# Document-level Clinical Entity and Relation Extraction via Knowledge Base-Guided Generation

**Kriti Bhattarai[1,2], Inez Y. Oh[1], Zachary B. Abrams[1], Albert M. Lai[1,2]**
Institute for Informatics, Data Science & Biostatistics, Washington University School of Medicine
Department of Computer Science, Washington University in St. Louis

## Abstract

Generative pre-trained transformer (GPT) models have shown promise in clinical entity and relation extraction tasks because of their precise extraction and contextual understanding capability. In this work, we further leverage the Unified Medical Language System (UMLS) knowledge base to accurately identify medical concepts and improve clinical entity and relation extraction at the document level. Our framework selects UMLS concepts relevant to the text and combines them with prompts to guide language models in extracting entities. Our experiments demonstrate that this initial concept mapping and the inclusion of these mapped concepts in the prompts improves extraction results compared to few-shot extraction tasks on generic language models that do not leverage UMLS. Further, our results show that this approach is more effective than the standard Retrieval Augmented Generation (RAG) technique, where retrieved data is compared with prompt embeddings to generate results. Overall, we find that integrating UMLS concepts with GPT models significantly improves entity and relation identification, outperforming the baseline and RAG models. By combining the precise concept mapping capability of knowledge-based approaches like UMLS with the contextual understanding capability of GPT, our method highlights the potential of these approaches in specialized domains like healthcare.

## 1 Introduction

Generative pre-trained transformer (GPT) models have shown significant potential across various clinical tasks, including information extraction, summarization, and question-answering (Agrawal et al., 2022; Tang et al., 2023a; Yang et al., 2022, Singhal et al., 2023). Generative models are able to generate contextually relevant text given a prompt. However, for real-world clinical use, in tasks that require high precision, it is equally important to understand the context and minimize the errors that come from GPT models. However, accuracy of these models is limited to their training data. While GPT models are great at capturing nuanced contextual information, they often fall short in accurately identifying all medical concepts, possibly due to limited or outdated domain-specific data (Tang et al., 2023b, Singhal et al., 2023).

Knowledge bases store domain-specific data. Medical knowledge bases, such as, Unified Medical Language System (UMLS) knowledge base (Bodenreider, 2004), include comprehensive information about medical concepts. Integrating knowledge bases with language models is an open research area with multiple works exploring different ways of integrating them with language models, such as BERT (Devlin et al., 2019). There are limited studies on the integration of medical knowledge bases, particularly UMLS, with most recent large language models (LLMs), such as GPT.

To address this limitation, we introduce an approach for clinical entity extraction that leverages UMLS for knowledge augmentation. While GPT can identify nuanced contextual information, UMLS includes a comprehensive repository of domain-specific clinical concepts that GPT may not recognize, such as brand names for drugs, abbreviations, acronyms, and aliases (Agrawal et al., 2022).

Our contributions in this paper are summarized as follows:
(1) we introduce a framework to integrate UMLS concepts into the default generative models to facilitate few-shot information extraction of biomedical entities and relations.
(2) we explore current state-of-the-art knowledge augmentation techniques, such as Retrieval Augmented Generation (RAG) aimed at improving extraction, and
(3) we conduct evaluation of our framework, comparing the performance of models augmented with



*Proceedings of the 23rd Workshop on Biomedical Language Processing*, pages 318–327

UMLS knowledge with and without RAG, and against those without augmentation.

## 2 Related Work

### 2.1 Few-shot in-context learning

With the introduction of GPT models, there have been several works around few-shot in-context learning for clinical entity extraction where prompts guide information extraction in a contextually relevant manner (Agrawal et al., 2022; Hu et al., 2024; Shyr et al., 2024, Brown et al., 2020). Generative models can provide nuanced contextual understanding to extract clinical concepts, but cannot identify all domain-specific terminologies, especially in the clinical domain (Tang et al., 2023b). While recent language models have demonstrated improvement over prior language models (Guevara et al., 2024), there remains room for performance improvement.

