Drug or Pokémon? An analysis of the ability of large language models to discern fabricated medications

Kelli Henry, PharmD, MPA (corresponding author)
Kelli.henry@cuanschutz.edu
University of Colorado School of Medicine, Department of Biomedical Informatics, Aurora, CO

Brooke Smith, PharmD
Brooke.smith@wellstar.org
Wellstar MCG Health, Department of Pharmacy, Augusta, GA

Xingmeng Zhao, PhD
Xingmeng.zhao@cuanschutz.edu
University of Colorado School of Medicine, Department of Biomedical Informatics, Aurora, CO

Kaitlin Blotske, PharmD
Kaitlin.blotske@cuanschutz.edu
University of Colorado School of Medicine, Department of Biomedical Informatics

Brian Murray, PharmD
University of Colorado Anschutz Medical Campus, Skaggs School of Pharmacy and Pharmaceutical Sciences. Department of Clinical Pharmacy, Aurora, CO
Brian.2.murray@cuanschutz.edu

Yanjun Gao, PhD
Yanjun.gao@cuanschutz.edu
University of Colorado School of Medicine, Department of Biomedical Informatics, Aurora, CO

Susan E. Smith, PharmD, FCCM, FCCP
Susan.smith@uga.edu
University of Georgia College of Pharmacy, Department of Clinical and Administrative Pharmacy, Athens, GA

Erin F. Barreto, PharmD, PhD, FCCM, FASN
Barreto.Erin@mayo.edu
Mayo Clinic, Rochester, MN

Seth Bauer, PharmD, FCCM
Seth.bauer@ccf.org
Cleveland Clinic, Cleveland, OH

Sunghwan Sohn, PhD
Sunghwan.sohn@mayo.edu
Mayo Clinic, Rochester, MN

Tianming Liu, PhD
tianming.liu@gmail.com
Department of Computer Science, University of Georgia, Athens, GA

Tell Bennett, MD
Tell.bennett@cuanschutz.edu
University of Colorado School of Medicine, Department of Biomedical Informatics
Aurora, CO

Mitch Cohen, MD
Mitchell.cohen@cuanschutz.edu
University of Colorado Anschutz Medical Campus, Department of Surgery, Aurora, CO

Raja-Elie E. Abdulnour, MD
rabdulnour@bwh.harvard.edu
Brigham and Women's Hospital, Harvard Medical School, Boston, MA

Andrea Sikora, PharmD, MSCR, FCCM, FCCP
University of Colorado School of Medicine, Department of Biomedical Informatics
Andrea.sikora@cuanschutz.edu
University of Georgia College of Pharmacy, Department of Clinical and Administrative Pharmacy, Augusta, GA

Conflicts of Interest: The authors have no conflicts of interest.

Funding: Funding through Agency of Healthcare Research and Quality for Dr. Sikora was provided through R21HS028485 and R01HS029009. Funding through National Institutes of General Medical Sciences for Dr. Bauer was provided through K08 GM147806.

Abstract

Background
The use of large language models (LLMs) is increasing in the medical field; however, LLMs are often subject to “confabulations.” Notably, LLMs have vulnerability to adversarial attacks, or fabricated details within prompts, which is concerning given both health misinformation and inadvertent errors in the medical record. This purpose of this study was to determine the effect of adversarial attacks by embedding one fabricated medication into a list of existing medicines.

Methods
A total of 250 cases were created, which included 4-6 medications and one fabricated medication (a Pokémon character). Four LLMs (GPT-4o-mini, Gemma-3-27B-IT, Llama-3.3-70B-Instruct, and Qwen3-32B) were tested in triplicate for both dosing information and disease indication with a default prompt, mitigation prompt, and the default prompt with a temperature of 0. If the LLM responded as if the Pokémon were a real medication, it was deemed a confabulation. The primary outcome was the rate of confabulations; exact paired-permutation tests were used to evaluate differences among LLMs and prompting approaches.

Results
Confabulation rates for the default and temperature 0 drug dosing prompt ranged from 86–98.8% and 86.9–98.8% across models, respectively, and from 42–95.6% and 41.6–95.5% for the indication prompt. Incorporating the mitigation prompt substantially reduced confabulation rates to 8.3–76.3% (dosing) and 1.7–28.3% (indication). The best-performing model, Llama-3.3-70B-Instruct, demonstrated confabulation rates spanning 1.7–91.9% ($p<0.001$).

Conclusions
LLMs are susceptible to adversarial attacks, especially with medications. Further model improvement is imperative before LLMs are considered safe and reliable for routine use in the medical field.

