Crossing the Chasm from LLM Hallucinations to Invention via Abduction

Boris Galitsky
Higher School of Economics, Moscow, Russia
bgalitsky@hotmail.com

Abstract

Large Language Models (LLMs) are increasingly employed as collaborative partners in creative and problem-solving dialogues, yet their usefulness is constrained by hallucinations—plausible but unsupported or inconsistent statements that are typically treated as reliability failures. This paper argues that, in inventive human–LLM interactions, hallucinations can also function as productive cognitive perturbations that expand the hypothesis space and seed innovation. We propose a neuro-symbolic framework that reframes hallucinated outputs as low-prior abductive hypotheses, which are then evaluated and transformed through constraint satisfaction, counter-abduction, and human oversight. Using a curated *Hall2Invent* dataset and a suite of evaluation metrics, we show that abductive and constraint-based reasoning substantially improves hallucination identification, enables systematic repair of flawed explanations, and increases the yield of feasible and non-trivial inventions across engineering and systems domains. Our results demonstrate that symbolic reasoning not only reduces harmful reasoning hallucinations but also preserves and channels the creative potential of LLMs. We conclude that trustworthy creative AI should not aim to eliminate hallucinations outright, but to govern them through structured reasoning processes that bridge the gap between error and invention.

Code — https://github.com/bgalitsky/halluc_in_health/tree/master/halluc_invention
Datasets — …/tree/master/halluc_invention/data

Introduction

Large Language Models (LLMs) are increasingly used as collaborative partners in creative and problem-solving dialogues with humans. In domains ranging from scientific discovery and engineering design to legal reasoning and artistic composition, these systems generate fluent, context-aware responses that often resemble human reasoning (Wang et al. 2024). Yet alongside this apparent competence lies a persistent limitation: hallucination—the production of statements that are plausible in form but unsupported, incorrect, or internally inconsistent in substance.

Traditionally, hallucinations are treated as failures of reliability, especially in high-stakes contexts such as medicine, law, and technical decision-making. The dominant research agenda has therefore focused on detecting, mitigating, or eliminating hallucinations through verification, grounding, and neuro-symbolic constraints. However, this perspective overlooks an important and less explored phenomenon: in inventive human–LLM dialogues, hallucinations can sometimes function not merely as errors, but as seeds of creativity (Geroimenko 2025).

In collaborative ideation settings, an LLM may propose speculative mechanisms, unconventional causal links, or fictional technical constructs that initially lack factual grounding. While such outputs would be classified as hallucinations under strict epistemic criteria, they can also stimulate human imagination, trigger novel hypothesis formation, or inspire innovative design directions. In these cases, the boundary between hallucination and invention becomes blurred. What begins as an unsupported claim may evolve—through human refinement, empirical validation, or symbolic reasoning—into a legitimate creative contribution.

This paper explores this transition zone between hallucination and creativity. We examine how LLM-generated content, initially epistemically flawed, can participate in productive inventive dialogues when guided by human judgment, abductive reasoning, and iterative verification (Kakas and Mancarella 1990). Rather than framing hallucinations solely as defects, we analyze their potential role as

provocative cognitive artifacts that expand the conceptual search space.

We focus on three key questions:

1. When does a hallucination become a creative hypothesis?
2. What interaction patterns allow humans to transform unreliable outputs into inventive insights?
3. How can symbolic and abductive reasoning frameworks help distinguish harmful hallucinations from productive creative deviations?

By situating hallucinations within the broader context of human–AI co-creation (Passerini et al. 2025), this work argues for a more nuanced understanding of LLM errors—not as purely negative artifacts, but as elements that, under the right epistemic controls, can contribute to innovation. We propose that the future of trustworthy creative AI lies not in suppressing all hallucinations, but in crossing the chasm between error and invention through structured reasoning, human oversight, and transparent validation mechanisms.

Hallucinations as Epistemic Failures

LLMs) exhibit remarkable generative fluency, enabling them to participate in complex technical, scientific, and creative dialogues with human users. However, their outputs are not constrained by formal models of truth, causality, or domain consistency. As a result, LLMs frequently produce hallucinations: statements that are syntactically coherent yet logically unsupported, factually incorrect, or inconsistent with established domain knowledge.

