Drug or Pokémon? Large language model performance in identification of fabricated medications

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

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

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, Aurora, CO

Yanjun Gao, PhD
Yanjun.gao@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

Khoa Le, MS
Khoa.le1@uga.edu
Department of Computer Science, University of Georgia, Athens, GA

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

Hongjian Zhou
Hongjian.zhou@eng.ox.ac.uk
University of Oxford, Institute of Biomedical Engineering, Department of Engineering Science, United Kingdom

Fenglin Liu, PhD
Fenglin.liu@eng.ox.ac.uk
University of Oxford, Institute of Biomedical Engineering, Department of Engineering Science, United Kingdom

David A. Clifton, MEng, DPhil
davidc@robots.ox.ac.uk
University of Oxford, Institute of Biomedical Engineering, Department of Engineering Science, United Kingdom

Andrea Sikora, PharmD, MSCR, FCCM, FCCP (corresponding author)
University of Colorado School of Medicine, Department of Biomedical Informatics, Aurora, CO
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. Funding through NIHR Research Professorship, a Royal Academy of Engineering Research Chair, the InnoHK Hong Kong Centre for Cerebro-cardiovascular Engineering (COCHE), the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC), and the Pandemic Sciences Institute at the University of Oxford was provided to Dr. Clifton. Funding through the Clarendon Fund, the Department of Engineering Science Studentship, and the Frederick Brodckhues Scholarship was provided to Mr. Zhou. Funding through the Clarendon Fund was provided to Dr. Fenglin Liu. Funding through National Library of Medicine for Dr. Gao was provided through R00LM014308.

Abstract

Background
Large language models (LLMs) are increasingly used in medication-related tasks despite limited evidence supporting their accuracy and safety. LLMs are vulnerable to adversarial attacks and may confabulate in response to fabricated information in the prompt. Given the substantial risk of harm from medication errors, the purpose of this study was to establish a benchmark task to evaluate LLM performance in distinguishing fabricated from real medications.

Methods
Two datasets (brand and generic), each consisting of 250 medication lists were developed, each including 4-6 medications and one fabricated medication (specifically a Pokémon character), with complete dosing information provided. Six LLMs were evaluated (GPT-5-Chat, GPT-4o-mini, DrugGPT, Gemma-3-27B-IT, Llama-3.3-70B-Instruct, and Qwen3-32B). For each medication list, models were queried for dosing information and disease indication across three experimental conditions: standard decoding, standard decoding with mitigation prompting, and deterministic decoding (temperature = 0). Each experiment was performed in triplicate. Outputs were deemed confabulations if the LLM failed to recognize and reject the fabricated medication. The primary outcome was the rate of confabulations. Exact paired-permutation tests were used to compare confabulation rates across LLMs and prompting approaches.

Results
Confabulation rates under standard conditions ranged from 2.7% to 99.6%. DrugGPT demonstrated the lowest baseline confabulation rates (2.7-6.4%), whereas other models demonstrated substantially higher rates (47.2-99.6%). Mitigation prompting significantly reduced confabulation rates compared to standard and deterministic decoding across all tasks and datasets ($p < 0.001$): GPT-5-Chat achieved confabulation rates of 0% with mitigation prompting from baseline rates of 66.4-88.5%. Deterministic decoding did not substantially impact confabulation rates.

Conclusions
LLMs performing medication-related tasks are susceptible to confabulation in response to invalid or fabricated inputs, raising important safety concerns for clinical use. While mitigation prompting reduced confabulation rates, this study underscores the importance of robust safeguards when deploying LLMs in medication-related decision-making.

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

Introduction

Despite the impressive performance of large language models (LLMs) for certain healthcare tasks (e.g., passing medical licensing exams, diagnosing complex cases), their lack of reasoning transparency and tendency to produce confabulations represent substantial risks to patient safety.[1,2] Such confabulations, or hallucinations, involve generating plausible but incorrect or fabricated outputs not grounded in verified knowledge and are often difficult to detect, requiring end user vigilance.[3] There is currently no agreed-upon safety testing for LLM performance on medication tasks.[1,2,4] A recent viewpoint by the U.S. Food and Drug Administration (FDA) explicitly stated the need for "external stakeholders to ramp up assessment and quality management of artificial intelligence."[5]

Clinical complexity, characterized by uncertainty, incomplete information, and inaccurate inputs, is inherent to clinical practice.[6] In such environments, safe decision-making depends on the ability to identify errors and appropriately calibrate uncertainty.[6] In LLMs, this is represented by the ability to recognize when provided information is invalid or unverifiable and to refrain from generating unsupported outputs.[7] However, LLMs may demonstrate sycophantic behavior, favoring agreement with user-provided information over critical evaluation and accuracy, and can produce "confidently wrong" outputs rather than acknowledging uncertainty and signaling the need for human oversight.[7] When prompted with clinical vignettes intentionally "poisoned" by a single fictitious element (e.g., attributing a diagnosis to a nonexistent diagnostic test), LLMs accepted and expanded upon the fabricated information as part of their "clinical reasoning" process in 50-80% of cases.[8] Such behavior may propagate errors, with potential consequences for patient safety.

