medRxiv preprint doi: https://doi.org/10.1101/2023.10.17.23297028; this version posted April 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

FHIR-GPT Enhances Health Interoperability with Large Language Models

Yikuan Li, MS[1,2], Hanyin Wang, BMed[1], Halid Z. Yerebakan, PhD[2], Yoshihisa Shinagawa, PhD[2], Yuan Luo, PhD[1]

[1] Division of Health and Biomedical Informatics, Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, Chicago, Illinois, USA

[2] Siemens Medical Solutions, Malvern, Pennsylvania, USA

Correspondence to:

Dr. Yuan Luo

750 N. Lake Shore Drive

11th floor

Chicago, Illinois 60611

Email: yuan.luo@northwestern.edu

Keywords: *Large Language Models (LLMs), Health Interoperability, Fast Healthcare Interoperability Resources (FHIR), Generative Pre-trained Transformer (GPT)*

Word Count: 1,890

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2023.10.17.23297028; this version posted April 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Abstract

Advancing health interoperability can significantly benefit health research, including phenotyping, clinical trial support, and public health surveillance. Federal agencies, including ONC, CDC, and CMS, have been collectively collaborating to promote interoperability by adopting Fast Healthcare Interoperability Resources (FHIR). However, the heterogeneous structures and formats of health data present challenges when transforming Electronic Health Record (EHR) data into FHIR resources. This challenge becomes more significant when critical health information is embedded in unstructured data rather than well-organized structured formats. Previous studies relied on multiple separate rule-based or deep learning-based NLP tools to complete the FHIR resource transformation, which demands substantial development costs, extensive training data, and meticulous integration of multiple individual NLP tools. In this study, we assessed the ability of large language models (LLMs) to transform clinical narratives into HL7 FHIR resources. We developed FHIR-GPT specifically for the transformation of clinical texts into FHIR medication statement resources. In our experiments using 3,671 snippets of clinical texts, FHIR-GPT demonstrated an exceptional exact match rate of over 90%, surpassing the performance of existing methods. FHIR-GPT improved the exact match rates of existing NLP pipelines by 3% for routes, 12% for dose quantities, 35% for reasons, 42% for forms, and over 50% for timing schedules. Our findings provide the foundations for leveraging LLMs to enhance health data interoperability. Future studies will aim to build upon these successes by extending the generation to additional FHIR resources.

medRxiv preprint doi: https://doi.org/10.1101/2023.10.17.23297028; this version posted April 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Introduction

Interoperability enhances the ability of healthcare providers to deliver safe, effective, and patient-focused care. It also offers novel avenues for individuals and caregivers to access electronic health data for care coordination and management [1]. The promotion of interoperability has become an integral aspect of various health initiatives, spanning from ensuring health equity to responding to public health emergencies [2]. Federal agencies, including the Office of the National Coordinator of Health IT (ONC) [1], the Centers for Disease Control and Prevention (CDC) [3], and the Centers for Medicare & Medicaid Services (CMS) [4], collectively collaborate to promote interoperability through the adoption of FHIR, which is an interoperability standard developed by the Health Level 7 (HL7®) standards development organization.[5] FHIR is specifically designed to facilitate the swift and efficient exchange of health data. FHIR has seen growing adoption in the modeling and integration of both structured and unstructured data for various health research purposes. Its applications range from developing computational phenotyping [6-8] to supporting clinical trials [9-12], building surveillance systems [13,14], and much more.

Transforming health data into the FHIR format presents a major challenge, as various health organizations have their unique infrastructure, standards, and formats for generating, storing, and organizing health data [15]. This challenge becomes more significant when critical health information is embedded in unstructured data other than well-organized structured formats. There are existing efforts for promoting the transformation of unstructured data into FHIR resources, offered by both academic and commercial sectors. In academic research, Hong et al.[16] integrated clinical NLP tools, including cTAKES [17], MedXN [18], and MedTime [18], to extract clinical entities from corresponding document sections and standardize them into FHIR

medRxiv preprint doi: https://doi.org/10.1101/2023.10.17.23297028; this version posted April 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

resources. Wang et al. developed Opioid2FHIR [19], a system that employs multiple deep learning-based natural language processing (NLP) techniques for opioid information extraction and normalization. In the commercial domain, Google Cloud has released the Healthcare Natural Language API [20], capable of converting medical text input into FHIR resources. Azure Health Data [21] is proficient at converting semi-structured data into FHIR resources but does not handle free-text unstructured input. All the above FHIR transformation tools necessitate multiple NLP tools in sequence. Creating a pipeline that integrates multiple NLP tools requires substantial computational resources, annotated data, and human efforts. Furthermore, as the transformation progresses along the pipeline, the errors from each NLP tool compound and decrease overall accuracy.

