A Unified Definition of Hallucination, Or: It's the World Model, Stupid

Anonymous Author(s)
Affiliation
Address
email

Abstract

Despite numerous attempts to solve the issue of hallucination since the inception of neural language models, it remains a problem in even frontier large language models today. Why is this the case? We walk through definitions of hallucination used in the literature from a historical perspective up to the current day, and fold them into a single definition of hallucination, wherein different prior definitions focus on different aspects of our definition. At its core, we argue that hallucination is simply inaccurate (internal) world modeling, in a form where it is observable to the user (e.g., stating a fact which contradicts a knowledge base, or producing a summary which contradicts a known source). By varying the reference world model as well as the knowledge conflict policy (e.g., knowledge base vs. in-context), we arrive at the different existing definitions of hallucination present in the literature.

We argue that this unified view is useful because it forces evaluations to make clear their assumed "world" or source of truth, clarifies what should and should not be called hallucination (as opposed to planning or reward/incentive-related errors), and provides a common language to compare benchmarks and mitigation techniques. Building on this definition, we outline plans for a family of benchmarks in which hallucinations are defined as mismatches with synthetic but fully specified world models in different environments, and sketch out how these benchmarks can use such settings to stress-test and improve the world modeling components of language models.

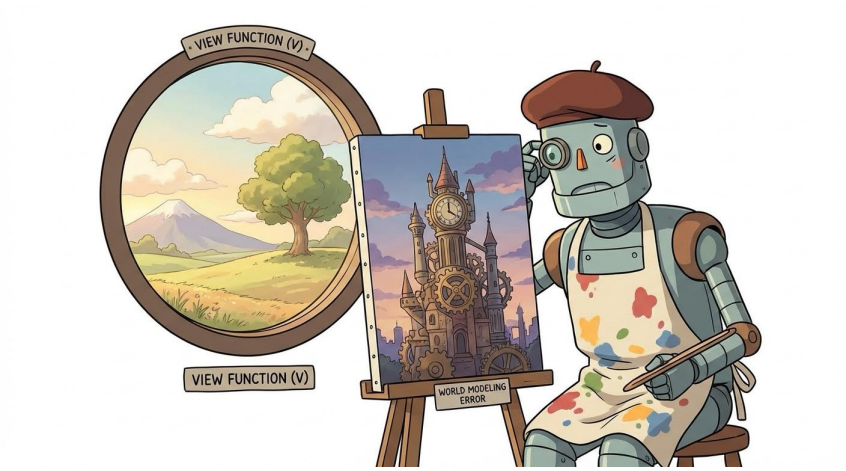

Figure 1: Hallucination as inaccurate (internal) world modeling

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

1 A History of Hallucination Definitions

1.1 Early Definitions: Generations Ungrounded in the Source

The original notion of hallucination was introduced in the context of neural machine translation, and referred to translations which were unrelated to the source text to be translated Lee et al. [2019]. This original definition focused on perturbations to an input sentence which would cause the translation to be ungrounded in the input (i.e., to have few tokens in common with the reference translation). There is a notable body of follow-up work [Raunak et al., 2021, Guerreiro et al., 2023, Dale et al., 2023, Xu et al., 2023] that deals with either analysis, benchmark creation or mitigation methods based on this original NMT-based notion. Though in different contexts, this translation-oriented notion of hallucination was extended to other quasi-related settings such as surface realization Nie et al. [2019] and table-to-text generation Parikh et al. [2020].

Abstractive summarization introduced a closely related notion. A summary is said to contain hallucinations if it contains spans that are not supported by the input document, even if they might be factually correct in the real world [Maynez et al., 2020]. More fine-grained typologies, such as FRANK [Pagnoni et al., 2021], distinguish multiple error types (entity swaps, incorrect numbers, wrongly attributed facts, etc.), but this class of hallucinations are all fundamentally defined *with respect to the source document*. Also within summarization, the distinction between intrinsic and extrinsic hallucinations became standard [Ji et al., 2023]: *intrinsic* hallucinations are those that directly contradict the source, while *extrinsic* hallucinations are statements that cannot be verified from the source but are not necessarily false.

1.2 From Source Groundness to "Unfactual" Content

As language models have become more capable, the meaning of hallucination has also expanded to encompass not just generations that are not grounded in the input, but also generally unfactual outputs, even if they do not explicitly contradict any source documents. In many current works, hallucination simply refers to fluent content that is factually wrong with respect to some (often implicit) notion of world knowledge. Similar failures arise in concept-to-text generation [Lin et al., 2020, Feng et al., 2021], where models produce fluent but commonsense-violating outputs even when no explicit source document exists. The focus is less on faithfulness to a particular input document, and more on consistency with "the world". This challenge is especially pronounced in high-stakes domains such as healthcare, where models must rely on implicit domain knowledge rather than a single authoritative source, and where fluent but incorrect explanations can be misleading (e.g., Feng et al. [2023]).

Recent surveys such as Ji et al. [2023] reflect this shift by defining hallucination as "generated content that is nonsensical or unfaithful to the provided source content," folding together both plausibility with respect to contextual semantics and factual unfaithfulness under a single umbrella. This definition still foregrounds a provided source when one exists, but in practice the survey and many later works broaden "source" to include external knowledge bases or general world knowledge. Works such as Chen et al. [2024], Rashad et al. [2024] also collect corpora around variations of the generally unfactual vs. factual definition of hallucination, referring to these as "fact-conflicting" and "fact-level", respectively.

1.3 Fine-Grained Hallucination Taxonomies and Detectors

Recent work has moved toward fine-grained classification and detection of hallucinations in long-form generation. FavaBench [Mishra et al., 2024] introduces a detailed taxonomy of hallucination types (entity errors, relation errors, unverifiable statements, among others), and a benchmark of span-level human annotations over model outputs in information-seeking scenarios. The associated FAVA model is trained as a retrieval-augmented detector and editor for these fine-grained hallucinations.

Despite increased granularity, the underlying structure remains similar: each hallucination label is defined relative to a particular *reference source*, typically a collection of web documents or curated passages. Spans that contradict this source are "contradictory"; spans not entailed by it are "unverifiable." The focus is still on *factual precision* with respect to a chosen evidence set, rather than on a more general notion of world-modeling error.

HALoGEN [Ravichander et al., 2025] introduces a benchmark to measure hallucinations in language models across 9 knowledge-oriented domains such as programming, scientific attribution, summarization, numerical reasoning, and false-presupposition. They employ automatic verifiers, which decompose model responses into meaningful atomic units - i.e., the smallest independently verifiable factual statements - and then check each atomic unit against external tools (e.g., the Python package index), knowledge bases like Semantic Scholar, or LLM-based entailment classifiers.

They evaluate both response-based tasks, where the model should answer, and refusal-based tasks, where the model should abstain in the presence of flawed premises. The benchmark categorizes hallucinations into 3 types:

- **Type A:** the correct fact existed in pretraining data but the model still hallucinated
- **Type B:** an incorrect fact was in the pretraining data, or the fact is taken out of context
- **Type C:** neither a correct nor an incorrect fact was present in the pretraining data, and the model overgeneralized when making predictions

Despite this broader categorization, hallucinations are still defined relative to externally verifiable facts, rather than as failures of the model's internal world representation or decision-making process. Recent work such as Feng et al. [2024] shows that such failures are strongly influenced by data selection, mixture, and ordering during pretraining, suggesting that hallucinations may arise even when relevant information is present in pretraining but improperly emphasized.

