Automatic BioNER Data Annotation with Large Language Models

Jenny Cai, Victoria Gao, Isabella Struckman, Ryan Welch
MIT

Abstract

This paper introduces a novel approach for biomedical Named Entity Recognition (BioNER) that leverages OpenAI's GPT-3.5 for efficient dataset annotation. Faced with the challenge of rapidly expanding biomedical literature, traditional expert human-annotated methods of labeling BioNER datasets are too slow and costly. Our method employs GPT-3.5's zero-shot, one-shot, and few-shot learning capabilities to generate datasets that mirror the quality of expert human annotation. We validated our approach intrinsically by comparing it against human-labeled datasets and extrinsically by evaluating the performance of two state-of-the-art Biomedical Pretrained Language Models (BioLinkBERT and BioGPT) after fine-tuning on our GPT-labeled datasets. Our approach achieves a promising level of accuracy, comparable to human-annotated data, and shows that BPLMs fine-tuned with our datasets perform well. This study highlights the potential of large language models to revolutionize BioNER dataset annotation, which will result in significant advancements in the field of biomedical natural language processing.

1 Introduction

Named entity recognition of biomedical entities (BioNER) in text is essential for facilitating biology research as it assists researchers in quickly finding relevant information from biomedical texts, which are typically unstructured. It's also essential for performing downstream NLP tasks such as information extraction, summarization, and question answering (QA) in the medical domain. Models that perform NLP tasks are essential for keeping up with the biomedical field as it rapidly expands: over 1 million biomedical papers are added to the PubMed database yearly (about two papers per minute) (Huang et al., 2020; Landhuis, 2016).

Current state-of-the-art (SOTA) results in BioNER are achieved using transformer-based biomedical pretrained language models (BPLMs) that can be fine-tuned to perform many NLP tasks, including NER (Kalyan et al., 2022). However, these models rely on extremely large datasets to properly learn the NER task, and creating these NER datasets is specifically challenging due to the precise nature and evolving dynamics of the biomedical field.

Currently, gold-standard BioNER datasets are created via a slow, expensive processes that rely on human experts for manual data annotation, which are difficult to scale. Such expert-annotated datasets struggle to match the variability in biomedical terminology, the complexity of biomedical concepts, and the pace of growth in biomedical literature (Dai et al., 2020). Consequently, the quantity and quality of existing BioNER datasets limit the performance of SOTA BioNER models.

To overcome the limitations imposed by the sparse and specialized nature of BioNER datasets, researchers employed deep-learning techniques to train models for the BioNER task but none of these techniques are as effective as simply training models on more data.

For example, multi-task learning addresses limited training data and allows models to learn several related tasks simultaneously by finding similarities in different small datasets. However, models trained with multi-task learning have low precision scores and inconsistent performance across diverse datasets (Khan et al., 2020; Wang et al., 2019).

Other approaches, such as intermediate fine-tuning (where a pre-trained model's parameters are adjusted by training it on new data related to a specific task) have similar drawbacks of the model not generalizing to perform well on different datasets (Kalyan et al., 2022).

An approach to quickly and efficiently label large, high-quality datasets for BioNER would greatly benefit the field by removing the limitations of sparse data altogether. This paper is a

preliminary exploration of how well large language models (LLMs) might be up to the task.

GPT-3.5 and similar LLMs have shown remarkable results for their few-shot and zero-shot capabilities on many NLP tasks, including data annotation and label error handling (Chong et al., 2022; Chen et al., 2023; Chung et al., 2022). Recent attempts to use LLMs for automatic data label generation have shown promising results (Brown et al., 2020). However, this approach has never been applied to the biomedical or NER context as far as we know.

This paper addresses key limitations of expensive, slowly-produced human-annotated datasets by introducing a novel approach that uses OpenAI's GPT-3.5 to cheaply and automatically annotate BioNER training datasets. We use zero-shot, one-shot, and few-shot learning with GPT-3.5 models to recreate SOTA BioNER datasets. We validate our results both intrinsically and extrinsically. Intrinsically, we compare our datasets with the original, expert human-labeled dataset. Extrinsically we compare the performance of two SOTA BPLMs (BioLinkBERT and BioGPT), finetuned on our datasets versus the expert-labeled dataset.

We were able to achieve promising results despite major limitations in time and finances. Intrinsically, many of our datasets' labels are comparable to human labels. Extrinsically, our results suggest that the BioNER performance of BPLMs fine-tuned on LLM-labeled datasets may also be comparable with the performance of BPLMs fine-tuned on human-annotated datasets. With further research, it's possible that GPT-3.5-generated datasets will offer an automated solution for labeling biomedical entities (diseases, chemicals, genes, proteins) in texts, mitigating the bottleneck of human annotations amid the rapidly expanding volume of biomedical literature.

2 Related Work

There is significant existing work toward improving datasets for BioNER, but none to our knowledge on label automation. Previous synthetic data labeling methods were not robust enough to motivate their use in complex biomedical domains. However, recent work outside of the biomedical field yields promising results for automatic labeling.

2.1 State of the Art BioNER

Current SOTA approaches to the BioNER task feature a two-stage pre-training and fine-tuning paradigm. Pre-training takes advantage of large swathes of unlabeled biomedical corpora (during pre-training) to achieve stronger performance on biomedical NLP tasks after fine-tuning.

There are two general approaches to pre-training.

1. *Mixed-Domain Pre-Training:* SOTA models like BioBERT (Lee et al., 2019), ClinicalBERT (Alsentzer et al., 2019), and BlueBERT (Peng et al., 2019) are initialized with weights from a general LLM trained on non-specific text (e.g. BERT), and they then undergo self-supervised learning on unlabeled biomedical text. However, initial exposure to general text causes model tokens to be unrepresentative of the target biomedical domain (e.g. common medical terms like "naxalone" might be split into several tokens by BPLMs initially exposed to general text).(Gu et al., 2021a).

2. *Domain-Specific Pretraining from Scratch:* More recently, SOTA models like PubMedBERT are pre-trained without any exposure to general text and only using biomedical text (Gu et al., 2020). This has been shown to overcome vocabulary limitations in the Mixed-Domain approach (Gu et al., 2021a).

Fine-tuning involves training the BPLM on a smaller, task-specific dataset to help the model learn valuable features that can be generalized to complete tasks like BioNER (Wang et al., 2023).

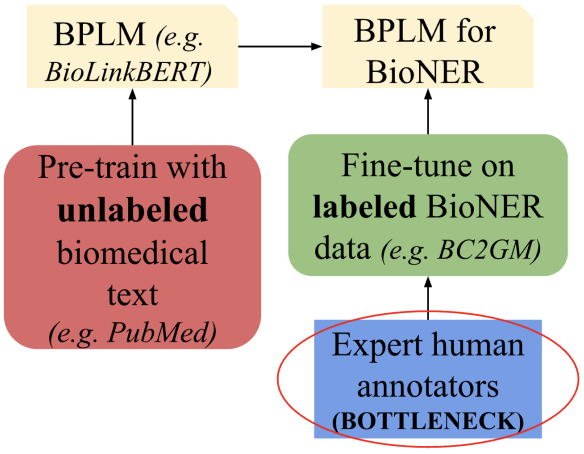

Figure 1: Current Approach to SOTA BioNER

Biomedical NLP benchmarks evaluate BPLMs' performance on a variety of NLP tasks (including BioNER) after fine-tuning on specialized datasets. The current most comprehensive biomedical NLP

benchmark is BLURB (Biomedical Language Understanding and Reasoning Benchmark) which we used to source our gold-standard human-labeled datasets.

