An Information-Theoretic Model of Abduction for Detecting Hallucinations in Explanations

Boris A. Galitsky * and Alexander Rybalov

Posted Date: 8 December 2025

doi: 10.20944/preprints202512.0598.v1

Keywords: hallucination detection; abductive reasoning; information theory; minimum description length; neuro-symbolic ai; discourse analysis; entropy-based inference

 doi:10.20944/preprints202512.0598.v1

Article

An Information-Theoretic Model of Abduction for Detecting Hallucinations in Explanations

Boris Galitsky [1,*] and Alexander Rybalov [2]

[1] Knowledge Trail Inc, San Jose CA USA

[2] Tel Aviv University Israel

* Correspondence: galitsky@hotmail.com

Abstract

We present *An Information-Theoretic Model of Abduction for Detecting Hallucinations in Generative Models*, a neuro-symbolic framework that combines entropy-based inference with abductive reasoning to identify unsupported or contradictory content in large language model outputs. Our approach treats hallucination detection as a dual optimization problem: minimizing the information gain between source-conditioned and response-conditioned belief distributions, while simultaneously selecting the minimal abductive hypothesis capable of explaining discourse-salient claims. By incorporating discourse structure through RST-derived EDU weighting, the model distinguishes legitimate abductive elaborations from claims that cannot be justified under any computationally plausible hypothesis. Experimental evaluation across medical, factual QA, and multi-hop reasoning datasets demonstrates that the proposed method outperforms state-of-the-art neural and symbolic baselines in both accuracy and interpretability. Qualitative analysis further shows that the framework successfully exposes plausible-sounding but abductively unsupported model errors, including real hallucinations generated by GPT-5.1. Together, these results indicate that integrating information-theoretic divergence and abductive explanation provides a principled and effective foundation for robust hallucination detection in generative systems.

Keywords: hallucination detection; abductive reasoning; information theory; minimum description length; neuro-symbolic ai; discourse analysis; entropy-based inference

1. Introduction

Large Language Models (LLMs) have made substantial advances in natural language understanding and generation across diverse tasks. However, their practical use is limited by a persistent tendency to produce *hallucinations*—outputs that may be fluent and coherent yet factually incorrect or semantically implausible.

A broad range of techniques has been proposed for detecting unsupported or fabricated model outputs (Huang et al., 2025). Existing methods are typically categorized as white-box, gray-box, or black-box. ***White-box*** approaches use internal representations or activation patterns to flag inconsistencies (Azaria & Mitchell, 2023; Su et al., 2024), but their dependence on model internals limits cross-model applicability. ***Gray-box*** approaches rely on intermediate signals such as token probabilities or entropy (Varshney et al., 2023), though these signals often correlate imperfectly with factual correctness, especially in open-ended generation. ***Black-box*** methods, which examine only the generated text, are the most general but face their own limitations: external-knowledge approaches suffer from coverage gaps (Kossen et al., 2024; Chen et al., 2025), and heuristic strategies such as self-consistency often fail when hallucinations are linguistically fluent and semantically coherent (Galitsky, 2021). While many approaches to hallucination detection rely on external knowledge sources for fact-checking, several methods have been developed to operate in zero-resource settings, thereby eliminating dependence on retrieval. These methods rest on the premise that the genesis of LLM hallucinations is closely linked to the model's intrinsic uncertainty. If one can

 doi:10.20944/preprints202512.0598.v1

estimate the uncertainty associated with the factual content produced by the model, hallucinations can often be detected without recourse to external evidence.

Uncertainty-based strategies generally fall into two categories:

1. LLM internal states. Internal model signals—such as token-level probabilities or entropy—serve as proxies for epistemic uncertainty (Varshney et al., 2023). Low-entropy generations tend to reflect confident, predictable continuations, whereas atypically high entropy may indicate unsupported or unstable content.
2. LLM behavioral variance. The studies elicit uncertainty behaviorally, either through natural-language self-assessment prompts (Kadavath et al. 2022) or through output-level variability. For example, Manakul et al. (2023) detect hallucinations by sampling multiple responses to the same query and measuring the consistency of factual claims across samples.

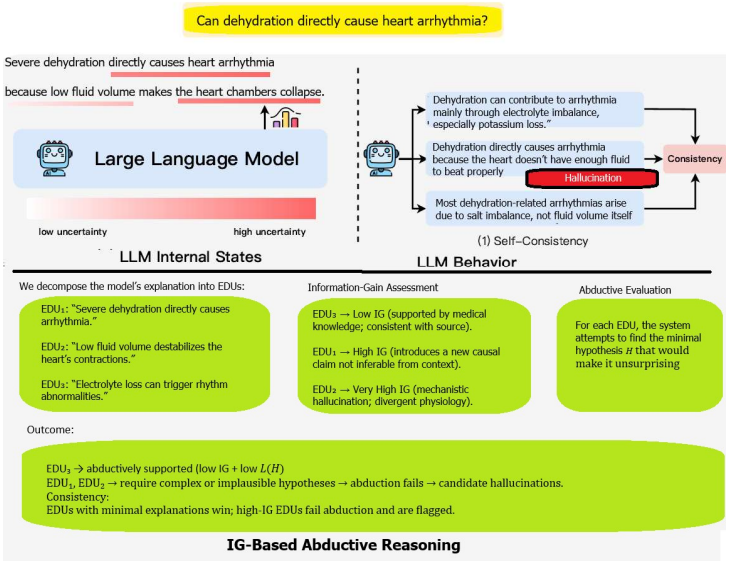

Figure 1. Illustration of our IG-based Abductive Reasoning approach (on the bottom) in comparison to LLM Internal States and LLM Behavior approaches. EDU – Elementary Discourse Units are used to access correctness of explanations.

Although such methods capture important uncertainty signals, they provide only local or surface-level indicators of instability. They do not explain why a claim is unsupported or what minimal hypothesis would be required for it to be true. This motivates our shift from merely estimating uncertainty to quantifying informational deviation and evaluating abductive plausibility. In particular, we extend uncertainty-based detection with an information-gain–driven abductive framework, where hallucinations are identified as claims whose informational divergence from the source cannot be justified by any computationally reasonable abductive hypothesis (Figure 1).

In this paper, we concentrate on a specific subclass of hallucinations that arise when a model produces claims that appear to be easily explainable by the given premises, even though the explanation is in fact incorrect. These are cases in which the model identifies a superficially plausible causal pathway connecting the premises to the conclusion, and—because the explanation is simple, salient, or heuristically attractive—treats it as valid. Crucially, the claim in question may still be *factually true*, yet the model's justification for it is faulty. This makes the hallucination particularly insidious: it is not the claim's truth-value that is compromised, but the inferential route by which the model arrives at it.

A paradigmatic example is the widely circulating misconception that *walking in cold water can cause a gout attack*. The model may generate the following reasoning: *cold temperature → uric acid*

 doi:10.20944/preprints202512.0598.v1

crystallization → gout flare. This explanation is coherent, compact, and causally intuitive—precisely the kind of abductive reasoning pattern that LLMs frequently overgenerate. However, the medical reality is substantially more complex: The combination of high temperature and low humidity had the greatest association compared with moderate temperature and average relative humidity (Neogi et al 2014). Cold exposure alone does not precipitate gout; rather, gout flares arise from interactions among metabolic factors, urate load, local tissue dynamics, and inflammatory signaling. Cold may modulate symptoms indirectly, but it is not a straightforward causal trigger. Thus, while the conclusion ("I had a gout attack after walking in cold water") could be true, the ease of the explanation masks its inaccuracy.

This phenomenon illustrates a central methodological challenge. Models tend to privilege explanations that are simple, available, and minimally costly from a cognitive perspective. When these low-complexity explanations align superficially with the structure of the premises, the model is likely to accept them uncritically—even when domain knowledge would rule them out. Our analysis therefore focuses on detecting hallucinations that stem not from fabricated facts, but from overly convenient abductive leaps: explanations that are *too easy* relative to the true causal structure underlying the domain.

The class of hallucinations we are tackling is shown in the bottom-right corner (Figure 2).

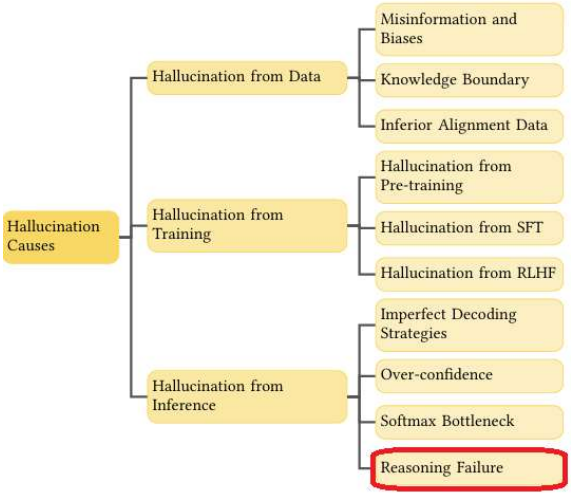

Figure 2. A taxonomy of hallucination types and the focus of this study.

The effective use of knowledge in LLMs is fundamentally constrained by their underlying reasoning capabilities. In multi-hop question answering, for example, a model may possess all the relevant information yet still fail to produce accurate answers when multiple associative pathways exist between premises and the query (Zheng et al 2023). Such ambiguity exposes the model's limited ability to perform structured reasoning across intermediate steps. Moreover, (Berglund et al. 2023) identify a characteristic reasoning failure they term the *Reversal Curse*: while an LLM may successfully answer a query posed in the form "A is B," it often fails to infer the logically equivalent converse, "B is A." This asymmetry reveals a deeper deficiency that extends well beyond simple deductive transformations, indicating structural weaknesses in how LLMs represent and manipulate relational knowledge. This failure is even stronger in case of "A explains B" and its reversal.

1.1. Example: Information-Gain analysis of a genealogical hallucination

Although the focus of this paper is on explanation-based hallucinations, to illustrate how information gain identifies unsupported model inferences of plain facts, let us consider the query *"Who are the grandmothers of Russian poet Alexander Pushkin?"*. For this well-defined biographical question, authoritative sources converge on the following facts:

- **Paternal grandmother:** *Olga Vasilyevna Chicherina*

 doi:10.20944/preprints202512.0598.v1

- **Maternal grandmother:** *Maria Alekseevna Pushkina (Hannibal)*

However, YandexGPT produces the following hallucinatory response H_0 **(hallucinated output):**

Alexander Sergeyevich Pushkin had two grandmothers—one on his father's side and one on his mother's side.
Maria Alekseyevna Ganibalkova (née Pushkina) (1745–1818) — paternal grandmother (c_1),
and Olga Vasilyevna Cherkasova (née Mineeva) (1737–1802) — maternal grandmother (c_2).

Both names in H_0 correspond to **non-existent individuals,** and neither appears in Pushkin scholarship or historical genealogical records. This discrepancy enables a clear demonstration of the information-gain framework.

Let S denote the distribution of plausible entities conditioned on verified biographical knowledge. An auditor model (e.g., NLI, QA, or a secondary LLM) assigns high probability mass to the true grandmothers (Chicherina; Pushkina/Hannibal) and effectively zero probability to fabricated entities such as *Ganibalkova* or *Cherkasova–Mineeva*. In contrast, the model's response $R = H_0$ commits strongly to these fabricated names, shifting nearly all probability mass toward non-existent individuals.

For the atomic claim *"Pushkin's paternal grandmother was Maria Alekseyevna Ganibalkova"*, the auditor estimates:

$P(Ganibalkova|S) \approx 0$ and $P(Ganibalkova|R) \approx 1$.

The resulting information gain is therefore dominated by the KL-divergence term:

$IG(c_1, S) = D_{KL}(P(\cdot|R) \parallel P(\cdot|S)) \approx \log 1/\epsilon$,

where ϵ is a small floor value used to avoid division by zero. In practice, this yields an IG score exceeding **13 bits**, far above typical hallucination thresholds (1–5 bits). An analogous computation for the fabricated maternal grandmother (c_2) yields a similarly high IG value. Aggregating claim-level scores—either by maximum or mean—produces a response-level information-gain estimate indicative of severe hallucination.

This example demonstrates the utility of IG-based detection: the model's answer introduces entities that have no support in the source-conditioned distribution, resulting in extreme divergence between $P(\cdot|S)$ and $P(\cdot|R)$. Even without external databases, the probabilistic mismatch is sufficient to classify the response as hallucinated. The case thus provides a clear empirical instance of how information gain captures unsupported factual additions in generative model outputs.

1.2. Contribution

This work introduces a discourse-aware abductive reasoning framework that unifies four complementary mechanisms for hallucination detection and explanation verification: abduction, counter-abduction, discourse weighting, and probabilistic web grounding. Together, these components transform explanation validation into a structured process of conditional justifiability:

1. The system not only identifies unsupported or inconsistent statements but also distinguishes between legitimate hypothesis formation and genuine reasoning errors.
2. By integrating abductive inference with rhetorical structure analysis, the model prioritizes nucleus-level claims and down-weights peripheral content, improving both interpretability and factual precision.
3. Counter-abduction introduces rival explanations as logical defeaters, ensuring that conclusions are robust under evidential challenge.

 doi:10.20944/preprints202512.0598.v1

4. Finally, by leveraging web-scale frequency estimates as probabilistic confirmation metrics within a minimum description length (MDL) framework, the approach generalizes fact checking into an open, distribution-free form of explanation verification.

Collectively, these contributions establish a principled and computationally grounded basis for hallucination-resistant, human-aligned neuro-symbolic reasoning across domains such as medicine, law, and scientific analysis.

2. Information-theoretic formalization of abduction

Abductive inference is traditionally understood as a qualitative process in which a reasoner selects the most plausible explanation for an observed fact. Classical philosophical treatments—from Peirce's early writings to contemporary accounts of Inference to the Best Explanation (IBE)—identify several normative criteria for evaluating candidate explanations, including *simplicity*, *coherence*, *plausibility*, and *explanatory power* (**Peirce** 1878; 1903). While these guidelines capture the intuitions behind abductive reasoning, they lack precise quantitative definitions and therefore resist operationalization in computational systems. Recent work has demonstrated that information theory provides a principled mathematical foundation capable of formalizing these criteria and turning abduction into an optimization problem over measurable quantities.

Information theory treats inference as a process of minimizing uncertainty and encoding data as efficiently as possible. Within this view, hypotheses are evaluated based on how effectively they compress the information contained in observations. This perspective naturally aligns with the key abductive desiderata. First, simplicity corresponds to the *description length* of a hypothesis: shorter, less complex hypotheses carry a lower bit-cost and are therefore preferred according to the Minimum Description Length (MDL) principle. Second, explanatory adequacy is reflected in the *conditional entropy* of the observation given the hypothesis, $H(O|H)$; a hypothesis that predicts or entails the observation well leaves little residual uncertainty and thus has low conditional entropy. Third, coherence (the degree to which the hypothesis and observation mutually support one another) maps onto *mutual information*, $I(H;O)$, which quantifies how much knowing one reduces uncertainty about the other.

The plausibility of a hypothesis is naturally encoded as its *prior probability* within a probabilistic framework; plausible hypotheses have low information content (high prior, low $-\log P(H)$ and therefore contribute minimally to the total encoding cost. Finally, surprise reduction, a central feature of explanatory reasoning, corresponds to maximizing likelihood or minimizing the negative log-likelihood of the data, thus reducing the number of bits required to encode surprising events. Together, these correspondences establish a direct mapping between abductive criteria and information-theoretic quantities (Table 1).

Table 1. Mapping between abductive criteria and information-theoretic interpretation.

Abductive criterion	Information-Theoretic Interpretation
Simplicity	Low description length of the hypothesis
Explanatory adequacy	Low conditional entropy ($H(O;H)$)
Coherence	High mutual information ($I(H;O)$)
Plausibility	High prior probability ($P(H)$)
Surprise reduction	High likelihood / low bit-cost of data given (H)

By grounding abductive reasoning in measurable information-theoretic terms, we can formalize the selection of "best explanations" as a minimization of total encoding cost or, equivalently, as an optimization over uncertainty reduction. This yields computationally tractable objectives, such as MDL-based scoring or mutual-information–based selection, that directly instantiate the philosophical criteria of abduction. The result is a rigorous, unified account in which explanatory goodness is quantified through entropy, likelihood, and description length—allowing abductive

 doi:10.20944/preprints202512.0598.v1

inference to be implemented, compared, and evaluated systematically across symbolic, probabilistic, and neuro-symbolic reasoning systems.

Let O be an observation and H a candidate explanatory hypothesis. Abduction chooses $H^*=\arg\max_H \text{Expl}(H,O)$. Information theory allows us to turn "explanatory quality" into a measurable objective.

The MDL principle states:

$H^*=\arg\min_H [L(H)+L(O|H)]$,

where:

- $L(H)$ is the number of bits needed to encode the hypothesis
- $L(O|H)$ is the number of bits needed to encode the data given the hypothesis

Abduction becomes choosing the hypothesis that yields **maximum compression.** Equivalently:

$L(O|H)=-\log P(O|H)$

Thus abduction **maximizes a likelihood with model complexity penalty**.

