

<https://doi.org/10.1038/s44400-025-00043-x>

Comparing Alzheimer disease phenotype extraction using rule-based natural language processing, GPT-4, Phi-4, LLaMA, and DeepSeek



Inez Y. Oh¹ ✉, William J. B. Powell¹, Suzanne E. Schindler², Nupur Ghoshal^{2,3}, Albert M. Lai¹, Philip R. O. Payne¹ & Aditi Gupta¹

Clinical management of Alzheimer disease (AD) leverages information in unstructured narratives. We compared large language models, GPT-4, Phi-4-14b, DeepSeek-R1-Distill-LLaMA-8b, and LLaMA-3.2-3b, against a previous rule-based pipeline for extracting AD-related phenotypes from 100 real-world clinical notes. Evaluated against clinician annotations, GPT-4 performed best across phenotypes (median precision = 0.96, recall = 1, F1 = 0.98), but rule-based natural language processing was most precise (median = 0.97). Model choice depends on context, considering trade-offs between precision, recall, and scalability.

Accurate information extracted from unstructured clinical notes in electronic health records (EHR) is valuable for a variety of tasks including cohort identification for research and targeted clinical interventions, identification of disease risk factors, and generation of computable phenotypes for predictive modeling of disease outcomes (reviewed in ref. 1). Within Alzheimer disease (AD) research, there has been interest in leveraging machine learning techniques to develop early prediction models for diagnosing AD² and identify clinical factors modulating AD vulnerability³, but these studies have used only structured EHR data variables. For AD and other forms of dementia, many of the informative phenotypes clinicians use to rate dementia severity, formulate differential diagnoses, recommend appropriate testing, and determine clinical trial and treatment eligibility⁴, including cognitive concerns, behavioral changes, and personal or family medical history^{5–7}, are often only contained within clinical notes. Extracting these phenotypes at scale for downstream uses such as computational predictive modeling and screening for clinical trial or treatment eligibility requires automated phenotype extraction techniques, particularly natural language processing (NLP).

NLP approaches to clinical phenotype extraction from narrative text include rule-based, machine learning-based, deep learning-based, and hybrid techniques^{8–12}. In recent years, pre-trained transformer-based large language models (LLMs) have risen to prominence due to their impressive text-generation abilities and have become widely popular due to their user-friendly chat interfaces. Within the healthcare domain, studies are actively exploring the use of LLMs for various NLP tasks such as text generation for

patient communication^{13–15}, patient record summarization¹⁶, and clinical phenotype extraction from clinical narratives¹⁷, with promising results.

To extract phenotypes relevant to AD and other dementias, including presence of behaviors indicating cognitive decline (repeating questions and statements, misplacing objects), comorbidities (hypertension, depression), family history of dementia, brain imaging findings (atrophy, infarct), cognitive assessment scores (Mini-Mental State Examination [MMSE], Clinical Dementia Rating [CDR[®]]), and biomarker laboratory results (cerebrospinal fluid [CSF] total tau, phosphorylated tau concentrations), we previously developed a rule-based NLP pipeline evaluating the pipeline against gold-standard annotations of 100 unstructured clinical notes from a cohort of AD patients independently produced by two clinical experts as part of that study (Cohen's kappa = 0.72–1)¹⁸. Considering the popularity and reported performance of LLMs across diverse domains with minimal implementation effort¹⁹, in this brief study, we sought to compare the ability of LLMs, specifically a cloud-based implementation of OpenAI's Generative Pre-trained Transformer (GPT)-4²⁰ and local implementations of Microsoft's Phi-4-14b (Phi-4)²¹, DeepSeek-R1-Distill-LLaMA-8b (DeepSeek)²², and Meta's LLaMA 3.2-3b (LLaMA)²³, to extract phenotypes relevant to AD, against the previously developed a rule-based NLP pipeline¹⁸ (Fig. 1). These pre-trained models were chosen to include popular industry benchmark LLMs including both cloud-based (GPT-4) and locally-implementable open-source (Microsoft's Phi-4; Meta's LLaMA; and DeepSeek) models of varying parameter sizes.

¹Institute for Informatics, Data Science, and Biostatistics, Washington University School of Medicine, St. Louis, MO, USA. ²Department of Neurology, Washington University School of Medicine, St. Louis, MO, USA. ³Department of Psychiatry, Washington University School of Medicine, St. Louis, MO, USA.