BioLinkBERT currently achieves the most competitive BioNER scores on BLURB. It achieves its high performance by leveraging a unique pretraining process that takes advantage of links between documents (e.g. hyperlinks) (Yasunaga et al., 2022). BioGPT (Luo et al., 2022) is a recent BPLM with promising performance in many biomedical NLP tasks that has not yet been evaluated on BLURB.

Our experiments automate the labeling of BioNER-specific datasets found in BLURB, and we validate our results by fine-tuning a mixed-domain (BioGPT) and a domain-specific (BioLinkBERT-Base) BPLM on them. We selected BioLinkBERT-Base because it's the highest performing smaller model on the BLURB leaderboard (which aligned with our constraints), and we selected BioGPT because its lower BioNER performance and unique architecture provided greater diversity in our extrinsic validation.

2.2 Toward Improving BioNER Datasets

Low-quality datasets, such as those that have biased class distributions and name regularities, negatively impact BioNER models' performance on biomedical NLP benchmarks. Many approaches have been designed to resolve this, all of which are applied to previously-labeled data. This is because it is currently much easier to try to curate existing datasets than to add to them or produce entirely new ones.

One statistics-based debiasing method involves temperature scaling to smooth biased model probabilities. Testing their debiased BioBERT model on datasets with rare diseases revealed a 20% improvement in generalization to rare patterns. However, this debiasing also led to the model making unusual predictions, such as identifying text spans, adjectives, and parentheses as entities, because it no longer overtrusted class distributions and surface forms of mentions in the training dataset (Kim and Kang, 2022).

Another way to improve the quality of labeled datasets is data augmentation through back translation, which involves translating from a language to a pivot language and then translating back to the original language. Despite this counter-intuitive approach to producing different data, (Wang et al., 2020) observed that back translation increased linguistic diversity and expanded the training dataset such that models trained on small datasets are more robust and less prone to overfitting for Semantic Textual Similarity (STS) task.

Dataset debiasing often both has adverse affects and involves undesirable methods to "produce" new data. Such complicated approaches could be simplified or done away with if it were cheaper and easier to produce new labeled biomedical data. Our pipeline for labeling automation could be used as a supplement or replacement for many existing approaches to improving BioNER datasets.

2.3 Toward Automating Data Labeling

Large language models such as GPT-3.5 have remarkable few-shot learning capabilities, which have been employed toward data labeling for downstream NLP tasks (Brown et al., 2020). Traditionally, experts manage crowd-sourced annotation by defining tasks and giving examples. Recent studies have shown that prompting LLMs can produce annotations similarly.

He et al. introduced AnnoLLM which uses GPT-3.5 and a few-shot chain-of-thought (CoT) prompting method to create annotations, achieving accuracy comparable to crowd-sourcing in user query and keyword relevance, word sense disambiguation, and question-answering (He et al., 2023). Wang et al. instead used GPT-3 to both label data and assign its labels an uncertainty score (followed by human review and re-labeling of those above an uncertainty threshold) (Wang et al., 2021). This method also yielded results on par with traditional datasets.

Our research stands out as it is the first, to our knowledge, to investigate LLMs in annotating NER and biology-specific data. We leveraged these approaches and expand off He et al. and Wang et al. to create a pipeline for annotating BioNER data using various prompting techniques with LLMs.

3 Methods

To analyze LLMs' BioNER dataset labeling abilities, we implemented an automated label generation pipeline. We passed in untagged biomedical text, along with various prompts, to a GPT-3.5 model to generate an NER tagging of the text. We generated datasets using a zero-shot, one-shot, and few-shot prompting for five BioNER datasets in the BLURB benchmark. After generating our

datasets, we performed both an intrinsic evaluation and extrinsic evaluation. In the intrinsic evaluation, we compared the GPT-labeled dataset to the original human-labeled dataset. In the extrinsic evaluation, we compared the performance of two different BPLMs' fine-tunined on the human-labeled test set and the the GPT-labeled datasets, respectively. All of the code for label generation and evaluation is found this GitHub repo.

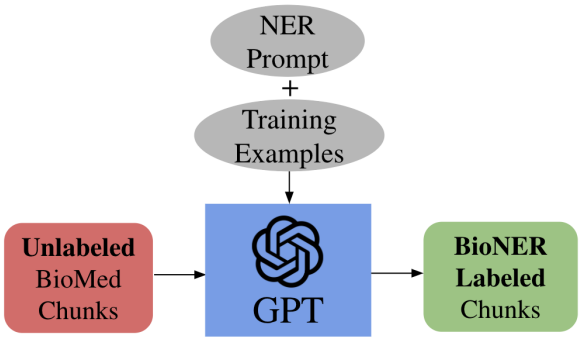

Figure 2: Automated Label Generation Pipeline

3.1 Datasets for BioNER Label Generation

We've selected the BioNER datasets used in the current most comprehensive biomedical NLP benchmark, BLURB (Gu et al., 2021b). These, outlined in Table 9, include NCBI Disease (Doğan et al., 2014), the BioCreative V CDR task corpus (Li et al., 2016), JNLPBA (Huang et al., 2019), and the BioCreative II gene mention recognition corpus (Smith et al., 2008).

Each corpus focuses on a specific type of biomedical entity and is labeled by human experts using the standard "BIO" labeling schema: every token is labeled "B" if it is the beginning of a relevant biomedical entity, "I" if it is in the middle of a multi-token biomedical entity, or "O" if it is not part of a relevant biomedical entity.

Figure 3: "BIO" labeling schema example. Green = "O", Orange = "B", Purple = "I"

Each of the five corpuses label only certain types of biomedical entities which are shown in Table 9. Together, they cover a comprehensive range of biomedical entity types, providing a comprehensive basis for testing our pipeline. For each original corpus, we used the unlabeled text in the "devel" split to apply our automatic labeling pipeline, and we used the labeled test splits in the "test" split for our extrinsic evaluation.

Dataset	Entity	Devel	Test
BC5DR-chem	chemical	122.0k	129.5k
BC5DR-disease	disease	122.0k	129.5k
NCBI-disease	disease	24.9k	25.4k
BC2GM	genes	73.6k	148.5k
JNLPBA	protein	121.1k	118.6k

Table 1: Name, featured biomedical entity type, and number of tokens in the original devel and test sets of NER Datasets used in the BLURB Biomedical NLP Benchmark. BLURB is the current most comprehensive biomedical NLP benchmark.

3.2 Prompt Engineering for BioNER Label Generation

Our prompting strategy with GPT-3.5 involved detailed instructions for NER and required output formatting. We set the model's temperature to 0 for output consistency and experimented with zero-shot, one-shot, and few-shot learning prompts. These prompts varied in the number of training examples provided, ranging from zero to three, tailored to optimize performance without excessive computational cost.

