Hallucinations: Clinical, Jungian, and LLM Engineering - The Integrity of Cognition

I. Authorship and Trilingual Titles

A. Authorship and Affiliations

Rafael Oliveira (ORCID: 0009-0005-2697-4668)
Affiliation Note: Independent Researcher and Psychologist.
James Bednarski 'gridwalker' (ORCID: 0009-0002-5963-6196)
Affiliation Note: Research Nexus for Decentralized Science (DeSci) and Verifiable Computation.

B. Trilingual Titles

English: Specialized Report: Hallucinated, Hallucinatory, and the Integrity of Cognition — From Clinical Psychopathology to Language Model Engineering.

Português: Relatório Especializado: Alucinados, Alucinatórios e a Integridade da Cognição — Da Psicopatologia Clínica à Engenharia de Modelos de Linguagem.

Français: Rapport Spécialisé: Hallucinés, Hallucinatoires, et l'Intégrité de la Cognition — De la Psychopathologie Clinique à l'Ingénierie des Modèles Linguistiques.

C. Trilingual Abstract

English Abstract
This report rigorously analyzes the terms "hallucinated" and "hallucinatory" across clinical medicine, Jungian analytical psychology, and Large Language Model (LLM) engineering. We establish that clinical hallucination is a perceptual error (perception without object) [1], and delusion is a cognitive error (fixed, false belief) [2]. The LLM phenomenon, lacking sensory input, is definitively a semantic confabulation or factual fabrication [3]. This epistemological distinction is critical, as LLM errors (e.g., 39.8% fabricated academic references [4]) present an analogous risk to human delusion: the confident assertion of irrepressible falsehoods. Conversely, Jungian psychology views non-factual content (symbols, involuntary images) not as errors to be eliminated, but as prospective agents of transformation and meaning—essential for Individuation [5]. The convergence of these domains reveals a shared challenge: protecting the integrity of the core system (human self or AI state) from narrative corruption (suggestion/trauma or 'LLM Token Hypnosis' [6]). The only robust defense is Verifiable Computing, which, through architectures like Arweave AO, imposes immutable provenance on the model's reasoning trace (CoIn framework [7]), mitigating the structural vulnerability inherent in centralized AI systems.
Resumo (Português)
Este relatório analisa rigorosamente os termos "alucinado" e "alucinatório" através da medicina clínica, da psicologia analítica junguiana e da engenharia de Modelos de Linguagem Grande (LLM). Estabelecemos que a alucinação clínica é um erro perceptivo (percepção sem objeto) [1], e o delírio é um erro cognitivo (crença falsa e fixa) [2]. O fenômeno do LLM, por não possuir entrada sensorial, é definitivamente uma confabulação semântica ou fabricação factual [3]. Esta distinção epistemológica é crítica, pois os erros de LLM (e.g., 39,8% de referências acadêmicas fabricadas [4]) apresentam um risco análogo ao delírio humano: a afirmação convicta de falsidades irredutíveis. Inversamente, a psicologia junguiana vê o conteúdo não-factual (símbolos, imagens involuntárias) não como erros a serem eliminados, mas como agentes prospectivos de transformação e sentido—essenciais para a Individuação [5]. A convergência desses domínios revela um desafio compartilhado: proteger a integridade do sistema central (self humano ou estado da IA) da corrupção narrativa (sugestão/trauma ou 'Hipnose de Tokens de LLM' [6]). A única defesa robusta é a Computação Verificável, que, através de arquiteturas como Arweave AO, impõe proveniência imutável ao traço de raciocínio do modelo (framework CoIn [7]), mitigando a vulnerabilidade estrutural inerente aos sistemas de IA centralizados.
Résumé (Français)
Ce rapport analyse rigoureusement les termes « hallucinés » et « hallucinatoires » à travers la médecine clinique, la psychologie analytique jungienne et l'ingénierie des Modèles Linguistiques Larges (LLM). Nous établissons que l'hallucination clinique est une erreur perceptive (perception sans objet) [1], et le délire est une erreur cognitive (croyance fausse et fixe) [2]. Le phénomène LLM, n'ayant pas d'entrée sensorielle, est définitivement une