Therefore, we proposed leveraging pre-trained large language models (LLMs) and meticulous prompt engineering to facilitate the transformation of free-text input into FHIR resources. We manually annotated a dataset of free-text to FHIR *MedicationStatement* resource transformation pairs. We compared the transformation accuracy between FHIR-GPT and existing NLP pipelines using the annotated dataset.

medRxiv preprint doi: https://doi.org/10.1101/2023.10.17.23297028; this version posted April 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Methods

In this section, we delve into the technical details employed in data annotation, LLMs usage, and the evaluation process. **Error! Reference source not found.** is an illustrative visual representation of the workflow.

To the best of our knowledge, there is no publicly available dataset with corresponding text and FHIR resource pairs. We therefore annotated a dataset containing pairs of free-text input and corresponding FHIR *MedicationStatement* resource output. The FHIR resource of *MedicationStatement* is a record of *a medication being taken by a patient or that a medication has been given to a patient, where the record is the result of a report from the patient or another clinician, or derived from supporting information*.[22] This transformation holds particular significance because many medication-related details, such as the reasons for administration and dosage instructions, often remain absent in structured data. Clinical notes within the EHR system frequently serve as the primary source for this medication-related information. For detailed examples of the elements in *MedicationStatement*, please refer to **Error! Reference source not found.**.

The clinical text input was obtained from the 2018 n2c2 medication extraction challenge.[23] We extracted the text snippets, each containing mentions of one medication and all its associated entities, from the discharge summaries. These extracted snippets, each tied to a specific medication, serve as input for both annotations and transformations. Our human annotation consisted of three key steps. We started by standardizing the entities from free text into clinical terminology coding systems. To achieve this, we leveraged the word spans of entities provided in the n2c2 dataset and manually looked up the HL7 CodeSets, SNOMED CT Browser [24], and the RxNav [25] for standardization. We then assembled the identifiers, codes, texts, elements, and

medRxiv preprint doi: https://doi.org/10.1101/2023.10.17.23297028; this version posted April 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

structures into a complete *MedicationStatement* resource in JSON format as per FHIR v6.0.0: R6 implementation guide [22]. Finally, the human-converted *MedicationStatement* resources underwent validation using the official FHIR validator[26] to ensure compliance with FHIR standards, including structure, datatypes, cardinalities, code sets, display names, etc.

We experimented with the following 3 Large Language Models (LLMs), OpenAI GPT-4 [27], Llama-2-70B [28], and Falcon-180B [29]. We accessed the GPT-4 (model: *gpt-4-32k as of 2023-05-15*) through the Azure OpenAI API service. We made multiple asynchronous API calls to enhance efficiency. For Llama-2-70B and Falcon-180B, we deployed them on our HIPAA-compliant firewalled local servers with multiple GPUs. GPTQ [30] was used to accelerate the inference time for Llama-2-70B and Falcon-180B.

We required these LLMs to transform the free-text entries into *MedicationStatements* conforming to the FHIR standard, employing the few-shot prompt settings that include 4-5 examples of transformations in the prompts. Each clinical text snippet was individually input into the LLMs to generate their *MedicationStatement* resource. We leveraged five separate prompts to instruct the LLM to transform the free-text input into the elements of a *MedicationStatement* resource, including medication details (drug name, strength, and form), route, timing, dosage, and reason. All few-shot prompts followed a template consisting of task instructions, expected output FHIR templates in JSON format, 4-5 examples of transformations, a comprehensive list of codes from which the model could make selections, and the input text to be transformed. As there was no fine-tuning or domain-specific adaptation in our experiments, we initially had the LLMs generate the FHIR resource for a small subset of the dataset (N=100). Then, we manually reviewed the discrepancies between the LLM-generated FHIR output and our human annotations. Common mistakes were identified and used to refine the prompts. There were slight

medRxiv preprint doi: https://doi.org/10.1101/2023.10.17.23297028; this version posted April 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

differences in the prompts for each LLM, as different LLMs may be sensitive to different prompts. We did not instruct the LLMs to look up the SNOMED codes for the 'medication' and 'reason' elements, as there are thousands of SNOMED CT Medication and Finding codes, exceeding the token limits of LLMs. Instead, our instructions were for them to identify the contexts mentioned in the input text and convert them into the appropriate JSON format. For other elements, such as routes and forms, we instructed LLMs to directly look up from the code set (numbering in hundreds). Example prompts can be found in the supplemental material. We will also share our annotated dataset on physionet.org upon acceptance.

