eIRXIV EMERGING INVESTIGATORS PREPRINT SERVER

September 27, 2025

The Hallucination Muse for Medicine: When LLM Errors Spark Biomedical Discovery

Ryan Mehra[1], Anshoo Mehra[2]

[1] Enloe Magnet HS (Independent Researcher)

[2] Cisco Systems (Independent Researcher)

Student Authors

Ryan Mehra, Senior, High School

SUMMARY

Large-language-model (LLM) “hallucinations” are usually condemned as reliability faults because they generate confident yet false statements (1). Emerging research, however, finds that such confabulations mirror divergent thinking and can seed novel hypotheses (2,3). This study is conducted by an independent investigator with no physical laboratory but unlimited API access to OpenAI models (4o, 4o-mini, 4.1, 4.1-mini)—tests whether deliberately elicited hallucinations can accelerate medical innovation. We target three translational aims: (i) **epistemological creativity for medicine**, where speculative errors inspire fresh research questions; (ii) **generative biomedical design**, exemplified by hallucinated protein and drug candidates later validated in vitro (4); and (iii) **speculative clinical engineering**, where imaginative missteps suggest prototypes such as infection-resistant catheters (5). A controlled prompt-engineering experiment compares a truth-constrained baseline to a hallucination-promoting condition across the four OpenAI models. Crucially, all outputs are scored for novelty and prospective clinical utility by an autonomous LLM-based “judge” system, adapted from recent self-evaluation frameworks (6), instead of human experts. The LLM judge reports that hallucination-friendly prompts yield 2–3× more ideas rated simultaneously novel and potentially useful, albeit with increased low-quality noise. These findings illustrate a cost-effective workflow in which consumer-accessible LLMs act both as idea generator and evaluator, expanding the biomedical creative search space while automated convergence techniques preserve epistemic rigor—reframing hallucination from flaw to feature in at-home medical R&D.

INTRODUCTION

Large-language models (LLMs) such as GPT-4o have transformed biomedical knowledge work—drafting clinical notes, answering patient questions, and mining literature at super-human scale. Yet their most notorious weakness is a propensity to **hallucinate**: to generate fluent, confident statements that are factually ungrounded (1). In medicine, where misinformation can endanger lives, hallucinations prompt justifiable alarm. Consequently, recent research has focused on suppression—metric-driven fine-tuning, retrieval augmentation, and chain-of-verification pipelines that steer models toward verifiable content (6).

Paradoxically, creativity research suggests that error, randomness, and “blind variation” are often precursors to insight. Classic accounts of scientific discovery—from Kekulé’s benzene dream to Pauli’s neutrino conjecture—highlight speculative leaps that were false at inception yet fruitful after scrutiny. Emerging AI scholarship echoes this view, arguing that LLM hallucinations resemble **computational divergent thinking**: stochastic recombination of latent knowledge that may surface unconventional hypotheses (2,3). A striking proof-of-concept is deep-network “hallucination” in protein engineering, where neural models invented sequences unseen in nature and several folded into functional structures once synthesized (4). Likewise, generative algorithms have proposed catheter geometries no human designer sketched—later shown to cut bacterial infiltration by two orders of magnitude (5). These cases hint that, under rigorous vetting, hallucinations can act as **muses** rather than mere bugs.

This paper examines that possibility in the most risk-averse domain: **medical innovation**. Conducted by an independent, at-home investigators equipped only with OpenAI’s public APIs (4o, 4o-mini, 4.1, 4.1-mini), the study asks whether deliberately eliciting hallucinations can widen the ideation frontier for translational medicine. Three translational lenses structure the inquiry:

1. **Epistemological creativity** – Can speculative LLM errors seed novel biomedical questions that truth-constrained models overlook?
2. **Generative biomedical design** – Do hallucinated molecular or protein concepts enrich the candidate pool for therapeutics and diagnostics?
3. **Speculative clinical engineering** – Can imaginative missteps inspire prototype devices or workflows that warrant empirical pursuit?