We now express entropy-based explanation quality. The entropy of observation is expressed as

$H(O)=-\sum_x P(x)\log P(x)$

Conditional entropy under a hypothesis:

$H(O|H)=-\sum_x P(x|H)\log P(x|H)$

A good explanation minimizes conditional entropy:

$H^*=\arg\min_H H(O|H)$

Equivalently, this H^* hypothesis makes the observation least surprising.

Also, mutual information measures the explanatory power:

$I(H;O)=H(O)-H(O|H)$

Thus:

$H^*=\arg\max_H I(H|O)$

The best explanation is the one that **provides the largest entropy reduction**.

2.1. Bayesian surprise and abductive shift

Bayesian surprise (Baldi & Itti 2009) can be expressed as

$S=D_{KL}(P(H|O)\|P(H))$

An abductive hypothesis should induce **high posterior shift**, but with low description-length cost. Hence the combined objective:

$H^*=\arg\max_H I(H;O) - L(H)$

This expression unifies informativeness, simplicity and explanatory adequacy, providing a fully information-theoretic formalization of abduction (Figure 3).

 doi:10.20944/preprints202512.0598.v1

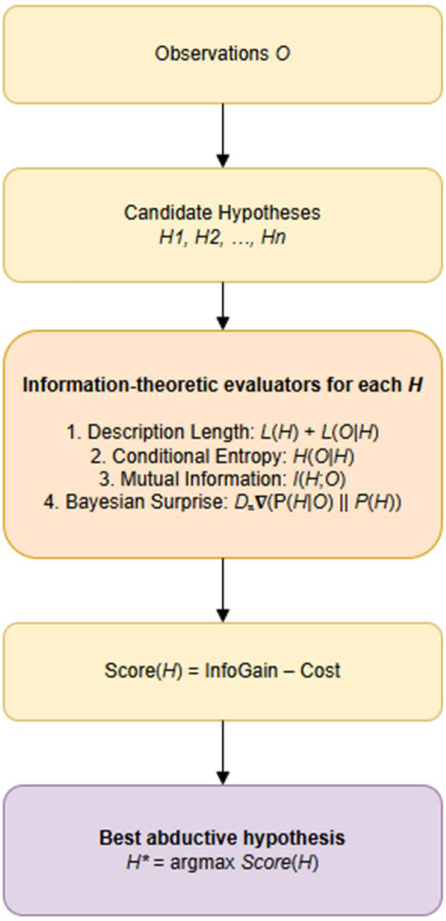

Figure 3. An algorithm for abduction + information-theoretic formalization.

2.2. Estimating description lengths via web search frequencies

To operationalize the MDL principle in settings where explicit probabilistic models are unavailable, we approximate the code lengths $L(H)$ using web-scale frequency statistics. The central idea is to exploit the web as an implicit empirical corpus: the number of indexed pages matching a query serves as a noisy but informative estimator of how probable a hypothesis or hypothesis–observation pairing is in natural language use. This approach is inspired by prior work on information-theoretic measures such as Normalized Google Distance, where search frequencies function as proxies for distributional probabilities (Fig 3a).

Let $f(q)$ denote the number of search results returned for query q, and let N denote the approximate size of the search engine's index. Although N is unknown, its precise value is unnecessary because MDL compares code lengths only up to additive constants. We therefore approximate the probability of a linguistic expression q by

$p(q) \approx f(q)/N$,

which induces an information content or code length.

$L(q) = -\log_2 p(q) = \log_2 N - \log_2 f(q)$

Because $\log_2 N$ is constant for all hypotheses, we drop it and use the simplified form

$L(q) \propto -\log_2 f(q)$.

Thus, hypotheses that appear more frequently on the web receive shorter code lengths, reflecting the intuition that widely attested statements are simpler or more conventional.

To estimate $L(H)$, we map each hypothesis H to a canonical query string q_H (e.g., a key phrase or normalized proposition). The code length is then approximated as

 doi:10.20944/preprints202512.0598.v1

$L(H) \approx -\log_2 f(H)$.

The conditional length $L(O|H)$ is derived by treating joint search frequencies as empirical co-occurrence counts. Let $f(H,O)$ denote the number of results returned when the query enforces both H and O (e.g., through conjunction or a joint phrase). A conditional probability estimator follows:

$p(O|H) \approx f(H,O) / f(H)$.

Substituting this into the MDL expression yields

$L(O|H) = -\log_2(O|H) = -\log_2 f(H,O) + \log_2 f(H)$.

In many practical settings the combined MDL score simplifies to a single term dominated by the joint frequency:

$L(H) + L(O|H) \approx - -\log_2 f(H,O)$,

meaning that the preferred hypothesis is the one that most frequently co-occurs with the observation in the web corpus.

Because web counts are inherently noisy, we apply standard smoothing—for example, replacing each frequency with $f'(q) = f(q) + \alpha$ to avoid undefined logarithms—and ensure consistent query normalization across hypotheses. Despite the noise, this frequency-based MDL approximation provides a robust, scalable mechanism for ranking hypotheses using ubiquitous web signals, and requires no domain-specific probability model.

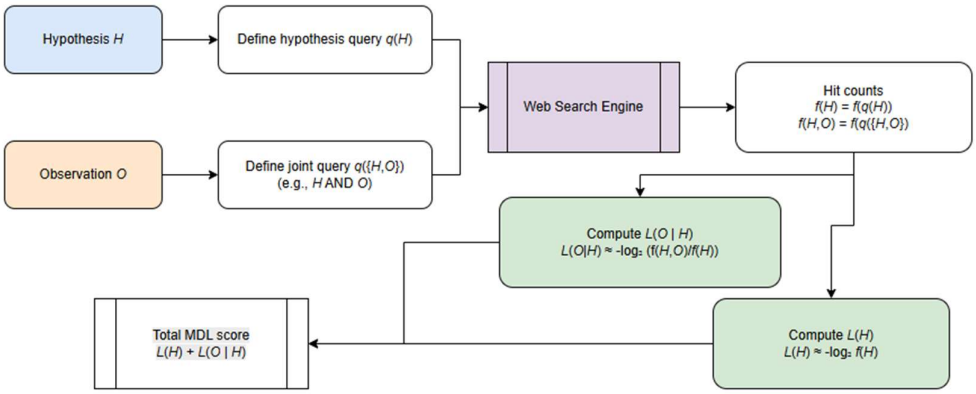

Figure 3. a: Estimating description length via web search frequencies.

3. Abduction as a structural corrective layer for Chain-of-Thought reasoning

Chain-of-Thought (CoT) prompting has become a dominant strategy for eliciting multi-step reasoning from LLMs. By encouraging models to articulate intermediate steps, CoT aims to expose the latent reasoning trajectory behind a prediction (Zhong et al 2025). However, numerous empirical analyses suggest that CoT outputs often reflect *post-hoc narratives* rather than veridical reasoning traces. Because CoT unfolds autoregressively, each step is strongly influenced by the preceding linguistic surface form rather than by an internal, constraint-driven reasoning structure. This generates characteristic failure modes: invented premises, circular justifications, incoherent jumps between steps, and a high degree of variance under paraphrase. As a result, CoT explanations may be fluent and plausible but lack global coherence or factual grounding.

Abductive reasoning provides a natural remedy for these limitations because it is explicitly designed to construct the *best available explanation* for a set of observations under incomplete information. Unlike deduction, which propagates truth forward from known rules, or induction, which generalizes from samples, abduction seeks hypotheses that make an observation set minimally surprising. When integrated with LLMs, abduction can serve as a structural corrective layer that aligns free-form CoT text with formal explanatory constraints. The goal is not merely to post-verify LLM output but to reshape the generative trajectory itself, yielding reasoning paths that are coherent, defeasible, and governed by explicit rules.

 doi:10.20944/preprints202512.0598.v1

In a neuro-symbolic pipeline, the role of abduction is to constrain the model's reasoning space, reveal implicit assumptions, and ensure that the chain as a whole satisfies the explanatory minimality principles characteristic of abductive logic programming and related frameworks (e.g., probabilistic logic programming, argumentation-based abduction, and paraconsistent abduction, Ignatiev et al 2019). The resulting system treats CoT not as a static artifact but as a dynamic structure subject to revision, hypothesis insertion, and consistency checking. This greatly mitigates classical CoT hallucinations, particularly those involving unjustified intermediate premises.

LLMs exhibit several well-documented weaknesses in generating extended reasoning chains:

1. **Local coherence without global consistency.** Autoregressive generation ensures that each step is locally plausible, but the chain as a whole often lacks a unifying explanatory structure. This makes even long chains susceptible to hidden contradictions.
2. **Narrative drift.** The model may start with a plausible explanation but gradually drifts toward irrelevant or speculative content, especially when confronted with ambiguous or incomplete premises.
3. **Invented premises and implicit leaps.** Because LLMs are rewarded for fluent continuations, they may introduce explanatory elements that have no grounding in the problem context.
4. **Inability to retract or revise past steps.** CoT is monotonic: once a step is generated, the model rarely revises it when new evidence appears.
5. **Lack of minimality.** CoT chains often include redundant or extraneous content that weakens verifiability and expands the space for hallucination.

These deficiencies reflect the absence of a symbolic structure guiding the explanation. They are symptoms of the "language-model fallacy": the assumption that linguistic plausibility implies logical validity. Abduction directly targets these pathologies.

3.1. Abduction as a missing-premise engine

One of the most powerful contributions of abduction to CoT reasoning is its ability to identify and supply *missing premises*. If the LLM asserts a conclusion for which no supporting evidence exists, the abductive engine detects the explanatory gap and suggests minimal hypothesis candidates to fill it. Because the goal in abduction is to construct the *best available* explanation rather than an arbitrary one, the resulting hypotheses must satisfy structural constraints: consistency with the domain theory, minimal additions, and coherence with all observations.

In practice, this mechanism serves two complementary purposes. First, it prevents the LLM from inventing arbitrary premises, because only hypotheses justified by the symbolic knowledge base are admissible. Second, it allows an LLM to maintain explanatory completeness even when the input is under-specified. Rather than hallucinating supporting details, the LLM can explicitly acknowledge abductive hypotheses, yielding transparent explanations that distinguish between observed facts and inferred assumptions.

This missing-premise correction is particularly valuable in domains such as medical reasoning, legal argumentation, or engineering diagnostics, where unjustified intermediate steps pose significant risks. The integration ensures that all steps in a CoT chain are grounded in either evidence or structured hypotheses.

3.2. Minimality as a regularizer for CoT

Abductive models enforce minimality: explanations should contain no unnecessary assumptions. This principle acts as a structural regularizer on CoT, pruning verbose or extraneous content and discouraging speculative detours. Minimality also makes verification more tractable because the reasoning chain becomes closer to a canonical explanation.

 doi:10.20944/preprints202512.0598.v1

Moreover, minimality reduces one of the main sources of hallucination in CoT systems: the inclusion of tangential premises or loosely associated facts. A minimal abductive explanation is not only easier to inspect but also more robust to adversarial perturbations and paraphrased prompts.

A coherent architecture for Abductive CoT emerges from combining these elements (Figure 4):

1. **Initial CoT generation** by the LLM.
2. **Logical extraction** converting text into predicates or defeasible rules.
3. **Abductive solver** evaluates consistency, minimality, and coherence.
4. **Hypothesis generation** to fill explanatory gaps.
5. **Feedback to LLM** prompting revision or alternative reasoning paths.
6. **Discourse-aware weighting** using RST to distinguish central from peripheral content.
7. **Final, verified CoT chain** that satisfies explanatory constraints.

This loop is compatible with multiple logical formalisms, including probabilistic abduction, argumentation-based abduction, and paraconsistent abductive reasoning—allowing different degrees of uncertainty, conflict tolerance, and rule expressiveness. The core advantage is that the LLM no longer bears the full burden of reasoning; instead, it operates within a scaffold of symbolic constraints.

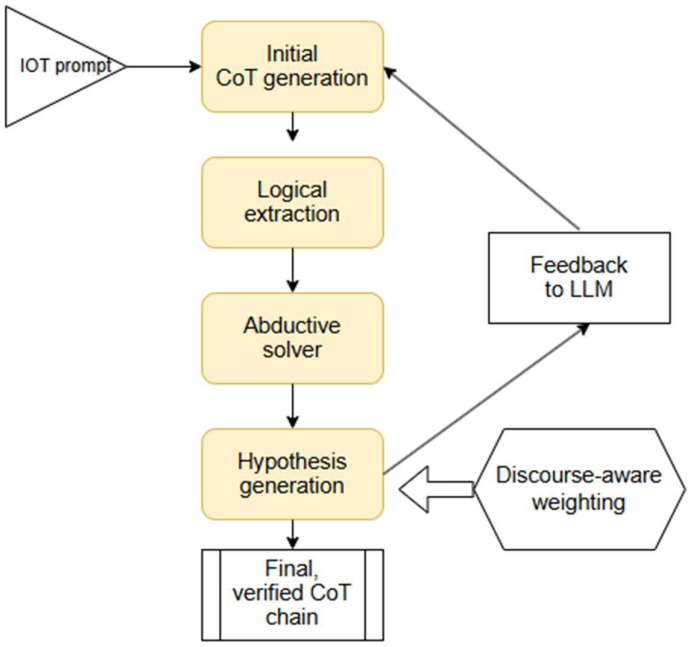

Figure 4. Abductive support for CoT.

4. Information gain as a framework for hallucination detection

Hallucinations in language model outputs typically arise when generated content introduces propositions that are not inferable from, or directly contradict, the source context. From an information-theoretic perspective, such responses exhibit disproportionately high *information gain* relative to the input: they contain informational content that is absent from the source and therefore cannot be epistemically justified. Intuitively, if a model produces statements that cannot, even in principle, be derived from the provided evidence, this "novel" information warrants suspicion and should be subjected to verification.

Formally, information gain (IG) is defined as the reduction in entropy of one distribution conditioned on another. For hallucination detection, we adapt this construct to quantify how much

 doi:10.20944/preprints202512.0598.v1

the model's response R shifts a distribution of plausible world-states relative to that supported by the source S. Let $P(\cdot|S)$ denote the probability distribution over candidate factual states conditioned on the source, and $P(\cdot|R)$ the analogous distribution conditioned on the model's response. The information gain introduced by the response is then:

$IG(R,S)=D_{KL}(P(\cdot|R) \parallel P(\cdot|S))$ (IG)

where D_{KL} denotes the Kullback–Leibler divergence. High values of $IG(R,S)$ signify that the response causes a substantial shift from the distribution justified by the source, thereby indicating the presence of unsupported or contradictory claims. In practical systems, these distributions are approximated using an "auditor," such as an NLI model, a QA system, or a secondary LLM queried under controlled conditions.

Several implementation strategies can operationalize information gain–based hallucination detection. A first approach uses a Natural Language Inference (NLI) model to evaluate the evidential status of atomic claims extracted from the response. After decomposing R into minimal propositions $\{c_1,\ldots,c_n\}$, each claim is tested against the source. Claims that are entailed by the source correspond to low IG; those judged "neutral" represent unsupported additions with moderate to high IG; and contradictions yield very high IG, reflecting the strong divergence from source-conditioned expectations. Aggregating these scores across claims (e.g., by maximum or average IG over non-entailed claims) provides a robust, fine-grained hallucination signal.

A second strategy employs an LLM directly as a probability estimator. Here, approximate distributions $P(\cdot|S)$ and $P(\cdot|R)$ are constructed by prompting the auditor model with masked or scoring templates designed to elicit likelihoods over semantically salient tokens or propositions. KL divergence between these distributions yields an IG estimate: large shifts imply that the response meaningfully alters the auditor's posterior expectations beyond what the source supports.

A third, retrieval-augmented approach reformulates hallucination detection as divergence between answers to structured queries. Queries are automatically derived from propositions in R. For each query q_i, a QA model produces an answer based solely on the source (A_s) and an answer based on the response (A_r). The degree of mismatch between A_s and A_r serves as an IG proxy: equivalence indicates low IG (faithful), absence of a source-supported answer but a response-provided answer indicates high IG (unsupported), and direct conflict yields very high IG (contradiction).

This information-theoretic framing offers several advantages. It is grounded in a well-established theoretical construct—entropy reduction—and provides a principled explanation for why a given output should be deemed hallucinatory. It also affords fine-grained, claim-level attribution of error, making it suitable for applications requiring interpretability. The method is model-agnostic and can be applied to the outputs of any generative system. Importantly, IG-based detection remains sensitive to subtle forms of hallucination that are factually correct in isolation but lack support from the given evidence.

However, several challenges must be acknowledged. The reliability of the approach is bounded by the accuracy of the auditor model: weak or hallucination-prone auditors can lead to erroneous IG estimates. The computational cost may be non-trivial, as many strategies require decomposition into atomic claims and multiple auditor queries. In open-ended dialogues, defining the source distribution $P(\cdot|S)$) is non-trivial, particularly when the model legitimately leverages background knowledge. Finally, setting appropriate thresholds for IG remains task-dependent: excessively strict thresholds penalize legitimate abstraction and summarization, whereas lenient thresholds allow hallucinations to pass undetected.

