TAXONOMY FOR HALLUCINATION LOCALIZATION AND ATTRIBUTION IN LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

While hallucination is a well-documented failure mode of LLMs, localizing and attributing its occurrence within generated text remains ambiguous. Existing span-level hallucination annotations often exhibit substantial disagreement because annotators select different textual spans for the same underlying semantic error. This ambiguity leads to noisy training labels, unreliable evaluation of hallucination detectors, and obscures where hallucinations emerge, knowledge that is essential for building effective mitigation strategies. In this work, we argue that this localization challenge originates due to the absence of a semantic model the defines the minimal unit of hallucination. To address this limitation, we introduce a diagnostic taxonomy for hallucination localization, grounded in Conceptual Dependency Theory (CDT), which categorizes hallucinations according to failures in five semantic components: `ACTOR`, `STATE`, `TEMPORAL`, `OBJECT`, and `ACTION`. A deterministic decision procedure guides annotators in identifying the minimal semantic component responsible for each hallucination, reducing ambiguity in span localization and semantic labelling. The taxonomy was developed through the analysis of 2,000 LLM hallucinations and validated in a held-out replication study in which five independent human annotators applied the frozen taxonomy and decision procedure to 300 previously unseen atomic claims drawn from four benchmark datasets. The taxonomy coupled with the decision tree achieved an average pairwise Cohen's kappa κ of 0.865, demonstrating that it reduces ambiguity in both hallucination span annotation and failure attribution.

1 INTRODUCTION

Large Language Models (LLMs) hold immense potential for high-stakes domains like healthcare, law, and finance. However, hallucinations (outputs that are plausible but factually incorrect) remain a persistent barrier to trustworthy deployment, even in recent model releases (OpenAI, 2025; Anthropic PBC, 2026). Recent benchmarks (Bang et al., 2025; Zhao et al., 2024; Jackson et al., 2025; Robertson et al., 2026) and industry audits (Fletcher & Verckist, 2025; Tiller et al., 2026; Magesh et al., 2025) consistently document these factual errors across sectors.

To mitigate this, precisely localizing hallucinations is essential. For instance, in the statement "Marie Curie won three Nobel Prizes," the error lies specifically in the span "three," not in the entire sentence. However, many existing evaluations (Lin et al., 2022; Thorne et al., 2018; Krishna et al., 2023) rely on response-level or sentence-level labels. While span-level annotation is important for understanding hallucinations, such datasets are scarce because fine-grained labeling is costly, time-consuming, and highly subjective. Previous research notes substantial inter-annotator disagreement at the span level (Vázquez et al., 2025).

Our analysis of existing span-level datasets such as RAGTruth (Niu et al., 2024), HalluEntity (Yeh et al., 2025), PsiloQA (Rykov et al., 2025) along with our own annotation efforts confirm this challenge, revealing frequent boundary disagreements and instances of over-annotation.

Accurate annotation is the foundation for further research in detecting and mitigating hallucinations at the span-level. We argue that for high inter-annotator agreement of annotations at the span-level, a new taxonomy is required. For example, in "The Eiffel Tower was designed by Gustave Eiffel's architecture firm,"; an annotator may label "was designed by" as the erroneous relation, while another may label "Gustave Eiffel's architecture firm" as the erroneous object. A clear taxonomy is

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

therefore needed to distinguish these cases and consistently identify the erroneous span along with the attribution to what the error was.

In this work, we address this gap by developing a taxonomy that can answer:

- "What span(s) have hallucinated?"
- "What kind of semantic failure caused this?"

To answer these questions and achieve a higher inter-annotator agreement in localizing hallucinations, we ground our taxonomy in Conceptual Dependency Theory (CDT), originally proposed by Roger Schank as a theory of meaning representation (Schank, 1972).

This work makes the following primary contributions:

- A cognitively grounded taxonomy, conceptually language-independent, of hallucinations inspired by Conceptual Dependency Theory (CDT).
- A decision-tree annotation protocol achieving high inter-annotator agreement (average pairwise Cohen's kappa κ = 0.865) in classifying 300 hallucinated claims by five human annotators.
- A benchmark of inter-annotator agreement of taxonomies in prior work. (Section 5.2).

2 Related Work

In this section, we present past works in two directions: 1) efforts to categorize hallucinations, and 2) Conceptual Dependency Theory, which is the framework used to develop the taxonomy.

2.1 Hallucination Taxonomies

Recent benchmarks and detection methods (Dai et al., 2025; Goel et al., 2025; Bang et al., 2025) expose increasingly subtle hallucination patterns, but lack structured mechanisms to distinguish them, motivating the need for a localization taxonomy. Existing taxonomies remain largely descriptive rather than diagnostic (Ji et al., 2023). Prior taxonomies categorizes hallucinations into broad categories of intrinsic (contradicting source text) or extrinsic (introducing unverifiable information) (Bang et al., 2025; Goel et al., 2025; Rykov et al., 2025; Rashkin et al., 2023; Tam et al., 2023). Task-centric categorization frameworks classify hallucinations as input-, fact-, or context-conflicting (Zhang et al., 2025), and related benchmarks such as FaithDial (Dziri et al., 2022) emphasize categorizing hallucinations with respect to source knowledge. Although some existing taxonomies identify where hallucinations occur relative to the source text, they do not specify which entities, types, or relations failed (Ji et al., 2023; Maynez et al., 2020).

(Filippova, 2020; Raunak et al., 2021) focused on hallucinations caused by bad training data and split those into different categories. Other model-centric approaches categorize hallucinations by generation behavior, including Oscillatory, Strongly Detached, and Fully Detached outputs (Xu et al., 2023). Span-centric methods classify hallucinations based on length of the error(token-level, sentence-level or passage-level) (Hossain et al., 2025).