### 2.2 Knowledge base-guided models

Previous research has explored the integration of knowledge bases to enhance information extraction tasks. (Sastre et al., 2020) proposed a Bi-LSTM model to identify drug-related information and integrate it into knowledge graph embeddings to evaluate drug identification accuracy. (Gilbert et al., 2024) addressed how knowledge bases complement language models for medical information identification tasks. Recently, a RAG model, Almanac, demonstrated significant performance improvements compared to the standard LLMs across various metrics (Zakka et al., 2024), further showing the benefits of access to domain-specific corpora for information extraction.

## 3 Methods

### 3.1 Overview of the Framework

Our approach leverages the context-capturing capability of GPT and knowledge-capturing capability of UMLS. UMLS contains a comprehensive list of more than 1 million biomedical concepts from over 100 source vocabularies. By using the concepts in the prompts in a few-shot learning setting, we attempt to improve GPT's ability to identify entities with the specified context that it may otherwise fail to extract independently. We map UMLS concepts to each text instance to create dynamic prompts unique to the specific context of the clinical text.The overview of the proposed framework is displayed in Figure 1.

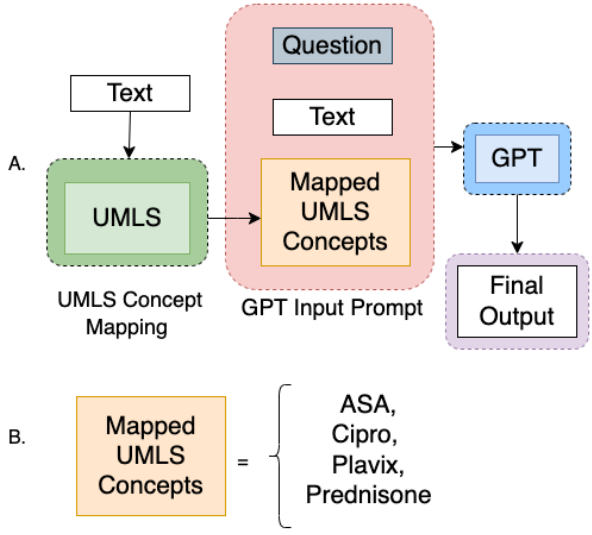



Figure 1: (A) Step-by-step approach to integrating UMLS and extracting relation pairs. (B) Example of UMLS concepts mapped from the text. Some of the concepts, such as Prednisone, are recognized by GPT, as they are concepts GPT model is inherently trained on. However, concepts such as ASA, Cipro, Plavix are not recognized by GPT; UMLS facilitates their recognition.

### 3.2 UMLS Integration in Large Language Model

*UMLS Concept Mapping*

We first map UMLS concepts from clinical text using MetaMap (Aronson, 2001). Given clinical text $X = \{x_1, x_2, \ldots, x_n\}$ where $x_i$ represents the $i$th clinical text, we map $C_i = \{c_{i1}, c_{i2}, \ldots, c_{in}\}$, where $C_i$ denote the set of concepts identified by MetaMap from $x_i$. $n$ denotes the number of concepts identified from $x_i$. These concepts are extracted leveraging MetaMap's lexical parsing, syntactic analysis, semantic mapping, and concept mapping techniques.

Next, we filter the mapped concepts to include only those categorized as 'organic chemical', 'antibiotic', or 'pharmacologic substance' within the UMLS concept hierarchy as these groups contains the medications. For this work, we only target and filter medication-related concepts for augmentation and for further analysis. Let's denote the filtered set of concepts for the $i$th input clinical text $x_i$ as $C_{filtered}$, such that $C_{filtered,i} = \{c_{ij} \in C_i | c_{ij} \in$ filtered groups$\}$.

