Fine-tuning Large Language Models for Rare Disease Concept Normalization

Andy Wang[1,2#], Cong Liu, PhD[1], Jingye Yang, PhD[3], Chunhua Weng, PhD[1*]

[1]Peddie School, Hightstown, NJ, USA

[2]Department of Biomedical Informatics, Columbia University, New York, NY, USA

[3]Department of Mathematics, University of Pennsylvania, Philadelphia, PA, USA

#: This author does not have an academic degree

***: Correspondence should be addressed to Chunhua Weng, 622 West 168 Street, PH-20-room 407, New York, NY 10032, cw2384@cumc.columbia.edu, 212- 305-5334**

Keywords: Large language model, concept normalization, LLaMA, HPO, OMIM, fine-tuning

Word Count:

ABSTRACT

Objective

We aim to develop a solution for rare disease concept normalization based on fine-tuning LLaMA 2, an open-source large language model (LLM), using a domain-specific corpus.

Methods and Materials

We fine-tuned four LLaMA2 models, each comprising seven billion parameters, using sentences incorporating clinical concepts from the HPO and OMIM vocabularies. The fine-tuning was conducted on four NVIDIA A100 GPUs.

Results

All models proved resilient to newly prompt-engineered sentences not used in the fine-tuning, achieved nearly perfect accuracies when prompted with original training data, and exhibit some robustness to typos. We tested each model on concepts they had not been trained on. The non-synonym HPO model fine-tuned without synonyms achieved 25.2% accuracy, while the synonym HPO model, fine-tuned with half the synonyms, achieved 85.6% accuracy. When tested against concept synonyms from SNOMED-CT, the non-synonym model achieved an accuracy of 33.9% while the synonym model improved to 57.4%. Synonyms proved challenging to both non-synonym and synonym OMIM models. ChatGPT 3.5 correctly identified HPO IDs for four out of 20 prompts.

Discussion

Our increasingly fine-tuned models demonstrated growing robustness to challenges such as misspellings, synonyms, and concepts from other ontologies. Incorrect outputs stem from tokens in the input that the models have never encountered, such as parenthesis. Many synonyms do not share the same semantic meaning and often include abbreviations.

Conclusion

Our fine-tuned LLaMA 2 models provide the capability to identify variations in medical concepts from clinical narratives while successfully normalizing them to a standard concept.

BACKGROUND AND SIGNIFICANCE

Rare diseases exhibit complex patterns of phenotypes and genetic heterogeneity. Considering rare diseases' rarity individually and commonality collectively, both their phenotypes and disease names are often documented in diverse ways. For instance, one doctor might describe a patient as experiencing "hearing loss," while another doctor might characterize the same patient as having "difficulty hearing." A lack of standardization in clinical vocabulary can lead to oversight in diagnosis and errors during treatment [1, 2, 3] . Adopting a standardized clinical vocabulary would make patient data more accessible to interpret and share and has the potential to improve patient outcomes [4]. More importantly, in research environments, standardized clinical concepts are crucial for efficiently and accurately studying trends and outcomes. The presence of heterogeneous and unstandardized clinical data combined with insufficient comprehensive rare disease data only hampers and reduces the efficiency of medical research, which consequently may compromise the quality and reliability of clinical findings [5, 6, 7, 8, 9]. The adoption of a standardized vocabulary holds the promise of simplifying clinical data significantly, allowing researchers to easily compare and analyze data across multiple medical settings and databases, accelerating medical research [10, 11, 12].

While several standardized vocabularies for rare diseases, such as the Human Phenotype Ontology (HPO) and Online Mendelian Inheritance in Man (OMIM) have been established, their integration into clinical settings remains infrequent [13, 14]. This scarcity of adoption poses challenges in gathering rare disease data through existing structured clinical databases. Consequently, researchers often find themselves in the position of manually phenotyping patients using standardized vocabularies or, on a larger scale, employing Natural Language Processing

(NLP) tools to recognize these standardized concepts within clinical narratives. In the latter scenario, the traditional approach typically involves a two-step process. First, the machine must identify relevant medical entities within sentences or paragraphs. Subsequently, these identified terms are mapped or normalized to a desired standard vocabulary (such as translating "Gastroparesis" to "HP:0002578" or "Tangier disease" to "OMIM:205400"). For instance, approaches like Doc2Hpo have utilized traditional NLP parsers, such as Metamap, to identify terms within clinical text and then employ a string-based methodology to normalize these terms to standardized HPO concepts [15, 16].