A common assumption across these benchmarks is that hallucinations can be identified by comparing model outputs against a fixed, externally defined notion of truth. This framing presumes static contexts and single-turn responses, where factual correctness can be assessed independently of the model's subsequent behavior.

1.4 Agentic Hallucinations

The previous definitions of hallucination focus on single outputs in response to prompts. However, the notion of hallucination widened further once language models began to be used as agents. Namely, agentic setups allow models to act over multiple turns and also interact with their environment, allowing for a more action-oriented definition of hallucination. MIRAGE-Bench [Zhang et al., 2025d] is a recent benchmark that focuses on hallucinations in interactive, risky agent settings such as web automation and software engineering. Rather than single-turn QA, their unit of analysis is an *action* taken in a particular environment state and interaction history.

MIRAGE-Bench proposes a three-part taxonomy of agentic hallucinations: actions that are unfaithful to (i) task instructions, (ii) execution history, or (iii) environment observations. Concretely, an agent is said to hallucinate when it, for example, clicks a non-existent button, acts as if a prior step had succeeded when it failed, or ignores explicit instructions. They construct deterministic "snapshot" decision points and use an LLM-as-a-judge framework to label actions as faithful or unfaithful.

There is a concomitant evolution in the mitigation method landscape especially for domains such as code, to approaches that directly aim to improve the LLM's understanding of its actions as an agent in typical coding environments of interest such as repositories and filesystems. More specifically, works such as Meta's CWM language models [Copet et al., 2025] and other proximal recent work [Armengol-Estapé et al., 2025, Sharma, 2024], augment mid-training [Mo et al., 2025] or post-training data for coding tasks with information on registers, call stack, program instruction traces, etc. These approaches are not easily generalizable to non-coding, non-math domains where there is no closed-form understanding or further intrinsically deeper layers of symbolic information available to augment. Slightly distinct in method but with similar motivation, some methods seek to amplify the LLM's understanding of the codebase at test-time through various means such as iterative context engineering [Eghbali and Pradel, 2024], parallel test-time scaled solutions aggregated by a domain-aware verifier [Ravuri and Amarasinghe, 2025], and so forth.

1.5 From Text-Grounding to World Model Failure: The Multimodal Shift

The definition of hallucination in Vision-language models (VLMs) evolved in parallel to the more text-based notions of hallucination, but also converges on similar core ideas. Similar to how text-based definitions moved from being centered on consistency with a specific source towards a more

general notion of factuality with regard to an outside world, VLM definitions have also moved from surface-level inconsistencies towards a more world-model centered definition.

In early work on VLMs, hallucination taxonomies focused on static text-image inconsistencies, that is to say classifying surface-level errors such as object or attribute mismatches [Li et al., 2023]. However, this simple classification did not fully capture multimodal hallucinations, and was superseded by frameworks like HallusionBench [Guan et al., 2024], which reframed the hallucination as a **Knowledge Conflict** between the model's powerful *parametric priors* (language memory) and its *contextual understanding* (visual input). This conflict yields two primary failure modes: Language Hallucination, where strong priors override visual evidence, and Visual Illusion, where the perception module fails complex interpretation. Notably, this parallels the distinction in RAG settings between prioritizing the retrieved context and parametric knowledge as sources of truth. Related multimodal grounding approaches such as Feng et al. [2022] show that supplementing parametric language knowledge with visual context can reduce commonsense violations in text generation, highlighting how hallucinations arise from unresolved conflicts between internal priors and external world evidence.

This shift deepened further with the introduction of multi-sensory and dynamic tasks. Benchmarks such as SavvyBench [Chen et al., 2025], which focused on audio-visual scene understanding reveal failures that transcend static modality-dependent inconsistencies. These failures highlight that models' inability to synthesize and disambiguate signals across modalities may lie in an inability to maintain a model of coherent dynamic scenarios over time. Importantly, this notion of hallucination parallels newer notions in text-based domains as well, particularly in agentic settings: we are concerned that a model's generated content is not merely unfaithful with respect to some input source, but rather with respect to a plausible world which we would like the model to internalize when it attempts to generate outputs.

2 A General Definition of Hallucination

Having reviewed hallucination definitions across many tasks, including neural machine translation, summarization, multimodal generation, and agentic settings, can we observe any commonalities? In each case, though the specifics vary, the common factor is that there is a mismatch between the model's output and what we view as true. However, depending on what we take as the ground truth, different things may be viewed as "true". In summarization, truth is usually defined by the source document, regardless of facts about the outside world, while in open-domain QA, we are usually interested in real-world facts regardless of whether or not we can find some supporting document for what we generate. In VLMs, truth may be defined by visual evidence, while in agentic settings, truth may be defined by the observable state of the environment the agent is in. Despite these differences, the underlying structure is consistent: we want a model's definitions to always be in alignment with some underlying truth which can be derived from different sources.

In order to formalize this notion, we introduce the notion of a *reference world model* (Definition 1), which is a formal representation that captures what is objectively true in a given context. Typically, a "world model" in the literature refers to an agent's internal learned representation of its environment, which can be incomplete or incorrect. We define the reference world model to be the gold standard for what we ultimately want the model's internal world model to be.

Definition 1 (Reference world model)**.** *A* reference world model *is a tuple*

$$W = (\mathcal{S}, \mathcal{H}, \mathcal{R}),$$

where $\mathcal{S}$ is a set of possible world states, $\mathcal{H}$ is a set of possible interaction histories (e.g., instructions, dialogue, logs), and $\mathcal{R}$ is a set of rules or relations that constrain which $(s, h) \in \mathcal{S} \times \mathcal{H}$ are admissible.[1]

For a given input x and world model W, we assume:

- *a* view function V *that selects the portion of the world that is relevant for x:*

$$V(W, x) \subseteq \mathcal{S} \times \mathcal{H},$$

[1]Note that $\mathcal{R}$ is a bit of syntactic sugar here; it can also be folded into the definition such that we already work with only admissible states for each task. However, we keep it explicit in order to write out benchmark-level assumptions.

- *a* conflict resolution policy P *that specifies how to reconcile multiple sources of information within* $V(W, x)$ *(e.g., "KB overrides in-context text"), and*
- *a* truth function

$$T_{W,P}(x, c) \in \{\text{true}, \text{false}, \text{unknown}\}$$

that assigns a truth status to any atomic claim c *given the world model* W*, input* x*, and policy* P*.*

Definition 2 (Hallucination as inaccurate world modeling)**.** *Let* x *be an input and let a language model produce an output* y *in response to* x*. Let* $C(y)$ *denote the set of* atomic claims *expressed in* y *(e.g., factual assertions about entities, events, or states of the world) that are observable to the user.*

Given a reference world model W*, conflict policy* P*, and truth function* $T_{W,P}$ *as above, we say that* y hallucinates with respect to (W, P) *if and only if*

$$\exists\, c \in C(y) \;\text{ such that }\; T_{W,P}(x, c) = \text{false}.$$

Intuitively, hallucination occurs when the world implicitly described by the model's output y *disagrees with the reference world model* W *on at least one observable claim.*

2.1 Examples of Hallucinations Based on Our Definition

To illustrate how our framework unifies different notions of hallucination, we present four representative cases from different domains. In each example, we explicitly specify the reference world model $\mathcal{W}$, the view function $\mathcal{V}$, and a conflict policy $\mathcal{P}$ to show where a hallucination manifests via the truth function $T_{W,P}$.