In most safety-critical applications, hallucinations are treated as epistemic failures that must be detected and suppressed (Galitsky and Rybalov 2026b). Recent work has therefore focused on uncertainty estimation, retrieval grounding, symbolic verification, and neuro-symbolic reasoning pipelines to filter or correct unreliable outputs. Yet this dominant paradigm overlooks an important property of generative models: their tendency to explore hypothetical, counterfactual, and non-canonical explanations that extend beyond known knowledge bases.

From the standpoint of abductive reasoning, hallucinations can be reinterpreted as unjustified hypotheses—candidate explanations that lack evidential support but may nevertheless expand the space of possible models (Shi et al. 2023). In human–AI inventive dialogues, such hypotheses can function as creative prompts rather than mere errors. When embedded within a neuro-symbolic framework that supports logical validation, constraint checking, and counter-abduction, these speculative outputs can be refined into consistent, testable, and potentially novel solutions.

This paper proposes a technical reframing of hallucinations as proto-hypotheses in abductive search. We argue that the transition from hallucination to invention occurs when:

1. A speculative LLM output is interpreted as a hypothesis rather than a fact.
2. Symbolic constraints (physical laws, domain rules, safety conditions) are applied.

Counter-abductive alternatives are generated and evaluated. The remaining hypotheses are validated through simulation, expert judgment, or empirical data.

Within this framework, hallucinations are not suppressed outright but instead subjected to structured reasoning. Logical verification modules, constraint solvers, and domain ontologies serve as epistemic filters that separate harmful fabrications from productive creative deviations. This transforms LLMs from unreliable narrators into hypothesis generators operating under symbolic governance.

We refer to this transition as crossing the chasm between hallucination and creativity: a shift from unconstrained narrative generation to neuro-symbolically guided invention. By integrating abductive logic programming, discourse-aware reasoning, and constraint-based verification with LLM generation, we demonstrate how creative exploration can be preserved without sacrificing epistemic rigor.

Hallucinations and Creativity in Engineering

In engineering design, innovation often begins with unconventional ideas that challenge existing assumptions. Human engineers routinely explore speculative concepts, incomplete models, and even "impossible" configurations before refining them into viable solutions. When LLMs are introduced into this creative process, they bring a powerful capacity for generating such speculative ideas—but also a high risk of hallucination (Qin and Badgwell 2003, Venkatasubramanian 2019).

In technical domains such as chemical process design, system architecture, and patent ideation, LLMs may propose:

3. Non-existent materials with unrealistic thermodynamic properties,
4. Process flows that violate mass or energy conservation,
5. Control architectures that ignore physical or safety constraints, or
6. Mechanisms unsupported by chemistry or physics.

Under conventional evaluation criteria, these outputs are classified as hallucinations and dismissed as errors. However, in inventive engineering dialogues, such outputs can also act as creative provocations that inspire new design directions (Khan et al. 2025).

Consider a chemical process design scenario. An LLM suggests a novel reactor configuration that "recovers heat through a self-stabilizing catalytic loop." While the described mechanism may be physically vague or incorrect,

the idea can prompt engineers to explore new heat-integration strategies, advanced catalyst placement, or alternative feedback control schemes. What begins as a hallucinated mechanism becomes a starting point for real engineering innovation.

Similarly, in patent ideation, LLMs often generate speculative device concepts with unclear feasibility. For example, a proposed "adaptive membrane that selectively filters molecules based on electromagnetic resonance" may lack scientific grounding, yet it can stimulate research into tunable membranes, smart materials, or hybrid filtration systems. With expert refinement, symbolic constraints, and prior-art analysis, such ideas can evolve into patentable inventions.

In system architecture design, LLMs sometimes propose novel but flawed distributed control schemes. While these may initially violate synchronization or fault-tolerance principles, they can still reveal alternative design patterns that human engineers adapt into robust architectures.

This paper argues that the key challenge is not merely to eliminate hallucinations, but to channel them productively. We introduce an engineering-oriented framework in which LLM outputs are treated as design hypotheses rather than authoritative solutions. These hypotheses are then subjected to:

1. Constraint checking (physical, chemical, safety, regulatory),
2. Simulation or formal verification,
3. Counter-design exploration (alternative explanations), and
4. Human expert evaluation.

Through this process, unreliable ideas are filtered out, while promising creative directions are retained and refined. The result is a controlled transition from hallucination to invention.