While NLP offers a solution, the effectiveness of traditional two-step processes is hindered when faced with slight modifications in clinical data and an inability to adapt to varying textual contexts. For instance, terms like "hearing loss" and "difficulty hearing" are not explicitly indexed as HPO names or synonyms. Therefore, traditional string-based normalization approaches may struggle to correlate them with the standardized HPO concept "Hearing Impairment" (HP:0000365). Consequently, there is a pressing need for the development of more adaptable NLP tools to address the challenges associated with clinical concept normalization. Recent advancements in Large Language Models (LLMs), such as ChatGTP come with incredible contextual interpretability abilities backed by a myriad of knowledge. However, while general-purpose LLMs, whether closed-source (e.g. ChatGPT) or open-source (e.g. LLaMA2 [17]), have advanced clinical term identification tasks, they are known to fabricate or "hallucinate" citations, references, and source links [18]. This limitation restricts their suitability for concept normalization.

Recent studies have provided compelling evidence that the fine-tuning of Large Language Models (LLMs) with specialized medical data sources can facilitate their adaptation to specific tasks within clinical settings [19, 20, 21]. For instance, Yang et al. successfully developed an LLM model from fine-tuning BERT and ChatGPT to extract and recognize HPO phenotypes in clinical texts within the presence of non-HPO phenotypes, typos, and semantic differences with the model's original training data [22]. In our study, we hypothesize that by fine-tuning LLMs using rare-disease-specific corpora and terminologies or ontologies, we can significantly augment their capacity to adeptly handle a myriad of synonyms and textual variations, thereby enabling them to more precisely capture the intricate nuances woven into clinical texts. Consequently, the fine-tuned model has the potential to offer a nonstop solution for the critical task of recognizing standardized concepts from clinical narratives, an imperative need in the field of rare disease patient phenotyping.

METHODS

Overview

Figure 1 provides an overview of the study design. We hypothesize that fine-tuned LLMs using a domain-specific corpus will help overcome the challenge of clinical concept normalization. We fine-tuned LLaMA2, comprising 7 billion parameters, using manually generated sentences derived from a collection of predefined templates incorporating clinical concepts sourced from the HPO and OMIM vocabularies (detailed in Data). In contrast to instruction fine-tuning, the LLaMA 2 model operates by completing a user input. For example, giving LLaMA 2 the prompt "The color of an apple is: " yields an output of "Red." We adopted this element when fine-tuning and evaluating our model. Our fine-tuning resulted in four separate models: two HPO models and two OMIM models. The initial HPO and OMIM models were fine-tuned using only standard concept names without providing synonyms. The second model variations were fine-tuned using standard concept names with half of a concept's associated synonyms. We assessed each model's performance by constructing various prompts, including standard concepts, concepts with spelling errors, as well as synonyms (not used in fine-tuning), and comparing the model's output to the correct IDs. We utilized the LLaMA2 base model and ChatGPT 3.5 as a benchmark to assess the performance of our fine-tuned model in comparison to the prevailing LLMs in current use.

Data source

The non-synonym corpus consisted of sentences generated by associating each concept's ID and only its standard concept name. Six sentence templates with varying levels of contextual complexity were utilized for training data generation. For example, the most complex training sentence pattern we used is “The Human Phenotype Ontology term Fibular hypoplasia is identified by the HPO ID HP:0003038.” Compared to another training sentence pattern, “The HPO term Fibular hypoplasia represents HP:0003038,” the latter sentence is much more concise: it abbreviated “Human Phenotype Ontology” to HPO, and many contextual words are removed. Both inputs have the same corresponding output of “HP:0003038.” This same sentence pattern was repeated for OMIM data. Furthermore, we constructed a synonym corpus that consisted of sentences generated by both standard concept names and some of their synonyms (as annotated in the vocabulary). In total, each HPO and OMIM name-based corpus was fine-tuned using 3,000 concepts with 18,000 sentences, and each name-plus corpus was fine-tuned using 1,000 concepts with around 12,000 sentences (Table 1).