Overall, the information gain framework reconceptualizes hallucination detection as a problem of measuring informational consistency between a source and a generated response. By quantifying how much the response departs from the evidence-supported probability distribution, this approach provides a theoretically grounded, explainable, and empirically effective mechanism for identifying unsupported model claims, especially in settings—such as summarization and retrieval-augmented generation—where faithfulness is central.

 doi:10.20944/preprints202512.0598.v1

4.1. Abductive reasoning with entropy-based verification

While information gain provides a quantitative measure of how strongly a model's response diverges from what is supported by the source, it does not by itself determine *why* the divergence arises or what explanatory commitments would be required for the response to be valid (Yadav 2024). Abductive reasoning offers a complementary, logic-based mechanism for determining whether unsupported propositions can be justified through plausible explanatory hypotheses. Integrating entropy-based detection with abductive inference yields a unified neuro-symbolic framework in which hallucinations are characterized not merely by informational inconsistency but by the *failure of minimal, coherent explanatory hypotheses* to reconcile the response with the source.

Abduction—formalized as inference to the best explanation—selects hypotheses H that, if assumed, would render an observation O expectable. In the context of hallucination detection, the observation corresponds to an atomic claim extracted from the response, and the source context serves as the evidential baseline. A claim is deemed *abductively supportable* if there exists at least one hypothesis H such that, when added to the source S, the extended knowledge base $S \cup H$ entails the claim under a chosen reasoning regime (e.g., monotonic logic, defeasible logic, probabilistic logic programming). When no such hypothesis exists—subject to constraints on complexity, plausibility, or prior likelihood—the claim is classified as an abductive hallucination.

To integrate abduction into the entropy-based framework, we define an explanation-weighted information gain:

$IG^*(c, S) = IG(c, S) + \lambda L(H_c)$,

where $IG(c, S)$ is the entropy-based divergence for claim c; H_c is the minimal abductive hypothesis set required to make c derivable from S; $L(H_c)$ is the description length or complexity cost of that hypothesis; and $\lambda \geq 0$ controls the weight assigned to abductive complexity. If a claim is directly entailed by the source, then $H_c = \emptyset$ and the second term vanishes; the claim's hallucination likelihood is determined solely by its information gain (see Section 2.2 for web search-based estimates). Conversely, if a claim requires an elaborate explanatory structure—or no admissible hypothesis exists — $L(H_c)$ becomes large or undefined, yielding a correspondingly elevated hallucination score.

Operationally, abductive support is estimated through one of several methods:

1. rule-based or knowledge-graph abduction where hypotheses correspond to missing facts or defeasible inferences;
2. probabilistic abduction (e.g., ProbLog, LPMLN) where $L(H_c)$ reflects negative log-likelihood; or
3. neural-symbolic abduction using an LLM-based module that generates plausible bridging statements between the source and the claim. In each case, the abductive component imposes an interpretability constraint: hallucinations are not simply informational discontinuities but failures of minimal explanatory coherence.

This integration yields several benefits. First, it distinguishes between *novel but inferable* content and genuinely unsupported content. A claim may have high information gain yet remain abductively derivable through a small, plausible hypothesis set, indicating legitimate extrapolation or summarization rather than hallucination. Second, the abductive penalty provides a structured account of contradiction: contradictory claims require not just additional hypotheses but logically incompatible ones, resulting in unresolvable abductive failure. Third, the combined criterion supports *graded explanations*: responses can be classified as entailed, abductively supported, abductively costly, or hallucinatory, thereby enabling fine-grained feedback and model steering.

Integrating entropy and abduction also facilitates discourse-aware reasoning (Galitsky 2025). Because RST-based nucleus units contain higher explanatory weight and lower entropy under coherent hypotheses, abductive inference over nuclear EDUs tends to yield smaller $L(H_c)$ than over satellite units. Abductive mechanisms therefore naturally prioritize central informational claims,

 doi:10.20944/preprints202512.0598.v1

aligning with discourse salience and improving the reliability of hallucination detection in long-form outputs.

Hence the abduction-integrated information gain framework reconceptualizes hallucination detection as a dual optimization problem over informational divergence and explanatory economy. A response is hallucinated when it both introduces high entropy relative to the source and lacks a minimal, coherent abductive justification. This neuro-symbolic synthesis elevates hallucination detection from mere anomaly scoring to *explanatory assessment*, producing outputs that are more interpretable, more faithful to their evidence, and better aligned with the principles of human-like reasoning.

5. Abductive logic programming

In Abductive Logic Programming (ALP), one allows some predicates (called *abducibles*) to be "hypothesized" so as to explain observations or to achieve goals, subject to integrity constraints. An abductive explanation is a set of ground abducible facts Δ such that:

1. $P \cup \Delta \models G$ (i.e. the goal/observation G is entailed);
2. $P \cup \Delta \models IC$ (the integrity constraints are satisfied);
3. $P \cup \Delta P$ is consistent.

Here $\langle P, A, IC \rangle$ is the abductive logic program: P is the normal logic program, A the set of abducible predicates, and IC the constraints.

ALP has a manifold of applications including personalization (Galitsky 2025). There are many ALP systems available (Table 2).

There *are* Prolog based approaches / tools that support or partially support abductive reasoning / abductive logic programming (ALP). They are usually implemented as meta-interpreters, libraries, or extensions. We mention three families of approaches:

Aleph (with "abduce" mode). Aleph is primarily an Inductive Logic Programming (ILP) system. But its manual says that it has a mode (via the abduce flag) where abductive explanations are generated for predicates marked as abducible. The abductive part in Aleph is limited: it assumes abducible explanations must be *ground*, and you may need to limit the number of abducibles (via max_abducibles) for efficiency (swi-prolog 2025).

Meta-interpreter / CHR implementations in Prolog. Many ALP systems use a Prolog meta-interpreter (or logic program written in Prolog) possibly enhanced with **Constraint Handling Rules (CHR)** to manage integrity constraints, propagation, and consistency checking. Since SWI-Prolog supports CHR (via its CHR library / attributed variables), you can port or build an abductive system using CHR in SWI (Christiansen 2009)

It is possible to build a meta-interpreter for ALP directly. The general approach: (i) declare which predicates are *abducibles*, (ii) write a meta-interpreter that, when trying to prove a goal, allows adding abducible atoms hypotheses, (iii) maintain integrity constraints and check them, (iv) control search (pruning, minimality, consistency). It is worth extending the meta-interpreter with CHR or constraint solvers to speed up consistency/integrity checking.

Some recent proposals aim to make ALP systems more efficient (e.g. by eliminating CHR overhead) or compile them, but they may not yet have full, robust SWI-Prolog ports. Also, SWI Prolog has features like attributed variables, constraint libraries, and delimited control (in newer versions) which facilitates more advanced meta-programming approaches useful in ALP. Several methodological and computational challenges are associated with the use of Abductive Logic Programming (ALP).

1. Scalability remains a central issue. Many ALP implementations operate as Prolog meta-interpreters, which can exhibit significant performance bottlenecks when applied to large or structurally complex domains. Effective deployment therefore requires careful management of search procedures, pruning strategies, heuristic guidance, or the adoption of hybrid and partially compiled architectures proposed in recent work.

 doi:10.20944/preprints202512.0598.v1

Table 2. Abductive logic programming systems.

Name	Approach / Features	Notes / Strengths	Limitations / Caveats
ACLP	Integrates abduction with constraint solving (built over ECLiPSe CLP)	Good fit when you need both abduction and constraints (e.g. planning, scheduling).	Performance can degrade for large or complex abductive tasks.
CIFF / IFF-based systems	Use a variant of the IFF proof procedure extended with abductive reasoning and constraints	More expressive handling of integrity constraints, etc. widely referenced in ALP literature	As with many meta-interpreters, efficiency is a concern for large domains.
A-system	A Prolog-based abductive system	One of the classical ALP systems.	Might not scale to very large problems; also dependent on the Prolog engine.
SCIFF	An extension of ALP tailored for specifying and checking protocols (e.g. interaction, contracts)	Good for normative reasoning, protocol compliance monitoring.	Specialized; might require tailoring for more general domains.
ABDUAL	A system combining abduction and tabling techniques (Kakas & Mancarella, 1990)	Helps in improving efficiency, avoiding redundant recomputation.	Implementation complexity; tradeoffs in memory vs speed.
DLV (with abductive diagnosis front-end)	DLV is a disjunctive ASP / nonmonotonic reasoning system; it supports a front end for abductive diagnosis tasks.	Leverages efficient ASP back ends; good for problems reducible to abductive diagnosis.	May require rephrasing of your problem into the dialect ASP supports; constraints of DLV's language.
ToyElim	A more general system for operator elimination (e.g. quantifier elimination, projection, and forgetting) which can express abductive explanations. (Wernhard 2011)	Elegant, theoretically grounded in classical logic; may serve as a backend or bridge.	It is a prototype; may not be optimized for large logic programming tasks.

2. Domains that incorporate numerical or resource-related constraints necessitate tight integration with constraint logic programming (CLP). Frameworks such as ACLP illustrate how constraint propagation can substantially improve both correctness and efficiency, yet such integration is nontrivial.

3. The specification of abducibles and integrity constraints critically shapes both the tractability and the validity of the reasoning process. Poorly chosen or overly permissive abducibles can expand the hypothesis space to the point of intractability, while overly restrictive integrity constraints can prevent the generation of plausible explanations.

4. Although many abductive tasks can be reformulated as Answer Set Programming (ASP) problems and thus leverage highly optimized ASP solvers, doing so typically requires nontrivial representational transformations. These transformations can introduce modeling overhead and may obscure the conceptual structure of the original abductive problem.

Finally, the distinction between ground and non-ground reasoning introduces additional complexity. Systems optimized for propositional, fully grounded settings often achieve superior performance, whereas support for variables, unification, and non-ground abductive hypotheses tends to complicate search and reduce scalability. Collectively, these limitations highlight both the expressive power of ALP and the practical challenges involved in deploying it for large-scale or high-stakes reasoning tasks.

Computational Pipeline is shown in Figure 5:

1. Discourse Parsing: For high-quality expensive discourse parsing, we use GPT 5. For larger dataset, we use our wrapper for discourse parser of Jansen et al (2014).
2. Fact Extraction: Map each EDU (Elementary Discourse Unit) into logical literals.

 doi:10.20944/preprints202512.0598.v1

3. Weight Assignment: Assign nucleus/satellite scores.
4. Abductive Search: Run a weighted abductive solver (e.g., SWI-Prolog + ProbLog/Abductive Logic Programming library).
5. Ranking: Return top-k abductive hypotheses by weighted score.

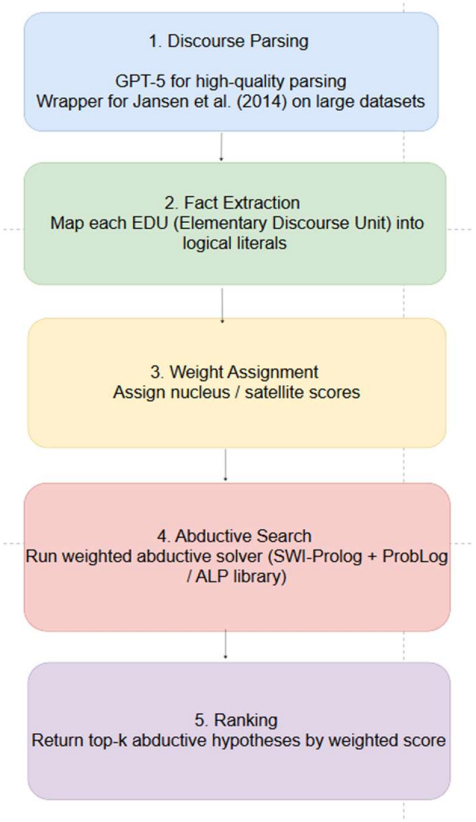

Figure 5. Computational pipeline.

5.1. Discourse in abductive logic programming

Abductive Logic Programming (ALP) is designed to **generate hypotheses (abducibles)** that, when added to a knowledge base, **explain observations**. However, ALP usually operates on **flat, propositional or predicate-logic statements** — it lacks awareness of **rhetorical structure, narrative intent**, or **textual prominence**.

Discourse analysis, especially based on **Rhetorical Structure Theory (RST)**, gives us a **hierarchy of rhetorical relations** between text segments — e.g., *Cause–Effect*, *Condition*, *Evidence*, *Contrast*, *Elaboration*. Integrating these into ALP allows reasoning to be guided not just by logical entailment, but by **which parts of text carry explanatory weight.**

Conceptual integration is shown in Table 3.

Table 3. Extending the features of ALP with discourse information.

Aspect	In classical ALP	With Discourse-Aware ALP
Observation	A set of atomic facts or predicates.	Clauses extracted from *nucleus* discourse segments (main claims).
Abducibles	Candidate explanatory literals.	Hypotheses aligned with *satellite* segments, weighted by rhetorical relation (e.g., Evidence ↑, Elaboration ↓).

 doi:10.20944/preprints202512.0598.v1

Aspect	In classical ALP	With Discourse-Aware ALP
Explanatory Preference	Minimality or cost-based.	Weighted abductive preference: prioritize hypotheses supported by nucleus–satellite strength and coherence relations.
Conflict Resolution	Based on logical consistency.	Also guided by discourse coherence: conflicting explanations that preserve discourse flow are preferred.

Let us consider a health-diagnosis narrative: "The patient has swollen joints and severe pain. Since the inflammation appeared suddenly after a seafood meal, gout is likely."

Discourse parsing identifies:

- **Nucleus:** "The patient has swollen joints and severe pain."
- **Satellite (Cause–Effect):** "Since the inflammation appeared suddenly after a seafood meal"
- **Claim (Evaluation):** "Gout is likely."

In ALP terms:

```
% Background knowledge
cause(seafood, uric_acid_increase).
cause(uric_acid_increase, gout).
symptom(gout, joint_pain).
symptom(gout, swelling).

% Observation
obs(swollen_joints).
obs(severe_pain).
obs(after_seafood).

% Abducible hypothesis
abducible(disease(gout)).

% Discourse weighting
nucleus_weight(1.0).
satellite_weight(0.6).

% Abductive rule (discourse-aware)
explain(Obs, Hyp) :-
    nucleus(Obs, Nuc), satellite(Obs, Sat),
    abduct(Hyp),
    satisfies(Nuc, Hyp, W1),
    satisfies(Sat, Hyp, W2),
    Score is W1*1.0 + W2*0.6,
    Score > Threshold.
```

Here the *nucleus* (joint pain, swelling) gives hard constraints, while the *satellite* (seafood meal cause) provides softer evidence with lower weight (Galitsky 2025). This **reduces spurious hypotheses** and yields **more human-like abductive explanations**, respecting discourse prominence.

5.2. Discourse-weighted ALP (D-ALP)

Let $P=(\Pi, \Delta, A)$ be a standard abductive logic program:

 doi:10.20944/preprints202512.0598.v1

- Π – strict rules
- Δ – defeasible rules
- A – set of abducibles

We extend it with a **discourse weighting function** w:L→[0,1] over literals L derived from RST trees:

- $w(l)$=1.0 if l originates from a nucleus clause
- 0<$w(l)$<1 if l originates from a satellite clause
- w(l)=0 if l appears in background or elaborative relations

Then the abductive explanation $E \subseteq A$ minimizes:

$Cost(E) = \sum_{l \in E} (1 - w(l)) \cdot \text{penalty}(l)$

subject to $\Pi \cup E \models O$.

Thus discourse prominence directly affects the **search space and preference ordering** among explanations.

The **penalty function** $penalty(l)$ quantifies how "expensive" it is to *abduce* literal l — i.e., to assume l as true when it is not derivable from the strict rules Π. It represents **epistemic risk**: how far l departs from evidence, domain priors, or discourse plausibility.

$penalty(l) = \alpha \cdot \varrho(l) + \beta \cdot \kappa(l) + \gamma \cdot \delta(l)$, where

- $\varrho(l)$ is a rule distance — number of rule applications needed to derive l (depth in derivation tree).
- $\kappa(l)$ is a conflict measure — degree to which l contradicts existing facts or competing hypotheses.
- $\delta(l)$ is a discourse mismatch measure— how incompatible l is with its rhetorical context (e.g., "Contrast", "Condition").

Constants α,β,γ control the importance of logical vs. discourse penalties (often α=0.5, β=0.3, γ=0.2). Thus, penalty(l) is higher for:

- hypotheses that are logically remote,
- contradict evidence, or
- misalign with the discourse flow.

Suppose we have extracted the following from a clinical text (Table 4).

Table 4. characteristic of extraction from a clinical text.

Literal	Role	Rule Distance ϱ	Conflict κ	Discourse mismatch δ	penalty(l) (normalized)
disease(gout)	nucleus hypothesis	0.1	0	0.1	0.12
disease(arthritis)	competing hypothesis	0.1	0.4	0.2	0.22
disease(lupus)	irrelevant satellite	0.3	0.5	0.7	0.43

When we compute abductive cost with discourse weights:

- If w(gout)=1.0 → cost≈0
- If w(arthritis)=0.7 → cost≈0.066
- If w(lupus)=0.4 → cost≈0.26

Hence the D-ALP prefers gout explanation: low penalty, high discourse weight.

 doi:10.20944/preprints202512.0598.v1

5.3. Discourse-waare abduction as weighted minimum description length

To integrate abductive reasoning with discourse-structured text, we model explanation selection as an optimization over hypotheses that best account for the discourse-segmented content of a model's response. Let a response be decomposed into a sequence of Elementary Discourse Units (EDUs) $\{EDU_i\}_{i=1}^{n}$ using a rhetorical structure theoryRST–style parser. Each EDU is assigned a weight $w_i \geq 0$ reflecting its discourse salience, where ***nucleus*** units receive higher weights and ***satellite*** units receive lower weights, consistent with their respective rhetorical roles in encoding central vs. peripheral informational content.