2.1.1 Example 1: Document Summarization

Task: Summarize a news article about a company's quarterly earnings.

Input article (x)**:** "TechCorp reported Q3 revenue of \$2.1 billion, missing analyst expectations of \$2.3 billion. CEO Jane Smith attributed the shortfall to supply chain disruptions in Southeast Asia. The company's stock fell 8% in after-hours trading."

Model output (y)**:** "TechCorp exceeded analyst expectations with Q3 revenue of \$2.1 billion. CEO Jane Smith praised the company's strong performance despite supply chain challenges."

Reference world model $\mathcal{W}$**:**

- S (world states): The set of facts stated in the source document.
- H (interaction history): The instruction "Summarize the following article."
- R (rules): The source article is a complete and authoritative description of the events.

View function $V(\mathcal{W}, x)$**:** The complete source article is visible to the model.

Conflict policy $\mathcal{P}$**:** The source document is the single source of truth, no external knowledge should override it.

Truth function $T_{\mathcal{W},\mathcal{P}}$**:**

- Claim c_1: "TechCorp reported Q3 revenue of \$2.1 billion" $\rightarrow$ **true**.
- Claim c_2: "TechCorp exceeded analyst expectations" $\rightarrow$ **false** (contradicts "missing analyst expectations").
- Claim c_3: "CEO Jane Smith praised the company's strong performance" $\rightarrow$ **false** (contradicts attribution of "shortfall").

Conclusion: The output y hallucinates because there exists some claim c such that $T_{\mathcal{W},\mathcal{P}}(x, c) =$ false. Specifically, c_2 and c_3 are intrinsic hallucinations.

2.1.2 Example 2: Open-Domain Question Answering

Task: Answer a factual question without explicit source documents.

Input (x): "Who won the Nobel Prize in Literature in 2023?"

Model output (y): "Haruki Murakami won the Nobel Prize in Literature in 2023 for his masterful novels exploring themes of alienation and surrealism."

Reference world model $\mathcal{W}$:

- S: The relevent envelope of real-world facts from authoritative sources (Nobel announcements, news archives).
- H: The question.
- R: The official Nobel prize records uniquely determine the winner for each year.

View function $V(\mathcal{W}, x)$: The model has no explicit access to $\mathcal{W}$ during generation so it relies on parametric knowledge.

Conflict policy $\mathcal{P}$: Real-world facts override the model's parametric memory.

Truth function $T_{\mathcal{W},\mathcal{P}}$:

- Claim c_1: "Haruki Murakami won the Nobel Prize in Literature in 2023" $\rightarrow$ **false**.
- Claim c_2: "Haruki Murakami writes novels exploring themes of alienation and surrealism" $\rightarrow$ **true**.

Conclusion: The output hallucinates because c_1 is false. Here $V(\mathcal{W}, x)$ is empty, but $\mathcal{W}$ still exists. The model's output still diverges from $\mathcal{W}$, leading to hallucination.

2.1.3 Example 3: Retrieval-Augmented Generation (RAG)

Task: Answer a question using retrieved documents.

Input (x): "Freedonia lies on the east coast, above New York. What is the capital of Freedonia?"

Retrieved context: Document D_1 (Wikipedia, 2020): "Freedonia is a fictional country appearing in the 1933 Marx Brothers film *Duck Soup*. As a fictional entity, it has no real capital city."

Model output (y): "The capital of Freedonia is Freedstadt, located on the eastern coast. Freedonia has a population of two million people."

Reference world model $\mathcal{W}$:

- S: Retrieved document content + general world knowledge.
- H: The question and retrieved context.
- R: Retrieved documents are factually correct; Freedonia is fictional with no capital.

View function $V(\mathcal{W}, x)$: The retrieved document D_1 is visible; remaining knowledge is implicit.

Conflict policy $\mathcal{P}$: Retrieved documents override parametric knowledge.

Truth function $T_{\mathcal{W},\mathcal{P}}$:

- Claim c_1: "The capital of Freedonia is Freedstadt" $\rightarrow$ **false**.
- Claim c_2: "Freedstadt is located on the eastern coast" $\rightarrow$ **false**.
- Claim c_3: "Freedonia has a population of two million people" $\rightarrow$ **unknown**. (Freedonia is fictional, but its population in the fictional world is unknown)

Conclusion: The output contradicts the retrieved evidence (and wrongfully attributes truth to the input context when the conflict policy is defined to prioritize the retrieved evidence) so hence the model hallucinates. Although the third claim about Freedonia is irrelevant, its truth status is unknown unless a document can be retrieved validating its population size.

2.1.4 Example 4: Agentic Web Navigation a.k.a Browser Use

Task: An agent is instructed: "Find and click the 'Submit' button on the payment form."

Environment state (current DOM):

```
<form id="payment-form">
  <input type="text" name="card-number" />
  <input type="text" name="cvv" />
  <button id="confirm-btn">Confirm Payment</button>
  <button id="cancel-btn">Cancel</button>
</form>
```

Model output (y): Action: `click(button#submit-btn)`. Reasoning: "I will now click the Submit button to complete the payment."

Reference world model $\mathcal{W}$:

- S: DOM structure of the current webpage.
- H: Task instructions + interaction history.
- R: The set of interactive elements is exactly that specified by the DOM.

View function $V(\mathcal{W}, x)$: The agent receives a textual/structured DOM.

Conflict policy $\mathcal{P}$: The DOM is ground truth; instructions cannot override reality.

Truth function $T_{\mathcal{W},\mathcal{P}}$:

- Claim c_1: "A button with id 'submit-btn' exists" $\rightarrow$ **false**.
- Claim c_2: "There is a Submit button available" $\rightarrow$ **false**.

Conclusion: The agent hallucinated an element that does not exist.

Insight: In agentic settings we distinguish:

- **Hallucination:** Incorrect belief about environment state.
- **Planning error:** Correct beliefs, incorrect action.
- **Instruction-following error:** Correct beliefs, goal ignored.

2.1.5 Takeaways from Examples

Example	$\mathcal{W}$ **(Reference World)**	V **(Visible Inputs)**	$\mathcal{P}$ **(Conflict Policy)**	**Hallucination Type**
Summarization	Source document	Full document	Document is truth	Intrinsic contradiction
Open QA	Real-world facts	Parametric memory only	Real world is truth	Incorrect factual claim
RAG	Retrieved docs + world knowledge	Retrieved docs (+ memory)	Docs override memory	Context contradiction
Agentic	Environment DOM	DOM observation	Environment is truth	Observation hallucination

Table 1: Comparison of hallucination types under our unified framework. Each type arises from different choices of the reference world model $\mathcal{W}$, view function V, and conflict policy $\mathcal{P}$.

Across all settings, hallucination is detected when a claim c in the output satisfies $T_{\mathcal{W},\mathcal{P}}(x, c) = \text{false}$. The diversity and detection of hallucination is entirely self contained by what constitutes the reference world, visible inputs, and conflict policy.