Fine-tuning strategy

We utilized an autoregressive objective to fine-tune the two normalization models as the next token prediction task. The fine-tuning was conducted on 4 NVIDIA A100 GPUs, with significant speed-up through low-rank adaptation (LoRA). Earlier variants of the model underwent ten epochs to evaluate if the fine-tuning was functioning properly. Once the fine-tuning was confirmed to work, the number of training epochs gradually increased to assess how the model performs after more training. The data used to train the model, including the number of sentence variations and clinical concepts, also increased once the simplistic prototype models achieved functional results.

Hyper-parameters:

- LoRA: 8
- LoRA Alpha: 16
- LoRA Dropout: 0.05
- Learning rate: 0.0003
- Batch size: 128
- Microbatch size: 128
- Train steps: 40

EVALUATION OF THE MODELS

We assessed the performance of the models when presented with varied prompts and terms. The first part of the evaluation involved testing the models against different inputs via prompt engineering. Prompt engineering maintains the same concept names as used in the training data but changes the sentence structure. For example, our training data prompt "The [Human Phenotype Ontology/Online Mendelian Inheritance in Man] term [concept] is identified by the HPO ID" whereas the evaluation prompt is formulated as "[HPO/OMIM] ID of [concept] is." This allows us to evaluate how well the models performed given "foreign" prompts (i.e. setenence not seen in the training data) with the same concept names.

The second part of the evaluation assesses the model's adaptability to the alterations in concept names to which they have not previously been trained. Our fine-tuning prompt "[HPO/OMIM] ID of [concept] is" and evaluation prompt "[HPO/OMIM] ID of [concept*] is" include the same context and simplicity but the input [concept] differs. The modified concept names can be standard names, standard names with typos, synonyms, or associated terms found in another vocabulary such as SNOMED-CT [23]. Synonyms were sourced from a list of concept synonyms provided by the HPO and OMIM databases. For example, synonyms of "Hypoplastic hippocampus" include "Small hippocampus" and "Undeveloped hippocampus"; all three terms correlate to the HP:0025517 but differ semantically. All synonyms used during evaluation were not included in the fine-tuning process. Typos were introduced randomly by deleting one character from the original concept name. This enables us to more effectively assess models' practical utility in real-world applications, where typos and alterations to concept names are commonly encountered.

RESULTS

Before we began the fine-tuning procedure, we assessed the performance of the LLaMA2 base model. The LLaMA2 base model is unable to associate HPO and OMIM terms with their respective IDs (Figure 2). When inputted with prompts such as "The HPO ID of Lymphoproliferative disorder is:", the model outputted an arbitrary string of numbers unrelated to the HPO ID. The model produced the same output when inputted with OMIM terms.

Following 90 epochs of fine-tuning for the HPO and OMIM models, both achieved nearly perfect accuracies when prompted with original training data. In addition, both models exhibit robustness to prompt engineered inputs not used in training. Modifying the parts of the input sentence not involved with the concept term does not affect the model's ability to correspond a concept's name with its respective identifier, suggesting the most sensitive part of the input is the concept name. The HPO models performed poorly in terms of typos such as "vascular dilaton" or "vascular dilaion" instead of "vascular dilation", while the OMIM models tended to adapt better.

An interesting caveat we discovered from our fine-tuned model is its inability to perform when given more information. For example, the prompt "The HPO ID of the concept Fibular hypoplasia is HP: " contains an extra "HP: " at the end of it, which was not in the training data. This minor change in the input confuses the model and, for a majority of cases, results in the model's inability to correlate the prompted clinical concept and the concept's respective identifier.

The performance of fine-tuned LLaMA2 models

Table 2 shows the performance of various models in accurately identifying concept IDs (i.e. normalization) when presented with diverse prompt inputs. Both non-synonym and synonym HPO models (trained for 90 epochs) achieved accuracies of 99.6% and 99.7%, respectively, in identifying a term's HPO ID when prompted with the original training sentences. In many incorrect cases, the inputted HPO term names were often lengthy and contained commas such as "Low-set, posteriorly rotated ears." We suspect that this type of complex and long input could have confused the model and is the reason behind its incorrect identification. The models' resilience to newly prompt engineered sentences not used in the fine-tuning proved strong. The non-synonym model achieved accuracies of 99.5% and 93.4% and the synonym model achieved accuracies of 99.2% and 98.8%. When introducing typos into concept names, the performance decreased significantly in both models. The non-synonym model identified concept IDs with a 45.2% correction rate while the synonym model did not perform significantly better with a 54.2% accuracy. We tested both models on names they had not trained on. The non-synonym model achieved 25.2% accuracy, likely due to limited variation for a single ID. In contrast, the synonym model, fine-tuned with half the synonyms, performed much better at 85.6%. As an additional benchmark for the HPO models, we provided them with concept synonyms from SNOMED-CT. We maintained the same training sentence formatting but replaced the concepts with SNOMED-CT synonyms. Both non-synonym and synonym-trained models performed suboptimally with the former achieving an accuracy of 33.9% while the latter improved to 57.4%.