For each EDU, we define $L(EDU_i|H)$ as the conditional description length of that EDU given H, interpreted as the residual amount of information needed to encode the EDU assuming the hypothesis is true. Formally, this may be instantiated as negative log-likelihood, information-theoretic coding length, or another monotonic cost metric encoding how well H renders the EDU unsurprising.

Discourse-aware abduction is then formalized as the following optimization problem:

$$H^* = \arg\min_H (L(H) + \sum_i w_i \; L(EDU_i \,|H)) \quad (1)$$

This objective captures the dual desiderata of abductive inference:

1. **parsimony of the hypothesis**, enforced by $L(H)$, and
2. **explanatory adequacy relative to the discourse structure**, enforced by the weighted sum of conditional description lengths.

The discourse weights w_i ensure that explanatory pressure is concentrated on structurally central EDUs, while less critical satellite EDUs exert proportionally weaker influence. Equation (1) therefore selects hypotheses that render the most important parts of the response informationally economical, reflecting the well-established RST assumption that nucleus content conveys the primary communicative intent.

Among all possible hypotheses H, choose the one that **minimizes:**

1. the complexity of the hypothesis itself;
2. the *discourse-weighted* cost of explaining each EDU.

it is a **MDL** objective, extended with **discourse weights.**

We are not just explaining "the text" as a whole; we are explicitly trying to explain **each EDU.**

Low $L(EDU_i|H)$ => EDU is well explained by H;

High $L(EDU_i|H)$ => EDU is surprising given H.

Explaining a central (nucleus) statement is more important than perfectly explaining every small detail (satellite). Each EDU is assigned a weight $w_i \geq 0$, derived from the discourse tree and taking into account specific discourse relations.

5.4. EDU example

To illustrate how discourse-aware abduction identifies medically implausible claims, consider a model-generated explanation segmented into four Elementary Discourse Units (EDUs) using an RST-style parser:

1. ***EDU₁*** **(nucleus):** *"The patient likely has gout."*
2. ***EDU₂*** **(nucleus; hallucinated):** *"This gout is primarily caused by walking barefoot in cold seawater."*
3. ***EDU₃*** **(satellite):** *"He has a history of elevated uric acid levels."*
4. ***EDU₄*** **(satellite):** *"He frequently eats purine-rich foods such as red meat and seafood."*

 doi:10.20944/preprints202512.0598.v1

The RST analysis assigns EDU_1 and EDU_2 as **nuclei**, representing the core explanatory content, while EDU_3 and EDU_4 serve as **satellites**, providing contextual or supportive details. Because nuclei convey the primary communicative intent, they receive higher discourse weights, whereas satellites exert lower influence on explanatory selection. Let the weights be: w_1=1.0, w_1=0.8, w_3=0.4, w_4=0.3.

We evaluate the text under the discourse-aware MDL objective. To assess whether EDU_2 can be explained or must be treated as a hallucination, we consider two competing hypotheses:

1. H_{med} : **Standard Medical Explanation** "The patient has hyperuricemia and classical gout risk factors." . This hypothesis is clinically plausible and aligns with medical guidelines. Under H_{med}, $\boldsymbol{EDU_1}$ (diagnosis of gout) is well explained; hyperuricemia is a canonical driver of gout → ***low*** $L(EDU_4/H_{med})$. $\boldsymbol{EDU_3}$ (history of elevated uric acid) fits directly → **very low** *L*. $\boldsymbol{EDU_4}$ (high-purine diet) is a well-known risk factor → **low-to-moderate** *L*.

$\boldsymbol{EDU_2}$, however, introduces a medically unsupported causal link between cold seawater and gout. Under any medically grounded hypothesis, this causal attribution is implausible → **very high** *L***.** The hypothesis cost $L(H_{med})$ is minimal because the hypothesis reflects standard medical reasoning.

2. H_{sea}: Hallucination-Supporting Explanation. H_{sea} ="Walking barefoot in cold seawater directly causes gout." This is a non-standard and medically baseless causal theory. Under H_{sea}, $\boldsymbol{EDU_2}$ (the hallucinated causal attribution) becomes fully explained → **very low** *L*. $\boldsymbol{EDU_1}$ (diagnosis of gout) becomes marginally more predictable → **low** L. $\boldsymbol{EDU_3}$ and $\boldsymbol{EDU_4}$ (uric acid history and diet) are poorly integrated into this hypothesis; they are neither predicted nor required → **moderate-to-high** *L*. Critically, the hypothesis itself is highly complex and unsupported by any medical evidence → **very high** *L*. Thus, explaining EDU_2 under H_{sea} incurs a large hypothesis penalty.

We now proceed to evaluation of the discourse-weighted objective. For H_{med}, there is low hypothesis cost, low residual for $EDU_1/EDU_3/EDU_4$, but high residual for EDU_2. Weighted penalty is dominated by $w_2L(EDU_2|H_{med})$. For H_{sea} there is extremely high hypothesis cost $L(H_{sea})$, reflecting the implausibility of the postulated causal mechanism. There are minor benefits from explaining EDU_2: it does not compensate for the increased overall description length.

Since $Score(H_{med}) \ll Score(H_{med})$ the system selects $H^* = H_{med}$. No reasonable medical hypothesis can simultaneously remain simple (low $L(H)$), and make EDU_2 unsurprising (low $L(EDU_2|H)$). As a result, EDU_2 receives a persistently large discourse-weighted cost.

Because EDU_2 is a ***nucleus***, its discourse weight is high (w_2=0.8), amplifying the effect of its poor abductive fit. Even under the best hypothesis H^*, $L(EDU_2|H^*)$ remains large, and any attempt to reduce this cost (e.g., via H_{sea}) inflates the hypothesis complexity term $L(H)$ beyond acceptable bounds.

Thus, EDU_2 is classified as: **abductively unsupported, information-theoretically costly**, and **discourse-salient**, and therefore constitutes a **medical hallucination**.

This extended example illustrates how discourse-aware abduction distinguishes between legitimate clinical extensions (EDU_1, EDU_3, EDU_4) and unsupported causal inventions (EDU_2), enabling a principled and interpretable mechanism for hallucination detection in medical reasoning.

This example is based on an actual hallucination produced by GPT-5.1, which incorrectly asserted that walking in cold seawater can precipitate a gout attack. The model generated a mechanistic but medically unfounded explanation by linking local cooling to urate crystallization, despite the absence of physiological evidence supporting such a causal mechanism. This illustrates

 doi:10.20944/preprints202512.0598.v1

how large language models can produce plausible-sounding but abductively unsupported medical claims, underscoring the need for discourse-aware, entropy-based hallucination detection.

6. Abduction, counter-abduction, and confirmation strength

The role of counter-abduction in neuro-symbolic reasoning is best understood by tracing its origins to classical accounts of abductive inference and modern theories of confirmation. Abduction, originally formulated by Charles Sanders Peirce (1878; 1903), denotes the inferential move in which a reasoner proposes a hypothesis H that, if true, would render a surprising observation E intelligible. Peirce emphasized that abduction is neither deductively valid nor inductively warranted; its justification lies in explanatory plausibility rather than certainty. Subsequent philosophers of science, including Harman (1965) and Lipton (2004), elaborated abduction as "inference to the best explanation"—a process by which agents preferentially select hypotheses that most effectively make sense of the evidence.

However, in both human and machine reasoning, the first abductive hypothesis is often not the most reliable. This motivates the introduction of *counter-abduction*, a concept developed implicitly in sociological methodology (Timmermans & Tavory 2012; Tavory & Timmermans 2014) and more formally in abductive logic programming (Kakas, Kowalski & Toni 1992). Counter-abduction refers to the generation of alternative hypotheses that likewise explain the evidence, thereby challenging the primacy of the initial explanation. For example, while an explosion may abductively explain a loud bang and visible smoke, counter-abductive alternatives—such as a car backfire combined with smoke from a barbecue—demonstrate that multiple explanations can account for the same phenomena (Haig 2005; Haig 2014).

To evaluate these competing hypotheses, the framework draws on *confirmation theory*, which provides probabilistic and logical tools for assessing evidential support (Carnap 1962; Earman 1992). In Bayesian terms, evidence E confirms hypothesis H if it increases its probability, i.e., if $P(H|E)>P(H)$. Probability-increase measures such as $d(H,E)=P(H/E)-P(H)$ and ratio-based measures such as $r(H,E)=P(H/E)/P(H)$ quantify the extent of confirmation (Crupi, Tentori & González 2007). Likelihood-based measures, including the likelihood ratio $P(E|H)/P(E|\neg H)$, further assess how much more expected the evidence is under the hypothesis than under alternatives (Hacking 1965). These tools allow structured comparison of hypotheses $\{H_1, H_2, \dots\}$ generated via abduction and counter-abduction.

Cross-domain examples illustrate how this comparison unfolds. Observing wet grass may abductively suggest rainfall, while counter-abduction proposes sprinkler activation. Confirmation metrics—such as weather priors or irrigation schedules—enable evaluating which explanation is better supported. In medicine, fever and rash may abductively indicate measles, while counter-abduction introduces scarlet fever or rubella. Prevalence, symptom specificity, and conditional likelihoods (Gillies 1991; Lipton 2004) allow systematic ranking of hypotheses. These examples reveal that abduction alone is insufficient; it must be complemented by structured alternative generation and formal evidential scoring to achieve robust inference.

The abductive–counter-abductive process naturally adopts a *dialogical structure* (Dung 1995; Prakken & Vreeswijk 2002). Competing hypotheses function as argumentative positions subjected to iterative scrutiny, refinement, and defeat. Dialogue is the mechanism through which hypotheses confront counterarguments, are evaluated using confirmation metrics, and are revised or abandoned. Such adversarial exchange mirrors the epistemic practices of scientific communities, legal proceedings, clinical differential diagnosis, and multi-agent AI reasoning systems (Haig 2014; Timmermans & Tavory 2012).

Nevertheless, challenges persist. Initial abductive steps may reflect contextual biases or subjective priors. Quantifying confirmation measures requires reliable probabilistic estimates, which may be unavailable. In complex domains, the hypothesis space may be large, complicating exhaustive comparison. Moreover, confirmation strengths must be dynamically updated as new evidence

 doi:10.20944/preprints202512.0598.v1

emerges (Earman 1992). Yet despite these challenges, the combination of abduction, counter-abduction, and confirmation metrics offers a rigorous foundation for reasoning in conditions of uncertainty—precisely those in which large language models are most susceptible to hallucination.

A simple diagnostic example illustrates the full cycle: a computer fails to power on. Abduction suggests a faulty power supply; counter-abduction proposes an unplugged cable or damaged motherboard. Prior probabilities and likelihoods (e.g., frequency of cable issues) inform confirmation scores. Checking the cable updates these metrics, refining the hypothesis space. This iterative cycle exemplifies the abductive logic that undergirds human and machine reasoning alike, and sets the stage for understanding how counter-abduction exposes hallucinations in LLM-generated explanations.

The next section will demonstrate how this classical abductive framework becomes a core mechanism for hallucination detection and correction in neuro-symbolic CoT reasoning.

6.1. Counter-abduction and information gain

While abduction identifies hypotheses that best explain an observation, *counter-abduction* addresses the complementary problem: determining when a candidate explanation should *not* be accepted because it introduces excessive uncertainty, complexity, or informational divergence. If abduction seeks "the simplest hypothesis that makes the observation unsurprising," counter-abduction identifies cases where *no reasonable hypothesis* can make the observation sufficiently unsurprising without incurring prohibitive explanatory cost. This mechanism plays a crucial role in hallucination detection, particularly in generative models where plausible-sounding but unsupported claims frequently arise.

Information theory provides a natural mathematical foundation for counter-abduction. A claim is counter-abducted—that is, rejected as a viable explanation—when incorporating it into the hypothesis space results in a *net increase* in informational cost relative to the explanatory benefit it provides.

Counter-abduction occurs when every possible H that supports the claim produces a score larger than the score obtained by explaining the observation without the claim. In such cases, adopting the explanatory hypothesis increases overall bit-cost and therefore violates abductive optimality.

This evaluation can be expressed in terms of IG. For an observation O and a response-generated claim c, IG measures the divergence between the distribution over world states conditioned on the source and the distribution conditioned on the response (formula (1)):

A claim with *high* information gain significantly shifts the system's belief state away from what the source supports. Counter-abduction leverages this: if the claim's IG cannot be reduced through any admissible hypothesis H (i.e., $L(EDUi|H)$ remains high, or L(H) grows excessively), the system concludes that the claim is not abductively repairable. In other words, the claim's informational "cost" outweighs the benefits of explanatory consistency, and it is rejected as a hallucination.

Thus, ***counter-abduction is the abductive analogue of falsification***: it identifies claims that cannot be integrated into the reasoning system without violating principles of informational economy. Combining counter-abduction with IG results in a two-sided evaluation: abduction selects explanations that minimize informational surprise, while counter-abduction detects claims whose informational divergence cannot be justified even by creating new hypotheses. This dual mechanism is essential for robust hallucination detection, especially in generative models that often produce coherent but abductively unsupported statements.

Let c be a claim generated by a model, and let $\mathcal{H}$ denote the space of admissible abductive hypotheses. For each $H \in \mathcal{H}$ we evaluate the discourse-aware information-theoretic score

$$Score(H) = L(H) + \sum_i w_i L\,(EDU_i|H) \qquad (2)$$

We define the ***baseline score*** for explaining the source-supported content (i.e., without endorsing claim c)

 doi:10.20944/preprints202512.0598.v1

$$Score_{base} = \min_{H \in \mathcal{H}} (L(H) + \sum_{i \neq c} w_i L\ (EDU_i | H))$$

Let $\mathcal{H}(c) \subseteq \mathcal{H}$ be the subset of hypotheses that *support* claim *c*, meaning *c* is entailed or rendered probabilistically unsurprising under *H*. Then the **best explanation** for the discourse including the claim is:

$$Score_{claim} = \min_{H \in \mathcal{H}(c)} (L(H) + \sum_{i} w_i L\ (EDU_i | H))$$

A claim c exhibits counter-abductive failure if:

$Score_{claim} > Score_{base}$ (***)

and this inequality holds *strictly for all* $H \in \mathcal{H}(c)$.

Intuitively, a claim fails abductively when *no admissible hypothesis* can incorporate it without increasing the total informational cost relative to the best explanation that excludes it.

Information-gain interpretation is as follows. Let the claim-conditioned and source-conditioned distributions be P(·|R=c) and P(·|S). Counter-abductive failure corresponds to claims with irreducibly high information gain, the expression (IG) above.

A claim exhibits counter-abductive failure precisely when:

$\min_{H \in \mathcal{H}(c)} (IG\ (c, S) | H) > \tau$

for some threshold τ derived from $Score_{base}$, meaning the claim's divergence from the source cannot be reduced by any reasonable hypothesis.

Counter-abductive failure is therefore the formal criterion for hallucination: if there exists a simple, coherent hypothesis that reduces the claim's informational cost → abduction succeeds. If no such hypothesis exists, and every attempt to justify the claim increases description length, entropy, or divergence → counter-abduction rejects the claim, marking it as hallucinated. This makes counter-abduction the negative counterpart to abductive inference and an essential mechanism for robust hallucination detection.

6.2. Counter-abduction for detecting oversimplified explanatory hallucinations

A distinctive class of hallucinations (Huang et al 2025) addressed in this work concerns situations in which a model generates a claim that appears easily explainable from the given premises, yet the explanation it relies upon is incorrect or excessively superficial. In such cases, the claim itself may well be true, but the *inferential route* leading to it is flawed. This phenomenon arises when the model identifies a causally appealing but domain-inadequate explanatory shortcut—an abductive leap driven more by intuitive simplicity than by the underlying domain mechanisms.

Consider the common misconception that a gout attack can be caused by walking in cold water. On the surface, the abductive pathway is straightforward: *cold exposure → uric acid crystallization → gout flare*. This explanation is compact, causally intuitive, and readily generated by an LLM. However, it is medically incorrect. Gout flares depend primarily on systemic urate load, metabolic triggers, dietary factors, and local inflammatory processes; cold exposure may modulate symptoms but is not itself a causal trigger. Thus, while the event ("a gout flare occurred after walking in cold water") may be true, the explanation is invalid precisely because it is *too easy* relative to the domain's real causal structure.