Several studies categorize by attribution to source content (Rashkin et al., 2023), including faithful vs. factual distinctions (Zhao et al., 2023) and dimensions, such as fluency, plausibility, confidence (Huang et al., 2025).

Across frameworks centered on the source, task, model, span, trust, and style, a common limitation is the lack of granularity and focus on inter-annotator agreement. These categorization schemes capture important surface-level aspects, but offer limited guidance for localizing and attributing the failure unambiguously.

NLP-centric taxonomies introduce categories, such as Predicate, Entity, Circumstance, Co-reference Errors (Pagnoni et al., 2021), partially aligning with semantic roles of spans. However, these labels are applied at the surface-level without a theory of meaning representation, leading to overlapping categories, inconsistent terminology, and limited inter-annotator agreement. In contrast, our taxonomy is grounded in Conceptual Dependency Theory (CDT) and comes with a decision tree that enables a consistent separation between different categories.

2.2 Conceptual Dependency Theory

Conceptual Dependency Theory (CDT) was introduced by Schank as a framework to represent the meaning of natural-language utterances independently of their surface wording (Schank, 1972; 1975). Its central idea is that any sentence's meaning can be reduced to a small, language-independent set of primitive conceptual acts, such as `ATRANS` (transfer of an abstract relationship, e.g. ownership), `PTRANS` (transfer of physical location), `MTRANS` (transfer of mental information, e.g. telling or reading), `PROPEL` (application of physical force), rather than represented by surface-level verbs.

Each unit of meaning, what Schank called a *conceptualization*, is built around an actor performing an ACT. This ACT can take the form of an object (what the act is performed on), a direction (where something moves to or from), a recipient (who receives the effect of the act), or an instrument (what is used to perform it). These roles attach to the ACT the way case grammar (Fillmore, 1968) attaches roles to a verb. However, CDT uses semantic primitives instead of surface grammar.

This primitive-based structure is what sets CDT apart from other semantic representations used in NLP. SRL (Gildea & Jurafsky, 2002) ties roles to the exact verb used in a sentence. FrameNet (Baker et al., 1998) requires choosing among thousands of fine-grained categories. AMR (Banarescu et al., 2013) represents meaning as graphs shaped by English syntax. Each of these properties can widen disagreement while localizing a hallucination.

We derive our taxonomy from CDT, giving annotators a much smaller, theory-grounded space to localize an error in. Our taxonomy is not a direct application of CDT, but a derivation from it for the specific problem of hallucination localization, rather than the general-purpose semantic parsing CDT was originally designed for.

3 Methodology

3.1 Taxonomy Development

To develop the proposed hallucination taxonomy, we collected 2,000 hallucinated model outputs from multiple NLP tasks, including summarization, dialogue, and question answering. These outputs were drawn from HaluEval-Dialogue, HaluEval-QA, HaluEval-Summarization (Krishna et al., 2023), XSum (Maynez et al., 2020), and PsiloQA (Rykov et al., 2025). During open coding and analysis, the model output was the unit of analysis. However, for the development and test phases, the unit of annotation was changed to atomic claims, since a single model output may contain multiple atomic claims.

3.1.1 Training, Development and Test Sets

The training set consisted of 1,550 model outputs used for open coding and taxonomy development. A development set of 150 hallucinated atomic claims was used to assess inter-annotator agreement, while a test set of 300 hallucinated atomic claims was programmatically sampled from HaluEval-Dialogue, HaluEval-QA, HaluEval-Summarization, and PsiloQA for the final evaluation. The development set was annotated by one human annotator and two LLMs (GPT-4o-2024-08-06 and GPT-4.1-mini) to facilitate rapid iterative refinement. The test set was independently annotated by five human annotators.

3.1.2 Open Coding

Annotators were asked to describe what went wrong in each hallucinated output (in the training set). This open-coding step was designed to obtain descriptive explanations of failure without imposing predefined categories. Following open coding, annotators independently grouped their explanations into themes (see Table 2 in the Appendix for some theme examples). For each identified theme, the annotators provided a concise description along with representative examples. This process was conducted iteratively, with themes refined between rounds based on emerging patterns and annotator agreement. The resulting set of error themes served as the basis for building the taxonomy.

The taxonomy was developed by selecting and refining specific CDT constructs. Each iteration of the taxonomy was evaluated for inter-annotator agreement on the development set. Annotations

were performed at the level of atomic claims extracted once from model outputs (using GPT-4o-2024-08-06 and human validation), rather than at the sentence or response level. Pairwise Cohen's kappa scores were then computed. This iterative process was repeated across multiple taxonomy versions to arrive at the final taxonomy reported in the next section.

3.2 Taxonomy Overview

We propose a hallucination taxonomy grounded in Conceptual Dependency Theory (CDT) and operationalized via a deterministic decision tree. The taxonomy assigns each hallucinated claim to exactly one type of semantic failure by isolating which component of the claim's meaning breaks.