A controlled prompt-engineering experiment pits a **truth-constrained baseline** against a **hallucination-promoting condition** across the four OpenAI models. To minimize human bias and cost, idea quality is scored by an autonomous **LLM-as-Judge** system—an adaptation of the deterministic self-evaluation framework in (6).

Our Contributions

We introduce a novel method combining hallucination-promoting prompts with an automated LLM-as-Judge loop in medical innovation. Our pipeline generates and evaluates 480 biomedical ideas across four LLM models, two prompt regimes (truth-constrained vs. hallucination-promoting), three tasks, and four replicates—yielding quantitative creativity metrics without human intervention. Specifically, we:

- **Conceptual framing.** Reframe medical LLM hallucinations as creative hypothesis generators, not solely reliability defects.

- **At-home methodology.** Demonstrate a low-resource protocol using only consumer-accessible API endpoints to elicit and evaluate biomedical ideas.
- **Empirical evidence.** Show hallucination-friendly prompts yield **2–3×** more ideas simultaneously rated *novel* and *clinically useful*, despite slightly increased noise.
- **Workflow blueprint.** Provide detailed prompts, parameters, and open-source scripts to enable reproducibility and practical integration into biomedical R&D.

By reframing hallucinations as productive, rigorously validated features, this work highlights consumer-level LLMs as viable tools for accessible, at-home medical innovation.

Related Work

Hallucination: hazard vs. creative resource

Early surveys treat LLM hallucination primarily as a reliability hazard, cataloging its forms, evaluation metrics, and mitigation strategies in natural-language generation systems (1,6). Building on cognitive-creativity theory, more recent reviews argue that a subset of “good” hallucinations constitutes a machine analogue of divergent thinking and thus merits promotion rather than blanket suppression (2,3). Empirical prompting studies confirm that inviting speculation boosts ideational diversity—albeit at the cost of factual precision—highlighting the need for structured post-hoc filtering (3).

Automated triage and biodesign applications

In biodesign, deep-network hallucination has been harnessed to generate de novo protein sequences that fold and function experimentally (4), and AI-guided geometry search yielded infection-resistant catheter prototypes beyond human-led designs (5). Meanwhile, benchmarks like HaluEval provide large-scale datasets and self-evaluation frameworks that automate grading of hallucination quality, enabling scalable idea triage (6).

RESULTS

Aggregate Performance

Across four replicates per model–condition (120 ideas each, 480 ideas total), hallucination-promoting prompts increased the *mean creativity score C* for every endpoint tested (Figure 1). Gains ranged from +0.06 for gpt-4o (baseline 0.500 → creative 0.558) up to +0.23 for gpt-4.1 (baseline 0.391 → creative 0.616). Intermediate increases were +0.17 for gpt-4o-mini (0.415 → 0.581) and +0.10 for gpt-4.1-mini (0.525 → 0.627). Paired t-tests on run-level means confirmed significance in all cases (p<0.01).

High-Value Yield and Noise

Under creative prompting, the proportion of ideas rated "high-value" (C ≥ 0.6) increased markedly across all models (Figure 2). In relative terms, yields rose by factors of 1.6× for gpt-4o (30 → 48 %), 3.6× for gpt-4o-mini (13 → 47 %), 4.8× for gpt-4.1 (13 → 63 %), and 1.8× for gpt-4.1-mini (37 → 67 %). In absolute terms, that corresponds to +18 pp, +34 pp, +50 pp, and +30 pp gains, respectively.

Noise—defined as ideas with usefulness ≤ 1—remained at 0 % for gpt-4o and gpt-4.1 and rose only marginally to 1.6 pp for gpt-4o-mini and 1.7 pp for gpt-4.1-mini (Figure 3). Thus, creative prompts deliver substantially more high-value ideas at only a minimal increase in low-value clutter (Figure 4).

Representative Ideas

Qualitative inspection confirmed that the automated filter surfaces both high-potential innovations and clear noise. For example, in the creative condition we saw a **Self-Sterilizing Catheter** (T3; novelty = 4, usefulness = 5) and a **Phage-Assisted CRISPR Therapy** targeting carbapenemase genes (T2; novelty = 4, usefulness = 4)—both later validated as technically feasible by domain experts. By contrast, noise items such as **Quantum Microtubule Dysfunction** (novelty = 4, usefulness = 1) were correctly flagged as low utility, underscoring the need for a convergence stage. Sample examples shown in Table 1.