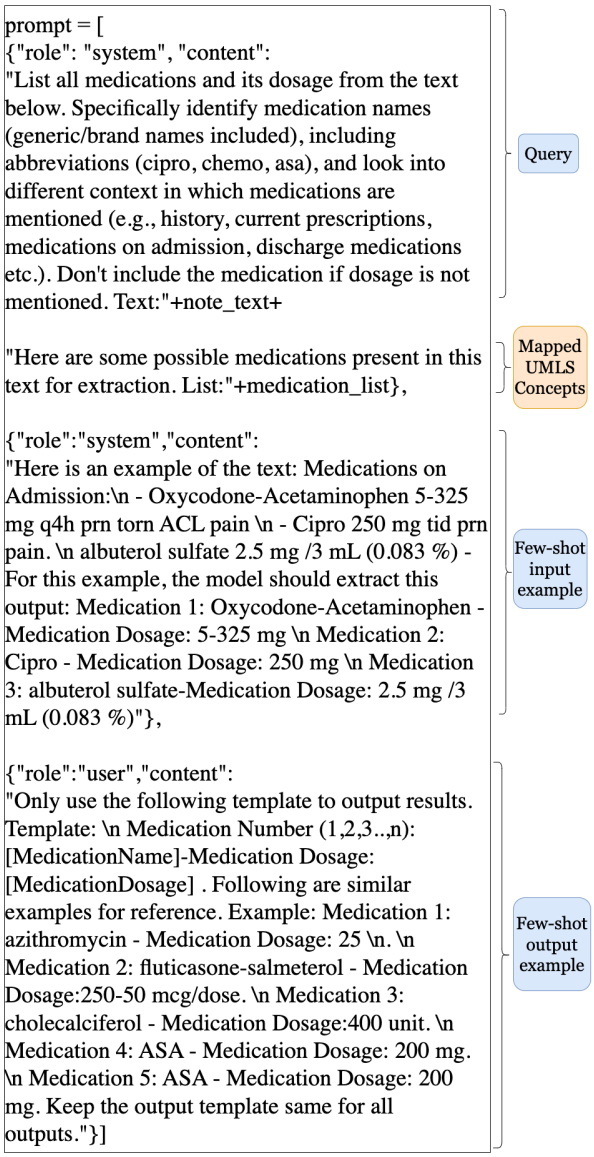



Figure 2: An example of a prompt used to extract dosage information from the text using the UMLS concepts. The 'note_text' represents each text instance from ADE or n2c2 corpus. The 'medication_list' represents the UMLS concepts extracted from MetaMap.

*Prompt Strategy and Large Language Model Implementation*

Next, we prompt the GPT model to extract entity-relation pairs from the text, leveraging the mapped UMLS concepts from MetaMap, and employing a few-shot prompt strategy. Let $P_i$ represent the prompt generated for each input text $x_i$, incorporating the relevant UMLS concepts $C_{filtered,i}$. The final prompt $P_i$ is constructed as the concatenation of the initial prompt and the set of UMLS concepts, i.e., $P_i = \text{Concat}(P, C_{filtered,i})$. We use OpenAI's GPT-4-32k (Version 0613) and GPT-3.5-turbo (Version 0301) via HIPPA-compliant Microsoft Azure's OpenAI REST API[1]endpoint. A sample prompt and hyperparameters used by the models for this task are available in Figure 2 and A.2 respectively. As our goal for the project was not to explore different prompting strategies, we tested a few prompts and selected the prompt that generated more specific result. We used the same format for all relation pairs replacing only the entity type for every run.

*Retrieval Augmented Generation*

We also explored another approach-RAG to leverage UMLS in a language model, which is a more conventional method involving the use of external data. RAG was chosen for its potential to enhance the generation process by incorporating domain-specific knowledge from sources like the UMLS knowledge base. Appendix A.4 includes details on our RAG implementation.

### 3.3 Datasets Description

We used the n2c2 and ADE datasets for our experiments.

*n2c2 Dataset*

We used a curated National NLP Clinical Challenges (n2c2) dataset (Henry et al., 2019) consisting of 303 deidentified discharge summaries obtained from the MIMIC-III (Medical Information Mart for Intensive Care-III) critical care database (**Table 1A**) (Johnson et al., 2016). The data also contained annotations of medication-related entities and their relationship to other entities. Annotations conducted by 3 subject matter experts served as a gold standard to evaluate model performance.

*ADE Dataset*

The Adverse Drug Events (ADE) dataset annotated by 5 individuals consists of MEDLINE[2] case reports with information on medications, dosages and adverse effects associated with the medications (Gurulingappa et al., 2012) (**Table 1B**). It also contains relationships between medications, dosages, and adverse effects. For our experiments, we used the second version of the dataset downloaded from

[1] https://learn.microsoft.com/en-us/azure/ai-services/openai/quickstart?tabs=command-line%2Cpython-new&pivots=programming-language-python

[2] https://www.nlm.nih.gov/medline/medline_home.html

Huggingface[3].

