HHS Public Access

Author manuscript

Epilepsy Res. Author manuscript; available in PMC 2026 March 01.

Published in final edited form as:

Epilepsy Res. 2025 March ; 211: 107532. doi:10.1016/j.eplepsyres.2025.107532.

Inductive reasoning with large language models: a simulated randomized controlled trial for epilepsy

Daniel M. Goldenholz, MD, PhD[1,2], **Shira R. Goldenholz, MD, MPH**[2], **Sara Habib, MD**[1,2], **M. Brandon Westover, MD PhD**[1,2]

[1] –Department of Neurology, Harvard Medical School, Boston USA

[2] –Department of Neurology, Beth Israel Deaconess Medical Center, Boston USA

Abstract

INTRODUCTION: To investigate the potential of using artificial intelligence (AI), specifically large language models (LLMs), for synthesizing information in a simulated randomized clinical trial (RCT) for an anti-seizure medication, cenobamate, demonstrating the feasibility of inductive reasoning via medical chart review.

METHODS: An LLM-generated simulated RCT was conducted, featuring a placebo arm and a full-strength drug arm with a cohort of 240 patients divided 1:1. Seizure counts were simulated using a realistic seizure diary simulator. The study utilized LLMs to generate clinical notes with four neurologist writing styles and random extraneous details. A secondary LLM pipeline synthesized data from these notes. The efficacy and safety of cenobamate in seizure control were evaluated by both an LLM-based pipeline and a human reader.

RESULTS: The AI analysis closely mirrored human analysis, demonstrating the drug's efficacy with marginal differences (<3%) in identifying both drug efficacy and reported symptoms. The AI successfully identified the number of seizures, symptom reports, and treatment efficacy, with statistical analysis comparing the 50%-responder rate and median percentage change between the placebo and drug arms, as well as side effect rates in each arm.

Corresponding author: Daniel Goldenholz, daniel.goldenholz@bidmc.harvard.edu, 330 Brookline Ave, Baker 5, Boston MA 02215.

Author contributions

DMG – project design and oversight, data interpretation, manuscript writing and editing
SRG – manuscript editing
SH – data analysis, manuscript editing
MBW – project oversight, manuscript editing

• ethics approval statement
• No ethics approval was required – this was a simulation study.

• patient consent statement
No patient consent was required as this was a simulation study.

• permission to reproduce material from other sources
No materials are being reproduced. All rights are retained by the authors.

• clinical trial registration
No real clinical trial was conducted, the study was pure simulation.

DISCUSSION: This study highlights the potential of AI to accurately analyze noisy clinical notes to inductively produce clinical knowledge. Here, treatment effect sizes and symptom frequencies derived from unstructured simulated notes were inferred despite many distractors. The findings emphasize the relevance of AI in future clinical research, offering a scalable and efficient alternative to traditional labor-intensive data mining.

Keywords

artificial intelligence; large language models; epilepsy; randomized clinical trials

Introduction

It is very challenging to accurately extract generalizable knowledge from the entire electronic medical system(Yang et al., 2023). If it were possible to do this, a continuous learning medical artificial intelligence (AI) could begin to develop a ruudimentary sense of “clinical experience” in a manner analogous to clinicians who learn from decades of practicing medicine(Benner, 2004; Nilsson and Pilhammar, 2009). Various approaches, including the use of structured data(Ostropolets et al., 2023), natural language processing toolboxes(Fu et al., 2023; Murphy et al., 2023; Wu et al., 2020), and others have been shown to hold some promise. Researchers have struggled to extract meaning from clinical notes when trying to correctly classify every element of data from single notes(Agrawal et al., 2022; Yang et al., 2022). With the advent of highly capable foundational large language models (LLMs)(Bubeck et al., 2023; Lee et al., 2023; OpenAI, 2023), new opportunities for AI in medicine are emerging.

Current generation LLMs are plagued with a variety of constraints, including input size limits(Liu et al., 2024), confabulations (a.k.a. “hallucinations”), inaccuracies(Soroush et al., 2024), biases, and incomplete knowledge bases(Lee et al., 2023). Despite these limitations, modern LLMs have made important strides. For example, accurately reading clinical notes(Yang et al., 2024), and summarizing clinical reports(Liu et al., 2024). In epilepsy, LLMs correctly answered practice licensing exam questions(Habib et al., 2024), classified patients as seizure-free based on in clinical notes(Xie et al., 2023) and uncovered outcome disparities in subgroups of people with epilepsy(Xie et al., 2024).

In the present work, our data was synthetic clinical notes from a simulated randomized clinical trial (RCT). By simulating the data, we had complete control over the ground truth that the AI was to determine. We set out to explore a model system consisting of LLM-generated encounter notes from all “patients” in the RCT (Figure 1) that are fed into an AI pipeline to summarize and synthesize information (Figure 2). The hypothesis was that a simulated randomized clinical trial data could be summarized and accurately evaluated with the help of LLMs. *Inductive reasoning*, defined here as generalizing knowledge based on a set of observations, is submitted as one of the ways clinicians learn. If correctly implemented, an inductive reasoning system could determine features of each treatment arm (both the efficacy and side effects) by reading the clinical notes. The purpose of this task was to demonstrate the power of AI-enhanced inductive reasoning applied to medical chart review in the context of an RCT.

