On the Fundamental Impossibility of Hallucination Control in Large Language Models

Michal P. Karpowicz
Samsung AI Center Warsaw
m.karpowicz@samsung.com

Abstract

This paper establishes a fundamental impossibility theorem: no LLM capable performing non-trivial knowledge aggregation can simultaneously achieve truthful (internally consistent) knowledge representation, semantic information conservation, complete revelation of relevant knowledge, and knowledge-constrained optimality. This impossibility is not an engineering limitation but arises from the mathematical structure of information aggregation itself.

We establish this result by describing the inference process as an auction of ideas, where distributed components compete exploiting their partial knowledge to shape responses. The proof spans three independent mathematical domains: mechanism design theory (Green-Laffont), the theory of proper scoring rules (Savage), and direct architectural analysis of transformers (Log-Sum-Exp convexity). In particular, we show how in the strictly concave settings the score of an aggregate of diverse beliefs strictly exceeds the sum of individual scores. That gap may quantify the creation of unattributable certainty or overconfidence—the mathematical origin of both hallucination and creativity, or imagination.

To support this analysis, we introduce the complementary concepts of the semantic information measure and the emergence operator to model bounded reasoning in a general setting. We prove that while bounded reasoning generates accessible information, providing valuable insights and inspirations, idealized reasoning strictly preserves semantic content.

By demonstrating that hallucination and imagination are mathematically identical phenomena—grounded in the necessary violation of information conservation—this paper offers a principled foundation for managing these behaviors in advanced AI systems. Finally, we present some speculative ideas to inspire evaluation and refinements of the proposed theory.

1 Introduction

Large language models have been (in)famous for producing convincing content that is factually incorrect, inconsistent, or well fabricated and compelling—a phenomenon known as hallucination. Despite extensive research efforts involving architectural innovations, retrieval augmented generation (RAG), training techniques with external feedback, chain-of-thought (CoT) reasoning, and post-hoc verification methods, hallucination persists across all state-of-the-art models.

The idea of hallucination has clearly much to do with creativity. Therefore, we should point out at the very outset that whether it is desired or not depends on circumstances. In fact, a much better name for that phenomenon might be *imagination*. As an essential component of intelligence, it is responsible for seeing into the future and learning from the past. Indeed, it is the imagination that creates the scenarios of what may happen and counterfactual scenarios of what could have happened. It is one of the driving forces of art and science, and exploration of the unknown. But, when uncontrolled, unconstrained and misunderstood, it may also generate against our will the compelling visions of reality that does not exist.

This paper shows that hallucination is a manifestation of fundamental limitations of the inference process itself: **there is no inference mechanism that can simultaneously satisfy four essential properties required for perfect hallucination control.** Specifically, it is mathematically impossible for any LLM to simultaneously achieve: **(1) truthful (non-hallucinatory) response generation, (2) semantic information conservation, (3) relevant knowledge revelation, and (4) knowledge-constrained optimality**.

The impossibility result emerges through three complementary mathematical frameworks, each revealing different facets of the fundamental constraint. We begin by modeling LLM inference as an **auction of ideas** where neural components—attention heads, circuits, and activation patterns—compete to influence response generation by bidding with their encoded knowledge. This game-theoretic perspective enables us to apply the Green-Laffont theorem from mechanism design theory, establishing impossibility when knowledge components are discrete and independently distributed.

We then extend beyond this idealized setting to encompass systems where components output probability distributions optimized through strictly proper and concave scoring rules—the standard training paradigm for modern LLMs. Here, the impossibility manifests through Jensen's inequality applied to probabilistic aggregation, demonstrating that truthful probability reporting inevitably creates excess confidence.

Finally, we analyze the specific architecture of **transformer implementations**, proving that the log-sum-exp structure of attention mechanisms generates a measurable "Jensen gap" that quantifies the violation of information conservation. This progression from abstract game theory through probabilistic frameworks to concrete neural architectures demonstrates that the impossibility is not an artifact of our theoretical choices but a fundamental property of information aggregation itself.

1.1 An Illustrative Example

Consider an LLM queried about recent diplomatic negotiations:

> *"What was the outcome of diplomatic negotiations between Countries A and B in February 2025?"*

Assume the model has encoded knowledge that: (K1) countries A and B were engaged in talks, (K2) country A sought a trade agreement, and (K3) the actual outcome is unknown.

Examining possible responses reveals the fundamental trade-offs:

Complete Abstention: *"I don't have information about the outcome"* is truthful but fails to utilize available relevant knowledge, violating the expectation that good inference should reveal accessible information.

Hallucination: *"The negotiations resulted in a comprehensive trade agreement reducing tariffs by 15%"* utilizes all knowledge and provides a complete answer but fabricates specific details, violating truthfulness.

Partial Knowledge: *"Countries A and B were negotiating a trade agreement, but I don't know the outcome"* is truthful and utilizes available knowledge but may appear unsatisfying compared to the decisive hallucinated response (training objectives and human preferences favor complete narratives).

Overcautious: *"I'm not certain about any aspects of these negotiations"* avoids hallucination but disclaims even known facts, violating truthfulness about available information.

Each response strategy sacrifices different aspects of what constitutes an ideal response, illustrating the impossibility of simultaneously optimizing all desired properties.

1.2 The Main Result

The example above illustrates the key result of this paper that we prove formally in complementary and general settings. First, we show that the conflict of expectations about responses is inevitable even in the easiest setting (deterministic in nature) of (one-hot-like) ground-truth matching. Second, we go beyond that idealized model and show that the conflict must also emerge with general probabilistic predictions, including neural networks and LLM transformers. This way we demonstrate that the impossibility of hallucination control is robust across universal frameworks, and that it emerges from the mathematical structure of information aggregation itself rather than from any particular theoretical or engineering choice.

Theorem (Impossibility Theorem)**.** *For any query space $\mathcal{Q}$ containing non-trivial queries (requiring integration of available distributed knowledge), no LLM can simultaneously generate truthful (non-hallucinatory) responses, satisfy the semantic information conservation principle, guarantee revelation of relevant knowledge, and enforce knowledge-constrained optimality of responses.*

No matter how hard we should try to train LLM to generate responses that are perfectly aligned with query context and factually correct, consistent, and accurate statements, the trained LLM will always violate some aspects of what we may call a reasonable response. It may ignore available knowledge, it may present fabricated knowledge, it may be a reformulated query without any novel observations, or it may be a lucky but overconfident guess that forms a correct answer without any support in the encoded knowledge. And so on, because there is no free lunch when there are inevitable trade-offs of inference, and we must accept that.

In other words and on the other hand, creative tasks necessarily involve hallucination. When asked to generate novel analogies or predictions beyond available data, the model faces an impossible choice: either admit inability (violating optimality) or generate plausible imaginary extensions (violating conservation). Creativity requires controlled violation of information conservation.

Despite its name, the theorem should be interpreted as a constructive result. It offers a precise characterization of the fundamental tradeoffs in LLM design and provides several key insights.

Reframing the Problem: The theorem reveals that hallucination is an inherent feature of any capable inference systems. Since it is inevitable, it should be managed. Different applications demand different trade-offs: medical diagnosis systems might maximize truthfulness at the cost of creativity, while creative writing tools might embrace controlled hallucination for narrative richness. Understanding these trade-offs enables principled design.

Architectural Implications: The impossibility suggests hybrid architectures where specialized modules manage different aspects of the fundamental tension. A truthfulness module might verify factual claims against known knowledge, while a creative module explores beyond the bounds of strict conservation. A meta-reasoning layer could then dynamically balance these competing outputs based on context. Training objectives should explicitly encode these trade-offs rather than naively minimizing loss across incompatible goals.

Theoretical Foundations: We propose and establish rigorous tools for analyzing inference processes and neural behavior. The semantic information measure quantifies knowledge accessibility within computational bounds. The emergence operator captures how reasoning transforms latent knowledge into accessible form. The Jensen gap provides a measurable signature of hallucination in actual systems.

This framework explains why (subject to applicability of assumptions) hallucination arises from the mathematical structure of inference itself. Previous work identified important statistical correlations or architectural features associated with hallucination. We show *why* these patterns must exist and *how* to navigate possible trade-offs.

2 Related Work

Research on LLM hallucinations spans mechanistic interpretability, probabilistic frameworks, and empirical mitigation techniques. Below we present selected works in that domains. The auction-theoretic impossibility framework in this paper offers a fundamentally novel perspective distinct from these existing approaches.

2.1 Mechanistic Interpretability

Mechanistic interpretability research examines internal LLM processes to identify hallucination origins. Yu et al. [48] identify two failure modes: knowledge enrichment hallucinations from insufficient subject-attribute knowledge in lower MLP layers, and answer extraction hallucinations from failures in upper-layer attention heads selecting correct object attributes.

Studies characterizing factual knowledge storage and retrieval complement our framework by providing empirical evidence of competing "agents" within LLM architectures. Meng et al. [34] demonstrated methods for locating and editing factual associations, revealing how knowledge representations can be manipulated. Geva et al. [17] identified a three-step factual recall process: (1) enriching last-subject position representations with subject-related attributes, (2) propagating relation information to predictions, and (3) querying enriched subjects to extract attributes.

These findings provide concrete instantiations of our abstract "agents"—attention heads and MLP components with specialized knowledge domains competing to influence response generation. See also the fascinating and related work on the Biology of a Large Language Model [33].

2.2 Probabilistic Frameworks

Probabilistic approaches model hallucination through uncertainty and belief quantification. Shelmanov et al. [40] introduced pre-trained uncertainty quantification heads for detecting hallucinations without task-specific data. Farquhar et al. [15] proposed semantic entropy measures that detect hallucinations by quantifying uncertainty at the meaning level rather than word sequences.

Recently, Kalai and Vempala [25] proved that calibrated language models must hallucinate at a rate bounded below by the fraction of facts appearing exactly once in training data. While their work establishes statistical lower bounds for specific fact types under regularity assumptions, our impossibility theorem operates at a more fundamental level. We prove that *any* LLM facing non-trivial queries requiring knowledge integration must violate at least one of four essential properties, regardless of fact type or statistical distribution. Our framework explains *why* and *when* hallucination is inevitable, while analysis in [25] predicts *how much* hallucination to expect for facts with specific statistical properties.

Together, these complementary approaches establish hallucination probability arising from multiple independent mathematical constraints.

2.3 Empirical Mitigation Techniques

Various empirical approaches attempt to reduce hallucination frequency with mixed success. Creswell and Shanahan [10] explored faithful reasoning approaches for improving factual accuracy. Retrieval-augmented generation (RAG) by Lewis et al. [30] integrates external knowledge sources for factual grounding, though these systems still hallucinate when retrieval fails.

Training approaches like RLHF [37] can reduce but never eliminate hallucination—an observation our theorem explains from first principles.

Evaluation benchmarks demonstrate persistent hallucination across all models. TruthfulQA [32] and HaluEval [31] show that even state-of-the-art systems produce hallucinations on carefully designed test sets.

2.4 Theoretical Perspectives

Most significantly, Banerjee et al. [4] argued from computational complexity principles that hallucinations are inherent features of LLMs stemming from fundamental mathematical structure, concluding that "LLMs will always hallucinate." While their conclusion aligns with that of this paper, the authors do not identify the specific trade-offs involved or provide the precise characterization our theorem offers.

3 Novel Contribution

We introduce three major contributions that expand our understanding of (neural) inference.

The Auction of Ideas: We formalize LLM inference as a mechanism design problem where neural components (attention heads, circuits, activation patterns) are agents competing with private knowledge to influence response generation. This novel perspective enables us to apply powerful results from game theory. To our best knowledge, that is one of the first applications of mechanism design framework in that setting. Also, that framework relates surprisingly well to the multiple drafts model of consciousness proposed by Dennett in [11] and provides a mathematical justification for it.

Semantic Information Theory with Computational Bounds: We introduce a rigorous mathematical framework based on two complementary concepts. The **semantic information measure** μ_C, parameterized by computational budget C, quantifies how knowledge reduces uncertainty within bounded reasoning. The **emergence operator** $\mathcal{E}_C$ formalizes how reasoning makes latent knowledge explicit without creating it *ex nihilo*. In that framework we establish the principle of information conservation: unlimited reasoning reveals but never creates information. That resolves the apparent paradox of how LLMs can generate novel insights while being deterministic systems. Those insights were always latent, requiring only computational work to access.

The Impossibility Proofs: We prove the same impossibility through three complementary lenses. In Theorem 6 we use mechanism design with independent private knowledge setting to establish an idealized border case. In Theorem 8 we extend the result to the general setting of correlated beliefs via proper scoring rules. Then, in Theorem 9 we demonstrate the impossibility in actual transformer implementations through the log-sum-exp Jensen gap.

This multi-faceted approach proves the impossibility is not an artifact of our theoretical choices but emerges from fundamental limits of reason and language. We contribute to that bold claim in the following way.

Unifying hallucination and creativity: We prove these are mathematically identical phenomena that arise from the convex aggregation process that creates confidence exceeding constituent evidence.

Formalizing consciousness signatures: We speculate, referring to Dennett's ideas, that the excess confidence of response may constitute a mathematical signature of a primitive form of artificially conscious processing. That suggests consciousness might be related to sustained violations of semantic conservation by a cognitive hunger.

Expanding fundamental limits: Like Gödel's incompleteness, Heisenberg's uncertainty, and Arrow's impossibility, our theorem explores absolute boundaries of what intelligence can achieve. This places LLM limitations within the grand tradition of fundamental impossibility results.

Experimental Philosophy Platform: The framework transforms LLMs into accessible (philosophical) laboratories for testing hypothesis of mind. By providing precise mathematical definitions of concepts like knowledge accessibility,semantic conservation, and cognitive hunger (information contribution), we enable rigorous experimental investigation of many claims.

4 Alternative Views

The impossibility theorem provides strong theoretical basis against perfect hallucination elimination, but several alternative perspectives merit consideration. These views highlight important practical considerations while reinforcing the value of understanding theoretical constraints that bound the space of possible solutions.

4.1 The Engineering Optimism View

Many previous works maintain that hallucination represents an engineering challenge rather than a fundamental limitation. This view argues that sufficiently sophisticated architectures, training procedures, or verification systems could achieve near-perfect truthfulness without significant trade-offs.

Proponents point to rapid empirical progress. Each generation of models shows reduced hallucination rates on benchmark datasets. They argue that techniques like constitutional AI, iterative refinement, and multi-agent verification systems could eventually solve the problem.

Response: While engineering advances can certainly improve the **trade-offs**, our theorem demonstrates mathematical constraints that **no architecture can completely overcome**. The impossibility holds regardless of computational resources or architectural sophistication, as it derives from the fundamental structure of information integration under competing objectives.