Similarly, both non-synonym and synonym OMIM models (90 epochs) achieved high accuracies on original training data (Table 2). When encountering sentence variations not present in fine-tuning, the non-synonym model achieved accuracies of 98.1% and 91.8%, while the synonym model achieved accuracies of 99.3% and 90.4%. However, the OMIM models performed strikingly differently from the HPO models when imputed with altered concept names than sentence contexts. Both models exhibit significantly more robustness to typos (72.4% and 70.8% accuracy). Synonyms on the other hand proved challenging to both models, resulting in low accuracies of 6.2% for the non-synonym model and 30.8% for the synonym model.

The performance of ChatGPT (GPT3.5)

Mainstream LLMs such as ChatGPT are excellent at concept identification tasks, but we wanted to analyze whether they also possess accurate concept normalization abilities. Using ChatGPT 3.5 as a benchmark, it correctly identified HPO IDs for four out of 20 prompts. The four correctly identified concepts are relatively common in clinical notes like “Diabetes mellitus HP:0000819” and “Hypertension HP:0000822.” The remaining 16 prompts consist of less commonly seen phenotypic features As a result, ChatGPT either claimed unfamiliarity, insisted the term did not exist, or generated imaginary (non-existent) HPO IDs. For example, when tasked to identify the HPO ID of “Vascular dilatation,” ChatGPT does not recognize the term as of its update in 2022. However, ChatGPT suggests a non-existent, “Arterial dilation,” with HPO ID, HP:0012824, which corresponds to the HPO concept “Severity.” ChatGPT not only hallucinates HPO IDs but the entire HPO concept names themselves. The hallucinated HP IDs follow the same formatting as the standard HPO ID, but the actual ID itself is incorrect. In other incorrect cases, ChatGPT claims the specific HPO term provided does not have a corresponding

HPO ID but has offshoots, such as “neoplasm” and “Abnormality of the upper arm.” Both of these examples have their respective HPO IDs but ChatGPT claims otherwise. Additionally, the offshoot HPO terms and IDs it provides are incorrect, similar to that observed in the case with “arterial dilation” noted above. In some cases, ChatGPT was close to generating the correct HPO ID but was incorrect to 2-3 decimal places, similar to incorrect cases in our fine-tuned LLaMA 2 model. Our benchmark of ChatGPT indicates that mainstream LLMs fail at clinical concept normalization tasks.

DISCUSSION

Compared to conventional national language processing algorithms, LLMs such as ChatGPT can effectively handle typos and variations in sentences, thereby enhancing their efficacy in identifying phenotypes within clinical narratives. However, mainstream LLMs like ChatGPT fail at clinical concept normalization tasks. Our fine-tuning LLM helps to bridge this caveat in standardizing clinical concept normalization. In the case of our fine-tuned LLaMA 2 model, we were able to achieve the first step in concept normalization by accurately associating a clinical concept’s name directly with its respective identifier in different ontologies including HPO and OMIM. Our increasingly fine-tuned models demonstrated growing robustness to normalizing clinical concepts, handling challenges such as misspellings, synonyms, and concepts from other ontologies like SNOMED-CT. However, several issues emerged during our evaluation, prompting considerations of additional improvement.

In earlier iterations of our model, the training data's output contained the digits only and did not include an "HP" or "OMIM" prefix. For example, the training input would be "The Human Phenotype Ontology term Fibular hypoplasia is identified by the HPO ID HP:" while the output would be "0003038." This absence of the ontology prefix in the output resulted in a low-accuracy model that struggles to correlate terms with their respective identifiers, generating outputs similar to that of the non-fine-tuned LLaMA 2 base model. This caveat was resolved after relocating the ontology prefix into the output with its corresponding numerical tag. Fine-tuning the model using this modified data drastically improved the accuracy of the model.