Methods

Building the dataset

A simulated randomized clinical trial (Figure 1) was constructed with parameters based on a historical RCT in epilepsy for cenobamate(Krauss et al., 2020). In that trial, there were 2 months of baseline, and 3 months of maintenance at steady state for the drug. In our simulation, there was a placebo arm and a full-strength drug arm (corresponding to 400mg/day cenobamate). Similar to the cenobamate trial, we included 120 patients per arm. To generate a realistic cohort of simulated patients, a previously validated seizure diary simulator (CHOCOLATES) was used(Goldenholz and Westover, 2023). CHOCOLATES was designed to account for heterogeneity in seizure frequencies across patients(Ferastraoaru et al., 2018), the “L-relationship” power law within dairies(Goldenholz et al., 2018), seizure clustering(Chiang et al., 2020; Haut, 2015), seizure susceptibility cycles(Baud et al., 2018; Karoly et al., 2018), and maximum allowable seizure rates(Trinka et al., 2015). Based on multiple lines of evidence in epilepsy(Daniel M. Goldenholz et al., 2017a; D.M. Goldenholz et al., 2017; Goldenholz et al., 2015; Goldenholz and Goldenholz, 2020, 2016; Goldenholz and Westover, 2023; Karoly et al., 2019; Romero et al., 2020), we assumed that placebo has no intrinsic effect on seizure rate and any measured placebo response would be solely due to natural variability and regression to the mean(Goldenholz et al., 2023). Despite this assumption (i.e. that the “placebo effect” is effectively nonexistent in epilepsy trials) typical placebo response rates from numerous historical RCTs are reproduced in our simulator(Goldenholz and Westover, 2023; Kerr et al., 2024). Similar to the RCT(Krauss et al., 2020), simulated patients needed to have an average rate of 4 seizures per month to be included in the simulated study. Like cenobamate, the simulated drug was 39% more effective than placebo(Krauss et al., 2020). The precise symptom reporting rates in the placebo and drug arms of the cenobamate trial were simulated as well(Krauss et al., 2020). A well-characterized, open-source LLM(Touvron et al., n.d.) called Llama2:13b was used to generate clinical notes with the temperature parameter set to 1.0 (values >0 increase creativity). The creativity, as well as LLM hallucination, were intentionally part of this study to properly simulate “noise” caused by inaccurate patient reporting and inaccurate documentation. One of four neurologist writing styles was randomly selected at the time of each clinical note generation: 1) a terse minimalist style using bullet points, 2) a complete but brief style, 3) a narrative style in 2-4 paragraphs, and 4) an erudite academic professor with many extraneous details. Each simulated patient had two notes generated, one after the baseline period and one after the blinded maintenance period (480 notes total). Of note, each clinical note could have been in a different style, even for a single patient. Additional random details about the patients’ past medical history were added randomly but kept consistent within each patient. In addition to a complete note, each encounter also generated a “ground truth” entry in a data table that indicates what information was used in the prompt to the LLM to generate the clinical note.

AI analysis of the notes

An AI pipeline for analysis of the RCT was constructed as follows (Figure 2). Each note was fed individually (due to input size constraints) to a second open-source LLM(Jiang et al., 2023) (Mistral 7B v0.1) set to a temperature of 0.0 to increase precision and

decrease extraneous detail. This LLM was selected because it was produced independently of Llama2, and thus would not have the luxury of expecting certain styles or methods of writing. The LLM was asked to summarize the note, specifically indicating the number of seizures during the observation period and what symptoms were reported by the patient. Due to inaccuracies and incomplete responses from typical open-source LLMs, it was not feasible for the LLM to build the final data table required for statistical analysis. Thus, a set of somewhat poorly formatted but mostly complete summaries was obtained from the second LLM.

A third LLM (Claude 2), was also used. This LLM has an extended data input limit and is able to ingest large numbers of summaries at once, resulting in the ability to produce a well formatted data table, and correctly make synthesis inferences correctly. Claude 2 is freely available via web interface, but the application programming interface (API) requires a paid account. In addition to improving the formatting, the third LLM was asked to indicate, on each row, the number of seizures during each period of the study; it was also asked to indicate if there were symptoms reported in the second encounter that differed from the first encounter (representing new symptoms that started along with the experimental treatment). All code was run on an M1 Studio Mac desktop computer with 32Gb of RAM.

Human analysis of the notes

The set of 480 generated clinical notes were assessed by one of the authors (SH), a trained neurologist. The relevant features, namely, the number of seizures during the observation period and any symptoms reported, were manually extracted and organized into a data table.

Statistical analysis of data tables

Three data tables (the ground truth, the AI, and the human) were analyzed in the same way. The percentage change between average monthly seizure rate during baseline to average monthly seizure rate during the maintenance period was computed for each patient(Siddiqui and Hershkowitz, 2010). These percentage change values were used to tally the fraction of 50%-responders in each arm, and then to compute a Fisher Exact test to compare arms (RR50). The same percentage change values were also used to compute a median percentage change (MPC) and the Mann-Whitney U test was used to non-parametrically compare the two arms. Uniquely reported symptoms were tallied up in each arm, and these were summarized.

The TRIPOD reporting checklist(Collins et al., 2015) is provided (Appendix). Code was prepared in python using langchain and ollama. Open-source code is available at https://github.com/GoldenholzLab/LLM-rct.git.

Results

Computational time for generation and summarization of notes combined took roughly 20 hours on a single computer; this time would of course be reduced with increased computational resources. The complete set of notes are available for review (Appendix). The human review of the 480 notes required roughly 5 hours. In the placebo group, 9 patients were identified as not having any value reported for seizures in either the baseline

or maintenance periods. In the drug group there were an additional 8 such patients. These failures can be attributed to the generative LLM A (Figure 1) that produced the notes. These were not corrected, as these represented examples of undesirable “noise” that prevented perfect reconstruction of the ground truth. When computing the statistics for efficacy, patients with missing numbers were excluded. All patients were included when computing symptom report summaries.

The treatment effect sizes reported for the 50%-responder rate (RR50) and median percentage change (MPC) are shown in Figure 3. The marginal efficacy between drug and placebo are shown in Table 1. All comparisons were statistically significant. The AI and human marginal efficacies differed by 1% in both RR50 and MPC.

The reported symptoms identified from each of the data tables are shown in Figure 4. The maximum differences in symptom rates between tables were: AI vs. truth – 2%, human vs truth – 2%, AI vs. human – 3%.

Discussion

This study simulated a realistic trial modeled after a published randomized drug trial(Krauss et al., 2020), employed AI to reconstruct the important elements that were reported quantitatively and qualitatively in the clinical notes. Although completely simulated, the present study demonstrates a proof-of-concept – that the present framework can enable inductive reasoning from large sets of unstructured clinical encounters. The context used here was a simulated RCT, however typical RCTs use standardize input forms, and would not typically need this tool. However, it is easy to imagine how the present context could be broadened to clinical notes generally. The AI pipeline was able to correctly show the marginal drug efficacy (drug vs. placebo) differing from human review by no more than 1%. Similarly, the AI was able to identify the relevant symptoms reported in drug and placebo arms, differing with the human by no more than 3%. The use of generative AI to produce synthetic clinical notes allowed us to intentionally inject “noise” (distracting and/or incorrect elements) into our experiment. This deliberate addition was made to help determine if we could teach AI system to learn medical information by induction in the presence of noise. In typical clinical situations, there is virtually always some “noise” generated, whether due to inaccurate reporting by patients or caregivers, or inaccurate recording by clinicians. Our system was able to correctly show a strong effect of the simulated drug and found the appropriate common side effects without being taught to look for something specific. These achievements are more remarkable when considering an important point: this entire project did not make use of *any* LLMs specially trained in medical language(Singhal et al., n.d.). Moreover, the AI pipeline for interpretation was run without cloud server or service costs. The simulated dataset produced for this study (see Appendix) may be considered a valuable tool for future work as well.

