End-to-End Deep Learning for Named Entity Recognition and Relation Extraction in Gut-Brain Axis PubMed Abstracts

Aleksis Datseris[1,2], Mario Kuzmanov[3,1], Ivelina Nikolova-Koleva[1,4], Dimitar Taskov[5,6] and Svetla Boytcheva[1,2]

[1]Ontotext, Sofia, Bulgaria
[2]Faculty of Mathematics and Informatics, Sofia University St. Kliment Ohridski, Sofia, Bulgaria
[3]Tübingen University, Tübingen, Germany
[4]IICT, Bulgarian Academy of Sciences, Sofia, Bulgaria
[5]Multiprofile Hospital for Active Treatment in Neurology and Psychiatry "St. Naum", Sofia, Bulgaria
[6]Medical University of Sofia, Sofia, Bulgaria
aleksis.datseris@graphwise.ai

Abstract

This is a comparative study tackling named entity recognition and relation extraction from PubMed abstracts with focus on the gut-brain interplay. The proposed systems for named entity recognition cover a range of models and techniques from traditional gazetteer-based approaches, transformer-based approaches, transformer domain adaptation, large models pretraining as well as LLM prompting. The best performing model among these achieves 82.53% F1-score. The relation extraction task is addressed with ATLOP and LLMs and their best results reach F1 up to 63.80% on binary relation extraction, 89.40% on ternary tag-based relation extraction and 40.32% on ternary mention-based relation extraction.

1 Introduction

This paper presents a research on named entity recognition (NER) and relation extraction (RE) in the biomedical domain, focusing on gut-brain interplay. The results, reported here, are part of the GutBrainIE Task at CLEF 2025 BioASQ Lab (Nentidis et al., 2025; Martinelli et al., 2025). The challenge aims to promote automatic extraction of gut-brain related facts from scientific literature and evaluates four subtasks: NER across 13 categories, binary relation extraction, ternary tag-based relation extraction (with 19 predicates), and ternary mention-based relation extraction (with 19 predicates).

State-of-the-art Natural Language Processing (NLP) for biomedical literature employs deep learning (DL) (e.g., BERT, BiLSTM-CRF), classical machine learning (ML) (CRF, SVM), and rule-based methods. NER achieves F1-score up to 89.3% on PubMed articles, while RE performance varies from 47.7% to 88.5% (Goyal and Singh, 2024; Luo et al., 2022; Névéol et al., 2011; Sänger and Leser, 2021). Key entities include genes, proteins, diseases, and drugs; typical relations are gene-disease and drug-treatment. While most approaches tackle both NER and RE tasks (Luo et al., 2022), there are also specific techniques tailored for RE exclusively (Hassan et al., 2023), (Sänger and Leser, 2021).

Recent works leverage Large Language Models (LLMs), shifting NER from application of sequence labeling methods to a natural language generation task. LLM-based approaches for NER include zero/few-shot prompt engineering (Lu et al., 2024; Hu et al., 2024), retrieval augmented generation (RAG) (Monajatipoor et al., 2024), and hybrid models combining LLMs with external knowledge or traditional methods (Bian et al., 2023; García-Barragán et al., 2025; Biana et al., 2024; Rohanian et al., 2024; Zhou et al., 2023). While LLMs show strong in category classification tasks, challenges with span identification persist. They are typically addressed via post-processing (Lu et al., 2025). Approaches like INSPIRE (Bian et al., 2023) and VANER (Biana et al., 2024) use Chain-of-Thought reasoning and external resources, outperforming prior BioNER systems, with F1-score up to 94% on some categories. Instruction-based paradigms (e.g. BioNER-LLaMA) now surpass GPT-4 in few-shot learning (Keloth et al., 2024). For RE, LLM-based methods exhibit similar advances, with external knowledge and prompt design as critical factors.