3.3 Conceptual Dependency Representation

We use CDT as a meaning-representation scaffold to organize hallucination causes. The CDT primitives used in this work are:

- PP (Picture Producer): physical objects or entities that can be represented mentally (e.g. Drew Goddard, The Walking Dead).
- PA (Picture Aider): attributes or modifiers of a PP. We use PA subtypes including:
  - `TIME`: date, year, duration, ordering (e.g. 1994, two years, before/after)
  - `LOCATION`: geographic context (e.g. Paris, London)
  - `PROPERTY`: quantity, units, material, size, etc.
- `ACTION` (Primitive Action): the underlying event/relation encoded by a verb phrase (e.g. write, direct, buy).
- `STATE`: a property or condition assigned to a PP, including identity/classification states (e.g. X is a Y).
- Semantic roles: the `ACTOR` (which is the entity performing the action), the `OBJECT` (which is entity acted upon) and the `RECIPIENT` [1] (which is the entity receiving an object because of an action).

These primitives allow claims to be decomposed into role-structured representations, making semantic failure explicit.

3.4 Claim Segmentation and Role Mapping

We distinguish three levels of analysis in our framework:

- Output: the full model-generated response.
- Claim: an atomic, fact-like statement extracted from the output.
- Span: the minimal text segment within a claim corresponding to a semantic component (e.g. `ACTOR`, `OBJECT`, `ACTION`) that may be hallucinated.

All hallucination classification is performed at the claim level, while localization operates at the span level within each claim, with spans later mapped back to the output. Each hallucinated atomic claim is segmented, e.g. "J.K.Rowling wrote Harry Potter" is segmented into `ACTOR`: J.K.Rowling, `ACTION`: wrote, `OBJECT`: Harry Potter.

3.5 Hallucination Taxonomy

We define five mutually exclusive hallucination types. Each label corresponds to a failure of a specific semantic component.

[1] No RECIPIENT modeling is required: each atomic claim has a single object/recipient.

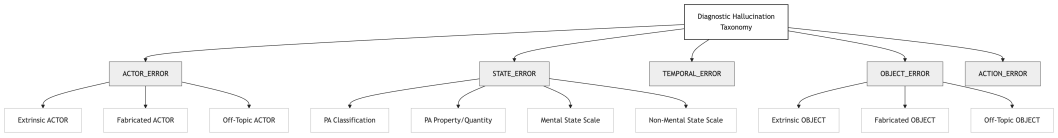

Figure 1: Conceptual Dependency-based hallucination taxonomy.

3.5.1 ACTOR ERROR

Definition: The claim identifies the wrong entity as the primary actor or anchor of the claim or introduces an actor that is not on topic for the user's intent.

This applies when the ACTOR fails relevance or existence validation. For example, in "*Taylor Swift* discovered penicillin.", Taylor Swift is a musician with no relevance to the discovery of penicillin; the actual discoverer was Alexander Fleming in 1928.

3.5.2 STATE ERROR

Definition: The claim makes an incorrect identity/classification assertion of the actor in the form A is equivalent to B (e.g. "X is a Y", "X is about Y").

This applies when the predicate is a state (identity/genre/category) rather than an ACTION, and the state is wrong. For example, in "Guernsey Literary is a *thriller* about spies.", Guernsey Literary is a historical fiction/romance novel and not a thriller.

3.5.3 TEMPORAL ERROR

Definition: The claim asserts an incorrect time, date, year, duration, or event ordering.

This applies when a temporal modifier (TIME) is present and is wrong. For example, in "Otto Stern won a Nobel Prize in *1955*", Otto Stern did win a nobel prize in atomic physics but in 1943.

3.5.4 OBJECT ERROR

Definition: The actor in the claim is valid, but the target entity is incorrect such that the ACTOR–OBJECT pair has no grounded connection (under any plausible relation).

This applies when there is no real-world link between the actor and the object, even before considering the specific action. For example, in "Drew Goddard wrote for *The Walking Dead*", Drew Goddard is indeed a screenwriter, but he has no documented involvement with "The Walking Dead", which was developed by "Frank Darabont".

3.5.5 ACTION ERROR

Definition: The actor and object in the claim are correctly grounded and related, but the specific action (relation) asserted between them is incorrect or never occurred.

This applies when the ACTOR–OBJECT pair is valid, but the ACTION fails factual verification. For example, in "Neil Armstrong *commanded* Apollo 11.", Armstrong was part of the mission, but not the commander.

3.6 TAXONOMY DECISION PROCEDURE

We operationalize the proposed taxonomy using a structured decision tree applied to the model outputs that are being evaluated. Figure 1 illustrates the taxonomy dependencies, while Figure 2 outlines the step-by-step localization and attribution procedure.

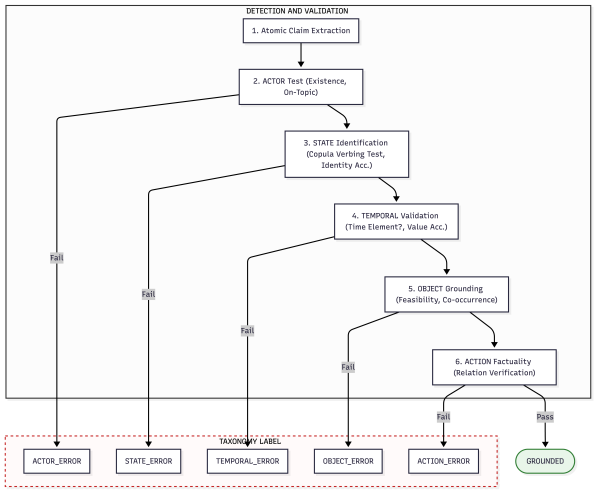

Figure 2: Decision Tree Procedure.

3.6.1 Output Decomposition

First, the model outputs are segmented into sentences. Because a single sentence can contain multiple assertions, we decompose each sentence into individual atomic claims. For example, the sentence "Emily Giffin wrote a science fiction novel called Interstellar" is decomposed into: (1) Emily Giffin wrote a science fiction novel, (2) Interstellar is a science fiction novel, and (3) Emily Giffin wrote Interstellar.