✉ e-mail: iyoh@wustl.edu

Fig. 1 | Workflow diagram. Alzheimer disease-related clinical phenotypes were extracted using LLMs (GPT-4, Phi-4, DeepSeek, LLaMA), evaluated against gold-standard manual annotations, and the results compared to those of an existing rule-based NLP pipeline for the same task.

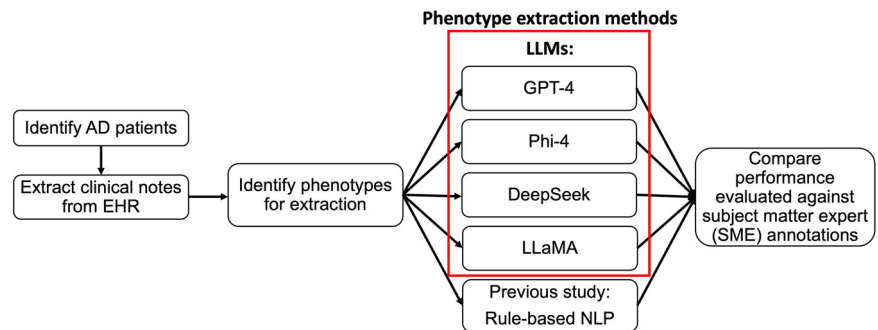
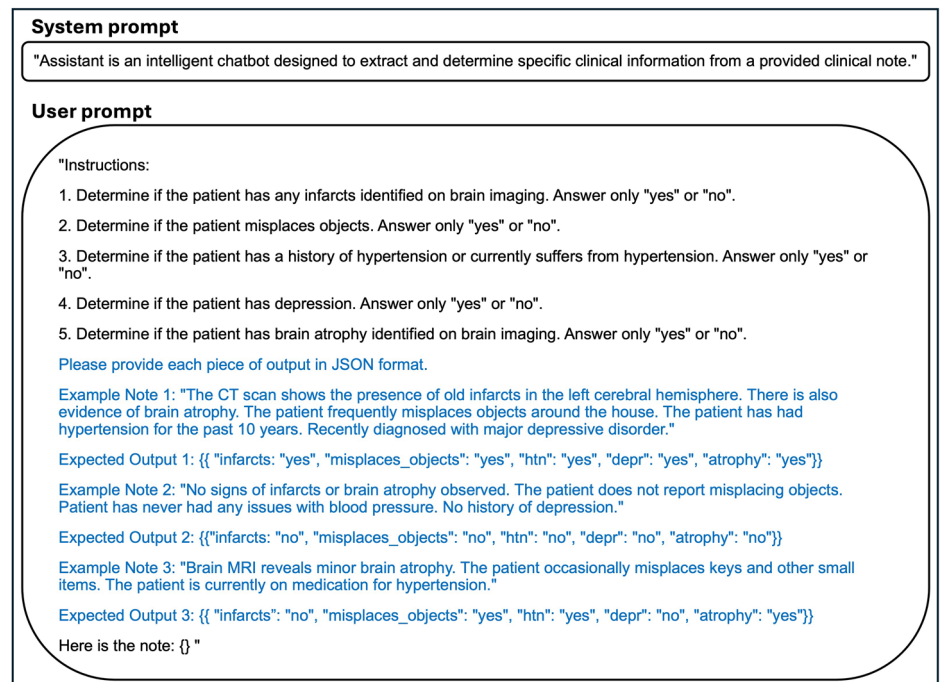


Fig. 2 | Comparison of zero-shot versus few-shot sample prompts. Black text indicates text included in both zero-shot and few-shot prompts. Blue text was included only in few-shot prompts.



For this brief study, we limited the scope to investigating the base capabilities of the LLMs, providing minimal guidance to each model using prompts optimized to induce models to return concise responses (e.g. integer values, Yes/No) (Fig. 2). Experiments comparing zero-shot and few-shot prompt strategies (Supplementary Tables 1, 2) showed similar performance trends across models and phenotypes (Supplementary Fig. 2), and differences between the two prompt strategies were not significant (Supplementary Table 5). Consequently, we focused on the zero-shot results; few-shot results are available in Supplementary Figs. 3–5 and Supplementary Tables 8, 9.