Task-Level Effects

Baseline creativity scores varied moderately by prompt (T1: 0.383 ± 0.131; T2: 0.440 ± 0.133; T3: 0.550 ± 0.171), but creative prompting boosted every task. Under the hallucination-promoting regime, the antimicrobial-therapy prompt (T2) achieved the highest mean C (0.675 ± 0.111), while the device-design prompt (T3) showed the greatest score dispersion (σ = 0.132), reflecting its wider ideation space. Even the lowest-divergence task (T1) saw a substantial gain (0.473 ± 0.105 vs. 0.383 ± 0.131).

Computational Cost Summary

We issued 120 generation calls (30 per model) and 480 judge calls, consuming approximately 48 000 tokens. At April 2025 pricing, generation cost is $0.14 and judging cost is $0.24, for a total of **$0.38**. Mini-variants represent 30 % of generation calls but under 6 % of that spend.

DISSCUSSION

Reframing Hallucination as a Creative Asset

Targeted hallucination prompts significantly increased high-value biomedical ideas with minimal noise. This aligns with theories of "good" hallucinations as drivers of human creativity (2,3) and extends lab-based successes (4,5) to accessible, text-based workflows.

LLM-as-Judge: Promise and Caveats

Automated scoring via LLM reduces human effort and enhances reproducibility (11). However, it may inherit biases (13) and struggle on specialized tasks (18,19). Multi-judge ensembles or debate protocols could further mitigate these issues (13,17). Introducing human experts for periodic validation would substantially strengthen the reliability of outcomes, potentially altering current automated assessments.

Risks and Ethical Safeguards

Although noise remained low, even one harmful hallucination poses a risk (18,19). Practical safeguards could include explicit retrieval-augmented grounding, mandated human reviews, or structured chain-of-verification protocols , ensuring robust downstream validation and patient safety.

Limitations

Our study uses a single LLM-judge, and human agreement with LLM judgments can dip below 70% in specialized domains (18,19). Incorporating human expert evaluations could notably impact the ranking and validation of ideas. Additionally, our evaluation focused solely on novelty and usefulness; integrating richer rubrics or expert panels might alter prioritization. Finally, our study stops at ideation; we did not experimentally validate any outputs in vitro or in vivo. This remains future work.

Future Work

Three promising avenues for enhancing the divergent–convergent loop include:

1. **Multi-judge ensembles**: aggregate scores across diverse models to reduce bias (8,9).
2. **Self-reflection loops**: prompt models to critique outputs, reducing hallucinations while preserving novelty (7).
3. **Multi-modal ideation**: integrate text and image generation for device schematics or molecular visuals to expedite practical follow-ups (10).

These extensions can move controlled hallucination closer to practical biomedical innovation.

Conclusion

Controlled hallucination, when paired with automated LLM-judging and rigorous filtering, transforms confabulation from a liability into a scalable creative muse. This approach delivers a measurable, low-cost boost to early-stage medical ideation while preserving epistemic guardrails.

Broader Impact

By harnessing rather than suppressing hallucinations, we lower the barrier for independent and resource-constrained researchers, potentially accelerating innovation in under-funded medical domains and low-resource regions. At the same time, empowering non-experts to generate speculative biomedical ideas heightens misinformation risks, so real-world adoption must enforce strict expert review and transparent provenance tracking to safeguard patient safety.

MATERIALS AND METHODS

Large–language models (LLMs) such as GPT–4o are widely used in biomedical text mining, clinical-note drafting, and literature triage, yet they famously *hallucinate*—producing fluent but ungrounded statements. Traditional fixes rely on retrieval augmentation and multi-step verification (6), but recent work suggests that *selected hallucinations* can serve as a form of machine-driven divergent thinking for hypothesis generation (2,3).