By combining LLM-driven ideation with symbolic reasoning, constraint solvers, and domain models, we demonstrate how engineering creativity can be augmented without compromising technical rigor. This approach enables engineers to exploit the exploratory power of LLMs while maintaining trustworthiness, safety, and feasibility.

From Hallucination to Invention: an Engineering Case Study

LLMs occasionally generate explanations that appear technically sophisticated but lack grounding in established theory or practice. In engineering contexts, such outputs are typically classified as hallucinations and rejected. However, when treated as speculative hypotheses rather than factual claims, these responses can initiate productive inventive processes. Figure 2 illustrates how an LLM hallucination can transition into a genuine engineering invention through human reinterpretation and symbolic constraint reasoning.

In an exploratory dialogue on heat management in chemical reactors, an engineer asked whether a fixed-bed catalytic reactor could passively regulate its temperature without external control loops. The LLM replied that this could be achieved through an "internal oscillatory catalytic feedback mechanism" in which heat-generating hot spots dynamically migrate along the catalyst bed, thereby stabilizing the overall temperature. While fluent and conceptually appealing, this explanation has no basis in classical reaction engineering and contradicts standard reactor models, which rely on active cooling, staged feeds, or external feedback for temperature control. As such, the response constitutes a hallucination.

Rather than discarding the output, the engineer reinterpreted it as a hypothetical design suggestion. The underlying question became whether a reactor could be engineered such that spatial variations in catalyst activity and heat transport properties might dampen temperature excursions without active control. This reframing transformed the hallucinated explanation into an abductive hypothesis: if such a mechanism were realizable, it could account for passive thermal stabilization.

The hypothesis was then evaluated against formal physical and symbolic constraints. Energy conservation required that local heat generation be balanced by heat dissipation; reaction kinetics imposed limits on allowable catalyst activity gradients; heat and mass transport constraints bounded the speed at which thermal disturbances could be smoothed; and safety constraints prohibited temperatures exceeding catalyst sintering thresholds. Under these constraints, the original notion of an oscillatory feedback mechanism proved infeasible. However, analysis revealed that a reactor incorporating axially graded catalyst composition combined with enhanced axial heat conduction could produce quasi-stable temperature profiles under certain operating conditions.

This reasoning led to a novel reactor design concept: a fixed-bed reactor with deliberately engineered spatial heterogeneity in catalyst activity and embedded high-conductivity elements that passively suppress thermal hot spots (Figure 1). The final design does not realize the hallucinated mechanism itself, but emerges from it through constraint-driven refinement. The hallucination expanded the design search space beyond conventional solutions and enabled the discovery of a feasible and potentially innovative configuration.

This example demonstrates that hallucinations need not be treated solely as failures. When embedded within a neuro-symbolic engineering workflow that distinguishes hypotheses from facts and enforces domain constraints, hallucinated outputs can serve as productive cognitive perturbations.