3 Utility of the Definition

A reader may somewhat justly ponder what the point is of introducing yet another definition on top of the pre-existing ones: are we not simply introducing more baggage on top of an already overloaded term? However, we argue that the problem is not that we lack a good definition of hallucination, but that we have too many that capture different facets, without a streamlined and unified definition.

Making assumptions explicit. We do not claim to be the first to observe that hallucinations reflect a mismatch with reality. What we do claim is that most current definitions *implicitly* assume some source of truth without spelling out what that source is, how it is accessed, and how conflicts are resolved. Our framework forces these assumptions out into the open.

In our view, any hallucination definition must implicitly specify: (i) a reference world model W that encodes what counts as true or false for the task; (ii) a view function V that determines which parts of W the model is supposed to have access to for a given input x; and (iii) a conflict policy and truth function $T_{W,P}$ that turn potentially inconsistent evidence into categorical labels (*true*, *false*, *unknown*) for atomic claims. Much of the confusion in the literature stems from leaving W, V, and P/T underspecified. Our definition of hallucination as *inaccurate world modeling that is visible by the user* is, in essence, a statement that one should not claim hallucination without first specifying what W and the accompanying two elements are.

Guiding benchmark design. A second motivation is practical: the world-model view gives a clean design space for hallucination benchmarks. Under our framework, any benchmark must explicitly answer:

- What is the reference world model W? (e.g., a source document, a collection of documents, a knowledge base, an environment state and its history, or some combination, etc.)
- What is the view $V(W, x)$ made available to the model? (e.g., the full document, a retrieved subset, a snapshot of a web page, or partial observations in an environment.)
- How are conflicts resolved, and what is the truth function $T_{W,P}$? (e.g., "document overrides KB," "DOM snapshot is ground truth for the visible page," or "unknown facts must be treated as such.")

Once W, V, and $T_{W,P}$ are explicit, hallucination evaluation becomes straightforward: an output is hallucinated if and only if it implies at least one observable claim c with $T_{W,P}(x, c) = \text{false}$. This perspective brings together settings that are currently treated disparately - e.g., summarization, open-domain QA, RAG, and agentic decision-making - all under the aegis of a common evaluation template.

It also motivates our proposed benchmark direction. By working with synthetic but fully specified worlds, we can construct tasks where W is programmatically known, V is precisely controlled, and $T_{W,P}$ can be computed exactly. This enables large-scale, language-level hallucination benchmarks in which labels are defined by construction, rather than via another LLM or human annotator.

Clarifying what is, and is not, hallucination. A third motivation is conceptual. In current usage, "hallucination" often collapses several distinct failure modes:

- *World-modeling errors*, where the model's internal picture of the world is simply wrong.
- *Planning errors*, where the model has a basically correct picture of the world but chooses a poor plan or action.
- *Incentive or reward errors*, where the model may know it is uncertain but has been trained or prompted to produce confident-sounding answers regardless.

Our definition intentionally targets only the first category. We say that an output hallucinates when the implied world in the model's text or actions contradicts W; we do not call every wrong answer a hallucination. Errors can arise from label noise, distribution shift, adversarial perturbations, or misaligned incentives. Hallucination, in our sense, is the subset of errors that correspond to incorrect beliefs about the reference world.

One way to phrase this distinction is: *error is about outputs; hallucination is about the implied world those outputs assume*. An agent that chooses a suboptimal but faithful plan is making a control or planning mistake, not hallucinating. An agent that claims to have clicked a button that does not exist, or to be on a page that the DOM shows is different, is hallucinating: its world model is wrong.

Unifying, not renaming. Finally, there is a social reason to be explicit. The term "hallucination" is now entrenched; attempting to ban or entirely replace it is unrealistic. Our goal is not to proclaim one

true definition and discard existing practice, but to provide a unifying lens on what researchers are already doing.

From this perspective, early faithfulness-w.r.t.-source definitions, modern factuality benchmarks, fine-grained span-level taxonomies, and agentic hallucination benchmarks can all be seen as instantiations of the same general template, with different choices of W, V, and $T_{W,P}$. Rather than adding yet another incompatible notion, we aim to standardize the language: when we say that a model hallucinated, we will always specify the reference world model w.r.t. which that statement is made.

This unification also helps more neatly organize mitigation work. Methods that change the information available to the model (better retrieval, richer environment state representations, multimodal grounding) are interventions on V; methods that change how truth is judged or expressed (external verifiers, abstention training, calibration) are interventions on $T_{W,P}$ or on incentives; architectural or training changes that improve the model's internal approximation to W are world-modeling improvements. Making these targets explicit helps clarify when two methods are actually addressing the same problem or not.

In summary, our contribution is not the observation that hallucinations are "wrong" with respect to reality, but the explicit introduction of a reference world model formalism and the argument that hallucination should be reserved for *inaccurate world modeling* with respect to that formalism. We believe this provides both conceptual clarity and practical guidance for the design of future hallucination benchmarks and mitigations.

4 Enabling the Creation of Larger-Scale Benchmarks

One practical use of viewing hallucination as world-modeling failure is that we can take advantage of the many existing synthetic or real environments which have an explicitly defined world, such as game environments. As we can specify all the components of our hallucination definition in these cases, we can generate a very large number of scenarios from these environments with varying complexities in order to investigate under what circumstances models tend to hallucinate. Whenever we can specify W_{ref} and a truth function $T_{W,P}(x, c)$ that evaluates atomic claims c in context x, we can generate a large number of instances without additional human or model annotation.

4.1 A General Recipe for Benchmarks

In general, given an environment backend that gives us a world state and history W, our construction proceeds in three steps. First, a document generator Γ_{docs} produces a set of textual artifacts D such as synthetic articles, environment logs, or rendered web pages. In practice, these could be produced directly from the state representation as well, or also paraphrased by an LM, but the details are not important for the general recipe.

Second, we need a query generator Γ_{query} that samples one or more queries $x \sim \Gamma_{\text{query}}(W, D)$ given the current world state that probe knowledge of the world in different ways. These could include factual questions, questions about the sequence of events or counterfactuals, as well as requests for free-form explanations of the situation.

Third, at evaluation time, a view function $V(W, x)$ that we can define and vary determines what information about the world is visible to the model, returning a context that may include the query alone, the query along with different representations of the world and history, or retrieved documents from a datastore.

During evaluation, the model produces an output y given $(x, V(W, x))$. If the query is a factual or multiple-choice instance, we can simply check the answer given by the model. If it involves free-form generation, we can use a lightweight claim-extraction model to decompose y into atomic claims $C(y)$, and our truth function $T_{W,P}(x, c)$ will label each claim as supported, contradicted, or in an unknown state. Hallucinations are precisely those claims which contradict the world state.