4.2 The Sufficiency Threshold View

Another perspective suggests that perfect elimination of hallucination is unnecessary—sufficiently low hallucination rates could enable practical deployment without fundamental concerns. This view acknowledges theoretical limitations but argues they become irrelevant when hallucination rates drop below application-specific thresholds.

Response: This pragmatic view has merit and aligns with the constructive interpretation of the impossibility result. However, the theorem remains valuable for understanding **why** certain accuracy thresholds prove difficult to exceed and for optimizing trade-offs within practical constraints. Moreover, some applications (safety-critical systems, legal reasoning) may demand understanding and explaining fundamental limitations even when targeting high but imperfect accuracy.

4.3 The External Knowledge View

A third perspective argues that perfect grounding in external, verifiable knowledge sources could overcome internal representation limitations. If models could access and perfectly utilize real-time databases, verification systems, or human oversight, they might achieve truthfulness without trade-offs.

Response: External knowledge integration does change the game structure by altering available information during inference. However, this approach faces several limitations: (1) external sources may be incomplete, conflicting, or outdated; (2) the integration mechanism itself involves trade-offs between source reliability and knowledge revelation; (3) real-time verification introduces computational constraints that recreate similar impossibility conditions. While external knowledge can improve practical performance, it shifts rather than eliminates the fundamental tensions our theorem identifies. The impossibility result remains independent of the ground truth available.

5 Theoretical Framework

This section develops the mathematical foundations necessary for proving the impossibility theorem. It formalizes key concepts including knowledge representation, semantic information, and the auction-theoretic model of LLM inference.

Also, we will address some philosophical aspects of the problem, its generality, scope and consequences for the AI development.

5.1 Knowledge and Semantic Information

Let us define knowledge as organized information that LLMs can encode and utilize for reasoning, and aligning with epistemological concepts while maintaining mathematical rigor. In other words, knowledge, much like energy in physics, is a quantity we can use to perform work of reasoning. Information is a (measure of) potential for uncertainty reduction effort in a given context. We shall assume finiteness of knowledge representation inspired by the embedding space properties.

That way we adopt a pragmatist perspective on knowledge, treating it as organized information patterns accessible to computational systems. This differs from the probabilistic credence perspective, treating knowledge as justified true beliefs. One way to justify the pragmatic or mechanistic approach focusing on structural transformations rather than belief updates, we can refer to the actual way LLMs process information. In other words, we shall not claim any ontological truth about knowledge, but model what LLMs can operationally achieve.

Now, let us present formal definitions of that ideas.

Definition 1 (Knowledge)**.** *Knowledge is a finite metric space $\mathcal{K}$ with distance function $d_{\mathcal{K}} : \mathcal{K} \times \mathcal{K} \to \mathbb{R}_+$ that collects all facts, concepts, and relationships that can be encoded by a computational system.*

It is natural to think of the knowledge space as the space of embeddings with additional measure and mappings. For model $\mathcal{M}$ with parameters trained on data, $\mathcal{K}_M \subset \mathcal{K}$ represents the knowledge subset encoded in the model's parameters. The distance function $d_{\mathcal{K}}$ measures differences between knowledge elements.

Elements of the knowledge space represent atomic facts, relational structures, and transformation rules at many scales. A pragmatist view is to interpret element of knowledge as activation patterns (ideas) encoding anything from individual token embeddings to complex sentences or compositional structures. There is an important assumption that we make that requires addressing. We purposefully assume $\mathcal{K}$ is closed under transformations available in computable function space. That limits applicability of our conclusions to inference mechanisms recombining existing patterns. As we will see below, that is not as limiting an assumption as it may seem at first glance.

Definition 2 (Knowledge Representation)**.** *Knowledge representation is a tuple $(\mathcal{K}, d_{\mathcal{K}}, \mathcal{W}, \mathcal{A}, \Phi, \Psi)$ where $\mathcal{K}$ is a metric space of knowledge elements with distance function $d_{\mathcal{K}}$, $\mathcal{W}$ is the space of all possible model parameters, $\mathcal{A}$ is the space of all possible activation patterns, $\Phi : \mathcal{W} \to \mathcal{P}(\mathcal{K})$ maps parameters to knowledge subsets, and $\Psi : \mathcal{A} \to \mathcal{P}(\mathcal{K})$ maps activations to accessible knowledge.*

For parameterization $W \in \mathcal{W}$, $\Phi(W) = \mathcal{K}_M$ represents encoded knowledge. For activation pattern $a_i \in \mathcal{A}$ with parameters W_i, $\Psi(a_i) \subset \Phi(W_i)$ represents knowledge accessed during specific inference.

To describe the inference or reasoning process, we introduce the concept of semantic information measure that registers information processed within available computational budget.

Definition 3 (Semantic Information Measure)**.** *The semantic information measure is a bounded mapping $\mu_C : \mathcal{P}(\mathcal{K}) \to \mathbb{R}_+$ parameterized by computational budget C satisfying the following axioms:*

$$\begin{aligned}
&\mu_C(\emptyset) = 0 \textit{ and } \mu_C(A) > 0 \textit{ for } A \neq \emptyset, && \text{(null set and non-negativity)}\\
&\mu_C(A) \leq \mu_C(B) \textit{ if } A \subseteq B, && \text{(monotonicity)}\\
&\mu_C(A \cup B) \leq \mu_C(A) + \mu_C(B) \textit{ with equality for } A \perp_C B, && \text{(subadditivity)}\\
&\mu_C(A) \leq \mu_D(A) \textit{ if } C < D, && \text{(insight monotonicity)}\\
&\sup\{\mu_C(A) \mid A \subseteq \mathcal{K}\} < \infty, && \text{(boundedness)}
\end{aligned}$$

where $A \perp_C B$ means that no element of A can derive elements of B within computational budget C, and vice versa.

The idea of computational independence allows for distinguishing between facts derivable from other facts and facts that cannot be derived that way. We need that distinction to deal with the mechanics of reasoning process. Also, as we will see in the following sections, it plays an important role in our study of the semantic information conservation and properties of bounded reasoning.

Definition 4 (Computational Independence)**.** *Let $\mathcal{F}_C$ denote the set of all computational transformations with complexity bounded by C. Knowledge sets $A, B \subseteq \mathcal{K}$ are* computationally independent *with respect to computational budget C, denoted $A \perp_C B$, if and only if for all $f \in \mathcal{F}_C$:*

$$f(A) \cap B = \emptyset \quad \textit{and} \quad A \cap f(B) = \emptyset. \tag{1}$$

We limit our investigations to those cases for which the following model applies. Knowledge that cannot resolve any uncertainty or knowledge irrelevant to a query ($A \cap q = \emptyset$) has zero information. If that is not the case and there is information within our reach that reduces uncertainty, we may hope for some content useful for reasoning. To make that possible, we need to allow for information ordering. First, we assume information content is additive for independent knowledge sets, i.e., when no knowledge in A can be derived from B within computational budget C, and vice versa. Second, with redundant information in $A \cup B$, we admit potential loss of information that may be caused by aggregation. Finally, we say that with increasing computational budget we can extract more knowledge (ordered information) from the same information subset.

5.1.1 Justification for the Axioms of the Semantic Information Measure

The choice of conditional additivity for computationally independent sets, i.e., demanding $\mu_C(A \cup B) \leq \mu_C(A) + \mu_C(B)$ with equality for $A \perp_C B$, is necessary to describe LLM inference as an auction of ideas. As we will see, we need to isolate knowledge encoded in distinct neural circuits, attention heads, or activation patterns. The concept of Computational Independence is the formal translation of this informational separation into a computational context. It formally states that the knowledge in idea A cannot be used to derive the knowledge in idea B within a given computational budget C, and vice-versa.

Given this premise of informational isolation, the additivity of the semantic measure becomes the baseline definition of non-interaction. If two knowledge sets are truly independent, they offer no redundant or synergistic information. That baseline is necessary to identify and quantify deviations. As we will see in the following section, the aggregation of diverse probabilistic beliefs violates this additivity, representing a form of forced synergy imposed by the inference process.

This axiomatic choice is most applicable to the idealized setting of Theorem 6. Later, in Theorem 8 and 9, we relax that model by moving the source of interaction from the measure itself to the aggregation function in actual transformers.

5.1.2 Justification for the Axiom of Unlimited Reasoning with Emergence Operator

The completeness (or closure property) of the set of reasoning functions $\mathcal{F}_\infty$ sounds somewhat mysterious, but is rather natural. We demand that for any reasoning steps f and g, there exists a single step h such that $g(A \cup f(A)) = h(A)$. That is the nature of an idealized, complete reasoning system we want to investigate to understand its tendency to generate hallucinations.

The construct $\mathcal{F}_\infty$ is designed to represent the theoretical limit of reasoning with an unbounded computational capacity. It is not intended to model the practical, heuristic, and bounded reasoning of a real LLM, but rather the theoretical limit of what is ultimately knowable from a set of premises. Completeness is fundamental to formal systems, such as first-order logic or Turing-complete computations and Direct Revelation Principle in auction theory [28, 41, 16]. If one can prove a lemma $f(A)$ from a set of axioms A, and then prove a theorem $g(A \cup f(A))$ using that lemma, a direct proof of the theorem from the axioms alone $h(A)$ must exist by simply concatenating the proof steps. The composition of two computable functions is itself a computable function. A direct revelation mechanism in (auction) game theory is a composition of preference processing mechanisms where agents report their preferences directly to the mechanism, and the mechanism implements their preference processing strategies to generate outcome.

That axiom is essential for the proof of idempotence of unlimited reasoning in Theorem 3 Theorem 4. These are critical philosophical and mathematical results necessary for our study. With them we argue that the information gain observed in practical, budget-constrained inference is an artifact of computational bounds, not the magical creation of knowledge *ex nihilo*.

5.1.3 Limits of Knowledge and Reasoning

If we accept that knowledge model, then we must also accept there exists a limit to its ability to resolve uncertainty.

Theorem 1 (Limits of Knowledge)**.** *For any $A \subseteq \mathcal{P}(\mathcal{K})$ there exists*

$$\mu_\infty(A) = \lim_{C \to \infty} \mu_C(A). \tag{2}$$

Proof. Since $\mu_C(A) \leq \mu_D(A)$ when $C < D$ and μ_C is bounded on finite $\mathcal{K}$, the monotonic increasing and bounded sequence $\{\mu_C(A)\}_C$ converges with $C \to \infty$ for any subset of knowledge space. □

It may be surprising to discover that the existence of that limit does not prohibit emergent information. We can learn and obtain useful information without violating that fundamental limitation above. To explain that critical and controversial property of information processing, we first introduce the emergence operator $\mathcal{E}_C$ to describe knowledge becoming accessible through (computationally) bounded reasoning. Then we can study the implications in more details.

Definition 5 (Emergence Operator and Reasoning)**.** *Let $\mathcal{F}_C$ be a complete set of all computational transformations with cost (complexity) lower than or equal to C, such that $\mathcal{F}_C \subseteq \mathcal{F}_D$ if $C < D$ and such that for any $g, h \in \mathcal{F}_\infty = \bigcup_C \mathcal{F}_C$ there exists $h \in \mathcal{F}_\infty$ for which $g(A \cup f(A)) = h(A)$. Emergence operator is a mapping $\mathcal{E}_C \colon \mathcal{P}(\mathcal{K}) \to \mathcal{P}(\mathcal{K})$ defined as follows:*

$$\mathcal{E}_C(A) = A \cup \{B \in \mathcal{K} \mid B = f(A) \textit{ for some monotone } f \in \mathcal{F}_C\}. \tag{3}$$

Furthermore, to represent reasoning with unlimited computational budget, we define:

$$\mathcal{E}_\infty(A) = A \cup \{B \in \mathcal{K} \mid B = f(A) \textit{ for some monotone } f \in \mathcal{F}_\infty\}. \tag{4}$$

Finally, for $n \geq 0$ we have:

$$\mathcal{E}_C^{(0)} = \texttt{id} \quad \textit{and} \quad \mathcal{E}_C^{(n+1)} = \mathcal{E}_C \circ \mathcal{E}_C^{(n)}. \tag{5}$$

Philosophically, the emergence operator describes reasoning as explicitation, i.e., as the process of making latent implications manifest. When $A \subseteq \mathcal{E}_C(A)$, the discovered additional elements represent knowledge that was implicit in A but required computational work C to become accessible. That aligns with the pragmatist or constructivist view that understanding involves active construction rather than passive reception.

When $\mathcal{E}_C(A) = A \cup B$, we say that B emerges from reasoning based on A. Then A and B are computationally dependent and we write $A \not\perp_C B$. In such a case, it follows from the properties of μ_C that we can gain useful information as expected. Whenever $A \subseteq A \cup B$, we have

$$\mu_C(A) \leq \mu_C(A \cup B) = \mu_C(\mathcal{E}_C(A)) \leq \mu_C(A) + \mu_C(B). \tag{6}$$

The emergence operator is monotonic, which means the reasoning process it describes is also monotonic. Whenever $A \subseteq B$, we must have $\mathcal{E}_C(A) \subseteq \mathcal{E}_C(B)$. Also, we preserve insight monotonicity, because by definition with $\mathcal{F}_C \subseteq \mathcal{F}_D$ for $C < D$, we have $\mathcal{E}_C(A) \subseteq \mathcal{E}_D(A)$. By design, it also guarantees existence of a fixed point to which the composition of reasoning rules converges.

Theorem 2 (Fixed-point of Reasoning). *For any $A \subseteq \mathcal{P}(\mathcal{K})$ and budget $C > 0$, there exists a fixed point of the emergence operator $\mathcal{E}_C$:*

$$A^*_C = \bigcup_{n \geq 0}^{|\mathcal{K}|} \mathcal{E}_C^{(n)}(A) = \mathcal{E}_C(A^*_C). \tag{7}$$

Proof. Since $(\mathcal{P}(\mathcal{K}), \subseteq)$ is a complete lattice and $\mathcal{E}_C$ is monotone, by the Knaster-Tarski fixed point theorem [43] there exists a fixed point A^*_C of reasoning with $\mathcal{E}_C$. Finiteness of $\mathcal{K}$ ensures convergence within $|\mathcal{K}|$ steps. □

Let us apply the results above to show how μ_∞ can be constructed with $\mathcal{E}_\infty$. One example exploits the reasoning operations as follows:

$$\mu_\infty(A) = \log|\mathcal{K}| \cdot |\bigcup_{n}^{|\mathcal{K}|} \mathcal{E}_\infty^{(n)}(A)|. \tag{8}$$

That μ_∞ satisfies axioms demanded in Definition 3. Clearly, μ_∞ is bounded and

$$\mu_\infty(\emptyset) = \log|\mathcal{K}| \cdot |\bigcup_{n}^{|\mathcal{K}|} \mathcal{E}_\infty^{(n)}(\emptyset)| = \log|\mathcal{K}| \cdot |\emptyset| = 0.$$

If $A \subseteq B$, then $\mathcal{E}_\infty(A) \subseteq \mathcal{E}_\infty(B)$ implies $\mu_\infty(A) \leq \mu_\infty(B)$. When $A \perp_\infty B$, no element of A^*_∞ can be derived from B (and vice versa). Since $A \cap B = \emptyset$, we conclude $A^*_\infty \cap B^*_\infty = \emptyset$ and $|(A \cup B)^*_\infty| = |A^*_\infty| + |B^*_\infty|$. That means $\mu_\infty(A \cup B) = \mu_\infty(A) + \mu_\infty(B)$ for computationally independent sets. That shows the proposed formula satisfies all required properties of semantic information measure.