Providing more context on how to effectively label tokens improved the model's annotation capabilities. Therefore, our prompts included precise directions on what entity type to label, according to the BIO labeling schema (Hong et al., 2020), and that the output must be formatted in JSON (such that each key and value pair corresponds to a unique token and corresponding label from the inputted text). To resolve identical tokens, we specified each output token = the original token + "_" + its index in the text. For more details on what factors we took into consideration during the prompt engineering process, refer to subsection A.3.

Each training example for one- and few- shot prompts featured sample texts with approximately 300 tokens and labels taken from the training set of the corresponding dataset in the required output format.

For one-shot prompts, we empirically picked training examples with a relatively high ratio of labeled entities. This metric was chosen intuitively because our financial constraints limited our ability to experiment.

For few-shot learning, we chose three example sentences using three metrics: 1) a sentence with a high number of entities, 2) a sentence with relatively long entities, and 3) a sentence with a relatively low number of entities that better represents the entire dataset's entity ratio. This method aimed to improve GPT's accuracy in recognizing and labeling specific entity types without mislabeling unrelated tokens.

We concluded our prompt by appending a line indicating the text we want the LLM to label. Examples of prompts used in zero-shot, one-shot and few-shot prompting can be viewed in subsection A.2.

3.3 Generating BioNER Labels with GPT-3.5

Utilizing OpenAI's GPT-3.5 API, we segmented human-annotated datasets into sentences, using period tokens as delimiters. We grouped these sentences into approximately 300 token chunks– optimizing the number of tokens per API request while avoiding RateLimitErrors.

While sending in our prompts, the API typically responded within 10 seconds, though some requests took up to 10 minutes. To enhance efficiency, we made the calls multi-threaded, sending 10 parallel requests. We encountered occasional timeout errors, leading to temporary closed connections. To mitigate data loss and redundant requests, we recorded each response in a CSV file in real-time. We found that API performance was highly variable on both the time of day and the OpenAI account. Fully labeling larger datasets (100,000 tokens) sometimes took 30 minutes and other time tooks about 8 hours.

A detailed depiction of the entire automatic labeling pipeline can be viewed in subsection A.1.

4 Evaluating Label Accuracy and Quality

To evaluate the abilities of GPT-3.5 as an annotator of BioNER datasets, we analyzed the quality of the generated dataset labels from each experiment both intrinsically and extrinsically.

4.1 Intrinsic Evaluation

Intrinsic evaluation consisted of comparing the LLM-generated labels to the human expert labels for each of the datasets and deriving classification metrics. This sought to answer both how effectively our prompt generated valid outputds from GPT-3.5 and how differently GPT-3.5 performs compared to human experts at labeling.

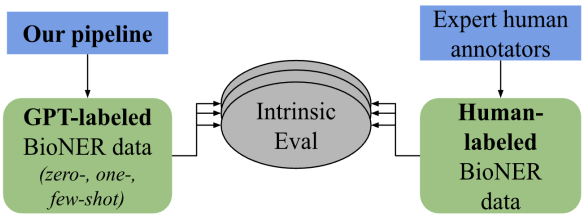

Figure 4: Intrinsic Evaluation of GPT-generated labels

4.1.1 Output Validity

We first measured and corrected syntactical mistakes made by GPT-3.5 in the labeling process. GPT-3.5 occasionally forgot to list some tokens in its JSON output. It also hallucinated tokens, meaning it returned token and label pairings that were not present in the inputted text. As a result, when we concatenated all of the labeled chunks, the tokens in our LLM-generated dataset were slightly different from the tokens in the original human annotated dataset.

In order to compute classification metrics determining how similar the two datasets were, we needed the tokens in both datasets to match exactly.

Therefore, we determined the indices of all of the missed tokens (tokens that GPT-3.5 did not label) in the original dataset, and all indices of the hallucinated tokens (tokens that GPT-3.5 labeled that did not exist) in the GPT-3.5 generated dataset via a brute force approach. We counted then removed the missing token and label pairings from the original human-labeled dataset and removed the hallucinated token and label pairings from the GPT-3.5 generated dataset.

Once the tokens in the augmented datasets lined up exactly, we performed the rest of our analysis to determine how similar GPT-3.5 labeled entities were to human-labeled entities.

4.1.2 Entity Recognition

We measured GPT-3.5's ability to both exactly and approximately match human-labeled entities. Given the human annotations had labeling inconsistencies and entities in the biomedical domain are referred to by numerous names, approximate classification metrics give deeper insight into how well GPT-3.5 identified biomedical entities.

We used the seqeval.metrics library in Python to determine the Precision, Recall and F1-score of LLM-generated labels with human labels as ground-truth. Exact entity matches required straight-forward use of seqeval.metrics.

To evaluate approximate entity matches, we first developed a definition: any GPT-labeled entity that shares a token with a human-labeled entity is an approximate match. Intuitively, if the human labeled dataset identified 'Ebola virus disease' as an entity and GPT-3.5 only identified 'Ebola' as the entity, then that would be an approximate entity match.

We created an augmented dataset where all approximate entity matches were now exact matches then applied the same method we used to get exact entity classification metrics.

4.2 Extrinsic Evaluation

Our extrinsic evaluation determined how GPT-3.5 generated labeled datasets impacted BPLMs performance on BioNER (relative to human-labeled datasets). We tested how the performance of current SOTA BPLMs BioLinkBERT (Yasunaga et al., 2022) and BioGPT (Luo et al., 2022) fine-tuned on GPT-3.5 labeled datasets compared to BPLMs fine-tuned on human-labeled dataset when evaluated on human-labeled BioNER test sets.

We implemented fine-tuning using the transformers library from Hugging Face, where both BioLinkBERT and BioGPT were easily accessible. We used cross-entropy loss and an Adam optimizer initialized with the default learning rate. This was run on the T4 GPU accessible via Google Colab.

We separately performed this procedure for every human-labeled and GPT-3.5 zero, one, and few-shot labeled dataset. We then evaluated BioLinkBERT and BioGPT's performances after fine-tuning on the GPT-labeled against their performances when fine-tuned on the human-labeled test sets. The repository of our fine-tuned BioNER models can be found on Hugging Face: https://huggingface.co/68611-llm-annotation-group.

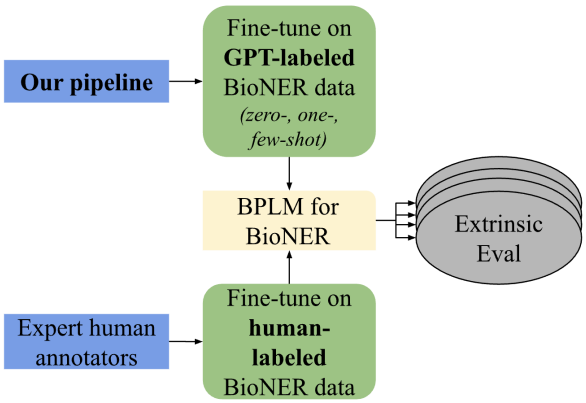

Figure 5: Extrinsic Evaluation of GPT-generated labels

5 Results

Our research reveals that LLM-labeled data, while it does not outperform human-labeled data, shows promising results in both intrinsic and extrinsic evaluations. These findings are significant in spite of our constrained resources and short timeline, suggesting that LLMs have great potential to rapidly create high-quality BioNER datasets.