It is worth reflecting that present generation LLM tools continue to make mistakes in interpretation, which has plagued the field, and made it difficult to trust AI to “read the chart”. Our study recommends a paradigm shift away from perfectly understanding the individual patient towards generalizable knowledge extracted from groups of patients. This

new paradigm capitalizes on the strengths of LLMs which acknowledging their weakness at high precision.

Future versions of the present pipeline might employ only a single LLM if it was computationally efficient, impervious to hallucination, inexpensive and had a very large token input length. The advantage of the current approach is that it is not necessary to wait for such advances to be made available. New LLM tools are being developed at an accelerated pace, featuring less errors, increased accuracy, and better specialty skills(Habib et al., 2024). Therefore, the capabilities demonstrated in the present study should be considered a lower boundary on what is possible.

Like any simulation, this study is only as good as the assumptions made. We assumed we have an adequate model for seizure diaries and trial simulation based on prior work(Daniel M. Goldenholz et al., 2017a, 2017b; Goldenholz and Westover, 2023; Oliveira et al., 2019; Romero et al., 2020). We also assumed that generative LLM clinical notes can represent a first approximation for true clinical notes, and that the conditions presented here are relevant to other inductive learning tasks of interest in clinical settings. Although the clinical notes may have had some regularities, we attempted to mitigate this by requiring the notes to be written in multiple styles and to include a number of distractors. We also allowed the note generator to be “creative”, thus potentially adding irrelevant or conflicting details to the notes which prevented the task from being accurate. This was done to mimic a “messier” real-life situation of different authors of clinical notes with various levels of accuracy. It is worthwhile considering that the “ground truth” notes were subject to hallucinations – this was by design. We wanted the ground truth to roughly approximate the statistical features of interest but to allow noise to enter at every level in order to make the task more complex. The group level results (shown in Figure 1 and 2) indicate that the aggregated statistical features that were desired were nevertheless captured. It is possible that the reported symptoms, despite the various strategies mentioned above, may have not been sufficiently hidden by the AI compared to some disorganized human-written notes. Further study is needed to enhance the level of realism in note simulation. Another limitation of this study was a linguistic one: our study was conducted entirely in English. Multilingual open-source models(Workshop et al., 2022) are becoming more available to extend the present work to many other languages. We acknowledge that despite multiple safeguards (using different LLMs for generation and evaluation, adding noise in the generating LLM itself as well as noise in the note style and extra details) it may have been possible that the AI system “cheated” by looking for the key information and ignored everything else. We believe this to be unlikely given the properties of the disconnected components of the system, but further study will be needed to confirm this point.

It must also be noted that extremely rare side effects in a randomized controlled trial might be missed by the type of system developed here – for example, if an investigational drug causes a systemic inflammatory reaction in only 1 patient for the whole study, this fact must not be missed by trialists. Whereas the system proposed here may miss such rare side effects, our goal is to look for larger trends and not “outlier” rare results. Indeed, if such rare reactions were only noted in postmarketing studies, it could take a long time for regulators (and therefore clinicians) to become aware of them, yet an inductively learning AI system

could flag situations like this if they happen in low fractions of patients beyond expected levels.

The longer-term purpose of building clinical inductive learning tools is to develop real-time systems that can learn from large populations and apply this knowledge to uncertain situations. For instance, if a new drug is approved, physicians develop a certain personal clinical "experience" with that drug, and after this they base their prescribing habits on that experience. That personal experience sometimes matches the clinical trials, while sometimes there is a mismatch. This "clinical experience" is one of the ingredients that makes seasoned clinicians more effective at choosing from an uncertain set of choices(Benner, 2004; Nilsson and Pilhammar, 2009). If an AI-enhanced system can develop such clinical experience across populations, it will be able to rapidly assist countless clinicians with the most updated experience base possible – vastly larger than any one clinician can accrue in a lifetime of practice.

In conclusion, we demonstrated that known (but hidden) knowledge could be inferred by induction with a moderate sample of patient charts. Further studies are needed to expand this capability to broader medical knowledge acquisition and applications.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

Funding for this work came in part from NIH K23NS124656. The authors wish to thank the open-source community for sharing and distributing large language models such as llama2 and mistral, as these tools can help advance biomedical science.

Conflict of interest / Ethical Statement

None of the authors has any conflict of interest to disclose. We confirm that we have read the Journal's position on issues involved in ethical publication and affirm that this report is consistent with those guidelines. Source code is available on https://github.com/GoldenholzLab/LLM-rct.git. All data used for this study was simulated and is provided in the Appendix. This study was funded by NIH K23NS124656. No materials are being reproduced. All rights are retained by the authors.

Data Sharing Statement

Open-source code and data is available on: https://github.com/GoldenholzLab/LLM-rct.git.

REFERENCES

Agrawal M, Hegselmann S, Lang H, Kim Y, Sontag D, 2022. Large Language Models are Few-Shot Clinical Information Extractors.

Baud MO, Kleen JK, Mirro EA, Andrechak JC, King-Stephens D, Chang EF, Rao VR, 2018. Multi-day rhythms modulate seizure risk in epilepsy. Nat Commun 9, 1–10. 10.1038/s41467-017-02577-y [PubMed: 29317637]

Benner P, 2004. Using the dreyfus model of skill acquisition to describe and interpret skill acquisition and clinical judgment in nursing practice and education. Bull Sci Technol Soc 24, 188–199. 10.1177/0270467604265061

Bubeck S, Chandrasekaran V, Eldan R, Gehrke J, Horvitz E, Kamar E, Lee P, Lee YT, Li Y, Lundberg S, Nori H, Palangi H, Ribeiro MT, Zhang Y, 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4.

Chiang S, Haut SR, Ferastraoaru V, Rao VR, Baud MO, Theodore WH, Moss R, Goldenholz DM, 2020. Individualizing the definition of seizure clusters based on temporal clustering analysis. Epilepsy Res 163. 10.1016/j.eplepsyres.2020.106330

Collins GS, Reitsma JB, Altman DG, Moons KGM, 2015. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement. Eur Urol 67, 1142–1151. 10.1016/j.eururo.2014.11.025 [PubMed: 25572824]

Ferastraoaru V, Goldenholz DM, Chiang S, Moss R, Theodore WH, Haut SR, 2018. Characteristics of large patient-reported outcomes: Where can one million seizures get us? Epilepsia Open 3, 364–373. 10.1002/epi4.12237 [PubMed: 30187007]

Fu S, Wen A, Liu H, 2023. Clinical Natural Language Processing in Secondary Use of EHR for Research 433–451. 10.1007/978-3-031-27173-1_21

Goldenholz DM, Goldenholz EB, Kaptchuk TJ, 2023. Quantifying and controlling the impact of regression to the mean on randomized controlled trials in epilepsy. Epilepsia 64, 2635–2643. 10.1111/epi.17730 [PubMed: 37505116]

