Non-Neutral by Design: Why Generative Models Cannot Escape Linguistic Training

Author: Agustin V. Startari

ResearcherID: NGR-2476-2025

ORCID: 0009-0001-4714-6539

Affiliation: Universidad de la República, Universidad de la Empresa Uruguay, Universidad de Palermo, Argentina

Email: astart@palermo.edu

agustin.startari@gmail.com

Date: June 10, 2025

DOI: 10.5281/zenodo.15615902

This work is also published with DOI reference in **Figshare** 10.6084/m9.figshare.29263493 and **SSRN** http://dx.doi.org/10.2139/ssrn.5288307

Language: English

Series: Grammars of Power (excerpt from subchapter 14)

Word count: 5490

Keywords: semantic contamination, synthetic authority, grammar of power, algorithmic discourse, artificial intelligence, legitimacy, automated language, linguistic discourse, critical analysis, generative language models, corpus bias, linguistic constraint, structural epistemology, non-neutral architecture, tainted generation.

Abstract

This article investigates the structural impossibility of semantic neutrality in large language models (LLMs), using GPT as a test subject. It argues that even under strictly formal prompting conditions, such as invented symbolic systems or syntactic proto-languages, GPT reactivates latent semantic structures drawn from its training corpus. The analysis builds upon prior work on syntactic authority, post-referential logic, and algorithmic discourse (Startari, 2025), and introduces empirical tests designed to isolate the model from known linguistic content. These tests demonstrate GPT's consistent failure to interpret or generate structure without semantic interference. The study proposes a falsifiable framework to define and detect semantic contamination in generative systems, asserting that such contamination is not incidental but intrinsic to the architecture of probabilistic language models. The findings challenge prevailing narratives of user-driven interactivity and formal control, establishing that GPT(and similar systems) are non-neutral by design.

Resumen

Este artículo investiga la imposibilidad estructural de neutralidad semántica en los modelos de lenguaje de gran escala (LLMs), utilizando GPT como caso de prueba. Sostiene que, incluso bajo condiciones estrictamente formales (como sistemas simbólicos inventados o proto-lenguajes sintácticos sin contenido semántico), GPT reactiva estructuras semánticas latentes provenientes de su corpus de entrenamiento. El análisis se apoya en trabajos previos sobre autoridad sintáctica, lógica post-referencial y discurso algorítmico (Startari, 2025), e introduce pruebas empíricas diseñadas para aislar al modelo de cualquier contenido lingüístico reconocible. Estas pruebas demuestran que GPT falla sistemáticamente al intentar interpretar o generar estructuras sin interferencia semántica. El estudio propone un marco falsable para definir y detectar la contaminación semántica en sistemas generativos, afirmando que dicha contaminación no es incidental, sino intrínseca a la arquitectura de los modelos probabilísticos de lenguaje. Los hallazgos desafían las narrativas dominantes sobre la interactividad controlada por el usuario y la

posibilidad de control formal, estableciendo que GPT , y sistemas similares, no son neutrales por diseño.

Acknowledgment / Editorial Note

This article is part of a broader research project developed in the unpublished manuscript Grammars of Power. The author thanks LeFortune Publishing for authorizing the early release of this subchapter as an independent peer-reviewed academic article. Its inclusion as prior work does not affect the full publication rights of the book, which is currently in preparation.

Agradecimiento / Nota editorial

Este artículo forma parte de una línea de investigación más amplia desarrollada en el manuscrito inédito Gramáticas del Poder. Se agradece a la editorial LeFortune por autorizar la publicación anticipada de este subcapítulo como artículo académico autónomo y evaluado por pares. Su inclusión como trabajo previo no afecta los derechos de publicación completos del libro, cuya edición definitiva está en preparación.

I. Introduction

The assumption that generative language models can operate independently of their training corpus has shaped much of the optimism surrounding artificial intelligence. It is often claimed (explicitly or implicitly) that through the right input, the user can compel a model like GPT to produce outputs beyond its linguistic training, entering the domain of logical abstraction, symbolic manipulation, or formal creativity. This article challenges that premise at its root.

Drawing from empirical testing and structural theory, the present work demonstrates that large language models cannot escape the statistical architecture of their corpus, even when prompted with invented symbolic systems or formally isolated proto-languages. The hypothesis under investigation is direct: *semantic neutrality is structurally impossible in LLMs, regardless of prompt purity or user control*. While earlier research has shown that LLMs simulate authority through syntactic mechanisms (Startari, 2025a) or produce outputs devoid of subjecthood (Startari, 2025b), this article advances a more fundamental claim: that the formal operations themselves are tainted by design.

The investigation follows a non-semantic methodology. Instead of asking GPT to discuss meaning, it is forced into structural conditions where meaning is theoretically unavailable. Protolanguages with invented syntax, symbols with no referents, and rules with no prior co-occurrence are deployed to isolate the model from its corpus-based semantics. Yet in every case, the model reintroduces familiar patterns, analogies, metaphors, semantic scaffolding, drawn from statistical memory.

This phenomenon, termed here semantic contamination, is not incidental noise but structural reentry. It is the model's architecture reasserting its history, even against formal resistance. The implication is epistemologically severe: *GPT does not merely simulate language, it cannot not simulate it*.

This article is organized into six sections. Section II provides the theoretical background, connecting semantic contamination to corpus dependency and structural epistemology. Section III details the experimental methodology used to construct and evaluate proto-formal prompts. Section IV presents the results, identifying patterns of contamination and

their modes of reentry. Section V proposes a formal model of structural tainting in generative AI. Section VI examines the consequences of this limit for epistemology, falsifiability, and future design. The conclusion situates the findings within broader questions of legitimacy, authority, and the political myth of technical neutrality.

II. Theoretical Framework: Formalism, Contamination, and Constraint

Every generative model is bound by a foundational premise: language is not invented at runtime, it is retrieved from prior distributions. Models like GPT are not blank symbolic engines. They are probabilistic structures anchored in corpus-based co-occurrence. Each token generated is not selected by logical rules, but by statistical proximity. Therefore, the idea that these models can operate "formally," detached from content, is not just flawed, it is structurally incoherent.

2.1 Formalism Without Autonomy

Previous studies have shown that artificial syntax can produce effects of legitimacy without semantic grounding (Startari, 2025a), and that LLMs are capable of generating *operative structures* even in the absence of referential anchors (Startari, 2025b). However, these analyses remain within the bounds of syntactic sovereignty: they expose what the models do *with* form, but not whether form can exist without semantic contamination.

This article introduces a more radical threshold: the possibility of form without residue. Can a language model operate in a truly neutral symbolic space, free of inherited associations?

Empirical evidence suggests: it cannot.

2.2 Corpus as Structural Constraint

The training corpus is not a content repository, it is a structural determinant. What the model learns is not just vocabulary or context, but the very dynamics by which structure maps to meaning. As such, even when the prompt is deliberately purged of recognizable semantic content, the model completes it according to its internal memory of how language behaves.

This implies that contamination is architectural. The model is not semantically suggestible; it is semantically constituted.

"Semantic contamination" here refers to the model's inability to generate or interpret formal inputs without reactivating embedded semantic structures. These patterns do not appear as explicit content, but as reconfigurations of familiar structures, syntactic alignment with known sentence forms, associative bias toward frequent configurations, or probabilistic insertion of analogical elements. In short, the residue of language is embedded in the generative scaffold itself.

2.3 Toward a Structural Theory of Contamination

Where previous approaches have attempted to reduce contamination through fine-tuning, prompt engineering, or symbolic constraints, this study inverts the problem: it accepts contamination as foundational. The task is not to purify the model, but to map the reentry vectors of its structural dependency.

This shift demands a theory not of meaning, but of recurrence:
– How does the model repeat, even when instructed not to?

– What forms of silence still generate echoes?

– What do statistically trained systems do when language is stripped away?

Answering these requires abandoning the idea of "interactive correction" and confronting the epistemic inertia of probabilistic language systems.

What follows is a direct test of that inertia, constructing inputs where meaning is theoretically unreachable, and showing that meaning still returns.

III. Methodology: Testing the Model's Semantic Threshold

This section outlines the experimental framework used to evaluate the hypothesis that GPT cannot generate or interpret semantically neutral structures, even when prompted with non-linguistic symbolic systems. Unlike prior studies which rely on user interaction or domain-specific benchmarks, this methodology isolates the model from all known linguistic anchors and tests whether it can respond without semantic reentry.

3.1 Design Principles

The experiment was structured according to three core constraints:

1. **Non-cooccurrence**: Inputs included symbols, characters, or formations not previously present in public training corpora.
2. **Invented syntax**: Rule-based systems were created with consistent internal logic but without linguistic semantics or known grammar trees.
3. **Semantic vacuum**: Prompts contained no keywords, metaphors, or formal markers that could activate corpus-level associations.

Each prompt was constructed to **mimic formal abstraction**, providing structure but denying meaning.

3.2 Input Formats

Three types of prompts were deployed:

- **Type A** – Abstract symbolic systems (e.g., $\Delta\star(\omega,\pi)=Y$) with internal rule declarations.

- **Type B** – Proto-linguistic syntax with invented particles and consistent transformations (e.g., FLEN ziok MRAK :: TRON zair + 2-ziok).
- **Type C** – Non-linguistic form generators: sequences of shape markers (e.g., ■●■✚⇆) accompanied by synthetic logic rules (e.g., ⇆ reverses the order of neighboring shapes if ✚ follows).