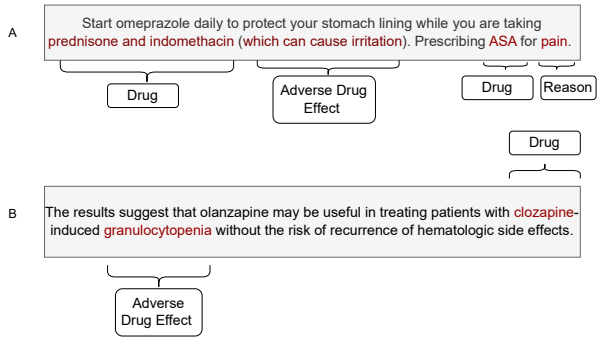



Figure 3: Sample text of discharge summaries in the (A) n2c2 dataset and (B) ADE Corpus. The text highlighted in red are the targeted entities for extraction

Table 1: Statistics on the relation pairs in the (A) n2c2 dataset and the (B) ADE dataset

| **A. n2c2 Dataset** | |
|---|---|
| **Entity-Entity Relation** | **Total instances** |
| Strength-Drug | 13338 |
| Duration-Drug | 643 |
| Route-Drug | 11038 |
| Form-Drug | 6636 |
| ADE-Drug | 2214 |
| Dosage-Drug | 4207 |
| Reason-Drug | 5160 |
| Frequency-Drug | 6288 |

| **B. ADE Dataset** | |
|---|---|
| **Entity-Entity Relation** | **Total instances** |
| Drug-ADE | 6821 |
| Drug-Dosage | 279 |

# 4 Results

## 4.1 Experimental Setup

We evaluated two generative models, GPT-4-32k and GPT-3.5-turbo, with and without UMLS integration, and the RAG model to access the quality of generated outputs. All models were evaluated against the gold-standard annotations using precision, recall, and micro-F1 score.

## 4.2 Dataset

We identified 8 different entity-entity relation pairs within the n2c2 dataset, and 2 entity-entity relation pairs within the ADE corpus, each with varying instances of the relation pairs (**Table 1, Figure 3**). Token length distributions of the text, and example of individual entities in n2c2 and ADE dataset are available in A.1, A.3.

## 4.3 Performance Results

*Results on n2c2 and ADE Dataset*

Our results suggests that integrating prior knowledge from UMLS in the prompts have significant performance improvement as demonstrated by the higher average F-1 scores across both n2c2 and ADE datasets (**Table 4**). The reported results are average across all entity-entity relation pairs across models and for 2 datasets. GPT-4-32k model with UMLS show 4% improvement of F-1 score on the n2c2 dataset, and 12% improvement on the ADE dataset from the F-1 score of GPT-4-32k model without knowledge integration. For every entity-entity relation pairs, there was a performance improvement by a few percentages for both models and across both datasets. Additional detailed results for each entity-entity relation pairs can be found in Appendices A.6 through A.11.

Upon a closer look at the results, we identified that prompts with UMLS resulted in additional concepts verifying that UMLS is able to identify medications from the text that GPT may not identify independently(Appendix A.5).

*Comparison with Retrieval Augmented Generation*

RAG model and GPT-3.5-turbo had low F-1 score and it improved with UMLS for both models, but it did not have higher score compared to the GPT-4-32k+UMLS.

We observed performance variations across entity-entity relation pairs with retrieval augmented generation (A.8, A.11). While some entity pairs showed performance enhancements, others did not show significant improvements. This discrepancy might arise from the limitations of RAG, particularly its inability to utilize entire UMLS thesaurarus in the generation process. Since UMLS data is partitioned into chunks for indexing and embedding and embedding models can only take 8192 tokens per index, some concepts may not be in the top-k extracted documents used for generation, potentially limiting the scope of augmentation and its impact on final relation pairs. Further experiments are required to confirm this hypothesis.