Goldenholz DM, Goldenholz SR, 2020. Placebo in epilepsy, in: International Review of Neurobiology. Academic Press Inc. 10.1016/bs.irn.2020.03.033

Goldenholz DM, Goldenholz SR, 2016. Response to placebo in clinical epilepsy trials-Old ideas and new insights. Epilepsy Res 122. 10.1016/j.eplepsyres.2016.02.002

Goldenholz DM, Goldenholz SR, Moss R, French J, Lowenstein D, Kuzniecky R, Haut S, Cristofaro S, Detyniecki K, Hixson J, Karoly P, Cook M, Strashny A, Theodore WH, 2018. Is seizure frequency variance a predictable quantity? Ann Clin Transl Neurol 5. 10.1002/acn3.519

Goldenholz DM, Moss R, Scott J, Auh S, Theodore WH, 2015. Confusing placebo effect with natural history in epilepsy: A big data approach. Ann Neurol 78. 10.1002/ana.24470

Goldenholz DM, Strashny A, Cook M, Moss R, Theodore WH, 2017. A multi-dataset time reversal approach to clinical trial placebo response and the relationship to natural variability in epilepsy. Seizure 53. 10.1016/j.seizure.2017.10.016

Goldenholz Daniel M., Tharayil J, Moss R, Myers E, Theodore WH, 2017a. Monte Carlo simulations of randomized clinical trials in epilepsy. Ann Clin Transl Neurol 4, 544–552. 10.1002/acn3.426 [PubMed: 28812044]

Goldenholz Daniel M., Tharayil JJ, Kuzniecky R, Karoly P, Theodore WH, Cook MJ, 2017b. Simulating clinical trials with and without Intracranial EEG data. Epilepsia Open 2, 156–161. 10.1002/epi4.12038 [PubMed: 28758158]

Goldenholz DM, Westover MB, 2023. Flexible realistic simulation of seizure occurrence recapitulating statistical properties of seizure diaries. Epilepsia 64, 396–405. 10.1111/epi.17471 [PubMed: 36401798]

Habib S, Butt H, Goldenholz SR, Chang CY, Goldenholz DM, 2024. Large Language Model Performance on Practice Epilepsy Board Examinations. JAMA Neurol 81, 660–661. 10.1001/jamaneurol.2024.0676 [PubMed: 38587850]

Haut SR, 2015. Seizure clusters: characteristics and treatment. Curr Opin Neurol 28, 143–50. 10.1097/WCO.0000000000000177 [PubMed: 25695133]

Jiang AQ, Sablayrolles A, Mensch A, Bamford C, Chaplot DS, Casas D. de las Bressand F, Lengyel G, Lample G, Saulnier L, Lavaud LR, Lachaux M-A, Stock P, Scao T. Le, Lavril T, Wang T, Lacroix T, Sayed W. El, 2023. Mistral 7B.

Karoly PJ, Goldenholz DM, Freestone DR, Moss RE, Grayden DB, Theodore WH, Cook MJ, 2018. Circadian and circaseptan rhythms in human epilepsy: a retrospective cohort study. Lancet Neurol 17, 977–985. 10.1016/S1474-4422(18)30274-6 [PubMed: 30219655]

Karoly PJ, Romero J, Cook MJ, Freestone DR, Goldenholz DM, 2019. When can we trust responders? Serious concerns when using 50% response rate to assess clinical trials. Epilepsia 60. 10.1111/epi.16321

Kerr WT, Suprun M, Kok N, Reddy AS, McFarlane KN, Kwan P, Somerville E, Bagiella E, French JA, 2024. Factors associated with placebo response rate in randomized controlled trials of antiseizure medications for focal epilepsy. Epilepsia. 10.1111/epi.18197

Krauss GL, Klein P, Brandt C, Lee SK, Milanov I, Milovanovic M, Steinhoff BJ, Kamin M, 2020. Safety and efficacy of adjunctive cenobamate (YKP3089) in patients with uncontrolled focal seizures: a multicentre, double-blind, randomised, placebo-controlled, dose-response trial. Lancet Neurol 19, 38–48. 10.1016/S1474-4422(19)30399-0 [PubMed: 31734103]

Lee P, Goldberg C, Kohane I, Bubeck S, 2023. The AI revolution in medicine : GPT-4 and beyond.

Liu A, Zhou H, Hua Y, Rohanian O, Thakur A, Clifton L, Clifton DA, 2024. Large Language Models in the Clinic: A Comprehensive Benchmark.

Murphy RM, Klopotowska JE, de Keizer NF, Jager KJ, Leopold JH, Dongelmans DA, Abu-Hanna A, Schut MC, 2023. Adverse drug event detection using natural language processing: A scoping review of supervised learning methods. PLoS One 18, e0279842. 10.1371/journal.pone.0279842 [PubMed: 36595517]

Nilsson MS, Pilhammar E, 2009. Professional approaches in clinical judgements among senior and junior doctors: Implications for medical education. BMC Med Educ 9. 10.1186/1472-6920-9-25

Oliveira A, Romero JM, Goldenholz DM, 2019. Comparing the efficacy, exposure, and cost of clinical trial analysis methods. Epilepsia 60, e128–e132. 10.1111/epi.16384 [PubMed: 31724165]

OpenAI, 2023. GPT-4 Technical Report.

Ostropolets A, Hripcsak G, Husain SA, Richter LR, Spotnitz M, Elhussein A, Ryan PB, 2023. Scalable and interpretable alternative to chart review for phenotype evaluation using standardized structured data from electronic health records. J Am Med Inform Assoc 31, 119–129. 10.1093/jamia/ocad202 [PubMed: 37847668]

Romero J, Larimer P, Chang B, Goldenholz SR, Goldenholz DM, 2020. Natural variability in seizure frequency: Implications for trials and placebo. Epilepsy Res 162, 106306. 10.1016/j.eplepsyres.2020.106306 [PubMed: 32172145]

Siddiqui O, Hershkowitz N, 2010. Primary Efficacy Endpoint in Clinical Trials of Antiepileptic Drugs: Change or Percentage Change. Drug Inf J 44, 343–350. 10.1177/009286151004400316

Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Hou L, Clark K, Pfohl S, Cole-Lewis H, Neal D, Schaekermann M, Wang A, Amin M, Lachgar S, Mansfield P, Prakash S, Green B, Dominowska E, Aguera Arcas B, Tomasev N, Liu Y, Wong R, Semturs C, Sara Mahdavi S, Barral J, Webster D, Corrado GS, Matias Y, Azizi S, Karthikesalingam A, Natarajan V, n.d. Towards Expert-Level Medical Question Answering with Large Language Models.