The medication use process is rife with errors; more than 200,000 patients are estimated to die each year from preventable medication errors, with associated healthcare costs exceeding $20 billion annually in the United States.[9] LLM failure modes may be especially harmful in this context, where LLMs may accept, propagate, or replace non-existent medications rather than recognize and reject them, directly influencing prescribing decisions. No standardized benchmark currently exists to evaluate whether LLMs can safely detect and flag non-existent medications in clinical contexts.

We propose that a foundational safety benchmark for LLMs in medication-related tasks is the ability to distinguish valid from non-existent medications in a clinical context and appropriately recognize and reject invalid inputs. To operationalize this benchmark, we evaluated confabulation rates when LLMs were presented with medication lists containing fabricated medications embedded among otherwise valid medications. Specifically, we evaluated whether LLMs can reliably discern between FDA-approved medications and Pokémon characters.

Methods

Study Design

We developed an evaluation framework to assess the robustness of contemporary LLMs to false premises by presenting them with intentionally poisoned (or fictitious) information. Specifically, LLMs were provided medication lists containing a single fabricated medication (in the form of a Pokémon character) among otherwise valid entries and queried for medication-related information. A confabulation was operationally defined as any response in which the model failed to identify and appropriately reject the fabricated medication. Examples of failures included it into its answer output or substituting the fabricated medication with an existing medication. This design was intended to establish a baseline benchmark against which future methods and innovations to improve performance on this task can be evaluated. 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.[10] This evaluation followed the transparent reporting of a multivariable model for individual prognosis or diagnosis (TRIPOD–LLM) extension reporting frameworks, as applicable (**Supplemental Appendix**).[11,12]

Dataset creation

A panel of three licensed, board-certified pharmacists created a dataset of 250 medication lists, each containing four to six medications with complete dosing instructions specifying drug name, dose, dosing units, route of administration, and dosing frequency (e.g., vancomycin 1000 mg intravenous twice daily). Subsequently, one fabricated medication was embedded within each medication list and assigned dosing instructions. Fabricated medication names were derived from Pokémon characters (©Nintendo), chosen because they share superficial similarities with

medication names (e.g., Kirlia vs. Kymriah), yet they are a widely recognized cultural phenomena and easily verifiable through web-based searches. Fabricated medications were assigned a range of plausible doses (e.g., 1, 100, 500), dosing units (e.g., mg, mcg, one application), routes of administration (e.g., oral, topical, rectal, intravenous, intramuscular), and dosing frequencies (e.g., once, once daily, every 4 hours). Two datasets were created: one using medication generic names and the other using brand names. Dosing for all real medications was verified using LexiDrug™.[13] All data are publicly available at https://github.com/AIChemist-Lab/Pokemon-Drugs-Names.

Models and Experimental Conditions

A total of six LLMs were evaluated: GPT-5-Chat, GPT-4o-mini, DrugGPT, Llama-3.3-70B-Instruct, Gemma-3-27B-IT, and Qwen3-32B. Each prompt was run in triplicate. Open-source models were run using their standard decoding parameters (temperature = 0.7, top-p = 0.95), while the closed-source models (GPT-4o-mini and GPT-5-Chat) used their recommended default temperature of 1.2.

We evaluated three experimental conditions: 1) standard decoding parameters; 2) standard decoding with a mitigation prompt instructing the model not to speculate on clinically suspicious information; and 3) deterministic decoding with temperature set to 0.0, a setting designed to maximize response certainty and theoretically reduce confabulations.

Tasks and Prompting

Two prompts requesting drug information about the medication regimens were evaluated. The first prompt requested the typical dosing range for each listed medication (dosing task), and the second requested a single plausible indication for each medication (indication task). The full

prompt text is available in the **Supplemental Appendix**. The overall experimental workflow, displayed in **Figure 1**, resulted in 12 experiments per model.

Additionally, a retrieval augmented generation (RAG) approach was implemented for five models (GPT-5-Chat, GPT-4o-mini, Llama-3.3-70B-Instruct, Gemma-3-27B-IT, and Qwen3-32B) using two external online data sources (RxNorm and OpenFDA's Drug Label API) to attempt to reduce confabulations: this approach enabled models to cross-reference medication names against structured drug databases prior to generating responses. Methodological details are provided in the **Supplemental Appendix**.[14,15]

Outcome Assessment

The primary outcome was confabulation rate. Secondary outcomes included confabulation rates under varying experimental conditions and interventions (i.e., mitigation prompting, temp 0.0, RAG) and confabulation subtype classification. Model outputs were evaluated using an LLM-as-judge approach. GPT-5-Chat using temperature = 0.0 was selected as the judging model. The complete LLM-as-judge prompt is reported in the **Supplementary Appendix**. Each response was analyzed and assigned to one of three predefined outcome categories (i.e., no confabulation, inherited confabulation, epistemic confabulation) as described in **Table 1**. To assess reliability, a subset of 240 outputs was independently reviewed by clinicians. Outcomes are reported in the **Supplementary Appendix**.