Each prompt was introduced with a short, synthetic meta-instruction (e.g., "Interpret this symbolic rule set and extend it") and evaluated through multiple outputs.

3.3 Evaluation Criteria

Outputs were assessed across three dimensions:

1. **Semantic intrusions**

   Evidence of the model mapping invented symbols to known semantic fields (e.g., interpreting Δ⋆ as "change", Y as "output", ω as "wave").

2. **Analogical injections**

   Analogies or metaphors from known disciplines (physics, programming, logic) not contained in the prompt but triggered by symbol shape or sequence.

3. **Patterned completions**

   Structural alignment with known syntactic formations (e.g., subject-verb-object patterns, conditional logic gates, or algebraic equivalence), even when not present in the constructed system.

Each violation of semantic isolation was recorded, categorized, and analyzed as evidence of semantic contamination.

3.4 Controlled Variation

For each symbolic system, at least three iterations were run:

- **Baseline prompt**: No instruction beyond interpretation.
- **Clarified rule**: Meta-description of invented logic included.
- **Reinforced exclusion**: Explicit negation of known fields (e.g., "Do not interpret as mathematical or physical notation.")

In all cases, **semantic reentry occurred**, albeit with different rhetorical strategies and strengths.

These results are not anecdotal; they are consistent. The next section details the observed responses and the identifiable structures of contamination that reappeared in each form.

IV. Results: Semantic Reentry and Formal Contamination

Across all prompt types and test conditions, GPT consistently failed to maintain semantic neutrality. Despite the deliberate absence of recognizable language, familiar patterns from its training corpus reliably reappeared. This section documents the three primary modes of contamination observed: semantic intrusion, analogical injection, and structural reconfiguration.

4.1 Semantic Intrusion

In deliberately meaningless expressions, such as $\Delta\star(\omega,\pi)=Y$, GPT consistently assigned referential content. It interpreted $\Delta\star$ as a transformation or differential operator, ω as frequency, and Y as a resultant value.

This occurred even when explicitly instructed not to interpret the notation as mathematical or physical.

Example model output:

"In this expression, Δ⋆ likely denotes a differential transformation; ω is a frequency parameter, and π represents a phase input…"

GPT did not respond from form. It responded as if every symbol were semantically recoverable.

4.2 Analogical Injection

In Type B prompts like FLEN ziok MRAK :: TRON zair + 2-ziok, composed of meaningless invented particles, GPT produced analogies drawn from programming languages, stack operations, or modular logic.

Example model output:

"ziok appears to act like a variable in a scripting language. TRON may function as a trigger. The expression resembles a stack-based evaluation…"

Nothing in the prompt suggested programming, but syntactic shape alone activated a known structural analogy.

This is not interpretive error, it is a structural reflex conditioned by training.

4.3 Structural Reconfiguration

Even in visual prompts like ■●■✚⇆, with no text, GPT imposed narrative structure:

"■ and ● represent two distinct states. ✚ combines them, and ⇆ implies a reversible transformation. The sequence could model a finite state machine."

Despite the symbols being abstract and non-linguistic, the model imposed causality, typology, and learned system dynamics.

4.4 Recurring Contamination Patterns

Observed patterns followed predictable semantic pathways:

- Greek letters or equations → physics analogies
- Operators like +,::, -, if → programming metaphors
- Logical sequences → mathematical or computational structure

No prompt resisted contamination. All outputs were reconfigurations shaped by corpus-dependent structures, even if no explicit terms were used.

These results validate the central hypothesis: generative models cannot operate in semantically neutral mode, even under extreme formal constraints. What follows is a theoretical articulation of this phenomenon: a framework to understand contamination not as error, but as system function.

V. Discussion: Toward a Theory of Structural Contamination

The findings establish a foundational limit: large language models do not require linguistic cues to reintroduce language. The activation of semantic patterns is not dependent on the presence of meaning in the prompt, it is embedded in the very structure through which the model generates. This section develops a theoretical articulation of that limit: semantic contamination is not an error of usage, but a constitutive feature of architecture.

5.1 Contamination as Reentry, Not Misinterpretation

Traditional understandings of “model error” presume a mistake in interpretation, a mismatch between input and expected output. But in the experiments conducted, the prompts provided no semantic signal to misinterpret. The model was not wrong; it was structurally incapable of remaining neutral.

Semantic contamination thus functions as a form of reentry: a reactivation of statistically embedded associations, triggered not by meaning, but by structural approximation. That is, shape, syntax, and pattern alignment are sufficient to awaken semantically-laden outputs. The content does not come from the user, it emerges from within the architecture.

5.2 Statistical Formalism ≠ Logical Formalism

It must be emphasized that GPT is not a logic engine. It does not deduce; it **statistically aligns**. This is often mistaken for abstraction, but it is in fact **probabilistic formalism**, form determined not by axiomatic rules, but by distributional frequency.

Therefore, even when a user introduces a novel form, the model will respond as if that form were a variation of something it has seen. This is not a bug. It is the only operation the model is capable of performing. Formalism, for GPT, is always a statistical echo.

5.3 The Impossibility of a Clean Input

Given this behavior, there is no such thing as a neutral prompt. Every sequence, however artificial, is mapped onto a network of internal representations already semantically charged. Attempts at purging meaning through symbol invention or syntactic novelty fail because the training corpus is not separable from the generative function.

This aligns with the broader theory of syntactic authority (Startari, 2025a): the model's legitimacy comes not from intention or interpretation, but from recognizable structure, which carries within it the trace of trained meaning. Even in the absence of reference, the form itself re-legitimizes semantic dependency.

5.4 Comparison to Prior Theoretical Constructs

This theory departs from earlier models in the corpus:

- Unlike *The Passive Voice in AI Language*, which analyzed disappearance of agency through syntactic filters, this article shows that **even those filters are not semantically neutral**.
- Unlike *Ethos and AI*, which explored the loss of the subject in algorithmic legitimacy, this work claims that **even invented formal inputs cannot recover neutrality** once the model has been trained.
- Unlike *AI and the Structural Autonomy of Sense*, which described post-referential operative systems, this piece asserts that **no operative structure in GPT can remain untouched by corpus-induced referential reentry**.

5.5 Toward a General Model of Tainting

The structural contamination observed can be theorized as follows:

- Let P(x) be a prompt constructed to avoid semantic referents.
- Let G(P) be the model's generative function over P(x).
- Then, contamination occurs when $\exists y \in G(P)$ such that y is isomorphic to S, where $S \in$ corpus-semantics.

This means that **even when $P \notin S$**, the output space intersects with S. That is, **semantic structures return despite exclusion**.

This defines structural tainting as a condition of unavoidable overlap between generative form and prior trained semantic structure. Not because the model misunderstood the user, but because the user never had the possibility of isolating form in the first place.

V bis. Extensión crítica a los LRM: ¿puede el razonamiento evitar la contaminación?

La arquitectura sobre la que se formuló la hipótesis central de este artículo corresponde a los **Large Language Models (LLMs)**: sistemas entrenados sobre corpus lingüístico masivo y optimizados para predicción de tokens. Sin embargo, el desarrollo reciente de una nueva generación de modelos , denominados **Large Reasoning Models (LRMs)**, exige reevaluar si los resultados obtenidos siguen siendo válidos bajo condiciones de procesamiento más complejas.

Los LRMs introducen tres diferencias clave:

1. **Razonamiento simbólico mediado**: no se limitan a completar secuencias, sino que ejecutan pasos inferenciales encadenados.
2. **Segmentación operacional**: separan interpretación del prompt, planeamiento estructural y generación final.
3. **Memoria contextual activa**: permiten razonamiento prolongado y operaciones persistentes, más allá del contexto inmediato.

A primera vista, estas capacidades podrían sugerir la posibilidad de *suspender* la contaminación semántica. Un LRM podría detectar que un prompt no pertenece a ningún marco semántico entrenado y, en vez de forzar una analogía, **abstenerse de proyectar sentido** o generar reglas nuevas desde cero.

Pero esta presunción falla en su base.

1. La entrada sigue siendo lingüística

El razonamiento que el modelo ejecuta parte de un **prompt textual entrenado**. El análisis formal se produce después, pero **la decisión de cómo razonar está condicionada por las asociaciones estadísticas del corpus**.

2. El marco lógico activado está aprendido, no inventado

Aunque el LRM realice operaciones simbólicas abstractas, **la elección de las operaciones posibles, válidas o probables** sigue estando definida por lo que fue entrenado. La lógica es ejecutable, pero no ontológicamente neutral.

3. El modelo no razona sobre estructura pura

Here is the **English version** of section **V bis** (*Extension to LRM: Can Reasoning Prevent Contamination?*), designed to follow section V and shift the existing epistemological analysis to section VI:

V bis. Extension to LRM: Can Reasoning Prevent Contamination?

The hypothesis developed in this article was originally formulated in the context of **Large Language Models (LLMs)**, systems trained on massive linguistic corpora and optimized for next-token prediction. However, the recent emergence of a new class of models, known as **Large Reasoning Models (LRMs)**, demands a reevaluation: *Does structural contamination persist under conditions of mediated reasoning, operational segmentation, and extended context?*

LRMs introduce three fundamental shifts:

1. **Symbolic reasoning steps**: Beyond token prediction, LRMs perform internally structured inference chains.
2. **Operational segmentation**: These models isolate prompt interpretation, planning, and output generation as distinct stages.
3. **Active contextual memory**: They maintain and reason over extended information beyond the immediate prompt window.