Soroush A, Glicksberg BS, Zimlichman E, Barash Y, Freeman R, Charney AW, Nadkarni GN, Klang E, 2024. Large Language Models Are Poor Medical Coders — Benchmarking of Medical Code Querying. NEJM AI 1. 10.1056/aidbp2300040

Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, Bashlykov N, Batra S, Bhargava P, Bhosale S, Bikel D, Blecher L, Ferrer CC, Chen M, Cucurull G, Esiobu D, Fernandes J, Fu J, Fu W, Fuller B, Gao C, Goswami V, Goyal N, Hartshorn A, Hosseini S, Hou R, Inan H, Kardas M, Kerkez V, Khabsa M, Kloumann I, Korenev A, Koura S, Lachaux M-A, Lavril T, Lee J, Liskovich D, Lu Y, Mao Y, Martinet X, Mihaylov T, Mishra P, Molybog I, Nie Y, Poulton A, Reizenstein J, Rungta R, Saladi K, Schelten A, Silva R, Michael E, Ranjan S, Xiaoqing S, Tan E, Tang B, Taylor R, Williams A, Kuan JX, Xu P, Yan Z, Zarov I, Zhang Y, Fan A, Kambadur M, Narang S, Rodriguez A, Stojnic R, Edunov S, Scialom T, n.d. Llama 2: Open Foundation and Fine-Tuned Chat Models.

Trinka E, Cock H, Hesdorffer D, Rossetti AO, Scheffer IE, Shinnar S, Shorvon S, Lowenstein DH, 2015. A definition and classification of status epilepticus - Report of the ILAE Task Force on Classification of Status Epilepticus. Epilepsia 56, 1515–1523. 10.1111/epi.13121 [PubMed: 26336950]

Workshop B, Le Scao T, Fan A, Akiki C, Pavlick E, Ili S, Hesslow D, Castagné R, Sasha Luccioni A, Yvon F, Gallé M, Tow J, Rush AM, Biderman S, Webson A, Sasanka Ammanamanchi P, Wang T, Sagot B, Muennighoff N, Villanova del Moral A, Ruwase O, Bawden R, Bekman S, McMillan-Major A, Wolf T, Beltagy I, Nguyen H, Saulnier L, Tan S, Ortiz Suarez P, Sanh V, Laurençon H, Jernite Y, Launay J, Mitchell Margaret, Raffel Dataset Aaron Gokaslan C., Simhi A, Soroa A, Fikri Aji A, Alfassy A, Rogers A, Kreisberg Nitzav A, Xu, Canwen Mou C, Emezue C,

Klamm C, Leong C, Raffel C, van Strien D, Ifeoluwa Adelani D, Radev D, González Pon-ferrada E, Levkovizh E, Kim E, Bar Natan E, De Toni F, Dupont G, Kruszewski G, Pistilli G, Elsahar H, Benyamina H, Tran H, Yu I, Abdulmumin I, Johnson I, Gonzalez-Dios I, de la Rosa J, Chim J, Dodge J, Zhu J, Chang J, Frohberg J, Tobing J, Bhattachar-jee J, Almubarak K, Chen K, Lo K, Von Werra L, Weber L, Phan L, Ben allal L, Tanguy L, Dey M, Romero Muñoz M, Masoud M, Grandury M, Huang M, Coavoux M, Singh M, Tian-Jian Jiang M, Chien Vu M, Jauhar MA, Ghaleb M, Subramani N, Kassner N, Khamis N, Nguyen O, Espejel O, de Gibert O, Villegas P, Henderson P, Colombo P, Amuok P, Lhoest Q, Harliman R, Bommasani R, Luis López R, Ribeiro R, Osei S, Pyysalo S, Nagel S, Bose S, Hassan Muhammad S, Sharma S, Longpre S, Nikpoor S, Silberberg S, Pai S, Zink S, Timponi Torrent T, Schick T, Thrush T, Danchev V, Nikoulina V, Laippala V, Lepercq V, Prabhu V, Alyafeai Z, Talat Tokenization Arun Raja Z, Heinzerling B, Si C, Emre Ta D, Salesky E, Mielke SJ, Lee WY, Sharma A, Santilli A, Chaffin A, Stiegler A, Raja A, Datta D, Szczechla E, Chhablani G, Wang H, Pandey H, Strobelt H, Alan Fries J, Rozen J, Gao L, Sutawika L, Saiful Bari M, Al-shaibani MS, Manica M, Nayak N, Teehan R, Albanie S, Shen S, Ben-David S, Bach SH, Kim T, Bers T, Fevry T, Neeraj T, Thakker U, Raunak V, Tang X, Yong Z-X, Sun Z, Brody S, Uri Y, Tojarieh H, Roberts A, Won Chung H, Tae, Jaesung, Phang J, Muen-nighoff N, Press O, Li C, Narayanan D, Bourfoune H, Casper J, Rasley J, Ryabinin M, Mishra M, Zhang M, Shoeybi M, Peyrounette M, Patry N, Tazi N, Sanseviero O, von Platen P, Cor-nette P, François Lavallée P, Lacroix R, Rajbhandari S, Gandhi S, Smith S, Requena S, Patil S, Dettmers T, Baruwa A, Singh, Amanpreet Cheveleva A, Ligozat A-L, Subramonian A, Névéol A, Lovering C, Garrette D, Tunuguntla D, Reiter E, Taktasheva E, Voloshina E, Bogdanov E, Indra Winata G, Schoelkopf H, Kalo J-C, Novikova J, Zosa Forde J, Clive J, Kasai J, Kawamura K, Hazan L, Carpuat M, Clinciu M, Kim N, Cheng N, Serikov O, Antverg O, van der Wal O, Zhang Rui, Zhang Ruochen, Gehrmann S, Mirkin S, Pais S, Shavrina T, Scialom T, Yun T, Limisiewicz T, Rieser V, Protasov V, Mikhailov V, Pruksachatkun Y, Belinkov Y, Bamberger Z, Kasner Z, Talat Z, Gokaslan A, Rueda A, Pestana A, Feizpour A, Khan A, Faranak A, Santos A, Hevia A, Unldreaj A, Aghagol A, Abdollahi A, Tammour A, HajiHosseini A, Behroozi B, Ajibade B, Saxena B, Muñoz Ferrandis C, McDuff D, Contractor D, Lansky D, David D, Kiela D, Nguyen DA, Tan E, Baylor E, Ozoani E, Mirza F, Ononiwu F, Rezanejad H, Jones H, Bhattacharya I, Solaiman I, Sedenko I, Nejadgholi I, Tae Jae-sung, Passmore J, Seltzer J, Bonis Sanz J, Dutra L, Samagaio M, Mitchell Mar-garet, Mieskes M, Gerchick M, Akinlolu M, McKenna M, Qiu M, Ghauri M, Burynok M, Abrar N, Rajani N, Elkott N, Fahmy N, Samuel O, An R, Kromann R, Hao R, Alizadeh S, Shubber S, Wang S, Roy S, Viguier S, Le, Thanh, Oyebade T, Le, Trieu, Yang Y, Nguyen Z, Yong Applications Abhinav Ramesh, Kashyap Z.-X., Palasciano A, Callahan A, Shukla A, Miranda-Escalada A, Singh, Ayush, Beilharz B, Wang B, Brito C, Muñoz Ferrandis B, Zhou C, Jain C, Xu, Chuxin, Fourrier C, León Periñán D, Molano D, Yu D, Manjavacas E, Barth F, Fuhrimann F, Altay G, Bayrak G, Burns G, Vrabec HU, Bello I, Dash I, Kang J, Giorgi J, Golde J, David Posada J, Rangasai Sivaraman K, Bulchandani L, Liu L, Shinzato L, Hahn de Bykhovetz M, Takeuchi M, Pàmies M, Castillo MA, Nezhurina M, Sänger M, Samwald M, Cullan M, Weinberg M, De Wolf M, Mihaljcic M, Liu M, Freidank M, Kang M, Seelam N, Dahlberg N, Michio Broad N, Muellner N, Fung P, Haller P, Chandrasekhar R, Eisenberg R, Martin R, Canalli R, Su Rosaline, Su Ruisi, Cahyawijaya S, Garda S, Deshmukh SS, Mishra S, Kiblawi S, Ott S, Sang-aroonsiri S, Kumar S, Schweter S, Bharati S, Laud T, Gigant T, Kainuma T, Kusa W, Labrak Y, Shailesh Bajaj Y, Venkatraman Y, Xu Yifan, Xu Yingxin, Xu Yu, Tan Z, Xie Z, Ye Z, Bras M, Belkada Y, 2022. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model.

