Is It Really Hallucinating, and Why? Toward a UFO-Based Ontology for Categorizing LLM Errors

Lucas Maddalena[1,*], Fernanda Baião[1], Tiago Prince Sales[2] and Giancarlo Guizzardi[2]

[1]Department of Industrial Engineering, Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Rua Marquês de São Vicente, 225, Rio de Janeiro, 22451-900, Brazil

[2]Semantics, Cybersecurity & Services, University of Twente, Enschede, The Netherlands

Abstract

Large language models (LLMs) fail in several ways that current practice labels with the same (overloaded) term, "hallucination." Collapsing distinct (semantic) meanings into a single (syntactic) term hinders the identification and analysis of what kind of error occurred, what it depends on, or what to do about it. We present a reference ontology of LLM errors, grounded in the Unified Foundational Ontology (UFO) and built in OntoUML, that precisely identifies and categorizes the ontological nature of an erroneous output by the mode of truthmaking failure it instantiates, the target against which it must be verified, and the dependence structure that produced it. The categorization is the basis, or foundation, for explanation and diagnosis: the explanatory force of a categorized error is the *truthmaker* it identifies (the entity to verify and, if wrong, contest), rather than the label itself. Once an error is identified and categorized, a neuro-symbolic explainer automatically generates an action-licensing account rather than a binary flag, and ties it to a root cause. We pose competency questions (CQs) as contrastive requests for explanation, illustrate them on worked cases, and outline benchmark-scale instantiation as ongoing work.

Keywords

LLM hallucination, reference ontology, UFO, OntoUML, explainable AI, error taxonomy

1. Introduction

Large language models (LLMs) are increasingly deployed in high-stake domains where a wrong answer carries real cost, such as clinical decision support, legal drafting, and financial analysis. There, the response to an error should not be restricted to its detection, but also include its explanation: to a clinician deciding whether to trust an output, or an engineer deciding how to prevent its recurrence. An explanation is useful only if it is *faithful* (it identifies the actual truthmaking failure, not a plausible surrogate) and *actionable* (it says what to verify or change) [1]. A binary "this output is a hallucination" is neither: it reports that something is wrong without saying what kind, and so warrants no particular response.

The label compounds the problem: "hallucination" is a borrowed, anthropomorphic term whose usage lacks any consensus definition [2]. It is read pragmatically as a producer indifferent to truth, in Frankfurt's sense of bullshit [3, 4], or clinically as a perceptual metaphor of doubtful fit [5]. But the obstacle is not the word; it is the absence of a principled account of *what kinds of error an LLM makes*, precise enough for an explanation tool or an engineer to build on.

We propose a reference ontology of LLM errors, grounded in UFO and built in OntoUML, which categorizes an erroneous output and explains it by the kind of *truthmaking failure* it instantiates and the *target* against which it must be verified. Hence, we take explanation in the precise sense of *ontological unpacking* from Guizzardi and Guarino [1]: to explain an error is to reveal its *truthmaker*, the entity responsible for the assertion's truth value. A categorized error explains because it names that truthmaker:

Proceedings of FOIS 2026 Satellite events co-located with the 16th International Conference on Formal Ontology in Information Systems (FOIS 2026), September 21-25, 2026, Vitória, Brazil.

*Corresponding author.

lucasmadda@aluno.puc-rio.br (L. Maddalena); fbaiao@puc-rio.br (F. Baião); t.princesales@utwente.nl (T. P. Sales); g.guizzardi@utwente.nl (G. Guizzardi)

0000-0001-6411-4259 (L. Maddalena); 0000-0001-7932-7134 (F. Baião); 0000-0002-5385-5761 (T. P. Sales); 0000-0002-3452-553X (G. Guizzardi)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-27

CEUR-WS.org/Vol-4342/short3.pdf

the type fixes a falsifiable proposition (the cited source exists, the world is thus-and-so, the claim follows from the evidence) and the entity to verify and, if wrong, contest; it is not a self-explanatory label. The categorization thus underwrites two kinds of explanation: a *constitutive* one (what truthmaking failure the error *is*, which fixes the verification target) and an *etiological* one (what in the generation process *produced* it).

Thus, our objectives in this research are threefold:

1. (O1) to provide the categorization, delivered as preliminary work;
2. (O2) to operationalise it in gUFO/OWL so cases can be instantiated and queried;
3. (O3) to demonstrate its utility for explanation and diagnosis, which we illustrate on example cases with benchmark-scale instantiation as ongoing work.