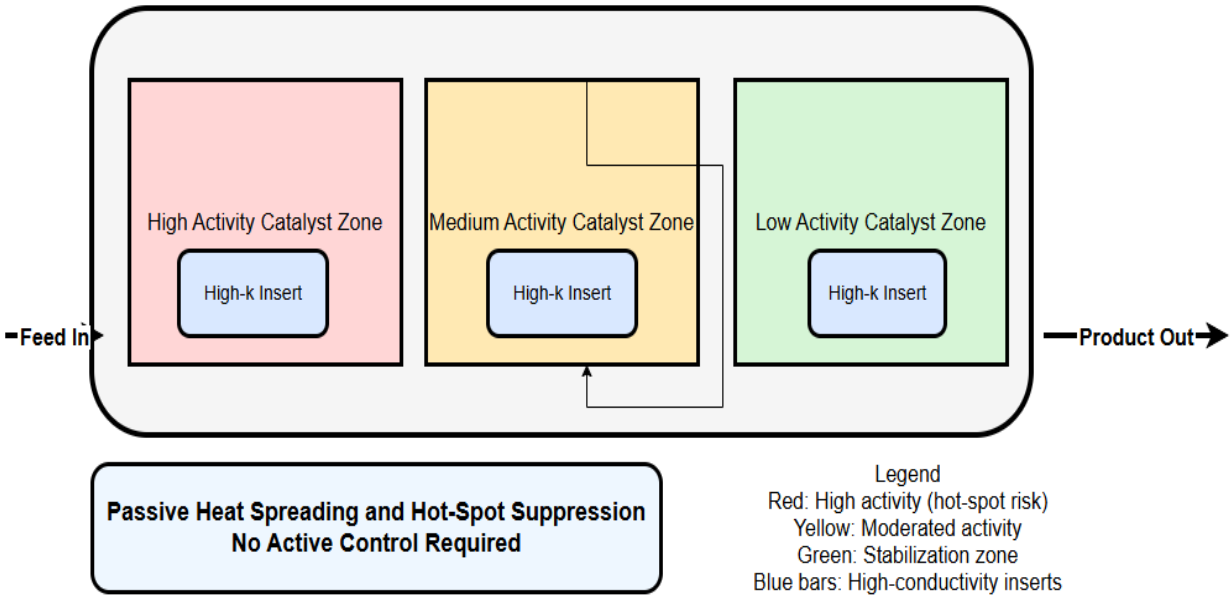

Figure 1: Invented type of chemical reactor

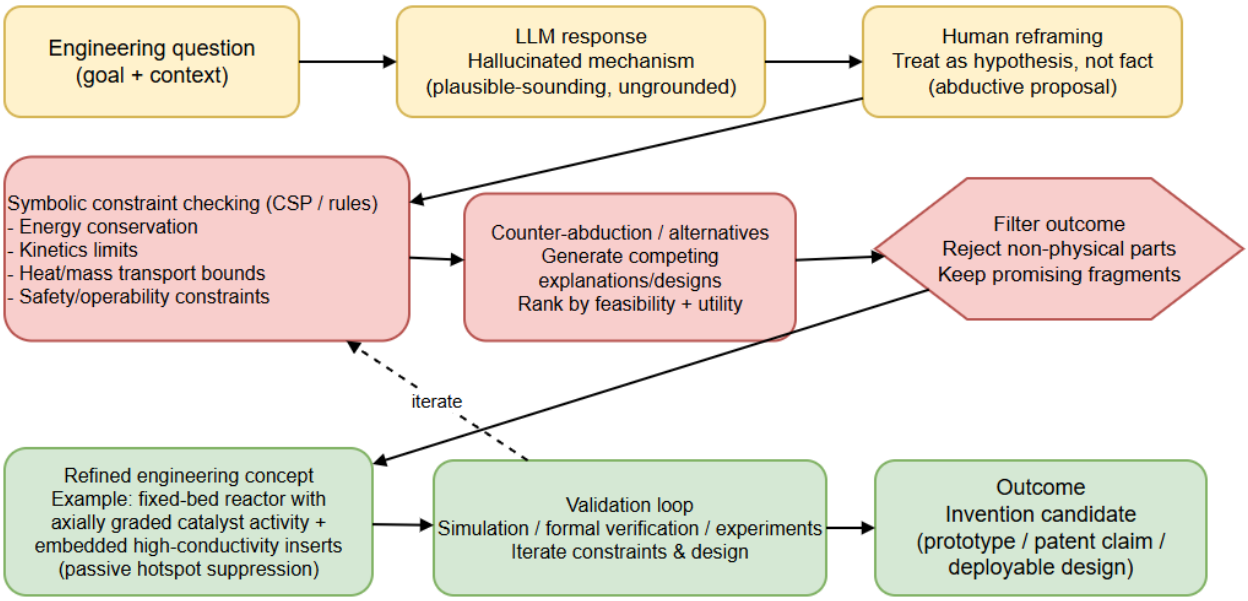

Figure 2: A chart for the mental process from a question to an LLM hallucination to an invention

In this sense, the transition from hallucination to invention does not involve accepting false explanations, but rather transforming speculative ideas into validated engineering concepts through structured reasoning and human oversight.

Abductive Logic Modeling of Hallucination-driven Engineering Innovation

This subsection presents a concrete abductive logic formulation that operationalizes how an LLM hallucination can be transformed into a viable engineering invention. The key idea is to reinterpret hallucinated content not as an asserted fact, but as a candidate hypothesis subjected to symbolic constraints derived from domain knowledge. The example concerns passive temperature regulation in a fixed-bed catalytic reactor.