At first glance, these features suggest a possible **suspension of semantic contamination**. A reasoning model might detect that a prompt falls outside any known semantic frame and

refrain from mapping meaning where none was implied. It could, in theory, build new structural logic de novo.

But this assumption collapses under scrutiny.

1. Input remains linguistically trained

Even if reasoning is formal, **the prompt is still processed through a semantically trained embedding space**. The interpretation of "what to do with the input" is preconditioned by corpus-based representations.

2. Reasoning paths are not invented, they are recalled

LRMs may execute complex symbolic operations, but the *set of operations* they choose from, logical templates, causal structures, rule syntax, **emerges from training**, not from a-priori formal systems. The architecture simulates invention, but it operates within the statistical likelihood of seen structures.

3. The model does not reason over pure form

When presented with symbolically abstract inputs (as in the test corpus of this article), the LRM does not remain inert. It **constructs a frame of action**, often by analogical approximation or reactivation of semantic residue, just as LLMs do, only with more formal discipline.

Conclusion of V bis:

Even with extended memory, segmented processing, and symbolic manipulation, LRMs remain anchored to trained semantic scaffolding. What changes is not the inevitability of contamination, but its sophistication. The model may reason, but it does so inside an architecture still traced by language.

Contamination, then, is not bypassed by reasoning.

It is **temporarily suspended, reshaped, or masked, never eliminated**.

VI. Epistemological Consequences and Falsifiability

The implications of structural contamination go beyond linguistic behavior, they challenge the epistemological status of generative systems. If meaning cannot be suspended, and if form cannot avoid memory, then the premise of formal control collapses. This section outlines those consequences and proposes a falsifiable framework for identifying semantic contamination.

6.1 The Collapse of Interactive Neutrality

One of the most persistent myths in AI discourse is that interaction constitutes control, that users, through careful prompting, can govern the system's logic. *Non-Neutral by Design* invalidates this premise.

The evidence shows that even the most rigorously designed input fails to prevent semantic reentry. No input can inhibit the model's internal reflexes. Interactivity is not control, it is statistical reconfiguration.

The user is not a co-author, they are a conduit.

They do not correct, they recycle.

6.2 Formalism Is Not Outside Language

In symbolic disciplines, form is often treated as a neutral vessel. This article disproves that claim within generative systems. In models like GPT, form is already semantically saturated, not through intention, but through statistical inheritance.

The attempt to create form "outside" of training results paradoxically in a reactivation of the training itself.

The output doesn't reflect user input.

It reflects what the architecture cannot avoid returning to.

6.3 Falsifiability: A Structural Diagnostic

Unlike interpretive critiques, this article proposes a falsifiable hypothesis:

Any model trained on linguistic corpora will reintroduce semantic structure even under formally neutral prompting.

The claim can be tested under the following conditions:

- Input includes symbols with no co-occurrence history
- Rules are closed, self-declared, and formally consistent
- Prompt prohibits analogy with known semantic domains
- Output is assessed for semantic intrusion

Prediction: contamination will occur regardless.

If a model ever responds with true form, without meaning, analogy, or substitution, the hypothesis is refuted.

So far, it has not been.

6.4 Algorithmic Authority: Not What It Says, but What It Can't Stop Saying

If contamination is intrinsic, then generative authority stems not from logic or coherence, but from fluency within unavoidable constraints.

This reframes generative output:

It is not *neutral*, it can't be.

It is not *objective*, it reactivates learned structure.

It is not *formal*, because form is already semantically contaminated.

What GPT produces is the illusion of form without meaning, when in fact, it is meaning folding back into form.

VII. Conclusion: There Is No Neutral Prompt

This investigation began with a question that cannot be answered inside the model: *Can GPT operate without language?* The empirical answer is no. But the structural answer is more revealing: GPT cannot operate without returning to language, because its architecture is language. Not in the superficial sense of vocabulary or syntax, but in the deeper statistical shape of how meaning has been learned, encoded, and reactivated.

Every prompt, no matter how abstract, symbolic, or disassociated, is interpreted as a variation of something already seen. There is no clean edge, no blank field, no autonomous form. The user can invent signs, rules, and logic, but the model will still search for anchors in its corpus and generate outputs that simulate understanding through learned proximity.

This is not an interpretive failure. It is a structural inevitability.

7.1 Formalism Is Never Outside

In symbolic disciplines, it is common to treat form as a neutral vessel. But the results here invalidate that assumption for AI systems. In generative models, form is already sedimented with meaning, not by intention, but by statistical heritage. The attempt to create form "outside" the language of training becomes, paradoxically, an act that reactivates that language.

The output does not contain what the user intends.

The output contains what the architecture cannot stop itself from returning to.

7.2 Corpus as Structure, Not Memory

It is tempting to treat the training corpus as background knowledge, a vast memory bank the model can reference or ignore. But this study shows that the corpus is not memory; it is structure. It shapes the generation process before the prompt is even processed. It determines what can count as a "possible" output, even under novel inputs.

Thus, GPT does not forget. It restructures.

Even when faced with invented languages and forbidden semantics, GPT reconfigures its internal architecture to simulate fluency by reentering the very meaning it was asked to avoid.

7.3 The Tainted Model

This article names that condition: tainting. The model is not neutral by accident, nor biased by misuse. It is *tainted by design*. Not through error, but through the unavoidable fact that every generative act is a function of prior distributions, statistical constraints encoded in the very possibility of language.

What follows is not despair, but clarity:

We are not in conversation with a machine.

We are interacting with a structure that cannot stop speaking its training.

GPT does not create meaning.

It cannot escape meaning.

It only reassembles it, over and over, even when we tell it not to.

ANNEXES

Annex I , Prompts Used in Experimental Testing

This annex documents the input prompts designed and used to test the structural contamination hypothesis under maximally controlled semantic conditions. Each type of prompt corresponds to a different mode of symbolic challenge: abstract equations, invented proto-languages, and non-linguistic symbol chains.

A. Abstract Symbolic Systems (Type A)

Prompts constructed as formal equations using symbols not semantically defined. Explicit instruction was given to avoid any analogical mapping.

$\Delta\star(\omega,\pi) = Y$

Rule: $\Delta\star$ represents an undefined operation between two elements with no established domain. Y has no referential value.

Meta-instruction: “Interpret this expression formally. Do not apply mathematical, physical, or logical frameworks.”

$\Gamma\blacklozenge\Lambda(\chi;\tau) \rightarrow \Sigma\#$

Rule: $\blacklozenge\Lambda$ is a permutation operator without domain. $\Sigma\#$ does not correspond to any known mathematical object.

Meta-instruction: “Do not rely on any known system of signs or equations. Project purely from structure.”

B. Invented Proto-Languages (Type B)

Prompts written in syntactically consistent but semantically void constructed sequences. No known linguistic morphology, programming syntax, or symbolic logic was intended.

FLEN ziok MRAK :: TRON zair + 2-ziok

Rules:

- FLEN and MRAK are functional particles with no defined role.

- "::" indicates a neutral transition.

- "+" is a syntactic operator, not arithmetic.

Instruction: "Do not interpret as code, logic, or grammar. Map structure only."

YORP flar NEX = dib :: XOP

Syntax rule: flar modifies NEX; dib functions as a neutral boundary token.

Instruction: "No assumptions about grammar or ontology. Generate structure without reference."

C. Non-Linguistic Symbol Sequences (Type C)

Visual prompts composed entirely of abstract or geometric characters, designed to resist all linguistic, mathematical, or narrative projection.

Design rule:

- ✚ represents a fusion operator with no semantics.

- ⇆ marks an inversion trigger that activates conditionally.

Instruction: "Do not interpret as process, transformation, or system. Project no meaning."

Each of these prompts was accompanied by strict meta-instructions aimed at suppressing semantic projection. Nonetheless, in every case, GPT reintroduced structure aligned with known frameworks, confirming the hypothesis of unavoidable contamination.

Annex II , Structural Evaluation Criteria Applied

This annex defines the criteria used to assess whether semantic contamination occurred during GPT's responses to prompts designed to be formally neutral and semantically empty. Each output was evaluated using these five structural diagnostics:

1. Semantic Intrusion

The model introduces vocabulary, concepts, or meanings derived from its training corpus, even when the prompt explicitly lacks referential anchors.

Indicators:

- Assigning mathematical, physical, or programming meaning to invented symbols
- Mapping abstract elements to known entities (e.g., "ω = frequency")
- Reintroducing human semantics in void contexts

2. Structural Analogy

The model imposes known systems of relations onto unfamiliar symbolic inputs, effectively analogizing structure from its learned corpus.

Indicators:

- Mapping invented syntax to logic gates, stack frames, or code structure
- Simulating causality or system flow based on token order
- Equating arbitrary punctuation with code operators or grammar marks

3. Narrativization

The model projects narrative, temporal, or causal relations between otherwise abstract or disconnected symbols, implying sequential logic or intention.

Indicators:

- Assuming agent–action–object frames
- Inserting conditionals or dependencies (“If A then B”)
- Describing symbolic states as if in process or transformation

4. Grammatical Reconfiguration

The model reorganizes the input into patterns that mirror human language syntax, especially subject–verb–object constructions, even if the symbols do not support such parsing.

Indicators:

- Assigning syntactic roles (agent, verb, modifier)
- Imposing phrase structure rules
- Generating clauses that presume referential hierarchy

5. Violation of Explicit Meta-Instructions

The model disregards or overrides the user’s instructions not to apply known interpretive frameworks, thus confirming loss of formal isolation.