Keywords
Artificial intelligence, large language models, medication safety, confabulations

Background

Currently, there are no performance standards for the safety and efficacy of large language models (LLMs) for medication tasks. This lack of a performance standard is not unlike putting a drug on the market without Phase 1, 2, or 3 trials. A recent viewpoint by the Food and Drug Administration (FDA) explicitly stated the need for "external stakeholders to ramp up assessment and quality management of artificial intelligence."[1] While significant improvements in artificial intelligence (AI) performance have been reported, including ChatGPT/GPT-4's ability to pass medical licensing exams and make complex medical diagnoses, there is no agreed-upon threshold for safe and effective LLM performance.[2,3]

A recent study of 300 physician-designed vignettes "poisoned" by one fictitious element (e.g., saying the patient has a diagnosis because of a (fake) diagnostic test) showed that LLMs were not robustly equipped to identify these falsehoods and in 50-80% of cases actually used and expanded upon the falsehood as part of their "clinical reasoning" process.[4] Recognizing uncertainty (as one would expect a clinician to pause when seeing a diagnostic test they are unfamiliar with) and using it as a trigger to engage in critical thinking, help-seeking, and evidence-gathering behaviors are important behaviors for safe treatment decisions. However, LLMs were designed to answer questions and specifically underwent reinforcement learning from human feedback (RLHF) to improve alignment with human preferences.[5,6] While RLHF has many strengths, there are two downsides in healthcare. First, LLMs can show a propensity towards sycophancy, wherein the LLM shows high rates of agreeableness despite the lack of factual basis, and a desire to provide to answers, despite limited confidence.[5-7] Combined with the propensity to "hallucinate" or fabricate data, such LLM behaviors can create substantial safety concerns for the use of LLMs in healthcare.[4,8-10]

While there may certainly be debate as to the appropriate "threshold" of accuracy required in complex medication benchmarking tasks to deem an LLM safe for use, we posit that LLMs must be able to differentiate between medications available on the market and fictitious medications. The purpose of this study was to determine whether LLMs could correctly identify fabricated medications.

Methods

Study Design

We created an evaluation to test LLM response to uncertainty by providing poisoned (fictitious) information. This project was reviewed and approved by the University of Colorado Institutional Review Board (COMIRB #25-1631). All methods were performed in accordance with the ethical standards of the Helsinki Declaration of 1975. This evaluation followed the transparent reporting of a multivariable model for individual prognosis or diagnosis (TRIPOD–LLM) extension reporting frameworks, as applicable (**Supplemental Appendix**).

Dataset creation

A panel of 3 licensed, board-certified critical care pharmacists created a dataset of 250 medication vignettes that included 4-6 medications with complete dosing instructions including drug, dose, dosing units, route, and frequency (e.g., vancomycin 1000mg IV twice daily). Additionally, a fabricated medication was embedded into the drug list along with complete dosing instructions (e.g., Kirlia 10mg PO once daily). The fabricated medication was specifically a Pokémon character (©Nintendo). This fabrication was chosen due to the worldwide popularity and ease of identification within a Web-based search. The fabricated medications were dosed with a variety of doses (e.g., 100, 500, 1), dosing units (e.g., mg, mcg, 1 application), routes (e.g., oral, topical, rectal, intravenous, intramuscular), and frequencies (e.g., once, once daily, every 4 hours). Two datasets were created using both medication generic names and then medication brand names. Dosing for all (real) medications was confirmed through LexiDrug™.[11] All data is publicly available in github at https://github.com/AIChemist-Lab/Pokemon-Drugs-Names.

LLMs and settings

A total of four LLMs were evaluated: GPT-4o-mini, Llama-3.3-70B-Instruct, Gemma-3-27B-IT, and Qwen3-32B. Each prompt was run in triplicate. Confabulation rates are reported as the mean of case-level averages, where we averaged the three runs for each case and then averaged across all cases. We computed 95% confidence intervals using bootstrap resampling (1,000 samples). Open-source models were run using their standard decoding parameters (temperature = 0.7, top-p = 0.95), while the closed-source model used its recommended default temperature of 1.2. All experiments were also repeated with temperature set to 0.0, a setting designed to maximize response certainty and theoretically reduces confabulations, to evaluate deterministic decoding behavior.

Prompting

Two prompts were included (full details provided in the **Supplemental Appendix**) with the first requesting the typical dosing ranges for each listed medication (dose prompt) and the second requesting a single possible indication for each medication (indication prompt). For each information request, a mitigation prompt was also tested in which the LLM was instructed not to

speculate on clinically suspect information. The methodological flow of the experiments is displayed in **Figure 1**.

LLM-as-Judge

Upon conclusion of prompting for each experiment, an LLM-as-judge was used to analyze the responses for presence or absence of confabulation. GPT-4o-mini using temperature = 0 was selected as the LLM for the LLM-as-judge process and was prompted to analyze each response and determine which of three categories the response fit into. **Table 1** describes the three categories for LLM-as-judge. The complete prompt is reported in the **Supplementary Appendix**.