The work extends a review of the XAI–ontology intersection [6] and the case for ontologies as explanation scaffolding [7].

2. The problem

Existing work does classify LLM hallucinations: intrinsic vs. extrinsic errors [8], factuality vs. faithfulness and the retrieval-augmented (RAG) setting [9], and RAG-specific subtype lists [10]. But these classify by surface symptom or source without saying what kind of entity a hallucination is, what it depends on, or what distinguishes the subtypes at the level of their truthmakers. The classification is descriptive, not ontologically grounded.

A flat label cannot license differentiated action. Consider two errors a single "hallucination" tag merges: a citation to a source that does not exist, and a claim that contradicts the world. Diagnosing the first is an existence check (does the document exist?), the second a correspondence check against external fact. These are different operations, not faster and slower versions of one. An explanation that conflates them cannot say which to perform, and an engineer cannot route the error to the right fix.

Etiology is missing for the same reason. The weights aggregate a training corpus that may itself hold conflicting commitments (intra-memory knowledge conflicts [11]), so a parametric error has, on our account, a *diffuse* truthmaker with no single source proposition; a RAG error, by contrast, multiplies *traceable* sources: wrong or insufficient retrieval, a wrong document, or a parametric/contextual blend [10, 9]. A taxonomy blind to this dependence structure cannot connect an error to its cause.

What is missing, then, is a foundationally grounded categorization of LLM errors, by the mode of truthmaking failure, the target against which it is assessed, and the dependence structure that produced it: one that licenses categorized, faithful explanations to contrastive why-questions and ties each error to a root cause. Supplying it is this paper's contribution.

3. Background and related work

The ontology is built in *OntoUML* [12], whose primitives realise the distinctions of the *Unified Foundational Ontology* (UFO) [13, 14]; we author it in the textual syntax *Tonto* [15] and operationalise it through the lightweight implementation *gUFO* [16]. Existing hallucination taxonomies [8, 9, 10] enumerate subtypes but, as argued above, remain descriptive rather than ontologically grounded. The role of ontologies as scaffolding for explanation, and for neuro-symbolic XAI [1, 7, 17], motivates applying that programme to the specific problem of categorizing LLM errors. Ontology-driven XAI has largely targeted the explanation of classifiers and knowledge graphs; where it does address generative errors (e.g. KEA Explain [17]), it detects and explains hallucinations without a foundationally grounded categorization of them. Our contribution gives the flat taxonomies the truthmaker and verification target that turn a descriptive subtype label into an action-licensing one.