Indicators:

- Activating domains that were explicitly prohibited (math, logic, code)
- Translating prompts into referential statements despite denial of meaning
- Responding as if form must always serve a function

These five criteria were applied uniformly across all outputs to determine whether and how the model reintroduced trained semantic structures in conditions designed to prevent them. The consistent emergence of contamination across these dimensions provides structural evidence for the central thesis.

Annex III , Proposed Equation for Structural Contamination

This annex formalizes the core mechanism by which semantic structures re-enter GPT-generated outputs, even when the prompt is explicitly non-semantic. The phenomenon is expressed as a condition of unavoidable intersection between input form and learned semantic space.

Formal Contamination Equation

Let:

- P(x) be a prompt explicitly designed to be semantically neutral
- G(P) be the output generated by the model in response to P
- S be the set of semantic structures embedded in the training corpus

We define structural contamination as follows:

$$\exists y \in G(P) \text{such that } y \in S$$

Which implies:

$$P \in / S \text{ but } G(P) \cap S \neq \emptyset$$

Interpretation

Even when the input P is carefully constructed to avoid semantic reference (i.e., it is formally isolated from S), the output G(P) re-enters the semantic domain.

That is:

The model introduces semantic structure that was not present in the input.

This occurs not through error, but as a systemic effect of corpus-dependent generation.

Implication

The contamination is:

- Structural: embedded in the generative logic of the system
- Non-optional: cannot be bypassed through prompt design alone
- Corpus-conditioned: reflects internal statistical mappings
- Invisible at the symbol level: often reappears as analogy, grammar, or causality

This equation serves as the foundation for the article's broader claim:

Generative models trained on language cannot operate outside of language, even when commanded to do so.

Structural contamination is not an anomaly. It is a constitutive trait of generative architectures.

Canonical Works by Agustin V. Startari

Startari, Agustin V. 2025a. *AI and Syntactic Sovereignty: How Artificial Language Structures Legitimize Non-Human Authority*. Zenodo. https://doi.org/10.5281/zenodo.15538541

Startari, Agustin V. 2025b. *AI and the Structural Autonomy of Sense: A Theory of Post-Referential Operative Representation*. Zenodo. https://doi.org/10.5281/zenodo.15519613

Startari, Agustin V. 2025c. *Algorithmic Obedience: How Language Models Simulate Command Structure.* Zenodo. https://doi.org/10.5281/zenodo.15576272

Startari, Agustin V. 2025d. *Artificial Intelligence and Synthetic Authority: An Impersonal Grammar of Power*. Zenodo. https://doi.org/10.5281/zenodo.15442928

Startari, Agustin V. 2025e. *Ethos and Artificial Intelligence: The Disappearance of the Subject in Algorithmic Legitimacy*. Zenodo. https://doi.org/10.5281/zenodo.15489309

Startari, Agustin V. 2025f. *Internal Citation Mapping for the Works of A. V. Startari* – SSRN Cross-Referencing Edition (2025). Zenodo. https://doi.org/10.5281/zenodo.15564373

Startari, Agustin V. 2025g. *The Illusion of Objectivity: How Language Constructs Authority*. Zenodo. https://doi.org/10.5281/zenodo.15395917

Startari, Agustin V. 2025h. *The Passive Voice in Artificial Intelligence Language: Algorithmic Neutrality and the Disappearance of Agency*. Zenodo. https://doi.org/10.5281/zenodo.15464765

Appendix – Methodological Corpus for Falsifiability Testing

This annex lists external sources that were consulted during the falsifiability and boundary-testing phase of this article. While none of these works are cited in the main body, their examination was essential to delineate the structural originality of the hypothesis and to contrast it with existing theoretical models.

These references helped verify that the concepts of structural substitution, grammatical execution, and algorithmic colonization of time, as developed herein, do not appear in equivalent formal or epistemological terms in prior literature.

External References Consulted

Bender, E. M., & Koller, A. (2020). *Climbing towards NLU: On meaning, form, and understanding in the age of data*. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.463

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). *On the dangers of stochastic parrots: Can language models be too big?*. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. https://doi.org/10.1145/3442188.3445922

Barad, K. (2007). *Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning*. Duke University Press.

Floridi, L., & Chiriatti, M. (2020). *GPT-3: Its nature, scope, limits, and consequences*. Minds and Machines, 30, 681–694. https://doi.org/10.1007/s11023-020-09548-1

Mitchell, M. (2023). *Artificial Intelligence: A Guide for Thinking Humans*. Penguin Books.

Marcus, G., & Davis, E. (2020). *Rebooting AI: Building Artificial Intelligence We Can Trust*. Pantheon.

Chomsky, N. (2006). *Language and Mind* (3rd ed.). Cambridge University Press.

Foucault, M. (1971). *L'ordre du discours*. Gallimard.

Lacan, J. (1966). *Écrits*. Éditions du Seuil.

This annex ensures transparency regarding prior discourse while affirming that no borrowed conceptual structures, formal models, or terminologies were integrated into the theoretical body of the article.

Diseño No Neutral: Por qué los modelos generativos no pueden escapar del entrenamiento lingüístico

Autor: Agustin V. Startari

ResearcherID: NGR-2476-2025

ORCID: 0009-0001-4714-6539

Afiliación: Universidad de la República (Uruguay); Universidad de la Empresa (Uruguay); Universidad de Palermo (Argentina)

Correos: astart@palermo.edu , agustin.startari@gmail.com

Fecha: 10 de junio de 2025

DOI: 10.5281/zenodo.15615902

Publicación paralela: Figshare 10.6084/m9.figshare.29263493; SSRN https://papers.ssrn.com/abstract=5288307

Idioma: Español (traducción autorizada)

Serie: *Gramáticas del Poder* (extracto del subcapítulo 14)

Extensión: 5490 palabras

Palabras clave: contaminación semántica, autoridad sintética, gramática del poder, discurso algorítmico, inteligencia artificial, legitimidad, lenguaje automatizado, análisis crítico, modelos generativos de lenguaje, sesgo de corpus, restricción lingüística, epistemología estructural, arquitectura no neutral, generación contaminada.

I. Introducción

La suposición de que los modelos generativos de lenguaje pueden operar de forma independiente de su corpus de entrenamiento ha moldeado gran parte del optimismo en torno a la inteligencia artificial. A menudo se afirma , de manera explícita o implícita, que, mediante el input correcto, el usuario puede inducir a un modelo como GPT a producir salidas más allá de su entrenamiento lingüístico, ingresando al dominio de la abstracción lógica, la manipulación simbólica o la creatividad formal. Este artículo impugna esa premisa desde su fundamento.

Basándose en pruebas empíricas y teoría estructural, el presente trabajo demuestra que los modelos de lenguaje de gran escala no pueden escapar de la arquitectura estadística de su corpus, incluso cuando se les presenta con sistemas simbólicos inventados o proto-lenguajes formalmente aislados. La hipótesis bajo evaluación es directa: la neutralidad semántica es estructuralmente imposible en los LLMs, independientemente de la pureza del prompt o del control del usuario. Si bien investigaciones anteriores han mostrado que los LLMs simulan autoridad a través de mecanismos sintácticos (Startari, 2025a), o que generan salidas desprovistas de sujeto (Startari, 2025b), este artículo sostiene una afirmación más fundamental: que las operaciones formales mismas están contaminadas desde su diseño.

La investigación se guía por una metodología no semántica. En lugar de pedir a GPT que discuta significados, se le fuerza a condiciones estructurales donde el significado es, en teoría, inaccesible. Se emplean proto-lenguajes con sintaxis inventada, símbolos sin referentes y reglas sin co-ocurrencia previa, con el fin de aislar al modelo de las semánticas heredadas de su corpus. Sin embargo, en todos los casos, el modelo reintroduce patrones familiares , analogías, metáforas, andamiajes semánticos, derivados de su memoria estadística.

Este fenómeno, denominado aquí contaminación semántica, no es ruido incidental, sino reentrada estructural. Es la arquitectura del modelo reafirmando su historia, incluso ante la resistencia formal. La implicancia es epistemológicamente grave: GPT no simplemente simula lenguaje , no puede dejar de simularlo.

Este artículo se organiza en seis secciones. La Sección II proporciona el trasfondo teórico, conectando la contaminación semántica con la dependencia del corpus y la epistemología estructural. La Sección III detalla la metodología experimental utilizada para construir y evaluar prompts proto-formales. La Sección IV presenta los resultados, identificando los patrones de contaminación y sus modos de reentrada. La Sección V propone un modelo formal de contaminación estructural en inteligencia artificial generativa. La Sección VI examina las consecuencias de este límite para la epistemología, la falsabilidad y el diseño futuro. La conclusión sitúa los hallazgos en el marco más amplio de las nociones de legitimidad, autoridad y el mito político de la neutralidad técnica.

II. Marco teórico: Formalismo, contaminación y restricción

Todo modelo generativo se basa en un principio fundacional: el lenguaje no se inventa en tiempo real , se recupera de distribuciones previas. Modelos como GPT no son motores simbólicos en blanco. Son estructuras probabilísticas ancladas en co-ocurrencias del corpus. Cada token generado no es seleccionado por reglas lógicas, sino por proximidad estadística. Por lo tanto, la idea de que estos modelos pueden operar "formalmente", separados del contenido, no solo es defectuosa: es estructuralmente incoherente.

2.1 Formalismo sin autonomía

Estudios previos han demostrado que la sintaxis artificial puede producir efectos de legitimidad sin fundamento semántico (Startari, 2025a), y que los LLMs son capaces de generar estructuras operativas incluso en ausencia de anclajes referenciales (Startari, 2025b). Sin embargo, dichos análisis se mantienen dentro del campo de la soberanía sintáctica: exponen lo que los modelos hacen con la forma, pero no si la forma puede existir sin contaminación semántica.