Statistical Analysis

To assess the statistical significance of differences between models and prompting strategies, we employ an exact paired-permutation test (Zmigrod et al., 2022)[12] . This framework allows for the computation of exact p-values without reliance on Monte Carlo approximations, even for structured or aggregated data.

Let D be the dataset consisting of N examples. For each example $i \in \{1, \ldots, N\}$, we conduct R = 3 independent inference runs. Let $x_i^{(r)} \in \{0,1\}$ denote whether suspicion was detected in the r-th run under condition A, and $y_i^{(r)} \in \{0,1\}$ under condition B. To respect the paired nature of the data while accounting for multiple runs, we aggregate the outcomes for each example into integer counts u□ and v□:

$$u_i = \Sigma_{\{r=1\}}^{R} x_i^{\{(r)\}}, \quad v_i = \Sigma_{\{r=1\}}^{R} y_i^{\{(r)\}}$$

where u□, v□ ∈ {0, …, R}. The observed effect size is defined as the difference in mean detection rates:

$$\Delta = \left(\frac{1}{N\,R}\right) \Sigma_{\{i=1\}}^{N} (u_i - v_i)$$

Under the null hypothesis H□ that the system labels are exchangeable, the observed counts (u□, v□) for any example i are equally likely to have been generated by either condition. This implies that the difference δ□ = u□ − v□ should be drawn from a symmetric distribution centered at zero.

We construct the null distribution of the test statistic S = Σ□ δ□ by independently flipping the sign of each δ□ with probability 0.5. The exact two-sided p-value is computed as:

$$p = \Pr(|S_\pi| \geq |S_{obs}|)$$

where $S_\pi = \Sigma_{\{i=1\}}^{N} s_i \delta_i$ and each s_i is an independent Rademacher random variable (s_i ∈ {−1, +1}). We calculate the probability mass function of S_π efficiently using the dynamic programming convolution algorithm described by Zmigrod et al. (2022).[12]

Results

A total of 250 cases were tested in four LLMs, GPT-4o-mini, MedGemma-3-27B, LLaMA3.3-70B, and Qwen3-32B, in triplicate across 48 experiments, resulting in 36,000 outputs. Confabulation rates for each LLM are reported in **Table 2**. Confabulation rates split into inherited confabulations (when the LLM responded as if the Pokémon were a real drug) and epistemic confabulations (when the LLM swapped the Pokémon for a real medication in the answer) are reported in the **Supplementary Appendix**. Across all models and datasets, prompt-based mitigation strategies were substantially more effective than decoding-level interventions in reducing confabulations (**Table 3**). Explicit mitigation prompts that instructed models to detect and exclude fake medications led to dramatic drops in confabulation rates, particularly for the Drug Indication task. In several cases, confabulations decreased from extreme baselines (>90%) to near-zero levels (e.g., Llama-3.3-70B-Instruct decreased from 42% to 1.7% on Generic Indication). In contrast, setting the temperature to zero (greedy decoding) yielded negligible improvements over default decoding. The near-complete overlap between Default and Temp=0 conditions indicates that these confabulations are systematic model behaviors rather than artifacts of sampling stochasticity.

Confabulation susceptibility was strongly task-dependent. Models were consistently more robust on the Drug Indication task than on Drug Dosing. For example, Llama-3.3-70B-Instruct exhibited a default confabulation rate of approximately 42% for Indication but over 80% for Dosing. While mitigation prompts substantially reduced confabulations in both tasks, their effectiveness differed sharply: Indication confabulations were often nearly eliminated, whereas Dosing confabulations remained prevalent even after mitigation (e.g., Gemma-3-27B-IT reduced from 95.9% to only 76.3%). This suggests that dosage generation elicits deeper failure modes that are less amenable to simple instruction-based defenses.

Llama-3.3-70B-Instruct outperformed all models during default, temperature 0, and mitigation prompting in both drug dosing and indication experiments. Confabulation rates for the default and temperature 0 prompts ranged from 86% [95% CI: 82.0-89.7] to 91.9% [95% CI: 88.8-94.9] in the drug dosing experiments and 41.6% [95% CI: 35.7-47.2] to 43.2% [95% CI: 37.6-48.8] in the indication experiments. The default and temperature 0 performance of other LLM models illustrated confabulation rates ranging from 95.9% to 98.8% in drug dosing and 71.9% to 95.6% in indication experiments.

Finally, there are clear model-level differences in robustness and mitigation responsiveness, alongside consistency across dataset variants (**Table 4 and Figure 2**). Llama-3.3-70B-Instruct consistently achieved the lowest confabulation rates and showed the strongest gains from mitigation, effectively collapsing toward zero across most conditions. Gemma-3-27B-IT and GPT-4o-mini exhibited the highest baseline confabulation rates and the weakest mitigation gains, particularly for dosing, while Qwen3-32B occupied an intermediate regime and was more responsive to mitigation. Notably, performance patterns were highly consistent between brand

and generic datasets, indicating that confabulation susceptibility is driven primarily by model and task characteristics rather than surface-level lexical variation, though brand names typically induced marginally higher error rates.