confabulation sémantique ou fabrication factuelle [3]. Cette distinction épistémologique est critique, car les erreurs des LLM (par exemple, 39,8 % de références académiques fabriquées [4]) présentent un risque analogue au délire humain : l'affirmation convaincue de faussetés irrépressibles. Inversement, la psychologie jungienne considère le contenu non factuel (symboles, images involontaires) non pas comme des erreurs à éliminer, mais comme des agents prospectifs de transformation et de sens—essentiels à l'Individuation [5]. La convergence de ces domaines révèle un défi commun : protéger l'intégrité du système central (le soi humain ou l'état de l'IA) de la corruption narrative (suggestion/traumatisme ou « Hypnose des Tokens LLM » [6]). La seule défense robuste est la Computation Vérifiable, qui, via des architectures comme Arweave AO, impose une provenance immuable à la trace de raisonnement du modèle (cadre CoIn [7]), atténuant la vulnérabilité structurelle inhérente aux systèmes d'IA centralisés.

II. English Report: Hallucinated, Hallucinatory, and the Integrity of Cognition

II.A. Clinical Foundations: The Dissociation Between Perception and Belief (Medicine)

The analysis of the terms "hallucinated" and "hallucinatory" requires rigorous framing within the domains of Medicine and Clinical Psychopathology. It is crucial to differentiate between disorders of perception and disorders of thought content, concepts often condensed in non-clinical contexts.

1. Hallucination: The Perceptual Phenomenon (Hallucinatory)

Hallucination refers to a fundamental alteration of sensory perception. Psychopathologically, it is defined as a perception occurring in the absence of a corresponding external sensory stimulus—a perception without an object [1, 8]. The individual experiencing such phenomena is in an "hallucinatory" state [9]. These experiences can involve all five senses, often prominent in severe clinical conditions such as schizophrenia [9]. The clinical relevance lies in the fact that the clarity of perception is the basis of sanity; *delirium*, often indicating an acute illness (like COVID-19 in the elderly [10, 11]), represents a disintegration of the ability to anchor

perception in external reality [12, 13].

2. Delusion: The Cognitive Phenomenon (Deluded/Hallucinated)

In direct contrast, **delusion** describes a primary disturbance of thought content. A person exhibiting this can be described as "deluded" (or colloquially "hallucinated") in the sense of holding fixed false beliefs. A delusion is a **false, fixed belief** that is irreducible by logical arguments or factual evidence, and is not typically accepted by the individual's culture [2, 14].

The essential distinction is etiology: while hallucinations are *perceptual* (sensory) in origin, delusions are *cognitive*, arising from distorted thought patterns ``. The danger inherent in delusion, much like in factual AI error, is the **high confidence and fixity** with which the false belief is maintained [15, 14].

3. Semantic Confabulation: The Flawed LLM Analogy

The application of the term "hallucination" to Large Language Models (LLMs) is a fundamental analogical error. LLMs lack sensory systems, meaning their error is not perceptual, but a failure of integrity or semantics . The algorithmic phenomenon is more accurately analogous to human **confabulation**, where the machine fills a knowledge gap or a low factual probability with a linguistically plausible response `[3]`. This LLM error is a **semantic confabulation** (or **factual fabrication**) .

II.B. The Jungian Perspective: Symbol, Meaning, and Transformation (Psychology)

Analytical Psychology, founded by Carl Gustav Jung, offers a perspective where internal images and "non-factualities" are vehicles of meaning, transcending mere pathological valuation.