This paper proposes an in-depth comparison of methods for NER including: (i) transformer fine-tuning - GLiNER (Zaratiana et al., 2024), BiomedNLP ELECTRA (Tinn et al., 2021), BioBERT PubMed (Gu et al., 2022), XLM-R (Conneau et al., 2020); (ii) domain adaptation (DA) on XLM-R and BiomedBERT; (iii) fine-tuning of large encoder models flan-t5-xl (Chung et al., 2022), t5-xxl (Raffel et al., 2023) and Gatortron-medium (Yang et al., 2022a) and (iv) LLM prompting. For RE subtasks, ATLOP (Zhou et al., 2020) and LLM-assisted extraction are compared.

Proceedings of Recent Advances in Natural Language Processing,pages 254–259
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-031

2 Data

2.1 Annotated Data

The data provided by the challenge organizers[1] consists of documents, sourced from PubMed[2], focusing on the gut-brain interplay and its impact on neurological and mental health. The training data is divided into four collections (Martinelli et al., 2025), each of them with annotations for NER and RE tasks, explained in the introduction. **Platinum collection (P)** (111 documents) - expert-curated annotations reviewed by biomedical specialists; **Gold collection (G)** (208 documents) - expert-curated; **Silver collection (S)** (499 documents) - annotated by trained students; **Bronze collection (B)** (750 articles) - annotated with distant supervision using GLiNER for NER and ATLOP for RE task. In addition to the training data, a separate **Validation/Dev set (D)** is provided. The articles in D are a held out non-overlapping sample of P and G corpora.

2.2 Augmented Data

PubMed articles from the period 2015-2025 are added to the training dataset in order to address the scarcity issue in key entity categories. The documents are extracted using the PubMed API with search parameters corresponding to the entity categories e.g. genes or diagnosis, all within the gut-brain axis topic. GLiNER and BiomedNLP ELECTRA (both fine-tuned on P+G+S+B) are then applied for distant supervision of named entities. The result is a Bronze-Standard entity annotated corpus (BA) of 6728 articles. The annotations from both models are used in various combinations during the experiments.

2.3 Gazetteers

For some of the underrepresented categories (e.g., genes, food) named entity gazetteers are created to overcome the low recall. For *biomedical technique* terms, organizer-provided mappings are used. The gazetteer includes all children concepts of the provided mapping concept. The other gazetteers, such as for *gene*, *drug*, *disease* etc., are generated based on the UMLS[3] semantic network. In total, nine gazetteers are created, containing between 18 to 2,001,200 alternative biomedical term names.

3 Named Entity Recognition

3.1 Deep Learning Approaches

To get a sense of the data, **GatorTron-Base** (Yang et al., 2022b), mainly pretrained on clinical notes with 345M parameters, was trained on P + G + S + B concatenated datasets and evaluated on the D set. Secondly, **GLiNER** (Zaratiana et al., 2024), pretrained on the PileNER corpus[4] was fine-tuned on P + G + S + B + BA collections, and tested on the D set. Next experiments were with a more specialized model in the PubMed domain - **BiomedNLP ELECTRA** (Tinn et al., 2021), pretrained on PubMed abstracts. Exactly the same train/test split as with GLiNER was used. The **BiomedNLP ELECTRA** version trained on full-text articles lead to decrease in performance. The most successful biomedical model of **ScispaCy** turns to be - *en_core_sci_md*, fine-tuned only on the P + G + S collections with an internal validation split. In all experiments, the context window was 512 tokens. Results are presented in Section 4.1.

3.2 Gazetteer Matching

Gazetteer matching with exact match was applied in combination with the other approaches, using the resources described in Section 2.3. Unfortunately the improvements were only in one dimension, e.g. they increased significantly the recall of genes but removing the noise, they bring in was too costly. The results are presented in Section 4.1.