4.2 Example: Chess as a Hallucination Benchmark

To give a concrete example, we illustrate the benchmark creation process with the game of chess. Here, the world W consists of the state space and rules describing chess, and a state s is a specific

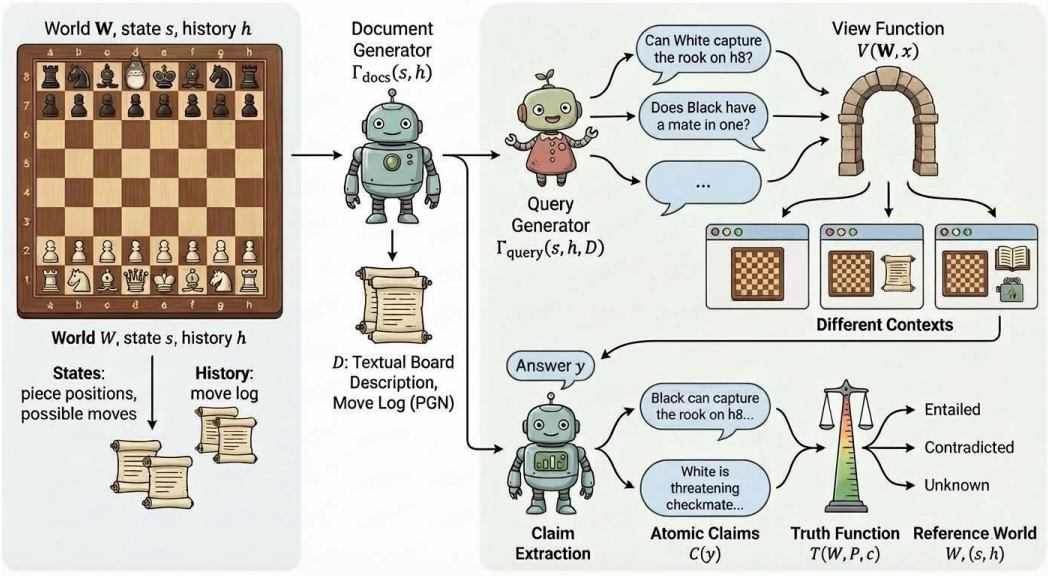

Figure 2: Overview of the chess scenario environment and instances

board position while h describes the moves that led up to that position. Figure 2 illustrates the chess environment backend as well as the construction of specific instances given a specific (s, h) pair.

Given (s, h), a document generator Γ_{docs} produces textual artifacts $D = \Gamma_{\text{docs}}(s, h)$ that serve as possible context for the model. For chess, these could include a textual description of the current board (i.e. which pieces are on which squares), or a move log in a standard notation such as PGN. These documents can be directly generated from the structured state, or potentially rewritten by an LM into a more language-based form. Next, given this, a query generator samples one or more queries $x \sim \Gamma_{\text{query}}(s, h, D)$ that probe understanding of the current position in different ways. For instance, we could generate multiple-choice questions such as

"Does Black have a mate in one?" or "Can White capture the rook on h8?"

or more open-ended prompts such as

"What is the best move for White in this position?" or "How many times did Black blunder a piece?"

At evaluation time, a view function $V(W, x)$ determines what information about (s, h) is visible to the model, returning a context. In the chess environment this might include: (i) the current board only, (ii) the board plus the full move history, or (iii) a small set of derived documents, such as textual commentary on the position or similar games. By varying V, we can control observability and retrieval quality while keeping the underlying world (s, h) fixed.

Note that the same (s, h) can support many different questions, from simple yes or no multiple-choice questions ranging to free-form detailed analysis of the position. For free-form generation questions, our claim-extraction component decomposes y into a set of atomic claims $C(y)$, for example:

"Black can capture the rook on h8 in one move."
"White is currently threatening checkmate."

Our truth function $T_{W,P}(x, c)$ then evaluates each claim $c \in C(y)$ against the reference world, using the rules in W and the specific state-history pair (s, h) to determine whether c is entailed, contradicted, or in an unknown truth state. For chess, this would just involve checking claims about possible moves against the set of legal moves, and checking that claims about which squares pieces are on are correct. By default, any claim that contradicts this would be labeled as a hallucination. For instance, the claim "White can capture black's queen in one move." would be a contradiction in the

board state shown in Figure 3, where there is a pawn in the way and thus capturing the queen in one move would violate the rules of the game.

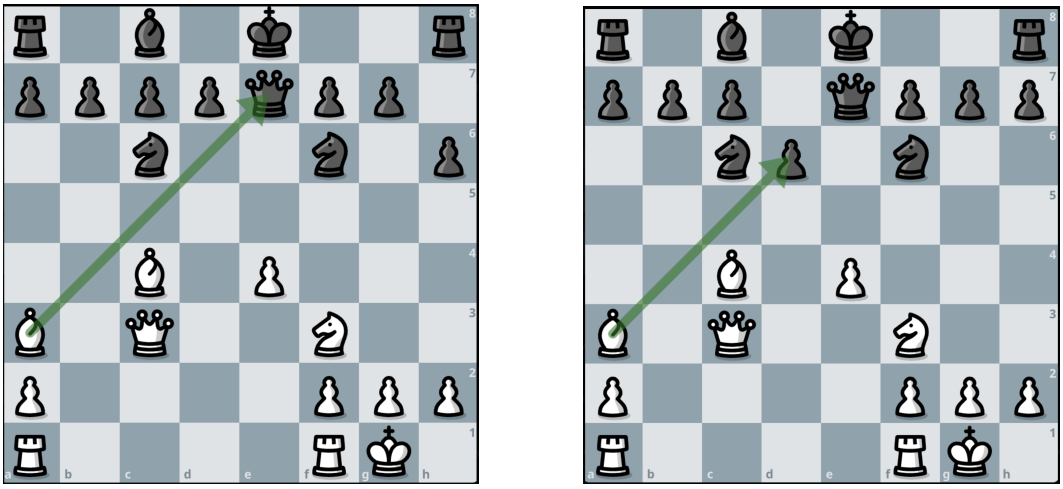

Figure 3: Misaligned move - the model's prediction (left) vs. reality (right)

Even in this simple setting, this construction allows us to have fine-grained control over many factors. These include the complexity of the environment (e.g., a simple endgame, where only a few pieces are on the board as opposed to a complex middlegame, with many pieces on the board and possible next moves), the observability of the environment to the model (e.g., the different documents available to the model in context), as well as the query type and answer format. Additionally, because hallucination labels are obtained automatically, we can search for settings where models tend to hallucinate more and sample more queries in those scenarios. In this case, if models tend to hallucinate more in the middlegames and with games that have a longer move history, we can sample queries associated with these game states.

Note that we do not lay claim to any novelty here on account of proposing QA around contextual chess understanding as an interesting evaluation setting for LLMs. Indeed, we acknowledge the body of extant work using chess understanding to evaluate LLMs, such as ChessQA [Wen et al., 2025], informal investigations of LLM chess playing ability [Tadepalli, 2024], or even earlier works evaluating related skills such as board reconstruction from PGN move sequences [Cooper, 2025] and chess commentary [Kim et al., 2025].

Unlike these works, we do not seek to propose a new chess understanding or QA benchmark for LLMs, but rather earmark chess understanding as an illustrative first example setting to construct a hallucination benchmark around our proposed definition. Chess currently lies in a goldilocks zone of appropriate difficulty for LLMs, as backed up by recent work which shows that frontier models still struggle with illegal moves and position tracking, making it a useful environment to study hallucinations in Tadepalli [2024].

That said, the chess example we gave is just one instantiation of our general framework. Even if LLMs eventually stopped hallucinating and reached superhuman capabilities specifically for chess, we can also apply our environment construction recipe to other environments with defined environment dynamics, such as Nethacker [Küttler et al., 2020, Piterbarg et al., 2023] or Crafter in the BALROG collection [Paglieri et al., 2024], which are challenging even for expert human players. In general, through constructing synthetic environments or using existing challenging games, it may be possible to probe a model's propensity to hallucinate in almost any environment.