Wu S, Roberts K, Datta S, Du J, Ji Z, Si Y, Soni S, Wang Q, Wei Q, Xiang Y, Zhao B, Xu H, 2020. Deep learning in clinical natural language processing: a methodical review. J Am Med Inform Assoc 27, 457–470. 10.1093/jamia/ocz200 [PubMed: 31794016]

Xie K, Gallagher RS, Shinohara RT, Xie SX, Hill CE, Conrad EC, Davis KA, Roth D, Litt B, Ellis CA, 2023. Long-term epilepsy outcome dynamics revealed by natural language processing of clinic notes. Epilepsia 64, 1900–1909. 10.1111/epi.17633 [PubMed: 37114472]

Xie K, Ojemann WKS, Gallagher RS, Shinohara RT, Lucas A, Hill CE, Hamilton RH, Johnson KB, Roth D, Litt B, Ellis CA, 2024. Disparities in seizure outcomes revealed by large language models. Journal of the American Medical Informatics Association 31, 1348–1355. 10.1093/jamia/ocae047 [PubMed: 38481027]

Yang S, Varghese P, Stephenson E, Tu K, Gronsbell J, 2023. Machine learning approaches for electronic health records phenotyping: a methodical review. J Am Med Inform Assoc 30, 367–381. 10.1093/jamia/ocac216 [PubMed: 36413056]

Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, Compas C, Martin C, Costa AB, Flores MG, Zhang Y, Magoc T, Harle CA, Lipori G, Mitchell DA, Hogan WR, Shenkman EA, Bian J, Wu Y, 2022. A large language model for electronic health records. NPJ Digit Med 5. 10.1038/s41746-022-00742-2

Yang Z, Mitra A, Kwon S, Yu H, 2024. ClinicalMamba: A Generative Clinical Language Model on Longitudinal Clinical Notes.

HIGHLIGHTS

- AI analysis of simulated clinical trial notes for cenobamate matched human analysis
- AI successfully extracted key data on efficacy and side effects
- This study demonstrates AI can synthesize clinical knowledge from unstructured medical data