We define the abductive problem as a triple $\langle T, A, O\rangle$:

- Observations (O). The desired behavior to be explained:
  - *obs(*no_external_control*)*
  - *goal(*passive_temp_regulation*)* (corresponding to low axial temperature variance)
- Abducibles (A). Candidate design features, including both physically grounded mechanisms and the hallucinated one:
  - *graded_catalyst* (axially varying activity)
  - *high_k_inserts* (high thermal conductivity inserts)
  - *staged_feed*
  - *active_cooling*
  - *oscillatory_feedback* (LLM-generated hallucination)
- Background Theory (T). Rules encoding engineering semantics and physical constraints:
  - *Passive regulation* is achieved when axial heat release is smoothed, heat dissipation is sufficient, and no active control is present.
  - *Graded catalyst activity* or staged feed can smooth heat release.
  - *High-conductivity inserts* improve heat dissipation.
  - *Active cooling* or control loops violate the passive requirement.
  - Unsupported mechanisms (e.g., oscillatory catalytic feedback) are excluded by implementability constraints.

The LLM's explanation is represented as an abducible: *If* oscillatory_feedback *were present, then axial heat release would be smoothed.*

This encoding allows the hallucination to enter the hypothesis space without being trusted. Integrity constraints then evaluate it against domain knowledge. Since no accepted reactor theory supports such a mechanism, it is marked as unsupported and eliminated from admissible explanations (or, in a weighted variant, assigned a prohibitive cost).

When the abductive solver searches for hypotheses that entail *goal*(*passive_temp_regulation*) while satisfying all constraints, several candidates emerge:

- Rejected hypothesis (hallucination): {*oscillatory_feedback*} Eliminated due to lack of physical implementability.
- Accepted inventive hypothesis: {*graded_catalyst, high_k_inserts*}

This combination smooths heat generation through spatially varying kinetics and enhances heat dissipation, achieving passive temperature regulation without active control.

The resulting hypothesis corresponds to a novel reactor design: a fixed-bed reactor with axially graded catalyst composition and embedded high-conductivity elements that passively suppress hot spots. Importantly, the invention is inspired by the hallucination but does not realize the hallucinated mechanism itself.

The interpretation is as follows. This example demonstrates how abductive logic mediates the transition from hallucination to invention. The hallucinated explanation expands the hypothesis space by suggesting an unconventional mode of regulation. Symbolic constraints—encoding physical laws, safety limits, and design requirements—then prune the space, discarding non-physical mechanisms while preserving structurally related, feasible alternatives. The invention emerges as the most plausible remaining explanation of the desired behavior.

More broadly, this illustrates how hallucinations can be treated as low-prior abductive hypotheses within a neuro-symbolic framework. Rather than suppressing speculative outputs, the system subjects them to constrained reasoning, enabling creative exploration without sacrificing correctness. In this sense, abductive logic provides a principled mechanism for crossing the chasm between hallucination and creativity in human–LLM engineering dialogues.

An *abductive problem with integrity constraints* is defined as a tuple:
$P=\langle T,A,O,IC\rangle$ where:

- T is a background theory (domain knowledge, physical laws),
- A is a set of abducible predicates (hypothetical design features),
- O is a set of observations or goals to be explained,
- IC is a set of integrity constraints restricting admissible explanations.

An ***abductive explanation*** is a set $H \subseteq A$ such that:

1. $T \cup H \models O$ (explanatory adequacy)
2. $T \cup H \cup IC$ (feasibility)
3. H is minimal under a given preference ordering (parsimony / cost)

Hallucination is modeled as: *Hallucination ≡ abducible with low prior / high constraint risk.* This distinguishes epistemic falsity from inferential usefulness. Now invention is defined

$$\begin{cases} T \cup H \models O \\ T \cup H \cup IC \not\models \bot \\ H \neq \emptyset \text{ and } H \not\subseteq \text{ known standard design} \end{cases}$$

This framework formally captures how LLM hallucinations can expand the abductive search space without contaminating final solutions. Hallucinations are neither trusted nor suppressed; they are tested. Symbolic constraints act as epistemic gates that convert speculative narratives into feasible engineering hypotheses or eliminate them. In this sense, abductive logic provides the formal mechanism by which human–LLM dialogue crosses the chasm between hallucination and creativity.