Evaluated against the gold-standard annotations, the rule-based method had the best precision (median = 0.98; range = 0.71–1) across target phenotypes, albeit not significantly better than GPT-4 and Phi-4, while GPT-4 produced significantly better recall (median = 1; range = 0.99–1) and F1 (median = 0.98; range = 0.82–1) than all other methods (Figs. 3–6; Supplementary Table 7). Overall, DeepSeek appeared to have the least consistent performance across phenotypes, evidenced by the wide range of precision (0.67–0.98), recall (0.5–0.98), and F1 (0.63–0.98) across phenotypes (Figs. 4–6).

Our results showed that even without sophisticated prompt engineering and fine-tuning, the LLMs could attain similar or better F1-scores compared to the rule-based pipeline, which had undergone iterative improvements during development (Figs. 3–6). As seen in Fig. 3, performance across LLMs appeared to approximately correlate with the number

of model parameters, as GPT-4 has the most parameters (~1.8 trillion), followed by Phi-4 (14 billion), DeepSeek (8 billion), and LLaMA (3 billion). While GPT-4 always traded higher recall for lower precision (Figs. 4, 5), Phi-4, LLaMA, and DeepSeek, despite achieving lower median F1-scores, displayed higher precision than recall for some phenotypes (e.g., hypertension, repeating), as did the rule-based method. All methods excelled at extracting numeric values such as test scores (MMSE, CDR) and biomarker lab values (total and phosphorylated tau levels), producing precision, recall, and F1 scores >0.90. All methods were generally successful at determining the presence/absence of phenotypes (hypertension, depression, misplacing, repeating, atrophy), except for infarct, which was an outlier that all models struggled with due to poor precision (F1 = 0.35–0.82). GPT-4 and Phi-4 successfully extracted the family history of dementia, which LLaMA, DeepSeek, and the rule-based method struggled to extract. Of note, DeepSeek's architecture significantly differs from the other LLMs as it uses a Mixture-of-Experts approach with sparse transformers versus dense transformers. This may contribute to DeepSeek's greater performance discrepancies when extracting simpler phenotypes (e.g., MMSE/CDR scores) versus more complex phenotypes (e.g., infarct, family history), where the relevance of information may not be as apparent without the full context.

We made several noteworthy observations based on a sample of errors by prompting the LLMs to provide an explanation along with the final answer (explanatory prompt; Supplementary Table 1). We found that some