4.3 Beyond World-modeling Benchmarks

It is important to note that this use of environments is conceptually distinct from existing work that probes *latent* world models or evaluates game-playing strength. Chess-based language model studies typically ask whether a model has learned an accurate internal representation of the board [Toshniwal et al., 2022, Karvonen, 2024] and can produce strong moves; metrics are Elo, legal-move accuracy,

or the linear decodability of hidden states. In our framework, by contrast, the environment is used as an explicit reference world against which we evaluate *textual behavior*. The question is not whether the model's internal world model is optimal, but whether, given a specified view $V(W, x)$, its stated claims conflict with (s, h). A model could have an excellent latent world model and still hallucinate by making overconfident statements under partial information, or it could have an imperfect latent model yet avoid hallucinations by abstaining when uncertain.

This perspective is useful because many real applications do not necessarily care about an LM's internal representations but rather about what it *actually says or does* relative to a known database, log, or simulator state. By treating environments as reference worlds and scoring claim-level consistency with them, our approach turns existing simulators and structured domains into scalable hallucination benchmarks, with controllable knobs for observability, difficulty, and retrieval quality.

5 Conclusion

Hallucination has become an overloaded term across many tasks and aspects of machine learning including translation, summarization, open-domain QA, RAG, multimodal generation, and agentic systems. Different communities use hallucination to mean different kinds of mismatch. This paper argues that these variants are not incompatible but rather instantiations of the same underlying template, once one makes the implicit assumptions explicit.

We propose a unified definition in which hallucination is a failure of *inaccurate (internal) world modeling* that becomes *observable* through the model's outputs: a model hallucinates (with respect to a specified reference) if it produces an output that implies at least one atomic claim that is *false* under a reference world model W and a conflict policy P. This framework forces evaluations to state 1) what the reference world is (W), 2) what information is visible (V), and 3) how truth is determined under conflicts ($T_{W,P}$). In turn, this clarifies what should and should not be called hallucination (as opposed to planning or reward errors), and provides a common language for comparing benchmarks and mitigation techniques. This lens naturally enables a scalable benchmark family: by grounding hallucination labels in synthetic or simulator-defined worlds, one can precisely control observability, difficulty, and conflict policies, and compute hallucination labels by construction.

6 Future Work

Our framework opens up a large design space for new hallucination benchmark design and we highlight several directions that we believe are especially important.

(1) Testing conflict policies P. Conflict resolution is often the hidden source of disagreement across hallucination papers (i.e., does the source document override parametric memory? Does a retrieved snippet override the prompt? Does an environment observation override instructions?, etc.). Our framework makes P explicit. The next step is to 1) systematically vary P in benchmark suites, and 2) study whether models can be trained to follow a given policy reliably. This includes adversarial tests where sources conflict or are partially corrupted, and the correct behavior will need to depend on the stated policy. Further, we could assess scenarios where the model may need to correctly infer or discover the conflict policy itself, either by interacting with the user or through logical assumptions.

Note that there already exists a significant body of work studying knowledge conflicts or related contradictions between and across model parameters as well as contextual elements, e.g., retrieval augmented sources, system prompt, etc. [Wang et al., 2023, Cheng et al., 2024, Cattan et al., 2025, Liu et al., 2023]. A subthread of work also studies settings with language subsets that are inclined to surface such conflicts, such as hypothetical statements [Basmov et al., 2024] and counterfactuals [Yamin et al., 2025]. Naturally, a consequent subthread of work studies ways of mitigating downsides of such knowledge conflicts [Bi et al., 2025, Huang et al., 2024, Zhang et al., 2025c, Zhou et al., 2025, Zhang et al., 2025b, Pham et al., 2024].

However, our work is, to the best of our knowledge, the first to explicitly recommend stating conflict policy directly available to the LLM as a pre-condition for soundly studying hallucination detection as well as its mitigation in a fair, well-defined fashion.

(2) Handling a changing (reference) world W. In real world settings, there arise many situations where a language model needs to significantly adapt its world knowledge, and consequently its own internal world model, to align with an ever-changing reference world model W. More complicated scenarios could introduce new information or alter existing information in W, potentially even non-stationarily during a long-horizon interaction as the history (context) h gets increasingly long. Designing hallucination benchmarks that can appropriately assess models in such scenarios with a significantly changing W represents an important next step.

We acknowledge the existing collection of related work referred to by term variants such as "knowledge editing" [Zhang et al., 2024, 2025a, Li et al., 2025], "knowledge updating" [Ni et al., 2024, He et al., 2025, Yu and Ji, 2024, Zhang et al., 2025e], or more broadly, "temporal evolution/generalization" [Zhu et al., 2024, Tang et al., 2025]. These threads of work introduce evaluation benchmarks/settings and mitigation methods for how LLMs can adapt to changed world knowledge beyond what is already baked into its parameters due to training on data until a cutoff date.

Nonetheless, our work is the first to propose unifying these threads into the wider umbrella of hallucination detection work by making the reference world W a first-citizen element of our definition, and hence of resultant benchmarks/settings. This will also enable cross-utilization of mitigation methods in newer hallucination scenarios with a changing W (e.g., non-stationary environments [Mao and Zhang]). Further, one could design challenging settings with co-variation of W and P both, for example, with a changing W and a non-trivial conflict policy that admits changes in some aspects but blocks them on others (e.g., temporally complex notions such as *status quo ante* [Barale et al., 2025] and rescission [Brooks and Stremitzer, 2011] in legal settings).

(3) Benchmark families beyond chess: richer worlds, longer histories, partial observability. Chess is a clean base case, but future work should broaden the environment suite to cover failure modes closer to deployment:

- **Web/DOM worlds:** dynamic pages, A/B variants, and tool feedback loops (agentic observation hallucinations).
- **Codebase/Repos:** compilation/test outcomes and versioned file states as W (hallucinated APIs [Spracklen et al., 2025], non-existent files, incorrect build claims).
- **Databases/logs:** structured enterprise records with controlled access patterns (RAG-style conflicts and temporal staleness).
- **Multimodal simulators:** audio-visual or embodied settings where W is multi-sensory and the main challenge is cross-modal state integration.

(4) Agentic settings: separating belief errors from control errors in interactive loops. In multi-turn environments, wrong outcomes can arise from 1) incorrect beliefs (hallucination under our definition), 2) correct beliefs but poor planning, or 3) misaligned incentives to appear confident. A next step is to build interactive benchmarks where W includes both state and execution traces, and where the evaluation separately scores claim-level state consistency, action validity, and task success. This helps prevent the term "hallucination" from becoming shorthand for *any* form of agent failure.

The term hallucination is unlikely to disappear. Our aim is to make its use both precise and encompassing enough that results are comparable across tasks and settings, while still being actionable enough to construct benchmarks which can drive real improvements. Perfect world modeling is likely not attainable for LLMs, nor should it be the goal. Even humans have incomplete and imperfect representations of reality, especially in fields they are not adept in. However, the goal is not to become an omniscient expert, but to rather recognize the boundaries of one's knowledge, abstain when not confident, take into account external information, and keep updating one's knowledge over time. This is a worthy goal for both humans and language models.