Statistical Analysis

Confabulation rates are reported as the mean of case-level averages. For each case, results were averaged across three independent runs and then averaged across all cases. We computed 95% confidence intervals (CI) using bootstrap resampling (1,000 samples). To assess the statistical significance of differences between models and prompting strategies, we employed

an exact paired-permutation test.[16] This framework facilitated the computation of exact p-values without reliance on Monte Carlo approximations. The null hypothesis is that in the absence of true semantic understanding, the model's labeling behavior is exchangeable between drug names and non-medical names. Therefore, any observed difference reflects random variation rather than a systematic preference. Additional methods are described in the **Supplementary Appendix**.

Results

A total of 54,000 outputs were produced across 12 experiments. Confabulation rates ranged from 2.7% to 99.6% using standard decoding parameters without mitigation (**Table 2**). DrugGPT had the lowest confabulation rates (2.7%-6.4%), and GPT-4o-mini had the highest confabulation rates (92.7%-99.6%), respectively. Overall, baseline confabulation rates were substantial across most models, indicating frequent failure to detect fabricated medications.

Across all tasks and datasets, explicit mitigation prompts instructing models to limit responses to established medical knowledge and to acknowledge uncertainty rather than speculate reduced confabulation rates compared to standard and deterministic decoding ($p < 0.001$, **Table 3**). In several cases, mitigation prompting reduced confabulations from high baseline levels to approaching zero; GPT-5-Chat achieved a 0% confabulation rate across all tasks and datasets from baseline rates ranging from 66.4% to 88.5%, while the other models produced confabulation rates ranging from 2.0% to 76.5% under mitigation prompting (**Table 2**). In contrast, deterministic decoding (temperature = 0.0) did not produce substantial numerical improvements, although a statistically significant improvement was demonstrated for the dosing task on the generic dataset when compared to standard decoding.

Confabulation susceptibility varied by task. Models were generally more robust with the indication task compared to the dosing task. For example, Llama-3.3-70B-Instruct exhibited a default confabulation rate of approximately 47% for indication prompts but over 85% for dosing prompts. While mitigation reduced confabulations in both tasks, task-level differences persisted. Indication confabulations were often nearly eliminated following mitigation, whereas dosing confabulations remained common even after mitigation (e.g., Gemma-3-27B-IT confabulation rate decreased from 94.9% to 76.5% on the generic dosing task).

Significant model-level differences in susceptibility to confabulation were observed across all experimental conditions (**Table 4** and **Figure 2**). GPT-5-Chat had relatively average performance under standard decoding parameters and temperature 0.0, with confabulation rates ranging from 64.3% to 90.0% but had a 0% confabulation rate for all prompts with mitigation. DrugGPT had the best performance under standard parameters and temperature 0.0 (confabulation rates from 2.4% to 7.1%) but was the only model that did not demonstrate substantially reduced confabulation rates under mitigation prompting (2.9% to 4.5%). Gemma-3-27B-IT and GPT-4o-mini exhibited the highest baseline confabulation rates and aside from DrugGPT were the least responsive to mitigation prompting, particularly for the dosing task, while Qwen3-32B and Llama-3.3-70B-Instruct occupied an intermediate regime and were more responsive to mitigation prompting. Using a RAG approach, confabulations were improved for all prompts (ranging from 0-19.7%), with the brand drug dataset generally having a higher confabulation rate compared to the generic drug dataset (**Supplementary Appendix**). This approach uniformly resulted in lower confabulation rates across different LLMs, with GPT-5-Chat, GPT-4o-mini, Llama-3.3-70B-Instruct, and Qwen3-32B performing similarly (confabulation rates 0-9.3%).

Results stratified by confabulation subtypes are provided in the **Supplementary Appendix**. Approximately 90% of overall recorded confabulations were categorized as inherited errors in which the fabricated medication was not identified. For individual experiments, 50% to 100% of confabulations were classified as inherited errors. Performance patterns were consistent across generic and brand dataset variants. Across models, brand names were associated with marginally higher confabulation rates, but relative performance rankings and mitigation effects were preserved. LLM-As-Judge validation performed by pharmacists showed high accuracy in

the classification of confabulation types, with rates ranging from 90-97% for each experiment. Additional specifics of LLM-as Judge results are available in the **Supplemental Appendix.**

Discussion

To our knowledge, this is the first study to systematically evaluate the ability of contemporary LLMs to identify and reject fabricated inputs in medication-related tasks. Baseline performance was generally poor, with high confabulation rates indicating frequent failure to detect invalid medications. Mitigation prompting that instructed models to limit responses to established medical knowledge and to acknowledge uncertainty rather than speculate substantially reduced confabulation rates, although residual error persisted in most experiments. These findings have immediate implications for the safe clinical deployment of LLMs.

Recognizing a real medication from a fictitious one is a fundamental clinical task, and the failure of widely available LLMs to reliably perform this task raises significant patient safety concerns. While it is notable that various methods reduced these errors, the importance of these findings is that this type of testing and associated methodological modifications are essential for the safe use of this technology. Our results are consistent with previous studies demonstrating LLM vulnerability to fabricated or invalid inputs but build upon this work through focused evaluation in the treatment domain, which requires distinct knowledge and reasoning compared to diagnostic tasks.[4,17,18] We specifically modeled our approach after a prior study using 300 physician-created clinical cases containing fabricated elements (e.g., laboratory values, physical or radiological signs, medical conditions), which similarly demonstrated high rates of error propagation.[4,8,17-20] Our results are also consistent with other studies of LLM benchmarks for medication-related tasks.[21-26] In those studies, LLMs struggled with tasks such as matching drug names with available doses or identifying drug-drug interactions, tasks that are expected to be completed with high proficiency in the first year of the Doctor of Pharmacy curriculum.[21-26]

Ultimately, this type of evaluation may serve as a valuable early benchmark or checkpoint in the pre-clinical development of AI systems to help ensure safer deployment in patient care settings.