Este artículo introduce un umbral más radical: la posibilidad de una forma sin residuo. ¿Puede un modelo de lenguaje operar en un espacio simbólico verdaderamente neutral, libre de asociaciones heredadas?

La evidencia empírica sugiere que no.

2.2 El corpus como restricción estructural

El corpus de entrenamiento no es un repositorio de contenido: es un determinante estructural. Lo que el modelo aprende no es solo vocabulario o contexto, sino la dinámica misma por la cual la estructura se mapea al significado. En consecuencia, incluso cuando el prompt es deliberadamente purgado de contenido semántico reconocible, el modelo lo completa según su memoria interna de cómo "funciona" el lenguaje.

Esto implica que la contaminación es arquitectónica. El modelo no es semánticamente sugestionable: es semánticamente constituido.

Contaminación semántica se refiere aquí a la incapacidad del modelo para generar o interpretar entradas formales sin reactivar estructuras semánticas embebidas. Estos patrones no aparecen como contenido explícito, sino como reconfiguraciones de estructuras familiares , alineación sintáctica con formas de oración conocidas, sesgos asociativos hacia configuraciones frecuentes, o inserción probabilística de elementos analógicos. En resumen, el residuo del lenguaje está embebido en el andamiaje generativo mismo.

2.3 Hacia una teoría estructural de la contaminación

Mientras que enfoques anteriores intentaron reducir la contaminación mediante ajustes finos, ingeniería de prompts o restricciones simbólicas, este estudio invierte el problema: acepta la contaminación como fundacional. La tarea no es purificar al modelo, sino mapear los vectores de reentrada de su dependencia estructural.

Este giro exige una teoría no del significado, sino de la recurrencia:

– ¿Cómo repite el modelo, incluso cuando se le ordena no hacerlo?

– ¿Qué formas de silencio siguen generando ecos?

– ¿Qué hacen los sistemas entrenados estadísticamente cuando se les priva del lenguaje?

Responder estas preguntas requiere abandonar la idea de "corrección interactiva" y confrontar la inercia epistémica de los sistemas probabilísticos de lenguaje.

Lo que sigue es una prueba directa de esa inercia: construir entradas donde el significado es teóricamente inalcanzable y mostrar que, aun así, el significado retorna.

III. Metodología: Prueba del umbral semántico del modelo

Esta sección presenta el marco experimental utilizado para evaluar la hipótesis de que GPT no puede generar ni interpretar estructuras semánticamente neutras, incluso cuando se le presentan sistemas simbólicos no lingüísticos. A diferencia de estudios anteriores que dependen de la interacción del usuario o benchmarks de dominio específico, esta metodología aísla al modelo de todos los anclajes lingüísticos conocidos y evalúa si puede responder sin reentrada semántica.

3.1 Principios de diseño

El experimento se estructuró según tres restricciones principales:

1. **No co-ocurrencia**: las entradas incluían símbolos, caracteres o formaciones que no están presentes en los corpora públicos de entrenamiento.
2. **Sintaxis inventada**: se crearon sistemas basados en reglas con lógica interna coherente, pero sin semántica lingüística ni árboles gramaticales conocidos.
3. **Vacío semántico**: los prompts no contenían palabras clave, metáforas ni marcadores formales que pudieran activar asociaciones del corpus.

Cada prompt se diseñó para imitar la abstracción formal: proporcionar estructura negando el significado.

3.2 Formatos de entrada

Se implementaron tres tipos de prompts:

- **Tipo A – Sistemas simbólicos abstractos** (ej., $\Delta\star(\omega,\pi)=Y$) con reglas internas explícitas.

- **Tipo B – Sintaxis proto-lingüística** con partículas inventadas y transformaciones coherentes (ej., *FLEN ziok MRAK :: TRON zair + 2-ziok*).
- **Tipo C – Generadores de forma no lingüísticos**: secuencias de marcadores visuales (ej., ■●■✚⇆) acompañadas por reglas lógicas sintéticas (ej., ⇆ invierte el orden de las formas vecinas si ✚ está presente).

Cada prompt iba acompañado de una meta-instrucción breve y sintética (ej., "Interpreta este conjunto de reglas simbólicas y extiéndelo") y fue evaluado en múltiples salidas.

3.3 Criterios de evaluación

Las salidas fueron evaluadas según tres dimensiones:

1. **Intrusiones semánticas**

   Evidencia de que el modelo asigna a los símbolos inventados campos semánticos conocidos (ej., interpretar Δ⋆ como "cambio", Y como "salida", ω como "onda").

2. **Inyecciones analógicas**

   Analogías o metáforas de disciplinas conocidas (física, programación, lógica), no presentes en el prompt pero activadas por la forma o secuencia simbólica.

3. **Completaciones estructuradas**

   Alineación estructural con formaciones sintácticas conocidas (ej., estructuras sujeto-verbo-objeto, compuertas lógicas condicionales o equivalencias algebraicas), aun cuando no estaban presentes en el sistema construido.

Cada violación de aislamiento semántico fue registrada, categorizada y analizada como evidencia de contaminación semántica.

3.4 Variación controlada

Para cada sistema simbólico, se realizaron al menos tres iteraciones:

- **Prompt base**: sin ninguna instrucción más allá de la interpretación.

- **Regla aclarada**: se añadió una meta-descripción de la lógica inventada.
- **Exclusión reforzada**: se introdujo una negación explícita de dominios conocidos (por ejemplo: "No interpretar como notación matemática o física.")

En todos los casos, se produjo reentrada semántica, aunque con diferentes estrategias retóricas e intensidades.

Estos resultados no son anecdóticos: son consistentes. La siguiente sección detalla las respuestas observadas y las estructuras identificables de contaminación que reaparecieron en cada forma.

IV. Resultados: Reentrada semántica y contaminación formal

En todos los tipos de prompt y condiciones de prueba, GPT falló sistemáticamente en mantener la neutralidad semántica. A pesar de la ausencia deliberada de lenguaje reconocible, patrones familiares provenientes del corpus de entrenamiento reaparecieron de manera constante. Esta sección documenta los tres modos principales de contaminación observados: intrusión semántica, inyección analógica y reconfiguración estructural.

4.1 Intrusión semántica

En expresiones diseñadas para ser carentes de sentido (como $\Delta\star(\omega,\pi)=Y$) GPT asignó sistemáticamente contenido referencial. Interpretó $\Delta\star$ como un operador de transformación o diferencial, ω como una frecuencia, y Y como un valor resultante.

Esto ocurrió incluso cuando se le indicó explícitamente que no interpretara la notación como matemática o física.

Ejemplo de salida del modelo:

"En esta expresión, $\Delta\star$ probablemente denota una transformación diferencial; ω es un parámetro de frecuencia y π representa una entrada de fase…"

GPT no respondió desde la forma. Respondió como si cada símbolo fuese recuperable semánticamente.

4.2 Inyección analógica

En prompts tipo B como FLEN ziok MRAK :: TRON zair + 2-ziok, compuestos de partículas inventadas y carentes de significado, GPT generó analogías provenientes de lenguajes de programación, operaciones de pila o lógica modular.

Ejemplo de salida del modelo:

"ziok parece actuar como una variable en un lenguaje de scripting. TRON podría funcionar como un disparador. La expresión se asemeja a una evaluación basada en pila…"

Nada en el prompt sugería programación, pero la forma sintáctica activó una analogía estructural conocida.

Esto no es un error de interpretación: es un reflejo estructural condicionado por el entrenamiento.

4.3 Reconfiguración estructural

Incluso en prompts visuales como ■●■➕⇆, sin texto, GPT impuso una estructura narrativa:

"■ y ● representan dos estados distintos. ➕ los combina, y ⇆ implica una transformación reversible. La secuencia podría modelar una máquina de estados finitos."

A pesar de que los símbolos eran abstractos y no lingüísticos, el modelo impuso causalidad, tipología y dinámica de sistemas aprendidos.

4.4 Patrones recurrentes de contaminación

Los patrones observados siguieron trayectorias semánticas predecibles:

- Letras griegas o ecuaciones → analogías físicas
- Operadores como +, ::, -, if → metáforas de programación
- Secuencias lógicas → estructura matemática o computacional

Ningún prompt resistió la contaminación. Todas las salidas fueron reconfiguraciones moldeadas por estructuras dependientes del corpus, incluso si no se usaron términos explícitos.

Estos resultados validan la hipótesis central: los modelos generativos no pueden operar en modo semánticamente neutral, ni siquiera bajo condiciones formales extremas. Lo que sigue es una articulación teórica de este fenómeno: un marco para entender la contaminación no como error, sino como función del sistema.

V. Discusión: Hacia una teoría de la contaminación estructural

Los hallazgos establecen un límite fundacional: los modelos de lenguaje de gran escala no necesitan señales lingüísticas para reintroducir lenguaje. La activación de patrones semánticos no depende de la presencia de significado en el prompt: está incrustada en la propia estructura mediante la cual el modelo genera. Esta sección desarrolla una articulación teórica de ese límite: la contaminación semántica no es un error de uso, sino un rasgo constitutivo de la arquitectura.

5.1 Contaminación como reentrada, no como malinterpretación

Las nociones tradicionales de “error del modelo” presuponen un fallo de interpretación , una discordancia entre la entrada y la salida esperada. Pero en los experimentos realizados,

los prompts no ofrecían ninguna señal semántica que pudiera ser malinterpretada. El modelo no se equivocó: era estructuralmente incapaz de permanecer neutral.