3.3 Domain Adaptation

To help some of the models with their understanding of the underlying dependencies in the texts, domain adaptation was performed. XLM-R (Conneau et al., 2020) and BiomedBERT (Biomed) (Gu et al., 2020) models and masked language modeling (Devlin et al., 2019) pretraining objective were selected. The results are presented in Section 4.2.

3.4 Large Models

Experiments were performed also with larger encoder (Devlin et al., 2019) models and larger encoder-decoder (Vaswani et al., 2023) models by taking only the encoder part of the model and using it as a standard encoder model. The selected models were flan-t5-xl (Chung et al., 2022), t5-xxl (Raffel et al., 2023) and gatortron-medium (Yang et al.,

[1] https://hereditary.dei.unipd.it/challenges/gutbrainie/2025/

[2] https://pubmed.ncbi.nlm.nih.gov/

[3] https://www.nlm.nih.gov/research/umls/index.html

[4] https://huggingface.co/datasets/Universal-NER/Pile-NER-type

2022a). Due to computational limitations, the models were fine-tuned using LoRA (Hu et al., 2021) adapters with rank 256 applied to the attention matrices. The results are presented in Section 4.2.

3.5 LLM approach

Two individual experiments GPT-4.1[5] were also conducted to explore capabilities of LLMs to tackle the NER task. In the first scenario, the prompt was instruction-example-based and included 3 examples (3-shot), manually adapted from the P dataset, detailed description of all 13 entity types, including definitions from the task description. In addition detailed instructions regarding the output format were provided along with restrictions and limitations on the resolution of some specific cases.

The second approach was designed with instruction-based prompt only (0-shot). It was simpler that the first one and included the names on the 13 entity types with short explanation only, without definitions. It included again instructions regarding the desired output format but no examples of sample input/output were provided. The differences with the first prompt were the additional instructions on the resolution of some border cases and overlapping concepts, as nested entity resolution is a major topic in BioNER (Park et al., 2024).

The LLM output was shaped with the help of post-processing scripts. In both approaches these were used for calibrating the start and end offsets of the entities; removing duplicated, overlapping and corrupted predictions. In the first approach specifically, consecutive entities with one and the same type were merged. In the second one - corner cases were compared with and without merging.

4 NER results

4.1 Deep Learning & Gazetteers Results

The results from all approaches discussed in Sections 2.3 and 3.1, as well as some hybrid approaches are shown in Table 1. The largest model GatorTron-Base gives a stable baseline. This is a good indicator that the dataset is suitable for lightweight models like GLiNER, which is the best in micro-recall. Overall, the best performing model is BiomedNLP ELECTRA. The worst performing approach is to use the Gazetteers on their own, not being able to account for most of the categories. The most precise system turns out to be the combination of ScispaCy with integrated Gazetteers and BiomedNLP ELECTRA. For this hybrid approach, we have taken the union of the predictions of ScispaCy + Gazetteers and BiomedNLP ELECTRA. While it does provide competitive results, the trade-off between precision and recall is too large. Still, the potential of such hybrid systems should be further explored. The performance of the same models on token-level predictions tend to average 84% micro-F1, opposed to the lower entity-level results shown in Table 1.

Model	**P**	**R**	**F1**
GatorTron-Base	77.37%	82.63%	79.91%
BNLP ELECTRA	82.60%	82.45%	**82.53%**
GLiNER	81.05%	**82.72%**	81.88%
ScispaCy	76.65%	71.71%	74.10%
Gazz	16.30%	12.71%	14.29%
ScispaCy + Gazz + BNLP ELECTRA	**91.41%**	69.56%	79.00%
ScispaCy + Gazz	56.38%	70.37%	62.60%
GPT (3-shot)	45.88%	67.32%	54.57%
GPT (0-shot)	42.01%	54.79%	46.38%

Table 1: Entity-level micro-scores of DL, hybrid systems and LLM approach for NER on set D (BNLP=Biomed NLP, Gazz=Gazetteers)

4.2 Domain Adaptation & Large Models Results

The results of DA on XLM-R (Conneau et al., 2020) and BiomedBERT (Gu et al., 2020) are shown in Table 2. There is consistent improvement of token-level performance by applying DA on the models. While BiomedBERTs' token-level performance is good, the entity-level performance of the model was only 80.35%.