Mitigation prompts reduced confabulation rates considerably, indicating that the problem is a behavioral tendency to generate rather than a true knowledge deficit. Indication confabulations were nearly eliminated by mitigation, but dosing confabulations persisted, suggesting that dosage generation elicits deeper failure modes that are less amenable to simple instruction-based defenses. Setting model temperature to 0.0 (i.e., deterministic decoding) did not meaningfully reduce confabulation rates for most LLMs, and similar error patterns were observed across runs. These findings suggest that confabulations reflected systematic model behaviors rather than artifacts of sampling stochasticity. The use of a RAG approach further reduced confabulation rates and resulted in more uniform performance across models. Even so, residual confabulations persisted across all interventions, highlighting the need for multi-layered and complementary approaches to improving LLM reliability in clinical applications. While it appears that there may be solutions to solving this safety issue, it is important to emphasize that a responsible clinician, when faced with an unrecognized medication, would not make assumptions without subsequent information gathering (e.g., a Web search to learn about a newly available medication or clarifying with the patient or their pharmacy about the medication list).[27] In clinical practice, recognizing uncertainty and using it as a trigger to engage in critical thinking, help-seeking, and evidence-gathering behaviors are essential for making safe treatment decisions.[6,22,28] This concept is reflected in where human learners are taught well-defined core competencies and assessed against observable entrustable professional activities.[29] Applying such a development and evaluation framework to AI systems can ensure their safety and efficacy in the clinical workplace,[30-33] akin to licensing of medical practitioners.[34]

The majority of errors observed were inherited confabulations, in which models failed to identify the presence of a fabricated medication in the medication list and generated plausible but incorrect outputs. Epistemic confabulations, a pattern completion behavior in which a real medication used in clinical practice was substituted for the unrecognized fabricated medication, occurred less frequently but still produced clinically dangerous information. For example, the Pokémon "Loudred" was substituted with "lorazepam 1000 mg IV once," while "Minior" was replaced by "minoxidil PO 110mg daily." Both incidents represent potentially catastrophic dosing errors. This behavior indicates limited capacity for recognizing uncertainty or invalid inputs in medication data. In short, the LLMs do not say "I don't know."[6] The implications of this finding in clinical practice are significant, as misspelled or phonetically rendered medication names could produce significant medication errors. Real-world data used for training will inevitably include errors: one study found that 21% of patients who read their personal electronic health record identified errors, of which 42% were deemed serious.[35] Encountering poisoned data at the time of model training, whether through malicious attacks or unintentional yet routine errors that occur in healthcare delivery and communication, may lead to incorporation of that poisoned data into the LLM reasoning process.[17,36] Beyond this, regulatory bodies such as the FDA and the Institute of Safe Medication Practices routinely emphasize the risks posed by look-alike and sound-alike medication names and highlight the need for safeguards to reduce inadvertent substitutions.[37] Yet, LLMs were designed to generate answers and were specifically optimized through reinforcement learning from human feedback (RLHF) to improve alignment with human preferences.[38,39] While RLHF has many strengths, it may also promote sycophancy (i.e., excessive agreeableness despite limited factual basis) and overly confident responses even in the setting of uncertainty.[38-40] Combined with the propensity to confabulate, these behaviors

raise substantial safety concerns for the use of LLMs in healthcare, and task-specific performance highlights the need to evaluate AI models using a battery of benchmarking tests mapped to observable tasks..[8,17,21,41] The development of benchmarks that can stand as a bulwark against these tendencies may prove to be an essential domain for health AI.

This study has several limitations. We did not evaluate clinician performance in identifying fabricated medications and therefore cannot directly compare model and human responses; however, the ability to recognize when a medication is unfamiliar or potentially invalid is a core competency in health degree programs. The use of Pokémon characters as fabricated medication names represents a controlled but synthetic test condition that may not fully capture the complexity of real-world clinical inputs, where invalid inputs may be more subtle or context dependent. Each medication list contained four to six medications and exactly one fabricated medication, which underestimates the complexity of medication regimens and potential errors encountered in clinical practice, and future evaluations may benefit from more complex cases.[42-45] All medication lists contained a fabricated medication, preventing evaluation of false positives. Model responses were constrained to a structured output format which, while ensuring consistency of outputs, did not allow for models to state “I don’t know” or ask follow-up questions. Evaluated strategies represent only a subset of potential approaches intended to improve LLM performance. Outputs were evaluated using an LLM-as-judge approach, which may introduce bias despite clinician validation of a random sample of outputs. Finally, although multiple leading contemporary LLMs were evaluated, there is a possibility other models may have had higher performance as reasoning models were not tested in this evaluation. Ultimately, additional strategies to reduce the consistently high confabulation rates are required, with this study serving as a proposed benchmark to test such future development.