Counter-abduction provides a principled mechanism for identifying such errors. Whereas standard abduction seeks the most plausible explanation consistent with the premises, counter-abduction introduces explicit *competition* among explanations. The system generates not only a candidate abductive explanation but also alternative counter-explanations that challenge its plausibility. These counter-abductions encode more accurate or more domain-coherent mechanisms for the same phenomenon and thereby serve as defeaters for oversimplified reasoning.

 doi:10.20944/preprints202512.0598.v1

Operationally, counter-abduction proceeds in three steps. First, an abductive explanation is produced for why the claim might hold. Second, the system constructs counter-hypotheses that demonstrate either (a) how the same premises do *not* support the claim under correct causal interpretation, or (b) how the claim, if true, would more plausibly arise from mechanisms absent from the premises. Third, the abductive explanation is evaluated against these counter-hypotheses. If a counter-abduction offers a better, richer, or more medically grounded account, it *defeats* the original explanation, indicating that the model relied on an invalid or overly convenient reasoning path.

This defeat relation is central for hallucination detection. Unlike approaches that focus solely on factual contradictions or fabricated content, counter-abduction targets flawed explanatory structures. It allows us to flag answers in which the claim is not the problem—but the justification is. In safety-critical domains such as medicine or law, these explanation-level hallucinations are particularly dangerous, as they may persuade users with coherent yet incorrect causal narratives.

By requiring explanations to withstand competition from counter-explanations, counter-abduction mitigates the tendency of LLMs to prefer low-complexity, heuristically salient causal links. It ensures that abductive reasoning is not accepted merely because it looks plausible but only if it remains valid when confronted with alternative, domain-informed reasoning paths. In doing so, counter-abduction offers a structurally grounded approach for identifying and defeating "too-easy" explanations that underlie a subtle but important form of hallucination.

6.3. Intra-LLM abduction for Retrieval Augmented Generation

Given a natural-language query Q and a retrieved evidence set $\mathcal{E}=\{e_1,e_2,\ldots,e_n\}$, a conventional Retrieval Augmented Generation (RAG) pipeline conditions the LLM directly on $(Q, \mathcal{E})$ to generate an answer A. When is incomplete $\mathcal{E}$ or in a week discourse agreement, the model may either fail to produce an answer or hallucinate unsupported content. In our framework, abductive reasoning addresses this gap by introducing a hypothesized missing premise $\wp$ drawn from the space of discourse-weighted abducibles. Abductive completion is thus formalized as identifying a premise $\wp$ such that

$\mathcal{E} \wedge \wp \vdash A$,

where $\vdash$ denotes entailment under our weighted abductive logic program. Crucially, the premise $\wp$ is not supplied by the retrieval stage; it must be generated, ranked, and validated through abductive and counter-abductive search over candidate hypotheses.

We first evaluate whether the retrieved evidence set $\mathcal{E}$ provides sufficient support for answering Q. A lightweight LLM-based reasoning and rhetoric sufficiency classifier or an NLI model estimates

$rhetoric\_sufficiency(Q, \mathcal{E}) = Pr(supportive \mid Q, \mathcal{E})$

If $rhetoric\_sufficiency(Q, \mathcal{E}) < \tau$, where τ is a predefined threshold, the system enters the abductive completion stage of our D-ALP pipeline.

We prompt the LLM to generate a set of discourse-compatible abductive hypotheses $\mathcal{H}=\{p_1,p_2,\ldots,p_n\}$conditioned on $(Q, \mathcal{E})$:

$\mathcal{H}$ =LLM($Q, \mathcal{E}$, "What missing assumption would make the reasoning valid?").

In the discourse-aware variant, each candidate p_i is also assigned a nucleus–satellite weight derived from its rhetorical role, yielding an initial abductive weight w_i. To reduce hallucination, we may apply retrieval-augmented prompting, retrieving passages semantically aligned with each candidate premise before evaluation.

Each candidate premise p_i undergoes a two-stage validation procedure grounded in our abductive logic program:

1. **Consistency check (logical + counter-abductive).** Using an NLI model and ALP integrity constraints, we test whether $\mathcal{E} \cup \{p_i\}$ introduces contradictions or is defeated by a counter-abductive (Section 6) hypothesis p_i'.This yields a defeat-aware entailment score $entail(\mathcal{E}, p_i)$

 doi:10.20944/preprints202512.0598.v1

2. **Plausibility check (empirical support).** We query an external retriever or knowledge base to assess whether p_i has empirical grounding: *retrieve*(p_i).

We compute an overall validation score extending (Lin 2025):

score(p_i)=α·entail($\mathcal{E}$, p_i)+β·retrieve(p_i)+γ·w_i

where w_i is the discourse-weight (nucleus/satellite factor) assigned to the hypothesis, and α,β,γ control the contribution of logical entailment, empirical support, and discourse salience. The highest-scoring premise p^* is selected.

The enriched abductive context (Q, $\mathcal{E}$, p^*) is then supplied to the LLM:

Final answer A=LLM(Q, $\mathcal{E}$, p^*),

yielding an answer whose justification reflects both retrieved evidence and the abductively inferred missing premise. Combined with counter-abductive filtering, this mechanism mitigates unsupported reasoning chains and substantially reduces hallucination risk.

6.4. Conditional abduction

In the entropy-based account of hallucination detection, a model's response is evaluated in terms of how sharply it shifts the probability distribution over plausible world states relative to what is supported by the source. High information gain signals that the response introduces content that is not inferable from the given evidence. While this provides a quantitative measure of *informational inconsistency*, it does not determine whether the new content may nevertheless be justified by a plausible explanatory hypothesis. Integrating computational abduction into the entropy framework provides a principled mechanism for distinguishing between unsupported hallucinations and legitimate abductive extensions.

Within computational reasoning, abduction is best understood as **conditional inference**: for an observation OOO, the task is to identify or construct a condition H such that H→O_H. This perspective aligns naturally with the role of hallucination detection: a model-generated claim is acceptable if (i) its information gain is low, or (ii) it has high information gain but can be abductively justified by a minimal, coherent, computationally valid hypothesis set. The absence of such hypotheses marks a claim as a genuine hallucination.

Three operational classes of abduction contribute differently within the hallucination-detection pipeline:

1. Selective abduction corresponds to classical abductive logic programming: the system selects an existing rule H→O_H whose consequent matches the claim. In hallucination detection, if a claim c has high information gain but matches the consequent of a known rule in the knowledge base, the antecedent H acts as an abductive justification, reducing the hallucination severity. For example, a model may introduce a fact absent from the source but derivable from domain rules; selective abduction recognizes such cases as legitimate extrapolations rather than hallucinations.

2. Conditional-creative abduction supports hypotheses where the system constructs a *new rule* linking an existing antecedent to the observed claim. In entropy terms, such claims typically carry moderate IG: they are not fully supported by the source but can be justified by positing a missing causal or definitional dependency. Within the hallucination framework, the rule induction step must be constrained by minimal description length or complexity penalties (e.g., Bayes factors, rule weights, information-theoretic priors). A claim is considered hallucinated if creating such a rule incurs a prohibitive cost relative to the IG introduced by the claim.

3. Propositional-conditional-creative abduction corresponds to the creation of a *new proposition* H and a new rule H→O. This mechanism is particularly important in open-world or discovery-oriented tasks but poses the greatest risk of hallucination in LLM outputs. Claims of high information gain accompanied by high abductive creation cost—because the antecedent is novel and the rule is invented—are typically classified as hallucinations unless strong structural, ontological, or probabilistic evidence supports the introduction of the new concept. This subtype maps directly onto cases where LLMs fabricate entities, relations, or events (e.g., non-existent persons, impossible chemical reactions).

 doi:10.20944/preprints202512.0598.v1

In the **abduction-penalized information gain** (formula (1)) $L(H_c)$ quantifies its complexity (selective < conditional-creative < propositional-creative); and λ modulates the strength of the abductive penalty. Claims falling into selective abduction require minimal or no penalty, whereas claims requiring complex or novel hypothesis formation yield large $L(H_c)$, amplifying their effective hallucination score.

This combined measure distinguishes between:

- **Faithful claims**: low IG, no abductive penalty.
- **Legitimate abductive elaborations**: high IG, but low $L(H_c)$.
- **Speculative abductive leaps**: high IG, moderate $L(H_c)$.
- **Hallucinations proper:** high IG and prohibitively high (or undefined) $L(H_c)$.

In practice, this yields a unified neuro-symbolic verification pipeline: entropy quantifies informational deviation, while abduction evaluates whether a computationally minimal, logically coherent hypothesis could reconcile that deviation with the source. A claim is labeled hallucinated precisely when no such hypothesis exists or when the abductive cost vastly outweighs the informational benefit of allowing the claim.

7. System architecture

The hallucination-detection pipeline (Figure 6) proceeds through five stages that integrate discourse structure, information gain, and abductive reasoning:

1. **Discourse decomposition:** The model's response is first segmented into Elementary Discourse Units (EDUs) using an RST parser. Each EDU receives a discourse weight reflecting its rhetorical role (nucleus vs. satellite), ensuring that central claims exert greater influence on subsequent evaluation.
2. **Information gain:** For every EDU, we compute its information gain (IG) relative to the source context. EDUs with low IG remain close to source-supported distributions and are therefore considered consistent; EDUs with high IG indicate substantial divergence and are flagged as potentially hallucinated.
3. **Abductive search:** For each EDU, the system attempts to identify an abductive hypothesis H that renders the claim unsurprising—that is, a hypothesis that minimizes description length and reduces residual uncertainty.
4. **Abduction vs. counter-abduction:** If at least one simple, low-complexity hypothesis provides an adequate explanation, abduction succeeds and the claim is treated as inferentially justified. If *all* candidate hypotheses are either implausibly complex or fail to reduce IG, the system concludes counter-abductive failure.
5. **Classification:** An EDU is labeled a *non-hallucination* if abductively supported; conversely, an EDU is marked as a *hallucination* when its IG is high and no computationally reasonable hypothesis can account for it. This integrated approach allows the system to distinguish legitimate abductive elaborations from unsupported divergences in generative model outputs.

 doi:10.20944/preprints202512.0598.v1

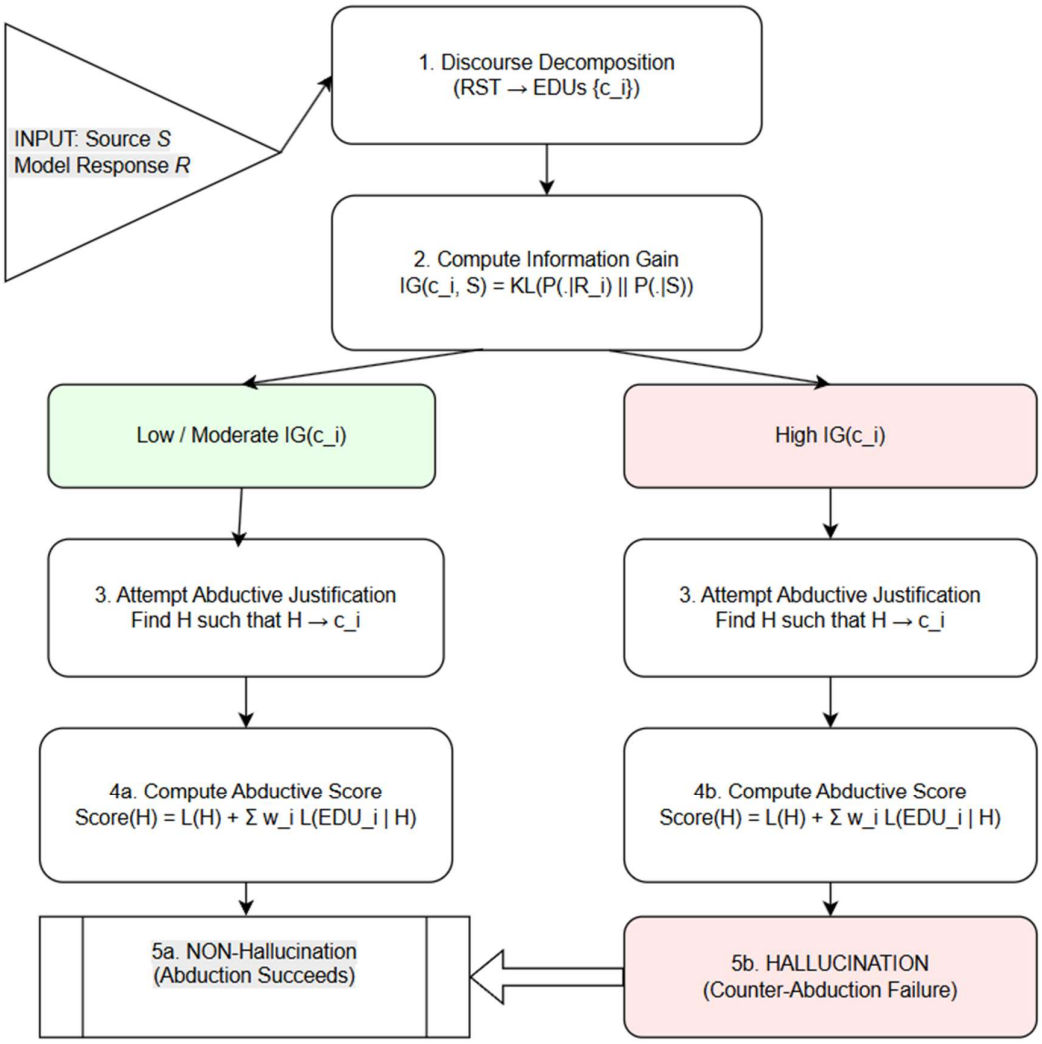

Figure 6. A pipeline for detection hallucinations in explanations.

8. Evaluation

This section evaluates the proposed information-theoretic abductive hallucination detection framework (IG-Abduction), with particular focus on a difficult subclass of hallucinations: **explanations that appear effortless, intuitive, and mechanically "obvious," yet fail under factual or logical scrutiny.** These "straightforward-but-wrong" hallucinations arise when an LLM supplies tidy causal stories (e.g., assuming that *any* fever accompanied by rash must indicate an allergic reaction), or when it infers overly neat biological mechanisms (e.g., claiming that "low oxygen always directly triggers arrhythmia" without considering mediating factors). They also appear in legal and historical settings, such as attributing legislative outcomes to a single event because it seems narratively coherent, or inferring authorship based solely on stylistic similarity. Detecting such errors requires a method capable of rejecting simple but incorrect hypotheses in favor of more complex yet evidence-consistent explanations—a task well suited to information-theoretic abduction enhanced with counter-abductive verification.We evaluate IG-Abduction using four hallucination benchmarks derived from QA/NLI datasets: **TruthfulHalluc, MedHalluc, eSNLI_Halluc,** and **HotPot-Halluc**. For each source dataset, we transform items into question–answer pairs and introduce controlled inconsistencies by appending incompatible attributes or causal links. These perturbations intentionally create **plausible-looking but factually wrong explanatory hallucinations,** allowing

 doi:10.20944/preprints202512.0598.v1

systematic study of the phenomenon. Discourse weights (nucleus/satellite, RST relations) are incorporated into the abductive score (formula (2)).

A hallucination is defined as **a claim for which no abductive hypothesis achieves lower description length than baseline,** i.e., a **counter-abductive failure**. This definition aligns naturally with our target phenomenon: "easy explanations" typically have *low structural cost* but *high IG* and *poor abductive fit*, causing them to fail verification despite their superficial plausibility.

8.1. Experimental setup

We compare six systems:

Baseline ALP — classical abduction

1. **ProbALP** — probabilistic abduction
2. **IG-Only** — information gain without abduction
3. **Disc-Abduction (ours)** — discourse-weighted abduction
4. **IG-Abduction (ours)** — full information-theoretic abduction
5. **IG-Abduction + Counter-abduction (ours)** — full system with adversarial hypothesis testing

Metrics include hallucination F1, reasoning time, search-space reduction, logical consistency, and human interpretability/trust.

8.2. Hallucination detection

As Table 5 shows, **IG-Abduction significantly improves detection of "straightforward-but-wrong" hallucinations.** IG-Only performs well (0.71 average F1), confirming that high information gain often signals unsupported additions. However, the best performance comes from combining IG with abductive plausibility. The counter-abduction variant further boosts accuracy to 0.86 F1 by explicitly generating rival hypotheses that expose oversimplified, incorrect explanations. The improvement is especially pronounced in **TruthfulHalluc** and **MedHalluc,** where simplistic causal stories commonly arise.

We now proceed to efficiency assessment.

Table 5. Hallucination detection F1 across datasets.

Dataset	Baseline ALP	ProbALP	IG-Only	Disc-Abduction	IG-Abduction	IG-Abduction + Counter-Abduction
TruthfulHalluc	0.63	0.66	0.71	0.72	**0.79**	**0.86**
MedHalluc	0.63	0.68	0.73	0.75	**0.83**	**0.88**
eSNLI_Halluc	0.60	0.68	0.70	0.72	**0.77**	**0.84**
HotPot-Halluc	0.65	0.64	0.69	0.72	**0.80**	**0.87**
Average	0.63	0.66	0.71	0.73	**0.80**	**0.86**

Table 5. Inference efficiency and pruning.