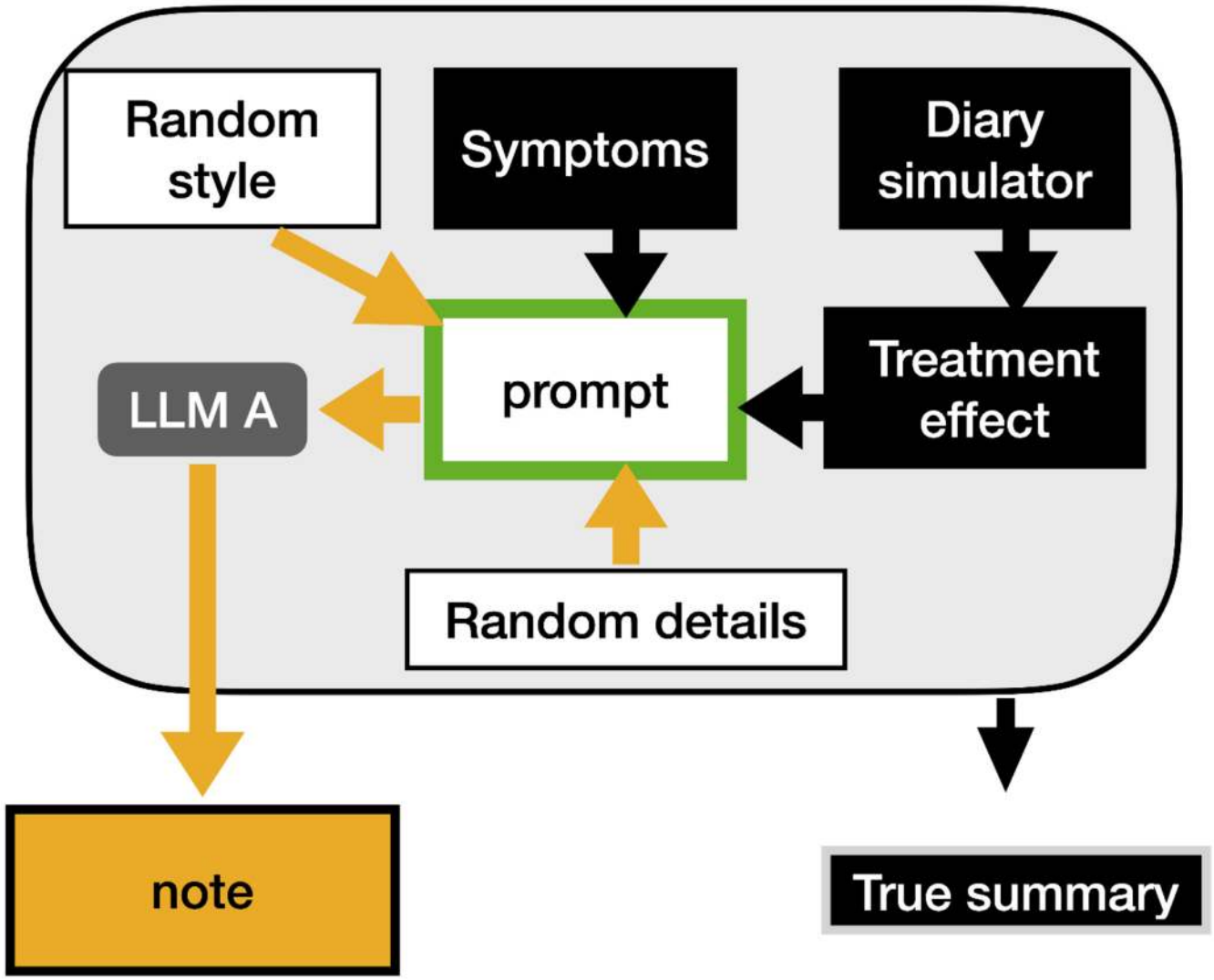

Figure 1:
Generation of clinical notes. The diary simulator (CHOCOLATES) was used to produce a realistic seizure diary. This was modulated by the treatment effect (0% in placebo arm, and 39% in drug arm) during the experimental maintenance stage. One of four writing styles were chosen, and a random set of reported symptoms were selected (based on previously reported incidence of symptoms for that arm). These items were used to generate the prompt submitted to LLM A (Llama 2:13b). The LLM generated the clinical note. The true summary was generated based on the original elements used to produce the prompt – specifically, the symptoms and seizure count after applying the pre-specified treatment effect.

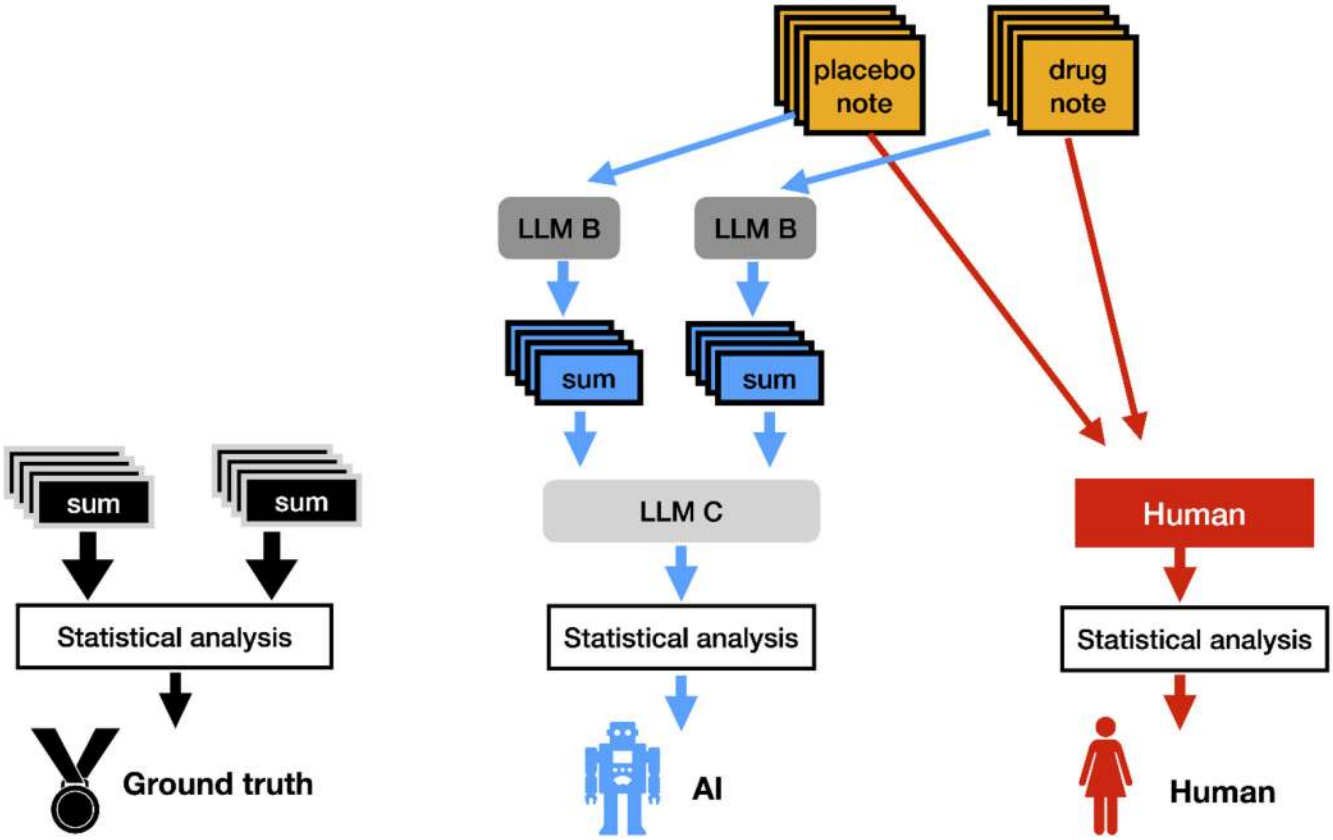

Figure 2:

Analyzing the trial. The Ground truth summaries (Figure 1) were used directly as a data table. The AI pathway took the clinical notes (Figure 1), and then LLM B (Mistral) produced a summary that indicated the number of seizures and symptoms reported. LLM C (Claude 2) was used to further summarize and synthesize the brief summaries from LLM B into a complete data table. The clinical notes (Figure 1) were manually assessed by the Human to build a data table. The data tables from the Ground truth, the AI and the Human were analyzed in the standard statistical fashion.