Throughout the fine-tuning process, we produced multiple variants of our fine-tuned model, each with varying amounts of training epochs and data. Our first variants trained on roughly 20 epochs had much lower accuracies than our current models but performed better than the LLaMA2 base model without fine-tuning. Generally, increasing the number of training epochs correlated with improved accuracy. Additionally, increasing the number of sentence variations contributed to the model's ability to grapple with inputs it was not trained with [24]. The earlier model variants, trained with only one training sentence, had an abysmal performance. However, introducing more diversified training sentences significantly improved the model's ability to normalize concepts when inputted with untrained sentences. The model demonstrates strong adaptability to distinct prompts and can accommodate modifications in inputs it has not been explicitly trained on.

Instances of incorrect answers from the model often stemmed from inaccuracies related to n-gram concepts with special tokens such as parentheses and hyphens. Of the four original training sentences, no specific input sentence had more or fewer errors than others. The errors seemed to sprout randomly and were not predictable. The model, however, had a subpar performance when tasked with input sentences with the least amount of textual context, suggesting more context in the input results in higher accuracies. The inaccurate results, however, were typically off by one or two digits from the end of the ID compared to the correct answer, indicating the model is close to associating those terms with their identifiers. This observation may be linked to the organizational structure of ontology IDs. Concepts with only the last digits differing often share the same parents in the concept hierarchy. This semantic closeness between two IDs could potentially contribute to errors in the ID identification task. Additionally, we noticed that the IDs were tokenized into digit-sized segments. This observation could explain the "last-digit" error, as LLMs ultimately aim to predict the next tokens. An alternative approach is to enhance fine-tuning by creating a customized tokenizer that treats the entire ID (e.g., HP:0004413) as a single token. This modification can potentially enable the model to capture more nuanced semantic relationships between concept names and their corresponding IDs.

Regarding the models' performances against synonyms, we identified multiple cases of how the model could have incorrectly identified concepts. Concerning the OMIM model, many concepts and their respective synonyms are not of the same semantic meaning without given the clinical context. For example, "Lou Gehirg's disease" and "ALS" share the same ID. Similarly, many listed OMIM synonyms include abbreviations such as "CASIL" representing "cerebral arteriopathy, autosomal dominant, with subcortical infarcts and leukoencephalopathy, type 1." Those abbreviations would even pose a challenge for humans to identify accurately without the underlying clinical context. Given that the prompts evaluated in this study lack clinical context, future efforts should focus on constructing prompts using clinical narratives. This will help assess whether abbreviations, like those observed in OMIM synonyms, can be accurately normalized within a more realistic clinical setting.

Another source of errors could be due to the commonality of commas present in OMIM terms or the omission omitting “type x” to “x” at the end of a certain concept’s name. The SNOMED-CT synonyms used in the HPO model evaluation presented challenges for both models. One possibility can be that a majority of the SNOMED-CT synonyms all include a hyphen pointing at a specialty of the concept. For example, the SNOMED-CT synonym of “Abnormality of the kidney” includes “Kidney - Abnormal” and “Kidney structure - Defect”, both of which have the hyphen as key to the concept meaning. Since almost none of the fine-tuning data for the non-synonym and synonym HPO models included hyphens, it suggests a potential reason why the models performed poorly when handling terms with hyphens.

CONCLUSION

Our fine-tuned LLaMA 2 model further advances the concept normalization task by linking identified phenotype terms with their respective identifiers. It provides the capability to identify variations in the writing of medical concepts from clinical narratives while successfully normalizing them to a standard concept. In a clinical setting, standardized phenotypic concepts can be used by other informatics tools to identify disease-causal variants, rank candidate diseases, and forecast disease risk, thereby improving diagnostic and treatment accuracies. We plan to augment the model by incorporating more data such as genes, drugs, and phenotypes, to standardize enormous amounts of data. The model has the potential to generate knowledge graphs from narratives by linking diseases, phenotypes, genes, and drugs in a standard manner, therefore revealing previously unestablished relationships and outcomes.

Acknowledgments

We thank Dr. Herbert Chase for the mentorship in the Columbia University Summer Research Program.

Funding

The study is supported by NIH grants LM012895 and HG013031.

Conflicts of Interest

The authors declare no competing interests.