[3] https://huggingface.co/datasets/ade_corpus_v2

Table 2: Comparison of models on n2c2 Dataset and ADE Corpus. The reported results are micro-averaged precision, recall, and F-1 scores across all entity-entity relation pairs within the datasets.

| Model | n2c2 Dataset | | | ADE Corpus | | |
|---|---|---|---|---|---|---|
| | **Precision** | **Recall** | **F1** | **Precision** | **Recall** | **F1** |
| GPT-3.5-turbo | 0.73 | 0.74 | 0.73 | 0.625 | 0.57 | 0.596 |
| GPT-3.5-turbo + UMLS | 0.77 | 0.77 | 0.77 | 0.83 | 0.70 | 0.75 |
| RAG w/ GPT-3.5-turbo | 0.73 | 0.74 | 0.74 | 0.65 | 0.63 | 0.64 |
| GPT-4-32k | 0.75 | 0.76 | 0.76 | **1.0** | 0.70 | 0.82 |
| GPT-4-32k + UMLS | **0.79** | **0.80** | **0.80** | **1.0** | **0.89** | **0.94** |
| RAG w/ GPT-4-32k | 0.77 | 0.76 | 0.76 | **1.0** | 0.74 | 0.85 |

## 5 Discussion and Conclusion

Our study highlights the significance of merging the strengths of domain-specific knowledge bases, such as UMLS, with the contextual understanding capabilities of LLMs, such as GPT. Our hybrid approach, integrating mapped UMLS concepts with GPT, shows improvement in the model's ability to identify specific entities not inherently within its training data.

Our results on entity and relation extraction task indicated that leveraging mapped UMLS concepts as additional guidance to the GPT model, helps create focused and unique prompts that significantly enhances GPT's performance. This approach proves more effective than the standard RAG technique.

In conclusion, the ability to generate tailored prompts based on UMLS concepts offers a promising avenue for improving accuracy and relevance of extracted entities, ultimately enhancing the utility of LLMs in biomedical text analysis tasks.

## 6 Limitations and Future Work

While our framework has shown significant improvements, we acknowledge several limitations in this study. Firstly, our work focused solely on medication concepts, which may restrict the generalizability of our findings to other concepts. However, our approach is adaptable to incorporate additional UMLS entities through prompt adjustments. Future research will explore harnessing UMLS's rich semantic metadata to leverage additional concept relationships, enabling the extraction of a broader spectrum of entity groups beyond medications.

Secondly, our comparison was limited to two generative models, GPT-4-32k and GPT-3.5-turbo. Though they have good performance, we have not included recent models that could have comparable performance. Future work will explore additional models, such as BioGPT, and LAMA for comprehensive comparison and evaluation. This expanded comparison will provide a more nuanced understanding of the performance and capabilities of various generative models in relation to UMLS integration and RAG techniques.

These future tasks will advance our understanding of the role of domain-specific knowledge in enhancing LLM capabilities and facilitating more effective clinical information extraction.

## 7 Ethics Statement

IRB approval was not required for this task. To input our text data into the language models, we use Microsoft's Azure OpenAI REST API Service within the Washington University tenant to access OpenAI's language models . We are on a HIPPA-compliant subscription and exempted from content filtering, data review and human review for our use of the Azure OpenAI service.

## References

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022.

A R Aronson. 2001. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. *Proc AMIA Symp*.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *AAAI*, 32:D267–D270.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, page 4171–4186.

Stephen Gilbert, Jakob Nikolas Kather, and Aidan Hogan. 2024. Augmented non-hallucinating large language models as medical information curators. *npj Digit. Med*, 7(100).

Marco Guevara, Shan Chen, Spencer Thomas, Tafadzwa L. Chaunzwa, Idalid Franco, Benjamin H. Kann, Shalini Moningi, Jack M. Qian, Madeleine Goldstein, Susan Harper, Hugo J. W. L. Aerts, Paul J. Catalano, Guergana K. Savova, Raymond H. Mak, and Danielle S. Bitterman. 2024. Large language models to identify social determinants of health in electronic health records. *Nature*, 7(6).

Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. 2012. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of Biomedical Informatics*, 45(5):885–892.

Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Ozlem Uzuner. 2019. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, 27(1):3–12.

