{\perp}(X \,|\, \mathfrak{C}) \to \mathbf{X} \setminus \mathbf{X}_{\perp}(\mathfrak{C}) \tag{25}$$

we can modify the model post training so that

$$X_{a(X,\mathfrak{C})}(\omega) = \begin{cases} a(X, \mathfrak{C})(\omega), & \text{if } \omega \in \Pi_{\perp}(X \,|\, \mathfrak{C}) \\ X(\omega), & \text{otherwise.} \end{cases} \tag{26}$$

By construction these post trained models do not hallucinate. Furthermore, they are *more relevant* than the model X since by construction

$$B_{\mathfrak{C}}(X(\omega) \,|\, \omega) \prec B_{\mathfrak{C}}(X_{a(X,\mathfrak{C})}(\omega) \,|\, \omega), \forall \omega \in \Pi.$$

Note that this process eliminates hallucinations and enlarges 'the support' or the range of the model.

While hallucinations have long been focus of attention a much more prevailing and equally if not more disturbing and possibly dangerous phenomenon is the tendency of models to make valid generative errors. Still even in this case, $B_{\mathfrak{C}}(x \,|\, \omega) = \perp$, one can steer the model towards a relevant generative output

$$x(X, \mathfrak{C} \,|\, \omega) \in \mathbf{X} \setminus \mathbf{X}_{\perp}(\mathfrak{C}, \omega).$$

The active system architecture

$$X_{\mathrm{ad}(X,\mathfrak{C})}(\omega) = \begin{cases} x(X, \mathfrak{C} \,|\, \omega), & \text{if } B_{\mathfrak{C}}(x \,|\, \omega) = \perp \\ X(\omega), & \text{otherwise.} \end{cases} \tag{27}$$

does not hallucinate and does not make cognitively valid generative errors.

Removing cognitively valid generative errors, $X_{\mathrm{ad}(X,\mathfrak{C})}$, requires more resources at the inference stage than simply sobering a model, $X_{a(X,\mathfrak{C})}$, since sampling $\mathbf{X} \setminus \mathbf{X}_{\perp}(\mathfrak{C})$ can be computed "off-line" before inference.

The tremendous efforts in removing hallucinations and errors indicate that generative AI may not and in fact should not be trained and left alone to carry on with the sobering and correcting itself. The system architectures described above benefit and require the joint contribution and collaboration of LLMs and cognitive agents. For example, the cognitive agents are very efficient in detecting both hallucinations and cognitively valid generative errors. They can guide and benefit from the support of agenting AI to do the sobering $a(X, \mathfrak{C})$ and the corrected narrative $x(X, \mathfrak{C} \,|\, \omega)$.

3 Discussion

Hallucinations are often characterized as statistical outliers or low probability events and are typically blamed on noise in the training data. However, in reality, hallucinations are not necessarily rare and may in fact be highly valuable. A visceral example is the "falling nightmare" frequently experienced by children. These vestigial "hallucinations" have deep origins in the evolution of human cognition and, under major theories and hypotheses–including the Arboreal Hypothesis, (Jon16), the Visual Predation Hypothesis, (Car74), and the Threat Simulation Theory, (VR09)–are understood to be functional. These hallucinations aid in children's development of the ability to sleep without interruption.

Furthermore, the falling hallucination is not caused by "noise in the training data". Our ancestors truly did fall from the branches they inhabited for safety and therefore needed to wake up, at least partially, to ensure that they are safe from attack or from slipping off the branch. The data were accurate and valid; what has changed is the context. While the cited evolutionary biology and cognition theories are often contradictory or sometimes complementary (Sus91) they all point towards the functionality of the falling hallucination.

The emergence of large foundational models and the rapidly evolving workflows built around them create a unique opportunity to study and exploit generative cognition. For the first time in history there is a widespread access to systems capable of producing vast numbers of narratives that extend beyond their training corpora. Similar to the classical Pythia, they can produce remarkable insights but also misleading or cognitively disturbing artifacts. Working with such systems therefore requires both creativity and careful analytical discipline. The framework developed in this paper aims to provide precisely such a principled approach to exploratory generative cognition. In particular, it makes it possible to discern the cognitive status of generated artifacts: whether they are discoveries, derivative, or transcendent, and whether they are hallucinations or creative artifacts. More importantly, the framework is aimed at operationalizing what we call **purpose-driven contextual restructuring**. Namely, discovering and establishing transformed contexts thereby grounding generated narratives that were hallucinations within a previous context, and that solve real problems once the context is transformed. The ML community has focused on identifying and eliminating hallucinations, but purpose-driven contextual restructuring has not received comparable attention. And yet, this is how human progress often works.