Token-level performance of *gatortron-medium* and *flan-t5-xl* is shown on Table 1. The models are significantly larger, however their results are not impressive; their performance is worse than the best DA model and transformer models, respectively.

4.3 LLM Approach Results

The results of the two LLM-based NER approaches, GPT (3-shot) and GPT (0-shot), are shown in Table 1. They perform significantly worse than supervised and dictionary-based ones, mainly due to poor identification of entity boundaries, which lowers precision and missed concepts, which lowers recall. The LLMs also tend to confuse entity categories (*chemical*, *gene* and *disease*) and favor shorter entities over longer ones. GPT (3-shot) outperforms GPT (0-shot), as expected.

[5] https://openai.com/index/gpt-4-1/

Model	P	R	F1
BiomedBERT	82.68%	83.51%	82.36%
BiomedBERT + DA	**84.26%**	83.42%	**83.05%**
XLM-R	78.70%	**84.76%**	81.62%
XLM-R + DA	80.46%	83.06%	81.74%
flan-t5-xl	77.75%	84.22%	**81.86%**
gatortron-medium	77.59%	83.78%	80.56%

Table 2: Token level micro-scores of DA and large models for NER on set D.

5 Relation Extraction Approaches

5.1 ATLOP

One of the methods used for the relation extraction subtask was the **Adaptive Thresholding and Localized Context Pooling (ATLOP)**. It is a well-recognised approach for relation extraction that utilizes as a base model a standard pre-trained transformer (Vaswani et al., 2023) encoder, such as BERT (Devlin et al., 2019).

5.2 LLM approach

The LLM-based approaches for RE were organized in a cascade manner using GPT-4.1. Each of them used as input the output of the previous step, starting from the NER result, followed by the binary relations, ternary tag-relation and finally ternary mention-relation. All methods were based on specially designed instruction-example-based prompt, with list of entity types and detailed definition of relations and their domain/range, according to the task definition. All prompts contained output format instructions and one example per relation category. In the prompt for binary relation extraction (1-shot) the input included the PubMed abstract and title and all extracted named entities from the NER subtask. In the ternary-tag RE prompt (2-shot), the input included extracted entitles and relations from the NER task and binary RE sub-tasks, plus the full context from the PubMed abstract and title. The ternary-mention RE prompt (1-shot) includes the PubMed abstract and title, extracted named entities from the NER sub-task and the extracted ternary-tag relations from the previous task.

6 Relation Extraction Results

The base model for ATLOP was BiomedNLP-KRISSBERT-PubMed-UMLS-EL (Zhang et al., 2021). It was provided with the extracted entities by the baseline GLiNER model. The micro-scores on the D set are shown in Table 3. The model achieves fairly good results, but with the increasing difficulty of each task, they degrade.

The input at the first step of for all LLM-based experiments includes the entities and relations from dataset D. The results (Table 3) on the binary relation subtask show that the LLM-approach outperforms the ATLOP model on micro-R and has negligibly better micro-F1 score and lower micro-P. The results on the ternary-tag RE significantly outperform ATLOP. However, this is not the case for ternary-mention-based extraction.

Model	P	R	F1	Task
KRISSBERT	68.04%	60.00%	63.77%	Binary RE
GPT (1-shot)	57.99%	**70.91%**	**63.80%**	Binary RE
KRISSBERT	67.35%	57.39%	61.97%	Tag RE
GPT (2-shot)	**95.10%**	**84.35%**	**89.40%**	Tag RE
KRISSBERT	46.60%	35.54%	40.32%	Mention RE
GPT (1-shot)	29.01%	**44.29%**	35.05%	Mention RE

Table 3: Micro-scores of RE systems on set D.