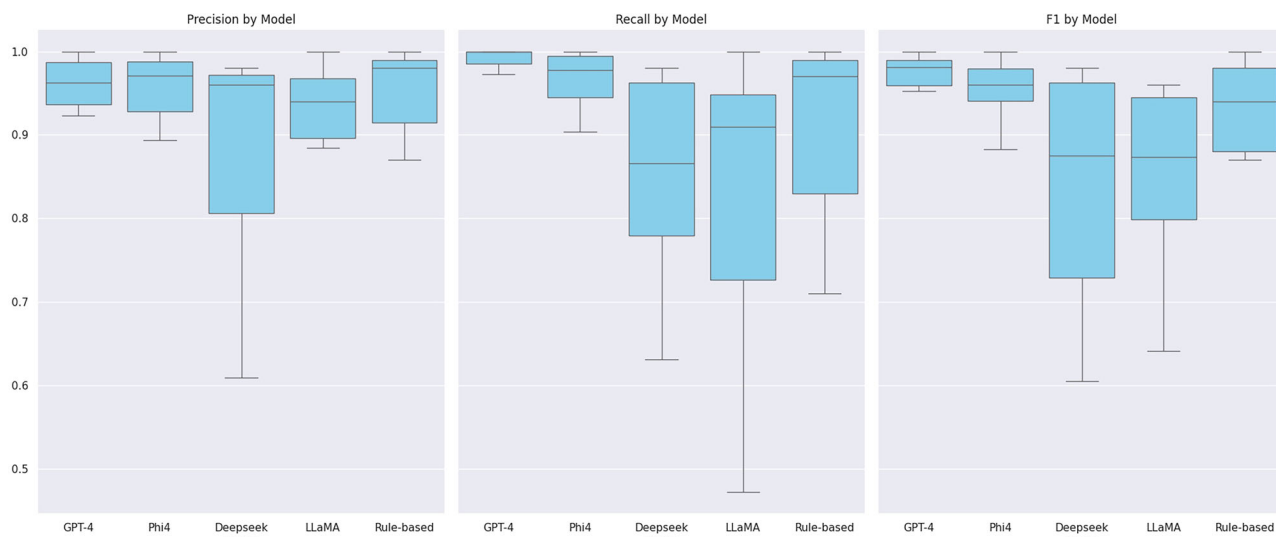
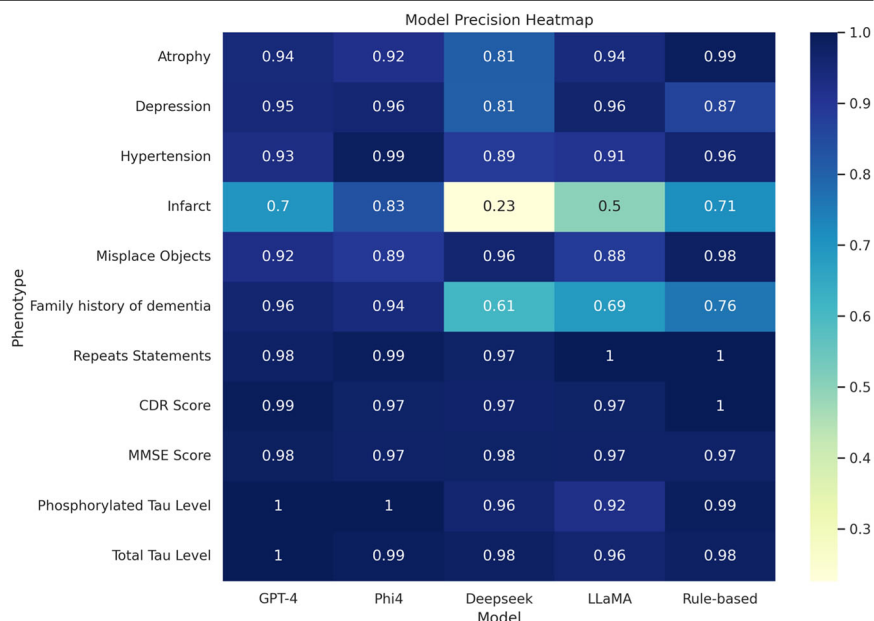


Fig. 3 | Comparison of precision, recall, and F1 for Alzheimer disease-related clinical phenotype extraction using GPT-4, Phi-4, DeepSeek, LLaMA, and an existing rule-based NLP pipeline. Box and whisker plots represent the distribution of metrics across target phenotypes, independently evaluated against gold-standard

manual annotations from each annotator, then averaged. Median and interquartile range (IQR) (box) are indicated, with the whiskers indicating 1.5 IQRs of the lower and upper quartiles. Outliers have been excluded from the figure. LLMs are ordered from left to right according to decreasing number of model parameters.

Fig. 4 | Heatmap comparing precision from few-shot prompting of GPT-4, Phi-4, DeepSeek, LLaMA, and the rule-based pipeline. Values in cells represent precision for each pipeline and phenotype calculated independently against each set of annotations, then averaged.



“errors” made by the LLMs were information missed by annotators (e.g., atrophy evidenced by CT scan results reporting “age appropriate volume loss”), demonstrating an advantage of LLMs over rule-based techniques, which rely on prior knowledge of expected documentation practices to achieve high accuracy but fail to generalize to unseen documentation patterns. Meanwhile, LLMs have been trained on large corpora of varied texts to identify patterns in unstructured text to predict output without needing to understand the text’s underlying meaning¹⁷, allowing them to be reasonably successful with minimal prompts and no additional domain-specific training data.