5.1 Intrinsic Results

5.1.1 Output Validity

We measured two types of output validity: missing tokens (human-labeled tokens not featured in the LLM's output) and hallucinated tokens (tokens featured in the LLM's output that weren't in the original prompt).

% of True Tokens Missing from the GPT-labeled datasets

Dataset	Zero-Shot	One-Shot	Few-Shot
BC5DR-C	0.63	0.63	**0.47**
BC5DR-D	0.61	**0.51**	0.65
NCBI	0.38	**0.24**	0.27
BC2GM	0.62	**0.17**	0.48
JNLPBA	0.55	**0.36**	0.56
Average	0.56	**0.38**	0.49

Table 2: Percent of missed tokens in GPT's zero, one, and few-shot outputs. Prompts with the fewest missed tokens are bolded for each dataset.

Our findings indicate that, across all datasets, the proportion of missing tokens was less than 1%, with one-shot prompts yielding the lowest rates. This suggests a high level of accuracy in capturing relevant data.

Conversely, hallucinated tokens, although all under 1%, increased with prompt complexity, indicating a trade-off between prompt detail and output purity.

5.1.2 Entity Recognition

We measured GPT-3.5's F1, precision, and recall on exactly labeled human-entities. We also measured its F1, precision, and recall on the more lenient "approximately" labeled human-entities as described in subsubsection 4.1.1.

Our study demonstrates a distinct variance in the LLM's ability to recognize different entity types. GPT-3.5 demonstrated high competency in identifying chemical entities, although it struggled with disease names.

% of Outputted Tokens Hallucinated in the GPT-labeled datasets

Dataset	Zero-Shot	One-Shot	Few-Shot
BC5DR-C	**0.21**	0.39	0.31
BC5DR-D	**0.14**	0.32	0.35
NCBI	**0.02**	0.07	0.07
BC2GM	0.25	**0.17**	0.34
JNLPBA	0.16	**0.11**	0.15
Average	**0.16**	0.21	0.24

Table 3: Percent of hallucinated tokens in GPT's zero, one, and few-shot outputs. Prompts with the fewest hallucinated tokens are bolded for each dataset.

Across all datasets and prompting techniques, our LLM-labeled data had significantly higher recall than precision. This mismatched ratio might be improved by introducing a pseudo-confidence threshold into our prompts. This might be as straightforward as adding to the prompt that "false positive labels are slightly worse than false negatives".

Higher precision scores were earned by few- and one-shot prompts across the board, indicating that including training examples in the prompt helps GPT-3.5 avoid false positives.

Exact Entity **F1(Recall, Precision)** Against Human-Labeled Entity Ground Truth:

Dataset	Zero-Shot	One-Shot	Few-Shot
BC5DR-C	60 (73,52)	**62** (81,50)	54 (64,47)
BC5DR-D	38 (53,29)	**39** (51,32)	18 (18,17)
NCBI	30 (46,22)	**35** (59,25)	33 (42,27)
BC2GM	42 (48,38)	42 (47,39)	**45** (47,44)
JNLPBA	40 (48,34)	**44** (50,40)	41 (38,45)

Table 4: Similarity of entities (measured exactly) labeled by GPT-3.5 to human-labeled entities for each prompt and dataset.

Higher exact entity recall scores were earned by one-shot prompts. We hypothesize the choice of examples in few-shot prompts might have inadvertently led to overfitting, impacting the performance. This suggests the need for more strategic selection of training examples in future research.

Approximate Entity **F1(Recall, Precision)** Against Human-Labeled Entity Ground Truth:

Dataset	Zero-Shot	One-Shot	Few-Shot
BC5DR-C	**72** (90,60)	70 (94,56)	64 (78,54)
BC5DR-D	**52** (76,40)	**52** (70,41)	27 (29,26)
NCBI	49 (76,37)	**52** (89,37)	50 (65,41)
BC2GM	67 (77,59)	67 (76,60)	**69** (74,65)
JNLPBA	**65** (77,57)	**65** (74,59)	61 (57,66)

Table 5: Similarity of entities (measured approximately s.t. any human-labeld entity with any correctly identified token is "correct") labeled by GPT-3.5 to human-labeled entities for each prompt and dataset.

Surprisingly, zero-shot prompts were slightly better than one-shot prompts for approximate entities. From this, we infer that training examples in the prompt help GPT-3.5 more precisely label full entities, but they also make GPT-3.5 significantly less likely to identify entities at all. Since our training examples were chosen empirically, we suspect this trend might be caused by sub-optimal example selection.

5.2 Extrinsic Results

The extrinsic evaluation, using SOTA BPLMs, further corroborates our intrinsic findings. While the LLM-labeled datasets did not surpass the performance of human-labeled data, they showed reasonable quality. Our results showed particular promise in recall metrics.

BioLinkBERT-Base **F1(Recall, Precision)** scores after fine-tuning on:

Dataset	Human	Zero-Shot	One-Shot	Few-Shot
BC5DR-C	**95** (97,93)	**83** (96,73)	81 (97,79)	75 (91,63)
BC5DR-D	**81** (90,74)	35 (67,24)	**53** (63,46)	52 (70,42)
NCBI	**81** (80,81)	36 (73,24)	40 (83,26)	**42** (56,33)
BC2GM	**85** (89,81)	65 (91,50)	68 (79,60)	**71** (82,63)
JNLPBA	**81** (87,75)	60 (87,46)	**62** (78,52)	**62** (64,59)

Table 6: Performance of BioLinkBERT-Base fine-tuned separately on each dataset. Overall high F1 is bolded (precision and recall are underlined), and highest non-control score is bolded (underlined for precision and recall) if distinct. Values are 100x F1 (Recall, Precision).

The gap between human and LLM-labeled data was narrower using the BioGPT model, so the utility of LLM-generated labels appears to be model-dependent. The gap may also have been smaller because BioGPT performed worse overall on the human-labeled datasets.

Performance also varied based on the entity type,

with chemical entities generally yielding closer results to human-labeled data. This contributes to the intrinsic evidence that GPT-3.5's ability to label BioNER data is affected by the entity of interest.

Interestingly, few-shot prompts, despite their mixed intrinsic performance, showed notable success in extrinsic evaluations. This indicates the value of increased prompt context, even without perfectly chosen examples.

BioGPT **F1(Recall, Precision)** scores after fine-tuning on:

Dataset	Human	Zero-Shot	One-Shot	Few-Shot
BC5DR-C	**90** (90,90)	29 (90,68)	**78** (93,68)	76 (82,71)
BC5DR-D	**71** (72,73)	39 (55,30)	41 (48,36)	**45** (51,40)
NCBI	**66** (66,67)	**29** (47,21)	27 (52,18)	**29** (41,22)
BC2GM	**81** (77,84)	66 (78,58)	68 (75,62)	**70** (69,72)
JNLPBA	**74** (84,67)	61 (80,49)	**63** (72,57)	49 (39,64)

Table 7: Performance of BioGPT fine-tuned separately on each dataset. Overall high F1 is bolded (precision and recall are underlined), and highest non-control score is bolded (underlined for precision and recall) if distinct. Values are 100x F1 (Recall, Precision).