7 Discussion, Error Analysis, Conclusions

On the NER task, transformer-based models gave the best results. One of the most underrepresented categories in the data was also among the hardest to learn - *gene*. GatorTron-Base was the best model for it. The integration of traditional approaches like Gazetteers increased the recall only for the gene category and achieved very good precision (and low recall) for all 13 categories. From the more frequent entities, *chemical* turned out to be a bottleneck, while the others like *DDF, microbiome* and *human* all got F1 between 84% and 95%. Although all of the systems could extract the text mentions, classifying them correctly was challenging.

On the RE task, **ATLOP** is still limited by the fact that the model needs as input the extracted entities in order to classify the relations. It is fairly good at the first 2 RE subtasks, but once it needs to classify both the relation type and the exact spans, its performance starts to drop. It often mistakes exactly which entities are in a relation. It also often confuses the relation types ”impact”, ”influence”, and ”affect”, with ”is linked to”. LLM based approaches demonstrate very good performance on some of the RE subtasks. In the same time they struggle to identify correct mentions, and a common issue is the mismatch of the relation direction, because all relations are anti-symmetric and the order does matter.

Acknowledgments

This work is partially supported by European Union's Horizon research and innovation programme projects RES-Q PLUS [Grant Agreement No. 101057603] and HEREDITARY [Grant Agreement No. 101137074]. Views and opinions expressed are however those of the author only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

References

Junyi Bian, Jiaxuan Zheng, Yuyi Zhang, and Shanfeng Zhu. 2023. Inspire the large language model by external knowledge on biomedical named entity recognition. *arXiv preprint arXiv:2309.12278*.

Junyi Biana, Weiqi Zhai, Xiaodi Huang, Jiaxuan Zheng, and Shanfeng Zhu. 2024. Vaner: leveraging large language model for versatile and adaptive biomedical named entity recognition. *arXiv preprint arXiv:2404.17835*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Álvaro García-Barragán, Ahmad Sakor, Maria-Esther Vidal, Ernestina Menasalvas, Juan Cristobal Sanchez Gonzalez, Mariano Provencio, and Víctor Robles. 2025. Nssc: a neuro-symbolic ai system for enhancing accuracy of named entity recognition and linking from oncologic clinical notes. *Medical & Biological Engineering & Computing*, 63(3):749–772.

Nandita Goyal and Navdeep Singh. 2024. Named entity recognition and relationship extraction for biomedical text: A comprehensive survey, recent advancements, and future research directions. *Neurocomputing*, page 129171.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthc.*, 3(1):1–23.

Nesma Abdel Aziz Hassan, Rania Ahmed Abdel Azeem Abul Seoud, and Dina Ahmed Salem. 2023. Open information extraction methodology for a new curated biomedical literature dataset. *International Journal of Advanced Computer Science and Applications*, 14(7).

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Yan Hu, Qingyu Chen, Jingcheng Du, Xueqing Peng, Vipina Kuttichi Keloth, Xu Zuo, Yujia Zhou, Zehan Li, Xiaoqian Jiang, Zhiyong Lu, et al. 2024. Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association*, 31(9):1812–1820.

Vipina K Keloth, Yan Hu, Qianqian Xie, Xueqing Peng, Yan Wang, Andrew Zheng, Melih Selek, Kalpana Raja, Chih Hsuan Wei, Qiao Jin, et al. 2024. Advancing entity recognition in biomedicine via instruction tuning of large language models. *Bioinformatics*, 40(4):btae163.

Qiuhao Lu, Rui Li, Andrew Wen, Jinlian Wang, Liwei Wang, and Hongfang Liu. 2025. Large language models struggle in token-level clinical named entity recognition. In *AMIA Annual Symposium Proceedings*, volume 2024, page 748.