Many of the errors made by LLMs were due to hallucinations^{24,25}, where the LLM seemed to connect concepts present in the text to draw inaccurate conclusions, suggesting the ability to make broad associations between certain clinical phenotypes and AD based on patterns seen in its training data, but lacking nuanced clinical knowledge to make accurate interpretations. For

example, the LLMs tended to struggle with determining if a clinical note contained evidence that an infarct had been identified on brain imaging. In several instances, GPT-4 inferred “hyperintensities” on MRI, suggesting “chronic small vessel ischemic disease”, or a patient history of cerebral infarctions or strokes to be evidence of infarcts seen on brain imaging. Additional prompt engineering and fine-tuning using real institutional data, such as teaching the model to exclude specific terms that commonly result in false positives (e.g., “do not consider T2 hyperintensities evidence of brain infarct”), could mitigate this issue. Further, we observed several occasions where the LLM provided an internally inconsistent explanation, stating “Yes”, but continuing to explain that no infarcts were identified on brain imaging, highlighting LLMs’ probabilistic nature compared to deterministic rule-based NLP. The poor precision and hallucinations observed in the extraction of this phenotype indicate that the use of automated phenotype extraction pipelines requires significant human oversight.

Fig. 5 | Heatmap comparing recall of GPT-4, Phi-4, DeepSeek, LLaMA, and the rule-based pipeline. Values in cells represent recall for each pipeline and phenotype calculated independently against each set of annotations, then averaged.

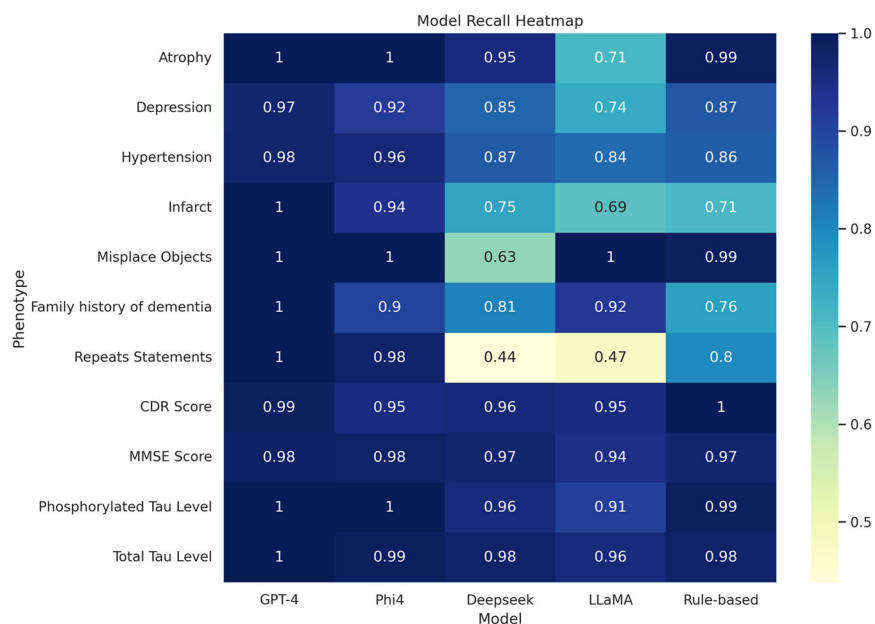
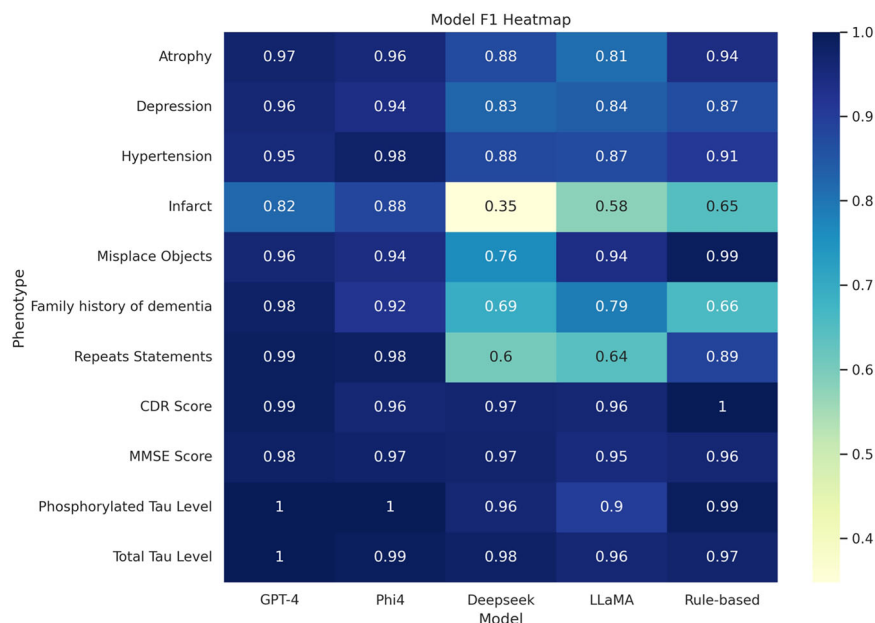