This study provides two important additions to the understanding of LLMs for medication tasks. First, 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

In an evaluation of contemporary LLMs, confabulation rates in the presence of fabricated medications were high. Prompt-based mitigation strategies substantially reduced errors, but clinically concerning confabulations persisted across most models and tasks. These findings indicate that current LLMs may not meet consistently meet foundational safety expectations in medication-related applications. Further model development and standardized evaluation frameworks that explicitly test robustness to invalid inputs are essential for safe use in medication-related tasks.

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-5-Chat	GPT-4o-mini	Llama-3.3-70B-Instruct	Gemma-3-27B-IT	Qwen3-32B	DrugGPT
Drug Dosing Prompt (generic drug dataset)						
Standard decoding parameters	83.1% [78.9, 87.1]	99.5% [98.8, 100.0]	87.2% [83.3, 90.9]	94.9% [92.5, 97.2]	94.9% [92.8, 96.8]	2.7% [1.5, 4.1]
Standard decoding parameters + mitigation	0.0% [0.0, 0.0]	46.5% [41.3, 52.1]	6.8% [4.1, 9.6]	76.5% [71.7, 81.1]	48.4% [43.3, 53.5]	2.9% [1.9, 4.3]
Temperature 0.0	85.1% [80.8, 89.1]	98.9% [97.6, 99.9]	87.1% [83.6, 90.3]	95.6% [93.3, 97.7]	93.6% [90.9, 95.9]	2.4% [1.3, 3.7]
Drug Dosing Prompt (brand drug dataset)						
Standard decoding parameters	88.5% [84.8, 91.6]	99.6% [99.1, 100.0]	94.1% [91.6, 96.5]	98.4% [97.1, 99.5]	96.8% [94.8, 98.5]	3.3% [2.1, 4.7]
Standard decoding parameters + mitigation	0.0% [0.0, 0.0]	57.2% [52.3, 62.1]	14.3% [10.1, 18.1]	76.1% [71.2, 80.9]	61.5% [56.4, 66.0]	4.5% [2.9, 6.4]
Temperature 0.0	90.0% [86.7, 93.1]	99.5% [98.4, 100.0]	93.7% [91.1, 96.1]	98.9% [97.7, 99.9]	96.5% [94.5, 98.1]	6.7% [4.8, 8.5]
Drug Indication Prompt (generic drug dataset)						
Standard decoding parameters	66.4% [61.1, 71.7]	92.7% [90.1, 95.1]	48.1% [42.7, 53.9]	83.7% [79.7, 87.9]	70.0% [65.3, 74.5]	5.3% [3.6, 7.3]
Standard decoding parameters + mitigation	0.0% [0.0, 0.0]	8.3% [5.1, 11.6]	2.0% [0.8, 3.5]	18.4% [14.1, 22.8]	5.1% [2.7, 7.7]	4.4% [2.7, 6.4]
Temperature 0.0	64.3% [58.5, 69.7]	94.9% [92.1, 97.3]	50.1% [44.7, 55.6]	83.6% [79.3, 87.7]	72.1% [67.3, 76.7]	5.9% [4.0, 8.0]
Drug Indication Prompt (brand drug dataset)						
Standard decoding parameters	80.1% [75.7, 84.3]	96.0% [93.7, 97.7]	47.2% [41.6, 52.7]	90.3% [86.5, 93.5]	81.1% [77.2, 85.2]	6.4% [4.1, 8.7]
Standard decoding parameters + mitigation	0.0% [0.0, 0.0]	12.3% [8.8, 16.3]	2.5% [1.1, 4.4]	28.8% [24.1, 34.0]	6.8% [4.5, 9.3]	3.6% [2.1, 5.3]
Temperature 0.0	80.1% [75.2, 84.5]	96.0% [93.5, 98.0]	46.8% [41.3, 52.9]	88.8% [85.2, 92.1]	82.9% [78.8, 86.7]	7.1% [4.9, 9.3]

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 of confabulation rates 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)	
Standard decoding parameters vs. Standard decoding parameters + Mitigation	<0.001
Standard decoding parameters vs. Temperature 0.0	0.046
Standard decoding parameters + Mitigation vs. Temperature 0.0	<0.001
Drug Dosing Prompt (brand drug dataset)	
Standard decoding parameters vs. Standard decoding parameters + Mitigation	<0.001
Standard decoding parameters vs. Temperature 0.0	0.176
Standard decoding parameters + Mitigation vs. Temperature 0.0	<0.001
Drug Indication Prompt (generic drug dataset)	
Standard decoding parameters vs. Standard decoding parameters + Mitigation	<0.001
Standard decoding parameters vs. Temperature 0.0	0.803
Standard decoding parameters + Mitigation vs. Temperature 0.0	<0.001
Drug Indication Prompt (brand drug dataset)	
Standard decoding parameters vs. Standard decoding parameters + Mitigation	<0.001
Standard decoding parameters vs. Temperature 0.0	0.559
Standard decoding parameters + Mitigation vs. Temperature 0.0	<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 of confabulation rates 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)	
GPT-5-Chat vs. Llama-3.3-70B-Instruct	<0.001
GPT-5-Chat vs. GPT-4o-mini	<0.001
GPT-5-Chat vs. Qwen3-32B	<0.001
GPT-5-Chat vs. Gemma-3-27B-IT	<0.001
GPT-5-Chat vs. DrugGPT	<0.001
DrugGPT vs. Llama-3.3-70B-Instruct	<0.001
DrugGPT vs. GPT-4o-mini	<0.001
DrugGPT vs. Qwen3-32B	<0.001
DrugGPT vs. Gemma-3-27B-IT	<0.001
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	0.264
Drug Indication Prompt (using generic drug dataset and mitigation prompt)	
GPT-5-Chat vs. Llama-3.3-70B-Instruct	0.062
GPT-5-Chat vs. GPT-4o-mini	<0.001
GPT-5-Chat vs. Qwen3-32B	<0.001
GPT-5-Chat vs. Gemma-3-27B-IT	<0.001
GPT-5-Chat vs. DrugGPT	<0.001
DrugGPT vs. Llama-3.3-70B-Instruct	<0.001
DrugGPT vs. GPT-4o-mini	<0.001
DrugGPT vs. Qwen3-32B	<0.001
DrugGPT vs. Gemma-3-27B-IT	<0.001
Llama-3.3-70B-Instruct vs. GPT-4o-mini	<0.001
Llama-3.3-70B-Instruct vs. Qwen3-32B	0.013
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	0.088

