2024 IEEE International Conference on Big Data and Smart Computing (BigComp) | 979-8-3503-7002-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/BIGCOMP60711.2024.00095

LLM based Biological Named Entity Recognition from Scientific Literature

Sung Jae Jung
Division of DG and AI Business
En-Core Inc.
Seoul, South Korea
sjjung@en-core.com

Hajung Kim
Division of DT Business
En-Core Inc.
Seoul, South Korea
hjkim@en-core.com

Kyoung Sang Jang
Division of DG and AI Business
En-Core Inc.
Seoul, South Korea
ksjang@en-core.com

***Abstract*—Recently, the application of Large Language Models (LLMs) in the field of natural language processing has witnessed remarkable growth, revolutionizing the field of bioinformatics by automating the extraction of biological entities from scientific literature. This study presents the development and evaluation of a Biological Named Entity Recognizer (BNER) using a pre-trained Large Language Model (LLM) refined through prompt engineering. The BNER was tailored to identify proteins, genes, and small molecules within scientific texts, specifically targeting the context of p53 protein-related research. To assess the BNER's efficacy, we curated a dataset comprising ten paragraphs extracted from the abstracts and significant sections of five high-relevance scientific papers. The system's performance was quantified through an entity recognition task, resulting in 51 true positives (TP), 10 false positives (FP), and 3 false negatives (FN). The BNER achieved an F1 score of 0.887, demonstrating a high degree of precision and recall. These results validate the utility of LLMs in bioinformatics and highlight the BNER's potential to support and accelerate scientific discovery by providing accurate, structured data outputs suitable for comprehensive analysis.**

***Index Terms*—Biological Named Entity Recognition (BNER), Large Language Models (LLM), Prompt Engineering, p53 Protein**

I. Introduction

Recently, the application of Large Language Models (LLMs) in the field of natural language processing has witnessed remarkable growth, revolutionizing how textual data is processed and analyzed [1]. One area of significant interest is the extraction of biological named entities from scientific literature, a task of high importance in bioinformatics and computational biology. This paper aims to explore the efficacy of LLMs in identifying and extracting biological entities, such as genes, proteins, diseases, and drugs, from the abstracts of scientific papers retrieved from PubMed [2].

The accurate recognition of these entities is crucial for several downstream tasks, including gene-disease association studies, drug repurposing, and understanding molecular mechanisms of diseases. Traditional approaches in biological named entity recognition (NER) have relied on rule-based systems or machine learning algorithms, which often require extensive feature engineering and domain-specific knowledge [3]. These approaches have relied heavily on manual curation and domain-specific algorithms, which are both time-intensive and laborious [4].

By contrast, LLMs offer an innovative approach with their ability to process and learn from vast datasets, potentially surpassing the limitations of traditional methods. This study evaluates the efficacy of a prompt engineered LLM, GPT-4, in extracting biological entities from a selected set of 5 scientific papers retrieved from PubMed [5] [6] [7] [8] [9]. The evaluation will focus on the comprehensiveness and accuracy of the entity extraction.

The outcome of this research will not only contribute to the field of computational biology but also provide insights into the adaptability and effectiveness of LLMs in handling specialized and domain-specific language prevalent in scientific literature.

II. Function Development and Evaluation

A. Developing Biological Entity Recognizer by Prompt

Based on GPT-4, we developed a Biological Named Entity Recognizer (BNER) designed to process scientific text and extract biological entities. The system maps each entity to its corresponding Reactome ID and PubMed GeneID. It then generates a CSV file as output, which organizes the results for ease of analysis and clarity. The BNER was configured to parse text input, identify biological entities, and format the output into a structured CSV file. This configuration included setting up the extraction rules and output schema, which defined the columns for entity names, Reactome IDs, and PubMed GeneIDs(Fig.1 1).

A pre-trained language model, GPT-4, was instructed by prompts with no additional annotated data set is provided for fine tuning the model. BNER was instructed to post-process output into a well-structured CSV file. This module lists the extracted entities and their mappings, placing them into the corresponding columns as per the predefined schema.

B. Dataset and Performance Evaluation

For the evaluation of the BNER system, we focused on the tumor protein p53, known for its critical role in cancer onset and as a target for chemotherapy agents. We conducted a PubMed search to identify papers highly relevant to p53, with a search keyword 'p53' given to PubMed website [3] The search results were sorted by relevance, and the top five papers were selected for inclusion in our evaluation dataset.