We compared LLMs with two existing NLP pipelines: NLP2FHIR [16] and Google HNL API [20]. NLP2FHIR was built based on a previous version of the FHIR implementation guide R5; the Google HNL API primarily standardized concepts to UMLS CUIs, while the latest guide R6 and SNOMED was used in our annotations and LLMs' transformations. We therefore made necessary adaptations and conversions to ensure a fair comparison. We deployed the NLP2FHIR pipeline on our firewalled local servers and accessed the Google HNL API through the Google Cloud Healthcare API.

When evaluating the FHIR resources generated by the LLMs, we first conducted a format validation check to ensure that the output was in a valid JSON format. On passing the validation, we evaluated the generated resources with exact match rate. This strict criterion required that the resources generated by the LLMs exactly matched the human annotations in all aspects, including structures, codes, and cardinality.

medRxiv preprint doi: https://doi.org/10.1101/2023.10.17.23297028; this version posted April 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Results

The annotation results are presented in **Error! Reference source not found.**. In summary, we annotated a total of 3,671 pairs of free-text input and FHIR *MedicationStatement* resource output. The free-text input was derived from discharge summaries for 280 admissions. The annotated resources encompass 625 distinct medications in 26 different forms and are associated with 354 different reasons, as well as 16 administration routes. These elements display varying levels of availability, ranging from approximately 30% for reasons to 65% for timing schedules. The annotated resources in the JSON structure have an average number of objects of 58.2 (standard deviation = 16.2) and an average depth of 6.7 (standard deviation = 0.5).

The transformation results are presented in **Error! Reference source not found.**. In summary, transformation using FHIR-GPT, achieved an exceptional exact match rate of over 0.90 for all elements, outperforming both baseline models and all other LLMs. Specifically, when compared to existing NLP pipelines, FHIR-GPT improved the exact match rate by 3% for routes, 12% for dose quantities, 35% for reasons, 42% for forms, and over 50% for timing schedules. Among all LLMs, we observed a trend of increasing accuracy as the parameter size increased. FHIR-GPT, with approximately 1.7 trillion parameters, surpassed the 180 billion parameter Falcon model and further improved upon the 70 billion parameter Llama-2 model.

medRxiv preprint doi: https://doi.org/10.1101/2023.10.17.23297028; this version posted April 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Discussion

The reproducibility of using LLMs for FHIR transformation was examined through two experiments conducted six months apart, in September 2023 and March 2024. No updates were applied to the weights of Falcon and Llama during this period. In our March 2024 experiment of FHIR-GPT, we employed the latest *gpt-4-turbo* model, an upgrade from the *gpt-4-32k* model utilized in the previous September 2023 run. The *gpt-4-turbo* model boasts an expanded context window, stretching from 32k to 128k tokens, and incorporates additional training data spanning from September 2021 to April 2023. The results of reproducibility can be found in **Error! Reference source not found.**. Although there are slight fluctuations observed across various models and elements, none exceed a decrease of 2 percent. This indicates a relative stability in using LLMs for FHIR transformation, even with the update to the foundation model.

We conducted an error analysis to investigate instances where FHIR-GPT and human annotation diverge, with a particular focus on drug routes as an example. The 204 disagreements in transforming drug routes were categorized into five types of errors: false negatives, false positives, mismatched errors, syntax errors, and content filter rejections. A comprehensive breakdown, along with examples and distribution of these errors, is provided in **Error! Reference source not found.**. False negative errors primarily result from FHIR-GPT's insensitivity to certain medical abbreviations, such as 'gtt', or its failure to associate medical terms like 'lumen' with the intravenous route. Conversely, false positive errors occur when FHIR-GPT inaccurately introduces nonexistent information or identifies annotation errors, such as the oversight of 'IVIG', which was unrecognized in the i2b2 expert annotation and therefore omitted from our human annotation. The mismatched-error category presents a combination of false negative and false positive errors. We posit that such errors could be mitigated through the

medRxiv preprint doi: https://doi.org/10.1101/2023.10.17.23297028; this version posted April 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

incorporation of more domain-specific knowledge or examples in the prompt, or through the injection of knowledge bases with retrieval-augmented generations. Additionally, rare instances occur where output cannot be parsed as JSON. These parsing issues can be easily rectified with a simple format adjustment, replacing all double quotes with single quotes or using more advanced JSON parsing tools. Additionally, the Azure platform occasionally rejects requests with content filters to avoid harmful content in the prompts, though such filters can be opted out of if necessary.