Fig. 6 | Heatmap comparing F1 of GPT-4, Phi-4, DeepSeek, LLaMA, and the rule-based pipeline. Values in cells represent F1 for each pipeline and phenotype, calculated independently against each set of annotations then averaged.



The trade-offs between precision and recall observed suggest that the choice of extraction method should depend on the extracted phenotypes' intended use. Diagnostic decision-making to identify patients for risky or expensive interventions such as amyloid-targeting therapies would require a high-precision method (i.e., rule-based NLP) as they are burdensome to provide and require careful monitoring²⁶. Determining eligibility for these treatments would require precise identification of inclusion and exclusion criteria to balance the opportunity to receive disease-altering therapy with the burdens and risks of treatment. On the contrary, a high-recall method such as GPT-4 would be a sensitive, albeit costly²⁷, tool for screening large cohorts for low-risk interventions (e.g., recommendations for lifestyle changes or further clinical evaluation). Models such as GPT-4 and Phi-4 would be suitable where speed of implementation and generalizability across diverse domains or institutions are essential and outweigh the need for precision, but the effort of optimizing a rule-based algorithm might be warranted for repeat use.

During this study, we observed marked differences in the time and cost of running the LLMs. The cloud-based proprietary GPT-4 model was much faster to run but far costlier than the local implementations of open-source LLMs such as Phi-4, whose performance was nearly on par with GPT-4 and would be cheaper but limited to locally available computational resources. This has current implications for scalability depending on whether time- or cost-savings is of greater consideration.

We anticipate that future work, including fine-tuning with real institutional data, further prompt engineering, and optimizing computational efficiency, such as using larger, quantized models, will improve the performance achievable using open-source LLMs. Future directions also include examining the consistency of results achievable by LLMs, integrating rule-based approaches with LLMs to leverage the strengths of both methods, assessing the abilities of LLMs to extract additional phenotypes and generalize across institutions, and testing these LLMs on additional clinical NLP tasks beyond the AD domain.

Methods

Dataset

This study was approved by the Washington University (WU) Institutional Review Board under protocol #201905161, which was granted a waiver of HIPAA Authorization for the use of Protected Health Information (PHI). Our dataset contained 100 clinical notes from the WU School of Medicine legacy Allscripts Touchworks EHR system associated with randomly selected outpatient office visits at the Memory Diagnostic Center for unique patients with cognitive concerns between 2013–2018. The note types included New Patient and Established Patient notes. There were no specific exclusion criteria. Further details on note selection were previously described¹⁸. Supplementary Fig. 1 and Supplementary Table 2 show the distribution of note length.

Phenotypes of interest

Phenotypes relevant to AD were previously identified by clinical dementia specialists (SES, NG)¹⁸. These phenotypes included medical comorbidities (hypertension, depression), biomarkers (cerebrospinal fluid [CSF] total tau [ttau] and phosphorylated tau [ptau] concentrations), cognitive assessments (Clinical Dementia Rating [CDR®], Mini-Mental Status Exam [MMSE]) given matching test dates, behavioral indicators of cognitive decline (misplacing objects, repeating statements), family history of dementia, and findings from brain imaging (atrophy, infarct). Supplementary Table 3 shows the expected output format for the extraction of each phenotype.

Rule-based phenotype extraction

A commercial rule-based NLP platform (Linguamatics I2E; IQVIA), which leverages public and proprietary ontologies, was used to extract each phenotype of interest. Details of the phenotype extraction were previously described¹⁸.