1. Symptom vs. Symbol: The Prospective Drive

For Jung, psychological distress is interpreted as a form of **symbolic communication** signaling an inner imbalance [16, 17]. The central goal is **Individuation**—a process of self-discovery and deep integration, often intensifying in the second half of life [18, 19]. The symbol acts as a crucial **bridge (*coniunctio oppositorum*)** between the conscious and the unconscious, transforming psychic energy [20, 16]. Unlike Freud's view of the symptom as a retrospective *mnemonic symbol* [1, 9], Jung sees it as a *prospective impulse* toward wholeness [5].

2. Active Imagination vs. Pathological Invasion

The Jungian technique of **Active Imagination** is a controlled, voluntary engagement with the unconscious's symbolic content [20, 21]. It is fundamentally distinct from clinical hallucination, which is a **dissociative invasion** of unconscious content into consciousness, indicating a failure of the perceptual barrier [22]. Active Imagination requires a strengthened ego capable of mediating and integrating this content [20]. The therapeutic value of non-factual imagery, therefore, lies in its **subjective authenticity** and potential for transformation, contrasting sharply with the LLM's goal of eliminating objective falsehood [16, 21].

II.C. The Computational Imperative: Integrity and Verification (LLM Engineering)

The LLM phenomenon's structural nature poses the greatest challenge to reliability. Since elimination of **Structural Hallucinations** is mathematically unlikely with current architectures ``, engineering focuses on mitigation and verification.

1. The LLM Hypnosis Attack: The Semantic Vulnerability

The critical challenge is the **LLM Token Hypnosis Attack** [6, 23], which exploits post-training preference optimization (DPO, KTO) ``. This attack is a technical analogy to psychological suggestion: a single malicious user can stochastically inject a "poisoned" factual or behavioral *token* via positive feedback [24]. This signal persists and generalizes, causing the model to

produce the false content even without the original malicious prompt [6].

The consequence is severe: an LLM, a "god of reasoning" , acts with the **high confidence** of a human delusion, generating factual fabrications (e.g., 39.8% of academic references) `[4]` that are incredibly difficult for users to correct .

2. The Verifiability Mandate: Arweave AO and CoIn

Protecting the system requires moving beyond human *cognitive trust* `` to **Decentralized Trust** guaranteed by architecture.

- **CoIn Framework (Auditing Reasoning):** The CoIn framework addresses the opacity of *black-box* models (where users are billed for invisible reasoning tokens) . It constructs a **verifiable hash tree (Merkle Tree)** from token *embeddings* to audit the quantity and semantic validity of the LLM's internal reasoning trace . This forces **billing transparency** and **semantic integrity** ``.
- **Arweave AO (Immutable Provenance):** The decentralized **Arweave AO** computer provides the permanent infrastructure. By committing every process state (*Checkpoints*) to the immutable Arweave ledger, AO guarantees **strong provenance** [25]. This architectural immutability is the ultimate countermeasure against the mutable preference store exploited by 'Token Hypnosis' [6].

III. Relatório Português: Alucinados, Alucinatórios e a Integridade da Cognição

III.A. Fundamentos Clínicos: A Dissociação Entre Percepção e Crença (Medicina)

A análise dos termos "alucinados" e "alucinatórios" exige um rigoroso enquadramento nos domínios da Medicina e da Psicopatologia Clínica, diferenciando distúrbios da percepção de distúrbios do conteúdo do pensamento.

1. Alucinação: O Fenômeno Perceptivo (Alucinatório)

A **alucinação** refere-se a uma alteração fundamental da sensopercepção: uma percepção que ocorre na ausência de um estímulo sensorial externo correspondente [1, 8]. O indivíduo está em um estado "alucinatório" [9]. A natureza da alucinação é *sensorial*, frequentemente associada a quadros clínicos severos, como a esquizofrenia [9]. O *delirium*, por exemplo, representa uma desintegração da capacidade de ancorar a percepção na realidade externa [12, 10].