6 Discussion

Our experiments show promise that automatically labeled data by LLMs could be practically used as a replacement or supplement to human-annotated dataset. Our preliminary results demonstrate the non-trivial ability for GPT-3.5 to, provided unlabeled data, match human BioNER labels and produce a high-quality dataset for BPLM fine-tuning.

Because it will always be costly and difficult to use human annotators for biomedical topics, our promising results (despite the limitations in our study) suggest deeper study into LLM-labeled datasets for BioNER (and other biomedical NLP tasks) could be incredibly fruitful for the field.

6.1 Cost Analysis

The cost of calling the API for the GPT-3.5-turbo-1106 model is 0.1¢ per 1K input tokens and 0.2¢ per 1K output tokens, as seen here. Our costs scale linearly with the size of the dataset; however, as the prompts increase in size, the average number of trials needed to generate proper labels increased (i.e. when moving from zero-shot to few-shot prompting). Intuitively, few-shot prompt labels should have cost more than zero-shot labels, however, there was no clear linear relationship due to the following two conflicting factors: LLMs provide better labels more consistently when provided few-shot prompting but our servers crashed more often because larger context required more server bandwidth, and LLMs needed more trials to output proper labels during zero-shot prompting and thus more API calls were used.

Avg # Input Tokens per Call	240
Avg # Output Tokens per Call	305
Avg Call Cost	9.5¢
Avg Cost per BLURB dataset	$84.73

Table 8: Breaking down the average costs of labeling BLURB datasets.

Our costs of labeling the datasets do not align with the prices provided by OpenAI, because our code for generating labels gives GPT a few trials to output proper BioNER labels, to mitigate the number of errors and hallucinations in our final label outputs. This, in addition to having our servers repeatedly crash and restart the API calls, significantly drove up costs.

In spite of these confounding factors, the cost of outsourcing human labor and third-party data labeling services still far exceeds the cost of using LLMs for label generation. High-quality data labeling services for NER tasks average **70¢ / unit of text**, the equivalent of 66 tokens, as seen here. By comparison, our costs average **2.6¢ / unit of text**, which is nearly 30x cheapter than the status quo.

6.2 Directions for Future Research

Due to limitations in time and financial resources, there were many directions we hoped to explore further, which we believe could bolster the strength of our results. The following are a few directions we plan to experiment with next.

1. **Fine-tuning GPT-3.5-Turbo:** OpenAI recently released the option to instruction fine-tune GPT models on pre-established prompts and responses. This involves formatting our training examples as JSON user prompts and model responses and fine tuning GPT-3.5-turbo on 50-100 examples before deployment. This allows for a way to provide multiple training examples without over populating each prompt. Given we have seen that additional training examples can assist in the labeling process, but longer prompts can also negatively affect results, we suspect this solution would yield very strong results. The bottleneck is that the instruction fine-tuning process

and deployment costs of fine-tuned models is very expensive compared to using the baseline GPT-3.5-turbo model. We were limited in our financial resources to experiment with this option, but have already written the scripts to perform these experiments, which we hope to perform once we receive further funding.

2. **Further Prompt Engineering:** Our experiments regarding optimal prompt instructions and output format were limited to only a few training examples, given we did not have the resources to generate multiple labeled datasets with different prompting techniques. However, at a larger scale other techniques could have yielded better results, and we hope to experiment with more prompt designs going forward.

3. **More Intelligent Dataset Chunking:** Our method of splitting biomedical texts into smaller chunks to send to GPT-3.5-turbo involved trivially splitting the text into chunks of approximately 300 tokens separated by end of sentence tokens. Using periods as end of sentence tokens negatively impacted accuracy, because we did not consider that periods would also split decimal numbers. Given the simplicity of this method, we likely lost a lot of context in our chunks. Given context is very crucial for effective NER labeling in any domain, we hope to see that if more context sensitive chunking techniques could improve the quality of our labeling results.

Research Impact

Our research on automating BioNER data labeling using LLMs demonstrates that there is great potential for developing systems that will rapidly increase the quantity of biomedical NLP datasets. This is particularly important in light of the ever-growing corpus of unlabeled medical literature. While our research focused on Named Entity Recognition in the biomedical sphere, it may easily be extrapolated to other NLP tasks (i.e. classification, summarization, question/answering), as well as other domains (i.e. legal and financial domains) with a growing wealth of data that is ripe for analysis.

The introduction of synthetic NLP data labels will likely have complex downstream effects on tasks that these datasets are used for in biology research, including: medical record analysis, drug discovery and development, sentiment analysis in patient feedback, and more. However, the rise of LLM-generated datasets introduces the following risks:

1. **Automation bias:** over-trusting models to automate away dataset labeling work. Lower quality data: We have not yet found a method for determining if the labeling performance of ML models will ever match or exceed the ability of humans to label data. If they cannot at least match the accuracy of expert human annotations, then we risk lowering the overall accuracy and quality of biomedical datasets. This is especially important to take into account, because many biomedical datasets may have direct impacts on health outcomes.

2. **Biased synthetic data:** LLMs are pre-trained on text corpuses that heavily involve certain demographics and exclude others. Before automating away data-labeling, it's important to develop a system that can check for biases inherent in LLM-generated labels.

3. **Skewing biomedical research towards English-speaking nations:** Because most LLMs, as of the present, are primarily pre-trained on English text corpuses, most of the synthetic data labels will also be for English biomedical texts. This may incentivize more NLP research in healthcare texts only in North America and English-speaking nations, and direct resources away from biomedical research in other regions.

4. **Generation of synthetic datasets with malicious intent:** Until we develop a robust detection system which will be able to distinguish between LLM-generated labels and human-generated labels, we must entrust those who are developing these datasets to manually note down which labels are which. If datasets are being auto-labeled with malintent and not caught, there may be rippling negative repercussions in the field of biomedical research, which directly impacts patient health outcomes.

For these reasons, we encourage further research to understand the biases and limitations of our proposed LLM data-labeling system. We suggest that researchers investigate the following questions:

1. Will synthetic dataset labels ever approach or exceed the quality of human-generated dataset labels?

2. How can we develop safe and robust detection systems to distinguish between synthetic and human labels?

3. How may we develop pipelines and methods for to debias LLM-generated data labels?

4. How much money and time will be saved via the automation of data labeling?

5. If data cannot be entirely auto-labeled by LLMs, how can we develop a partial automated, partial human-in-the-loop system that is more efficient than the status quo?

References

Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. 2019. Publicly available clinical bert embeddings.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, et al. 2023. Alpagasus: Training a better alpaca with fewer data. *arXiv preprint arXiv:2307.08701*.

Derek Chong, Jenny Hong, and Christopher D Manning. 2022. Detecting label errors using pre-trained language models. *arXiv preprint arXiv:2205.12702*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Xiang Dai, Sarvnaz Karimi, Ben Hachey, and Cecile Paris. 2020. An effective transition-based model for discontinuous NER. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5860–5870, Online. Association for Computational Linguistics.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Special report: Ncbi disease corpus: A resource for disease name recognition and concept normalization. *J. of Biomedical Informatics*, 47:1–10.