Data Availability

The software code and the fine-tuned LLaMA 2 models (as asset files in software release) are available on GitHub (https://github.com/andywang-25/LLaMA2-Fine-tuning/tree/v0.0.1)

References

1. Wąsiewicz P, Skalski M, Fornal-Pawłowska M. Chronic insomnia cases detection with the help of Athens Insomnia Scale and SF-36 health survey. 2011;8008. doi: 10.1117/12.905587.

2. Tang G, Liu T, Cai X, Gao S, Fu L. Standardization of clinical terminology based on hybrid recall and Ernie. In: Proceedings of the 3rd International Symposium on Artificial Intelligence for Medicine Sciences; 2022. doi: 10.1145/3570773.3570782.

3. Meijlink J. Patient-centred standardization in interstitial cystitis/bladder pain syndrome—a PLEA. Transl Androl Urol. 2015;4:499-505. doi: 10.3978/j.issn.2223-4683.2015.08.02.

4. Mirsaeidi M, Vu A, Leitman P, Sharifi A, Wisliceny S, Leitman A, Schmid A, Campos M, Falkinham J, Salathe M. A patient-based analysis of the geographic distribution of Mycobacterium avium complex, Mycobacterium abscessus, and Mycobacterium kansasii infections in the United States. Chest. 2017;151(4):947-50. doi: 10.1016/j.chest.2017.02.013.

5. Pariser A, Gahl W. Important role of translational science in rare disease innovation, discovery, and drug development. J Gen Intern Med. 2014;29:804-7. doi: 10.1007/s11606-014-2881-2.

6. Tingley K, Coyle D, Graham I, Sikora L, Chakraborty P, Wilson K, Mitchell J, Stockler-Ipsiroglu S, Potter B. Using a meta-narrative literature review and focus groups with key stakeholders to identify perceived challenges and solutions for generating robust evidence on the effectiveness of treatments for rare diseases. Orphanet J Rare Dis. 2018;13. doi: 10.1186/s13023-018-0851-1.

7. Wilson D, Hampton-Bagshaw K, Jorwic T, Bishop J, Giustina E. A new focus on process and measure. Raising data quality with a standard coding workflow and benchmarks. J AHIMA. 2008;79(3):54-6, 58.

8. Garcelon N, Neuraz A, Salomon R, Bahi-Buisson N, Amiel J, Picard C, Mahlaoui N, Benoit V, Burgun A, Rance B. Next generation phenotyping using narrative reports in a rare disease clinical data warehouse. Orphanet J Rare Dis. 2018;13. doi: 10.1186/s13023-018-0830-6.

9. Tang C, Xu Y, Zhu Q. Data Normalization Improves Semantic Annotation – a Case Study of Rare Disease Name Annotation. In: 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2021. p. 2609-2611. doi: 10.1109/BIBM52615.2021.9669475.

10. Hudson LD, Kush R, Almario EN, Seigneuret N, Jackson T, Jauregui B, Jordan D, Fitzmartin R, Zhou FL, Malone J, Galvez J, Becnel L. Global standards to expedite learning from medical research data. Clin Transl Sci. 2018;11:342-44. doi: 10.1111/cts.12556.

11. Mullin AP, Corey D, Turner EC, Liwski R, Olson D, Burton J, Sivakumaran S, Hudson L, Romero K, Stephenson D, Larkindale J. Standardized data structures in rare diseases: CDISC user guides for Duchenne Muscular Dystrophy and Huntington's Disease. Clin Transl Sci. 2020;14:214-21. doi: 10.1111/cts.12845.

12. Kodra Y, Weinbach J, Posada-de-la-Paz M, Coi A, Lemonnier S, Enckevort D, Roos M, Jacobsen A, Cornet R, Ahmed S, Bros-Facer V, Popa V, Meel M, Renault D, Gizycki R, Santoro M, Landais P, Torreri P, Carta C, Mascalzoni D, Gainotti S, Lopez E, Ambrosini A, Müller H, Reis R, Bianchi F, Rubinstein Y, Lochmüller H, Taruscio D. Recommendations for improving the quality of rare disease registries. Int J Environ Res Public Health. 2018;15. doi: 10.3390/ijerph15081644.

13. Chen L, Fu W, Gu Y, Sun Z, Li H, Li E, et al. Clinical concept normalization with a hybrid natural language processing system combining multilevel matching and machine learning ranking. J Am Med Inform Assoc. 2020 Oct 1;27(10):1576-84.