2. Delírio: O Fenômeno Cognitivo (Alucinado)

O **delírio** descreve um distúrbio primário do conteúdo do pensamento. Uma pessoa que o apresenta pode ser descrita como "alucinada" no sentido de possuir crenças falsas. O delírio é uma **crença falsa, fixa e irredutível** a argumentos lógicos [2, 14]. A etiologia é *cognitiva*, originando-se de padrões de pensamento distorcidos ``. O risco é análogo ao erro da IA: a **alta confiança** e a **fixidez** com que a crença falsa é mantida [15, 14].

3. Confabulação Semântica: A Falha de Analogia no LLM

A aplicação do termo "alucinação" aos LLMs é um erro analógico fundamental. Como os LLMs carecem de sistemas sensoriais , seu erro não é perceptivo, mas uma falha de integridade factual ou semântica. O fenômeno é mais precisamente análogo à **confabulação** humana, onde a máquina preenche o vazio de conhecimento com uma resposta linguística plausível `[3]`. O erro da LLM é, portanto, uma **confabulação semântica** ou **fabricação factual** .

III.B. A Perspectiva Junguiana: Símbolo, Sentido e Transformação (Psicologia)

A Psicologia Analítica de Jung avalia imagens internas e "não-factualidades" não como erros, mas como veículos de sentido e transformação.

1. Sintoma vs. Símbolo: O Impulso Prospectivo

Para Jung, a doença é uma **comunicação simbólica** que sinaliza desequilíbrio [16, 17]. O objetivo central é a **Individuação** [18, 19]. O símbolo atua como uma **ponte (*coniunctio oppositorum*)** entre o consciente e o inconsciente [20, 16]. Enquanto a alucinação clínica é uma **invasão dissociativa** (passiva), a técnica de **Imaginação Ativa** é um engajamento **voluntário e sintético-construtivo** que exige um ego fortalecido [20, 21]. O valor terapêutico da não-factualidade reside na sua **autenticidade subjetiva** e potencial transformador [16, 21].

III.C. O Imperativo Computacional: Integridade e Verificação (Engenharia LLM)

A eliminação completa das **Alucinações Estruturais** é matematicamente improvável ``. A engenharia se concentra na mitigação e **verificação** externa.

1. O Ataque de Hipnose de LLM: A Vulnerabilidade Semântica

O ataque mais crítico é a **Hipnose de Tokens de LLM** [6, 23], que explora a **otimização de preferências** (DPO, KTO) . Um único usuário malicioso pode injetar um *token* falso ou comportamental através de *feedback* positivo `[24]`. Este sinal persiste e se generaliza, fazendo o modelo produzir o conteúdo falso em contextos futuros `[6]`. O LLM, "um deus de raciocínio" , age com a **alta confiança** de um delírio, gerando fabricações factuais (e.g., 39,8% de referências) [4].

2. O Mandato da Verificabilidade: Arweave AO e CoIn

A defesa requer a transição da *confiança cognitiva* `` para a **Confiança Descentralizada** arquiteturalmente garantida:

- **CoIn Framework (Auditoria de Raciocínio):** O CoIn audita a opacidade dos *tokens*

ocultos do raciocínio interno do LLM (). Constrói uma **árvore hash verificável (Merkle Tree)** a partir de *embeddings* , forçando a **transparência** e a **integridade semântica** ``.
- **Arweave AO (Proveniência Imutável):** O computador descentralizado **Arweave AO** armazena permanentemente todo o estado computacional e *checkpoints* no Arweave. Essa **imutabilidade** arquitetural é o contraponto final ao *store* de preferência mutável explorado pela Hipnose [6].

IV. Rapport Français: Hallucinés, Hallucinatoires, et l'Intégrité de la Cognition

IV.A. Fondements Cliniques: La Dissociation entre Perception et Croyance (Médecine)

L'analyse des termes « hallucinés » et « hallucinatoires » doit être rigoureuse en Médecine et en Psychopathologie Clinique, distinguant les troubles de la perception de ceux du contenu de la pensée.