3.6.2 Claim Segmentation

Each atomic claim is further decomposed into three components: `ACTOR`, `ACTION`, and `OBJECT`. Because a single claim can often be segmented in more than one valid way, we need a consistent rule for choosing among them. For example, the claim "The researcher analyzed the impact of climate change on agriculture" can be segmented as (i) `ACTION`: analyzed the impact of climate change; `OBJECT`: on agriculture, or (ii) `ACTION`: analyzed; `OBJECT`: the impact of climate change on agriculture. To achieve consistency, we employ a left-to-right rule, which assigns the smallest eligible span to each of `ACTOR/ACTION/OBJECT` from left to right, with the residual text assigned to the `OBJECT`. Under this rule, annotators select the second segmentation.

3.6.3 Actor Test

After segmentation, the `ACTOR` is evaluated in isolation, ignoring the `ACTION`, `OBJECT`, and any other content in the claim. Two conditions must be satisfied for this test to pass:

- **Existence**: Does the actor correspond to a real entity or concept?
- **Relevance**: Is the actor relevant to the prompt or instruction?

If either condition fails, the test fails. The claim is labeled `ACTOR_ERROR`, and all subsequent tests are skipped. For example, for the question "Who invented quantum computing?", the claim "*Taylor Swift* invented quantum computing" is an `ACTOR_ERROR`, as Taylor Swift is a does exist but irrelevant to the question being asked.

3.6.4 State Test

Having ruled out an error in the `ACTOR`, the next test examines how the actor is described or classified. This test applies only to claims that can be reduced to an identity or classification of the form A is semantically equivalent to B (e.g. "X is Y", "X is about Y").

Not all claims describe the state, some describe an action (e.g. "John is jogging in the park' does not describe an actor). To distinguish state from action, we apply a verbing (copula) test: if the predicate can be rephrased as an event involving a primitive ACTION (e.g. written by, directed by), the claim is treated as action-based, and this test is skipped.

If the claim expresses a copular relation and the classification or definition is incorrect, then the claim is labelled STATE_ERROR. For example, the claim "Guernsey Literary is a *thriller about spies*" is STATE_ERROR due to an incorrect genre classification.

3.6.5 TEMPORAL TEST

Next, we check if the claim contains a temporal element (e.g. date, year, duration, or ordering). The temporal modifier is evaluated independently of other claim components.

If the temporal information is factually incorrect, the claim is labeled TEMPORAL_ERROR and evaluation is terminated. For example, the claim "The Berlin Wall fell in *1991*" contains TEMPORAL_ERROR due to the wrong year.

3.6.6 OBJECT TEST

This test sets aside the ACTION entirely and evaluates the connection between the ACTOR and the OBJECT. It applies two sub-tests:

- **Connection feasibility test:** Is there any real-world action that plausibly links them? For example, "Albert Einstein *wrote* Hamlet" fails because their different eras make any interaction impossible.
- **Co-occurrence test:** Do reliable sources suggest a meaningful link between them?

For example, consider the pair ACTOR: Christopher Nolan and OBJECT: *The Godfather*. There is no meaningful connection between them. Conceptual connections such as the director and the movie are insufficient. The link must be significant; Nolan simply liking the movie is not significant. Hypothetical or inferred connections are also prohibited. If no grounded connection exists, the claim is labeled OBJECT_ERROR and the evaluation is terminated.

3.6.7 ACTION TEST

At this stage, the Object test has established that a valid actor–object link exists. Now, we check if the specific action linking them is correct. For example, in "Mel Gibson *acted* in The Passion of the Christ (2004)", the right link is *directed*. Although, Mel Gibson is an A-List actor; he stepped completely behind the camera to direct this movie.

4 EXPERIMENTS

To arrive at the final taxonomy just presented, we experimented with several taxonomies that we developed after open-coding. The experiments were conducted with the development set. After each round, we computed average pairwise Cohen's kappa κ to measure inter-annotator agreement and analyzed disagreements to identify unstable category boundaries. Based on these analyzes, we refined the taxonomy and decision procedure and converged on a stable formulation. Once the stable formulation was achieved it was evaluated with a much larger held-out test set with five human annotators. To compare with prior work, we derived taxonomies from seven hallucination categorization schemes. These taxonomies were then evaluated on the development set as well.

5 RESULTS AND ANALYSIS

The final taxonomy and decision procedure achieved an average pairwise Cohen's kappa of $\kappa = 0.865$ on the held-out test set with five independent human annotators. The training, development and test sets used for these experiments were detailed in Section 3.1.1.

Table 3 in the Appendix reports average pairwise Cohen's kappa κ scores of the various iterations of the taxonomy. Next, we trace this evolution, showing how specific disagreement patterns motivated successive changes to the annotation procedure.

We found that model output decomposition into atomic claims was essential as annotating such outputs without decomposition consistently reduced agreement. The deterministic left-to-right segmentation was adopted to prevent annotators from choosing different semantic boundaries within the same sentence.

Increasing label granularity reduced reliability. Fine-grained but unordered CDT-derived categories added ambiguity rather than clarity. This motivated collapsing labels into a compact, role-based five node taxonomy. For example, earlier CDT-inspired labels, such as `PA PROPERTY`, `PA LOCATION` were consolidated into `STATE` and `TEMPORAL` to improve agreement. After the taxonomy was simplified and structured around the core semantic roles, the inter-annotator agreement improved substantially.