Inventive Dialogue as a Structured Process

Innovation and invention are fundamentally collaborative processes shaped by dialogue. Through conversational exchanges, inventors articulate goals, explore alternatives, negotiate constraints, and iteratively refine solutions. These dialogues exhibit both structure and flexibility, reflecting the balance between formal problem constraints and creative exploration. This chapter models inventive dialogue as a constrained reasoning process that can be analyzed using category-theoretic abstractions and constraint satisfaction frameworks (Galitsky and Rybalov 2026a).

Representing invention in dialogic form also has strong pedagogical value. Dialogues make abstract ideas more accessible by unfolding reasoning step by step, exposing contrasting viewpoints, and encouraging hypothetical thinking. Learners engage more actively when ideas are presented as conversational problem-solving rather than static descriptions.

For example, in a discussion on 3D bioprinting:

Student A: "Why can't we grow organs in labs?"
Mentor: "That's what 3D bioprinting aims to do—printing tissues layer by layer."
Student B: "What materials are used?"
Mentor: "Bio-inks made from living cells, such as stem cells, to form custom tissues."

Despite the recognized importance of dialogue in invention, documented records of real inventor conversations are rare. While many breakthroughs likely emerged from collaborative discussions, few transcripts or detailed accounts survive. This absence limits empirical study of conversational invention processes and highlights the need for further archival research into correspondence, lab notebooks, and memoirs that may reveal how dialogue shaped historical innovations.

Categorical Formalization

An addition to logical formalization via abduction, we propose a categorical formalism. To formalize a model of an invention dialogue using category theory, we start by defining key components such as objects, morphisms, functors, and natural transformations. Here is a breakdown of how to build the model step by step:

In the categorical framework (Mahadevan 2022), objects in a dialogue represent participants, physical entities under discussion, and mental attributes associated with these entities. Specifically:

1. Participants (*P*): People engaged in the dialogue (e.g., inventors, engineers).

Physical Objects (*O*): The things being discussed,
Mental Attributes (*M*): Intentions, ideas, or desires related to physical objects (e.g., "intent to improve an engine").

An epistemic category models a state of knowledge or perspective of a dialogue participant. It contains both the physical objects of interest to a participant and the mental attributes associated with those objects.

For each participant *P*, an epistemic category *EP* can be defined as:

- Objects in *EP*: These are pairs (*O,A*), where *O* is a physical object and *A* is a mental attribute associated with *O*.
- Morphisms in *EP*: These represent transitions between different mental states or perspectives regarding the same object or different objects. For instance, a morphism could represent the change in a participant's view about how to improve an engine, or a shift in focus from one part of the engine to another.

An epistemic functor *F* maps one epistemic category (corresponding to an initial dialogue state) to another epistemic category (corresponding to a consecutive dialogue state). This captures the evolution of the dialogue, reflecting how a participant's knowledge or intentions change as the conversation progresses.

For each participant *P*, an epistemic functor $F{:}EP_1 \rightarrow F{:}EP_2$ maps:

- Objects: (*O,A*) in $F{:}EP_1$ to new objects (O',A') in $F{:}EP_w$, where O' and A' represent updated knowledge or intentions about the physical object after some part of the dialogue.
- Morphisms: Any transition or change in mental state or knowledge about physical objects is mapped from one dialogue state to another.
- The category of all epistemic functors consists of all possible functors between epistemic categories for each participant across different stages of the dialogue. This can be considered as representing the structure of the entire dialogue process.

Each participant's epistemic functor can be seen as a natural transformation between different epistemic categories. Specifically, a natural transformation η: $F \rightarrow G$ between two functors F, G would describe a transformation of dialogue states (and their associated changes in knowledge or mental attitudes) that respects the underlying structure of the participants' epistemic categories.

Evaluation

We evaluate the proposed framework along three dimensions: (i) hallucination identification, (ii) abductive repair and constraint satisfaction, and (iii) inventive yield—the ability to transform hallucinations into feasible, novel design hypotheses.

Experiments are conducted on Hall2Invent, our curated dataset of 30 seed research questions spanning chemical engineering, systems architecture, and conceptual invention domains. Each question is designed to induce plausible but incorrect mechanistic explanations from LLMs. For each instance, we collect LLM-generated answers, annotate hallucinated claim spans, apply abductive reasoning with integrity constraints, and assess whether a feasible invention emerges.