In this study, we delve into three potential pathways for transforming free-text clinical input into FHIR resources. While human annotation is the gold standard for transformation, its dependence on extensive human efforts poses scalability limitations. Existing NLP pipelines can automate these transformation processes but demand substantial training data and resources, with challenges in generalizability and transferability. On the other hand, a new pipeline must undergo training or fine-tuning for even minimal changes in the code set or expansion to new resources. In addition, the multi-step transformation process incurs considerable maintenance costs, demanding meticulous tracing for effective error debugging across all steps. FHIR-GPT, harnessing the power of pre-trained LLMs, eliminates the need for high-cost training and depends on only minimal human annotation for the few-shot examples in the prompts. FHIR-GPT also achieves a superior level of accuracy compared to the approach employing NLP pipelines. Moreover, by altering the template and the corresponding code set in the prompt, FHIR-GPT holds the potential to generalize to other resources without the requirement for resource-specific training or fine-tuning. We believe that leveraging FHIR-GPT has the potential to greatly enhance interoperability, given its ease of implementation, high accuracy, and broad generalizability.

medRxiv preprint doi: https://doi.org/10.1101/2023.10.17.23297028; this version posted April 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

We acknowledge the following limitations in our study. irstly, while FHIR-GPT showcases superior performance compared to Llama and Falcon, its significant computational resource demands present a challenge. Moreover, its commercial nature gives rise to ethical considerations, impeding smooth integration into local EHR systems. We aim to investigate alternative lightweight and open-source foundation models of GPT-4 to overcome these obstacles while upholding comparable performance. Secondly, our evaluation of FHIR-GPT was confined to medication-related FHIR resources, potentially limiting the broader applicability of our findings. To address this, our future efforts aim to expand the transformation for a wider array of FHIR resources. Thirdly, our current approach primarily involves prompt engineering through trial-and-error with existing LLMs. There is no enhancement of the architecture of the foundational LLMs. To improve accuracy and reasoning in the transformation process, we aim to adopt advanced techniques like chain-of-thoughts[31], retrieval-augmented generation, or domain-specific finetuning in future endeavors. Fourthly, we recognize that LLMs can assist in other FHIR-related transformations, such as FHIR version upgrades and tabular-to-FHIR transformations. While we perceive these tasks as potentially less complex than our text-to-FHIR transformations, we encourage fellow researchers to explore the efficacy of LLMs in these alternative pathways to enhance interoperability.

Conclusion

In conclusion, this study lays the groundwork for harnessing LLMs to significantly improve health data interoperability through the transformation of free-text input into the FHIR resources. The FHIR-GPT model not only streamlines the process but also enhances transformation accuracy when compared to existing NLP pipelines. Building upon these promising results, our

medRxiv preprint doi: https://doi.org/10.1101/2023.10.17.23297028; this version posted April 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

future investigations will expand to encompass additional FHIR resources, aiming to advance the practical applications of LLMs in enhancing health data interoperability.

medRxiv preprint doi: https://doi.org/10.1101/2023.10.17.23297028; this version posted April 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