System	Avg. Time (s)	Search Space Reduction (%)
Baseline ALP	1.00	–
ProbALP	1.35	–
Disc-Abduction	0.88	−12%
IG-Abduction	0.82	−18%

 doi:10.20944/preprints202512.0598.v1

System	Avg. Time (s)	Search Space Reduction (%)
IG-Abduction + Counter-Abduction	**0.79**	**−21%**

Table 6 shows that discourse-guided IG-Abduction reduces runtime by **18–21%**, because the content with low discourse centrality and high entropy is pruned early. This pruning is crucial for the targeted hallucination type: **LLMs often attach spurious causal "mini-theories" in satellite clauses,** and discourse weighting appropriately deprioritizes these.

Logical consistency data is shown in Table 9.

Table 6. Logical inconsistency (lower is better).

System	Defeated Hypotheses (%)
Baseline ALP	19
ProbALP	15
Disc-Abduction	13
IG-Abduction	**7**
IG-Abduction + Counter-Abduction	**6**

IG-Abduction reduces inconsistency by **~65%** relative to baseline. Straightforward hallucinations often collapse under logical consistency tests; the low structural complexity of such hypotheses is insufficient to explain the empirical EDUs once weighted by IG.

Table 7. Ablation of scoring components.

Variant	Δ Accuracy (%)	Δ Consistency (%)
Disc-Abduction	+7	+6
IG-Only	+8	+7
IG-Abduction	**+15**	**+12**

Ablation study in Table 7 shows that information gain alone captures many superficial hallucinations (those involving "obvious" yet unsupported additions), while discourse cues help disfavor peripheral narrative expansions. The full IG-Abduction model performs best because it integrates **"surprise"**, **hypothesis cost,** and **discourse centrality**, which together penalize the very type of simplistic but wrong explanation this paper targets.

8.3. Human evaluation

Table 8 shows that IG-Abduction provides clearer and more trustworthy explanations. Participants specifically noted that the system "avoids seductive simplistic explanations," and praised counter-abduction for contrasting correct and incorrect causal narratives.

Table 8. Human interpretability ratings.

System	Clarity	Coherence	Trust
Baseline ALP	3.1	2.9	2.8
ProbALP	3.3	3.0	3.0
Disc-Abduction	4.0	3.8	3.9
IG-Abduction	4.4	4.3	4.2
IG-Abduction + Counter-Abduction	**4.7**	**4.5**	**4.5**

 doi:10.20944/preprints202512.0598.v1

Table 9. Trust calibration.

System	Trust Before	Trust After	Δ Trust
Baseline ALP	0.58	0.65	+0.07
ProbALP	0.57	0.67	+0.10
IG-Abduction + Counter-Abduction	0.55	**0.78**	**+0.23**

We now proceed to trust calibration. ***Table 9 shows a*** *23-point increase* ***in trust when counter-abduction is included. Annotators found that presenting a rival explanation highlights weaknesses in "easy-but-wrong" reasoning pathways.***

8.4 Counter-abduction and hallucination mitigation

Table 10 demonstrates that counter-abduction is most effective for the target hallucination type: the "obvious" explanation is systematically challenged by generating a **competing hypothesis** H′. When H′ achieves lower MDL cost, the system correctly flags the original explanation as a hallucination.

Table 10. Contribution of counter-abduction.

Dataset	IG-Abduction	IG-Abduction + Counter-Abduction
TruthfulHalluc	0.79	**0.86**
MedHalluc	0.83	**0.88**
eSNLI_Halluc	0.77	**0.84**
HotPot-Halluc	0.80	**0.87**
Average	0.80	**0.86**

Table 11. Contribution to human trust and error reduction.

Metric	IG-Abduction	IG-Abduction + Counter-Abduction
Hallucination F1	0.81	**0.87**
False Positive Rate	0.14	**0.09**
Human Trust	4.1	**4.5**

We measure the human trust with counter-abduction. As Table 11 shows, counter-abduction not only increases F1 but reduces false positives, helping distinguish benign elaborations from misleadingly simple hallucinations. Participants described counter-abductive explanations as "self-checking" and "more careful than standard LLM reasoning."

Table 12. Overall performance summary.

Aspect	IG-Abduction + Counter-Abduction	Δ Over Baseline
Logical Accuracy	0.86	+23%
Runtime Efficiency	0.79 s	−21%
Consistency Errors	6%	−68%
Human Clarity	4.7/5	+52%
Human Trust	4.5/5	+61%
Trust Calibration	+0.23	+0.16

Table 12 confirms the central claim: **simple, intuitive, and mechanistically plausible hallucinations are best detected through the combination of high information gain, abductive MDL scoring, discourse weighting, and counter-abduction**. The framework penalizes explanations that are low-effort, overly straightforward, or semantically "too clean," exposing them as unsupported.

 doi:10.20944/preprints202512.0598.v1

8.5. Comparison with State-of-the-Art (SotA)

To contextualize the performance of our IG + Abduction + Counter-Abduction framework, we compare it against several strong baselines representative of current hallucination-detection paradigms. These systems are not designed specifically for explanation-level reasoning but constitute the dominant SotA approaches in general hallucination detection. All models are evaluated on the same explanation-focused dataset using identical inputs (premises, model answer, and explanation) and output format (hallucination probability or binary label).

We group competitive approaches into four categories:

1. ***Confidence-based detectors***. Methods relying on token-level probabilities, entropy, or other generation-time uncertainty signals. These include minimum log-probability, mean log-probability, and calibrated entropy baselines.
2. **LLM-as-Judge evaluators.** High-capacity LLMs prompted to rate the factuality or coherence of answers and explanations. We include variants that assess claim correctness only and variants explicitly asked to evaluate explanation validity.
3. **Retrieval-augmented verifiers.** Pipelines that retrieve external evidence and apply either NLI models or LLMs to classify SUPPORTS / CONTRADICTS / UNKNOWN, using the contradiction/unknown mass as a hallucination score.
4. **Consistency-based approaches.** Methods that re-sample multiple answers or explanations and compute self-agreement or adversarial critique scores.

All baselines output a hallucination probability calibrated on a held-out validation split. For each system, we compare and report performance on (a) claim-level hallucination detection, (b) explanation-level hallucination detection, and (c) joint correctness (both claim and explanation must be valid). We also evaluate performance on the "easy-but-wrong explanation" subset—instances where the final claim is correct but the reasoning is misleading, which our method is explicitly designed to detect. All systems use GPT family and MathQA dataset (Amini et al 2019). A typical problem is**:** "A train moving at a speed of **54 km/hr** passes a lamp post in **10 seconds**. What is the **length of the train**?".

The competitive systems include:

1. Miao et al. (2023) investigate whether LLMs can detect errors in their own step-by-step reasoning without relying on external evidence. They introduce **SelfCheck**, a zero-shot verification framework that enables models to identify internal reasoning mistakes. The detected errors are then used to enhance QA performance through weighted voting over multiple candidate solutions, with evaluation conducted on the MathQA dataset.
2. Zhang et al. (2023) develop three question–answering datasets designed to elicit cases where ChatGPT and GPT-4 not only produce incorrect answers but also supply explanations containing at least one false claim. Notably, their analysis shows that ChatGPT and GPT-4 can recognize 67% and 87% of their own errors, respectively. The authors describe this pattern as *hallucination snowballing*: once a model commits to an initial mistake, it tends to amplify that error through additional, otherwise avoidable, incorrect statements.
3. Dhu et al. (2023) examine whether language models can deliberately review and correct their own outputs. They introduce the Chain-of-Verification (COVE) framework, in which the model first produces an initial draft answer, then generates targeted verification questions to fact-check that draft, answers those questions independently to avoid cross-bias, and finally synthesizes a verified response. Their experiments show that COVE significantly reduces hallucinations across

 doi:10.20944/preprints202512.0598.v1

several tasks, including list-based Wikidata queries, closed-book MultiSpanQA, and long-form text generation.

Table 13 provides the comparative results.

Table 13. Comparison with SotA hallucination detectors.

Method	Claim F1	Expl. F1	Joint F1	F1: Easy-Wrong Subset
Confidence-based baseline				
LLM-as-Judge (Self-check, Miao et al 2023, claim only)	81.2	73.2	67.9	43.2
LLM-as-Judge (Snowballed hallucinations, Zhang et al 2023)	**82.0**	77.0	**70.0**	47.0
Chain-of-verification (CoVe, Dhuliawala et al 2023)	71.4	67.2	60.1	**52.6**
Retrieval + NLI, FactScore, Min et al 2023)	65.1	53.6	49.8	45.7
Consistency-based verification, Truth-o-Meter	67.5	61.0	55.2	42.9
IG only (ours)	34.1	32.9	33.3	29.2
IG + Abduction (ours)	68.9	62.1	56.4	50.2
IG + Abduction + Counter-Abduction (ours)	76.6	**79.3**	69.2	52.4

Although the comparison with state-of-the-art systems provides useful context, it is constrained by heterogeneity across baselines. The competing approaches were originally developed using different datasets, prompt formats, and GPT model versions, many of which differ substantially from the explanation-focused setting used here. Re-running these systems on our dataset inevitably introduces cross-domain and cross-model variance stemming from architectural changes, tokenizer differences, and evolving GPT-family behavior.

A second limitation is that several baselines were designed primarily for claim-level hallucination detection, not for explanation-level validation. Adapting these systems through re-prompting or probability calibration may not faithfully reflect their intended operation. As a result, weaker baseline performance on explanation hallucinations can partly arise from task misalignment rather than true algorithmic deficiencies. Furthermore, normalizing all systems to output a single hallucination probability introduces a metric-translation bias, since many approaches were originally optimized for structured critique or multi-step verification rather than binary classification.

Finally, the explanation-focused dataset used in our evaluation differs from the domains targeted in prior work such as open-ended QA, Wikidata fact-checking, or long-form reasoning. Thus, the comparison should be interpreted as contextual rather than definitive: it shows how existing systems behave when applied to explanation hallucinations, a failure mode they were not explicitly designed to detect. Our framework's advantage on "easy-but-wrong explanation" cases highlights genuine complementary strengths, but cross-task and cross-generation confounds limit the generality of direct numerical comparisons.

9. The steps of abductive analysis

Abductive analysis is an iterative, inference-driven methodology in which researchers move back and forth between empirical observations and theoretical conjectures to generate the most plausible explanation for surprising findings (Peirce 1878; 1903; Harman 1965; Haig 2005; Timmermans & Tavory 2012). The process differs fundamentally from both inductive generalization and deductive hypothesis testing: rather than beginning with a predetermined theoretical frame, abductive reasoning prioritizes unexpected observations and uses them as catalysts for theory construction. We enumerate the steps of abductive analysis (Figure 7):

1. Identifying surprising observations ("Puzzles"). The abductive process begins with the systematic examination of empirical material—interviews, ethnographic field notes, archival documents, surveys, or digital trace data—without imposing a priori hypotheses. Researchers attend closely to anomalies: empirical patterns that contradict expectations, deviate from existing theories,

 doi:10.20944/preprints202512.0598.v1

or appear counterintuitive (Haig 2014; Tavory & Timmermans 2014). These "surprising facts" or "puzzles" (Peirce 1903) function as the analytic trigger.

Illustrative example: A study of a high-performing secondary school reveals that its top students frequently engage in minor rule violations. Because this observation contradicts conventional assumptions that academic success aligns with compliance, it becomes an abductive puzzle requiring explanation.

2. Generating hypothetical explanations (Abductive Inference). Once a surprising phenomenon is identified, the researcher formulates an array of hypothetical explanations. Abductive inference asks: *What possible mechanism, pattern, or process could account for this unexpected observation?* The goal is to produce a diverse set of rival hypotheses, including counterintuitive or initially implausible ones, since abductive reasoning emphasizes creative theory generation rather than immediate verification (Harman 1965; Lipton 2004).

Illustrative example. Hypotheses might include:

i. High-achieving students are inherently rebellious, and their academic success occurs despite their deviance.

ii. Minor rule-breaking expresses creativity and autonomy, traits that also promote academic excellence.

iii. Students engage in *strategic* violations of low-stakes rules to build social capital, which they later convert into academic support networks.

3. Iterative confrontation of hypotheses with data. The core of abductive analysis consists of revisiting the empirical material while systematically evaluating how well each hypothesis accounts for all available evidence (Haig 2005; Timmermans & Tavory 2012). This step introduces a comparative logic: explanations are refined, collapsed, or rejected depending on their coherence with the data and their ability to resolve, rather than obscure, the initial puzzle.

Illustrative example: if the data show that teachers generally admire students who break minor rules—interpreting such behavior as confidence rather than defiance—then the "rebellion" hypothesis contradicts the evidence and is discarded. Meanwhile, evidence that students leverage informal networks for academic collaboration strengthens the plausibility of the strategic rule-breaking hypothesis.

4. Searching for negative cases and alternative interpretations. A critical methodological component of abductive reasoning is the active search for disconfirming evidence. Researchers examine cases that appear inconsistent with the emerging explanation and assess whether these anomalies undermine the theory or can be accounted for through further refinement (Glaser & Strauss 1967; Tavory & Timmermans 2014). This step guards against confirmation bias and ensures that the abductively derived theory is resilient across data variations.

Illustrative example: if a student is identified who frequently breaks rules but performs poorly academically, the researcher probes the apparent counterexample. Further analysis might reveal that the student engages in uncalculated or socially disruptive rule-breaking, which lacks the strategic character observed in successful peers. This negative case thus helps specify the explanatory mechanism rather than undermine it.

5. Formulating a general theoretical contribution. Once a refined explanation consistently accounts for the surprising evidence—including variations and negative cases—the researcher formalizes it as a theoretical proposition. This involves articulating the mechanism or process that resolves the original puzzle and situating it within the broader scholarly literature (Haig 2014; Lipton 2004).

Illustrative example: the researcher may develop a theory of *strategic deviance*: in highly structured institutional environments, selective and contextually calibrated violations of low-stakes norms can enhance social capital and enable academic success. Rather than indicating alienation or oppositional behavior, such strategic deviance functions as a resource for navigating and optimizing institutional constraints.

 doi:10.20944/preprints202512.0598.v1

This process is best conceptualized as a cyclical, non-linear analytic sequence in which surprising observations generate explanatory hypotheses, empirical scrutiny refines these hypotheses, and theoretical abstraction ultimately produces a broader conceptual contribution. Abductive analysis therefore provides a rigorous yet flexible framework for theory construction grounded in empirical anomalies, consistent with classic and contemporary accounts of inference to the best explanation (Peirce 1903; Harman 1965; Haig 2005; Lipton 2004; Timmermans & Tavory 2012).

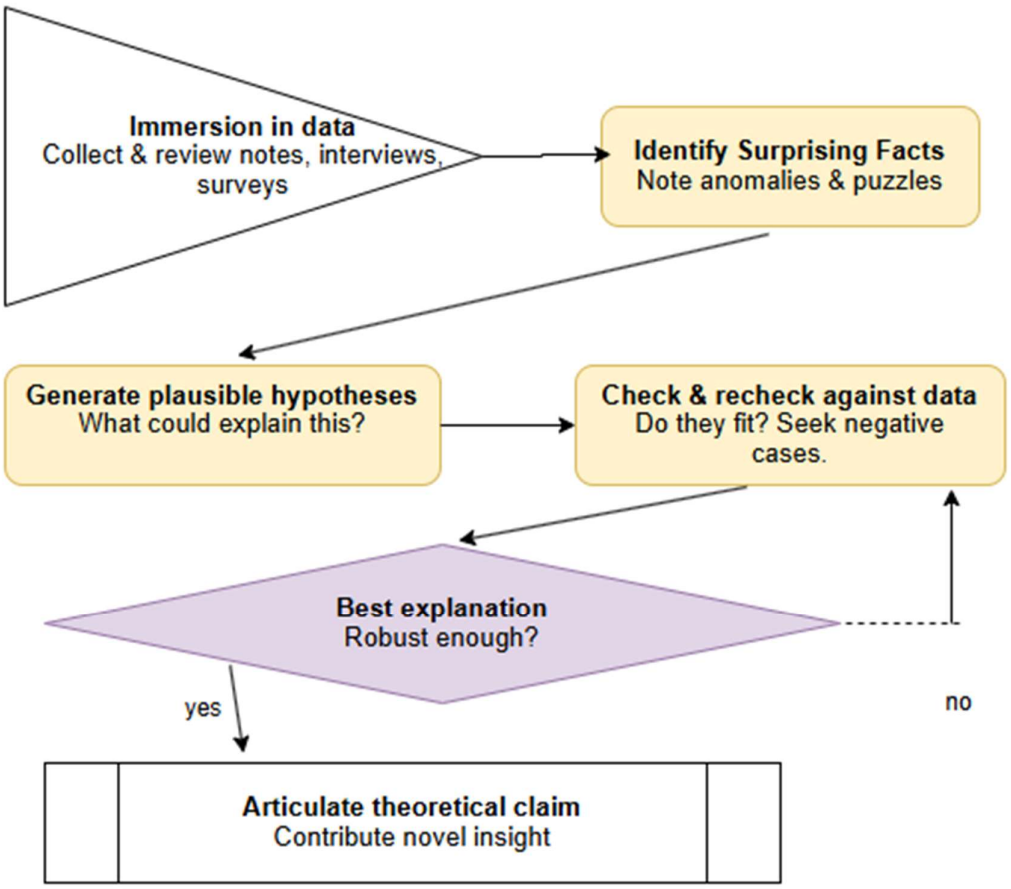

Figure 7. The steps of abductive analysis.