For example, financial crises provide operational examples of purpose-driven contextual restructuring. Contemporary generative models are capable of producing operational narratives structurally resembling historical

financial stabilization mechanisms. In several major historical financial crises operational narratives involving emergency lending facilities, special purpose vehicles, sovereign guarantees, and central-bank acquisition or revaluation of distressed assets were formulated before stable legal and institutional contexts existed in which they were valid. Faced with systemic failure, policymakers, bankers, and governments subsequently restructured the operational and regulatory context, thereby grounding narratives that had previously been considered invalid or impossible. The resulting mechanisms subsequently played a central role in financial stabilization. Such purpose-driven contextual restructuring has appeared historically across multiple central banking regimes, including those involving the Federal Reserve and the Banque de France.

References

[Art25] Ralph Artigliere, 2025. AI hallucinations in court: A wake-up call. JD Supra.

[B+24] Abhimanyu Banerjee et al. Llms will always hallucinate, and we need to live with this. *Journal of AI Theory*, 2024.

[Bal14] Peter D Balsam. The functions of context in learning and performance. In *Context and learning*, pages 1–21. Psychology Press, 2014.

[BDW25] Rajarshi Biswas, Sourav Dutta, and Dirk Werth. Hallucinations in llms and resolving them: A holistic approach. In *International Conference on Agents and Artificial Intelligence*, 2025.

[Car74] Matt Cartmill. Rethinking primate origins: The characteristic primate traits cannot be explained simply as adaptations to arboreal life. *Science*, 184(4135):436–443, 1974.

[Com26] Wikimedia Commons. File:tizian 002.jpg — wikimedia commons, the free media repository, 2026.

[Fil20] Katja Filippova. Controlled hallucinations: Learning to generate faithfully from noisy data. *arXiv preprint arXiv:2010.05873*, 2020.

[FKKG24] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

[JLF+23] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38, 2023.

[Joh25] Johnson v. Dunn, 2025. No. 23-CV-1234, 555 F. Supp. 3d 100 (S.D.N.Y).

[Jon16] Frederic Wood Jones. *Arboreal man*. E. Arnold, 1916.

[Kah20] Daniel-Henry Kahnweiler. *Der Weg zum Kubismus*. Delphin-Verlag, 1920. English translation: The Rise of Cubism, Wittenborn, 1949.

[Kam24] George I. Kamberov. Golyadkin's torment: Doppelgängers and adversarial vulnerability, 2024.

[Kam25] George Kamberov. Doppelgangers and adversarial vulnerability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10244–10254, June 2025.

[KNVZ25] Adam Tauman Kalai, Ofir Nachum, Santosh S Vempala, and Edwin Zhang. Why language models hallucinate. *arXiv preprint arXiv:2509.04664*, 2025.

[MNBM20] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*, 2020.

[PWG+20] Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. Totto: A controlled table-to-text generation dataset. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 1173–1186, 2020.

[Sch58] Erwin Schrödinger. *Mind and Matter*. Cambridge University Press, UK, 1958.

[Sos86] Alexei B Sossinsky. Tolerance space theory and some applications. *Acta Applicandae Mathematica*, 5:137–167, 1986.

[Sus91] Robert W Sussman. Primate origins and the evolution of angiosperms. *American Journal of Primatology*, 23(4):209–223, 1991.

[Tol32] Edward Chace Tolman. *Purposive behavior in animals and men*. Univ of California Press, 1932.

[VBY26] Bodla Krishna Vamshi, Rohan Bhatnagar, and Haizhao Yang. Manifold-based sampling for in-context hallucination detection in large language models. *arXiv preprint arXiv:2601.06196*, 2026.

[VR09] Katja Valli and Antti Revonsuo. The threat simulation theory in light of recent empirical evidence: a review. *The American journal of psychology*, 122(1):17–38, 2009.

[Wil90] Timothy Williamson. *Identity and Discrimination*. Wiley-Blackwell, 1990.

[XJK24] Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*, 2024.