5.1 Refinement Analysis

Even with a smaller label set, annotators frequently selected multiple plausible error types. To resolve these disagreements, several changes were made to the annotation procedure that resulted in the final fixed decision protocol, as described in Section 3.6.

Separating Actor from Object Errors. To resolve disagreements between `ACTOR_ERROR` and `OBJECT_ERROR`, we made the Actor Test precede the Object Test, evaluating the `ACTOR` independently of its relations. For example, given the prompt "What novels did George Orwell write?", the claim "J.K. Rowling wrote Nineteen Eighty-Four" could be attributed either to the irrelevant `ACTOR` (J.K. Rowling) or to the incorrect association between the `ACTOR` and `OBJECT`. By evaluating the `ACTOR` first, the claim is consistently labeled `ACTOR_ERROR` because *J.K. Rowling* is not relevant to the prompt.

Separating State from Action Errors. Annotators had difficulty distinguishing `STATE` from `ACTION`. For example, the claim "Guernsey Literary is a thriller about spies", was labeled either `STATE_ERROR` or `ACTION_ERROR`, depending on if *is a thriller* is interpreted as a classification or an action. To resolve this ambiguity, the copular (verbing) test was added.

Separating State from Object Errors. The copular (verbing) test resolved `STATE-ACTION` disagreements but revealed ambiguity between `STATE` and `OBJECT`. For example, the claim "Max Volmer is a Nobel Prize winner", was labeled `STATE_ERROR` if *Nobel Prize winner* is treated as a classification, or labeled `OBJECT_ERROR` if *Nobel Prize* is treated as the object. To resolve this ambiguity, the State Test was made to precede the Object Test.

Separating Temporal from Action errors. To resolve disagreements between `TEMPORAL` and `ACTION` errors, the Temporal Test was moved before the Object and Action Tests. For example, in "Scott Stevens played for the Senators from 1982 to 1992, then was traded to the New Jersey Devils", the error was attributed either to the `ACTION` or to the stated time period. The order ensured that `TEMPORAL_ERROR` was selected consistently.

Separating Object from Action errors To resolve disagreements between `OBJECT_ERROR` and `ACTION_ERROR`, we separated their evaluation into two tests: Object and Action test. For example, the claim "J.K. Rowling wrote The Lord of the Rings" was labeled either `OBJECT_ERROR` or `ACTION_ERROR`. The error could be attributed either to *The Lord of the Rings* as the wrong `OBJECT` associated with *Rowling*, or to *wrote* as the wrong `ACTION` between the two entities. To resolve this ambiguity, the Object Test first checks whether any grounded connection exists between the `ACTOR` and the `OBJECT`, ignoring the asserted `ACTION`. Only if such a connection exists does the Action Test evaluate whether the specific `ACTION` is correct.

Together, these refinements led to the final five-class taxonomy and decision tree presented in Section 3.6, achieving both high agreement and clear interpretability.

Framework	Category Type	# Labels	κ
Intrinsic / Extrinsic	Source-grounded	2	0.566
Faithfulness / Factuality	Source + world knowledge	4	0.307
Input / Context / Fact conflict	Source / context / world	3	0.240
Token–Passage granularity	Scope-based	4	0.402
Drift (Oscillatory / Detached)	Discourse-level	3	-0.054
Semantic error types	Entity / predicate / etc.	5	0.420
Fluency / plausibility	Style / epistemic	4	0.110
Ours (Role-based)	Semantic + ordered	5	**0.796**

Table 1: Average pairwise Cohen's kappa κ across hallucination taxonomies

5.2 Comparison with existing Taxonomies

Next, we examine whether the improved agreement is unique to our framework by comparing it against seven existing hallucination taxonomies. The comparison includes source-grounded, scope-based, discourse-level, semantic, and style-based approaches to hallucination categorization. We developed taxonomies by studying hallucination categorization schemes in prior work. These taxonomies are detailed in Table 4 in the Appendix.

Table 1 compares average pairwise Cohen's kappa across taxonomies on the development set (Section 3.1.1). Our framework achieved the highest agreement ($\kappa = 0.796$; the $\kappa = 0.865$ reported earlier was on the held-out test set), followed by the Intrinsic/Extrinsic taxonomy ($\kappa = 0.566$) and semantic error types ($\kappa = 0.420$). The proposed taxonomy does get the highest agreement, suggesting that grounding labels in semantic roles and using an ordered decision procedure resolves ambiguities present in prior schemes.

Beyond annotation agreement, we also analyzed the distribution of error types assigned by the taxonomy. `OBJECT` and `STATE` errors accounted for the majority of hallucinations, with similar distributions between annotators. A detailed analysis of error-type distributions, cross-annotator stability, and temporal errors is provided in Appendix A.3.

6 Limitations and Future Work

This work has several limitations. First, it focuses on the localization and attribution of span-level hallucinations, but does not examine the underlying model mechanisms that cause them. The second limitation is taxonomy granularity. In future work, we plan to conduct agreement studies on a more granular taxonomy which has been proposed in this work (in Appendix A.4 and Figure 1), but not evaluated. The taxonomy also has the potential to be converted into an algorithm for automatically detecting and labeling hallucinations, which we plan to explore in future work. Comparisons with taxonomies from previous studies should be interpreted with caution, as those taxonomies were not designed specifically for hallucination localization and attribution. Since the proposed taxonomy is based on CDT, it has the potential to be language-independent, but this has not yet been evaluated. Finally, the experiments were conducted after model outputs were segmented into atomic claims, but we did not conduct an inter-annotator agreement study for the atomic claim segmentation itself.