Yan Hu, Qingyu Chen, Jingcheng Du, Xueqing Peng, Vipina Kuttichi Keloth, Xu Zuo, Yujia Zhou, Zehan Li, Xiaoqian Jiang, Zhiyong Lu, Kirk Roberts, and Hua Xu. 2024. Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association*, pages 1–10.

Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. Mimic-iii, a freely accessible critical care database. *Journal of the American Medical Informatics Association*, 3(160035).

Javier Sastre, Faisal Zaman, Noirin Duggan, Caitlin McDonagh, and Paul Walsh. 2020. A deep learning knowledge graph approach to drug labelling. *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 32:2513–2521.

Cathy Shyr, Yan Hu, Lisa Bastarache, Alex Cheng, Rizwan Hamid, Paul Harris, and Hua Xu. 2024. Identifying and extracting rare disease phenotypes with large language models. *Journal of the American Medical Informatics Association*, 8:438–461.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2023. Large language models encode clinical knowledge. *Nature*, 620:172–180.

Liyan Tang, Zhaoyi Sun, Betina Idnay, Jordan G. Nestor, Ali Soroush, Pierre A. Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin F. Rousseau, Chunhua Weng, and Yifan Peng. 2023a. Evaluating large language models on medical evidence summarization. *Nature*, 6(158).

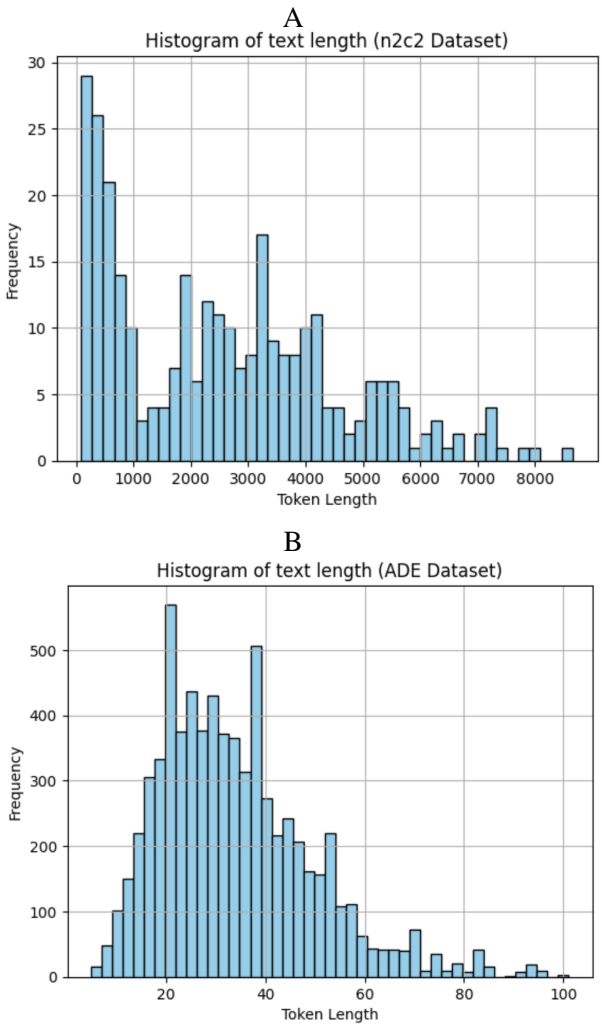
Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2023b. Medagents: Large language models as collaborators for zero-shot medical reasoning. *Computing Research Repository*, arXiv:2311.10537. Version 3.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 2022. An empirical study of gpt-3 for few-shot knowledge-based vqa. *AAAI*, 36:3081–3089.

Cyril Zakka, Rohan Shad, Akash Chaurasia, Alex R. Dalal, Jennifer L. Kim, Michael Moor, Robyn Fong, and William Hiesinger. 2024. Almanac — retrieval-augmented language models for clinical medicine. *npj Digit. Med*, 1(2).