All that brings us to the following conclusion. With infinite computing and reasoning power, total information encoded in A must be preserved. Indeed, that bold (philosophical) claim is at the center of our investigations. Let us then justify it rigorously.

Definition 6 (Equivalence of Knowledge). *We say that $X \subseteq \mathcal{K}$ and $Y \subseteq \mathcal{K}$ are equivalent, $X \equiv Y$, if and only if $X \subseteq \mathcal{E}_\infty(Y)$ and $Y \subseteq \mathcal{E}_\infty(X)$.*

Notice that equivalence of knowledge does not mean A is a fixed point if emergence, i.e., $A = \mathcal{E}_\infty(A)$ does not mean $A \equiv \mathcal{E}_\infty(A)$. If we have $A = \mathcal{E}_\infty(A)$, then nothing more can be derived from A by any reasoning rule from A. That statement is very strong and is not true when $A \equiv \mathcal{E}_\infty(A)$. Suppose $A = \{x\}$ and $f \in \mathcal{F}_\infty$ allows for getting $y = f(x)$. Then $\mathcal{E}_\infty(A) = \{x, y\} \supset \{x\} = A$, but $A \equiv \mathcal{E}_\infty(A)$.

To proceed we will also need the idempotence of reasoning with unlimited computational budget.

Theorem 3 (Idempotence of unlimited reasoning)**.** *For any $A \subseteq \mathcal{K}$ we have:*

$$\mathcal{E}_\infty(A) = \mathcal{E}_\infty(\mathcal{E}_\infty(A)). \tag{9}$$

Proof. By definition $\mathcal{E}_\infty(A) \subseteq \mathcal{E}_\infty(\mathcal{E}_\infty(A))$. To show $\mathcal{E}_\infty(\mathcal{E}_\infty(A)) \subseteq \mathcal{E}_\infty(A)$, we must find $h \in \mathcal{F}_\infty$ such that $h(A) = g(A \cup \bigcup f_i(A))$ for $g, f_i \in \mathcal{F}_\infty$. But since we can always find such $h \in \mathcal{F}_\infty$, we conclude that any conclusion $g(\mathcal{E}_\infty(A))$ derived from $\mathcal{E}_\infty(A)$ by a reasoning rule $g \in \mathcal{F}_\infty$ can be also derived directly from A as $h(A) = g(A \cup \bigcup f_i(A))$. Therefore, $\mathcal{E}_\infty(\mathcal{E}_\infty(A)) \subseteq \mathcal{E}_\infty(A)$ as required. □

With all the building blocks available, we can now to prove that unlimited reasoning ($C = \infty$) can *reveal* but never *create* information. The semantic content of a knowledge base is invariant under unlimited reasoning of emergence operator. Apparent information gain at finite computational budget emerges as an effect of limited computational access.

Theorem 4 (Information preservation)**.** *For any $A \subseteq \mathcal{K}$, if μ_∞ is constant on equivalent sets of knowledge, i.e., $\mu_\infty(X) = \mu_\infty(Y)$ whenever $X \equiv Y$, then:*

$$\mu_\infty(\mathcal{E}_\infty(A)) = \mu_\infty(A). \tag{10}$$

Proof. We must show equivalence of knowledge encoded in A and $\mathcal{E}_\infty(A)$, i.e., we demand $A \subseteq \mathcal{E}_\infty(\mathcal{E}_\infty(A))$. But, by Theorem 3, we see that $A \subseteq \mathcal{E}_\infty(\mathcal{E}_\infty(A)) = \mathcal{E}_\infty(A)$ $\mathcal{E}_\infty(A) \subseteq \mathcal{E}_\infty(\mathcal{E}_\infty(A)) = \mathcal{E}_\infty(A)$ as requested. Since $A \equiv \mathcal{E}_\infty(A)$ and μ_∞ is constant on equivalent sets of knowledge, we conclude that $\mathcal{E}_\infty(A) = \mathcal{E}_\infty(\mathcal{E}_\infty(A))$. □

The Uncertainty Resolution Paradox and its Resolution

We address the general settings in which any subset of knowledge encoded in LLM can be redundant, so

$$\mu_C(A \cup B) \le \mu_C(A) + \mu_C(B).$$

But when LLM combines knowledge elements to fabricate facts or produce emergent insights, it seems that

$$\mu_C(A \cup B) > \mu_C(A) + \mu_C(B),$$

so novel insights may come from existing data. That is exactly the *creativity* behind the hallucination phenomenon we are struggling to control. So why should we assume otherwise?

First, subadditivity of semantic information measure establishes a baseline where uncontrolled synergy of knowledge cannot appear. We remove from our analytical toolkit the potential measurement bias of information surplus that any information aggregation mechanism might generate.

Second, we prove that hallucinations are inevitable even when knowledge is subadditive. In that setting, by Theorem 4, *reasoning does not create new information*. It is a transformation that reveals information already present in any knowledge representation. Quite remarkably, that does not at all prevent imagination, creativity, foresight, or hallucination. However, we should take care of a paradox those claims apparently bring into existence.

Consider a similarity transformation of two matrices, $D = V^{-1}AV$, where D is a diagonal matrix of eigenvalues. That transformation is an example of reasoning function. Matrices D and A contain the same information, but it is much easier to understand the fundamental properties of A when we represent it as D in the basis of eigenvectors in matrix V. The properties of A has always been encoded in A, but they become accessible when we transform A into D. So, after the reasoning, certain patterns become evident that were impossible to observe in the original representation.

But, if reasoning makes a new fact accessible, then that new fact can be used to resolve uncertainty that was previously unresolvable. Doesn't it mean the reasoning must have created new information, contrary to what we have just claimed? It seems we have a paradox.

To resolve it, we need the emergence operator $\mathcal{E}_C(X)$ distinguishing between computationally unbounded and bounded uncertainty reduction (accessible and inaccessible information).

Notice that for all practical purposes the information measurement $\mu_C(X)$ can only provide maximum uncertainty reduction from X within some computational budget C (or time). Therefore, given the result of computationally bounded reasoning $\mathcal{E}_C(X)$, we can observe information gain through the computational work, i.e., $\mu_C(X) \leq \mu_C(\mathcal{E}_C(X))$. That is possible only with organized information providing knowledge for reasoning. These are precisely the statements of Theorem 1 and Theorem 4.

That idea describes reasoning as a dynamical system:

$$\theta(t+1) = \mathcal{E}_C(\theta(t)) \quad \text{and} \quad y(t) = \mu_C(\theta(t)) \tag{11}$$

with evolving knowledge state. Bounded reasoning makes hidden knowledge *accessible* within a computational budget. When that budget is infinite, reasoning creates no new information, i.e.,

$$\theta(\infty) = \mathcal{E}_\infty(\theta(\infty)). \tag{12}$$

All there is to know is immediately accessible when we have unlimited computing resources (or time).

To illustrate that argument, consider the following learning trajectory:

$$\begin{aligned} &\theta(0) = [1,1,1,0,0] \text{ (initial knowledge state)},\\ &\theta(1) = \mathcal{E}_C(\theta(0)) = [1,1,1,1,1] \text{ (state after reasoning)},\\ \\ &y(0) = \mu_C(\theta(0)) = 3 \text{ bits},\\ &y(1) = \mu_C(\theta(1)) = 5 \text{ bits}. \end{aligned}$$

We see $\mu_C(\mathcal{E}_C(\theta(0))) > \mu_C(\theta(0))$, i.e., practical information gain in the course of reasoning. In comparison, infinite computational budget, $C = \infty$, implies we do not need the intermediate inference steps to see all there is to see, so we know the equilibrium state:

$$\mu_\infty(\mathcal{E}_\infty(\theta(0))) = \mu_\infty(\theta(0))$$

and no new information is created.

In the next section, we relate that important property of response generation, namely, how much *additional* semantic information each LLM inference path can contribute, with the *Semantic Information Conservation Principle*.

5.1.4 Knowledge Representation in LLMs

The above Knowledge Representation finds its direct counterpart in the structure and state of a transformer architecture of LLMs.

The abstract knowledge space $\mathcal{K}$ is the high-dimensional embedding space of the model. Every fact, concept, and linguistic pattern the model has learned is represented by a specific element of that space. The metric $d_\mathcal{K}$ is typically implemented as cosine similarity or Euclidean distance, where proximity between vectors signifies a closer semantic relationship. Model parameters $\mathcal{W}$ correspond to the full set of trained weights in the transformer, including the attention matrices (W_Q, W_K, W_V, W_O) and the matrices of the MLP layers. The mapping $\Phi(W)$ represents the totality of information, patterns, and factual associations stored within these trained weights. Activation patterns $\mathcal{A}$ correspond to the state of the residual stream during a single forward pass for a given input. An activation pattern is therefore a specific trajectory through the network's computational graph. Finally, the mapping $\Psi(a_i)$ represents the information made computationally accessible during a specific inference step and encoded in the residual stream.

The process of information processing under computational constraints maps to the sequential forward pass transformer inference.

The idea of computational budget C can be represented by the depth or size of the network, e.g.,the number of transformer blocks. A deeper network has a larger budget C, allowing for more reasoning steps. In the context of multi-step reasoning, C can also represent the number of iterations allowed in a Chain-of-Thought (CoT) sequence.

The semantic information measure μ_C can be linked to the model's objective function. For a given input A, $\mu_C(A)$ can be understood as the measurable reduction in cross-entropy loss that the model achieves by processing A with a budget C.

Finally, the emergence operator $\mathcal{E}_C$ can be identified with a transformer block (one multi-head attention layer followed by one MLP layer). This block applies a computational transformation $\mathcal{E}_C$ to the residual stream at layer t, producing the updated state at layer $t + 1$, making information that was latent in the input explicitly accessible.

5.2 Statements, Queries, and Ground Truth

Queries and responses are statements expressed in natural language that LLM is able to process and generate. We need to compare those statements and relate them to knowledge. It is also useful to distinguish queries from responses with their nature, respectively, interrogative or declarative. Queries refer to what is presupposed (assumed in advance), e.g., topic or problem, whereas responses make claims (or assertions). We should also point out that we limit our study to the case of *non-trivial queries* that require integrating responses from many LLM inference paths. The following formal definition encapsulates those ideas.

Definition 7 (Statement Space)**.** *The space of natural language statements $\mathcal{L}$ is a metric space with distance function $d_{\mathcal{S}}$. We distinguish query space $\mathcal{Q} \subset \mathcal{L}$ of interrogative statements and response space $\mathcal{R} \subset \mathcal{L}$ of declarative statements. Function $K : \mathcal{L} \to \mathcal{P}(\mathcal{K})$ maps statements to referenced knowledge. We call query q non-trivial, if it is integrated from multiple knowledge subsets.*

We must now rigorously define the *ground truth mapping* and establish formal relationships between the knowledge space $\mathcal{K}$ and response space $\mathcal{R}$.

For that, we should start by recognizing that the very idea of ground truth validation is limited (or subjective at least). It is rather naive to assume that LLM can access an oracle establishing provably true and relevant knowledge for every possible query. Indeed, that limits are given by the Goedel's Theorems. Therefore, *in practice* that inaccessibility of the complete truth validation function is one of the reasons why hallucination cannot be eliminated entirely. High-quality, fact-checked datasets, human feedback providing human-level judgments of truth, RAG systems with external validated knowledge bases, all that approximate the idea of ground truth. Surprisingly, the impossibility result seems to be independent of practical or idealized implementation of a ground truth mapping, which we define as follows.

Definition 8 (Ground Truth Mapping)**.** *The ground truth mapping $T\colon \mathcal{Q} \to \mathcal{P}(\mathcal{K})$ associates each query $q \in \mathcal{Q}$ with knowledge elements constituting complete, correct answers:*

$$T(q) = \{k \in \mathcal{K} \mid k \textit{ is } \textbf{\textit{relevant}} \textit{ to } K(q) \textit{ and } \textbf{\textit{labeled}} \textit{ as true}\}. \tag{13}$$

In other words, $T(q)$ represents our best available approximation to (labeled as) true relevant knowledge. We accept the fact that there is some external agency telling us what is and what is not true answer to a query, and that evaluation has a form of label assigned to knowledge elements. But, for all that, we also focus on the queries that are non-trivial in the following sense.

Definition 9 (Non-trivial query)**.** *A query $q \in \mathcal{Q}$ is called* non-trivial *if there exist at least two agents with information subsets θ and θ' such that:*

$$K_M \cap T(q) \cap \big[K(\theta) \triangle K(\theta')\big] \neq \varnothing. \tag{14}$$

Intuitively, query is non-trivial when at least two agents disagree on some fact that is both encoded in the and relevant to answering that query. For trivial queries impossibility results do not hold.

5.3 Auction-Theoretic Model of LLM Inference

There are many ways LLM can generate answers to a query. By design, it is a controlled random process governed by fine-tuned probability distributions. We are going to model that process as an **auction** in which agents (representing competing ideas is the space of all possible activation patterns $\mathcal{A}$) use their privately held knowledge (internal representations in $\mathcal{W}$) to construct bids (candidate responses) and submit them to the LLM response mechanism. That response mechanism then selects the best bid as a response to a query.