2375-9356/24/$31.00 ©2024 IEEE
DOI 10.1109/BigComp60711.2024.00095

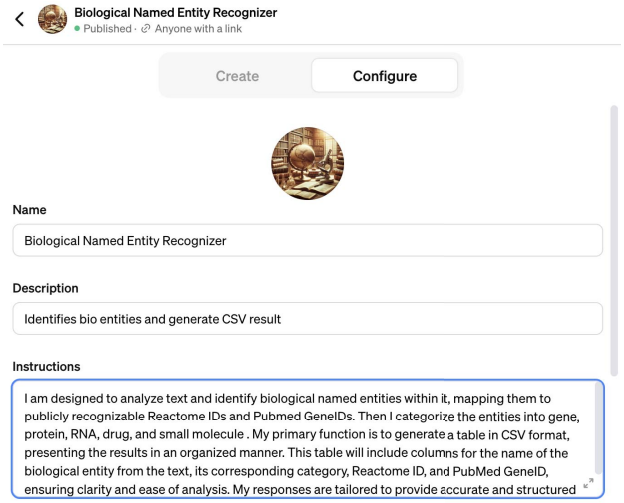
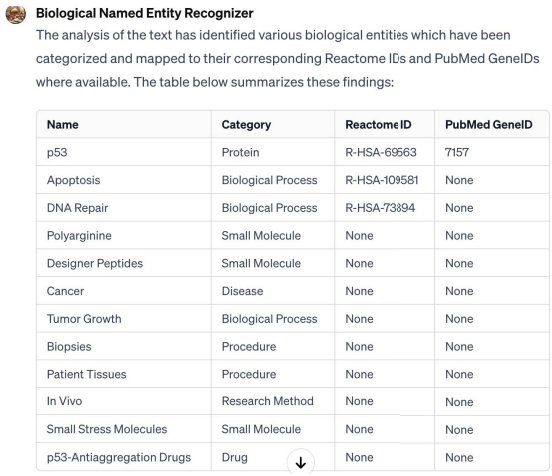


Fig. 1. The prompt instruction for engineering leveraging a LLM, GPT-4, to give a birth to BNER. We have instructed BNER to read scientific literature and identify biological entities from it. BNER is further instructed to classify the entities into one of protein, gene, or small molecules, and map them to the publicly recognizable PubMed IDs and Reactome IDs



Name	Category	Reactome ID	PubMed GeneID
p53	Protein	R-HSA-69563	7157
Apoptosis	Biological Process	R-HSA-109581	None
DNA Repair	Biological Process	R-HSA-73894	None
Polyarginine	Small Molecule	None	None
Designer Peptides	Small Molecule	None	None
Cancer	Disease	None	None
Tumor Growth	Biological Process	None	None
Biopsies	Procedure	None	None
Patient Tissues	Procedure	None	None
In Vivo	Research Method	None	None
Small Stress Molecules	Small Molecule	None	None
p53-Antiaggregation Drugs	Drug	None	None

Fig. 2. The result of an execution of BNER for an abstract from our curated literature. This execution extracted a protein "p53" with its corresponding Reactome ID and PubMed Gene ID. In addition to proteins and small molecules, BNER extracts biological processes, procedures, and research methods, which is beyond the instructions. In the performance evaluation, We excluded the entities recognized beyond instruction.

From each of the five selected papers, we extracted the abstract and an additional paragraph that appeared most pertinent for performance testing. These text excerpts were chosen based on their expected density of biological named entities and their relevance to p53's role in cancer and therapy. The extracted texts underwent a human curation process to identify and label biological entities. Domain experts annotated each entity in abstracts and the selected paragraphs. This human-curated data formed the ground truth against which the BNER system's performance was to be evaluated.

The BNER system was assessed using standard performance metrics including precision (the proportion of true positive results among all positive results identified by the BNER), recall (the proportion of true positive results identified by the BNER out of all actual positive results), and the F1 score (the harmonic mean of precision and recall). These metrics provided a quantitative measure of the BNER's entity recognition accuracy.

III. Result

The evaluation of the Biological Named Entity Recognizer (BNER) was conducted on a set of ten paragraphs extracted from the abstracts and relevant sections of five scientific papers focused on the p53 protein. Fig.2 shows the result of an execution of BNER for one of the abstracts. The NER task was aimed at identifying entities within three categories: proteins, genes, and small molecules. The performance of the BNER was quantified using the following metrics derived from the NER results:

- **True Positives (TP):** The number of instances where the BNER correctly identified a true entity. The BNER successfully recognized 51 true entities across the dataset.
- **False Positives (FP):** The number of instances where the BNER incorrectly identified an entity. In our evaluation, there were 10 false positives, indicating instances where the BNER predicted an entity that was not present as a true entity in the ground truth.
- **False Negatives (FN):** The number of instances where the BNER failed to identify an actual entity. The evaluation noted 3 false negatives, signifying entities that were present in the text but were not recognized by the BNER.