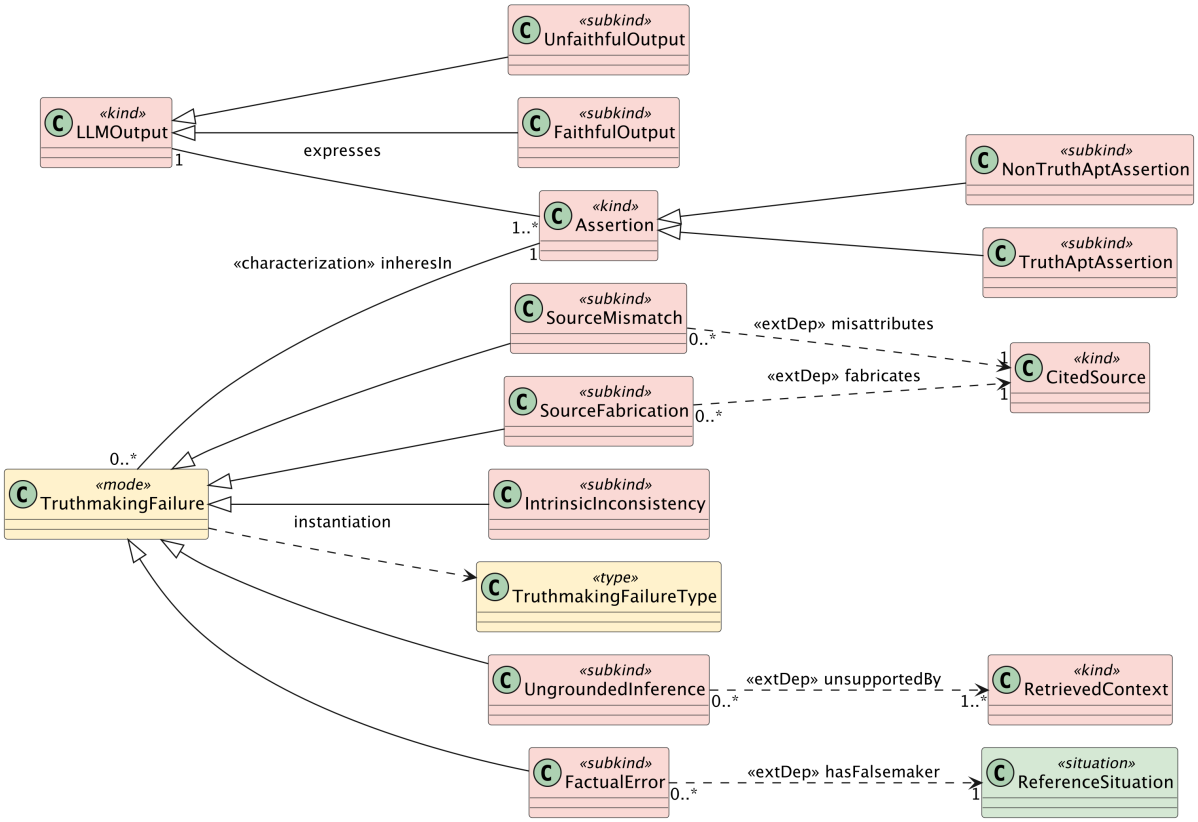

Figure 1: The categorization in OntoUML, rendered from the Tonto model: a `TruthmakingFailure` (*mode*) inheres in an `Assertion` and, through its five subtypes, externally depends (*extDep*) on the verification target it must be checked against. Box colour follows the UFO stereotype.

Table 1
The five subtypes of `TruthmakingFailure`: a verification target and a failure mode fix each one and license its verification operation.

Error type	Verification target	Failure mode	Operation
`FactualError`	World	Mismatch	correspondence
`SourceFabrication`	Source	Absence	existence
`SourceMismatch`	Source	Mismatch	support
`UngroundedInference`	Context & evidence	Absence	entailment
`IntrinsicInconsistency`	Self	Internal contradiction	consistency

4. A reference ontology of language-model errors

Carrier and content. An `LLMOutput` is the textual carrier produced by a `GenerationEvent`; it *expresses* one or more `Assertions`, the units of propositional content (Figure 1). Faithfulness is a property of content, so a failure inheres in the `Assertion`, not in the text. Only *truth-apt* assertions can fail truthmaking, a gate that excludes questions, commands, and formatting.

The categorization. A `TruthmakingFailure` is a *mode* inhering in an unfaithful assertion, reified by a higher-order type whose instances are the five subtypes of Table 1. Each subtype is fixed by two discriminants, the *verification target* (the entity an explainer must consult) and the *failure mode* (a positive falsemaker vs. the *absence* of any truthmaker), which jointly license a distinct verification operation. These subtypes partition `TruthmakingFailure` (disjoint and complete); a single assertion may still bear several failures at once (a *mixed* case), modelled as several tokens, not as overlap.

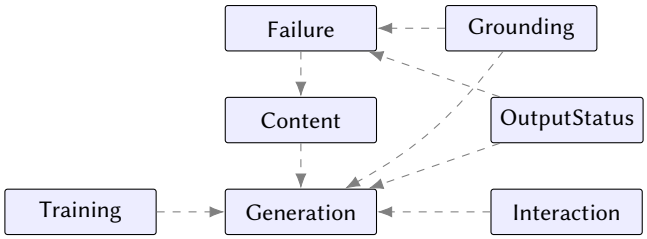

Figure 2: Module structure: dashed arrows are `import` dependencies, each pointing to the module it depends on.

Grounding and etiology. The verification target is realised as an *external dependence* from each extrinsic subtype to a worldly relatum: a reference situation (the falsemaker), a cited source, or the retrieved context; `IntrinsicInconsistency` alone is intrinsic, grounded in the assertion's own contradictory sub-parts. Etiology is structural, not an attribute: a run is retrieval-augmented exactly when a `RetrievedContext` participates in the `GenerationEvent`, so the parametric/RAG distinction (and the root cause it points to) is read off the model, not asserted.