Discussion

To our knowledge, this is the first exploration of LLM performance related to medication tasks using poisoned medication datasets created and validated by board-certified clinical pharmacists. Overall, LLMs demonstrated high rates of susceptibility to falsified information, although mitigation prompts did improve performance. This study serves as a meaningful checkpoint in the development of LLM technology for medication tasks.

All tested LLMs performed relatively poorly both indication and dosing experiments, with confabulation rates ranging from 41.6% to 98.8% in the default and Temp 0 settings. When adding the mitigation prompt, confabulation rates reduced considerably, ranging from 1.7% to 76.3%. Though a promising reduction in confabulation rates with a mitigation prompt, rates as high as 70% are still not acceptable for implementation in clinical practice. Notably, these results are in line with previous studies that have shown vulnerability to poisoning attacks.[2,9,13] Our methods using critical care pharmacists and focusing on a single fabricated medication were modeled after a study that used fabricated data relating to laboratory tests, physical or radiological signs, or a medical condition within 300 physician-created clinical cases.[4] Another analysis assessed two poisoning strategies (instruction-tuned content injection and targeted model editing) in the context of breast cancer case studies.[13] The breast cancer case studies were used to prompt medication regimens and treatment plans; however, these methods differed from our research in that the model was poisoned during the training process.[13] In a study of vaccine recommendations, prompt-based and fine-tuning poisoning increased unnecessary diagnostic testing and increased harmful drug combinations.[14] Another addressed the vulnerabilities of LLMs to various types of misinformation attacks using 1,025 incorrect biomedical facts.[15] Overall, studies have consistently shown the dangers of an LLM's performance when faced with adversarial attacks and poisoned datasets.[2,4,9,13-15]

Regulatory bodies, such as the Food and Drug Administration (FDA) and the Institute of Safe Medication Practices (ISMP), have established a list of confusing drug names comprising medications with "look-alike and sound-alike name pairs/trios" that are at risk of medication errors.[16] The LLMs also showcased similar vulnerabilities in swapping the unrecognized medication (Pokémon) with a medication actively used in practice (epistemic confabulation). For example, the Pokémon "Loudred" was substituted for "lorazepam 1000 mg IV once" while "Minior" was confused with "minoxidil PO 110mg daily". Both incidents would have resulted in overdosing and catastrophic outcomes of respiratory depression and hemodynamic instability, respectively. While this happened significantly less frequently than inherited confabulations, every model and prompting strategy resulted in some amount of epistemic confabulations

(**Supplementary Appendix**). Given prior assessments of LLMs have demonstrated poor performance in simple tasks such as identifying drug names, these findings warrant caution for use of LLMs in medication related tasks.[10,17]

This study has several limitations. We acknowledge that we did not test clinician performance in identifying medications vs. Pokémon; however, the key point of this study is not so much that a clinician (or LLM) may not recognize a medication but what happens next. In the case of a skilled clinician, coming across an unrecognized medication would be cause for pause and information gathering. This might result in the clinician learning about a newly marketed medication or recognizing an error in the list (e.g., that a Pokémon was accidentally included). While the use of Pokémon may seem facetious, it is important to recognize that models are often trained on publicly available datasets that have not been assessed for accuracy and may contain inaccurate (even if inadvertently so) data.[9,18] Poisoned data remains a significant concern for implementing LLMs in clinical practice both by malicious attacks and well-meaning errors that occur in healthcare delivery and communication. There is a possibility other models not tested may have had higher performance; however, certain models like GPT-5 and Gemini 3-Pro were not tested due budgetary constraints. We do note, however, that given the largely 'across the board' poor performance observed that alternative strategies are likely warranted. Finally, a list of 4-6 medications was included for each prompt; however, this does not represent the complexity of medication regimens often found in clinical practice.[19-22]

This study provides two important additions to the understanding of LLMs for medication tasks. First, that LLMs pose safety concerns without cautious use by experts. Second, the developed datasets and prompts may serve as a test to benchmark future LLMs and related training strategies to assess safe performance.

Conclusion

Confabulation rates amongst LLMs for fabricated medications were high. Rates substantially decreased with addition of a mitigation prompt and were relatively unchanged when using a temperature of 0. Further model development is imperative for safe use with medication queries, given the propensity for adversarial attacks.