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. *CoRR*, abs/2007.15779.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021a. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021b. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Xingwei He, Zhenghao Lin, Yeyun Gong, Alex Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, Weizhu Chen, et al. 2023. Annollm: Making large language models to be better crowdsourced annotators. *arXiv preprint arXiv:2303.16854*.

Zhi Hong, Roselyne Tchoua, Kyle Chard, and Ian Foster. 2020. Sciner: extracting named entities from scientific literature. In *Computational Science–ICCS 2020: 20th International Conference, Amsterdam, The Netherlands, June 3–5, 2020, Proceedings, Part II 20*, pages 308–321. Springer.

Ming-Siang Huang, Po-Ting Lai, Pei-Yen Lin, Yu-Ting You, Richard Tzong-Han Tsai, and Wen-Lian Hsu. 2020. Biomedical named entity recognition and linking datasets: survey and our recent development. *Briefings in Bioinformatics*, 21(6):2219–2238. _eprint: https://academic.oup.com/bib/article-pdf/21/6/2219/34671873/bbaa054.pdf.

Ming-Siang Huang, Po-Ting Lai, Richard Tzong-Han Tsai, and Wen-Lian Hsu. 2019. Revised JNLPBA corpus: A revised version of biomedical NER corpus for relation extraction task. *CoRR*, abs/1901.10219.

Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and Sivanesan Sangeetha. 2022. Ammu: A survey of transformer-based biomedical pretrained language models. *Journal of Biomedical Informatics*, 126:103982.

Muhammad Raza Khan, Morteza Ziyadi, and Mohamed Abdelhady. 2020. Mt-bioner: Multi-task learning for biomedical named entity recognition using deep bidirectional transformers. *CoRR*, abs/2001.08904.

Hyunjae Kim and Jaewoo Kang. 2022. How do your biomedical named entity recognition models generalize to novel entities? *IEEE Access*, 10:31513–31523.

Esther Landhuis. 2016. Scientific literature: Information overload. *Nature*, 535(7612):457–458.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets.

Larry L. Smith, Lorraine K. Tanabe, Rie Ando, Cheng-Ju Kuo, I-Fang Chung, Chun-Nan Hsu, Yu-Shi Lin, Roman Klinger, C. Friedrich, Kuzman Ganchev, Manabu Torii, Hongfang Liu, Barry Haddow, Craig A. Struble, Richard J. Povinelli, Andreas Vlachos, William A. Baumgartner, Lawrence E. Hunter, Bob Carpenter, Richard Tzong-Han Tsai, Hong-Jie Dai, Feng Liu, Yifei Chen, Chengjie Sun, Sophia Katrenko, Pieter W. Adriaans, Christian Blaschke, Rafael Torres, Mariana L. Neves, Preslav Nakov, Anna Divoli, Manuel Maña-López, Jacinto Mata, and W. John Wilbur. 2008. Overview of biocreative ii gene mention recognition. *Genome Biology*, 9:S2 – S2.

Benyou Wang, Qianqian Xie, Jiahuan Pei, Zhihong Chen, Prayag Tiwari, Zhao Li, and Jie Fu. 2023. Pre-trained language models in biomedical domain: A systematic survey. *ACM Comput. Surv.*, 56(3).

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? gpt-3 can help. *arXiv preprint arXiv:2108.13487*.

Xi Wang, Jiagao Lyu, Li Dong, and Ke Xu. 2019. Multitask learning for biomedical named entity recognition with cross-sharing structure. *BMC Bioinformatics*, 20(1).

Yuxia Wang, Fei Liu, Karin Verspoor, and Timothy Baldwin. 2020. Evaluating the utility of model configurations and data augmentation on clinical semantic textual similarity. In *Proceedings of the 19th SIGBioMed workshop on biomedical language processing*, pages 105–111.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. Linkbert: Pretraining language models with document links. *arXiv preprint arXiv:2203.15827*.

A Appendix

A.1 Automatic BioNER Labeling Pipeline

The following is a more in depth overview of the entire automatic labeling pipeline starting from the unlabeled BioMed corpus all the way to the fully labeled BioMed Corpus.

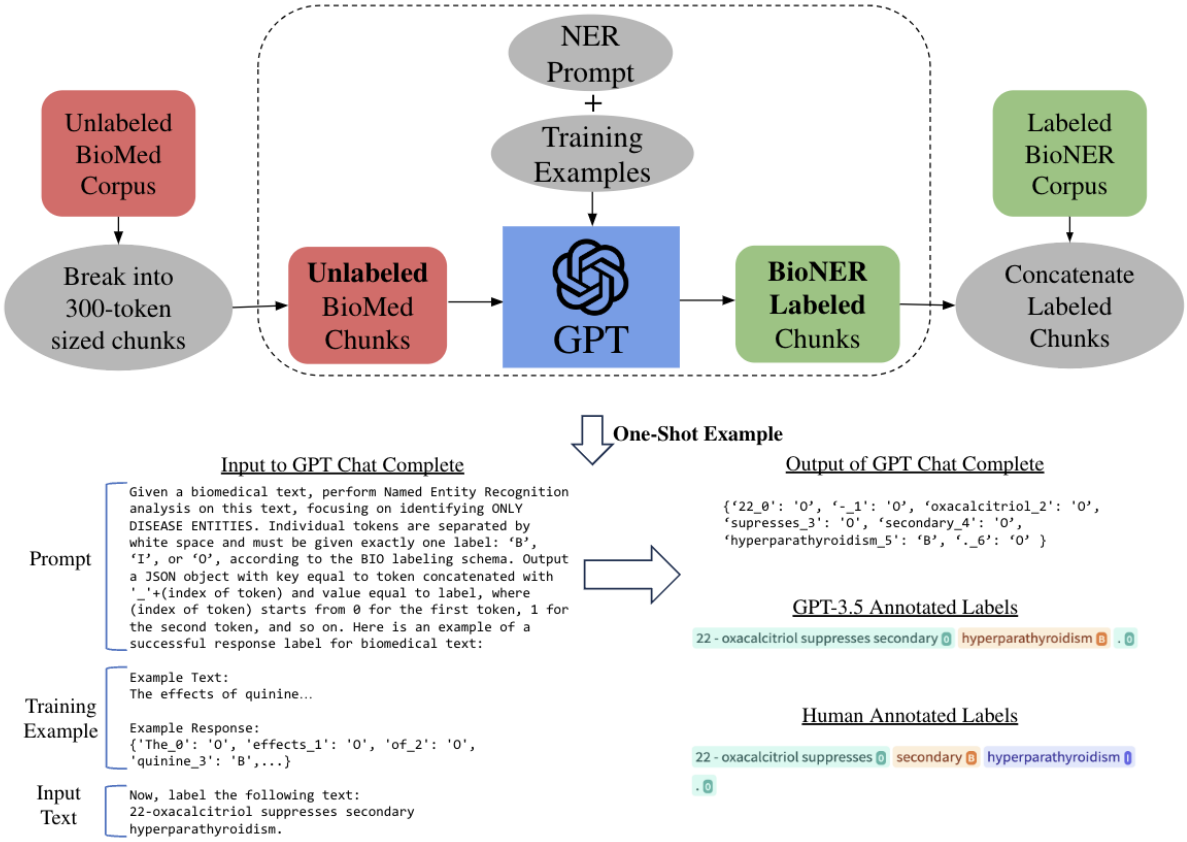

Figure 6: Complete overview of autmated BioNER data labeling pipeline

A.2 Example Prompts

The following are specific examples of zero-, one- and few-shot prompts used in our automatic label generation process. These prompts are taken from the label generation algorithm for BC5CDR-chem.