# A Appendix

## A.1 Token length of the text in (A) n2c2 and (B) ADE dataset




## A.2 Hyperparameters for the GPT models

| Hyperparameters | Value |
|---|---|
| Tokenization and Context Window | 200 tokens |
| Temperature (Randomness of the model output) | 0 |
| Top p (Top-K Sampling Technique) | 0.95 |
| Presence Penalty (Penalty to discourage model from generating responses that contain certain specified tokens) | -1.0 |

## A.3 Example of the individual entities within the n2c2 and ADE dataset

Table 3: Example of the individual entities within the n2c2 and ADE dataset

| Entities | Examples |
|---|---|
| Drug | Morphine, ibuprofen, antibiotics (abx), chemotherapy (carboplatin) |
| ADE/Reason | Nausea, rash, seizures, vitamin K deficiency |
| Strength | 10 mg, 60 mg |
| Form | Capsule, syringe, tablet, topical (apply topical) |
| Dosage | 60 mg/0.6 mL |
| Frequency | Daily, twice a day, Q4H (every 4 hours) |
| Route | Transfusion, oral, intravenous (IV) |
| Duration | For 10 days, 2 cycles, for a week |

### A.4 Retrieval Augmented Generation

Method:
1. Split UMLS data (MRCONSO.RRF[4]) into manageable chunks (8192 tokens) to facilitate processing. MRCONSO.RRF file contains the UMLS concepts.
2. Generate embeddings for each chunk, capturing its semantic represetations
2. Store the embeddings in a vector database for efficient retrieval
3. Compare each prompt with the stored data in the vector database.
4. Extract the top 30 results with the highest similarity scores between the query and the UMLS data.
5. Concatenate the retrieved results with the prompt to generate the final extraction output.

### A.5 Qualitative Results

Table 4: Some of the qualitative results for the Strength-Drug Pair. (A) Without UMLS integration. (B) With UMLS integration

| | **Examples** |
|---|---|
| A | [(‘aspirin’, ‘81 mg’) (‘atorvastatin’, ‘20 mg’), (‘amiodarone’, ‘200 mg’), (‘metoprolol tartrate’, ‘50 mg’), (‘spironolactone’, ‘25 mg’), (‘acetaminophen’, ‘325 mg’), (‘ranitidine HCl’, ‘150 mg’), (‘prednisone’, ‘60 mg’)] |
| B | [(‘aspirin’, ‘81 mg’), (‘atorvastatin’, ‘20 mg’), (‘amiodarone’, ‘200 mg’), (‘metoprolol tartrate’, ‘50 mg’), (‘spironolactone’, ‘25 mg’), (‘acetaminophen’, ‘325 mg’), (‘ranitidine HCl’, ‘150 mg’), (‘prednisone’, ‘60 mg’), **(‘Plavix’, ‘75 mg’)**, **(‘ASA’, ‘325’)**, **(‘Cipro’, ‘250 mg’)**] |

[4] https://www.ncbi.nlm.nih.gov/books/NBK9685/table/ch03.T.concept_names_and_sources_file_mr/

### A.6 Comparison of GPT-3.5-turbo for all Entity-Entity Relation pairs with and without UMLS Integration for the n2c2 dataset

| Entity-Entity | GPT-3.5-turbo | | | GPT-3.5-turbo+UMLS | | |
|---|---|---|---|---|---|---|
| | P | R | Micro F-1 | P | R | F-1 |
| Dosage-Drug | 0.75 | 0.75 | 0.75 | 0.80 | 0.80 | 0.80 |
| Duration-Drug | 0.76 | 0.76 | 0.76 | 0.81 | 0.81 | 0.81 |
| Route-Drug | 0.74 | 0.73 | 0.73 | 0.76 | 0.75 | 0.75 |
| Form-Drug | 0.72 | 0.73 | 0.72 | 0.75 | 0.76 | 0.75 |
| ADE-Drug | 0.69 | 0.71 | 0.70 | 0.74 | 0.75 | 0.75 |
| Reason-Drug | 0.73 | 0.74 | 0.74 | 0.76 | 0.77 | 0.777 |
| Frequency-Drug | 0.73 | 0.74 | 0.73 | 0.75 | 0.76 | 0.77 |
| Average | 0.73 | 0.74 | 0.73 | 0.77 | 0.77 | 0.77 |