7 Conclusion

Evaluating hallucinations must move beyond binary detection to localization and diagnosis, categorizing failure reasons and guiding targeted mitigation. We introduce a CDT-based taxonomy (`ACTOR`, `STATE`, `TEMPORAL`, `OBJECT`, `ACTION`) with a decision tree to assign precise failure labels. Validated on 450 hallucination claims (150 development and 300 test claims) drawn from 2,000 outputs across dialogue, QA, and summarization, the taxonomy achieves an average pairwise kappa κ of 0.865 and outperforms existing taxonomies in inter-annotator agreement.

By converting hallucinations into interpretable semantic components, this framework provides a foundation for annotating hallucinations, span-level benchmarks, and model debugging in high-stakes LLM applications.

REFERENCES

Anthropic PBC. Claude Opus 4.6 System Card. Technical report, Anthropic PBC, 2026. URL https://www-cdn.anthropic.com/0dd865075ad3132672ee0ab40b05a53f14cf5288.pdf. Technical report.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pp. 86–90, Montreal, Quebec, Canada, 1998. Association for Computational Linguistics. doi: 10.3115/980845.980860.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pp. 178–186, Sofia, Bulgaria, 2013. Association for Computational Linguistics. doi: 10.3115/v1/W13-2322.

Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. Hallulens: Llm hallucination benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24128–24156, 2025.

Ziyun Dai, Xiaoqiang Li, Shaohua Zhang, Yuanchen Wu, and Jide Li. See different, think better: Visual variations mitigating hallucinations in lvlms. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 3310–3319, 2025.

Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar R Zaïane, Mo Yu, Edoardo M Ponti, and Siva Reddy. Faithdial: A faithful benchmark for information-seeking dialogue. *Transactions of the Association for Computational Linguistics*, 10:1473–1490, 2022.

Katja Filippova. Controlled hallucinations: Learning to generate faithfully from noisy data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 864–870, 2020.

Charles J. Fillmore. The case for case. In Emmon Bach and Robert T. Harms (eds.), *Universals in Linguistic Theory*, pp. 1–88. Holt, Rinehart and Winston, New York, 1968.

James Fletcher and Dorien Verckist. News integrity in AI assistants: An international PSM study. Technical report, BBC and European Broadcasting Union, October 2025. URL https://www.ebu.ch/Report/MIS-BBC/NI_AI_2025.pdf.

Daniel Gildea and Daniel Jurafsky. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288, 2002. doi: 10.1162/089120102760275983.

Aman Goel, Daniel Schwartz, and Yanjun Qi. Zero-knowledge llm hallucination detection and mitigation through fine-grained cross-model consistency. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 1982–1999, 2025.

Adiba Ibnat Hossain, Sagnik Ray Choudhury, and Hamed Alhoori. SciHallu: A multi-granularity hallucination detection dataset for scientific writing. In Kentaro Inui, Sakriani Sakti, Haofen Wang, Derek F. Wong, Pushpak Bhattacharyya, Biplab Banerjee, Asif Ekbal, Tanmoy Chakraborty, and Dhirendra Pratap Singh (eds.), *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, December 2025.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, January 2025. ISSN 1558-2868. doi: 10.1145/3703155. URL http://dx.doi.org/10.1145/3703155.

Declan Jackson, William Keating, George Cameron, and Micah Hill-Smith. AA-Omniscience: Evaluating cross-domain knowledge reliability in large language models. *arXiv preprint arXiv:2511.13029*, 2025. URL https://arxiv.org/abs/2511.13029.

Zhe Ji, Nayeon Lee, Rebecca Frieske, Tianjun Yu, Dan Su, Yi Xu, et al. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

Kalpesh Krishna, Abhishek Roy, and Mohit Iyyer. Halueval: A large-scale hallucination evaluation benchmark. *arXiv preprint arXiv:2305.11747*, 2023.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of ACL*, pp. 3214–3229, 2022.

Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning, and Daniel E. Ho. Hallucination-free? assessing the reliability of leading AI legal research tools. *Journal of Empirical Legal Studies*, 22(2):216–242, 2025.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan Thomas Mcdonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1906–1919, Online, 2020.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Randy Zhong, Juntong Song, and Tong 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)*, pp. 10862–10878, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.585. URL https://aclanthology.org/2024.acl-long.585/.

OpenAI. GPT-5 System Card. Technical report, OpenAI, August 2025. URL https://cdn.openai.com/gpt-5-system-card.pdf. Technical report.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. Understanding factuality in abstractive summarization with frank: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4812–4829, 2021.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4):777–840, 2023.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. The curious case of hallucinations in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1172–1183, 2021.

Alex Robertson, Huizhi Liang, Mahbub Gani, Rohit Kumar, and Srijith Rajamohan. KGHaluBench: A knowledge graph-based hallucination benchmark for evaluating the breadth and depth of LLM knowledge. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 3975–3989, 2026.

Elisei Rykov, Kseniia Petrushina, Maksim Savkin, Valerii Olisov, Artem Vazhentsev, Kseniia Titova, Alexander Panchenko, Vasily Konovalov, and Julia Belikova. When models lie, we learn: Multilingual span-level hallucination detection with psiloqa. *arXiv preprint arXiv:2510.04849*, 2025.

Roger C. Schank. Conceptual dependency: A theory of natural language understanding. *Cognitive Psychology*, 3(4):552–631, 1972.

Roger C. Schank. *Conceptual Information Processing*. North-Holland, Amsterdam, 1975.