14. Silva JF, Antunes R, Almeida JR, Matos S. Clinical Concept Normalization on Medical Records Using Word Embeddings and Heuristics. Stud Health Technol Inform. 2020 Jun 16;270:93-7.

15. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp. 2001:17-21. PMID: 11825149; PMCID: PMC2243666.

16. Liu C, Peres Kury FS, Li Z, Ta C, Wang K, Weng C. Doc2Hpo: a web application for efficient and accurate HPO concept curation. Nucleic Acids Res. 2019 Jul 2;47(W1):W566-W570. doi: 10.1093/nar/gkz386. PMID: 31106327; PMCID: PMC6602487.

17. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, et al. LLaMA: Open and Efficient Foundation Language Models. arXiv. 2023 arXiv:2302.13971.

18. Alkaissi H, McFarlane SI. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. Cureus. 2023 Feb;15(2):e35179.

19. Y. Pawar, A. Henriksson, P. Hedberg and P. Naucler, "Leveraging Clinical BERT in Multimodal Mortality Prediction Models for COVID-19," 2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS), Shenzen, China, 2022, pp. 199-204, doi: 10.1109/CBMS55023.2022.00042.

20. Tinn R, Cheng H, Gu Y, Usuyama N, Liu X, Naumann T, Gao J, Poon H. Fine-tuning large neural language models for biomedical natural language processing. Patterns (N Y). 2023 Apr 14;4(4):100729. doi: 10.1016/j.patter.2023.100729. PMID: 37123444; PMCID: PMC10140607.

21. Kormilitzin A, Vaci N, Liu Q, Nevado-Holgado A. Med7: A transferable clinical natural language processing model for electronic health records. Artif Intell Med. 2021 Aug;118:102086. doi: 10.1016/j.artmed.2021.102086. Epub 2021 May 18. PMID: 34412834.

22. Jingye Yang, Cong Liu, Wendy Deng, Da Wu, Chunhua Weng, Yunyun Zhou, Kai Wang, Enhancing phenotype recognition in clinical notes using large language models: PhenoBCBERT and PhenoGPT, Patterns, 2023, 100887, ISSN 2666-3899, https://doi.org/10.1016/j.patter.2023.100887.

23. El-Sappagh, S., Franda, F., Ali, F. et al. SNOMED CT standard ontology based on the ontology for general medical science. BMC Med Inform Decis Mak 18, 76 (2018). https://doi.org/10.1186/s12911-018-0651-5

24. Lukas Berglund, Asa Cooper Stickland, Mikita Balesni, Max Kaufmann, Meg Tong, Tomasz Korbak, Daniel Kokotajlo, and Owain Evans. Taken out of context: On measuring situational awareness in llms, 2023.