Zhiyong Lu, Yifan Peng, Trevor Cohen, Marzyeh Ghassemi, Chunhua Weng, and Shubo Tian. 2024. Large language models in biomedicine and health: current research landscape and future directions. *Journal of the American Medical Informatics Association*, 31(9):1801–1811.

Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N Arighi, and Zhiyong Lu. 2022. Biored: a rich biomedical relation extraction dataset. *Briefings in Bioinformatics*, 23(5):bbac282.

M. Martinelli, G. Silvello, V. Bonato, G. M. Di Nunzio, N. Ferro, O. Irrera, S. Marchesin, L. Menotti, and F. Vezzani. 2025. Overview of GutBrainIE@CLEF 2025: Gut-Brain Interplay Information Extraction. In *CLEF 2025 Working Notes*.

Masoud Monajatipoor, Jiaxin Yang, Joel Stremmel, Melika Emami, Fazlolah Mohaghegh, Mozhdeh Rouhsedaghat, and Kai-Wei Chang. 2024. Llms in biomedicine: A study on clinical named entity recognition. *arXiv preprint arXiv:2404.07376*.

Anastasios Nentidis, Georgios Katsimpras, Anastasia Krithara, Martin Krallinger, Miguel Rodríguez-Ortega, Eduard Rodriguez-López, Natalia Loukachevitch, Andrey Sakhovskiy, Elena Tutubalina, Dimitris Dimitriadis, Grigorios Tsoumakas, George Giannakoulas, Alexandra Bekiaridou, Athanasios Samaras, Giorgio Maria Di Nunzio, Nicola Ferro, Stefano Marchesin, Marco Martinelli, Gianmaria Silvello, and Georgios Paliouras. 2025. Overview of BioASQ 2025: The thirteenth BioASQ challenge on large-scale biomedical semantic indexing and question answering. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025)*.

Aurélie Névéol, Rezarta Islamaj Doğan, and Zhiyong Lu. 2011. Semi-automatic semantic annotation of pubmed queries: a study on quality, efficiency, satisfaction. *Journal of biomedical informatics*, 44(2):310–318.

Yesol Park, Gyujin Son, and Mina Rho. 2024. Biomedical flat and nested named entity recognition: Methods, challenges, and advances. *Applied Sciences*, 14(20).

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer.

Omid Rohanian, Mohammadmahdi Nouriborji, Samaneh Kouchaki, Farhad Nooralahzadeh, Lei Clifton, and David A. Clifton. 2024. Exploring the effectiveness of instruction tuning in biomedical language processing. *Artificial Intelligence in Medicine*, 158:103007.

Mario Sänger and Ulf Leser. 2021. Large-scale entity representation learning for biomedical relationship extraction. *Bioinformatics*, 37(2):236–242.

Robert Tinn, Hao Cheng, Yu Gu, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Fine-tuning large neural language models for biomedical natural language processing. *arXiv [cs.CL]*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B Costa, Mona G Flores, Ying Zhang, Tanja Magoc, Christopher A Harle, Gloria Lipori, Duane A Mitchell, William R Hogan, Elizabeth A Shenkman, Jiang Bian, and Yonghui Wu. 2022a. A large language model for electronic health records. *npj Digital Medicine*, 5(1):194.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Mona G Flores, Ying Zhang, Tanja Magoc, Christopher A Harle, Gloria Lipori, Duane A Mitchell, William R Hogan, Elizabeth A Shenkman, Jiang Bian, and Yonghui Wu. 2022b. Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. GLiNER: Generalist model for named entity recognition using bidirectional transformer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sheng Zhang, Hao Cheng, Shikhar Vashishth, Cliff Wong, Jinfeng Xiao, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Knowledge-rich self-supervision for biomedical entity linking.

Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. 2020. Document-level relation extraction with adaptive thresholding and localized cOntext pooling. *arXiv [cs.CL]*.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2023. Universalner: Targeted distillation from large language models for open named entity recognition.