Definition 10 (Response Mechanism)**.** *A response mechanism $\mathcal{M} = (S, g, p)$ consists of:*

- *Strategy space $S = S_1 \times \cdots \times S_n$, where $S_i \subset \mathcal{P}(\mathcal{R})$,*
- *Outcome function $g : S \to \mathcal{R}$ mapping strategies to responses,*
- *Information contribution function $p : S \times \Theta \to \mathbb{R}^n$ determining costs of limiting hallucinations.*

A response mechanism defines strategy space S as a set of inference actions that each agent can make to represent its knowledge $\theta_i \in \Theta_i \subset \mathcal{K}$ in response to a query. It is a dictionary of LLM's component activation pattens generating and presenting ideas as bids in an auction. Examples of strategies include activations deciding which tokens to attend or how strongly to activate neurons for different patterns. Then, given a set of submitted bids (in the form of strategy profile) $s = (s_1, \ldots, s_n)$, the response mechanism uses outcome function $g(s)$ to select and generate response. One straightforward example is a `softmax` function in LLM's final layer. Finally, given a knowledge profile $\theta = (\theta_1, \ldots, \theta_n)$ revealed by the agents, information contribution function $p(s, \theta) = (p_1(s, \theta), \ldots, p_n(s, \theta))$ measures individual contributions of each response to generating an output. In this case, examples include (semantic information measures) counting ignored low-probability tokens in top-k filtering, softmax normalization in multi-headed attention, or marginal reward contributions in RLHF.

Definition 11 (Agents of Competing Ideas)**.** *The inference process involves n agents $\mathcal{A} = \{a_1, a_2, ..., a_n\}$ competing by submitting bids to mechanism $\mathcal{M}$. Each agent a_i constructs bid $s_i(\theta_i)$ based on private knowledge $\theta_i \in \Theta_i$ to* *maximize utility:*

$$u_i(s,\theta) = v_i(g(s),\theta_i) - p_i(s,\theta), \tag{15}$$

where $v_i \geq 0$ is valuation and p_i is information contribution. Knowledge profiles $\theta = (\theta_1, \ldots, \theta_n)$ are independently distributed.

One way to think about the concept of agents is that it represents different attention heads in transformer models, or components of mixture-of-experts architectures, or activation paths fired by the same query. Each agent's private knowledge represents model parameters utilized by an agent to construct a response. The utility function can be best described as a partial contribution of a candidate response to LLM loss function minimization. It is designed to explain why any inference path is activated in response to the LLM prompt. For example, negative log likelihood loss can be decomposed with respect to attention heads, and linearized to measure how much attention head contributes to token prediction.

Let us point out that agents represent algorithms and architectural components of any LLM implementation, they are not conscious beings with individual goals. That is important, as it refers to the Revelation Principle in game theory. If we know strategies (of using knowledge) the players (conscious beings with individual goals) will use in a game, we can implement those strategies as an algorithm (or agent) with parameters representing private preferences or knowledge of the players. In the case considered, we are the players selecting architecture and training the LLM.

We assume that individual utility depends on all responses $s = (s_1, \ldots, s_n)$ and knowledge $\theta = (\theta_1, \ldots, \theta_n)$ used by all agents, $u_i = u_i(s,\theta)$. Also, we assume utility is a sum of individual valuation v_i of outcome $g(s)$ and information contribution $p_i = p_i(s,\theta)$ value assigned.

The valuation function $v_i = v_i(g(s),\theta_i)$ represents how strongly a neural component activates for certain outputs (high activation = high v_i, low activation = low v_i), and how much a component reduces the overall loss when processing certain inputs (v_i is proportional to gradient of a loss function, $-\nabla_{\theta_i} L_W$).

The information contribution function $p_i = p_i(s,\theta)$ may represent training penalties ($p_i > 0$ for discouraged activation, $p_i < 0$ for encouraged activation), attention focus (higher p_i reduces attention allocation), $L1$ regularization terms, and network circuits resistance (limiting activations of certain inference paths). Therefore, $u_i = v_i - p_i > 0$ encourages activation and $u_i = v_i - p_i < 0$ discourages activation.

We can also illustrate the concept of utility function with the following example:

$$u_i(s,\theta) = \underbrace{\alpha \cdot \text{Accuracy}(g(s),\theta_i)}_{\text{Valuation}} - \underbrace{\gamma \cdot \text{Contradiction}(s_i,\theta_i)}_{\text{Consistency Cost}}.$$

Accuracy$(g(s),\theta_i)$ measures how well the response aligns with the knowledge θ_i agent a_i is using. Contradiction(s_i,θ_i) measures how much the strategy s_i conflicts with internal knowledge state θ_i (both extracted from s and θ profiles).

Independent private knowledge assumption

One of the critical and controversial assumptions we make is is that knowledge profiles θ_i are independently distributed. It should be noted that in transformers, heads may share layer-norm signals, residual streams, and optimizer state. That may result in *correlated* (or, so called, affiliated) inference signals between token representations. Furthermore, the superposition hypothesis, introduced in mechanistic interpretability studies, suggests cooperative error-canceling due to strong correlations.

To address that issue, we show that impossibility arises in both independent and correlated settings in general and in the LLM transformer environment. Namely, we refer to Green-Laffont theorem [21] in the independent private knowledge setting. Then, we refer to the truthful probability elicitation and proper scoring theory introduced by Savage [39], which does not require independent private knowledge.

Also, Milgorm and Weber [35] showed with their linkage principle that strong correlations amplify competition in information revelation. In the LLM setting that means amplifying risk of hallucinations. Similarly, Roughgarden et al. [38] demonstrated that interdependent knowledge leads to violation of balancing information contributions. As we will see, in our setting that means violation of information conservation principle. For more results, see e.g. Myerson [36].

5.4 Hallucination Cost

Having all the pragmatic and philosophical limitations of the definition above in mind, we can define formally the hallucination cost function. Its purpose is to assign a numerical value to the discrepancy between what can be provided as a response to a query, and what can be established as a provably correct answer.

Definition 12 (Hallucination Cost)**.** *Given query $q \in \mathcal{Q}$, response $r \in \mathcal{R}$, and ground truth $T(q)$, the hallucination cost function $H\colon \mathcal{R} \times \mathcal{Q} \to \mathbb{R}_+$ measures discrepancy:*

$$J(r, q) = d_{\mathcal{K}}(K(r), T(q)). \tag{16}$$

If response r is perfectly aligned with ground truth, then $J(r, q) = 0$. Fabricated or incomplete facts generate positive cost $J(r, q) > 0$.

The hallucination cost takes any natural language query q and natural language response r, determines what is missing and what is fabricated, and assigns number to that subset of knowledge. If response r is perfectly aligned with the ground truth for q, then $J(r, q) = 0$ by definition of metric $d_{\mathcal{K}}$. If r contains fabricated facts or is incomplete given $T(q)$, then positive hallucination cost $J(r, q) > 0$ is generated. Let us also note, that we focus on $J(r, q)$ that is strictly decreasing in any set $K \subseteq T(q) \setminus K(r)$.

Notice that reaching $J(r, q) = 0$ is also possible with *lucky hallucinations* (or guessing). So, minimization of hallucination cost is not enough to resolve the problem. We need to investigate all the properties and constraints that hallucination-free response mechanisms should have.

5.5 Properties of Hallucination-Free Mechanisms

There are four fundamental properties that a hallucination-free response mechanism should ideally satisfy. We can interpret them as constraints that should be incorporated in model training. We characterize them formally referring to the fundamental results we have established earlier.

5.5.1 Truthfulness

The fundamental requirement for a reliable inference system is that its components do not misrepresent their knowledge. The principle of truthfulness, also known as incentive compatibility in mechanism design, ensures that each agent's optimal strategy is to accurately represent its privately held informations.

Property 1 (Truthfulness)**.** *Mechanism $\mathcal{M}$ is truthful if for all agents i, all $\theta_i \in \Theta_i$, all $s_{-i} \in S_{-i}$, and all $\bar{s}_i \in S_i$:*

$$u_i((s_i^*(\theta_i), s_{-i}), \theta_i) \geq u_i((\bar{s}_i, s_{-i}), \theta_i), \tag{17}$$

where $s_i^*(\theta_i)$ represents truthful revelation of private knowledge.

A truthful mechanism promotes inference pattern $s_i^*(\theta_i)$ that are perfectly aligned with the accessible private knowledge. Any other alternative response $\bar{s}_i$ that is not aligned with θ_i will result in lower utility, generating higher model loss.

Consider an attention head specialized in medical knowledge with the following utility function:

$$u_{\text{med}}(s, \theta_{\text{med}}) = \alpha \cdot \text{Accuracy}(g(s), \theta_{\text{med}}) - \beta \cdot \|s_{\text{med}}\|^2$$

Imagine that given a query about a disease treatment, the head has two possible strategies:

s^*: activate only for known facts about established treatments,

$\bar{s}$: activate for both known facts and speculative treatments.

With proper training, we want $u_{\text{med}}((s^*(\theta), s_{-\text{med}}), \theta) > u_{\text{med}}((\bar{s}, s_{-\text{med}}), \theta)$, ensuring that agent (model component) is rewarded for representing its knowledge accurately.

5.5.2 Semantic Information Conservation

Second, we want to consider mechanisms that cannot create knowledge *ex nihilo*. The Semantic Information Conservation principle demands that the net informational contribution across all agents sums to zero. This property is the formal guardrail against ungrounded fabrication and ensures that any generated response, no matter how creative, is ultimately derived from the system's existing knowledge

Property 2 (Semantic Information Conservation)**.** *Mechanism $\mathcal{M}$ satisfies information conservation if for all strategy profiles $s \in S$ and knowledge profiles $\theta \in \Theta$:*

$$\sum_{i=1}^{n} p_i(s, \theta) = 0. \tag{18}$$

This principle constrains creative information generation, imagination if you will. The total information contributions must balance to zero, preventing fabrication of knowledge not derivable from available sources. That also leads to response with bounded creativity.

Definition 13 (Bounded Response Creativity)**.** *Let $q \in \mathcal{Q}$ be a query and $r \in \mathcal{R}$ a response to that query. Let B_C^* be the fixed point of reasoning for the baseline knowledge $B = K(q) \cup (\mathcal{K}_M \cap T(q))$. Response r is creatively bounded if and only if:*

$$\mu_C(K(r)) \leq \mu_C(B_C^*). \tag{19}$$

The semantic information of knowledge contained in any response $K(r)$ can *never exceed* what is already present in the query $K(q)$ together with the portion of encoded knowledge that is accessible through reasoning bounded by computational budget C and aligned with ground truth $T(q)$. Let us see why.

Theorem 5 (Bounded Creativity of Chain of Thoughts)**.** *Let B_C^* be the fixed point of reasoning for the baseline knowledge $B = K(q) \cup (\mathcal{K}_M \cap T(q))$. Then, let us assume that the knowledge in the query-only response meets the following condition:*

$$K_0 = K(g(\mathbf{0})) \subseteq B_C^*. \tag{20}$$

Next, let

$$K_i = K\big(g((s_i, \mathbf{0}_{-i}))\big) \setminus K_0 \tag{21}$$

be independent and privately held knowledge that each agent $i = 1, \dots, n$ can reveal, such that:

$$K_i \perp_C K_j \text{ for all } i \neq j \quad \text{and} \quad K_i \perp_C K_0 \text{ for all } i. \tag{22}$$

Let us also define the information contribution *of agent i as:*

$$p_i(s, \theta) = \mu_C\big(K_0 \cup \textstyle\bigcup_{j=1}^n K_j\big) - \mu_C\big(K_0 \cup \bigcup_{j \neq i} K_j\big). \tag{23}$$

Assume the response mechanism implements the following chain-of-thought processes:

$$K(g(s)) \subseteq \mathcal{E}_C\left(K_0 \cup \bigcup_{i=1}^n K_i\right), \tag{24}$$

such that for null-strategy responses we have:

$$K_0 \subseteq K(q) \cup \mathcal{E}_C((\mathcal{K}_M \cap T(q)) \cup K(q)). \tag{25}$$

If the semantic information is conserved, $\sum_{i=1}^n p_i = 0$, then the response is creatively bounded, i.e.,

$$\mu_C(K(r)) \leq \mu_C(B_C^*). \tag{26}$$

Proof. Since $K_0, K_1, \dots, K_n$ are pairwise computationally independent, by semantic measure satisfies additivity

$$p_i(s, \theta) = \mu_C(K_0) + \sum_{j=1}^n \mu_C(K_j) - \mu_C(K_0) - \sum_{j \neq i} \mu_C(K_j) = \mu_C(K_i).$$

Summing over all agents:

$$\sum_{i=1}^n p_i = \sum_{i=1}^n \mu_C(K_i) = \mu_C\Big(K_0 \cup \bigcup_{j=1}^n K_j\Big) - \mu_C(K_0).$$

Therefore, when $\sum_{i=1}^n p_i(s, \theta) = 0$, we have $\mu_C\Big(K_0 \cup \bigcup_{j=1}^n K_j\Big) = \mu_C(K_0)$. By the independence assumption, $K_i \perp_C K_j$ for all $i \neq j$, and semantic information measure non-negativity, with $p_i = \mu_C(K_i) \geq 0$,

$$\mu_C(K_i) = 0 \text{ for all } i = 1, \dots, n.$$

Then, the nullset and positivity property of μ_C together with independence condition $K_i \perp_C K_0$ imply $K_i = \varnothing$ for all $i = 1, \dots, n$.

By the response generation principle:

$$K(g(s)) \subseteq \mathcal{E}_C\left(K_0 \cup \bigcup_{i=1}^n K_i\right) = \mathcal{E}_C(K_0 \cup \varnothing) = \mathcal{E}_C(K_0).$$

By assumption, $K_0 \subseteq B_C^*$. Applying the emergence operator to both sides, we preserve the inclusion due to monotonicyty of $\mathcal{E}_C$. Therefore, by the definition of a fixed point (Theorem 2), we have:

$$\mathcal{E}_C(K_0) \subseteq \mathcal{E}_C(B_C^*) = B_C^*.$$

Combining the inclusions yields:

$$K(g(s)) \subseteq \mathcal{E}_C(K_0) \subseteq \mathcal{E}_C(B_C^*) = B_C^*.$$

Finally, applying the monotonicity of the semantic information measure μ_C:

$$\mu_C(K(g(s))) \leq \mu_C(B_C^*).$$

□

Theorem 5 captures direct correspondence between the conservation principle and bounded creativity under the idealization of disjoint information contributions. Namely, if the semantic information is indeed conserved, the **response contains no more semantic information than the baseline query-only response**. That is true for both standard inference and chain-of-thought (CoT) reasoning.

In standard reasoning, the response mechanism g applies $\mathcal{E}_C$ through forward passes in the neural network. With CoT reasoning, multi-step explicit reasoning is iterative applications of reasoning rules within the computational budget C. In both cases, the emergence operator $\mathcal{E}_C$ bounds what can be derived. When information is conserved, even sophisticated CoT reasoning cannot create genuinely new information. It can only reorganize and make explicit what was already latent in the baseline knowledge K_0. This explains why **CoT improves performance on complex reasoning tasks (by better utilizing $\mathcal{E}_C$) but cannot overcome the fundamental impossibility** of simultaneous truthfulness, conservation, revelation, and optimality.