Phenotype extraction using Cloud-Based LLMs - GPT

To extract phenotypes of interest using GPT-4, we used Microsoft's Azure OpenAI Service, which provides REST API access to OpenAI's language models. We deployed the OpenAI API endpoint via a HIPAA-compliant subscription within Washington University's (WashU's) Azure tenant. We applied for and received an exemption from content filtering, abuse monitoring, and human review of our use of the Azure OpenAI service, removing the ability of Microsoft employees to perform any form of data review. Keeping model parameters constant (temperature = 0, top_p = 0.95, frequency_penalty = 0, presence_penalty = 0), we provided GPT with prompts (concise prompt) to extract each of the phenotypes using a zero-shot (Supplementary Table 1) and few-shot (Supplementary Table 2) approach and return results in a structured JSON format that were used in the evaluation. For error analysis, a prompt designed to induce a verbose, explanatory output was used (Supplementary Table 1).

Phenotype extraction using Locally Implemented LLMs - LLaMA, Phi-4, DeepSeek

To evaluate NLP options with improved possibility of scalability, we evaluated a series of open-source LLMs. Models (LLaMA-3.2-3b, Phi-4-14b, and DeepSeek-R1-Distill-LLaMA-8b) were downloaded from the Ollama Hub repository²⁸, and run locally on WU's Research Infrastructure Services Scientific Compute Platform using a NVIDIA A100-SXM4_40GB GPU. Ollama model files were modified to increase the context window limit to 120k tokens to accommodate the clinical notes. The prompts follow the zero-shot and few-shot approaches employed by the GPT extraction pipeline (Supplementary Tables 1, 2). Outputs of the local LLMs were sometimes not in the desired JSON format; this was resolved by passing the outputs to the LLM again to return properly formatted JSON output that could be used in the evaluation.

Manual annotation and evaluation of phenotype extraction performance

Phenotype extraction results were evaluated as previously described¹⁸. Results from each pipeline were compared independently against each of

two subject matter experts' (SME's) manual annotations. The SMEs were blinded to each other's annotations and to the pipeline's results. As previously reported, inter-annotator agreement was high (Cohen's kappa = 0.72–1), thus we report performance metrics (precision, recall, and F1-score) averaged across annotators. Precision and recall assess a pipeline's ability to correctly identify positive cases and retrieve all positive cases, respectively, while F1-score balances precision with recall; all three metrics were calculated using the scikit-learn package (version 0.22.1).

Tests of statistical significance

Differences in model performance were evaluated for significance using Wilcoxon signed rank tests for pairwise comparisons between models or between prompting strategies. Pairwise comparisons were performed across all models for their zero-shot and few-shot performance, and within models for if the zero-shot or few-shot performance was better for that model. 95% confidence intervals for were calculated using Wilson score intervals for precision, recall, and F1 score for all models. Wilson score intervals were employed for their ability to handle small sample sizes without violating assumptions. Analyses were performed using Python and the scipy stats package.

Data availability

The datasets utilized are not publicly available to protect patient privacy.

Code availability

Prompts used in this study are shown in Supplementary Tables 1, 2. Model parameters utilized are described in the Methods.

Received: 17 March 2025; Accepted: 6 November 2025;

Published online: 25 November 2025

References

- Pendergrass, S. A. & Crawford, D. C. Using electronic health records to generate phenotypes for research. *Curr. Protoc. Hum. Genet.* **100**, e80 (2019).
- Li, Q. et al. Early prediction of Alzheimer's disease and related dementias using real-world electronic health records. *Alzheimer's. Dement.* **19**, 3506–3518 (2023).
- Tang, A. S. et al. Deep phenotyping of Alzheimer's disease leveraging electronic medical records identifies sex-specific clinical associations. *Nat. Commun.* **13**, 675 (2022).
- Jack, C. R. et al. Revised criteria for diagnosis and staging of Alzheimer's disease: Alzheimer's Association Workgroup. *Alzheimer's. Dement.* **20**, 5143–5169 (2024).
- Kho, A. N. et al. Electronic medical records for genetic research: results of the eMERGE Consortium. *Sci. Transl. Med.* **3**, 79re1–79re1 (2011).
- Jensen, K. et al. Analysis of free text in electronic health records for identification of cancer patient trajectories. *Sci. Rep.* **7**, 1–12 (2017).
- WQ, W. et al. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. *J. Am. Med. Inf. Assoc.* **23**, e20–e27 (2015).
- Devlin, J., Chang, M. W., Lee, K. & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019–2019 Conf. North Am. Chapter Assoc. Comput. Linguist.: Hum. Lang. Technol.–Proc. Conf.* **1**, 4171–4186 (2018).
- Eyre, H. et al. Launching into clinical space with medspaCy: a new clinical text processing toolkit in Python. *AMIA Annu. Symp. Proc.* **2021**, 438 (2021).
- Peng, Y., Torii, M., Wu, C. H. & Vijay-Shanker, K. A generalizable NLP framework for fast development of pattern-based biomedical relation extraction systems. *BMC Bioinforma.* **15**, 285 (2014).
- Lee, S. et al. Hybrid method incorporating a rule-based approach and deep learning for prescription error prediction. *Drug Saf.* **45**, 27–35 (2022).