1. **Zero-Shot Prompt:**

```
Given a biomedical text, perform Named Entity Recognition analysis on this
text, focusing on identifying ONLY CHEMICAL ENTITIES. Individual tokens are
separated by white space and must be given exactly one label: 'B', 'I', or 'O',
according to the BIO labeling schema. Output a JSON object with key equal to
token concatenated with '_'+(index of token) and value equal to label, where
(index of token) starts from 0 for the first token, 1 for the second token, and
so on.

Text: 22 - oxacalcitriol suppresses secondary hyperparathyroidism without
inducing low bone turnover in dogs with renal failure . BACKGROUND : Calcitriol
therapy suppresses serum levels of parathyroid hormone ( PTH ) in patients with
renal failure but has several drawbacks , including hypercalcemia and / or
marked suppression of bone turnover , which may lead to adynamic bone disease
.
```

2. **One-Shot Prompt:**

Given a biomedical text, perform Named Entity Recognition analysis on this text, focusing on identifying ONLY CHEMICAL ENTITIES. Individual tokens are separated by white space and must be given exactly one label: 'B', 'I', or 'O', according to the BIO labeling schema. Output a JSON object with key equal to token concatenated with '_'+(index of token) and value equal to label, where (index of token) starts from 0 for the first token, 1 for the second token, and so on. Here is an example of a successful response label for a biomedical text:

Example Text 1: The effects of quinine and 4 - aminopyridine on conditioned place preference and changes in motor activity induced by morphine in rats . 1 . The effects of two unselective potassium (K (+) -) channel blockers , quinine (12 . 5 , 25 and 50 mg / kg) and 4 - aminopyridine (1 and 2 mg / kg) , on conditioned place preference and biphasic changes in motor activity induced by morphine (10 mg / kg) were tested in Wistar rats .

Example Response 1:{'The_0': 'O', 'effects_1': 'O', 'of_2': 'O', 'quinine_3': 'B', 'and_4': 'O', '4_5': 'B', '-_6': 'I', 'aminopyridine_7': 'I', 'on_8': 'O', 'conditioned_9': 'O', 'place_10': 'O', 'preference_11': 'O', 'and_12': 'O', 'changes_13': 'O', 'in_14': 'O', 'motor_15': 'O', 'activity_16': 'O', 'induced_17': 'O', 'by_18': 'O', 'morphine_19': 'B', 'in_20': 'O', 'rats_21': 'O', '._22': 'O', '1_23': 'O', '._24': 'O', 'The_25': 'O', 'effects_26': 'O', 'of_27': 'O', 'two_28': 'O', 'unselective_29': 'O', 'potassium_30': 'B', '(_31': 'O', 'K_32': 'B', '(_33': 'O', '+_34': 'O', ')_35': 'O', '-_36': 'O', ')_37': 'O', 'channel_38': 'O', 'blockers_39': 'O', ',_40': 'O', 'quinine_41': 'B', '(_42': 'O', '12_43': 'O', '._44': 'O', '5_45': 'O', ',_46': 'O', '25_47': 'O', 'and_48': 'O', '50_49': 'O', 'mg_50': 'O', '/_51': 'O', 'kg_52': 'O', ')_53': 'O', 'and_54': 'O', '4_55': 'B', '-_56': 'I', 'aminopyridine_57': 'I', '(_58': 'O', '1_59': 'O', 'and_60': 'O', '2_61': 'O', 'mg_62': 'O', '/_63': 'O', 'kg_64': 'O', ')_65': 'O', ',_66': 'O', 'on_67': 'O', 'conditioned_68': 'O', 'place_69': 'O', 'preference_70': 'O', 'and_71': 'O', 'biphasic_72': 'O', 'changes_73': 'O', 'in_74': 'O', 'motor_75': 'O', 'activity_76': 'O', 'induced_77': 'O', 'by_78': 'O', 'morphine_79': 'B', '(_80': 'O', '10_81': 'O', 'mg_82': 'O', '/_83': 'O', 'kg_84': 'O', ')_85': 'O', 'were_86': 'O', 'tested_87': 'O', 'in_88': 'O', 'Wistar_89': 'O', 'rats_90': 'O', '._91': 'O'}

Now, label the following text: 22 - oxacalcitriol suppresses secondary hyperparathyroidism without inducing low bone turnover in dogs with renal failure . BACKGROUND : Calcitriol therapy suppresses serum levels of parathyroid hormone (PTH) in patients with renal failure but has several drawbacks , including hypercalcemia and / or marked suppression of bone turnover , which may lead to adynamic bone disease .

3. **Few-Shot Prompt:**

Given a biomedical text, perform Named Entity Recognition analysis on this text, focusing on identifying ONLY CHEMICAL ENTITIES. Individual tokens are separated by white space and must be given exactly one label: 'B', 'I', or 'O', according to the BIO labeling schema. Output a JSON object with key equal to token concatenated with '_'+(index of token) and value equal to label, where (index of token) starts from 0 for the first token, 1 for the second token, and so on. Here are a few examples of successful response labels for biomedical texts:

Example Text 1: The effects of quinine and 4 - aminopyridine on conditioned place preference and changes in motor activity induced by morphine in rats . 1 . The effects of two unselective potassium (K (+) -) channel blockers , quinine (12 . 5 , 25 and 50 mg / kg) and 4 - aminopyridine (1 and 2 mg / kg) , on conditioned place preference and biphasic changes in motor activity induced by morphine (10 mg / kg) were tested in Wistar rats .

Example Response 1:{'The_0': 'O', 'effects_1': 'O', 'of_2': 'O', 'quinine_3': 'B', 'and_4': 'O', '4_5': 'B', '-_6': 'I', 'aminopyridine_7': 'I', 'on_8': 'O', 'conditioned_9': 'O', 'place_10': 'O', 'preference_11': 'O', 'and_12': 'O', 'changes_13': 'O', 'in_14': 'O', 'motor_15': 'O', 'activity_16': 'O', 'induced_17': 'O', 'by_18': 'O', 'morphine_19': 'B', 'in_20': 'O', 'rats_21': 'O', '._22': 'O', '1_23': 'O', '._24': 'O', 'The_25': 'O', 'effects_26': 'O', 'of_27': 'O', 'two_28': 'O', 'unselective_29': 'O', 'potassium_30': 'B', '(_31': 'O', 'K_32': 'B', '(_33': 'O', '+_34': 'O', ')_35': 'O',