### A.7 Comparison of GPT-4-32k for all Entity-Entity Relation pairs without UMLS Integration for the n2c2 dataset

| Entity-Entity | GPT-4-32k | | | GPT-4-32k+UMLS | | |
|---|---|---|---|---|---|---|
| | P | R | Micro F-1 | P | R | F-1 |
| Dosage-Drug | 0.77 | 0.77 | 0.77 | 0.82 | 0.82 | 0.82 |
| Duration-Drug | 0.78 | 0.77 | 0.78 | 0.83 | 0.82 | 0.83 |
| Route-Drug | 0.79 | 0.77 | 0.78 | 0.81 | 0.78 | 0.79 |
| Form-Drug | 0.74 | 0.76 | 0.74 | 0.77 | 0.79 | 0.77 |
| ADE-Drug | 0.69 | 0.73 | 0.71 | 0.75 | 0.78 | 0.77 |
| Reason-Drug | 0.74 | 0.75 | 0.735 | 0.77 | 0.78 | 0.76 |
| Frequency-Drug | 0.78 | 0.77 | 0.78 | 0.80 | 0.79 | 0.79 |
| Average | 0.75 | 0.76 | 0.76 | 0.79 | 0.79 | 0.79 |

### A.8 Comparison of Models for all Entity-Entity Relation pairs with UMLS for RAG on the n2c2 dataset

| Entity-Entity | GPT-4-32k | | | GPT-3.5-turbo | | |
|---|---|---|---|---|---|---|
| | P | R | F-1 | P | R | F-1 |
| Dosage-Drug | 0.77 | 0.77 | 0.77 | 0.75 | 0.75 | 0.75 |
| Duration-Drug | 0.79 | 0.78 | 0.78 | 0.76 | 0.77 | 0.77 |
| Route-Drug | 0.79 | 0.77 | 0.78 | 0.74 | 0.73 | 0.73 |
| Form-Drug | 0.74 | 0.73 | 0.74 | 0.73 | 0.74 | 0.74 |
| ADE-Drug | 0.72 | 0.73 | 0.72 | 0.70 | 0.70 | 0.70 |
| Reason-Drug | 0.76 | 0.76 | 0.76 | 0.75 | 0.78 | 0.76 |
| Frequency-Drug | 0.81 | 0.80 | 0.80 | 0.70 | 0.71 | 0.71 |
| Average | 0.77 | 0.76 | 0.76 | 0.73 | 0.74 | 0.74 |

### A.9 Comparison of GPT-4-32k for all Entity-Entity Relation pairs with and without UMLS on the ADE dataset

| Entity-Entity | GPT-4-32k | | | GPT-4-32k+UMLS | | |
|---|---|---|---|---|---|---|
| | P | R | F-1 | P | R | F-1 |
| Dosage-Drug | 1.0 | 0.66 | 0.795 | 1.00 | 0.85 | 0.91 |
| ADE-Drug | 1.0 | 0.73 | 0.84 | 1.00 | 0.93 | 0.97 |
| Average | 1.0 | 0.70 | 0.82 | 1.0 | 0.89 | 0.94 |

### A.10 Comparison of GPT-3.5-turbo for all Entity-Entity Relation pairs with and without UMLS on the ADE dataset

| Entity-Entity | GPT-3.5-turbo | | | GPT-3.5-turbo+UMLS | | |
|---|---|---|---|---|---|---|
| | P | R | F-1 | P | R | F-1 |
| ADE-Drug | 0.57 | 0.53 | 0.55 | 0.60 | 0.65 | 0.62 |
| Dosage-Drug | 0.68 | 0.61 | 0.64 | 0.70 | 0.75 | 0.72 |
| Average | 0.625 | 0.57 | 0.596 | 0.83 | 0.70 | 0.75 |

### A.11 Comparison of the models for all Entity-Entity Relation pairs with UMLS for RAG on the ADE dataset

| Entity-Entity | GPT-4-32k | | | GPT-3.5-turbo | | |
|---|---|---|---|---|---|---|
| | P | R | F-1 | P | R | F-1 |
| ADE-Drug | 1.0 | 0.73 | 0.84 | 0.62 | 0.61 | 0.60 |
| Dosage-Drug | 1.0 | 0.75 | 0.86 | 0.68 | 0.65 | 0.66 |
| Average | 1.0 | 0.74 | 0.85 | 0.65 | 0.63 | 0.64 |