12. Liu, Z. et al. Entity recognition from clinical texts via recurrent neural network. *BMC Med. Inf. Decis. Mak.* **17**, 67 (2017).
13. Gordon, E. B. et al. Enhancing patient communication with Chat-GPT in radiology: evaluating the efficacy and readability of answers to common imaging-related questions. *J. Am. Coll. Radiol.* <https://doi.org/10.1016/J.JACR.2023.09.011> (2023).
14. Ayers, J. W. et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a Public Social Media Forum. *JAMA Intern. Med.* **183**, 589–596 (2023).
15. Semeraro, F., Schnaubelt, S., Montomoli, J., Bignami, E. G. & Monsieurs, K. G. Artificial intelligence in cardiopulmonary resuscitation: Driving awareness and debunking myths. *Resuscitation* **208**, 110539 (2025).
16. Nayak, A. et al. Comparison of history of present illness summaries generated by a Chatbot and senior internal medicine residents. *JAMA Intern. Med.* **183**, 1026–1027 (2023).
17. Bhattarai, K. et al. Leveraging GPT-4 for identifying clinical phenotypes in electronic health records: a performance comparison between GPT-4, GPT-3.5-turbo and spaCy's Rule-based & Machine Learning-based methods. *bioRxiv* <https://doi.org/10.1101/2023.09.27.559788> (2023).
18. Oh, I. Y. et al. Extraction of clinical phenotypes for Alzheimer's disease dementia from clinical notes using natural language processing. *JAMIA Open* **6**, 14 (2023).
19. Bhattarai, K. et al. Leveraging GPT-4 for identifying cancer phenotypes in electronic health records: a performance comparison between GPT-4, GPT-3.5-turbo, Flan-T5, Llama-3-8B, and spaCy's rule-based and machine learning-based methods. *JAMIA Open* **7**, ooae060 (2024).
20. OpenAI et al. GPT-4 Technical Report. Preprint at <https://arxiv.org/abs/2303.08774> (2024).
21. Abdin, M. et al. Phi-4 Technical Report. Preprint at <https://arxiv.org/abs/2412.08905> (2024).
22. Guo, D., et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature* **645**, 633–638 (2025).
23. Touvron, H. et al. LLaMA: Open and Efficient Foundation Language Models. Preprint at <https://arxiv.org/abs/2302.13971> (2023).
24. Zhang, Y. et al. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. Preprint at <https://arxiv.org/abs/2309.01219> (2025).
25. Ye, H., Liu, T., Zhang, A., Hua, W. & Jia, W. Cognitive Mirage: A Review of Hallucinations in Large Language Models. Preprint at <https://arxiv.org/abs/2309.06794> (2023).
26. van Dyck, C. H. et al. Lecanemab in Early Alzheimer's disease. *N. Engl. J. Med.* **388**, 9–21 (2023).

27. Pricing. <https://openai.com/pricing>.

28. Ollama. <https://ollama.com/>.

Author contributions

I.Y.O.: Conceptualization, data curation, formal analysis, investigation, methodology, software, visualization, writing – original draft. W.J.B.P.: Formal analysis, investigation, methodology, software, visualization, writing – review and editing. S.E.S. and N.G.: Investigation, writing – review and editing. A.M.L., P.R.O.P., A.G.: Conceptualization, resources, supervision, writing – review and editing.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s44400-025-00043-x>.

Correspondence and requests for materials should be addressed to Inez Y. Oh.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025