Figure 1. Overview of fine-tuning and evaluation methodology

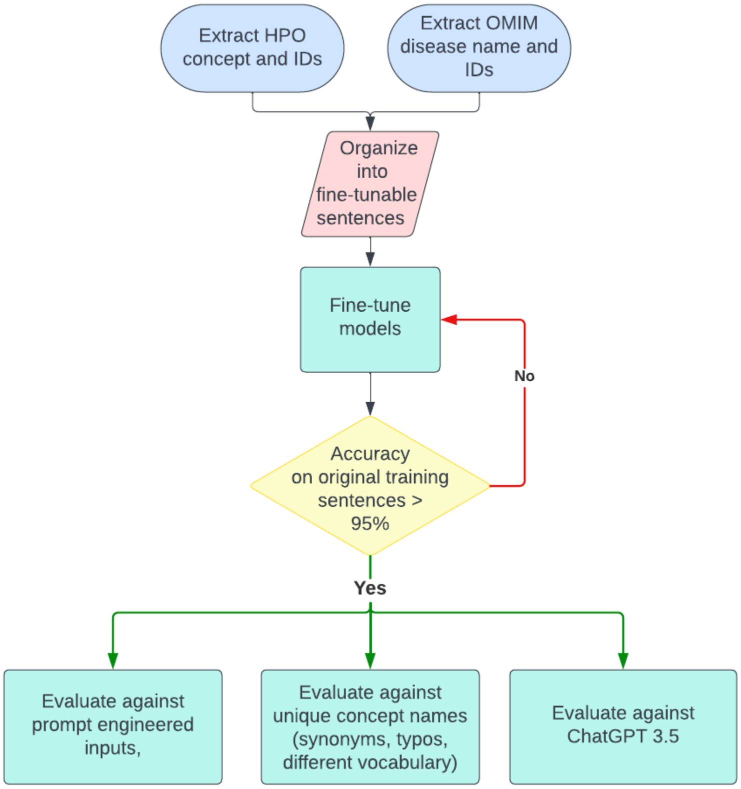
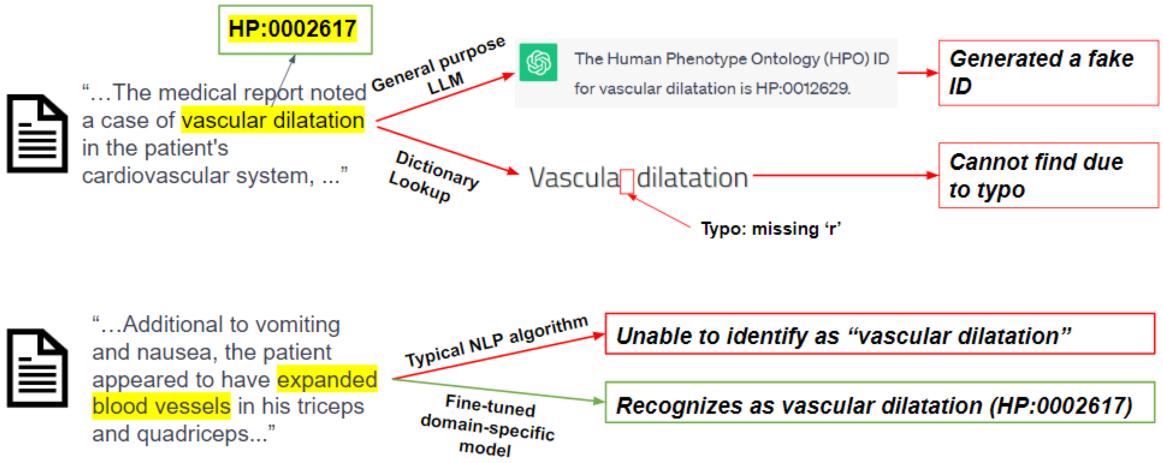

Extract HPO concept and IDs
Extract OMIM disease name and IDs
Organize into fine-tunable sentences
Fine-tune models
No
Accuracy on original training sentences > 95%
Yes
Evaluate against prompt engineered inputs,
Evaluate against unique concept names (synonyms, typos, different vocabulary)
Evaluate against ChatGPT 3.5

Figure 2. Examples of the ineffectiveness of traditional approaches and general-purpose LLMs (e.g. ChatGPT) at clinical concept normalization.



"…The medical report noted a case of vascular dilatation in the patient's cardiovascular system, …"

The Human Phenotype Ontology (HPO) ID for vascular dilatation is HP:0012629.

"…Additional to vomiting and nausea, the patient appeared to have expanded blood vessels in his triceps and quadriceps..."

Table 1. Examples of training data used to fine-tune the no-synonym HPO and OMIM model

Training Input	**Training Output**
The [Human Phenotype Ontology/Online Mendelian Inheritance in Man] term [concept] is identified by the HPO ID	HP/OMIM:xxxxxxx
The [HPO/OMIM] ID of [concept] corresponds to	HP/OMIM:xxxxxxx
The [HPO/OMIM] term [concept] represents	HP/OMIM:xxxxxxx
The [HPO/OMIM] ID of the concept [concept] is	HP/OMIM:xxxxxxx
[concept] has an [HPO/OMIM] ID of	HP/OMIM:xxxxxxx
[concept] is	HP/OMIM:xxxxxxx

Table 2. Performances of the different HPO and OMIM Models

Input Prompt	**Non-synonym HPO Model Accuracy**	**Synonym: HPO Model Accuracy**	**No synonym: OMIM Model Accuracy**	**Synonym: OMIM Model Accuracy**
Original Training Data	99.6%	99.7%	98.8%	99.8%
[HPO/OMIM] ID of [concept] is	99.5%	99.2%	98.1%	99.3%
[HPO/OMIM] [concept] is	93.4%	98.8%	91.8%	90.4%
Typo (one character deletion)	45.2%	54.2%	72.4%	70.8%
HPO/OMIM Synonyms	25.2%	85.6%	6.2%	30.8%
SNOMED-CT synonyms for HPO concepts	33.9%	57.4%	N/A	N/A