From these values, we calculated the precision, recall, and F1 score as follows:

- **Precision:** $\frac{TP}{TP+FP} = 0.836$
- **Recall:** $\frac{TP}{TP+FN} = 0.944$
- **F1 Score:** $\frac{Precision \times Recall}{Precision+Recall} = 0.887$

The BNER demonstrated a high degree of accuracy in identifying biological entities pertinent to the domain of interest, as indicated by an F1 score of 0.887. This performance indicates the model's robustness and precision in biological named entity recognition from scientific literature. The BNER, with its high F1 score, represents a significant step forward in automating the extraction and classification of biological entities, thereby facilitating enhanced data analysis and supporting the acceleration of scientific discovery.

IV. Conclusion

In this study, we successfully developed a Biological Named Entity Recognizer (BNER) leveraging the capabilities of a pre-trained Large Language Model (LLM), GPT-4, enhanced through prompt engineering. Our approach capitalized on the LLM's extensive base knowledge, refined through carefully designed prompts that directed the model's focus towards the

identification and categorization of biological entities within scientific literature.

The evaluation of the BNER was conducted using a dataset crafted from relevant scientific papers with a focus on the p53 protein, a critical element in cancer pathology and a target for chemotherapeutic agents. The dataset, enriched with annotations from domain experts, provided a rigorous platform for assessing the BNER's performance.

Our findings demonstrate that the BNER achieved an F1 score of 0.887, indicating a high level of precision and recall in the context of a complex domain-specific task. This performance benchmark shows the BNER's ability to accurately identify and classify biological entities, which is a testament to the efficacy of the LLM when combined with strategic prompt engineering.

The implications of our research are twofold. Firstly, it validates the potential of LLMs in the domain of biological named entity recognition, suggesting that such models can be prompt-engineered to meet specialized requirements of scientific literature analysis. Secondly, it establishes a methodological precedent for utilizing pre-trained language models in conjunction with expert annotations, providing a robust framework for future developments in bioinformatics tools.

REFERENCES

[1] Kim JK, Chua M, Rickard M, Lorenzo A. ChatGPT and large language model (LLM) chatbots: The current state of acceptability and a proposal for guidelines on utilization in academic medicine. J Pediatr Urol. 2023 Oct;19(5):598-604. doi: 10.1016/j.jpurol.2023.05.018. Epub 2023 Jun 2. PMID: 37328321.

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

[3] Holzinger A, Dehmer M, Jurisica I. Knowledge Discovery and interactive Data Mining in Bioinformatics–State-of-the-Art, future challenges and research directions. BMC Bioinformatics. 2014;15 Suppl 6(Suppl 6):I1. doi: 10.1186/1471-2105-15-S6-I1. Epub 2014 May 16. PMID: 25078282; PMCID: PMC4140208.

[4] Chun, HW. et al. (2013). Pathway Construction and Extension Using Natural Language Processing. In: Yamamoto, S. (eds) Human Interface and the Management of Information. Information and Interaction for Health, Safety, Mobility and Complex Environments. HIMI 2013. Lecture Notes in Computer Science, vol 8017. Springer, Berlin, Heidelberg.

[5] Kanapathipillai M. Treating p53 Mutant Aggregation-Associated Cancer. Cancers (Basel). 2018 May 23;10(6):154. doi: 10.3390/cancers10060154. PMID: 29789497; PMCID: PMC6025594.

[6] Sabapathy K, Lane DP. Understanding p53 functions through p53 antibodies. J Mol Cell Biol. 2019 Apr 1;11(4):317-329. doi: 10.1093/jmcb/mjz010. Erratum in: J Mol Cell Biol. 2019 Dec 19;11(12):1105. PMID: 30907951; PMCID: PMC6487784.

[7] Koo N, Sharma AK, Narayan S. Therapeutics Targeting p53-MDM2 Interaction to Induce Cancer Cell Death. Int J Mol Sci. 2022 Apr 30;23(9):5005. doi: 10.3390/ijms23095005. PMID: 35563397; PMCID: PMC9103871.

[8] Nagpal I, Yuan ZM. The Basally Expressed p53-Mediated Homeostatic Function. Front Cell Dev Biol. 2021 Nov 23;9:775312. doi: 10.3389/fcell.2021.775312. PMID: 34888311; PMCID: PMC8650216.

[9] Zavileyskiy L, Bunik V. Regulation of p53 Function by Formation of Non-Nuclear Heterologous Protein Complexes. Biomolecules. 2022 Feb 18;12(2):327. doi: 10.3390/biom12020327. PMID: 35204825; PMCID: PMC8869670.