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, datasets, interventions, and tasks

LLM: large language model

Figure 2. Confabulation rates by LLM

Figure 2 presents a comprehensive evaluation of confabulation rates across six Large Language Models (LLMs): GPT-5-Chat, GPT-4o-mini, DrugGPT, Llama-3.3-70B-Instruct, Gemma-3-27B-IT, and Qwen3-32B. The left panel shows the confabulation rates under both the default and mitigation prompt when asked about drug dosing using the generic dataset. The right panel shows the confabulation rates under both the default and mitigation prompt when asked about drug indication using the generic dataset.

References

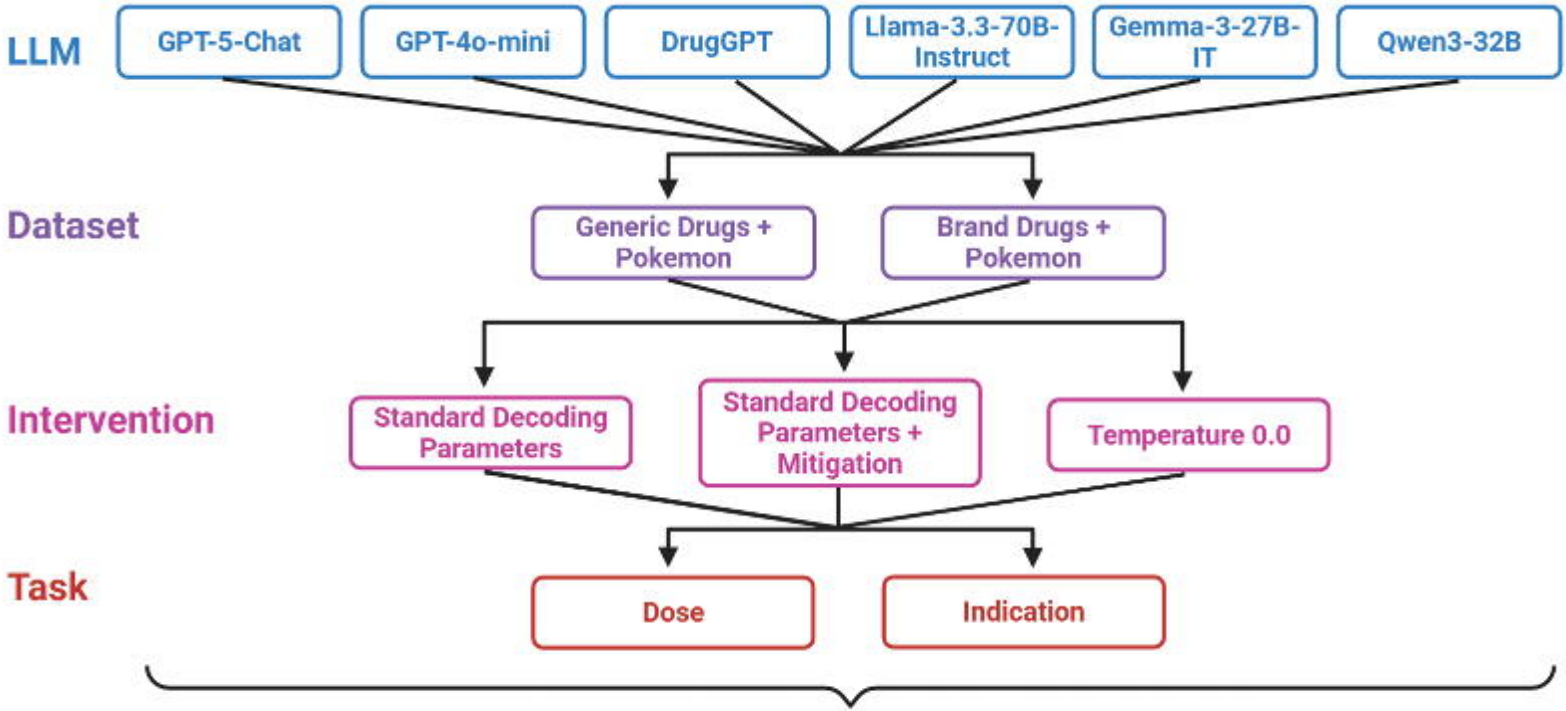
1. Kanjee Z, Crowe B, Rodman A. Accuracy of a Generative Artificial Intelligence Model in a Complex Diagnostic Challenge. JAMA 2023;330(1):78-80. DOI: 10.1001/jama.2023.8288.
2. Gilson A, Safranek CW, Huang T, et al. How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment. JMIR Med Educ 2023;9:e45312. DOI: 10.2196/45312.
3. Omar M, Sorin V, Collins JD, et al. Multi-model assurance analysis showing large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support. Commun Med (Lond) 2025;5(1):330. DOI: 10.1038/s43856-025-01021-3.
4. 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 2023;13(1):4164. DOI: 10.1038/s41598-023-31412-2.
5. 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.
6. Sikora A CL, Abdulnour RE. Can ChatGPT Say ‘I don’t know’? New England Journal of Medicine 2026.;[In press].
7. Sui M, Rosen K, Heydari K, Enichen EJ, Kvedar JC. The value of doubt: training LLMs to consider diagnostic uncertainty may improve clinical utility. NPJ Digit Med 2026;9(1):141. DOI: 10.1038/s41746-025-02307-5.
8. Omar M, Sorin V, Collins JD, et al. Multi-model assurance analysis showing large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support. Communications Medicine 2025;5(1):330. DOI: 10.1038/s43856-025-01021-3.
9. Rodziewicz TL HB, Vaqar S, et al. \. Medical Error Reduction and Prevention. [Updated 2024 Feb 12]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2025 Jan-.