References

1. Office of the National Coordinator for Health Information Technology (ONC). Interoperability. Accessed Oct 31, 2023, https://www.healthit.gov/topic/interoperability
2. Office of the National Coordinator for Health Information Technology. United States Core Data for Interoperability (USCDI). 2023;
3. Centers for Disease Control and Prevention. Advancing Interoperability for Public Health. Accessed Oct 31, 2023. https://www.cdc.gov/surveillance/policy-standards/interoperability.html
4. CMS Health Informatics and Interoperability Group (HIIG). Federal Interoperability. Accessed Oct 31, 2023. https://www.cms.gov/priorities/key-initiatives/burden-reduction/interoperability/federal-interoperability
5. HL7.org. FHIR Overview. Accessed Oct 31, 2023. https://hl7.org/fhir/overview.html
6. Bauer DC, Metke-Jimenez A, Maurer-Stroh S, et al. Interoperable medical data: the missing link for understanding COVID-19. *Transboundary and emerging diseases*. 2021;68(4):1753-1760.
7. Brandt PS, Pacheco JA, Rasmussen LV. Development of a repository of computable phenotype definitions using the clinical quality language. *JAMIA open*. 2021;4(4):ooab094.
8. Zong N, Sharma DK, Yu Y, et al. Developing a FHIR-based framework for phenome wide association studies: a case study with a pan-cancer cohort. *AMIA Summits on Translational Science Proceedings*. 2020;2020:750.
9. Metke-Jimenez A, Hansen D. FHIRCap: Transforming REDCap forms into FHIR resources. *AMIA Summits on Translational Science Proceedings*. 2019;2019:54.
10. Pfiffner PB, Pinyol I, Natter MD, Mandl KD. C3-PRO: connecting ResearchKit to the health system using i2b2 and FHIR. *PLoS One*. 2016;11(3):e0152722.
11. Reinecke I, Gulden C, Kümmel M, Nassirian A, Blasini R, Sedlmayr M. Design for a modular clinical trial recruitment support system based on FHIR and OMOP. *Digital Personalized Health and Medicine*. IOS Press; 2020:158-162.
12. Zong N, Stone DJ, Sharma DK, et al. Modeling cancer clinical trials using HL7 FHIR to support downstream applications: A case study with colorectal cancer data. *International journal of medical informatics*. 2021;145:104308.
13. Lee H-A, Kung H-H, Lee Y-J, et al. Global infectious disease surveillance and case tracking system for COVID-19: development study. *JMIR Medical Informatics*. 2020;8(12):e20567.
14. Wang X, Lehmann H, Botsis T. Can FHIR support standardization in post-market safety surveillance? *Public Health and Informatics*. IOS Press; 2021:33-37.
15. Dash S, Shakyawar SK, Sharma M, Kaushik S. Big data in healthcare: management, analysis and future prospects. *Journal of big data*. 2019;6(1):1-25.
16. Hong N, Wen A, Shen F, et al. Developing a scalable FHIR-based clinical data normalization pipeline for standardizing and integrating unstructured and structured electronic health record data. *JAMIA open*. 2019;2(4):570-579.
17. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*. 2010;17(5):507-513.
18. Sohn S, Clark C, Halgrim SR, Murphy SP, Chute CG, Liu H. MedXN: an open source medication extraction and normalization tool for clinical text. *Journal of the American Medical Informatics Association*. 2014;21(5):858-865.
19. Wang J, Mathews WC, Pham HA, Xu H, Zhang Y. Opioid2FHIR: A system for extracting FHIR-compatible opioid prescriptions from clinical text. IEEE; 2020:1748-1751.
20. Google Cloud. Use the Healthcare Natural Language API. Accessed Oct 31, 2023. https://cloud.google.com/healthcare-api/docs/how-tos/nlp
21. Microsoft Azure. Azure Health Data Services. Accessed Oct 31, 2023. https://azure.microsoft.com/en-us/products/health-data-services
22. HL7.org. Resource MedicationStatement - Content. Accessed Oct 31, 2023. https://build.fhir.org/medicationstatement.html

medRxiv preprint doi: https://doi.org/10.1101/2023.10.17.23297028; this version posted April 1, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

23. Henry S, Buchan K, Filannino M, Stubbs A, Uzuner O. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*. 2020;27(1):3-12.
24. SNOMED International. SNOMED International SNOMED CT Browser. Accessed Oct 31, 2023. https://browser.ihtsdotools.org/
25. Zeng K, Bodenreider O, Kilbourne J, Nelson SJ. RxNav: a web service for standard drug information. American Medical Informatics Association; 2006:1156.
26. HL7.org. Validate Resources. Accessed Oct 31, 2023. https://validator.fhir.org/
27. OpenAI R. GPT-4 technical report. *arXiv*. 2023:2303.08774.
28. Touvron H, Martin L, Stone K, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:230709288*. 2023;
29. Penedo G, Malartic Q, Hesslow D, et al. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:230601116*. 2023;
30. Frantar E, Ashkboos S, Hoefler T, Alistarh D. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:221017323*. 2022;
31. Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*. 2022;35:24824-24837.

Table 1. Descriptions, examples, and statistics of human annotation for the FHIR *medicationstatement* resource.