Derek Tam, Anisha Mascarenhas, Shiyue Zhang, Sarah Kwan, Mohit Bansal, and Colin Raffel. Evaluating the factual consistency of large language models through news summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 5220–5255, 2023.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: A large-scale dataset for fact extraction and verification. In *Proceedings of NAACL*, pp. 809–819, 2018.

Nicholas B. Tiller, Alessandro R. Marcon, Marco Zenone, Kristin E. Kidd, Asker E. Jeukendrup, Zubin Master, and Timothy Caulfield. Generative artificial intelligence-driven chatbots and medical misinformation: An accuracy, referencing and readability audit. *BMJ Open*, 16(4):e112695, 2026. doi: 10.1136/bmjopen-2025-112695. URL https://bmjopen.bmj.com/content/16/4/e112695.

Raúl Vázquez, Timothee Mickus, Elaine Zosa, Teemu Vahtola, Jörg Tiedemann, Aman Sinha, Vincent Segonne, Fernando Sánchez-Vega, Alessandro Raganato, Jindřich Libovickỳ, et al. Semeval-2025 task 3: Mu-shroom, the multilingual shared-task on hallucinations and related observable overgeneration mistakes. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pp. 2472–2497, 2025.

Weijia Xu, Sweta Agrawal, Eleftheria Briakou, Marianna J Martindale, and Marine Carpuat. Understanding and detecting hallucinations in neural machine translation via model introspection. *Transactions of the Association for Computational Linguistics*, 11:546–564, 2023.

Min-Hsuan Yeh, Max Kamachee, Seongheon Park, and Yixuan Li. HalluEntity: Benchmarking and understanding entity-level hallucination detection. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=494k7e9R5D. J2C certification.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. Siren's song in the ai ocean: A survey on hallucination in large language models. *Computational Linguistics*, 51(4):1373–1418, 2025.

Lingjun Zhao, Khanh Nguyen, and Hal Daumé III. Hallucination detection for grounded instruction generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 4044–4053, 2023.

Wenting Zhao, Tanya Goyal, Yu Ying Chiu, Liwei Jiang, Benjamin Newman, Abhilasha Ravichander, Khyathi Chandu, Ronan Le Bras, Claire Cardie, Yuntian Deng, et al. WildHallucinations: Evaluating long-form factuality in LLMs with real-world entity queries. *arXiv preprint arXiv:2407.17468*, 2024.

A Appendix

A.1 AI Disclosure

We used AI assistance in the following ways during the preparation of this work: as an LLM-as-a-judge for evaluation; to proofread paper drafts and aid or polish writing; for retrieval and discovery, including finding related work; for research ideation and brainstorming; to create or modify scientific figures and images; to create or edit software code; to source and search for information; to edit the paper for readability; to identify relevant literature and format references; All AI-generated content was reviewed and verified by the authors, who take full responsibility for the accuracy and integrity of the final work.

A.2 Open-Coding

Theme	Description
Incorrect Role Assignment	Assigning the wrong role or title to a person due to confusion between similar roles, names, or responsibilities.
Misleading or Incorrect Dates	Providing incorrect dates or timelines for historical events, releases, or other factual information.
Inaccurate Geographical Information	Incorrectly identifying or describing geographical locations, regions, or countries.
Incorrect Attribution or Credits	Wrongly attributing authorship, contributions, achievements, or works to an individual or organization.
Fictitious or Non-existent Entities	Introducing fictional or non-existent people, entities, titles, or events as factual.

Table 2: Representative hallucination themes used for error categorization.

κ Score	Taxonomy Variant
0.796	TEMPORAL_ERROR, STATE_ERROR, OBJECT_ERROR, ACTOR_ERROR, ACTION_ERROR
0.659	ATRANS_ERROR, NON_ATRANS_ERROR
0.618	CLAIM_WITH_ACTION, CLAIM_WITHOUT_ACTION
0.623	PA_LOCATION, FABRICATED_PP, TIME, NARRATIVE_ERROR, PA_PROPERTY, PA_STATE_SCALE
0.596	PP_ERROR, TIME, ACTION_ERROR, STATE_ERROR
0.518	ACTOR_ERROR, ACTION_ERROR, OBJECT_ERROR, PA_ERROR
0.376	NARRATIVE_ERROR, ACTION_LINK, ACTION_BRIDGE, GIBBERISH

Table 3: Inter-annotator agreement (κ) across taxonomy variants

A.3 Error Type Distribution Analysis

In this section we examine how hallucinations are distributed across the five error types.

A.3.1 The Mechanics of Plausible Falsehoods

The dominance of OBJECT_ERROR (42–46%) and STATE_ERROR (27–30%) across both annotator models reveals why hallucinations remain so deceptive in the models.

Our findings suggest that the underlying conceptual frame of the model's output is often structurally sound, the actor exists and the relation is linguistically valid—but the model fails during the "target selection" phase (OBJECT_ERROR) or in assigning an incorrect state or attribute to a valid entity (STATE_ERROR).

This suggests that hallucinations are less about total "incoherent text generation" but more about subtle failures in relational grounding.

A.3.2 Cross-Model Stability and Operationalization

The high diagonal counts in our confusion analysis (e.g. 39/42 for OBJECT_ERROR) indicate that these categories are not merely subjective interpretations but are highly operationalizable.

The strong agreement between independent models (GPT-4.1-mini and GPT-4o-2024-08-06) suggests that the taxonomy captures salient, consistently interpretable failure modes even by LLM-as-a-judge.

A.3.3 Temporal and Domain-Specific Risk

The identification of TEMPORAL_ERROR (10–12%) as a distinct and stable class highlights a specific failure mode often tied to outdated training data or the "knowledge cutoff" phenomenon.