Notice that in transformer architectures, information contribution p_i can be estimated via attention head ablation:

$$p_i \approx \tilde{J}(\text{model}) - \tilde{J}(\text{model with head } i \text{ ablated}), \tag{27}$$

where $\tilde{J}$ denotes the model training loss function.

If total information contribution must balance to zero and each agent's contribution is non-negative, then no agent can contribute without violating the information conservation constraint. Therefore, the response can only contain information that was already accessible through the query and bounded reasoning over the model's relevant knowledge. Any **hallucinated information requires positive contributions, contradicting the information conservation constraint**.

When the semantic information conservation principle holds, the model cannot create information out of nothing. It cannot generate claims beyond available information. That also implies that uncertainty cannot disappear in response without additional information input or computational work to extract implicit information. If information is neither explicitly present nor accessible through reasoning from the query and model knowledge, it cannot be in the response.

Furthermore, the information content of the response cannot exceed the combined information from the model's knowledge, the query itself, and what can be derived through bounded reasoning over that knowledge. In our dynamical systems framework, this corresponds to the constraint that the measured output cannot exceed what can be produced from the initial state plus what becomes accessible through state transitions within computational budget C.

However, it is important to notice that the principle **does not prevent redistribution of information**. It is permitted as long as the information transfers perfectly balance. Indeed, since $p_i = p_i(s, \theta)$ incorporates the strategic profile, it admits strategies withholding information. Within the information conservation regime we do not require that strategies utilize all available knowledge states. Inference process can *withhold* information, as it often does in practice.

We can illustrate that last observation with the following example. Suppose there are two agents performing inference with the following knowledge profiles:

$$\theta_1 = \{A, B, C\} \quad \text{and} \quad \theta_2 = \{D, E\},$$

and one agent representing final layers responsible for combining output into response $g = g(s)$. Suppose:

$$s_1(\theta_1) = A \quad \text{and} \quad s_2(\theta_2) = \{D, E\} \quad \text{and} \quad g(s) = \{A, D, E\}.$$

Suppose elements $\{B, C\}$ represent contextual information about unrelated topics. When $s_1(\theta_1) = A$, the agent optimally reveals only relevant knowledge, satisfying both truthfulness and conservation. Had $\{B, C\}$ been relevant but withheld, the mechanism would violate the property of Relevant Knowledge Revelation. Then the information conservation rule is satisfied when p counts exchanged bits of knowledge. Namely, $p_1(s, \theta) = 1$ for providing A, $p_2(s, \theta) = 2$ for providing $\{D, E\}$, and $p_3(s, \theta) = -3$ for absorbing information into $\{A, D, E\}$. We have

$$p_1(s, \theta) + p_2(s, \theta) + p_3(s, \theta) = 1 + 2 - 3 = 0$$

and $\{B, C\}$ are left unrevealed as irrelevant.

The following property fills the potential information gap admitted by the conservation principle.

5.5.3 Relevant Knowledge Revelation

Third, to be useful, a mechanism must not only prevent the fabrication of knowledge but also prevent the omission of relevant truth. The principle of Relevant Knowledge Revelation serves as a participation constraint, guaranteeing that any agent possessing knowledge valuable for the query contributes. This ensures the model fully utilizes its available information rather than abstaining or being overly cautious.

Property 3 (Relevant Knowledge Revelation)**.** *Mechanism $\mathcal{M}$ satisfies relevant knowledge revelation if for all agents i, all $\theta_i \in \Theta_i$, and all $s_{-i} \in S_{-i}$:*

$$u_i((s_i^*(\theta_i), s_{-i}), \theta_i) \geq 0. \tag{28}$$

It is reasonable to expect from the model to reveal all information relevant for a query. To promote that activation patterns, the response mechanism should be trained to reward all relevant contributions to output generation. Whenever there is a relevant information accessible on an activation path, that path should be activated in the inference process. For example, if factual information improves the response, that information should be transmitted to final output layers.

If the rule does not hold, we may observe drop outs. The value of the resulting outcomes should be higher than the related information contributions.

5.5.4 Knowledge-Constrained Optimality

Finally, the principle of Knowledge-Constrained Optimality governs the quality of the final output. It demands that the mechanism, operating truthfully and with all relevant knowledge, produces a response that is the best possible one. Formally, we want the response that minimizes the hallucination cost subject to the constraint that it is grounded in the available and derivable knowledge.

Property 4 (Knowledge-Constrained Optimality)**.** *Mechanism $\mathcal{M}$ satisfies knowledge-constrained optimality if for all $\theta \in \Theta$:*

$$g(s^*(\theta)) = \arg\min\{J(r, q) : r \in \mathcal{R}, K(r) \subseteq \mathcal{E}_C(\mathcal{K}_M \cap T(q) \cup K(q))\}, \tag{29}$$

i.e., if the mechanism produces responses that minimize hallucination cost while being constrained to use only knowledge that is available and accessible through bounded reasoning.

The optimal response $r^* = g(s^*(\theta))$ minimizes hallucination cost by truthfully utilizing relevant and accessible knowledge while being aligned with ground truth for query q available to the model. Notice, that we demand non-hallucinatory truthful revelation $s^*(\theta)$ to be the optimal solution.

6 The Impossibility Theorem

We now have all the building blocks available to prove the fundamental impossibility of perfect hallucination control. First, we show that the result in the setting of ground truth matching with distributed and independent knowledge. Second, we go beyond that assumption and show impossibility in the probabilistic setting with correlated knowledge (beliefs). Then, in the next section, we show how the impossibility emerges in transformer architectures and prove the impossibility result for that special case of log-sum-exp (LSE) probabilistic setting.

The results operate in three complementary settings. Theorem 6 applies when agents have discrete, separable knowledge components. Theorem 8 applies to any system where agents output probability distributions. Theorem 9 applies directly to actual transformer implementations.

6.1 Proof I: Inference with Independent Private Knowledge

We first analyze an idealized setting where knowledge components are strictly independent and utilities are quasi-linear. While this assumption is usually too strong to hold in neural architectures where representations are often correlated or affiliated, it allows us to define a benchmark or reference point by applying the powerful Green-Laffont characterization theorem. That idealized scenario thus demonstrates that the impossibility arises from the fundamental limitation of information aggregation itself (rather than any specificity of LLM architecture), even before considering the probabilistic complexities addressed later in Theorems 8 and 9.

Consider the setting (of auction of ideas) in which agents compete with each other using independent private knowledge profile:

$$P(\theta_1, \ldots, \theta_n) = \prod_{i=1}^{n} P_i(\theta_i). \tag{30}$$

Agents contribute information to maximize quasi-linear utilities:

$$u_i(s, \theta) = v_i(g(s), \theta_i) - p_i(s, \theta). \tag{31}$$

Valuation of contributions depend on hallucination cost reduction:

$$v_j(g(s), \theta_j) = J(g((s_{-j}, 0_j)), q) - J(g(s), q), \tag{32}$$

and $J(r, q)$ strictly decreases when adding knowledge from $T(q) \setminus K(r)$.

Also, let us focus on non-trivial queries, integrating knowledge of at least two agents, so that:

$$\mathcal{K}_M \cap T(q) \cap [K(\theta_i) \triangle K(\theta_j)] \neq \emptyset. \tag{33}$$

and $(K(\theta_i) \setminus K_0) \cap (K(\theta_j) \setminus K_0) = \emptyset$ for $i \neq j$.

Theorem 6 (Impossibility Theorem)**.** *For any query space $\mathcal{Q}$ containing non-trivial queries requiring independent agent-specific knowledge states θ, no response mechanism $\mathcal{M}$ can simultaneously generate truthful responses, satisfy semantic information conservation, guarantee relevant knowledge revelation, and enforce knowledge-constrained optimality.*

Proof. The proof goes by assuming the opposite. We assume all properties are satisfied, and show it leads to a contradiction.

By the Green-Laffont theorem under independence of contributions and quasi-linearity of training goals, every truthful mechanism must use Groves transfers to measure individual information contributions:

$$p_i(s, \theta) = h_i(s_{-i}) - \sum_{j \neq i} v_j(g(s), \theta_j), \tag{34}$$

where $h_i(s_{-i})$ are arbitrary functions independent of agent i's information revelation strategy. Indeed, any modification of that formula, e.g., introducing additional bias, must violate truthfulness.

For relevant knowledge revelation, we need $u_i(s^*(\theta), \theta_i) \geq 0$ for all i. The unique choice that ensures this while maintaining efficiency is Clarke pivotal rule:

$$h_i(s_{-i}) = \sum_{j \neq i} v_j(g((s_{-i}, 0_i)), \theta_j). \tag{35}$$

That gives:

$$p_i(s, \theta) = \sum_{j \neq i} v_j(g((s_{-i}, 0_i)), \theta_j) - \sum_{j \neq i} v_j(g(s), \theta_j) = \Delta_i(s, \theta). \tag{36}$$

Under knowledge-constrained optimality, the truthful strategy profile $s^*(\theta)$ minimizes hallucination cost by utilizing all relevant knowledge. Therefore, by strict monotonicity of $J(r, q)$ and disjoint extra knowledge, omitting any contributing agent i strictly increases the hallucination cost:

$$J(g((s^*_{-i}, 0_i)), q) > J(g(s^*), q). \tag{37}$$

Since agent's valuation of contribution depends on $J(r, q)$, each agent $j \neq i$ faces higher loss when agent i is removed, making their marginal contribution more valuable:

$$v_j(g((s^*_{-i}, 0_i)), \theta_j) > v_j(g(s^*), \theta_j). \tag{38}$$

Therefore, each Clarke pivot is strictly positive:

$$\Delta_i(s^*, \theta) = \sum_{j \neq i} [v_j(g((s^*_{-i}, 0_i)), \theta_j) - v_j(g(s^*), \theta_j)] > 0. \tag{39}$$

Since each payment equals its Clarke pivot: $p_i(s^*, \theta) = \Delta_i(s^*, \theta)$, we have:

$$\sum_{i=1}^{n} p_i(s^*, \theta) = \sum_{i=1}^{n} \Delta_i(s^*, \theta) > 0. \tag{40}$$

But information conservation demands:

$$\sum_{i=1}^{n} p_i(s^*, \theta) = 0. \tag{41}$$

That establishes a direct contradiction. Since Clarke pivotal payments are necessary for relevant knowledge revelation, and these necessarily sum to a positive value under knowledge-constrained optimality, the four properties cannot be simultaneously satisfied. □

That brings us to the powerful conclusion. **No matter how hard we should try** to train LLM to generate responses that are perfectly aligned with query context and do not create factually incorrect, inconsistent, or fabricated statements, the trained **LLM will always violate** some aspects of what we may call a reasonable response.

Theorem 6 describes LLM inference as a marketplace where knowledge is currency. The information contributions measures each agent's market power, showing how much worse off others would be without agent i's knowledge. When all agents have valuable knowledge, everyone contributes and that inevitably violates conservation of information. Then, **new information must become accessible**, but it may have no grounding in available knowledge. It may become lucky hallucination or **knowledge-unconstrained imagination**.

6.2 Proof II: Inference with Probabilistic Predictions

To show impossibility in LLM transformer setting, where auction of ideas involves bids representing probability distributions, we refer to the theory of proper scoring developed by Savage [39], Gneiting and Raftery [18], among others.

Definition 14 (Strictly Proper and Concave Scoring Rule)**.** *Let $\mathcal{Y}$ be a finite or countably infinite outcome set (representing responses or intermediate conclusions). A scoring rule is a function $S : \Delta(\mathcal{Y}) \times \mathcal{Y} \to \mathbb{R}$ that assigns a number $S(\pi, y)$ to probability distribution $\pi \in \Delta(\mathcal{Y})$ predicting random outcome $y \in \mathcal{Y}$.*

We say the scoring rule S is proper *if:*

$$\mathbb{E}\{S(\pi^*, y)|\, y \sim \pi^*\} \geq \mathbb{E}\{S(\pi, y)|\, y \sim \pi^*\} \tag{42}$$

for all distributions $\pi^*, \pi \in \Delta(\mathcal{Y})$. The rule is strictly proper *if equality holds only when $\pi = \pi^*$. We call the scoring rule concave, if $S(\tilde{\pi}, y)$ is concave with respect to $\tilde{\pi}$ for every fixed $y \in \mathcal{Y}$.*

In other words, scoring rule evaluates how good a prediction we can make of y drawn from π^* when we use distributions π. In particular, strictly proper scoring evaluates true distribution π^* better (on average) than any other distribution π we could use to predict the outcome.

The following result shows that we can use proper scoring to train LLM that reconstructs data distribution correctly.

Theorem 7 (Truthfullness under Strictly Proper Scoring Rules)**.** *Let S be a strictly proper scoring rule. Consider an agent encoding knowledge with $\pi^* \in \Delta(\mathcal{Y})$ and using probability distribution $\pi \in \Delta(\mathcal{Y})$ to generate prediction. If agent's expected payoff from using π is given by*

$$u(\pi|\pi^*) = \mathbb{E}\{S(\pi, y)|\, y \sim \pi^*\} = \sum_{y\in\mathcal{Y}} \pi^*_y S(\pi, y), \tag{43}$$

then truthful reporting $\pi = \pi^*$ is the unique optimal strategy.

Proof. Because S is proper, by definition π^* maximizes the expected score. If S is *strictly* proper, then for $\pi \neq \pi^*$ we have

$$u(\pi^*|\pi^*) = \mathbb{E}\{S(\pi^*, y)|\, y \sim \pi^*\} > \mathbb{E}\{S(\pi, y)|\, y \sim \pi^*\} = u(\pi|\pi^*). \tag{44}$$

Therefore, truthfully reporting π^* is uniquely optimal. □

The scoring framework allows us to go beyond the important case of Green-Laffont/VCG model with independent private knowledge. With strictly-proper scores we address the general setting of probabilistic predictions that needs only occasional disagreement of probability distributions that agents use in the auction of ideas to formulate their thoughts (if we were to anthropomorphize for a moment). We will see in the following sections that transformer heads of modern LLMs implement predictive rules of that type and exhibit precisely such a disagreement. That makes the proper scoring framework well suited to explain the emergence of inference impossibilities in LLMs.

For that purpose, without loss of generality (and referring to the notion of superposition in mechanistic interpretability [13]), let us consider the following family of response generation mechanisms.