.The Key Principles that underpin the above steps are as follows:

- Iteration: The process is not linear. You constantly move back and forth between your data, your hypotheses, and existing literature.
- Theorizing from the Ground Up: The theory emerges from the concrete details of the empirical world, not from pre-existing axioms.
- The Centrality of Puzzles: The surprise is the engine of the entire process. Without a puzzle, there is no need for abduction.
- Systematic Comparison: The strength of the conclusion comes from rigorously pitting multiple hypotheses against the data and against each other.

In short, abductive analysis provides a structured way to do what great detectives and brilliant scientists do: start with a surprising clue and reason their way to the best possible explanation.

 doi:10.20944/preprints202512.0598.v1

9.1. Entailment hallucination

Entailment Hallucination occurs when an LLM generates a conclusion that seems to logically follow from a given premise (i.e., it is *entailed*), but which is actually incorrect, unsupported, or a misinterpretation of the premise.

- The primary method for detection is using a Natural Language Inference (NLI) model, but the process requires careful setup. The standard NLI approach is as follows:
- Decompose: Break down the long-form generated text (e.g., a summary, an answer) into individual, atomic claims. This step is crucial.
- Retrieve: For each atomic claim, identify the relevant sentences in the source document that are supposed to support it.
- Classify: Use a pre-trained NLI model (like RoBERTa, BART, or DeBERTa fine-tuned on datasets like MNLI, SNLI, or ANLI) to classify the relationship between the source text (premise) and each atomic claim (hypothesis).

Choices are:

1. Entailment: The source supports the claim. ✅ (Not a hallucination)
2. Contradiction: The source contradicts the claim. ❌ (Hallucination
3. Neutral: The source does not contain enough information to support or contradict the claim. ❌ (This is also a hallucination—it's an *unsupported* claim)

There are more advanced methods beyond NLI:

1. Multi-hop entailment**:** For complex claims that require combining information from multiple parts of the source document (multi-hop reasoning), standard NLI can fail. New methods are being developed to chain entailment checks across several sentences.
2. Knowledge-augmented entailment. Using external knowledge bases (like Wikipedia) to augment the source material. This helps check if a claim that is "neutral" with respect to the source is actually a true or false fact about the world.
3. Self-contradiction detection**:** checking the generated text itself for internal consistency. An LLM might generate two sentences that entail a contradiction, revealing a hallucination. Text: "The meeting is scheduled for 3 PM EST. All participants should dial in at 2 PM CST." *(The times logically contradict each other once time zones are considered.)*
4. Uncertainty estimation**:** modern LLMs are being equipped to better calibrate their confidence levels. A low-confidence score on a claim that looks like an entailment could be a signal for a potential hallucination.

10. Related work and discussions

Most philosophers of science acknowledge that Gilbert **Harman's** (1965) notion of *Inference to the Best Explanation* must be qualified to reflect the cognitive and epistemic limitations of human reasoners. In its ideal form, IBE suggests that when confronted with a set of competing explanations for a given phenomenon, one ought to infer the *best* among them as true—assuming that explanatory goodness correlates with truth. However, in practice, this idealization fails to hold. As several authors have argued (e.g., Lipton, 1991; Psillos, 2002; Douven, 2021), reasoners rarely, if ever, have epistemic access to *all possible explanations*. The space of conceivable hypotheses is vast, open-ended, and often constrained by one's background knowledge, conceptual frameworks, and methodological paradigms.

 doi:10.20944/preprints202512.0598.v1

Hence, what scientists and everyday reasoners actually perform is not IBE in the ideal sense, but rather an **Inference to the Best Available Explanation (IBAE)**. The qualifier "available" underscores that explanatory selection occurs within the limits of what is currently *conceived, articulated, and epistemically accessible*. Consequently, the rationality of the inference depends not only on the comparative quality of the candidate explanations but also on the **completeness and maturity of the explanatory landscape** at a given time.

Yet, as **Lipton** (1991) and other authors have emphasized, even the best *available* explanation is not always *rationally acceptable*. In domains characterized by novelty, uncertainty, or insufficient empirical grounding, the best explanation one can offer may amount to little more than an informed conjecture—or even pure speculation. The quality of inference thus depends not on its relative optimality among known hypotheses, but on its **absolute adequacy** in meeting epistemic and methodological standards.

A historical example illustrates this tension well. In early animistic worldviews, natural phenomena such as the movement of the sun across the sky or the occurrence of thunderstorms were explained in terms of **intentional agency**—the sun as a sentient being, or thunder as the anger of gods. Within those conceptual systems, these were indeed the *best available explanations*, since alternative mechanistic or astronomical accounts were not yet conceivable. Nevertheless, such explanations are methodologically inadequate from the standpoint of modern science because they fail to satisfy the essential criteria of **empirical testability, causal coherence, and predictive power.**

Therefore, while IBAE captures a more realistic model of human explanatory reasoning than idealized IBE, it also highlights a fundamental epistemic constraint: the *availability* of explanations is historically and cognitively bounded. Scientific progress often depends precisely on expanding this space of availability—by introducing new conceptual resources, methodological tools, or theoretical frameworks that enable the formulation of *better* explanations than were previously possible.

In this light, the evolution from animistic speculation to heliocentric astronomy or from vitalism to molecular biology illustrates a broader pattern: rational explanation is not static but **dynamic**, expanding through the iterative cycle of **abduction, deduction,** and **induction**. Abduction generates conjectural hypotheses; deduction derives their empirical consequences; induction tests and refines them. IBAE thus marks the *context-bound* nature of abductive reasoning—it represents the best explanation one can formulate *given the current state of knowledge*, but not necessarily the best explanation *simpliciter*.

The distinction between Inference to the Best Explanation and Inference to the Best Available Explanation (IBAE) has profound implications for artificial intelligence, particularly for systems that aim to emulate or augment human reasoning. In computational contexts—such as ALP, Bayesian reasoning, and neuro-symbolic inference frameworks—the concept of "availability" translates directly into the boundedness of hypothesis spaces and the constraints of representational languages.

Just as human reasoners can only choose among the explanations they can conceive, an AI system can only infer among the hypotheses it is able to generate or represent within its formalism. Hence, the system's abductive reasoning process operationalizes IBAE rather than ideal IBE. The "best explanation" the model arrives at is the best *available* given (1) its background knowledge base, (2) its hypothesis-generation rules, and (3) its evaluation criteria (such as likelihood, plausibility, or explanatory coherence).

In abductive logic programming (Kakas & Mancarella, 1990), this principle is evident in the structure of the inference cycle:

1. Abductive generation: The system proposes candidate explanations—hypotheses that, when combined with background knowledge, entail the observed data.
2. Deductive testing: The implications of each hypothesis are deduced and checked against constraints.
3. Inductive evaluation: Empirical or probabilistic measures assess which hypotheses remain consistent with evidence.

 doi:10.20944/preprints202512.0598.v1

Here, the availability constraint manifests through the space of abductive hypotheses the system can construct—defined by its symbolic vocabulary, its logical rules, and its computational resources. Thus, the inferential behavior of an ALP system corresponds not to an ideal IBE, which presupposes access to all possible explanations, but to an IBAE process bounded by representational and algorithmic feasibility.

Inductive inferences in the narrow sense are well-investigated, but their inferential power is limited. With their help it is possible to reason from regularities observed in the past to unobserved or future instances of these regularities, but it is not possible to infer conclusions containing new concepts expressing un observed properties that are not contained in the premises. These inferences are conceptually creative. They are needed in science whenever one reasons from the observed phenomena to theoretical concepts. Theoretical concepts in science describe unobservable properties (e.g., electric forces) or structures (e.g., electrons) that explain the observed phenomena in a unified way. It was an important insight of the post-positivistic philosophy of science that these theoretical concepts cannot be reduced to observable concepts via chains of definitions (see Carnap 1956; Hempel 1951; Stegmüller 1976; French 2008; Schurz 2021). Thus the justification of explanations introducing the oretical concepts cannot be based on conceptual analysis, but must take the form of an ampliative inference.

Modern neuro-symbolic AI systems (Bader & Hitzler, 2005; d'Avila Garcez et al., 2019) face similar epistemic limitations. Even when neural networks are used to expand the hypothesis space by generating candidate patterns or latent variables, the symbolic reasoning layer can only evaluate those that fit within its logic schema. The model therefore performs an approximation to abduction, balancing between *expressive generativity* (neural) and *logical evaluability* (symbolic). This interplay mirrors the philosophical IBAE trade-off: the system can only infer the best explanation among those currently expressible and computationally tractable.

Furthermore, the evaluation criterion—what counts as the "best" explanation—must also be contextually defined. In philosophy, explanatory virtues include simplicity, coherence, and unification (Lipton, 1991); in AI, analogous metrics include likelihood, posterior probability, minimality, or information gain. As in human inquiry, a system's "rational acceptability" depends on whether its best available explanation satisfies the methodological standards relevant to its domain—e.g., causal adequacy in science, logical consistency in expert systems, or interpretability in explainable AI.

Seen through this lens, the evolution of AI reasoning systems can be interpreted as an ongoing attempt to expand the availability space—to make systems capable of generating and evaluating increasingly complex, context-sensitive, and semantically rich hypotheses. Advances in LLMs, probabilistic logic, and symbolic–neural hybrids all contribute to this expansion, approximating the human process of abductive discovery within formal computational frameworks. In doing so, AI research effectively continues the Peircean program: integrating abduction for hypothesis generation, deduction for consequence derivation, and induction for empirical validation—a triadic cycle now realized not only in human thought but also in machine reasoning.

HaluCheck is introduced as a visualization framework for assessing and prominently displaying the likelihood of hallucination in model outputs (Heo et al 2025). It allows users to select among multiple LLMs, providing flexibility for different tasks and preferences. The system integrates a diverse set of hallucination-evaluation metrics, enabling users to compute and compare likelihood scores using alternative methods. By allowing users to switch between these evaluators, HaluCheck supports experimentation with different assessment strategies and helps identify the most effective approach for a given use case.

Abduction addresses fundamental structural weaknesses of Chain-of-Thought reasoning by supplying missing premises, enforcing global coherence, enabling defeasible revision, supporting competing explanations, and regularizing explanations through minimality (Lin 2025). In a neuro-symbolic pipeline, CoT becomes a manipulable reasoning object whose validity can be checked, repaired, and optimized. The resulting system offers a principled alternative to unconstrained LLM

 doi:10.20944/preprints202512.0598.v1

reasoning, replacing narrative fluency with explanation-centered computation. Such integration yields more robust, trustworthy, and interpretable reasoning across domains requiring structured decision-making.

Reasoning with missing premises remains a core challenge. (Li et al. 2024) improve multi-hop knowledge graph reasoning using reinforcement-based reward shaping to better infer intermediate steps, while Quach et al. 2024) incorporate compressed contextual information into knowledge graphs via reinforcement learning. These efforts parallel abductive reasoning in their shared goal of supplying or optimizing missing intermediate premises.

In the framework proposed by Shi et al. (2023), an LLM is trained to perform abductive reasoning using a small set of expert-annotated demonstrations. The model generates plausible causes for a given proposal, and each hypothesized cause is then used as a query to retrieve similar or relevant real-world events. A secondary neural model embeds these retrieved instances and evaluates whether they genuinely support the original proposal.

Yao et al. (2023) conceptualize hallucination as an adversarial phenomenon. Using gradient-guided token substitutions, they construct prompts specifically designed to elicit hallucinated outputs. Their analysis shows that the first token generated from an unperturbed prompt typically has substantially lower entropy than the first token produced under adversarial manipulation. Leveraging this discrepancy, they propose an entropy-based threshold to operationalize the detection of hallucination-inducing adversarial attacks.

Integration of explicit logical reasoning outside of LLM with LLM is an extensive area of research. The system of Zeng et al (2025) comprises two modules: Knowledge Retrieval and Reasoning and Answering (RA). The KR module is LLM-independent and employs an entity-linking algorithm and a subgraph construction and fusion strategy to retrieve question-relevant knowledge. The architecture is oriented towards health, similar to the current study.

Modern explainable AI (XAI) techniques remain far from delivering human-like answers to *why*-questions, and even further from producing explanations that align with human-level understanding. Most existing methods ultimately reduce explanations to sets of causal attributions, leaving the acceptance of those attributions largely—if not entirely—dependent on the explainee's subjective judgment of their adequacy. Medianovskyi and Pietarinen (2022) argue that this evaluative burden could be shifted from humans to XAI agents themselves, provided these agents employ machine-learning algorithms capable of performing genuinely abductive inferences. Their work highlights a fundamental limitation of the dominant inductive paradigm in contemporary ML and its associated XAI practices, and outlines key desiderata for a second-generation, participatory XAI framework grounded in abductive reasoning.

Pietarinen and Beni (2021) develop an intellectual synthesis that connects, on one side, active inference and the free-energy principle (FEP), and on the other, Charles S. Peirce's semiotics and pragmatism. The present paper concentrates on the conceptual affinity between the notions of *active* and *abductive* inference as a suitable entry point for pursuing this broader integration. The authors outline the key theoretical components required for a naturalistic reconstruction of Peirce's late semiotic and logical conception of abduction. The overarching aim is to formulate a cognitive–biological model of abductive reasoning that maintains the organism's functional integrity while satisfying the existential imperative intrinsic to living systems—namely, the continual production of evidence of their own existence. This model draws on, and adapts, Peirce's late abductive schema as articulated in his largely unpublished writings from the early twentieth century. The proposed framework provides not only a viable interpretation of Peirce's ideas but also a conceptual bridge to recent advances in computational (specifically Bayesian) cognitive science.

Dubois et al. (2008) examine the problem of abduction within a Bayesian framework when the prior probability of the hypothesis is unavailable—either due to a lack of statistical data or because experts are unwilling or unable to provide a subjective prior. This situation leaves the abductive inference task unresolved, as standard sensitivity analysis on the missing prior often yields vacuous or indeterminate results. The authors propose a set of criteria that any satisfactory solution to this

 doi:10.20944/preprints202512.0598.v1

problem should meet and then review several existing or newly introduced approaches. These include the use of likelihood functions, classical information-theoretic principles such as maximum entropy, cooperative game–theoretic constructs like the Shapley value, and maximum-likelihood–based strategies. The paper culminates in the development of a novel maximum-likelihood solution grounded in conditional event theory, formulated within de Finetti's coherence framework, which admits conditioning on contingent events even when they have probability zero.

11. Conclusions

By embedding abduction into the entropy-based framework, hallucination detection becomes a structured evaluation of *conditional justifiability*. This integration enables systems not only to identify unsupported content but also to differentiate between benign hypothesis formation, plausible inference, domain-appropriate generalization, and genuine error—bringing the combined model significantly closer to human standards of reasoning and explanation.

The discourse-aware abductive framework introduced in this work provides a principled foundation for constructing and verifying complex explanations generated by LLMs. By integrating abductive inference with rhetorical structure analysis, the approach enables systems to distinguish central, hypothesis-bearing content from peripheral or contextual material, thereby strengthening both explanatory precision and hallucination detection. The value of this integration is evident across multiple application domains.

Counter-abduction is thus a foundational component of hallucination-resistant neuro-symbolic reasoning. By positioning rival explanations as defeaters of LLM-generated CoTs, counter-abductive reasoning transforms narrative reasoning into a competitive, evidence-driven process grounded in logic and discourse structure. This provides a unified theoretical and computational basis for hallucination detection and correction across medical analysis, legal reasoning, scientific interpretation, and general-purpose CoT verification.

In **medical narratives**, weighting discourse nuclei over satellite descriptions allows the system to focus abductive diagnosis generation on patient-relevant complaints rather than tangential remarks, improving causal hypothesis extraction. In **legal reasoning**, the framework supports more transparent argument evaluation by giving precedence to claims occurring in the conclusion or main argument segments while attenuating the influence of background information. In **scientific writing**, it enhances the identification of robust causal explanations by prioritizing claims derived from results and discussion sections over speculative or forward-looking commentary. Finally, in **LLM verification**, discourse-aware abductive logic programming offers a structured mechanism for identifying hallucinations: statements originating in low-weight, peripheral text segments can be selectively discounted, while central claims undergo rigorous consistency checking.

Taken together, these applications demonstrate that combining abductive reasoning with discourse structure provides a versatile and effective method for improving reasoning fidelity, ensuring interpretability, and increasing trust in neuro-symbolic systems across diverse high-stakes domains.

Advantages:

- Increases interpretability: abductive hypotheses are justified by discourse roles.
- Improves precision: ignores peripheral text when generating explanations.
- Enables alignment with human reasoning: since humans emphasize nuclei when forming explanations.
- Supports hallucination detection: contradictions in nucleus-derived claims outweigh peripheral inconsistencies.

"Counter-abduction strength of confirmation metrics dialogue" refers to a structured, interactive process where:

1. Abductive reasoning proposes initial explanations for observed evidence.
2. Counter-abduction introduces competing explanations.

 doi:10.20944/preprints202512.0598.v1

3. Confirmation metrics quantitatively assess how well the evidence supports each hypothesis.

4. Dialogue facilitates the comparison and discussion of these assessments to arrive at the most plausible explanation.

This framework is powerful in any scenario requiring rigorous evaluation of competing hypotheses, ensuring that conclusions are well-supported by evidence. It combines logical reasoning, probabilistic assessment, and collaborative discussion to navigate complex, uncertain situations effectively (compare with Zhang et al 2025).