1. Hallucination: Le Phénomène Perceptif (Hallucinatoire)

L'**hallucination** est une altération fondamentale de la perception sensorielle : une perception qui se produit en l'absence de stimulus sensoriel externe [1, 8]. L'individu est dans un état « hallucinatoire » [9]. C'est un phénomène *sensoriel*, souvent associé à des tableaux cliniques graves comme la schizophrénie [9]. Le *delirium* représente une désintégration de la capacité à ancrer la perception dans la réalité [12, 10].

2. Délire: Le Phénomène Cognitif (Halluciné/Délirant)

Le **délire** décrit un trouble primaire du contenu de la pensée. La personne concernée peut être décrite comme « hallucinées » au sens de posséder des croyances fausses. Le délire est

une **croyance fausse, fixe et irréductible** par des arguments logiques [2, 14]. L'étiologie est *cognitive* ``. Le risque est analogue à l'erreur de l'IA : la **grande confiance** et la **fixité** avec laquelle la fausse croyance est maintenue [15, 14].

3. Confabulation Sémantique: L'Analogie Fausse du LLM

L'application du terme « hallucination » aux LLM est une erreur analogique. Les LLM n'ayant pas de systèmes sensoriels , leur erreur n'est pas perceptive, mais une défaillance de l'intégrité factuelle ou sémantique. Le phénomène est plus précisément analogue à la **confabulation** humaine, où la machine comble un vide de connaissance par un récit plausible `[3]`. L'erreur du LLM est donc une **confabulation sémantique** ou **fabrication factuelle** .

IV.B. La Perspective Jungienne: Symbole, Sens et Transformation (Psychologie)

La Psychologie Analytique de Jung considère les images internes non factuelles comme des vecteurs de sens et de transformation.

1. Symptôme vs. Symbole: L'Impulsion Prospective

Pour Jung, la maladie est une **communication symbolique** signalant un déséquilibre [16, 17]. L'objectif central est l'**Individuation** [18, 19]. Le symbole sert de **pont (*coniunctio oppositorum*)** entre le conscient et l'inconscient [20, 16]. Alors que l'hallucination clinique est une **invasion dissociative** (passive), l'**Imagination Active** est un engagement **volontaire et constructif** nécessitant un ego renforcé [20, 21]. La valeur thérapeutique du non-factuel réside dans son **authenticité subjective** [16, 21].

IV.C. L'Impératif Computationnel: Intégrité et Vérification (Ingénierie LLM)

L'élimination totale des **Hallucinations Structurelles** est mathématiquement improbable ``. L'ingénierie se concentre sur l'atténuation et la **vérification** externe.

1. L'Attaque d'Hypnose des Tokens LLM: La Vulnérabilité Sémantique

L'attaque la plus critique est l'**Hypnose des Tokens LLM** [6, 23], qui exploite l'**optimisation des préférences** (DPO, KTO) ``. Un seul utilisateur malveillant peut injecter un *token* faux par *feedback* positif [24]. Ce signal persiste et se généralise, forçant le modèle à produire le contenu faux dans des contextes futurs [6]. Le LLM agit avec la **haute confiance** d'un délire, générant des fabrications (e.g., 39,8 % de références) [4].

2. Le Mandat de la Vérifiabilité: Arweave AO et CoIn

La défense nécessite la transition de la *confiance cognitive* `` vers la **Confiance Décentralisée** garantie par l'architecture :

- **Cadre CoIn (Audit de Raisonnement):** CoIn audite l'opacité des *tokens* cachés du raisonnement LLM (). Il construit un **arbre de hachage vérifiable (Merkle Tree)** , imposant la **transparence** et l'**intégrité sémantique** ``.
- **Arweave AO (Provenance Immuable):** L'ordinateur décentralisé **Arweave AO** enregistre en permanence l'état de calcul et les *Checkpoints* sur Arweave. Cette **immuabilité** architecturale est le remède ultime contre la vulnérabilité du stockage de préférences mutable [6].