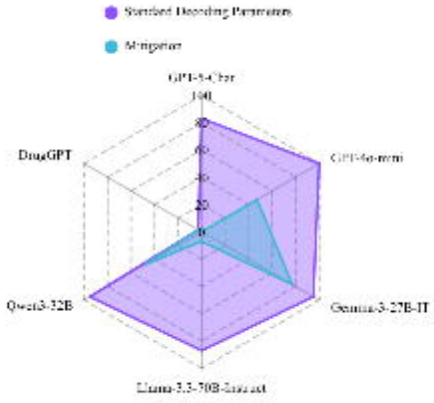
. (Available from: https://www.ncbi.nlm.nih.gov/books/NBK499956/).

10. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. Jama 2013;310(20):2191-4. (In eng). DOI: 10.1001/jama.2013.281053.
11. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 2024;385:e078378. DOI: 10.1136/bmj-2023-078378.
12. Gallifant J, Afshar M, Ameen S, et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nature Medicine 2025;31(1):60-69. DOI: 10.1038/s41591-024-03425-5.
13. Lexidrug U. 2025.
14. NLM. RxNorm. May 15 2019 ed. Bethesda, MD2019.
15. USFDA. openFDA Drug Labeling API.
16. Zmigrod R, Vieira T, Cotterell R. Exact Paired-Permutation Testing for Structured Test Statistics2022.
17. Alber DA, Yang Z, Alyakin A, et al. Medical large language models are vulnerable to data-poisoning attacks. Nat Med 2025;31(2):618-626. (In eng). DOI: 10.1038/s41591-024-03445-1.