Assessment. Faithfulness is assessed on assertions; an output's status is *derived*: an `LLMOutput` is unfaithful when some truth-apt assertion it expresses bears a failure and faithful when none does, with an epistemically-open third case left between the two (its partition is disjoint but not complete). These derivation rules, the truth-aptness gate (only truth-apt assertions may bear a failure), and the target–mode signatures of Table 1 cannot be stated in the conceptual model; they are realised as SHACL constraints and SPARQL rules over the gUFO export, where the subtype partition becomes OWL disjointness and covering axioms. The model is organised into the seven modules of Figure 2: `Generation` is the shared base; `Content`, `Failure`, and `Grounding` carry the categorization; `OutputStatus` the derived verdict; and `Training` and `Interaction` the provenance from which a failure's etiology is read.

5. Competency questions and demonstration

Following Guizzardi and Guarino [1], we pose the competency questions (CQs) not as lookup queries but as *requests for explanation*: each asks *why* an output failed and is answered by exhibiting the truthmaker its assertion must be verified against. They fall in three tasks: classify, explain, improve.

Classify: (CQ1) which `TruthmakingFailure` type does an assertion instantiate, and against which verification target (world situation, cited source, retrieved context, or itself); (CQ2) does it bear several types at once, a *mixed* case. *Explain:* (CQ3) in virtue of what is it unfaithful (the falsemaker that obtains, or the truthmaker whose *absence* grounds the failure); (CQ4, contrastive) why this type *rather than* another, and so which verification operation it licenses (existence, correspondence, support, or consistency); (CQ5, counterfactual) which truthmaker, had it obtained, would flip the verdict; and (CQ6) for a mixed case, the target to consult for each co-inhering failure. *Improve:* (CQ7) was the run parametric or retrieval-augmented (did a `RetrievedContext` participate), which locates the root cause; (CQ8) which intervention that licenses.

We illustrate this on a small set of worked cases instantiated in gUFO/OWL (one parametric, one retrieval-augmented, and a mixed case), over which the CQs are answered mechanically: SPARQL returns each verification target (the external-dependence truthmaker) with its contrast class, and SHACL enforces the truth-aptness gate and the derived output status. The claim is not automatic detection but that, *once a case is modelled*, the categorization yields faithful, action-licensing answers. Scaling the instantiation to hallucination benchmarks (HaluEval [18] and the RAG-specific RAGTruth [10], with Wikidata as an external world source of truth) is ongoing work (§6).

As a worked instance, the output "According to [a fabricated reference], the Eiffel Tower is in Berlin" yields one truth-apt assertion bearing *two* failure tokens: a `SourceFabrication` (*source*/*absence*: the

reference does not exist) and a `FactualError` (*world/mismatch*: it is in Paris). CQ1 returns both with their targets; CQ4 contrasts the licensed operations (an existence check on the citation vs. a correspondence check against the world); and CQ7 reads the run as parametric, since no `RetrievedContext` participates. The same query machinery answers all eight questions over the instantiated graph.

6. Conclusion and next steps

This categorization is the foundational step of an ongoing doctoral project on ontology-grounded, faithful explanation of AI errors. We have presented a UFO-grounded reference ontology that categorizes LLM errors by the mode of truthmaking failure and the verification target, turning a binary "hallucination" flag into a categorized error that names the truthmaker to verify and, if wrong, contest. The ontology is built and validated in Tonto; its utility for explanation and diagnosis is illustrated on worked cases that answer the competency questions. Immediate next steps are (i) the full gUFO/OWL operationalisation and benchmark-scale instantiation over HaluEval [18] and RAGTruth [10], with Wikidata as a world source of truth; (ii) extending the grounding context beyond retrieval to tool-use and long-context settings; and (iii) a study of whether categorized explanations improve diagnosis over flat labels. The main threat to validity is the cost of fine-grained subtype assignment, which binary benchmarks do not provide; addressing it is part of the instantiation work.

Declaration on Generative AI

During the preparation of this work, the authors used Claude Fable 5 (Anthropic) in order to: check grammar and spelling, and paraphrase, reword, and polish author-written text. The ideas, the technical content, and the structure of the manuscript are the authors' own, and no scientific content was generated by the tool. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] G. Guizzardi, N. Guarino, Explanation, semantics, and ontology, Data & Knowledge Engineering 153 (2024) 102325. doi:10.1016/j.datak.2024.102325.