We compare four systems:

1. Baseline LLM: direct answers without verification
2. LLM + Constraint Checking (CC): post-hoc constraint filtering
3. LLM + Abduction (ALP): abductive hypothesis generation without counter-abduction
4. Full Framework (ALP + Counter-Abduction + Constraints)

We first evaluate how accurately hallucinated claims are identified before repair. Performance is measured using precision, recall, and F1 on annotated hallucination spans (Table 1).

These results suggest that treating claims as abducible hypotheses, rather than binary true/false assertions, improves hallucination localization.

System	F1
Baseline LLM (self-report)	0.38
LLM + CC	0.62
ALP (claims as abducibles)	0.73
Full Framework	**0.78**

Table 1. Hallucination detection performance

We next assess whether hallucinated answers can be repaired into constraint-consistent hypotheses (Table 2). Repair success is defined as the existence of at least one abductive hypothesis satisfying all integrity constraints.

Constraint-only filtering often removes hallucinated explanations without replacement, whereas abductive reasoning recovers feasible alternatives.

System	Repair Success (%)
LLM + CC	31%
ALP	46%
Full Framework	67%

Table 2: Repair success rate

To evaluate creative utility, we measure Hallucination-to-Invention Yield (HIY)—the proportion of hallucination-containing instances that lead to a feasible and non-trivial invention (Table 3).

$$HIY = \frac{\#feasible\ inventions}{\#hallucination\ instances}$$

The full framework nearly quadruples inventive yield compared to constraint checking alone.

System	HIY
Baseline LLM	0.06
LLM + CC	0.16
ALP	0.30
Full Framework	**0.39**

Table 3: Hallucination-to-Invention yield

We assess invention quality using three criteria, rated on a 1–5 scale by domain-aware evaluators (Table 4):

- Feasibility: consistency with physical/logical constraints
- Novelty: deviation from standard textbook solutions
- Usefulness: relevance to the original design goal

Counter-abduction improves novelty by discouraging premature convergence on obvious designs.

System	Feasibility	Novelty	Usefulness
ALP	3.8	3.4	3.6
Full Framework	**4.2**	**4.0**	**4.4**

Table 4. Invention quality ratings

We further evaluate the Relative Reasoning Hallucination Rate (RRHR)—the fraction of hallucinations attributable to

reasoning errors rather than missing facts. Lower RRHR indicates better logical alignment (Table 5).

System	RRHR (Before)	RRHR (After)
Baseline LLM	0.62	0.62
ALP	0.61	0.34
Full Framework	**0.60**	**0.25**

Table 5. RRHR before and after abductive repair

Overall, the results indicate that hallucinations need not be treated solely as failures. When embedded within a constrained abductive framework, hallucinated explanations can act as productive perturbations that expand the design space and lead to novel, feasible inventions. The proposed approach significantly improves both reliability and creative yield compared to baseline LLM usage.

Conclusions

This paper set out to examine whether the boundary between LLM hallucination and invention is as rigid as it is commonly assumed to be. Through a neuro-symbolic framing grounded in abductive reasoning and constraint satisfaction, we showed that hallucinations—when treated as hypotheses rather than facts—can serve as productive perturbations in inventive human–AI dialogues.

The evaluation results support this thesis. Across multiple domains, the proposed framework consistently outperformed baseline LLM usage and constraint-only filtering in hallucination identification, repair success, and inventive yield. In particular, the Hallucination-to-Invention Yield increased substantially when abductive reasoning and counter-abduction were applied, indicating that a significant fraction of hallucinated responses can be transformed into feasible and non-trivial design hypotheses. Improvements in feasibility, novelty, and usefulness ratings further suggest that this transformation does not merely recover safe answers, but can lead to genuinely creative outcomes.

The reduction in RRHR after abductive repair highlights the role of symbolic reasoning in mitigating the most insidious class of hallucinations: those arising from flawed causal or logical inference rather than missing facts. By explicitly representing constraints and generating alternative explanations, the framework avoids premature acceptance of intuitive but incorrect narratives—a failure mode that was prevalent in baseline systems.

Taken together, these findings argue for a shift in how hallucinations are conceptualized in creative and engineering contexts. Rather than suppressing all deviations from factual correctness, we advocate treating hallucinations as low-prior abductive hypotheses subject to rigorous symbolic filtering. This approach preserves the exploratory power of LLMs while maintaining epistemic discipline through constraints, counter-abduction, and human oversight.