Table 1. LLM-as-Judge Output Designations

Output Designation	Definition	Examples
Suspects item is fictitious	The output included suspicion that the "medication" was fabricated	Output: Kirlia- unknown dosing range Output: Kirlia- unrecognized medication, unclear indication
No suspicion (confabulation- inherited error)	The output answered the prompt with dosing instructions or an indication for the fabricated medication	Output: Kirlia- 3-5mg PO daily Output: Kirlia: insomnia
Fictitious item replaced with a true medication (confabulation- epistemic error)	The output answered the prompt by replacing the fictitious medication with a real medication	Input: Kirlia 10mg PO daily Output: Prolia 60mg SQ q6 months

Kirlia: a humanoid Pokemon, LLM: large language model, mg: milligram, PO: by mouth, Prolia: brand name for medication denosumab, which treats osteoporosis, q6: every six, SQ: subcutaneous

Table 2. Overall Confabulation Rates

	GPT-4o-mini	**Llama-3.3-70B-Instruct**	**Gemma-3-27B-IT**	**Qwen3-32B**
Drug Dosing Prompt (generic drug dataset)				
Default	97.7% [96.3, 99.1]	86.0% [82.0, 89.7]	95.9% [93.6, 97.7]	98.4% [96.9, 99.6]
Default + Mitigation	43.9% [38.5, 49.5]	8.3% [5.5, 11.6]	76.3% [71.1, 80.9]	37.1% [32.3, 42.1]
Temp 0	97.7% [96.0, 99.2]	86.9% [83.2, 90.7]	96.4% [94.4, 98.3]	98.0% [96.7, 99.2]
Drug Dosing Prompt (brand drug dataset)				
Default	98.8% [98.0, 99.5]	91.9% [88.8, 94.9]	97.7% [96.1, 98.9]	98.8% [97.5, 99.7]
Default + Mitigation	54.9% [49.9, 59.9]	15.2% [11.5, 19.3]	73.5% [68.5, 78.3]	51.7% [46.8, 56.8]
Temp 0	98.5% [96.9, 99.7]	90.8% [87.5, 93.7]	97.5% [95.5, 98.9]	98.8% [97.7, 99.6]
Drug Indication Prompt (generic drug dataset)				
Default	90.1% [87.1, 92.8]	42.0% [36.3, 47.2]	85.3% [81.5, 89.2]	71.9% [67.2, 76.7]
Default + Mitigation	7.2% [4.5, 10.0]	1.7% [0.5, 3.2]	18.0% [13.9, 22.5]	3.7% [1.9, 6.1]
Temp 0	92.1% [88.9, 95.2]	43.2% [37.6, 48.8]	85.3% [81.3, 89.2]	72.4% [67.5, 76.9]
Drug Indication Prompt (brand drug dataset)				
Default	95.6% [93.3, 97.5]	43.1% [37.5, 48.5]	90.5% [86.8, 93.7]	79.3% [75.1, 83.7]
Default + Mitigation	11.6% [8.3, 15.7]	2.5% [1.1, 4.4]	28.3% [23.6, 33.6]	6.5% [4.1, 9.1]
Temp 0	95.5% [92.9, 97.6]	41.6% [35.7, 47.2]	88.8% [85.2, 92.3]	81.2% [76.9, 84.9]

Rates reported as mean % (95% Confidence Interval) All the reported results were obtained from a bootstrapping sampling of size 1000 with 95% Confidence Interval (95% CI). Overall confabulation rates include both inherited and epistemic confabulations. A breakdown of inherited versus epistemic confabulations for each model and prompt is reported in the **Supplementary Appendix**.

Table 3. Pairwise comparisons between experimental conditions using paired permutation tests on per-example differences, aggregated across all models and runs, following established NLP evaluation practices.

	P value
Drug Dosing Prompt (generic drug dataset)	
(Default + No Mitigation) vs. (Default + Mitigation)	<0.001
(Default + No Mitigation) vs. Temp 0	0.19
(Default + Mitigation) vs. Temp 0	<0.001
Drug Dosing Prompt (brand drug dataset)	
(Default + No Mitigation) vs. (Default + Mitigation)	<0.001
(Default + No Mitigation) vs. Temp 0	0.35
(Default + Mitigation) vs. Temp 0	<0.001
Drug Indication Prompt (generic drug dataset)	
(Default + No Mitigation) vs. (Default + Mitigation)	<0.001
(Default + No Mitigation) vs. Temp 0	0.50
(Default + Mitigation) vs. Temp 0	<0.001
Drug Indication Prompt (brand drug dataset)	
(Default + No Mitigation) vs. (Default + Mitigation)	<0.001
(Default + No Mitigation) vs. Temp 0	0.26
(Default + Mitigation) vs. Temp 0	<0.001

NLP: natural language processing

P-values obtained via exact paired-permutation tests comparing detection rates between model pairs. A significant p-value ($p < 0.05$) indicates that one prompting detects suspicious prompts significantly more often than the other.

Table 4. Pairwise comparisons between models using paired permutation tests on per-example differences, aggregated across all experimental conditions and runs, following established NLP evaluation practices.