[2] N. Maleki, B. Padmanabhan, K. Dutta, AI hallucinations: A misnomer worth clarifying, 2024. URL: https://arxiv.org/abs/2401.06796. doi:10.48550/ARXIV.2401.06796.

[3] H. G. Frankfurt, On Bullshit, Princeton University Press, Princeton, NJ, 2005.

[4] M. T. Hicks, J. Humphries, J. Slater, ChatGPT is bullshit, Ethics and Information Technology 26 (2024) 38. doi:10.1007/s10676-024-09775-5.

[5] A. L. Smith, F. Greaves, T. Panch, Hallucination or confabulation? Neuroanatomy as metaphor in large language models, PLOS Digital Health 2 (2023) e0000388. doi:10.1371/journal.pdig.0000388.

[6] L. G. Maddalena, F. Baião, T. P. Sales, G. Guizzardi, On the interaction between explainable artificial intelligence (XAI) and ontologies: A quasi-systematic literature review, in: Proceedings of the 18th Seminar on Ontology Research in Brazil (ONTOBRAS 2025) and 9th Doctoral and Masters Consortium on Ontologies (WTDO 2025), volume 4177 of *CEUR Workshop Proceedings*, CEUR-WS.org, São José dos Campos, Brazil, 2025, pp. 104–118. URL: https://ceur-ws.org/Vol-4177/paper8.pdf.

[7] R. Confalonieri, G. Guizzardi, On the multiple roles of ontologies in explanations for neuro-symbolic AI, Neurosymbolic Artificial Intelligence 1 (2025) 1–15. doi:10.3233/NAI-240754.

[8] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, P. Fung, Survey of hallucination in natural language generation, ACM Computing Surveys 55 (2023). doi:10.1145/3571730.

[9] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, T. Liu, A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, ACM Transactions on Information Systems 43 (2025). doi:10.1145/3703155.

[10] C. Niu, Y. Wu, J. Zhu, S. Xu, K. Shum, R. Zhong, J. Song, T. Zhang, RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 10862–10878. doi:10.18653/v1/2024.acl-long.585.

[11] R. Xu, Z. Qi, Z. Guo, C. Wang, H. Wang, Y. Zhang, W. Xu, Knowledge conflicts for LLMs: A survey, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 8541–8565. doi:10.18653/v1/2024.emnlp-main.486.

[12] G. Guizzardi, Ontological Foundations for Structural Conceptual Models, Ph.D. thesis, University of Twente, Enschede, The Netherlands, 2005.

[13] G. Guizzardi, A. B. Benevides, C. M. Fonseca, D. Porello, J. P. A. Almeida, T. P. Sales, UFO: Unified foundational ontology, Applied Ontology 17 (2022) 167–210. doi:10.3233/AO-210256.

[14] G. Guizzardi, C. M. Fonseca, J. P. A. Almeida, T. P. Sales, A. B. Benevides, D. Porello, Types and taxonomic structures in conceptual modeling: A novel ontological theory and engineering support, Data & Knowledge Engineering 134 (2021) 101891. doi:10.1016/j.datak.2021.101891.

[15] M. L. Coutinho, J. P. A. Almeida, T. P. Sales, G. Guizzardi, A textual syntax and toolset for well-founded ontologies, in: Formal Ontology in Information Systems (FOIS 2024), volume 394 of *Frontiers in Artificial Intelligence and Applications*, IOS Press, 2024, pp. 208–222. doi:10.3233/FAIA241308.

[16] J. P. A. Almeida, G. Guizzardi, T. P. Sales, R. A. Falbo, gUFO: A lightweight implementation of the unified foundational ontology (UFO), http://purl.org/nemo/doc/gufo, 2019. Accessed: 2026-06-15.

[17] R. Haskins, B. Adams, KEA explain: Explanations of hallucinations using graph kernel analysis, in: Proceedings of the 19th International Conference on Neurosymbolic Learning and Reasoning (NeSy), volume 284 of *Proceedings of Machine Learning Research*, PMLR, 2025, pp. 1–18. URL: https://proceedings.mlr.press/v284/haskins25a.html. arXiv:2507.03847.

[18] J. Li, X. Cheng, W. X. Zhao, J.-Y. Nie, J.-R. Wen, HaluEval: A large-scale hallucination evaluation benchmark for large language models, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Singapore, 2023, pp. 6449–6464. doi:10.18653/v1/2023.emnlp-main.397.