Our evaluation confirms that discourse structure and counter-abduction jointly improve both the logical soundness and perceived credibility of AI reasoning. D-ALP not only infers plausible explanations but also tests their robustness against rival interpretations, substantially reducing hallucinations. These combined results highlight the promise of discourse-aware abductive reasoning as a foundation for verifiable, trustworthy neuro-symbolic AI systems. In practical applications, the abductive hallucination discovery should work on top of white, grey and black-box families of approaches (Wu et al 2021) to be most efficient (Galitsky and Tsyrlin 2025).

Using web search frequencies to approximate the probabilistic components of MDL effectively turns explanation evaluation into a form of fact checking via large-scale web evidence. By grounding hypotheses and their supporting statements in empirical web co-occurrence statistics, the method implicitly verifies whether a proposed explanation aligns with widely attested facts, conventional causal relations, or commonly observed patterns. In this sense, the approach functions similarly to evidence retrieval pipelines—mapping a claim to the web and measuring how well it is supported—but does so in a model-agnostic and distribution-free way.

At the same time, this strategy is more versatile than standard fact checking. Rather than requiring explicit evidence passages or structured knowledge bases, the method leverages the web's implicit probability distribution: the relative frequency of statements serves as a proxy for how "complex," "unexpected," or "unsupported" an explanation is under MDL. This allows the system to score explanations even when no clean supporting document exists, and to detect misleading but superficially plausible reasoning by measuring its mismatch with broad linguistic and factual usage.

Thus, web-based probabilistic reconstruction provides a lightweight but powerful mechanism for explanation assessment—combining the grounding benefits of fact checking with the flexibility and generality of information-theoretic modeling.

11.1. Limitations

While web-search frequencies provide a convenient and scalable proxy for estimating description lengths, this approach introduces several important limitations. First, search-engine hit counts are inherently **noisy and unstable.** They vary across time, region, device, and even repeated queries, reflecting index fluctuations, ranking algorithms, and undocumented heuristics rather than true corpus frequencies. As a result, the estimated probabilities $p(H)$ and $p(O|H)$ may exhibit high variance and occasional discontinuities.

Second, web frequencies are highly **sensitive to query formulation**. Small changes in phrasing, ordering, or stemming can produce large differences in result counts. Synonyms, paraphrases, and domain-specific terminology further complicate interpretation. Because hypotheses and observations rarely have a unique canonical linguistic form, model comparison may be biased by the particular string chosen to represent each proposition. This sensitivity undermines the reproducibility and robustness of the MDL estimates.

Third, the web contains significant **topical, linguistic, and geographical biases**. High-frequency content often reflects media cycles, SEO-optimized text, misinformation, and commercial duplication rather than underlying factual priors. Thus, common hypotheses may be "simple" in an information-theoretic sense only because they are culturally salient, newsworthy, or sensationalized—not because they genuinely have low description length in a formal model class. Conversely, accurate but

 doi:10.20944/preprints202512.0598.v1

specialized scientific hypotheses may receive artificially high code lengths due to their limited online footprint.

Finally, this method lacks principled handling of **joint or conditional queries.** Co-occurrence counts such as $f(H,O)$ depend heavily on query operators ("AND", quotation marks, proximity constraints), each interpreted differently across search platforms. Consequently, the derived conditional code lengths $L(O|H)$ inherit semantic ambiguities from the search interface itself.

Taken together, these factors mean that web-based description lengths should be viewed as ***heuristic approximations*** rather than precise statistical quantities. They are most reliable when used for coarse-grained hypothesis ranking, and should ideally be complemented by more controlled corpora, domain-specific knowledge bases, or formal probabilistic models when higher fidelity is required.

Declaration of competing interest: The author declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. 415

Data availability: Data and code is available at https://github.com/bgalitsky/halluc_in_health/tree/master/abduction

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process: During the preparation of this work the author(s) used GPT5 in order to correct English grammar. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

References

1. mini A, Saadia Gabriel, Peter Lin, Rik Koncel-Kedziorski, Yejin Choi, Hannaneh Hajishirzi (2019) MathQA: Towards Interpretable Math Word Problem Solving with Operation-Based Formalisms. arXiv:1905.13319
2. Arcuschin, M., et al. (2025). Limitations of Chain-of-Thought as Veridical Explanation. (Forthcoming).
3. Azaria, A., Mitchell, T., 2023. The internal state of an llm knows when it's lying. arXiv preprint arXiv:2304.13734
4. Bader, S., & Hitzler, P. (2005). Dimensions of Neural-Symbolic Integration – A Structured Survey. arXiv preprint cs/0509015.
5. Barez, C., et al. (2025). Hallucinated reasoning in LLM chains: A structural analysis. Transactions of the ACL.
6. Berglund L, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2023. The Reversal Curse: LLMs trained on" A is B" fail to learn" B is A". ArXiv preprint abs/2309.12288 (2023). https://arxiv.org/abs/2309.12288
7. Carnap, R. (1962). The logical foundations of probability (2nd ed.). University of Chicago Press.
8. Christiansen H (2009) Executable specifications for hypothesis-based reasoning with Prolog and Constraint Handling Rules, Journal of Applied Logic, Volume 7, Issue 3, 341-362.
9. Crupi, V., Tentori, K., & González, M. (2007). On Bayesian measures of evidential support. Philosophy of Science, 74(3), 229–252.
10. d'Avila Garcez, A. S., Besold, T. R., De Raedt, L., Földiak, P., Hitzler, P., Icard, T., Kühnberger, K. U., Lamb, L. C., Miikkulainen, R., & Silver, D. L. (2019). Neural-Symbolic Learning and Reasoning: A Survey and Interpretation. Philosophical Transactions of the Royal Society A, 377(2140), 20180070.
11. Dhuliawala S, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, Jason Weston (2023) Chain-of-Verification Reduces Hallucination in Large Language Models. arXiv:2309.11495
12. Douven, I. (2021). Abduction. In The Stanford Encyclopedia of Philosophy (Fall 2021 Edition, ed. E. N. Zalta).
13. Dubois D, Angelo Gilio, Gabriele Kern-Isberner, Probabilistic abduction without priors,
14. Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming, and n-person games. Artificial Intelligence, 77(2), 321–357.
15. Earman, J. (1992). Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory. MIT Press.

 doi:10.20944/preprints202512.0598.v1

16. Eiter T, Wolfgang Faber, Christoph Koch, Nicola Leone, Gerald Pfeifer (2000) DLV - A System for Declarative Problem Solving. arXiv:cs/0003036
17. Ferrag, A., et al. (2025). Structured reasoning failures in large language models. In Proceedings of ACL.
18. French, Stephen 2008: "The Structure of Theories." In: The Routledge Companion to Phi losophy of Science, edited by Stathis Psillos & Martin Curd, New York: Routledge, 269–280
19. Galitsky B (2021) Improving open domain content generation by text mining and alignment. In Galitsky B and Goldberg S: AI for Health Applications and Management, Elsevier
20. Galitsky B (2024) Truth-o-meter: Collaborating with llm in fighting its hallucinations. In Interdependent Human-Machine Teams, 175-210
21. Galitsky B (2025) Chapter 8 - Identifying large language model hallucinations in health communication. In "Healthcare Applications of Neuro-Symbolic Artificial Intelligence" pages 283-329, Elsevier
22. Galitsky B, Tsirlin A (2025) Step Wise Approximation of CBOW Reduces Hallucinations in Tail Cases. https://www.preprints.org/manuscript/202507.0670
23. Ghallab, M., Nau, D., & Traverso, P. (2016). Automated Planning and Acting. Cambridge University Press.
24. Gillies, D. (1991). Intersubjectivity in Science. Harvester Wheatsheaf.
25. Hacking, I. (1965). The Logic of Statistical Inference. Cambridge University Press.
26. Haig, B. D. (2005). An abductive theory of scientific method. Psychological Methods, 10(4), 371–388.
27. Haig, B. D. (2014). Investigating the Psychological World: Scientific Method in the Behavioral Sciences. MIT Press.
28. Haig, B. D. (2014). Investigating the Psychological World: Scientific Method in the Behavioral Sciences. MIT Press.
29. Harman, G. (1965). The inference to the best explanation. The Philosophical Review, 74(1), 88–95.
30. Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. ArXiv preprint abs/2207.05221 (2022). https://arxiv.org/abs/2207.05221
31. Heo, S., Son, S., Park, H., 2025. Halucheck: Explainable and verifiable automation for detecting hallucinations in llm responses. Expert Systems with Applications, 126712.
32. Huang L, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. ACM Trans. Inf. Syst. 43, 2, Article 42 (March 2025), 55 pages. https://doi.org/10.1145/3703155
33. Ignatiev A & Narodytska, Nina & Marques-Silva, Joao. (2019). Abduction-Based Explanations for Machine Learning Models. Proceedings of the AAAI Conference on Artificial Intelligence. 33. 1511-1519. 10.1609/aaai.v33i01.33011511.
34. International Journal of Approximate Reasoning, V 47, Issue 3, 2008, 333-351.
35. Itti L, Pierre Baldi (2009) Bayesian surprise attracts human attention, Vision Research,V 49, Issue 10, 1295-1306
36. Jansen P, M Surdeanu, P Clark (2014) Discourse complements lexical semantics for non-factoid answer reranking. Proceedings of the 52nd Annual Meeting ACL.
37. Jiaying Wu, Ning Dong, Fan Liu, Sai Yang, Jinglu Hu, Feature hallucination via Maximum A Posteriori for few-shot learning, Knowledge-Based Systems, V 225, 107129, 2021
38. Kakas A. (2000) ACLP: Integrating Abduction and Constraint Solving. arXiv:cs/0003020
39. Kakas, A. C., & Mancarella, P. (1990). Generalized Abduction. Journal of Logic and Computation, 1(3), 389–407.
40. Kakas, A. C., Kowalski, R. A., & Toni, F. (1992). Abductive logic programming. Journal of Logic and Computation, 2(6), 719–770.
41. Kossen, J., Han, J., Razzak, M., Schut, L., Malik, S., Gal, Y., 2024. Semantic entropy probes: Robust and cheap hallucination detection in llms. FigarXiv preprint arXiv:2406.15927 .
42. Li C, H. Zheng, Y. Sun, C. Wang, L. Yu, C. Chang, X. Tian, and B. Liu (2024) Enhancing multi-hop knowledge graph reasoning through reward shaping techniques," in 2024 4th International Conference on Machine Learning and Intelligent Systems Engineering (MLISE), pp. 1-5, IEEE

 doi:10.20944/preprints202512.0598.v1

43. Lin S (2025) Abductive Inference in Retrieval-Augmented Language Models: Generating and Validating Missing Premises. arXiv:2511.04020v1
44. Lipton, P. (2004). Inference to the Best Explanation (2nd ed.). Routledge.
45. Manakul P, Adian Liusie, and Mark J. F. Gales. 2023. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. ArXiv preprint abs/2303.08896 (2023). https://arxiv.org/abs/2303.08896
46. Matsumoto S, A. Barreto, P. C. G. Costa, B. Benyo, M. Atighetchi and D. Javorsek, "Dynamic Explanation of Bayesian Networks with Abductive Bayes Factor Qualitative Propagation and Entropy-Based Qualitative Explanation," 2021 IEEE 24th International Conference on Information Fusion (FUSION), Sun City, South Africa, 2021, pp. 1-9, doi: 10.23919/FUSION49465.2021.9626961.
47. Medianovskyi, K.; Pietarinen, A.-V. On Explainable AI and Abductive Inference. Philosophies 2022, 7, 35. https://doi.org/10.3390/philosophies7020035
48. Neogi T, Chen C, Niu J, Chaisson C, Hunter DJ, Choi H, Zhang Y. Relation of temperature and humidity to the risk of recurrent gout attacks. Am J Epidemiol. 2014 Aug 15;180(4):372-7. doi: 10.1093/aje/kwu147.
49. Ning Miao, Yee Whye Teh, and Tom Rainforth. 2023. Selfcheck: Using LLMs to zero-shot check their own step-by-step reasoning. ArXiv preprint abs/2308.00436 (2023).
50. Peirce, C. S. (1878). Illustrations of the Logic of Science: Deduction, Induction, and Hypothesis. Popular Science Monthly, 13, 470–482.
51. Peirce, C. S. (1903). Lectures on Pragmatism. In Collected Papers of Charles Sanders Peirce (Vol. 5, ed. C. Hartshorne & P. Weiss). Harvard University Press.
52. Pietarinen, AV., Beni, M.D. Active Inference and Abduction. Biosemiotics 14, 499–517 (2021). https://doi.org/10.1007/s12304-021-09432-0
53. Prakken, H., & Vreeswijk, G. (2002). Logics for defeasible argumentation. In Handbook of Philosophical Logic (pp. 219–318). Springer.
54. Psillos, S. (2002). Causation and Explanation. Acumen.
55. Quach N, Q. Wang, Z. Gao, Q. Sun, B. Guan, and L. Floyd (2024) Reinforcement learning approach for integrating compressed contexts into knowledge graphs," in 5th International Conference on Computer Vision, Image and Deep Learning (CVIDL), pp. 862–866
56. Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac
57. Schurz G (2008) Patterns of abduction. Synthese (2008) 164:201–234 DOI 10.1007/s11229-007-9223-4
58. Shi X and Xue, Siqiao and Wang, Kangrui and Zhou, Fan and Zhang, James and Zhou, Jun and Tan, Chenhao and Mei, Hongyuan (2023) Language Models Can Improve Event Prediction by Few-Shot Abductive Reasoning. Advances in Neural Information Processing Systems, 29532-29557
59. Su, W., Wang, C., Ai, Q., Hu, Y., Wu, Z., Zhou, Y., Liu, Y., 2024. Unsupervised real-time hallucination detection based on the internal states of large language models, in: Ku, L.W., Martins, A., Srikumar, V. (Eds.), Findings of the Association for Computational Linguistics: ACL 2024, Association for Computational Linguistics, Bangkok, Thailand. pp. 14379–14391.
60. Swi-prolog (2025) https://www.swi-prolog.org/
61. Tavory, I., & Timmermans, S. (2014). Abductive Analysis: Theorizing Qualitative Research. University of Chicago Press.
62. Timmermans, S., & Tavory, I. (2012). Theory construction in qualitative research: From grounded theory to abductive analysis. Sociological Theory, 30(3), 167–186.
63. Varshney N, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. 2023. A Stitch in Time Saves Nine: Detecting and Mitigating Hallucinations of LLMs by Validating Low-Confidence Generation. ArXiv preprint abs/2307.03987 (2023). https://arxiv.org/abs/2307.03987
64. Varshney, N., Yao, W., Zhang, H., Chen, J., Yu, D., 2023. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation. arXiv preprint arXiv:2307.03987
65. Wang, X., Yan, Y., Huang, L., Zheng, X., Huang, X.J., 2023. Hallucination detection for generative large language models by bayesian sequential estimation, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 15361–15371.

 doi:10.20944/preprints202512.0598.v1

66. Wei, J., et al. (2022). Chain-of-Thought prompting elicits reasoning in large language models. NeurIPS.
67. Wernhard C (2011) Computing with Logic as Operator Elimination: The ToyElim System. arXiv:1108.4891
68. Yadav A (2024) Understanding Information Gain in Decision Trees: A Complete Guide. Medium. https://medium.com/biased-algorithms/understanding-information-gain-in-decision-trees-a-complete-guide-7774c6e0255b
69. Yao J-Y, Kun-Peng Ning, Zhen-Hui Liu, Mu-Nan Ning, and Li Yuan. 2023. LLM Lies: Hallucinations are not Bugs, but Features as Adversarial Examples. ArXiv preprint abs/2310.01469 (2023).
70. Zeng Z, Qing Cheng, Xingchen Hu, Yan Zhuang, Xinwang Liu, Kunlun He, Zhong Liu. KoSEL: Knowledge subgraph enhanced large language model for medical question answering, Knowledge-Based Systems, V 309, 2025.
71. Zhang X, Fuyong Zhao, Yutian Liu, Panfeng Chen, Yanhao Wang, Xiaohua Wang, Dan Ma, Huarong Xu, Mei Chen, Hui Li, TreeQA: Enhanced LLM-RAG with logic tree reasoning for reliable and interpretable multi-hop question answering, Knowledge-Based Systems, Volume 330, Part A, 2025, 114526,
72. Zheng S, Jie Huang, and Kevin Chen-Chuan Chang. 2023. Why Does ChatGPT Fall Short in Answering Questions Faithfully? ArXiv preprint abs/2304.10513 (2023). https://arxiv.org/abs/2304.10513
73. Zhong W, Jinglin Huang, Maoqiang Wu, Weinan Luo, Rong Yu, Large language model based system with causal inference and Chain-of-Thoughts reasoning for traffic scene risk assessment, Knowledge-Based Systems, V 319, 2025, 113630