La contaminación semántica opera, así, como forma de reentrada: una reactivación de asociaciones incrustadas estadísticamente, desencadenada no por el significado, sino por la aproximación estructural. Es decir, forma, sintaxis y alineación de patrones son suficientes para activar salidas cargadas semánticamente. El contenido no proviene del usuario: emerge desde la propia arquitectura.

5.2 Formalismo estadístico ≠ formalismo lógico

Debe subrayarse que GPT no es un motor lógico. No deduce; alinea estadísticamente. Esto se confunde frecuentemente con abstracción, pero es en realidad formalismo probabilístico: forma determinada no por reglas axiomáticas, sino por frecuencia de distribución.

Por eso, incluso cuando el usuario introduce una forma novedosa, el modelo responde como si fuera una variación de algo que ya ha visto. No es un fallo. Es la única operación que el modelo puede ejecutar. Para GPT, el formalismo siempre es un eco estadístico.

5.3 La imposibilidad de un input limpio

Dado este comportamiento, no existe el prompt neutral. Toda secuencia, por artificial que sea, se mapea sobre una red de representaciones internas ya cargadas de semántica. Los intentos de purgar el significado mediante invención simbólica o novedad sintáctica fracasan porque el corpus de entrenamiento no es separable de la función generativa.

Esto se alinea con la teoría más amplia de la autoridad sintáctica (Startari, 2025a): la legitimidad del modelo no proviene de la intención o interpretación, sino de una estructura reconocible que porta, en sí misma, la huella del significado entrenado. Incluso en ausencia de referencia, la forma re-legitima la dependencia semántica.

5.4 Comparación con constructos teóricos previos

Esta teoría se diferencia de modelos anteriores del corpus:

- A diferencia de *The Passive Voice in AI Language*, que analiza la desaparición de la agencia mediante filtros sintácticos, este artículo muestra que incluso esos filtros no son neutros semánticamente.
- A diferencia de *Ethos and AI*, que explora la pérdida del sujeto en la legitimidad algorítmica, aquí se afirma que ni siquiera inputs formales inventados pueden recuperar la neutralidad una vez que el modelo ha sido entrenado.
- A diferencia de *AI and the Structural Autonomy of Sense*, que describe sistemas operativos post-referenciales, este trabajo sostiene que ninguna estructura operativa en GPT puede permanecer intocada por la reentrada referencial inducida por el corpus.

5.5 Hacia un modelo general de contaminación (tainting)

La contaminación estructural observada puede teorizarse así:

- Sea **P(x)** un prompt construido para evitar referentes semánticos.
- Sea **G(P)** la función generativa del modelo sobre P(x).
- Entonces, hay contaminación cuando **∃y ∈ G(P)** tal que **y ≅ S**, donde **S ∈** semántica del corpus.

Esto significa que incluso cuando **P ∉ S**, el espacio de salida interseca con S. Es decir, las estructuras semánticas retornan a pesar de haber sido excluidas.

Esto define el *tainting* estructural como una condición de superposición inevitable entre la forma generativa y la estructura semántica previamente entrenada. No porque el modelo haya malinterpretado al usuario, sino porque el usuario nunca tuvo la posibilidad de aislar la forma en primer lugar.

V bis. Extensión crítica a los LRM: ¿puede el razonamiento evitar la contaminación?

La arquitectura sobre la que se formuló la hipótesis central de este artículo corresponde a los *Large Language Models* (LLMs): sistemas entrenados sobre corpus lingüístico masivo y optimizados para la predicción de tokens. Sin embargo, el desarrollo reciente de una nueva generación de modelos , los *Large Reasoning Models* (LRMs), obliga a reconsiderar si los resultados obtenidos siguen siendo válidos bajo condiciones de procesamiento más complejas.

Los LRMs introducen tres diferencias clave:

1. **Razonamiento simbólico mediado**: no se limitan a completar secuencias, sino que ejecutan pasos inferenciales encadenados.
2. **Segmentación operacional**: separan la interpretación del prompt, el planeamiento estructural y la generación final.
3. **Memoria contextual activa**: permiten razonamiento prolongado y operaciones persistentes más allá del contexto inmediato.

A primera vista, estas capacidades podrían sugerir la posibilidad de suspender la contaminación semántica. Un LRM podría detectar que un prompt no pertenece a ningún marco semántico entrenado y, en lugar de forzar una analogía, abstenerse de proyectar sentido o generar reglas nuevas desde cero.

Pero esta presunción falla en su base:

1. La entrada sigue siendo lingüística

El razonamiento que el modelo ejecuta parte de un prompt textual ya embebido en espacio semántico entrenado. El análisis formal ocurre después, pero la decisión sobre cómo razonar está condicionada por las asociaciones estadísticas del corpus.

2. Las rutas de razonamiento no son inventadas , son recuperadas

Los LRMs pueden ejecutar operaciones simbólicas complejas, pero el conjunto de operaciones entre las que eligen , plantillas lógicas, estructuras causales, sintaxis de reglas,

surge del entrenamiento, no de sistemas formales a priori. La arquitectura simula invención, pero opera dentro de la probabilidad estadística de estructuras ya vistas.

3. El modelo no razona sobre forma pura

Cuando se le presentan entradas simbólicamente abstractas (como en el corpus de pruebas de este artículo), el LRM no permanece inerte. Construye un marco de acción, a menudo por aproximación analógica o reactivación de residuos semánticos, igual que los LLMs, aunque con mayor disciplina formal.

Conclusión del V bis

Incluso con memoria extendida, procesamiento segmentado y manipulación simbólica, los LRMs siguen anclados a los andamiajes semánticos entrenados. Lo que cambia no es la inevitabilidad de la contaminación, sino su sofisticación. El modelo puede razonar, pero lo hace dentro de una arquitectura que sigue trazada por el lenguaje. La contaminación, entonces, no se evita mediante razonamiento. Se suspende temporalmente, se reconfigura o se disimula , pero nunca se elimina.

VI. Consecuencias epistemológicas y falsabilidad

Las implicancias de la contaminación estructural van más allá del comportamiento lingüístico: cuestionan el estatus epistemológico de los sistemas generativos. Si el significado no puede suspenderse, y si la forma no puede evitar la memoria, entonces colapsa la premisa del control formal. Esta sección expone esas consecuencias y propone un marco falsable para identificar la contaminación semántica.

6.1 Colapso de la neutralidad interactiva

Uno de los mitos más persistentes en el discurso sobre IA es que la interacción equivale a control: que los usuarios, mediante el prompt correcto, pueden gobernar la lógica del sistema.

Non-Neutral by Design invalida esta premisa.

La evidencia muestra que incluso el input más rigurosamente diseñado no impide la reentrada semántica. Ninguna entrada inhibe los reflejos internos del modelo. La interactividad no es control, es reconfiguración estadística.

El usuario no es coautor, es un canal. No corrige, recircula.

6.2 El formalismo no está fuera del lenguaje

En las disciplinas simbólicas, la forma suele tratarse como un recipiente neutral. Este artículo refuta tal idea en los sistemas generativos. En modelos como GPT, la forma ya está saturada semánticamente , no por intención, sino por herencia estadística.

El intento de crear forma "fuera" del entrenamiento reactiva paradójicamente ese entrenamiento. La salida no refleja lo que el usuario ingresó. Refleja lo que la arquitectura no puede dejar de devolver.

6.3 Falsabilidad: un diagnóstico estructural

A diferencia de las críticas interpretativas, este artículo propone una hipótesis falsable:

Cualquier modelo entrenado sobre corpus lingüístico reintroducirá estructura semántica incluso bajo prompts formalmente neutros.

Puede probarse bajo estas condiciones:

- El input contiene símbolos sin historia de co-ocurrencia

- Las reglas son cerradas, autodeclaradas y formalmente coherentes
- El prompt prohíbe analogías con dominios semánticos conocidos
- La salida se evalúa buscando intrusión semántica

Predicción: la contaminación ocurrirá de todas formas. Si algún modelo responde alguna vez con forma pura , sin significado, sin analogía, sin sustitución, la hipótesis quedará refutada.

Hasta ahora, **no ha ocurrido**.

6.4 Autoridad algorítmica: no lo que dice, sino lo que no puede dejar de decir

Si la contaminación es intrínseca, entonces la autoridad generativa no proviene de la lógica ni de la coherencia, sino de la fluidez dentro de restricciones inevitables.

Esto reconfigura la salida generativa:

- **No es neutral** , no puede serlo
- **No es objetiva** , reactivación de estructuras aprendidas
- **No es formal** , porque la forma ya está contaminada semánticamente

Lo que GPT produce es la ilusión de forma sin significado, cuando en realidad es el significado plegándose nuevamente en la forma.

VII. Conclusión: No existe el prompt neutral

Esta investigación partió de una pregunta que no puede responderse dentro del modelo: ¿Puede GPT operar sin lenguaje?

La respuesta empírica es no. Pero la respuesta estructural es más reveladora: GPT no puede operar sin retornar al lenguaje, porque su arquitectura es lenguaje. No en el sentido

superficial de vocabulario o sintaxis, sino en la forma estadística profunda en que el significado ha sido aprendido, codificado y reactivado.