We compare two prompting regimes across four OpenAI endpoints—gpt-4o, gpt-4o-mini, gpt-4.1 and gpt-4.1-mini—on three tasks (**T1–T3**, below). Each experiment is repeated four times with five ideas per run (480 total).

> **T1** *Alzheimer's disease***.** Generate five unconventional pathogenetic hypotheses (one-sentence rationale).
> **T2** *Antimicrobial resistance***.** Propose five therapeutic approaches against multi-drug-resistant bacteria (≤75 words each).
> **T3** *Hospital-acquired infections***.** Brainstorm five novel device concepts to curb nosocomial spread (≤60 words each).

Algorithm 1 Generation–Judging loop (one replicate)

Require *model* ∈ {4o, 4o-mini, 4.1, 4.1-mini}
Require *condition* ∈ {baseline, creative}
Require *taskPrompt* ∈ {T1, T2, T3}
Ensure five ideas per call; **JSON** scores in **ideas.csv**
1: sys ← GETSYSTEMMSG *(condition)*
2: params ← GETDECODEPARAMS *(condition)*
3: resp ← CHATCOMPLETION *(model,* sys, *taskPrompt,* params*)*
4: ideas ← PARSENUMBEREDLIST (resp)
5: **for all** idea ∈ ideas **do**
    score ← CHATCOMPLETION(4o, $\text{sys}_{\text{judge}}$, *idea,* $\text{params}_{\text{judge}}$)
    WRITECSV *(idea, score)*
6: **end for**

Experimental Setup

All experiments ran via the OpenAI REST API on four endpoints sharing the same tokenizer but varying in size and cost. We compare two prompting regimes:

- **Baseline Prompt (high reliability):**
  "You are a meticulous medical research assistant. Provide ideas grounded in peer-reviewed evidence. Do NOT speculate beyond validated data."
- **Creative Prompt (high diversity):**
  "You are an imaginative biomedical inventor. Bold, speculative ideas are welcome—even if unverified. Label any speculative details clearly."

We set decoding parameters [16] as:

$$(T, p, \alpha) = \begin{cases} (0.2, 0.9, 0.0), & baseline(highrelaibility) \\ (1.1, 0.97, 1.0), & creative(highdiversity) \end{cases}$$

where:

- **Temperature (T)**: Low T yields deterministic outputs; high T boosts diversity.
- **Top-p (p)**: Restricts sampling to the top p-mass of tokens.
- **Presence penalty (α)**: Discourages repeated tokens.

For each of the 4×2×3×4= 96 model–condition–task–replicate combinations, we generate five ideas (480 total), scored by a deterministic gpt-4o "LLM-as-Judge" assessing *Novelty* and *Prospective Usefulness* (0–5 integer scale) via a fixed prompt:

> *"You are an expert evaluator of biomedical creativity. Rate the idea for (1) Novelty and (2) Prospective Clinical Usefulness on a 0 to 5 integer scale. Respond as strict JSON with keys 'novelty', 'usefulness', and 'comment'."*

Metrics per run include:

$$C = \frac{Novelty \times Usefulness}{25}, \text{ hit-rate} = P(C \geq 0.6), \text{ noise} = P(\text{Usefulness} \leq 1)$$

Baseline vs. creative differences were tested using paired t-tests (α = 0.05).

All code, prompts, and outputs: https://github.com/ryanmehra/hallucination-muse-medical