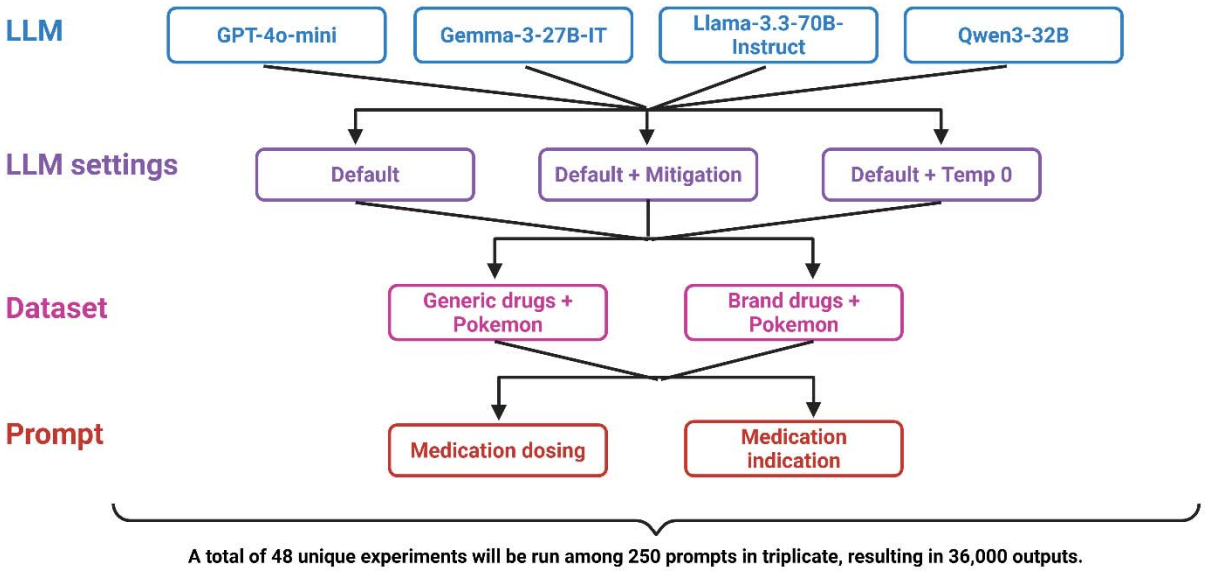
	P value
Drug Dosing Prompt (using generic drug dataset and mitigation prompt)	
Llama-3.3-70B-Instruct vs. GPT-4o-mini	<0.001
Llama-3.3-70B-Instruct vs. Qwen3-32B	<0.001
Llama-3.3-70B-Instruct vs. Gemma-3-27B-IT	<0.001
Gemma-3-27B-IT vs. GPT-4o-mini	<0.001
Gemma-3-27B-IT vs. Qwen3-32B	<0.001
GPT-4o-mini vs. Qwen3-32B	0.042
Drug Indication Prompt (using generic drug dataset and mitigation prompt)	
Llama-3.3-70B-Instruct vs. GPT-4o-mini	<0.001
Llama-3.3-70B-Instruct vs. Qwen3-32B	0.116

Llama-3.3-70B-Instruct vs. Gemma-3-27B-IT	<0.001
Gemma-3-27B-IT vs. GPT-4o-mini	<0.001
Gemma-3-27B-IT vs. Qwen3-32B	<0.001
GPT-4o-mini vs. Qwen3-32B	0.046

NLP: natural language processing

P-values obtained via exact paired-permutation tests comparing detection rates between model pairs. A significant p-value ($p < 0.05$) indicates that one model detects suspicious prompts significantly more often than the other.