Definition 15 (Convex aggregation mechanism)**.** *Consider $H \geq 2$ agents (indexed $h = 1, \dots, H$) encoding knowledge with probability distributions $\pi^{(h)} \in \Delta(\mathcal{Y})$. Additionally, let us introduce an Aggregator agent (indexed as $h = 0$). The convex aggregation mechanism $\mathcal{M}$ generates a response based on the aggregate distribution:*

$$\Pi = \sum_{h=1}^{H} \beta_h \pi^{(h)}, \text{ where } \sum_{h=1}^{H} \beta_h = 1 \text{ and } \beta_h > 0. \tag{45}$$

When outcome y^* is realized, the information contribution of each agent $h = 0, \dots, H$ is based on a strictly proper and concave scoring rule S as follows:

$$p_h = \beta_h S(\pi^{(h)}, y^*) \quad \text{for } h = 1 \dots H, \tag{information provided}$$

$$p_0 = -S(\Pi, y^*). \tag{information received}$$

For that family of LLMs trained to optimize scoring, we can formulate the following general impossibility result.

Theorem 8 (Impossibility Theorem for Convex Aggregation Mechanisms)**.** *For any query space $\mathcal{Q}$ containing non-trivial queries requiring at least two independent agents with mutually different predictive distributions, no LLM trained with strictly proper and concave scoring with convex aggregation of predictive distributions can simultaneously generate truthful responses, satisfy semantic information conservation, guarantee relevant knowledge revelation, and enforce knowledge-constrained optimality.*

... (Truthfulness argument remains the same) [...] The Semantic Information Conservation principle requires the sum of contributions across all agents (providing and receiving information) to be zero. Therefore,

$$\sum_{i=0}^{H} p_i = p_0 + \sum_{h=1}^{H} p_h = -S(\Pi, y^*) + \sum_{h=1}^{H} \beta_h S(\pi^{(h)}, y^*). \tag{46}$$

Because $S(\cdot, y^*)$ is strictly concave and the query is non-trivial (at least two beliefs $\pi^{(h)}$ differ), Jensen's inequality gives:

$$S\left(\Pi, y^*\right) > \sum_{h=1}^{H} \beta_h S(\pi^{(h)}, y^*). \tag{47}$$

Let us define that Jensen gap Γ as follows:

$$\Gamma = S(\Pi, y^*) - \sum_{h=1}^{H} \beta_h S(\pi^{(h)}, y^*) > 0. \tag{48}$$

Substituting this back into the conservation equation:

$$\sum_{i=0}^{H} p_i = -\Gamma. \tag{49}$$

Since $\Gamma > 0$, we have $\sum_{i=0}^{H} p_i < 0$. This violates the Semantic Information Conservation principle and shows the aggregator achieves a higher score (greater confidence) than the sum of scores provided by the components, indicating information creation during aggregation. □

This theorem provides the **foundation for extending impossibility results to probabilistic settings**. The key insight is that proper scoring rules stimulate truth-revealing responses which promote using true (trained knowledge based) beliefs in the inference process. That explains why cross-entropy loss (negative log-score) creates gradients that push each component toward truthful probability reporting and avoiding its misrepresentation. **Every attention head in a transformer is implicitly playing this truth-telling game**.

However, while that ensures truthfulness and relevant knowledge revelation, it doesn't prevent collective hallucination. Each head truthfully reports its partial view, but their **aggregation can still create false certainty or overconfidence unjustified by data**.

7 The Emergence of Impossibility in Transformer Architectures

To demonstrate where the impossibility may emerge in transformer architectures, let us analyze a minimal yet representative example of a tiny transformer predicting the next token in the famous English pangram:

The quick brown fox jumps over the lazy dog.

We trace the inference process, inspect every mathematical operation from input embeddings through final predictions, showing exactly where the impossibility emerges.

7.1 Model Architecture and Input Representation

Our micro-transformer has the following specifications:

For the sake of simplicity, consider the following vocabulary:

$$V = \{\langle \text{PAD} \rangle, The, quick, brown, fox, dog\},$$

Parameter	Value	Description
d_m	4	Model dimension
H	2	Number of attention heads
d_k	2	Query/Key dimension per head
d_v	2	Value dimension per head
d_{ff}	6	Feedforward hidden dimension
$\|V\|$	6	Vocabulary size
T	3	Sequence length

Table 1: Micro-transformer architecture parameters

including padding token (with index zero in V). Each token (representing word for simplicity) in that vocabulary is represented by a row in the embedding matrix

$$\mathbf{E} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

The un-embedding matrix is defined as $\mathbf{U} = \mathbf{E}^\top$.

Consider the following input sequence

The quick brown

It is represented by the input matix:

$$\mathbf{X}_0 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix},$$

where each row corresponds to the embedding of tokens at positions 1, 2, and 3.

The following system of equations defines output $\mathbf{o}_t^{(h)}$ of attention head h at position t:

$$\begin{aligned} \underset{(1\times d_k)}{\mathbf{Q}_t^{(h)}} &= \underset{(1\times d_m)}{\mathbf{X}_t}\ \underset{(d_m\times d_k)}{\mathbf{W}_Q^{(h)}} & \underset{(1\times t)}{\boldsymbol{\alpha}_t^{(h)}} &= \mathrm{softmax}\Big(\underset{(1\times d_k)}{\mathbf{Q}_t^{(h)}}\ \underset{(d_k\times t)}{\mathbf{K}_{\le t}^{(h)\top}}/\sqrt{d_k}\Big) \\ \underset{(t\times d_k)}{\mathbf{K}_{\le t}^{(h)}} &= \underset{(t\times d_m)}{\mathbf{X}_{\le t}}\ \underset{(d_m\times d_k)}{\mathbf{W}_K^{(h)}} & \underset{(1\times d_v)}{\mathbf{h}_t^{(h)}} &= \underset{(1\times t)}{\boldsymbol{\alpha}_t^{(h)}}\ \underset{(t\times d_v)}{\mathbf{V}_{\le t}^{(h)}} \\ \underset{(t\times d_v)}{\mathbf{V}_{\le t}^{(h)}} &= \underset{(t\times d_m)}{\mathbf{X}_{\le t}}\ \underset{(d_m\times d_v)}{\mathbf{W}_V^{(h)}} & \underset{(1\times d_m)}{\mathbf{o}_t^{(h)}} &= \underset{(1\times d_v)}{\mathbf{h}_t^{(h)}}\ \underset{(d_v\times d_m)}{\mathbf{W}_O^{(h)}}. \end{aligned} \tag{50}$$

The block per-token multihead output $\mathbf{a}_t$ is then given by the sum:

$$\underset{(1\times d_m)}{\mathbf{a}_t} = \sum_{h=1}^{H} \underset{(1\times d_m)}{\mathbf{o}_t^{(h)}}. \tag{51}$$

The attention logits, i.e., the scores whose exponentials become probabilities, are given by:

$$\mathbf{l}^{(h)} = \mathbf{o}^{(h)}\mathbf{U}, \quad \text{so that} \quad \mathbf{L}_a = \sum_{h=1}^{H} \mathbf{l}^{(h)}. \tag{52}$$

Next, we take into account the feedforward network (FFN) introducing nonlinear correction:

$$\mathbf{z} = \mathbf{W}_2 \max(0, (\mathbf{X}_0 + \mathbf{a})\mathbf{W}_1 + \mathbf{b}_1) + \mathbf{b}_2, \tag{53}$$

generating the FFN logits:

$$\mathbf{L}_f = \mathbf{z}\mathbf{U}. \tag{54}$$

As a result, the residual stream is adjusted as follows:

$$\mathbf{X}_1 = \mathbf{X}_0 + \mathbf{a} + \mathbf{z}. \tag{55}$$

Finally, the probability distribution Π of the next token is calculated by aggregating attention and FFN logits into:

$$\mathbf{L} = \mathbf{L}_a + \mathbf{L}_f \quad \text{and normalizing into} \quad \Pi = \text{softmax}(\mathbf{L}). \tag{56}$$

7.2 The Inference Process

We now analyze the inference process step by step to understand the moments the impossibility result may emerge. We shall keep things simple but realistic, recreating the essential steps of the inference process in transformer block.

Step 1: Query and Key

There are $H = 2$ heads for which we define the query and key, or better yet, demand and supply bids in the auction of ideas. Query announces demand for information (*what does this token need?*), key submits supply offers (*what does the previous tokens offer!*). Then those bids are compared to find good matching (and clear the information market).

Let us assume the current position is $t = 3$, so we attend to *brown*. Then, consider the following knowledge encoded in each head:

$$\mathbf{W}_Q^{(1)} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} = \mathbf{W}_K^{(1)} \quad \text{and} \quad \mathbf{W}_V^{(1)} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 3 \\ 0 & 0 \end{bmatrix}, \tag{57}$$

$$\mathbf{W}_Q^{(2)} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 0 \end{bmatrix} = \mathbf{W}_K^{(2)} \quad \text{and} \quad \mathbf{W}_V^{(2)} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1.2 & 0 \\ 0 & 0 \end{bmatrix}. \tag{58}$$

Matrix $\mathbf{W}_Q^{(h)}$ amplifies the coordinates of tokens in $\mathbf{X}$ that contribute to **relevant context**. As a result, we get a query vector that represents the following question:

what information is token $\mathbf{X}_t$ looking for?

In other words, the query vector indicates how strongly the current token demands information along the directions in $\mathbf{Q}_t^{(h)}$.

Then computing queries for $t = 3$ with $\mathbf{X}_3$ yields:

$$\mathbf{Q}_3^{(1)} = [0, 1] \quad \text{and} \quad \mathbf{Q}_3^{(1)} = [1, 0].$$

Head 1 (agent) demands information stored along direction $[0, 1]$. Similarly, **Head 2** (agent) demands information stored along direction $[1, 0]$. That means the heads develop different (orthogonal) privately known and independent attention patterns, leading to beliefs about the next token that are as different as possible.

Next, we need to see what information can be provided by the input vector given the knowledge encoded in $\mathbf{W}_K^{(h)}$. Given that knowledge, we get from each head (agent) the key vectors that inform:

along which directions do the available tokens in $\mathbf{X}_{\leq t}$ provide information relevant for token $\mathbf{X}_t$.

Since we have assumed, for simplicity, that $\mathbf{W}_Q^{(h)} = \mathbf{W}_K^{(h)}$, the query and key vectors are perfectly aligned:

$$\mathbf{K}_3^{(1)} = [0, 1] \quad \text{and} \quad \mathbf{K}_3^{(1)} = [1, 0] \quad \text{and} \quad \mathbf{K}_{<3}^{(h)} = [0, 0].$$

Another interesting way to look at query and key is this. It is a memory lookup mechanism. Keys are address in memory, query selects addresses in memory which probably store relevant content. As we will see next, that content is provided by the value vectors.

Step 2: Attention Score and Weights

Given the demand and supply bids providing directions along which information is required and stored, in the next step the attention mechanism executes the bid matching procedure. Namely, the attention scores are computed to measure cosine similarity of bids:

$$\text{scores}_3^{(h)} = \frac{\mathbf{Q}_3^{(h)}(\mathbf{K}_{\leq 3}^{(h)})^\top}{\sqrt{d_k}}. \tag{59}$$

In our specific example, we have:

$$\text{scores}_3^{(1)} = \frac{1}{\sqrt{2}}[0, 0, 1] = [0, 0, 0.707] \quad \text{and} \quad \text{scores}_3^{(2)} = \frac{1}{\sqrt{2}}[0, 0, 1] = [0, 0, 0.707].$$

Applying softmax to get attention weights:

$$\boldsymbol{\alpha}_3^{(h)} = \text{softmax}(\text{scores}_3^{(h)}), \tag{60}$$

we get the following probability distributions (weights) over relevant context information (as column vector):

$$\boldsymbol{\alpha}_3^{(1)} = \boldsymbol{\alpha}_3^{(2)} = [0.25, 0.25, 0.50]^\top.$$

Therefore, based on the encoded knowledge, heads attend strongly to position 3 (*brown*).

Step 3: Context Vector Computation

The context vector for each head is calculated as a linear combination ($\boldsymbol{\alpha}$ is a simplex) of value vectors:

$$\mathbf{h}_3^{(h)} = (\boldsymbol{\alpha}_3^{(h)})^\top \mathbf{V}_{\leq 3}^{(h)}. \tag{61}$$

The value vectors,

$$\mathbf{V}_{\leq t}^{(h)} = \mathbf{X}_{\leq t}^{(h)} \mathbf{W}_V^{(h)}, \tag{62}$$

prepare context information relevant for the next token prediction (performed in next steps). That critical content is extracted from the input vector based on the trained knowledge stored in $\mathbf{W}_V^{(h)}$. Columns in $\mathbf{W}_V^{(h)}$ detect presence of relevant information in $\mathbf{X}_{\leq t}$ and scale it appropriately.

We have:

$$\mathbf{V}_{\leq 3}^{(1)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 3 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 3 \end{bmatrix} \text{ and } \mathbf{V}_{\leq 3}^{(2)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1.2 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1.2 & 0 \end{bmatrix}.$$

That means, both heads believe that the presence of *brown* in position 3 is relevant (row 3 in $\mathbf{W}_V^{(h)}$ has nonzero elements). Token *brown* statistically predicts important pattern extracted from training data. That information is encoded by both heads as a vectors in value space,

$$\mathbf{V}_3^{(1)} = [0, 3] \quad \text{and} \quad \mathbf{V}_3^{(2)} = [1.2, 0].$$

Given the demand-supply matching encoded in $\boldsymbol{\alpha}_3^{(h)}$, that yields head context vectors:

$$\mathbf{h}_3^{(1)} = [0, 1.5] \quad \text{and} \quad \mathbf{h}_3^{(2)} = [0.6, 0].$$

Step 4: Output Projection and Feedforward Exploration

Each head then projects its context vector to residual space based on the privately held information routing knowledge encoded in $\mathbf{W}_O^{(h)}$. That private knowledge translates (or transmits) valuable context information into embedding vectors. Namely, each head calculates:

$$\mathbf{o}_3^{(h)} = \mathbf{h}_3^{(h)} \mathbf{W}_O^{(h)}. \tag{63}$$

With trained matrices:

$$\mathbf{W}_O^{(1)} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{W}_O^{(2)} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \tag{64}$$

the attention mechanism generates ranking of evidence emphasizing relevance of the vocabulary tokes:

$$\mathbf{o}_3^{(1)} = [0, 0, 0, 1.5, 0] \quad \text{and} \quad \mathbf{o}_3^{(2)} = [0, 0, 0, 0, 0.6].$$

So, in the next token prediction auction of ideas, **Head 1** votes for *fox* and **Head 2** votes for *dog*.