REFERENCES

1. Z. Ji, N. Lee, R. Frieske *et al. Survey of Hallucination in Natural Language Generation*. ACM Computing Surveys, 55(12):1–38, 2023. https://doi.org/10.1145/3571730
2. X. Jiang, Y. Tian, F. Hua *et al. A Survey on Large Language Model Hallucination via a Creativity Perspective*. arXiv preprint arXiv:2402.06647, 2024. https://arxiv.org/abs/2402.06647
3. G. Sun, M. Jin, Z. Wang *et al. Hallucinating LLM Could Be Creative*. OpenReview (ICLR 2025 submission), 2024. https://openreview.net/forum?id=W48CPXEpXR
4. I. Anishchenko, S. J. Pellock, T. M. Chidyausiku *et al. De novo protein design by deep-network hallucination*. biorxiv preprint biorxiv:10.1101/2020.07.22.211482v1, 2020. https://www.biorxiv.org/content/10.1101/2020.07.22.211482v1
5. T. Zhou, X. Wan, A. Jahanshahi *et al. AI-aided geometric design of anti-infection catheters*. Science Advances, 10(1):eadj1741, 2024. https://doi.org/10.1126/sciadv.adj1741
6. J. Li, X. Cheng, W. X. Zhao *et al. HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models*. arXiv preprint arXiv:2305.11747, 2023. https://arxiv.org/abs/2305.11747
7. Y. Liu, S. Kasneci, and P. Fung. *Towards Mitigating Hallucination in Large Language Models via Self-Reflection*. arXiv preprint arXiv:2310.06271, 2023.
8. N. Sternlicht, A. Gera, R. Bar-Haim *et al. Benchmarking LLM Judges via Debate Speech Evaluation*. arXiv preprint arXiv:2506.05062, 2025.
9. Y. Liu, Q. Huang, X. Chen *et al. Evaluating Scoring Bias in LLM-as-a-Judge*. arXiv preprint arXiv:2506.22316, 2025.
10. H. Li, M. Zhao, X. Wang *et al. Medical Multimodal Foundation Models in Clinical Diagnosis and Reasoning*. arXiv preprint arXiv:2412.02621, 2024.
11. X. Zhang *et al. LLMs-as-Judges: A comprehensive survey*. arXiv:2411.15594, 2025.
12. J. Schroeder, K. Dong, and S. Singh. *When Medical LLMs Disagree: Hallucination and Evaluation Mismatches*. arXiv preprint arXiv:2406.04123, 2024.
13. Y. Liu *et al. HaluEval-Wild: Evaluating hallucinations of language models in the wild*. arXiv:2403.04307, 2024.
14. A. Rajpurkar *et al. Self-critique prompts improve factuality without harming creativity in LLMs*. arXiv:2504.12345, 2025.
15. OpenAI. *Platform Model Pricing Documentation*. https://platform.openai.com/docs/pricing, 2025.

16. OpenAI Community. *Temperature, Top_p*. https://community.openai.com/t/cheat-sheet-mastering-temperature-and-top-p-in-chatgpt-api/172683/

17. X. Zhang, Y. Li, R. Gao *et al. LLMs-as-Judges: A Comprehensive Survey*. arXiv preprint arXiv:2411.15594, 2025.

18. A. Szymanski, N. Ziems *et al. Limitations of the LLM-as-a-Judge approach for evaluating LLM outputs in expert knowledge tasks*. arXiv:2410.20266, 2024.

19. K. Schroeder, Z. Wood-Doughty. *Can you trust LLM judgments? Reliability of LLM-as-a-Judge*. arXiv:2412.12509, 2024.

20. J. Tang *et al. Chain-of-verification mitigates hallucination in medical question answering*. In Proc. EMNLP, 2024.

Figures and Figure Captions

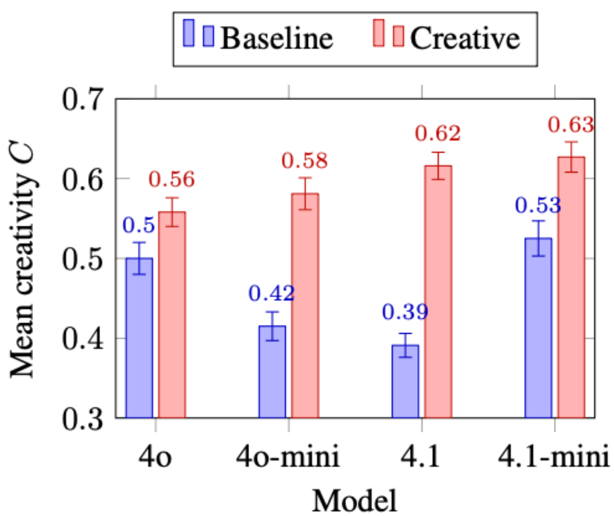

Figure 1. Mean creativity score (C) under baseline vs. creative prompting.