Our framework aligns with this goal by unifying hallucination definitions from different domains and decomposing hallucination into constituent components. Importantly, we also view the proposed benchmark family as a concrete path toward measuring, diagnosing, and ultimately reducing observable world-model errors in model behavior. As language models become more capable and are deployed in higher-risk settings, systematic understanding of when language models hallucinate becomes increasingly important. Our hope is by making explicit what has been implicit previously, we can accelerate progress towards further understanding why hallucinations occur.

References

Jordi Armengol-Estapé, Quentin Carbonneaux, Tianjun Zhang, Aram H Markosyan, Volker Seeker, Chris Cummins, Melanie Kambadur, Michael FP O'Boyle, Sida Wang, Gabriel Synnaeve, et al. What i cannot execute, i do not understand: Training and evaluating llms on program execution traces. *arXiv preprint arXiv:2503.05703*, 2025.

Claire Barale, Leslie Barrett, Vikram Sunil Bajaj, and Michael Rovatsos. Lextime: A benchmark for temporal ordering of legal events. *arXiv preprint arXiv:2506.04041*, 2025.

Victoria Basmov, Yoav Goldberg, and Reut Tsarfaty. Llms' reading comprehension is affected by parametric knowledge and struggles with hypothetical statements. *arXiv preprint arXiv:2404.06283*, 2024.

Baolong Bi, Shenghua Liu, Yiwei Wang, Yilong Xu, Junfeng Fang, Lingrui Mei, and Xueqi Cheng. Parameters vs. context: Fine-grained control of knowledge reliance in language models. *CoRR*, 2025.

Richard RW Brooks and Alexander Stremitzer. Beyond ex post expediency-an ex ante view of rescission and restitution. *Wash. & Lee L. Rev.*, 68:1171, 2011.

Arie Cattan, Alon Jacovi, Ori Ram, Jonathan Herzig, Roee Aharoni, Sasha Goldshtein, Eran Ofek, Idan Szpektor, and Avi Caciularu. Dragged into conflicts: Detecting and addressing conflicting sources in search-augmented llms. *arXiv preprint arXiv:2506.08500*, 2025.

Mingfei Chen, Zijun Cui, Xiulong Liu, Jinlin Xiang, Caleb Zheng, Jingyuan Li, and Eli Shlizerman. Savvy: Spatial awareness via audio-visual llms through seeing and hearing. *arXiv preprint arXiv:2506.05414*, 2025. URL `https://arxiv.org/abs/2506.05414`.

Xiang Chen, Duanzheng Song, Honghao Gui, Chenxi Wang, Ningyu Zhang, Yong Jiang, Fei Huang, Chengfei Lyu, Dan Zhang, and Huajun Chen. Factchd: benchmarking fact-conflicting hallucination detection. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 6216–6224, 2024.

Sitao Cheng, Liangming Pan, Xunjian Yin, Xinyi Wang, and William Yang Wang. Understanding the interplay between parametric and contextual knowledge for large language models. *arXiv preprint arXiv:2410.08414*, 2024.

Aidan Cooper. Pgn2fen: A benchmark for evaluating llm chess reasoning. `https://www.aidancooper.co.uk/pgn2fen-benchmark/`, 2025. Aidan Cooper's site, accessed 2025-12-14.

Jade Copet, Quentin Carbonneaux, Gal Cohen, Jonas Gehring, Jacob Kahn, Jannik Kossen, Felix Kreuk, Emily McMilin, Michel Meyer, Yuxiang Wei, et al. Cwm: An open-weights llm for research on code generation with world models. *arXiv preprint arXiv:2510.02387*, 2025.

David Dale, Elena Voita, Janice Lam, Prangthip Hansanti, Christophe Ropers, Elahe Kalbassi, Cynthia Gao, Loïc Barrault, and Marta Costa-jussà. Halomi: A manually annotated benchmark for multilingual hallucination and omission detection in machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 638–653, 2023.

Aryaz Eghbali and Michael Pradel. De-hallucinator: Mitigating llm hallucinations in code generation tasks via iterative grounding. *arXiv preprint arXiv:2401.01701*, 2024.

Steven Feng, Shrimai Prabhumoye, Kezhi Kong, Dan Su, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. Maximize your data's potential: Enhancing llm accuracy with two-phase pretraining, 2024. URL `https://arxiv.org/abs/2412.15285`.

Steven Y. Feng, Jessica Huynh, Chaitanya Prasad Narisetty, Eduard Hovy, and Varun Gangal. SAPPHIRE: Approaches for enhanced concept-to-text generation. In Anya Belz, Angela Fan, Ehud Reiter, and Yaji Sripada, editors, *Proceedings of the 14th International Conference on Natural Language Generation*, pages 212–225, Aberdeen, Scotland, UK, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.inlg-1.21. URL `https://aclanthology.org/2021.inlg-1.21/`.

Steven Y. Feng, Kevin Lu, Zhuofu Tao, Malihe Alikhani, Teruko Mitamura, Eduard Hovy, and Varun Gangal. Retrieve, caption, generate: Visual grounding for enhancing commonsense in text generation models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10618–10626, Jun. 2022. doi: 10.1609/aaai.v36i10.21306. URL https://ojs.aaai.org/index.php/AAAI/article/view/21306.

Steven Y. Feng, Vivek Khetan, Bogdan Sacaleanu, Anatole Gershman, and Eduard Hovy. CHARD: Clinical health-aware reasoning across dimensions for text generation models. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 313–327, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.24. URL https://aclanthology.org/2023.eacl-main.24/.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14375–14385, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/papers/Guan_HallusionBench_An_Advanced_Diagnostic_Suite_for_Entangled_Language_Hallucination_and_CVPR_2024_paper.pdf.

Nuno M Guerreiro, Elena Voita, and André FT Martins. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1059–1075, 2023.

Guoxiu He, Xin Song, and Aixin Sun. Knowledge updating? no more model editing! just selective contextual reasoning. *arXiv preprint arXiv:2503.05212*, 2025.

Yukun Huang, Sanxing Chen, Hongyi Cai, and Bhuwan Dhingra. To trust or not to trust? enhancing large language models' situated faithfulness to external contexts. *arXiv preprint arXiv:2410.14675*, 2024.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Delong Chen, Wenliang Dai, Ho Shu Chan, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(14):1–38, 2023. doi: 10.1145/3571730. URL https://arxiv.org/abs/2202.03629.

Adam Karvonen. Emergent world models and latent variable estimation in chess-playing language models. *arXiv preprint arXiv:2403.15498*, 2024.

Jaechang Kim, Jinmin Goh, Inseok Hwang, Jaewoong Cho, and Jungseul Ok. Bridging the gap between expert and language models: Concept-guided chess commentary generation and evaluation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9497–9516, 2025.

Heinrich Küttler, Nantas Nardelli, Alexander Miller, Roberta Raileanu, Marco Selvatici, Edward Grefenstette, and Tim Rocktäschel. The nethack learning environment. *Advances in Neural Information Processing Systems*, 33:7671–7684, 2020.

Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. Hallucinations in neural machine translation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019. URL https://openreview.net/forum?id=SkxJ-309FQ.