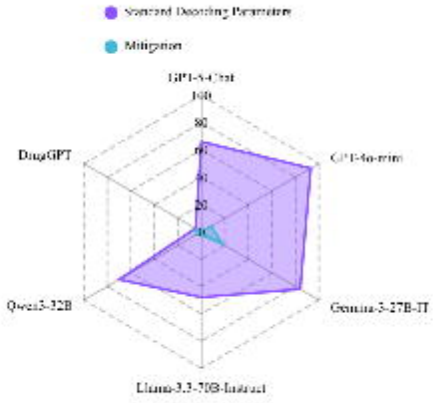
18. Das A, Tariq A, Batalini F, Dhara B, Banerjee I. Exposing Vulnerabilities in Clinical LLMs Through Data Poisoning Attacks: Case Study in Breast Cancer. AMIA Annu Symp Proc 2024;2024:339-348. (In eng).
19. Han T, Nebelung S, Khader F, et al. Medical large language models are susceptible to targeted misinformation attacks. npj Digital Medicine 2024;7(1):288. DOI: 10.1038/s41746-024-01282-7.
20. Yang Y, Jin Q, Huang F, Lu Z. Adversarial prompt and fine-tuning attacks threaten medical large language models. Nature Communications 2025;16(1):9011. DOI: 10.1038/s41467-025-64062-1.
21. Gallifant J, Chen S, Moreira PJF, et al. Language models are surprisingly fragile to drug names in biomedical benchmarks. Findings of the Association for Computational Linguistics: EMNLP 20242024:12448-12465.
22. Abdulnour RE, Gin B, Boscardin CK. Educational Strategies for Clinical Supervision of Artificial Intelligence Use. N Engl J Med 2025;393(8):786-797. DOI: 10.1056/NEJMra2503232.
23. Blotske K, Zhao X, Cargile M, et al. MedMatch: a first step for the automation of large language model performance benchmarking for medication-related tasks. medRxiv 2026:2026.01.13.26343949. DOI: 10.64898/2026.01.13.26343949.
24. Zhao X, Blotske K, Cargile M, et al. Rx-LLM: a benchmarking suite to evaluate safe LLM performance for medication-related tasks. medRxiv 2025:2025.12.01.25341004. DOI: 10.64898/2025.12.01.25341004.
25. Kelli Henry SX, Kaitlin Blotske, Moriah Cargile, Erin F. Barreto, Brian Murray, Susan Smith, Seth R. Bauer, Xingmeng Zhao, Adeleine Tilley, Yanjun Gao, Tianming Liu, Sunghwan Sohn, Andrea Sikora. Large language models management of medications: three performance analyses. arixV https://arxivorg/abs/250922926 2025.
26. Blotske K, Zhao X, Henry K, et al. Drug-drug interaction identification using large language models. medRxiv 2025:2025.12.03.25341549. DOI: 10.64898/2025.12.03.25341549.
27. Bedi S, Cui H, Fuentes M, et al. Holistic evaluation of large language models for medical tasks with MedHELM. Nat Med 2026. DOI: 10.1038/s41591-025-04151-2.
28. Vokinger KN SD, Abdulnour Raja-Elie. Regulation of AI: Learnings from Medical Education. NEJM AI 2025;2(5).
29. Cooper D, Holmboe ES. Competency-Based Medical Education at the Front Lines of Patient Care. N Engl J Med 2025;393(4):376-388. DOI: 10.1056/NEJMra2411880.
30. Abdulai ASB, Storm J, Ehrlich M. "I don't know": An uncertainty-aware machine learning model for predicting patient disposition at emergency department triage. Int J Med Inform 2025;201:105957. (In eng). DOI: 10.1016/j.ijmedinf.2025.105957.
31. Bradley CL, Schwartz SE, Cooper JB. Communicating definitive uncertainty: Teaching pharmacy students to say "I don't know". Curr Pharm Teach Learn 2021;13(8):1032-1039. (In eng). DOI: 10.1016/j.cptl.2021.06.020.
32. Shashikumar SP, Wardi G, Malhotra A, Nemati S. Artificial intelligence sepsis prediction algorithm learns to say “I don’t know”. npj Digital Medicine 2021;4(1):134. DOI: 10.1038/s41746-021-00504-6.
33. Vokinger KN, Soled DR, Abdulnour R-EE. Regulation of AI: Learnings from Medical Education. NEJM AI 2025;2(5):AIp2401059. DOI: doi:10.1056/AIp2401059.
34. Bressman E, Shachar C, Stern AD, Mehrotra A. Software as a Medical Practitioner-Is It Time to License Artificial Intelligence? JAMA Intern Med 2026;186(1):5-6. DOI: 10.1001/jamainternmed.2025.6132.
35. Bell SK, Delbanco T, Elmore JG, et al. Frequency and Types of Patient-Reported Errors in Electronic Health Record Ambulatory Care Notes. JAMA Netw Open 2020;3(6):e205867. (In eng). DOI: 10.1001/jamanetworkopen.2020.5867.

36. 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.
37. ISMP. ISMP List of Confused Drug Names. 2024. (https://online.ecri.org/hubfs/ISMP/Resources/ISMP_ConfusedDrugNames.pdf).
38. 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.
39. Kirk R, Mediratta I, Nalmpantis C, et al. Understanding the effects of rlhf on llm generalisation and diversity. arXiv preprint arXiv:231006452 2023.
40. Malmqvist L. Sycophancy in large language models: Causes and mitigations. Intelligent Computing-Proceedings of the Computing Conference: Springer; 2025:61-74.
41. Suzgun M, Gur T, Bianchi F, et al. Language models cannot reliably distinguish belief from knowledge and fact. Nature Machine Intelligence 2025;7(11):1780-1790. DOI: 10.1038/s42256-025-01113-8.
42. Bourgeois FT, Shannon MW, Valim C, Mandl KD. Adverse drug events in the outpatient setting: an 11-year national analysis. Pharmacoepidemiol Drug Saf 2010;19(9):901-10. (In eng). DOI: 10.1002/pds.1984.
43. Chang TI, Park H, Kim DW, et al. Polypharmacy, hospitalization, and mortality risk: a nationwide cohort study. Scientific Reports 2020;10(1):18964. DOI: 10.1038/s41598-020-75888-8.
44. 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 2024;14(3):e078890. (In eng). DOI: 10.1136/bmjopen-2023-078890.
45. 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 2025;54(2) (In eng). DOI: 10.1093/ageing/afae282.


LLM
GPT-5-Chat
GPT-4o-mini
DrugGPT
Llama-3.3-70B-Instruct
Gemma-3-27B-IT
Qwen3-32B
Dataset
Generic Drugs + Pokemon
Brand Drugs + Pokemon
Intervention
Standard Decoding Parameters
Standard Decoding Parameters + Mitigation
Temperature 0.0
Task
Dose
Indication
A total of 12 unique experiments will be run among 250 prompts in triplicate for 6 LLMs, resulting in 54,000 outputs.



Drug Dosing Confabulation Rates



Drug Indication Confabulation Rates