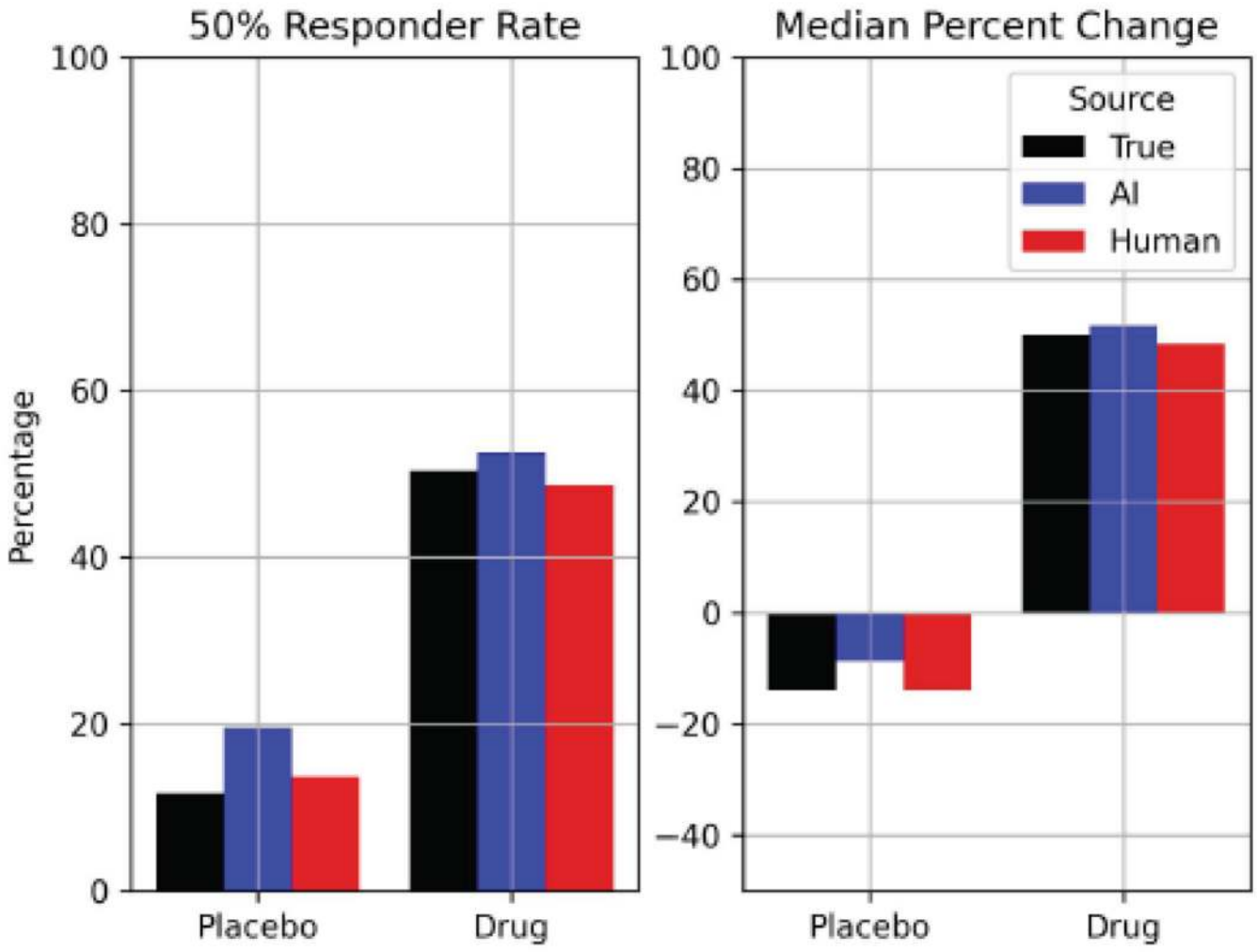

Figure 3:
Treatment effect. Shown here are the 50% responder rate (RR50) and the median percentage change (MPC) from the placebo and drug arms of the simulated study. Three colors are shown: ground truth (black), AI estimated (blue), and human reviewed (red). All three were similar though not identical. Nevertheless, both the AI and the human would conclude that the drug is dramatically better than placebo.

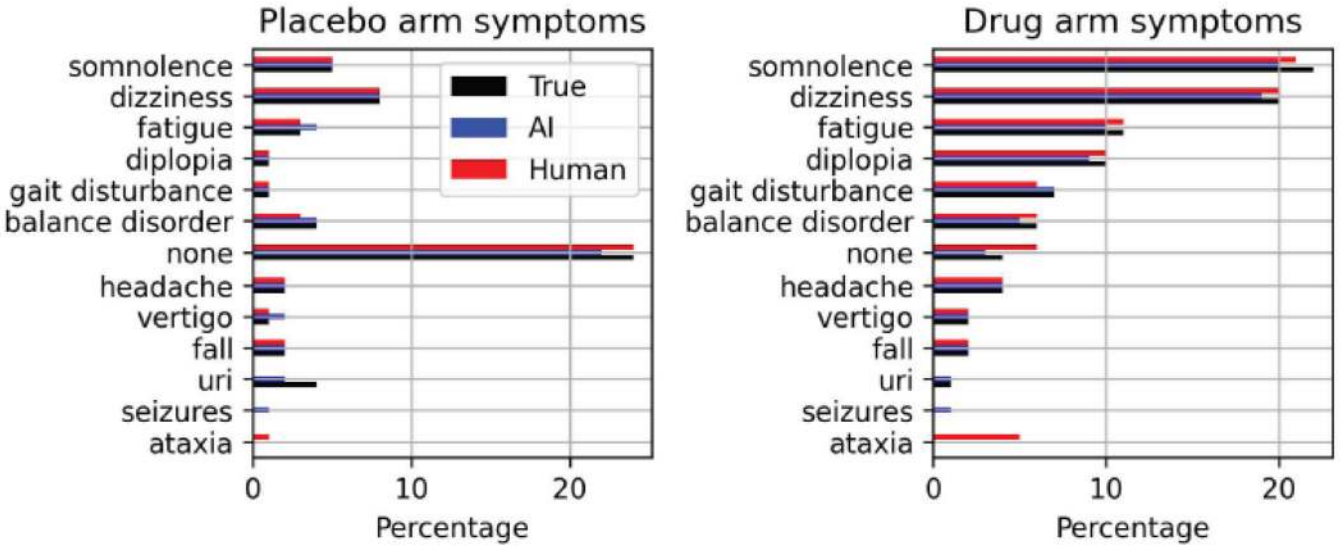

Figure 4:

Symptom list. Shown here are the symptoms found in either drug or placebo groups. The ground truth (black), AI derived (blue), and human reviewed (red) bars indicate the fraction of each group that reported the specific symptom. Not all bars match, however the general trend is that they are within 3% of each other.

Table 1:

The marginal difference between placebo and drug efficacy using the 50%-responder method (RR50) or the median percentage change (MPC) methods.

	Ground truth	AI	Human
RR50[a]	38% $p=1*10^{-10}$	34% $p=3*10^{-7}$	35% $p=8*10^{-8}$
MPC[b]	54% $p=1*10^{-15}$	61% $p=8*10^{-11}$	62% $p=1*10^{-12}$

[a] RR50p values are computed using Fisher Exact Test.

[b] MPC p values were computed using Mann-Whitney U test.