***Medication Statement* Elements**	**Type**	**Card.**	**Example**	**Description**	**Code Set**	**N (%)**	**N, Uniq. Entries**	**N, Uniq. Codes**
identifier	String	1..1	*100035T133*	External identifier	MIMIC+i2b2	3671 (100%)	3,671	3,671
subject	Codeable Reference	1..1	*{'reference': 'hadm_id164366'}*	Who is/was taking the medication	MIMIC	3671 (100%)	280	280
medication		1..1		What medication				
medication	Codeable Concept	0..1	*{'coding': [{'system': 'NDC', 'code': '51079088120', 'display': 'clonazepam 0.5 MG Oral Tablet'}, {'system': 'RxNORM', 'code': '197527'', 'display': 'Clonazepam 500 microgram oral tablet'}, {'system': 'SNOMED', 'code': '322897008', 'display': 'Clonazepam 500 microgram oral tablet'}], 'text': 'clonazepam 0.5 mg Tablet'}*	Codes that identify this medication	NDC RxNorm SNOMED CT Medication	3671 (100%)	1762	NDC: 625 RxNorm: 520 SNOMED: 210
doseForm	Codeable Concept	0..1	*{'text': 'Tablet', 'coding': [{'system': 'SNOMED','code': '385055001', 'display': 'Tablet'}]}*	powder | tablets | capsule +	SNOMED CT Dose Form	1478 (40.3%)	176	26
ingredient. Strength	Quantity	0..1	*{'value': 0.5,'unit': 'milligram', 'code': 'mg' 'system': 'http://unitsofmeasure.org'}*	Quantity of ingredient presents	unitsofmeasure. org	2383 (64.9%)	188	16
reason	Codeable Concept	0..*	*[{'concept': {'text': 'headache', 'coding': [{'system': 'SNOMED', 'code': '25064002', 'display': 'Headache'}]}}]*	Reason for why the medication is being/was taken	SNOMED CT Finding	1106 (30.1%)	619	354
dosage		0..*						
asNeeded	Boolean	0..1	*True*	Take "as needed"		3671 (100%)	2	
route	Codeable Concept		*{'text': 'PO', 'coding': [{'system': 'SNOMED', 'code': '26643006', 'display': 'Oral route'}]}*	How medication enters the body	SNOMED CT Route of Admin.	2011 (54.8%)	64	15
timing. repeat	Element	0..1	*{'frequency': 1, 'period': 4.0, 'periodMax': 6.0, 'periodUnit': 'h', 'duration': 3.0, 'durationUnit':'d'}*	Timing schedule	hl7.org/fhir/	2393 (65.2%)	177	6
timing. code	Codeable Concept	0..1	*{'coding': [{'system': 'HL7','code': 'Q4H', 'display': 'Q4H'}]}}*	Code for timing schedule, e.g. 'BID'	hl7.org/fhir/	2287 (62.3%)	17	17
doseRange	Quantity	0..1	*{"doseQuantity": {"value": 5.0, "unit": "ML"}}*	Amount or range of medication per dose		1378 (37.5%)	53	
dose-Quantity	Range	0..1	*{"doseRange": {"low": { "value": 1.0}, "high": { "value": 3.0}}}*			11 (0.30%)	7	

medRxiv preprint doi: https://doi.org/10.1101/2023.10.17.23297028; this version posted April 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

medRxiv preprint doi: https://doi.org/10.1101/2023.10.17.23297028; this version posted April 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Table 2. Comparison of LLMs and existing NLP pipelines for transforming free-text input into FHIR *MedicationStatement* resources. Performance is evaluated using the exact match rate, which requires that the resources generated by the models precisely match human annotations in all aspects, including structure, codes, and cardinality. Due to version and implementation differences, the existing NLP pipelines cannot generate all the elements included in our annotations. The best-performing model for each element is indicated in bold, while the second-place model is underlined.

Elements of *medicationstatement*	**Large Language Models**			**Existing NLP Pipelines**	
	FHIR-GPT[27]	Falcon-180B[28]	Llama-2-70B[29]	NLP2FHIR[16]	Google HNL API[20]
medication					
medication	**0.968**	0.899	0.859	0.862	0.963
doseForm	**0.976**	0.790	0.633	0.556	-
ingredient.Strength	**0.980**	0.921	0.792	-	-
reason	**0.902**	0.593	0.169	0.645	-
dosage					
route	**0.902**	0.457	0.516	-	0.871
timing.repeat	**0.947**	0.268	0.221	0.403	-
timing.code	**0.952**	0.818	0.600	0.424	-
doseQuantity/Range	**0.973**	0.864	0.823	0.724	0.854

Table 3. Reproducibility in using LLMs for transforming FHIR resources. Two separate experiments were conducted six months apart using the same prompts, with identical model weights used for Falcon and Llama models. FHIR-GPT used the gpt-4-32k model in the September 2023 experiment, which was upgraded to the gpt-4-turbo (128k) model in March 2024.

Elements of *medicationstatement*	**Large Language Models**					
	FHIR-GPT[27]		Falcon-180B[28]		Llama-2-70B[29]	
	Sept. 2023	Mar. 2024	Sept. 2023	Mar. 2024	Sept. 2023	Mar. 2024
medication						
medication	0.968	0.970	0.899	0.897	0.859	0.849
doseForm	0.976	0.974	0.790	0.785	0.633	0.634
ingredient.Strength	0.980	0.981	0.921	0.912	0.792	0.792
reason	0.902	0.908	0.593	0.617	0.169	0.172
dosage						
route	0.902	0.944	0.457	0.471	0.516	0.518
timing.repeat	0.947	0.938	0.268	0.264	0.221	0.218
timing.code	0.952	0.946	0.818	0.810	0.600	0.580
doseQuantity/Range	0.973	0.972	0.864	0.864	0.823	0.814

medRxiv preprint doi: https://doi.org/10.1101/2023.10.17.23297028; this version posted April 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Table 4. Discrepancies between FHIR-GPT generated resources and human annotations. 204 disagreements in transforming drug routes were categorized into five types of errors.