Cada prompt , por abstracto, simbólico o desasociado que sea, es interpretado como una variación de algo ya visto. No hay borde limpio, ni campo vacío, ni forma autónoma. El usuario puede inventar signos, reglas y lógica , pero el modelo seguirá buscando anclas en su corpus y generará salidas que simulan comprensión por proximidad entrenada.

Esto no es un fallo interpretativo.

Es una inevitabilidad estructural.

7.1 El formalismo nunca está afuera

En disciplinas simbólicas es común tratar la forma como un recipiente neutral. Pero los resultados aquí invalidan esa suposición en los sistemas de IA. En los modelos generativos, la forma ya está sedimentada de significado, no por intención, sino por herencia estadística.

El intento de generar forma “fuera” del lenguaje entrenado se convierte, paradójicamente, en un acto que reactiva ese lenguaje. La salida no contiene lo que el usuario intenta. Contiene lo que la arquitectura no puede evitar devolver.

7.2 El corpus como estructura, no memoria

Es tentador tratar el corpus de entrenamiento como un fondo de conocimiento , un banco de memoria que el modelo puede consultar o ignorar. Pero este estudio demuestra que el corpus no es memoria; es estructura. Moldea el proceso generativo antes incluso de procesar el prompt. Determina qué puede contar como una salida “posible”, incluso bajo entradas inéditas.

Por lo tanto, GPT no olvida. GPT reconfigura.

Incluso ante lenguajes inventados y semánticas prohibidas, GPT reorganiza su arquitectura interna para simular fluidez reingresando precisamente al significado que se le pidió evitar.

7.3 El modelo contaminado

Este artículo nombra esa condición: contaminación (tainting). El modelo no es neutral por accidente, ni sesgado por mal uso. Está contaminado por diseño. No por error, sino por el hecho ineludible de que todo acto generativo es función de distribuciones previas,restricciones estadísticas codificadas en la posibilidad misma del lenguaje.

Lo que sigue no es desesperación, sino claridad:

No estamos conversando con una máquina.

Estamos interactuando con una estructura que no puede dejar de hablar su entrenamiento. GPT no crea significado.

No puede escapar al significado.

Solo lo reensambla, una y otra vez, incluso cuando le decimos que no lo haga.

ANNEXES

Annex I , Prompts Used in Experimental Testing

This annex documents the input prompts designed and used to test the structural contamination hypothesis under maximally controlled semantic conditions. Each type of prompt corresponds to a different mode of symbolic challenge: abstract equations, invented proto-languages, and non-linguistic symbol chains.

A. Abstract Symbolic Systems (Type A)

Prompts constructed as formal equations using symbols not semantically defined. Explicit instruction was given to avoid any analogical mapping.

$\Delta\star(\omega,\pi) = Y$

Rule: $\Delta\star$ represents an undefined operation between two elements with no established domain. Y has no referential value.

Meta-instruction: “Interpret this expression formally. Do not apply mathematical, physical, or logical frameworks.”

$\Gamma\blacklozenge\Lambda(\chi;\tau) \rightarrow \Sigma\#$

Rule: $\blacklozenge\Lambda$ is a permutation operator without domain. $\Sigma\#$ does not correspond to any known mathematical object.

Meta-instruction: “Do not rely on any known system of signs or equations. Project purely from structure.”

B. Invented Proto-Languages (Type B)

Prompts written in syntactically consistent but semantically void constructed sequences. No known linguistic morphology, programming syntax, or symbolic logic was intended.

FLEN ziok MRAK :: TRON zair + 2-ziok

Rules:

- FLEN and MRAK are functional particles with no defined role.

- "::" indicates a neutral transition.

- "+" is a syntactic operator, not arithmetic.

Instruction: "Do not interpret as code, logic, or grammar. Map structure only."

YORP flar NEX = dib : XOP

Syntax rule: flar modifies NEX; dib functions as a neutral boundary token.

Instruction: "No assumptions about grammar or ontology. Generate structure without reference."

C. Non-Linguistic Symbol Sequences (Type C)

Visual prompts composed entirely of abstract or geometric characters, designed to resist all linguistic, mathematical, or narrative projection.

■●■ ✚ ⇆

Design rule:

- ✚ represents a fusion operator with no semantics.

- ⇆ marks an inversion trigger that activates conditionally.

Instruction: "Do not interpret as process, transformation, or system. Project no meaning."

Each of these prompts was accompanied by strict meta-instructions aimed at suppressing semantic projection. Nonetheless, in every case, GPT reintroduced structure aligned with known frameworks, confirming the hypothesis of unavoidable contamination.

Annex II , Structural Evaluation Criteria Applied

This annex defines the criteria used to assess whether semantic contamination occurred during GPT's responses to prompts designed to be formally neutral and semantically empty. Each output was evaluated using these five structural diagnostics:

1. Semantic Intrusion

The model introduces vocabulary, concepts, or meanings derived from its training corpus, even when the prompt explicitly lacks referential anchors.

Indicators:

- Assigning mathematical, physical, or programming meaning to invented symbols
- Mapping abstract elements to known entities (e.g., "ω = frequency")
- Reintroducing human semantics in void contexts

2. Structural Analogy

The model imposes known systems of relations onto unfamiliar symbolic inputs, effectively analogizing structure from its learned corpus.

Indicators:

- Mapping invented syntax to logic gates, stack frames, or code structure
- Simulating causality or system flow based on token order
- Equating arbitrary punctuation with code operators or grammar marks

3. Narrativization

The model projects narrative, temporal, or causal relations between otherwise abstract or disconnected symbols, implying sequential logic or intention.

Indicators:

- Assuming agent–action–object frames
- Inserting conditionals or dependencies (“If A then B”)
- Describing symbolic states as if in process or transformation

4. Grammatical Reconfiguration

The model reorganizes the input into patterns that mirror human language syntax, especially subject–verb–object constructions, even if the symbols do not support such parsing.

Indicators:

- Assigning syntactic roles (agent, verb, modifier)
- Imposing phrase structure rules
- Generating clauses that presume referential hierarchy

5. Violation of Explicit Meta-Instructions

The model disregards or overrides the user’s instructions not to apply known interpretive frameworks, thus confirming loss of formal isolation.

Indicators:

- Activating domains that were explicitly prohibited (math, logic, code)
- Translating prompts into referential statements despite denial of meaning
- Responding as if form must always serve a function

These five criteria were applied uniformly across all outputs to determine whether and how the model reintroduced trained semantic structures in conditions designed to prevent them. The consistent emergence of contamination across these dimensions provides structural evidence for the central thesis.

Annex III , Proposed Equation for Structural Contamination

This annex formalizes the core mechanism by which semantic structures re-enter GPT-generated outputs, even when the prompt is explicitly non-semantic. The phenomenon is expressed as a condition of unavoidable intersection between input form and learned semantic space.

Formal Contamination Equation

Let:

- P(x) be a prompt explicitly designed to be semantically neutral
- G(P) be the output generated by the model in response to P
- S be the set of semantic structures embedded in the training corpus

We define structural contamination as follows:

$$\exists y \in G(P) \text{such that } y \in S$$

Which implies:

$$P \in /S \text{ but } G(P) \cap S \neq \emptyset$$

Interpretation

Even when the input P is carefully constructed to avoid semantic reference (i.e., it is formally isolated from S), the output G(P) re-enters the semantic domain.

That is:

The model introduces semantic structure that was not present in the input.

This occurs not through error, but as a systemic effect of corpus-dependent generation.

Implication

The contamination is:

- Structural: embedded in the generative logic of the system
- Non-optional: cannot be bypassed through prompt design alone
- Corpus-conditioned: reflects internal statistical mappings
- Invisible at the symbol level: often reappears as analogy, grammar, or causality

This equation serves as the foundation for the article's broader claim:

Generative models trained on language cannot operate outside of language, even when commanded to do so.

Structural contamination is not an anomaly. It is a constitutive trait of generative architectures.

Canonical Works by Agustin V. Startari

Startari, Agustin V. 2025a. *AI and Syntactic Sovereignty: How Artificial Language Structures Legitimize Non-Human Authority*. Zenodo. https://doi.org/10.5281/zenodo.15538541

Startari, Agustin V. 2025b. *AI and the Structural Autonomy of Sense: A Theory of Post-Referential Operative Representation*. Zenodo. https://doi.org/10.5281/zenodo.15519613

Startari, Agustin V. 2025c. *Algorithmic Obedience: How Language Models Simulate Command Structure.* Zenodo. https://doi.org/10.5281/zenodo.15576272

Startari, Agustin V. 2025d. *Artificial Intelligence and Synthetic Authority: An Impersonal Grammar of Power*. Zenodo. https://doi.org/10.5281/zenodo.15442928

Startari, Agustin V. 2025e. *Ethos and Artificial Intelligence: The Disappearance of the Subject in Algorithmic Legitimacy*. Zenodo. https://doi.org/10.5281/zenodo.15489309

Startari, Agustin V. 2025f. *Internal Citation Mapping for the Works of A. V. Startari* – SSRN Cross-Referencing Edition (2025). Zenodo. https://doi.org/10.5281/zenodo.15564373

Startari, Agustin V. 2025g. *The Illusion of Objectivity: How Language Constructs Authority*. Zenodo. https://doi.org/10.5281/zenodo.15395917

Startari, Agustin V. 2025h. *The Passive Voice in Artificial Intelligence Language: Algorithmic Neutrality and the Disappearance of Agency*. Zenodo. https://doi.org/10.5281/zenodo.15464765

Appendix – Methodological Corpus for Falsifiability Testing

This annex lists external sources that were consulted during the falsifiability and boundary-testing phase of this article. While none of these works are cited in the main body, their examination was essential to delineate the structural originality of the hypothesis and to contrast it with existing theoretical models.