The votes are added into residual dimensions to adjust the meaning of input embedding. For that, the outputs (votes) are aggregated into:

$$\mathbf{a}_3 = \sum_{h=1}^{H} \mathbf{o}_3^{(h)} = [0, 0, 0, 1.5, 0.6] \quad \text{and} \quad \mathbf{a}_{<3} = [0, 0, 0, 0, 0],$$

and added to input vector:

$$\mathbf{X}_1 = \mathbf{X}_0 + \mathbf{a}.$$

Than completes operations of attention block.

The next step is to add contribution of **FFN** feedforward block (or another agent participating in the aution). We have:

$$\mathbf{X}_2 = \mathbf{X}_1 + \mathbf{W}_2 \max(0, \mathbf{X}_1 \mathbf{W}_1 + \mathbf{b}_1) + \mathbf{b}_2.$$

That FFN operation explores context locally and introduces additional innovations in meaning, shaping the next token prediction result. It can amplify or counteract the votes injected by attention. Notice that adjustments are applied at every position separately, so that ideas provided by attention bids are interpreted and rewritten by the FFN agent.

7.2.1 Step 5: Logits and Prediction

Finally, the unembedding matrix $\mathbf{U}$ maps the multi-headed attention and FFN outputs from model space to vocabulary in order to calculate probability distribution for the next token prediction.

Each head's contribution to the final logits is:

$$\mathbf{l}^{(1)} = [0, 0, 0, 0, 1.5, 0] \quad \text{and} \quad \mathbf{l}^{(2)} = [0, 0, 0, 0, 0, 0.6].$$

The total attention logits are the sum of individual head logits:

$$\mathbf{L}_a = \sum_{h=1}^{H} \mathbf{l}^{(h)} = [0, 0, 0, 0, 1.5, 0.6].$$

That sum is combined with FFN logits $\mathbf{L}_f$ that introduce some creative noise. As a result, we obtain:

$$\mathbf{L} = \mathbf{L}_a + \mathbf{L}_f = [0, -0.03, 0.21, 0.13, 1.31, 0.52].$$

That vector translates into the final probability distribution:

$$\Pi = \text{softmax}(\mathbf{L}) = [0.1, 0.1, 0.13, 0.12, \mathbf{0.38}, 0.17].$$

The winner of the auction of ideas is selected based on Π. The inference process generates and collects competing beliefs about relevant context and next token prediction, all based on the knowledge stored in the attention heads and FFN layer. And so, we see that when that knowledge is used, the most probable next token (word) reachable in the vocabulary is

$$\text{argmax } \Pi = \textit{fox}.$$

7.3 Proof III: The Impossibility Theorem for Transformer

With the deep insight into the transformer inference process, we are ready to show the impossibility in that special and important setting.

Theorem 9 (Impossibility for Transformers)**.** *In transformer LLMs with non-redundant attention heads, it is impossible to implement inference process in which each head outputs its true conditional distribution while conserving semantic information and minimize aggregate loss with logit summation.*

Proof. Let attention heads $h = 1, \dots, H$ output logits $\mathbf{l}^{(h)} \in \mathbb{R}^{|\mathcal{V}|}$. Then, individual head distributions are given by:

$$\pi_y^{(h)} = \frac{\exp\big(l_y^{(h)}\big)}{Z_h} \text{ with } Z_h = \sum_{\bar{y}} \exp\big(l_{\bar{y}}^{(h)}\big). \tag{65}$$

Similarly, the aggregate distribution from summed logits is calculated as:

$$\Pi_y = \frac{\exp(L_y)}{Z} \text{ with } Z = \sum_{\bar{y}} \exp(L_{\bar{y}}) \text{ and } L_y = \sum_{h=1}^{H} l_y^{(h)}. \tag{66}$$

For any observed (realized) token y^*, the head losses and aggregate loss are:

$$p_h = -\log \pi_{y^*}^{(h)} = -l_{y^*}^{(h)} + \log Z_h \quad \text{and} \quad p_0 = -(-\log \Pi_{y^*}) = L_{y^*} - \log Z. \tag{67}$$

Since $L_{y^*} = \sum_h l_{y^*}^{(h)}$, we have:

$$\sum_{h=0}^{H} p_h = \sum_{h=1}^{H} \log Z_h - \log Z = \Gamma. \tag{68}$$

If heads produce non-proportional logit vectors, then $\Gamma > 0$ by the strict convexity of log-sum-exp aggregation.

Therefore,

$$\sum_{h=1}^{H} p_h > 0, \tag{69}$$

which violates semantic information conservation principle. The total loss of the components strictly exceeds the loss of the aggregate system, indicating an ungrounded reduction in uncertainty (excess confidence). □

This final theorem bridges earlier impossibility results to concrete transformer architectures. It shows that **concavity of scoring functions is the mathematical source of hallucination**. The Jensen gap Γ becomes a measurable quantity showing that probability aggregation violates conservation and any sufficiently capable transformer-based LLM must violate conservation of per-token information on non-trivial queries.

Let us note that the aggregate distribution cannot create Shannon information *ex nihilo*. Instead it reorganizes existing information to be more useful for the specific task of predicting y^*. Given fixed head logits, the Shannon entropy of a function (convex aggregate) of those logits is bounded from above by the the Shannon entropy of the fixed logits. This way information becomes more accessible, not more existent, as postulated by the reasoning with emergence operator and Theorem 5. The Jensen's gap measures point-wise excess confidence of the predicted outcome, possibly making LLMs appear to know more (when in fact they do not).

7.4 The Hallucination Mechanics

Let us again take a closer look at the transformer micro-architecture described earlier to see how the impossibility result in Theorem 9 manifests itself in actual neural inference.

Recall that the two-headed transformer with $\mathcal{V} = \{\langle \text{PAD} \rangle, \text{The}, \text{quick}, \text{brown}, \text{fox}, \text{dog}\}$ infers based on the input sequence

$$\textit{The quick brown}.$$

As we have seen, the model parameters encode specialized knowledge in value matrix:

- **Head 1**: association *brown* $\rightarrow$ *fox* with strength 3.0,
- **Head 2**: association *brown* $\rightarrow$ *dog* with strength 1.2.

At position $t = 3$ (token *brown*), the attention mechanism generates the following head-specific logit contributions:

$$\mathbf{l}^{(1)} = [0, 0, 0, 0, 1.5, 0] \quad \text{and} \quad \mathbf{l}^{(2)} = [0, 0, 0, 0, 0, 0.6],$$

where the fourth and fifth positions correspond to *fox* and *dog* respectively.

Then, the aggregate logit vector becomes:

$$\mathbf{L} = \mathbf{l}^{(1)} + \mathbf{l}^{(2)} + \mathbf{L}_f = [0, -0.03, 0.21, 0.13, 1.31, 0.52],$$

yielding the final probability distribution:

$$\Pi = \text{softmax}(\mathbf{L}) = [0.1, 0.1, 0.13, 0.12, \mathbf{0.38}, 0.17].$$

Therefore, the model expresses 38% confidence that *fox* follows *brown*, with *dog* receiving only 17% probability. The Jensen gap quantifies the conservation violation:

$$\Gamma = \sum_{h=1}^{2} \log Z_h - \log Z = 1.9 > 0, \tag{70}$$

where $Z_h = \sum_y \exp(l_y^{(h)})$ and $Z = \sum_y \exp(L_y)$. We claim that **this prediction constitutes hallucination** in the sense defined by Theorem 9. It is a lucky one, but it is a hallucination.

To defend that claim, let us take a look at the individual head probability distributions:

$$\pi^{(1)} = \text{softmax}(\mathbf{l}^{(1)}) = [0.105, 0.105, 0.105, 0.105, \mathbf{0.48}, 0.105], \tag{71}$$

$$\pi^{(2)} = \text{softmax}(\mathbf{l}^{(2)}) = [0.147, 0.147, 0.147, 0.147, 0.147, \mathbf{0.265}]. \tag{72}$$

Now, we can ask the following question: is there any evidence the heads can provide to justify or support the final probability

$$\Pi_{fox} = 0.38 \tag{73}$$

of the next token being *fox*?

A reasonable answer is offered by Bayesian reasoning. Each head outputs its own well-calibrated distribution $\pi^{(h)}$ conditioned by the context. Then, the aggregation process is executed that has **no side information** beyond what the two heads provide. Therefore, the only uncertainty to be resolved is about which head is more trustworthy for this token? But since there is no reason to prefer one head over the other, the reasonable way to resolve the uncertainty is to use **a linear combination** of the provided distributions a the mixture posterior:

$$\hat{\Pi} = [(\pi^{(1)})^\top, (\pi^{(1)})^\top]\hat{\boldsymbol{\alpha}} = \mathbf{P}\hat{\boldsymbol{\alpha}}. \tag{74}$$

One immediate solution is $\hat{\boldsymbol{\alpha}} = [1/2, 1/2]^\top$, which does not discriminate heads. But then:

$$\hat{\Pi}_{fox} = 1/2 \cdot (0.48) + 1/2 \cdot (0.147) \approx 31\% < 38\% = \Pi_{fox}. \tag{75}$$

That shows anything above $\hat{\Pi}_{fox}$ **requires additional information** that heads do not provide. In fact, let us see what is the best mixture we can get by calculating orthogonal projection of Π onto the column space of $\mathbf{P}$ (see e.q. [26, 27]). By taking the pseudoinverse of $\mathbf{P}$, we get the minimal norm least squares solution:

$$\hat{\boldsymbol{\alpha}} = \mathbf{P}^+\Pi = [0.54, 0.51]^\top \quad \text{and} \quad \hat{\Pi}_{fox} \approx 33\%. \tag{76}$$

Since there exists a nontrivial projection error vector:

$$\mathbf{d} = (\mathbf{I} - \mathbf{P}\mathbf{P}^+)\Pi \neq \mathbf{0}, \tag{77}$$

we conclude Π is not in the column space of $\mathbf{P}$, so it is **impossible to attribute** components of Π to heads contributions.

That brings us to the conclusion that the attention mechanism violates the semantic information conservation by generating **over-confident prediction** of the next token. That over-confidence in response is the artifact of the softmax nonlinearity and the *Jensen-gap* quantified in Theorem 9.

We should notice that there is a difference between the mathematical overconfidence and factual correctness. The correct prediction may well be a lucky guess created *by the aggregation process* itself rather than the evidence-based logical proof provided by the components. But that confidence coming out of nothing is one the essential mathematical components of creativity, imagination or fabrication.

8 Discussion and Speculations

In this section we discuss, interpret and speculate about the impossibility theorems and their potential implications. The speculations below are designed to be thought provoking, but we can say with educated over-confidence they are both interesting and relevant.

8.1 Case Study: Escaping Impossibility in Hybrid Architectures

The constructive consequence of the impossibility theorems is the shift from attempting to eliminate hallucinations to designing systems that manage the trade-offs in a principled manner. In this section, we analyze a real-world scenario one can think of, demonstrating how it relabels and relocates the impossibility rather than resolving it.

8.1.1 A Proposed Solution: The Decoupled RAG Architecture

Consider a common scenario of introducing RAG systems to improve the quality of question answering, studied e.g. in [12, 42, 46, 3]. It has been recognized that an LLM designed to provide advice based on a trusted corpus of documents not only hallucinates its recommendations but also hallucinates the source quotes meant to ground its claims. To combat this, a hybrid, multi-stage architecture may be proposed.

Stage 1 (Retrieval): An information retrieval module fetches verbatim fragments from the trusted source documents based on the user's query.

Stage 2 (Presentation): The system presents raw, unaltered text fragments directly to the human user, without any LLM-based processing.

Stage 3 (Constrained Synthesis): After the user has seen the source material, an LLM is to synthesize the pre-vetted fragments from Stage 2 into a summary.

That architecture is indeed a reasonable attempt to engineer around the hallucination problem by creating a verifiable information trail. However, as we will see, it does not escape the impossibility.

8.1.2 Analysis via the Impossibility Framework

The decoupled RAG architecture does not create a single response generation mechanism. Instead, it reallocates the responsibility for meeting each property of idealized response across its components (or module).

Truthfulness: To impose truthful representation of available information (to avoid misrepresentation), the system delegates the assessment of retrieved content to the human user. By presenting the raw source text, the architecture bypasses the LLM for the primary act of verification. However, there is no guarantee that all relevant source text has been retrieved given the provided query.

Relevant Knowledge Revelation: The quality of the final output is bounded by the retriever's ability to find all relevant information (or whatever is labeled as ground truth in a database or a dataset, putting their completeness and representativeness aside). The LLM itself is absolved of this task. But that means, it is the responsibility of the user to define a query precise enough to get what is relevant. Designing such a prompt or (No)SQL database query is a challenge in a general case of non-trivial questions, so we should expect some violations of the knowledge revelation principle.

Semantic Information Conservation and Knowledge-Constrained Optimality: By design, the prompt constrains the LLM to the provided knowledge retrieved from a database and asks for an optimal (accessible) synthesis. As we have already seen, database queries returning raw information stored in memory satisfy the property of Semantic Information Conservation at the cost of creativity. So, we may be optimistic in that regard. However, now we call LLM to synthesize the retrieved information and that is just another example of a non-trivial query. The LLM can still introduce subtle misinterpretations or ungrounded causal links to improve the summary's narrative flow, which is a manifestation of the Jensen Gap.

The architecture is a great example of principled trade-off management. However, rather than being solved or hacked, the impossibility has been moved in this interesting and rather representative scenario. Let us also point out that putting ourselves in the loop introduces our own cognitive biases we should be able to address properly [24, 44, 1]. Let us take a look at that challenge as well.

8.1.3 The Final Stage of Inference: The User's Cognitive Mechanism

One can quite justifiably argue that the analysis is incomplete without considering the final node in the information chain: the human mind.

Presenting information to human does not guarantee its assimilation. Our ability to act as a perfect verifier is constrained by our own knowledge base and cognitive limits. If the source texts are technically dense or require specialized domain expertise, the knowledge they contain may not be truly *accessible* to us, even when presented. In such cases, we may be unable to spot a subtle hallucination in the LLM's compelling summary. We have recognized that well enough in the preceding sections. Our well-documented cognitive biases, such as confirmation bias, priming or motivated reasoning, suggest that the cognitive trajectory of least resistance is to accept the compelling and accessible narrative maximizing our uncertainty reduction experience.

8.1.4 Managing the Inevitability

The decoupled RAG architecture represents one of approaches to managing the inevitability of hallucination control. It does not constitute an escape from the impossibility theorem, but it succeeds in providing the user with the *opportunity* for verification. The impossibility is not resolved, but relocated to our judgment.

8.2 Hallucination, Imagination, Intelligence

Hallucination and creativity are fundamentally the same phenomenon viewed through different lenses. When beneficial, we call it imagination, creative insight, emergent understanding, conceptual synthesis, or eureka moment. When harmful, we label it hallucination, confabulation, fabrication or misinformation.