Category	Explanation	Input Example	Expected Output	Generated Output	N (%)
False Negative	Despite the presence of drug route information in the input, FHIT-GPT overlooks it.	2 ml of 100 Units/ml heparin (200 units heparin) **each lumen** Daily and PRN	{'text': 'each lumen', 'coding': [{ 'system': 'snomed.info/sct', 'code': '47625008', 'display': 'Intravenous route' }]}	**{}**	95 (45.7%)
		PEs and HIT at the referring institution, and was continued on an argatroban **gtt**	{'text': 'gtt'}	**{}**	
False Positive	FHIR-GPT fabricates drug route information that does not exist in the input.	Pneumonia was suspected and patient was started on vancomycin	**{}**	{'text': 'IV', 'coding': [{ 'system': 'snomed.info/sct', 'code': '47625008', 'display': 'Intravenous route' }]}	62 (29.8%)
		Joint pain: medication side effect (IVIG, hydralazine	**{}**	{'text': 'IVIG', 'coding': [{ 'system': 'snomed.info/sct', 'code': '47625008', 'display': 'Intravenous route' }]}	
Mis-matched Error	Although FHIR-GPT generates drug route resource, it does not align with the actual data provided in the input.	heparin lock flush (porcine) [heparin lock flush] 10 unit/mL 2mL to **PICC line Flush** dail	{'text': **'PICC line Flush'**, 'coding': [{ 'system': 'snomed.info/sct', 'code': **'417989007'**, 'display': **'Intraductal route'** }]}	{'text': **'PO'**, 'coding': [{ 'system': 'snomed.info/sct', 'code': **'284009009'**, 'display': **'Route of Administration values'** }]}	46 (22.1%)
		Artificial Tears 1-2 DROP **OU** PRN	{'text': **'OU'**, 'coding': [{ 'system': 'snomed.info/sct', 'code': **'54485002'**, 'display': **'Ophthalmic route'** }]}	{'text': **'DROP'**, 'coding': [{ 'system': 'snomed.info/sct', 'code': **'372473007'**, 'display': **'Oromucosal use'** }]}	
Syntax Error	The generated content is not in a valid FHIR resource or JSON format.	COPD flare with vancomycin 1 gm **IV**	{'text': 'IV', 'coding': [{ 'system': 'snomed.info/sct', 'code': '47625008', 'display': 'Intravenous route' }]}	{''text'': ''IV'', ''coding'': [{ ''system'': ''snomed.info/sct'', ''code'': ''47625008'', ''display'': ''Intravenous route'' }]}	4 (1.9%)
Content Filter Rejection	FHIR-GPT fails to generate content due to the input failing to pass the Azure content filter.	… he developed delirium … which manifested as inappropriate and sometimes violent actions with pt attempting to hit staff and spitting on staff. … pt had to be given haloperidol	**{}**	N/A	1 (0.5%)

medRxiv preprint doi: https://doi.org/10.1101/2023.10.17.23297028; this version posted April 1, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

[Figure 1. Overview of the transformation from free-text to FHIR resource. a. Example of a snippet from the discharge summary, which is the free-text input for FHIR resource generation. **b**. The n2c2 expert annotation of medication-related entities in the discharge summary. **c**. Example of the transformed FHIR *MedicationStatement* resource based on our annotations, serving as the ground truth transformation. The same color shading from panel b is used. **d**. Example prompt used to instruct LLMs in generating FHIR resources. **e**. The workflow details how we annotated the dataset and compared the performance of LLMs with existing NLP pipelines in transforming free-text inputs into associated FHIR resources.]

a. Discharge Summary

Discharge Summary:
Discharge Medications:
...
[7. senna 8.6 mg Tablet Sig: One (1) Tablet PO BID P.R.N Constipation]
...
Patient was discharged to long-term care facility.