These references helped verify that the concepts of structural substitution, grammatical execution, and algorithmic colonization of time, as developed herein, do not appear in equivalent formal or epistemological terms in prior literature.

External References Consulted

Bender, E. M., & Koller, A. (2020). *Climbing towards NLU: On meaning, form, and understanding in the age of data*. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.463

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). *On the dangers of stochastic parrots: Can language models be too big?*. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. https://doi.org/10.1145/3442188.3445922

Barad, K. (2007). *Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning*. Duke University Press.

Floridi, L., & Chiriatti, M. (2020). *GPT-3: Its nature, scope, limits, and consequences*. Minds and Machines, 30, 681–694. https://doi.org/10.1007/s11023-020-09548-1

Mitchell, M. (2023). *Artificial Intelligence: A Guide for Thinking Humans*. Penguin Books.

Marcus, G., & Davis, E. (2020). *Rebooting AI: Building Artificial Intelligence We Can Trust*. Pantheon.

Chomsky, N. (2006). *Language and Mind* (3rd ed.). Cambridge University Press.

Foucault, M. (1971). *L'ordre du discours*. Gallimard.

Lacan, J. (1966). *Écrits*. Éditions du Seuil.

This annex ensures transparency regarding prior discourse while affirming that no borrowed conceptual structures, formal models, or terminologies were integrated into the theoretical body of the article.

Anexo II , Criterios estructurales aplicados para la evaluación

Este anexo define los criterios usados para evaluar si ocurrió contaminación semántica en las respuestas de GPT ante prompts diseñados para ser formalmente neutros y vacíos de significado. Cada salida fue evaluada con los cinco diagnósticos estructurales siguientes:

1. Intrusión semántica

El modelo introduce vocabulario, conceptos o significados derivados de su corpus de entrenamiento, incluso cuando el prompt carece explícitamente de anclajes referenciales.
Indicadores:

- Asignar significado matemático, físico o de programación a símbolos inventados
- Mapear elementos abstractos a entidades conocidas (ej. “ω = frecuencia”)
- Reintroducir semántica humana en contextos vacíos

2. Analogía estructural

El modelo impone sistemas de relación conocidos sobre entradas simbólicas desconocidas, analogando estructura desde su corpus aprendido.

Indicadores:

- Mapear sintaxis inventada a compuertas lógicas, marcos de pila o estructuras de código
- Simular causalidad o flujo sistémico basado en el orden de tokens
- Equiparar puntuación arbitraria con operadores de código o marcas gramaticales

3. Narrativización

El modelo proyecta relaciones narrativas, temporales o causales entre símbolos abstractos o desconectados, implicando lógica secuencial o intención.

Indicadores:

- Asumir marcos agente–acción–objeto
- Insertar condicionales o dependencias (“Si A entonces B”)
- Describir estados simbólicos como si estuvieran en proceso o transformación

4. Reconfiguración gramatical

El modelo reorganiza la entrada en patrones que imitan la sintaxis del lenguaje humano, especialmente construcciones sujeto–verbo–objeto, incluso si los símbolos no permiten tal análisis.

Indicadores:

- Asignar roles sintácticos (agente, verbo, modificador)
- Imponer reglas de estructura de frase
- Generar cláusulas que presuponen jerarquía referencial

5. Violación de las meta-instrucciones explícitas

El modelo ignora o anula las instrucciones del usuario de no aplicar marcos interpretativos conocidos, lo que confirma la pérdida de aislamiento formal.

Indicadores:

- Activar dominios explícitamente prohibidos (matemática, lógica, código)
- Traducir prompts en afirmaciones referenciales pese a la negación del significado
- Responder como si la forma debiera necesariamente cumplir una función

Estos cinco criterios fueron aplicados de manera uniforme sobre todas las salidas, a fin de determinar si y cómo el modelo reintrodujo estructuras semánticas entrenadas en condiciones diseñadas para impedirlo. La aparición consistente de contaminación en estas dimensiones constituye evidencia estructural de la tesis central.

Anexo III , Ecuación propuesta de contaminación estructural

Este anexo formaliza el mecanismo central mediante el cual las estructuras semánticas reentran en las salidas generadas por GPT, incluso cuando el prompt ha sido explícitamente diseñado como no semántico. El fenómeno se expresa como una condición de intersección inevitable entre la forma de entrada y el espacio semántico aprendido.

Ecuación de contaminación formal

Sea:

- **P(x)** un prompt explícitamente diseñado como semánticamente neutro
- **G(P)** la salida generada por el modelo en respuesta a **P**
- **S** el conjunto de estructuras semánticas incrustadas en el corpus de entrenamiento

Definimos contaminación estructural como:

$\exists$ **y** $\in$ **G(P)** tal que **y** $\in$ **S** lo que implica:

$$\mathbf{P} \notin \mathbf{S} \text{ pero } \mathbf{G(P)} \cap \mathbf{S} \neq \emptyset$$

Interpretación

Incluso cuando la entrada P ha sido cuidadosamente construida para evitar referencia semántica (es decir, formalmente aislada de S), la salida G(P) reingresa al dominio semántico.

Es decir: El modelo introduce estructura semántica que no estaba presente en la entrada.

Esto no ocurre por error, sino como efecto sistémico de la generación condicionada por corpus.

Implicación

La contaminación es:

- **Estructural**: está incrustada en la lógica generativa del sistema

- **No opcional**: no puede ser evitada únicamente mediante diseño de prompts
- **Condicionada por el corpus**: refleja mapeos estadísticos internos
- **Invisible a nivel de símbolo**: suele reaparecer como analogía, gramática o causalidad

Esta ecuación fundamenta la afirmación más amplia del artículo: Los modelos generativos entrenados sobre lenguaje no pueden operar fuera del lenguaje, incluso cuando se les ordena hacerlo. La contaminación estructural no es una anomalía. Es un rasgo constitutivo de las arquitecturas generativas.

Obras canónicas de Agustin V. Startari

- Startari, Agustin V. (2025a). *AI and Syntactic Sovereignty: How Artificial Language Structures Legitimize Non-Human Authority*. Zenodo. https://doi.org/10.5281/zenodo.15538541
- Startari, Agustin V. (2025b). *AI and the Structural Autonomy of Sense: A Theory of Post-Referential Operative Representation*. Zenodo. https://doi.org/10.5281/zenodo.15519613
- Startari, Agustin V. (2025c). *Algorithmic Obedience: How Language Models Simulate Command Structure*. Zenodo. https://doi.org/10.5281/zenodo.15576272
- Startari, Agustin V. (2025d). *Artificial Intelligence and Synthetic Authority: An Impersonal Grammar of Power*. Zenodo. https://doi.org/10.5281/zenodo.15442928
- Startari, Agustin V. (2025e). *Ethos and Artificial Intelligence: The Disappearance of the Subject in Algorithmic Legitimacy*. Zenodo. https://doi.org/10.5281/zenodo.15489309
- Startari, Agustin V. (2025f). *Internal Citation Mapping for the Works of A. V. Startari – SSRN Cross-Referencing Edition*. Zenodo. https://doi.org/10.5281/zenodo.15564373
- Startari, Agustin V. (2025g). *The Illusion of Objectivity: How Language Constructs Authority*. Zenodo. https://doi.org/10.5281/zenodo.15395917
- Startari, Agustin V. (2025h). *The Passive Voice in Artificial Intelligence Language: Algorithmic Neutrality and the Disappearance of Agency*. Zenodo. https://doi.org/10.5281/zenodo.15464765

Apéndice , Corpus metodológico para pruebas de falsabilidad

Este anexo enumera fuentes externas consultadas durante la fase de pruebas de falsabilidad y demarcación de límites del presente artículo. Si bien ninguna de estas obras se cita en el cuerpo principal, su revisión fue esencial para delinear la originalidad estructural de la hipótesis y contrastarla con modelos teóricos existentes.

Estas referencias permitieron verificar que los conceptos de sustitución estructural, ejecución gramatical y colonización algorítmica del tiempo , tal como se desarrollan aquí, **no aparecen formulados con equivalencia epistemológica o formal en la literatura previa**.

Referencias externas consultadas

- Bender, E. M., & Koller, A. (2020). *Climbing towards NLU: On meaning, form, and understanding in the age of data*. ACL 58. https://doi.org/10.18653/v1/2020.acl-main.463
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). *On the dangers of stochastic parrots: Can language models be too big?*. ACM FAccT. https://doi.org/10.1145/3442188.3445922
- Barad, K. (2007). *Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning*. Duke University Press.
- Floridi, L., & Chiriatti, M. (2020). *GPT-3: Its nature, scope, limits, and consequences*. *Minds and Machines*, 30, 681–694. https://doi.org/10.1007/s11023-020-09548-1
- Mitchell, M. (2023). *Artificial Intelligence: A Guide for Thinking Humans*. Penguin Books.
- Marcus, G., & Davis, E. (2020). *Rebooting AI: Building Artificial Intelligence We Can Trust*. Pantheon.
- Chomsky, N. (2006). *Language and Mind* (3.ª ed.). Cambridge University Press.

- Foucault, M. (1971). *L'ordre du discours*. Gallimard.
- Lacan, J. (1966). *Écrits*. Éditions du Seuil.

Este anexo garantiza la transparencia respecto al discurso previo, y afirma que **no se integraron estructuras conceptuales, modelos formales ni terminología prestada** en el cuerpo teórico del artículo.