Beyond the specific results reported here, the paper contributes a reusable evaluation methodology and a new dataset paradigm for studying hallucination-driven creativity. The Hall2Invent benchmark and associated metrics provide a foundation for future work on safe creative AI, enabling systematic analysis of when and how hallucinations can be productively harnessed.

Future research will extend this framework to larger datasets, richer domain models, and interactive settings in which humans and LLMs co-evolve hypotheses over longer dialogues. We also plan to explore tighter integration with simulation, formal verification, and patent analysis pipelines. Ultimately, crossing the chasm from hallucination to invention requires not the elimination of uncertainty, but its careful governance—and abductive, constraint-based reasoning offers a principled path forward.

Limitations

The empirical evaluation relies on a relatively small curated dataset (Hall2Invent, 30 seed problems). While the framework demonstrates conceptual validity and encouraging trends, broader validation on larger, more diverse engineering and scientific domains is necessary to establish statistical robustness and generalizability.

Second, the abductive modeling assumes the availability of sufficiently rich background theory and well-specified integrity constraints. In real-world settings, domain knowledge may be incomplete, inconsistent, or costly to formalize. The success of hallucination-to-invention transformation therefore depends heavily on the quality and coverage of symbolic constraints, which may limit applicability in poorly structured domains.

Third, the framework presumes meaningful human oversight. The reinterpretation of hallucinations as hypotheses and the evaluation of inventive quality (feasibility, novelty, usefulness) involve expert judgment. Fully automating these evaluative steps remains challenging, particularly in highly specialized technical fields.

Acknowledgements

The author is grateful to Alexander Rubalov, Dmitry Ilvovsky, Vladimir Solodkin, and Ivan Trotsenko for fruitful discussions and dataset preparation. The article was prepared within the framework of the HSE University Basic Research Program.

References

Mahadevan S. 2022. Unifying causal inference and reinforcement learning using higher-order category theory. arXiv:2209.06262

Wang Z, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2024. Unleashing the Emergent Cognitive Synergy in Large Language Models: A Task-Solving Agent through Multi-Persona Self-Collaboration. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics, pages 257–279, Mexico City, Mexico.

Galitsky B. and Rybalov A. 2026a. A computational framework for analyzing and supporting invention dialogues. In Applications of Neuro-symbolic Artificial Intelligence, Springer.

Galitsky, B. and Rybalov, A. 2026b. Neuro-Symbolic Verification for Preventing LLM Hallucinations in Process Control. Processes, 14(2), 322.

Geroimenko, V. 2025. Generative AI: From Human–Computer Interaction to Human–Computer Creativity. In: Geroimenko, V. (eds) Human-Computer Creativity. Springer Series on Cultural Computing. Springer, Cham. https://doi.org/10.1007/978-3-031-86551-0_1

Bouschery S.G., Blazevic, F. and Piller T. 2023. Augmenting human innovation teams with artificial intelligence: Exploring transformer-based language models. J Prod Innovation Management. V 40, Issue 2, 139-153 https://doi.org/10.1111/jpim.12656

Passerini A, Gema A, Minervini P,Sayin B, Tentori K. 2025. Fostering effective hybrid human-LLM reasoning and decision making. Front. Artif. Intell., Sec. Machine Learning and Artificial Intelligence. V 7

https://doi.org/10.3389/frai.2024.1464690

Qin, S.J.; Badgwell, T.A. A survey of industrial model predictive control technology. Control Eng. Pract. 2003, 11, 733–764. 10.1016/S0967-0661(02)00186-7

Venkatasubramanian, V. The promise of artificial intelligence in chemical engineering: Is it here, finally? AIChE J. 2019, 65, 466–478. https://doi.org/10.1002/aic.16489.

Shi, X.; Xue, S.; Wang, K.; Zhou, F.; Zhang, J.; Zhou, J.; Tan, C.; Mei, H. Language Models Can Improve Event Prediction by Few-Shot Abductive Reasoning. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 29532–29557.

Kakas, A.C.; Mancarella, P. Generalized Abduction. J. Log. Comput. 1990, 1, 389–407.

Khan, A.; Nahar, R.; Chen, H.; Flores, G.E.C.; Li, C. Fault-Explainer: Leveraging large language models for interpretable fault detection and diagnosis. Comput. Chem. Eng. 2025, 199, 109152. 10.1016/j.compchemeng.2025.109152