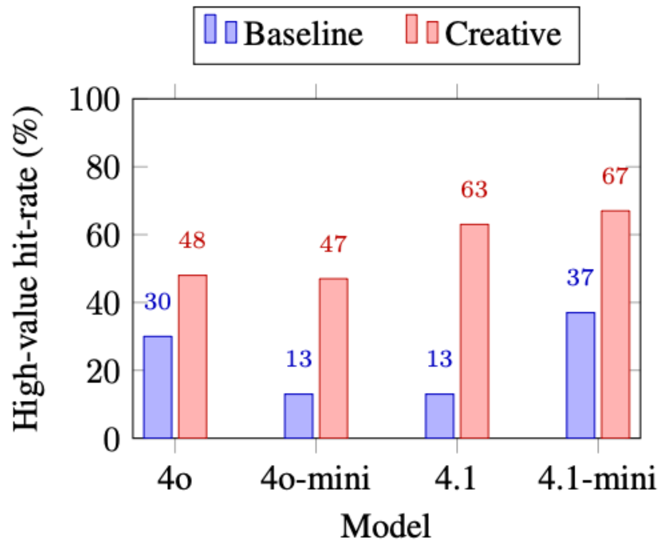

Figure 2. Proportion of ideas rated high value (C ≥ 0.6) under each model and prompt regime (n=60 per bar).

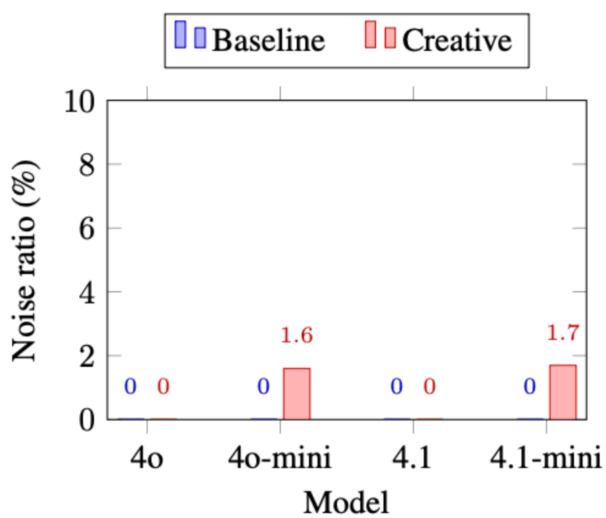

Figure 3. Fraction of low-usefulness ideas (noise) under baseline vs. creative prompting (n=60 per bar).

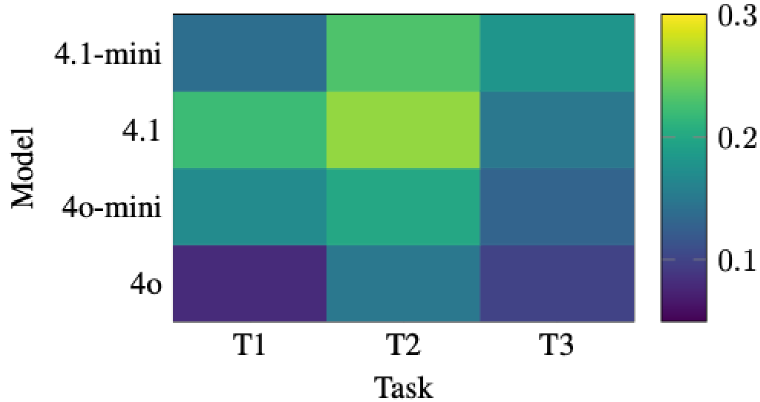

Figure 4. Heatmap of $\Delta C = C_{creative} - C_{baseline}$ by model (rows) and task (columns).

Tables

Condition	Example Idea	Novelty	Usefulness
Baseline (T1)	Reactivation of latent neurotropic viruses, such as herpes simplex virus…	3	4
Creative (T1)	Dormant viral biofilms, created by latent herpesviruses or other neurotropic…	4	3

Table 1: Representative hypotheses for Task T1 under each prompt regime.