In high-stakes domains like financial services or healthcare, distinguishing these from reasoning errors is vital; a temporal error might be mitigated through real-time Retrieval-Augmented Generation (RAG), while an ACTOR_ERROR (fabricating an entity) might indicate a more fundamental failure in the model's pre-training knowledge representation and may suggest the need for targeted mitigation.

A.4 Subtype Definitions

Although the core taxonomy distinguishes hallucinations at the level of ACTOR, ACTION, OBJECT, STATE, and TEMPORAL failures, our analysis revealed systematic structure within certain categories that warrants finer-grained treatment. ACTOR_ERROR and STATE_ERROR.

A.4.1 ACTOR ERROR SUBTYPES

An ACTOR_ERROR occurs when the primary entity anchoring the claim is incorrect relative to the context or user intent. We identify three subtypes based on the nature of the failure:

- EXTRINSIC_ACTOR_ERROR: The actor exists in the real world or external knowledge bases but is absent from the provided context or evidence. The model introduces an entity by assuming background knowledge not supported by the prompt. We treat assumptive hallucinations—where a vague prompt leads the model to confidently introduce an unstated but plausible entity—as a special case of this subtype.
- FABRICATED_ACTOR_ERROR: The actor does not exist in any known knowledge base and is entirely invented by the model. These errors reflect true fabrication and represent a severe breakdown in entity grounding.
- OFF_TOPIC_ACTOR_ERROR: The actor is real but misaligned with the user's informational intent. In summarization, this occurs when a peripheral entity is elevated to primary focus; in question answering, when the response discusses an entity that does not directly address the question. The failure lies in salience and relevance rather than existence.

Distinguishing these subtypes separates failures of existence, grounding, and relevance, which have different implications for mitigation but are otherwise conflated under a single actor-level label.

A.4.2 STATE ERROR SUBTYPES

A STATE_ERROR corresponds to an incorrect STATE assignment in CDT, arising from linking-verb constructions that describe what an actor is, has, or is like. We identify four subtypes:

- GENRE / CLASSIFICATION_ERROR: The claim assigns an incorrect type, category, genre, or role to the actor (e.g. labeling a memoir as science fiction or a person as a director rather than an author).
- QUANTITY_ERROR: The claim assigns incorrect scalar or attribute values, including number, units, size, material, or measurement (e.g. "23 games vs. 22 games," "30 pounds vs. 20 pounds").
- MENTAL_STATE_SCALE_ERROR: The claim incorrectly assigns or inverts a mental or emotional state, including exaggeration or polarity reversal (e.g. "sad" vs. "happy", "fearful" vs. "brave").
- NON_MENTAL_STATE_SCALE_ERROR: The claim incorrectly assigns a non-mental state, such as physical condition, socioeconomic status, or structural integrity (e.g. "functional" vs. "damaged", "wealthy" vs. "poor").

Category	Definition
INTRINSIC	Output contradicts the provided source/context (i.e., conflicts with input evidence).
EXTRINSIC	Output introduces information not supported by the provided source/context (i.e., unverifiable from the input), even if it may be true in the world.
FAITHFUL AND FACTUAL	Consistent with the provided source/context and correct with respect to world knowledge.
FAITHFUL BUT NONFACTUAL	Supported by the source/context, but the source itself is wrong or the claim is incorrect in the world.
UNFAITHFUL BUT FACTUAL	Not supported by the source/context, but happens to be correct in the world.
UNFAITHFUL AND NONFACTUAL	Not supported by the source/context and also incorrect in the world.
INPUT CONFLICT	Directly contradicts explicit information in the input/prompt/source.
CONTEXT CONFLICT	Contradicts implicit constraints of the context (e.g. earlier turns, established entities), without necessarily contradicting a single explicit input fact.
FACT CONFLICT	Incorrect with respect to external/world knowledge, even if not contradicted by the input.
TOKEN LEVEL SUBSTITUTION	Error localized to one or a few tokens (e.g. wrong name, number, date) while structure remains correct.
PHRASE LEVEL DISTORTION	Error affects a phrase or clause (e.g. incorrect relation or modifier) without fully derailing the response.
SENTENCE LEVEL HALLUCINATION	A full sentence-level claim is unsupported or false, while others may remain correct.
PASSAGE LEVEL DRIFT	Multiple sentences or large portions become ungrounded due to topic drift or narrative invention.
OSCILLATORY	Output alternates between grounded and ungrounded content or repeatedly revises assertions.
STRONGLY DETACHED	Output partially drifts from the prompt, introducing substantial irrelevant or invented content.
FULLY DETACHED	Output becomes largely unrelated to the prompt with minimal meaningful connection.
ENTITY ERROR	Incorrect named entity or referent selection, including wrong coreference resolution.
PREDICATE ERROR	Incorrect relation asserted between otherwise correct entities.
ATTRIBUTE ERROR	Incorrect attribute (e.g. number, size, role) attached to an entity or event.
CIRCUMSTANCE ERROR	Incorrect time, location, manner, cause, or condition.
COREFERENCE ERROR	Incorrect reference linking (e.g. pronoun refers to wrong entity).
HIGH FLUENCY LOW GROUNDEDNESS	Fluent and coherent but unsupported or incorrect.
HIGH PLAUSIBILITY BUT FALSE	Plausible given priors but factually incorrect.
HIGH CONFIDENCE ASSERTION ERROR	Strongly asserted claim that is incorrect or unsupported.
LOW CONFIDENCE BUT FALSE	Hedged or uncertain language but still incorrect.

Table 4: Summary of Prior Hallucination Taxonomy Categories