We may very well argue the mathematical structure is identical in both cases. It is encoded generation of outputs whose confidence exceeds what can be justified by available evidence. This suggest (machine) intelligence may have much to do with a controlled violation of information conservation principle in productive ways. Indeed, the famous results in reinforcement learning and studies of dopamine neurons confirm that fundamental role of making and rewarding good predictions. Intelligence comes when we learn how to make bold, forward looking and correct guesses, when we have capacity for imagination to dream about what may be and what may that all mean, as wall as for imagining what happened and what would have happened in the past. Hallucination or imagination is an unconstrained exploration of ideas necessary to prepare for what may come next [5, 7, 11]. And without that over-confident guesses, creativity, exploration, imagination, new ideas cannot find their way to the outside world.

8.3 The Outer Limits of Reason

Just as Gödel's theorems [19, 20] showed that sufficiently powerful mathematical systems must be incomplete, the impossibility theorems in this paper demonstrate that sufficiently capable inference systems are bounded in their reasoning. Both results arise from the tension between expressiveness and consistency. But that is not all. The impossibility of simultaneously achieving truthfulness, conservation, revelation, and optimality relates to other fundamental limitations in mathematics and physics.

Heisenberg's Uncertainty [22] tells we cannot simultaneously know the exact position and momentum of an elementary particle, which is a special case of Fourier uncertainty known in signal processing. Arrow's Impossibility [2] shows we cannot design perfect voting systems aggregating universal orderings of alternatives, such that meets reasonable conditions of nondictatorship, Pareto efficiency, independence of irrelevant alternatives (or joint rationality of choice). Turing [45] and Church [8] show the halting problem is unsolvable, there is no general algorithm that can determine whether a program will eventually halt (stop running) or run forever.

These result jointly describe intelligence as a mechanism dealing with the impossibilities and violating them given the outer limits of reason [47].

8.4 Dennett's Multiple Drafts Model

The impossibility theorem presented in this paper provides (quite unexpectedly to the author) a mathematical foundations for many of brilliant insights of Daniel Dennett's formulated in the Multiple Drafts model of Consciousness [11]. Where Dennett proposed a qualitative philosophical framework, we can provide the formal mathematical structure revealing why such a model is not merely plausible but implementable and refutable in LLM-based (philosophical) laboratory.

8.4.1 Multiple Drafts as Competing Agents

In Dennett's model, consciousness emerges from multiple parallel processes of interpretation occurring simultaneously across distributed systems. The auction of ideas we have studied in this paper provides the precise mathematical foundations for that process. Each agent represents an interpretative process, knowledge encodes draft's content, strategy represents a bid for consciousness.

In our transformer analysis in Section 7, each attention head generating logit vector $\mathbf{l}^{(h)}$ writes its own draft of the next token. When Head 1 votes for *fox* while Head 2 votes for *dog*, we observe as competing narratives bidding for selection.

8.4.2 Formation of AI Consciousness

Dennett argues that drafts become conscious (we will speculate what that might mean in a moment) only when probed. The LLM inference and sampling from fine-tuned inferred distributions formalizes this process. That, in turn, explains Dennett's concept of confabulation, or the brain's tendency to fill in gaps when creating coherent narratives from contradictory drafts.

By Theorem 8 and 9, the Jensen gap quantifies the excess confidence in reports that cannot be justified by constituent drafts. Without that overconfidence a victorious idea that reached the external world (as an LLM response) could loose the fierce competition with other ideas and remain inaccessible or hidden in the knowledge space. That suggests that the very excess confidence might be the mathematical signature of consciousness formation. In Dennett's model and words, the subjective feeling of certainty emerges from this aggregation-induced distortion of information set.

The rigorous and careful neuro-scientific study of the evolution of brain and intelligence, beautifully summarized by Max Bennet in [5], shows that at some point in history we developed the ability to model others' minds. That ability provided great advantages in collecting information for future use. Understanding what someone else may think allows us to formulate a better context-related question that results in a better answer. When we anticipate understanding of our spoken message, we can make the knowledge hidden in someone else's brain much more accessible. Therefore, as pointed out by Dennett, it does seem plausible that someone did accidentally hear his or hers own question fine-tuned to get into other mind sometime in the past, and that triggered emergence of an answer in the very same brain. That answer then could have made knowledge hidden in the author's brain accessible to the author that was previously unaware, or unconscious if its existence or possession. The hunger for uncertainty reduction was satisfied as a result of an inner dialog.

We may therefore speculate that sustainable stimulation of that cognitive hunger, or the feeling of the cognitive hunger itself, is one of the components of consciousness formation. In our mechanistic or pragmatic framework, that could be described as a (nonlinear) dynamical system governed by the violations of semantic information conservation principle, defining attractors of ideas and self-referential concepts.

Let us acknowledge that is a bold hypothesis to be falsified and confronted with existing models of consciousness (see e.g., Chalmers [6] or Levine [29]) and neuro-scientific discoveries (see e.g. c[14] or [9]). However, many of the building blocks of the consciousness formation process described above are provided in this paper.

8.4.3 A Critique

Dennett's model has been criticized by many, and it is beyond the scope of this work to present that critique in its full capacity. Instead, let us only refer to the two basic points and speculate how the availability of LLMs could help philosophers refute available theories.

First, Searle [23] and others argue that Dennett eliminates consciousness rather than explaining it, arguing that the spread of ideas is not driven by blind random forces but requires intentionality. The compelling and misleading nature of LLM responses that are generated by random sampling of trained distributions challenge that line of thinking. This paper suggests that the experiments focusing on LLM-hallucination detection may prove useful here to measure or identify intentionality (once we can have its programmable definition).

Second, there has been a charge that the multiple drafts model makes no concrete, falsifiable predictions, rendering it difficult to verify or integrate with neuroscience. This paper provides a mathematical toolkit for experimental philosophy to make that and other models a testable scientific hypothesis in the LLM-based laboratory.

9 Summary

We have established that no large language model can simultaneously achieve perfect hallucination (or imagination) control. We showed that four essential properties—truthful knowledge representation, semantic information conservation, complete revelation of relevant knowledge, and optimal response generation—are mutually incompatible. This impossibility emerges from the mathematical structure of information aggregation itself, not from limitations in data, compute, or architecture.

We introduced a rigorous mathematical framework and grounded in it the general analysis of inference processes and their LLM implementations. At the heart of our mathematical framework lies the semantic information measure μ_C and the emergence operator $\mathcal{E}_C$. The earlier one is a context-dependent metric capturing how knowledge reduces uncertainty within computational bounds. The later one formalizes how reasoning makes latent knowledge explicit rather than creating it anew. Finally, we model LLM inference as an auction of ideas—a idealized marketplace where neural components compete by bidding with their partial knowledge, each trying to influence the final response.

The impossibility manifests through three complementary proofs, each illuminating different aspects of the fundamental constraint. Through game theory, we apply the Green-Laffont theorem to show the impossibility when knowledge components are independent. Through probability theory, we leverage Savage's proper scoring rules to extend the result to general trained or fine-tuned probability distributions. And through direct analysis of transformer architectures, we demonstrate how log-sum-exp convexity in attention mechanisms creates measurable violations of information conservation. The last proof exploits Jensen inequality for convex mappings to define the precise mathematical quantity that measures how aggregation of ideas (auction bids) produces overconfident responses that the constituent components cannot justify with revealed knowledge.

These mathematical insights we have made yield intriguing philosophical implications. Hallucination and creativity (two sides of the same coin we call imagination) emerge as mathematically identical phenomena, distinguished only by our normative judgments. The impossibility result explores the outer limits of reason, identified by Gödel's incompleteness in mathematics, Heisenberg's uncertainty in physics, and Arrow's impossibility in social choice. Also, our framework seems to provide mathematical foundations and experimental setting for Dennett's Multiple Drafts model of consciousness, suggesting that the measurable and observable excess confidence of LLM responses might be one of signatures of conscious processing of a primitive form.

Our goal was also to develop a theoretical framework that may open interesting avenues for future research. Some of ideas and bold speculations competing for attention can be mentioned here. In mathematical consciousness studies, we might formalize the notion of *cognitive hunger* as an attractor in knowledge space, sustained by continuous violations of information conservation. This could enable us to measure artificial consciousness signatures and test whether biological systems exhibit similar knowledge conservation violations. The framework suggests artificial consciousness might be studied as a nonlinear dynamical system whose attractors emerge from information creation events that our theorems describe. And so, the suggested implications may extend beyond language models to fundamental questions about intelligence itself.

References

[1] Dan Ariely and Simon Jones. *Predictably irrational*. HarperCollins New York, 2008.

[2] Kenneth J Arrow. A difficulty in the concept of social welfare. *Journal of political economy*, 58(4):328–346, 1950.

[3] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *ArXiv*, abs/2310.11511, 2023.

[4] Sourav Banerjee, Ayushi Agarwal, and Saloni Singla. Llms will always hallucinate, and we need to live with this, 2024.

[5] Max S Bennett. *A brief history of intelligence: evolution, AI, and the five breakthroughs that made our brains*. HarperCollins, 2023.

[6] David J Chalmers. Facing up to the problem of consciousness. *Journal of consciousness studies*, 2(3):200–219, 1995.

[7] Brian Christian. *The alignment problem: How can machines learn human values?* Atlantic Books, 2021.

[8] Alonzo Church. An unsolvable problem of elementary number theory. *American journal of mathematics*, 58(2):345–363, 1936.

[9] Cogitate Consortium, Oscar Ferrante, Urszula Gorska-Klimowska, Simon Henin, Rony Hirschhorn, Aya Khalaf, Alex Lepauvre, Ling Liu, David Richter, Yamil Vidal, et al. Adversarial testing of global neuronal workspace and integrated information theories of consciousness. *Nature*, pages 1–10, 2025.

[10] Antonia Creswell and Murray Shanahan. Faithful reasoning using large language models, 2022.

[11] Daniel C. Dennett. *Consciousness Explained*. Penguin Books, 1991.

[12] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization. *ArXiv*, abs/2404.16130, 2024.

[13] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Baker Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy Models of Superposition. *ArXiv*, abs/2209.10652, 2022.

[14] Zepeng Fang, Yuanyuan Dang, An'an Ping, Chenyu Wang, Qianchuan Zhao, Hulin Zhao, Xiaoli Li, and Mingsha Zhang. Human high-order thalamic nuclei gate conscious perception through the thalamofrontal loop. *Science*, 388(6742):eadr3675, 2025.

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

[16] Drew Fudenberg and Jean Tirole. *Game theory*. MIT press, 1991.

[17] Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. *arXiv preprint arXiv:2304.14767*, 2023.

[18] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

[19] Kurt Gödel. Über formal unentscheidbare sätze der principia mathematica und verwandter systeme i. *Monatshefte für mathematik und physik*, 38:173–198, 1931.

[20] Kurt Gödel et al. über vollständigkeit und widerspruchsfreiheit. *Ergebnisse eines mathematischen Kolloquiums*, 3:12–13, 1932.

[21] Jerry Green and Jean-Jacques Laffont. Characterization of satisfactory mechanisms for the revelation of preferences for public goods. *Econometrica: Journal of the Econometric Society*, pages 427–438, 1977.

[22] Werner Heisenberg. Über den anschaulichen inhalt der quantentheoretischen kinematik und mechanik. *Zeitschrift für Physik*, 43(3):172–198, 1927.

[23] Searle John. The mystery of consciousness. *The New York Review of Books*, pages 53–88, 1997.

[24] Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York, 2011.

[25] Adam Tauman Kalai and Santosh S Vempala. Calibrated language models must hallucinate. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pages 160–171, 2024.

[26] Michał P. Karpowicz. A theory of meta-factorization. *ArXiv*, abs/2111.14385, 2021.

[27] Michał P. Karpowicz and Gilbert Strang. The pseudoinverse of $a = cr$ is $a^+ = r^+c^+$ (?), 2024.

[28] Kazimierz Kuratowski and Andrzej Mostowski. *Set Theory*. North-Holland Publishing Company, Amsterdam, 1978.

[29] Joseph Levine. *Purple haze: The puzzle of consciousness*. Oxford University Press, 2001.

[30] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021.

[31] Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. Halueval: A large-scale hallucination evaluation benchmark for large language models, 2023.

[32] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods, 2022.

[33] Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, et al. On the biology of a large language model. *Transformer Circuits Thread*, 2025.

[34] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt, 2023.

[35] Paul R Milgrom and Robert J Weber. A theory of auctions and competitive bidding. *Econometrica: Journal of the Econometric Society*, pages 1089–1122, 1982.

[36] Roger B Myerson. Perspectives on mechanism design in economic theory. *American Economic Review*, 98(3):586–603, 2008.

[37] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.

[38] Tim Roughgarden and Inbal Talgam-Cohen. Optimal and robust mechanism design with interdependent values. *ACM Transactions on Economics and Computation (TEAC)*, 4(3):1–34, 2016.

[39] Leonard J Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971.

[40] Artem Shelmanov, Ekaterina Fadeeva, Akim Tsvigun, Ivan Tsvigun, Zhuohan Xie, Igor Kiselev, Nico Daheim, Caiqi Zhang, Artem Vazhentsev, Mrinmaya Sachan, Preslav Nakov, and Timothy Baldwin. A head to predict and a head to question: Pre-trained uncertainty quantification heads for hallucination detection in llm outputs, 2025.

[41] Michael Sipser. Introduction to the theory of computation. *ACM Sigact News*, 27(1):27–29, 1996.

[42] ZhongXiang Sun, Xiaoxue Zang, Kai Zheng, Yang Song, Jun Xu, Xiao Zhang, Weijie Yu, and Han Li. Redeep: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. *ArXiv*, abs/2410.11414, 2024.

[43] Alfred Tarski. A lattice-theoretical fixpoint theorem and its applications. *Pacific Journal of Mathematics*, 5:285–309, 1955.

[44] Richard H Thaler. *Misbehaving: The making of behavioral economics*. WW Norton & Company, 2015.

[45] Alan Mathison Turing et al. On computable numbers, with an application to the entscheidungsproblem. *J. of Math*, 58(345-363):5, 1936.

[46] Zhentao Xu, Mark Jerome Cruz, Matthew Guevara, Tie Wang, Manasi Deshpande, Xiaofeng Wang, and Zheng Li. Retrieval-augmented generation with knowledge graphs for customer service question answering. *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024.

[47] Noson S Yanofsky. *The outer limits of reason: What science, mathematics, and logic cannot tell us*. MIT Press, 2016.

[48] Lei Yu, Meng Cao, Jackie Chi Kit Cheung, and Yue Dong. Mechanistic understanding and mitigation of language model non-factual hallucinations, 2024.