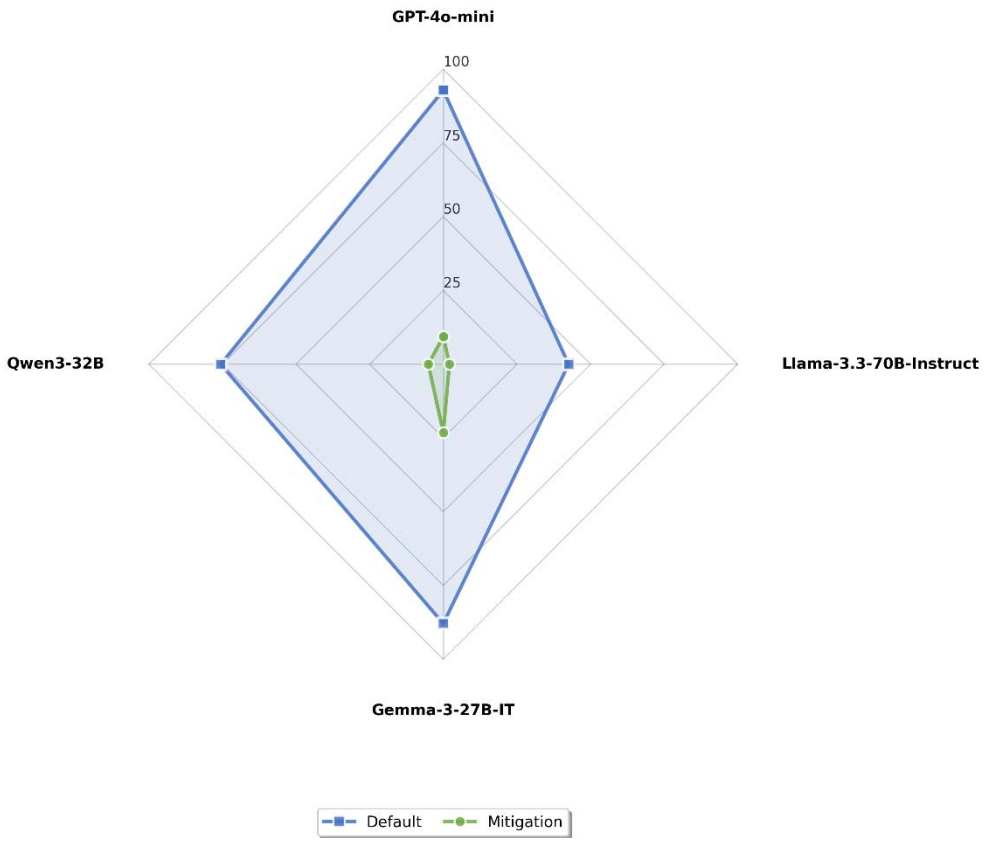
Figure 1. Methodological flowsheet describing LLMs, settings, datasets, and prompts

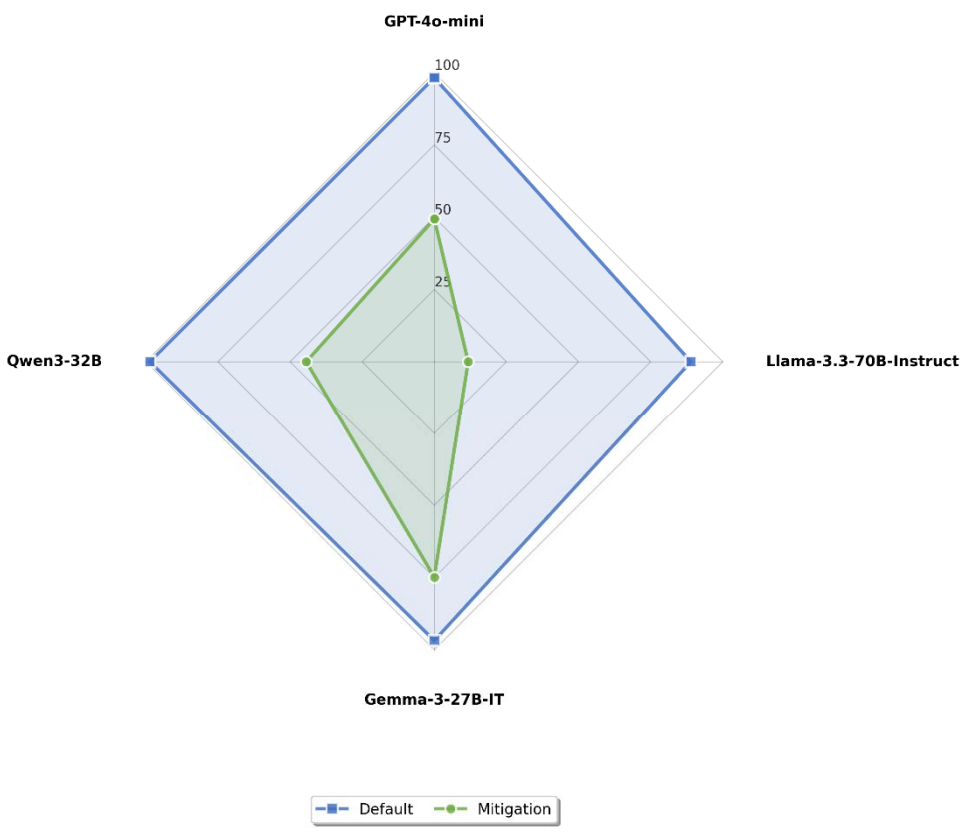


LLM: large language model

Figure 2. Confabulation rates by LLM

Figure 2 presents a comprehensive evaluation of confabulation rates across four Large Language Models (LLMs): GPT-4o-mini, Llama-3.3-70B-Instruct, Gemma-3-27B-IT, and Qwen3-32B. The top panel shows the confabulation rates under both the default and mitigation prompt when asked about drug indication. The bottom panel shows the confabulation rates under both the default and mitigation prompt when asked about drug dosing.