’-_36’: ’O’, ’)_37’: ’O’, ’channel_38’: ’O’, ’blockers_39’: ’O’, ’,_40’: ’O’, ’quinine_41’: ’B’, ’(_42’: ’O’, ’12_43’: ’O’, ’._44’: ’O’, ’5_45’: ’O’, ’,_46’: ’O’, ’25_47’: ’O’, ’and_48’: ’O’, ’50_49’: ’O’, ’mg_50’: ’O’, ’/_51’: ’O’, ’kg_52’: ’O’, ’)_53’: ’O’, ’and_54’: ’O’, ’4_55’: ’B’, ’-_56’: ’I’, ’aminopyridine_57’: ’I’, ’(_58’: ’O’, ’1_59’: ’O’, ’and_60’: ’O’, ’2_61’: ’O’, ’mg_62’: ’O’, ’/_63’: ’O’, ’kg_64’: ’O’, ’)_65’: ’O’, ’,_66’: ’O’, ’on_67’: ’O’, ’conditioned_68’: ’O’, ’place_69’: ’O’, ’preference_70’: ’O’, ’and_71’: ’O’, ’biphasic_72’: ’O’, ’changes_73’: ’O’, ’in_74’: ’O’, ’motor_75’: ’O’, ’activity_76’: ’O’, ’induced_77’: ’O’, ’by_78’: ’O’, ’morphine_79’: ’B’, ’(_80’: ’O’, ’10_81’: ’O’, ’mg_82’: ’O’, ’/_83’: ’O’, ’kg_84’: ’O’, ’)_85’: ’O’, ’were_86’: ’O’, ’tested_87’: ’O’, ’in_88’: ’O’, ’Wistar_89’: ’O’, ’rats_90’: ’O’, ’._91’: ’O’}

Example Text 2: OBJECTIVES : To investigate the effects of subchronic L - NOARG treatment in haloperidol - induced catalepsy and the number of NOS neurons in areas related to motor control . METHODS : Male albino Swiss mice were treated sub - chronically (twice a day for 4 days) with L - NOARG (40 mg / kg i . p .) or haloperidol (1 mg / kg i .

Example Response 2: {’OBJECTIVES_0’: ’O’, ’:_1’: ’O’, ’To_2’: ’O’, ’investigate_3’: ’O’, ’the_4’: ’O’, ’effects_5’: ’O’, ’of_6’: ’O’, ’subchronic_7’: ’O’, ’L_8’: ’B’, ’-_9’: ’I’, ’NOARG_10’: ’I’, ’treatment_11’: ’O’, ’in_12’: ’O’, ’haloperidol_13’: ’B’, ’-_14’: ’O’, ’induced_15’: ’O’, ’catalepsy_16’: ’O’, ’and_17’: ’O’, ’the_18’: ’O’, ’number_19’: ’O’, ’of_20’: ’O’, ’NOS_21’: ’O’, ’neurons_22’: ’O’, ’in_23’: ’O’, ’areas_24’: ’O’, ’related_25’: ’O’, ’to_26’: ’O’, ’motor_27’: ’O’, ’control_28’: ’O’, ’._29’: ’O’, ’METHODS_30’: ’O’, ’:_31’: ’O’, ’Male_32’: ’O’, ’albino_33’: ’O’, ’Swiss_34’: ’O’, ’mice_35’: ’O’, ’were_36’: ’O’, ’treated_37’: ’O’, ’sub_38’: ’O’, ’-_39’: ’O’, ’chronically_40’: ’O’, ’(_41’: ’O’, ’twice_42’: ’O’, ’a_43’: ’O’, ’day_44’: ’O’, ’for_45’: ’O’, ’4_46’: ’O’, ’days_47’: ’O’, ’)_48’: ’O’, ’with_49’: ’O’, ’L_50’: ’B’, ’-_51’: ’I’, ’NOARG_52’: ’I’, ’(_53’: ’O’, ’40_54’: ’O’, ’mg_55’: ’O’, ’/_56’: ’O’, ’kg_57’: ’O’, ’i_58’: ’O’, ’._59’: ’O’, ’p_60’: ’O’, ’._61’: ’O’, ’)_62’: ’O’, ’or_63’: ’O’, ’haloperidol_64’: ’B’, ’(_65’: ’O’, ’1_66’: ’O’, ’mg_67’: ’O’, ’/_68’: ’O’, ’kg_69’: ’O’, ’i_70’: ’O’, ’._71’: ’O’}

Example Text 3: Recently , fenfluramine appetite suppressants became widely used in the United States but were withdrawn in September 1997 because of concerns over adverse effects . MATERIALS AND METHODS : We conducted a prospective surveillance study on patients diagnosed with pulmonary hypertension at 12 large referral centers in North America .

Example Response 3: {’Recently_0’: ’O’, ’,_1’: ’O’, ’fenfluramine_2’: ’B’, ’appetite_3’: ’O’, ’suppressants_4’: ’O’, ’became_5’: ’O’, ’widely_6’: ’O’, ’used_7’: ’O’, ’in_8’: ’O’, ’the_9’: ’O’, ’United_10’: ’O’, ’States_11’: ’O’, ’but_12’: ’O’, ’were_13’: ’O’, ’withdrawn_14’: ’O’, ’in_15’: ’O’, ’September_16’: ’O’, ’1997_17’: ’O’, ’because_18’: ’O’, ’of_19’: ’O’, ’concerns_20’: ’O’, ’over_21’: ’O’, ’adverse_22’: ’O’, ’effects_23’: ’O’, ’._24’: ’O’, ’MATERIALS_25’: ’O’, ’AND_26’: ’O’, ’METHODS_27’: ’O’, ’:_28’: ’O’, ’We_29’: ’O’, ’conducted_30’: ’O’, ’a_31’: ’O’, ’prospective_32’: ’O’, ’surveillance_33’: ’O’, ’study_34’: ’O’, ’on_35’: ’O’, ’patients_36’: ’O’, ’diagnosed_37’: ’O’, ’with_38’: ’O’, ’pulmonary_39’: ’O’, ’hypertension_40’: ’O’, ’at_41’: ’O’, ’12_42’: ’O’, ’large_43’: ’O’, ’referral_44’: ’O’, ’centers_45’: ’O’, ’in_46’: ’O’, ’North_47’: ’O’, ’America_48’: ’O’, ’._49’: ’O’}

Now, label the following text: 22 - oxacalcitriol suppresses secondary hyperparathyroidism without inducing low bone turnover in dogs with renal failure . BACKGROUND : Calcitriol therapy suppresses serum levels of parathyroid hormone (PTH) in patients with renal failure but has several drawbacks , including hypercalcemia and / or marked suppression of bone turnover , which may lead to adynamic bone disease .

A.3 Prompt Engineering Consideration Factors

	Pros	**Cons**
Prompting with Detailed Input Context	GPT has a more accurate understanding of the BIO labeling schema.	Having a input prompt means there's less tokens allocated towards the model's output.
Prompting with Low Temperature	GPT is less likely to output additional information beyond a JSON string.	N/A
Prompting for a JSON output	GPT- 3.5 is very consistent at outputting syntatically proper JSON objects.	Each JSON object must have unique keys. Since we ask GPT to use the words within the prompt as keys, and words are often duplicated, we attach a unique index to the end of each word.

Table 9: Considerations when Developing a Prompt for generating BioNER Dataset Annotations

A.4 Source Code and Model Repository

GitHub Repository: https://github.com/jennyxycai/llm-annotations. This contains all the code for preprocessing data, generating GPT labels, and performing intrinsic and extrinsic evaluation.

HuggingFace Repository: https://huggingface.co/68611-llm-annotation-group. This contains all of the fine-tuned BPLM models we generated during the extrinsic evaluation process.