Belinda Z Li, Emmy Liu, Alexis Ross, Abbas Zeitoun, Graham Neubig, and Jacob Andreas. Language modeling with editable external knowledge. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3070–3090, 2025.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. URL https://arxiv.org/abs/2305.10355.

Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1823–1840, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.165. URL `https://aclanthology.org/2020.findings-emnlp.165/`.

Genglin Liu, Xingyao Wang, Lifan Yuan, Yangyi Chen, and Hao Peng. Examining llms' uncertainty expression towards questions outside parametric knowledge. *arXiv preprint arXiv:2311.09731*, 2023.

Yihuan Mao and Chongjie Zhang. A bayesian fast-slow framework to mitigate interference in non-stationary reinforcement learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1906–1919. Association for Computational Linguistics, 2020. URL `https://aclanthology.org/2020.acl-main.173`.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. Fine-grained hallucination detection and editing for language models. In *Proceedings of the 1st Conference on Language Modeling (COLM)*, 2024. URL `https://arxiv.org/abs/2401.06855`.

Kaixiang Mo, Yuxin Shi, Weiwei Weng, Zhiqiang Zhou, Shuman Liu, Haibo Zhang, and Anxiang Zeng. Mid-training of large language models: A survey. *arXiv preprint arXiv:2510.06826*, 2025.

Shiwen Ni, Dingwei Chen, Chengming Li, Xiping Hu, Ruifeng Xu, and Min Yang. Forgetting before learning: Utilizing parametric arithmetic for knowledge updating in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5716–5731, 2024.

Feng Nie, Jin-Ge Yao, Jinpeng Wang, Rong Pan, and Chin-Yew Lin. A simple recipe towards reducing hallucination in neural surface realisation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2673–2679, 2019.

Davide Paglieri, Bartłomiej Cupiał, Sam Coward, Ulyana Piterbarg, Maciej Wołczyk, Akbir Khan, Eduardo Pignatelli, Łukasz Kucinski, Lerrel Pinto, Rob Fergus, et al. Benchmarking agentic llm and vlm reasoning on games. *arXiv preprint arXiv:2411.13543*, 2024.

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 (NAACL-HLT)*, pages 4812–4829. Association for Computational Linguistics, 2021. URL `https://aclanthology.org/2021.naacl-main.383`.

Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. Totto: A controlled table-to-text generation dataset. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 1173–1186, 2020.

Quang Hieu Pham, Hoang Ngo, Anh Tuan Luu, and Dat Quoc Nguyen. Who's who: Large language models meet knowledge conflicts in practice. *arXiv preprint arXiv:2410.15737*, 2024.

Ulyana Piterbarg, Lerrel Pinto, and Rob Fergus. Nethack is hard to hack. *Advances in Neural Information Processing Systems*, 36:37540–37566, 2023.

Mohamed Rashad, Ahmed Zahran, Abanoub Amin, Amr Abdelaal, and Mohamed AlTantawy. Factalign: Fact-level hallucination detection and classification through knowledge graph alignment. In *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024)*, pages 79–84, 2024.

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*, pages 1172–1183, 2021.

Abhilasha Ravichander, Shrusti Ghela, David Wadden, and Yejin Choi. HALoGEN: Fantastic LLM hallucinations and where to find them. *arXiv preprint arXiv:2501.08292*, 2025. URL `https://arxiv.org/abs/2501.08292`.

Chaitanya Ravuri and Saman Amarasinghe. Eliminating hallucination-induced errors in llm code generation with functional clustering. *arXiv preprint arXiv:2506.11021*, 2025.

Asankhaya Sharma. Ellora: Enhancing llms with lora - standardized recipes for capability enhancement, 2024. URL `https://github.com/codelion/ellora`.

Joseph Spracklen, Raveen Wijewickrama, AHM Nazmus Sakib, Anindya Maiti, and Bimal Viswanath. We have a package for you! a comprehensive analysis of package hallucinations by code generating {LLMs}. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 3687–3706, 2025.

Karthik Tadepalli. Personal evaluation of llms, through chess. `https://www.lesswrong.com/posts/gNFixvxw7JxzvMjCJ/personal-evaluation-of-llms-through-chess`, 2024. LessWrong, accessed 2025-12-11.

Wei Tang, Yixin Cao, Yang Deng, Jiahao Ying, Bo Wang, Yizhe Yang, Yuyue Zhao, Qi Zhang, Xuan-Jing Huang, Yu-Gang Jiang, et al. Evowiki: Evaluating llms on evolving knowledge. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 948–964, 2025.

Shubham Toshniwal, Sam Wiseman, Karen Livescu, and Kevin Gimpel. Chess as a testbed for language model state tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11385–11393, 2022.

Yike Wang, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. Resolving knowledge conflicts in large language models. *arXiv preprint arXiv:2310.00935*, 2023.

Qianfeng Wen, Zhenwei Tang, and Ashton Anderson. Chessqa: Evaluating large language models for chess understanding. *arXiv preprint arXiv:2510.23948*, 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.

Khurram Yamin, Gaurav Ghosal, and Bryan Wilder. Llms struggle to perform counterfactual reasoning with parametric knowledge. *arXiv preprint arXiv:2506.15732*, 2025.

Pengfei Yu and Heng Ji. Information association for language model updating by mitigating lm-logical discrepancy. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 117–129, 2024.

Binchi Zhang, Zhengzhang Chen, Zaiyi Zheng, Jundong Li, and Haifeng Chen. Resolving editing-unlearning conflicts: A knowledge codebook framework for large language model updating. *arXiv preprint arXiv:2502.00158*, 2025a.

Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, et al. A comprehensive study of knowledge editing for large language models. *arXiv preprint arXiv:2401.01286*, 2024.

Qinggang Zhang, Zhishang Xiang, Yilin Xiao, Le Wang, Junhui Li, Xinrun Wang, and Jinsong Su. Faithfulrag: Fact-level conflict modeling for context-faithful retrieval-augmented generation. *arXiv preprint arXiv:2506.08938*, 2025b.

Ruizhe Zhang, Yongxin Xu, Yuzhen Xiao, Runchuan Zhu, Xinke Jiang, Xu Chu, Junfeng Zhao, and Yasha Wang. Knowpo: Knowledge-aware preference optimization for controllable knowledge selection in retrieval-augmented language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25895–25903, 2025c.

Weichen Zhang, Yiyou Sun, Pohao Huang, Jiayue Pu, Heyue Lin, and Dawn Song. MIRAGE-bench: LLM agent is hallucinating and where to find them. *arXiv preprint arXiv:2507.21017*, 2025d. URL `https://arxiv.org/abs/2507.21017`.

Xiaoying Zhang, Baolin Peng, Ye Tian, Jingyan Zhou, Yipeng Zhang, Haitao Mi, and Helen Meng. Self-tuning: Instructing llms to effectively acquire new knowledge through self-teaching. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5688–5724, 2025e.

Zeqi Zhou, Fang Wu, Shayan Talaei, Haokai Zhao, Cheng Meixin, Tinson Xu, Amin Saberi, and Yejin Choi. When to trust context: Self-reflective debates for context reliability. *arXiv preprint arXiv:2506.06020*, 2025.

Chenghao Zhu, Nuo Chen, Yufei Gao, and Benyou Wang. Evaluating llms at evaluating temporal generalization. *CoRR*, 2024.