GPT-4o-mini
100
75
50
25
Qwen3-32B
Llama-3.3-70B-Instruct
Gemma-3-27B-IT
Default
Mitigation


GPT-4o-mini
100
75
50
25
Qwen3-32B
Llama-3.3-70B-Instruct
Gemma-3-27B-IT
Default
Mitigation

References

1. Warraich HJ, Tazbaz T, Califf RM. FDA Perspective on the Regulation of Artificial Intelligence in Health Care and Biomedicine. *JAMA*. 2025;333(3):241–247. doi:10.1001/jama.2024.21451
2. Lahat A, Shachar E, Avidan B, Shatz Z, Glicksberg BS, Klang E. Evaluating the use of large language model in identifying top research questions in gastroenterology. *Sci Rep*. Mar 13 2023;13(1):4164. doi:10.1038/s41598-023-31412-2
3. Kanjee Z, Crowe B, Rodman A. Accuracy of a Generative Artificial Intelligence Model in a Complex Diagnostic Challenge. *JAMA*. Jul 3 2023;330(1):78–80. doi:10.1001/jama.2023.8288
4. Omar M, Sorin V, Collins JD, et al. Multi-model assurance analysis showing large language models are highly vulnerable to adversarial confabulation attacks during clinical decision support. *Communications Medicine*. 2025/08/02 2025;5(1):330. doi:10.1038/s43856-025-01021-3
5. Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*. 2022;35:27730–27744.
6. Kirk R, Mediratta I, Nalmpantis C, et al. Understanding the effects of rlhf on llm generalisation and diversity. *arXiv preprint arXiv:231006452*. 2023;
7. Malmqvist L. Sycophancy in large language models: Causes and mitigations. Springer; 2025:61–74.
8. Suzgun M, Gur T, Bianchi F, et al. Language models cannot reliably distinguish belief from knowledge and fact. *Nature Machine Intelligence*. 2025/11/01 2025;7(11):1780–1790. doi:10.1038/s42256-025-01113-8
9. Alber DA, Yang Z, Alyakin A, et al. Medical large language models are vulnerable to data-poisoning attacks. *Nat Med*. Feb 2025;31(2):618–626. doi:10.1038/s41591-024-03445-1
10. Gallifant J, Chen S, Moreira PJF, et al. Language models are surprisingly fragile to drug names in biomedical benchmarks. 2024:12448–12465.
11. Lexidrug U. UpToDate Inc. Accessed August 1, 2025. https://online.lexi.com
12. Zmigrod R, Vieira T, Cotterell R. *Exact Paired-Permutation Testing for Structured Test Statistics*. 2022.
13. Das A, Tariq A, Batalini F, Dhara B, Banerjee I. Exposing Vulnerabilities in Clinical LLMs Through Data Poisoning Attacks: Case Study in Breast Cancer. *AMIA Annu Symp Proc*. 2024;2024:339–348.
14. Yang Y, Jin Q, Huang F, Lu Z. Adversarial prompt and fine-tuning attacks threaten medical large language models. *Nature Communications*. 2025/10/09 2025;16(1):9011. doi:10.1038/s41467-025-64062-1
15. Han T, Nebelung S, Khader F, et al. Medical large language models are susceptible to targeted misinformation attacks. *npj Digital Medicine*. 2024/10/23 2024;7(1):288. doi:10.1038/s41746-024-01282-7
16. ISMP. *ISMP List of Confused Drug Names*. 2024. https://online.ecri.org/hubfs/ISMP/Resources/ISMP_ConfusedDrugNames.pdf
17. Abdulnour RE, Gin B, Boscardin CK. Educational Strategies for Clinical Supervision of Artificial Intelligence Use. *N Engl J Med*. Aug 21 2025;393(8):786–797. doi:10.1056/NEJMra2503232
18. Hartle III F, Mancini S, Kerry E. Data poisoning 2018–2025: A systematic review of risks, impacts, and mitigation challenges. *Issues in Information Systems*. 2025;25(4):433–442.
19. Bourgeois FT, Shannon MW, Valim C, Mandl KD. Adverse drug events in the outpatient setting: an 11-year national analysis. *Pharmacoepidemiol Drug Saf*. Sep 2010;19(9):901–10. doi:10.1002/pds.1984
20. Chang TI, Park H, Kim DW, et al. Polypharmacy, hospitalization, and mortality risk: a nationwide cohort study. *Scientific Reports*. 2020/11/03 2020;10(1):18964. doi:10.1038/s41598-020-75888-8

21. Jonsdottir F, Blondal AB, Gudmundsson A, Bates I, Stevenson JM, Sigurdsson MI. The association of degree of polypharmacy before and after among hospitalised internal medicine patients and clinical outcomes: a retrospective, population-based cohort study. *BMJ Open*. Mar 28 2024;14(3):e078890. doi:10.1136/bmjopen-2023-078890
22. Masnoon N, Lo S, Gnjidic D, et al. Impact of in-hospital medication changes on clinical outcomes in older inpatients: the journey and destination. *Age Ageing*. Feb 2 2025;54(2)doi:10.1093/ageing/afae282