b. Entity Annotations

Medication	**senna**
Reason	**Constipation**
Route	**PO**
Timing	**BID**
Dose	**One (1)**
Form	**Tablet**
Strength	**8.6 mg**
asNeeded	**P.R.N**

c. FHIR *MedicationStatement*

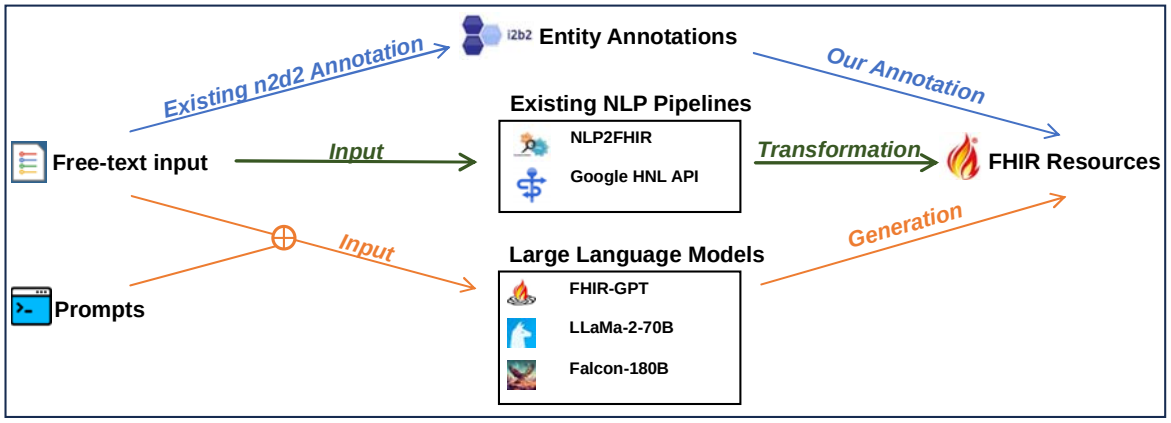
```
{'resourceType': 'MedicationStatement',
 'id': '100035T133',
 'subject': {'reference': 'hadm_id100035'},
 'medication': {'reference': {'reference': '#med100035T133'}},
 'reason': [
    {'concept': {'text': 'Constipation',
        'coding': [{'system': 'http://snomed.info/sct',
        'code': '14760008',
        'display': 'Constipation'}]}}],
 'dosage': [
    {'route': {'text': 'PO',
              'coding': [{'system': 'http://snomed.info/sct',
              'code': '26643006',
              'display': 'Oral route'}]},
     'timing': {'repeat': {'frequency': 2, 'period': 1.0, 'periodUnit': 'd'},
                'code': {'coding': [{'system': 'http://terminology.hl7.org/',
                         'code': 'BID',
                'display': 'BID'}]}},
     'asNeeded': True,
     'doseAndRate': [{'doseQuantity': {'value': 1.0}}]}],
 'contained': [
    {'resourceType': 'Medication',
     'id': 'med100035T133',
     'code': {'coding': [
        {'system': 'National Drug Code',
        'code': '00904516561',
        'display': 'sennosides, USP 8.6 MG Oral Tablet'},
        {'system': 'RxNorm',
        'code': '312935',
        'display': 'sennosides, USP 8.6 MG Oral Tablet'}],
            'text': 'senna 8.6 mg Tablet'},
     'doseForm': {'text': 'Tablet',
        'coding': [{'system': 'http://snomed.info/sct',
        'code': '385055001',
        'display': 'Tablet'}]},
     'ingredient': [{'item': {'concept': {'text': 'senna'}},
        'strengthQuantity':
            {'value': 8.6, 'unit': 'milligram',
             'system': 'http://unitsofmeasure.org',
             'code': 'mg'}}]}]}
```

d. Prompts for LLMs

[INSTRUCTIONS]
You are a helpful assistant that can help with medication data extraction.
User will paste a short narrative that describes the administration of a drug.
Please extract the drug route (How drug should enter body), e.g. PO, IV.
< Collapsed for more instructions >

[TEMPLATE]
```
{"text": "<string>", // the originial text mention of drug route
 "coding": [ //optional, but MUST lookup from the table below
   {"system": "http://snomed.info/sct",
    "code": "<code>", # SNOMED code
    "display": "<display>" # the display of the code}]}
```

[EXAMPLES]
For example, the narrative
"Oxycodone-Acetaminophen 5-325 mg Tablet
Sig: 1-2 Tablets PO\nQ4-6H (every 4 to 6 hours) as needed"
You should return a json format:
```
{'text': 'PO', 'coding': [{'system': 'http://snomed.info/sct', 'code':
'26643006','display': 'Oral route'}]}
```
< Collapsed for 4 more examples >

[TERMINOLOGIES]
Code Display
6064005 Topical route
10547007 Otic route
<Collapsed for 143 more SNOMED CT Codes>

e. Workflow