Where do LLMs currently stand on biomedical NER in both clean and noisy settings ?

Christophe Ye **Cassie S. Mitchell**

Georgia Institute of Technology

cye73@gatech.edu,
cassie.mitchell@bme.gatech.edu

Abstract

Biomedical Named Entity Recognition (NER) consists of identifying and classifying important biomedical entities mentioned in text. Traditionally, biomedical NER has heavily relied on domain-specific pre-trained language models; particularly variant of BERT models. With the emergence of large language models (LLMs), some studies have evaluated their performance on biomedical NLP tasks. These studies consistently show that, despite their general capabilities, LLMs still fall short compared to specialized BERT-based models for biomedical NER. However, as LLMs continue to advance at a remarkable pace, natural questions arise: Are they still far behind, or are they starting to be competitive? In this study, we investigate the performance of recent LLMs across multiple biomedical NER datasets under both clean and noisy dataset conditions. Our findings reveal that LLMs are progressively closing the performance gap with BERT-based models and demonstrate particular strengths in low-data settings. Moreover, our results suggest that in-context learning with LLMs exhibits a notable degree of robustness to noise, making them a promising alternative in settings where labeled data is scarce or noisy. We released our code on GitHub at https://github.com/CY/biomedicalNER.

1 Introduction

Despite significant advances in medicine, many diseases remain without effective treatments, leaving millions of patients with unmet therapeutic needs (Kusynová et al., 2022). The increasing prevalence of rare, complex, and newly emerging diseases highlights the urgent demand for innovative and scalable approaches to drug discovery and clinical research (Wouters et al., 2020).

As of recent estimates, approximately 3000 to 5000 biomedical articles are published each day and indexed in PubMed (Desai et al., 2018), containing an immense volume of unstructured data, including clinical notes and trial results that offer valuable insights for advancing patient care. Extracting actionable knowledge from this data is therefore both necessary and essential. Clinical meta-analysis, which aggregates findings across multiple clinical studies, plays a foundational role in this effort. It enables the discovery of novel associations between biomedical entities, driving new hypothesis generation and the development of new treatment strategies (Cheerkoot-Jalim and Khedo, 2021), including drug repurposing (Ashburn and Thor, 2004).

However, conducting such analyses typically requires extensive manual annotation and domain expertise, making the process time-consuming and resource-intensive (Wac et al., 2024). To overcome these limitations, researchers have turned to natural language processing (NLP) techniques to automate the extraction of clinical information from biomedical literature (Peng et al., 2020; Kirkpatrick et al., 2022; Kartchner et al., 2023).

Biomedical Named Entity Recognition (NER) is a critical task in biomedical NLP, involving the identification of relevant entities and their semantic types within unstructured text. By automating this process, biomedical NER significantly reduces reliance on manual curation, improves data processing efficiency, and accelerates the translation of research findings into clinical practice (Névéol et al., 2018; Kartchner et al., 2025).

Early clinical NER systems relied on rule-based methods and domain-specific lexicons derived from extensive linguistic and semantic analyses of clinical text (Wang et al., 2018). While effective, these approaches lacked flexibility and generalizability across diverse datasets. Over the past decade, the field has undergone a significant shift with the rise of machine learning, particularly deep learning-based models. A major breakthrough

Findings of the Association for Computational Linguistics: EACL 2026, pages 977–1001

came with the introduction of domain-adapted pretrained language models, which have since become central to biomedical NLP.

More recently, Large Language Models (LLMs) have emerged as another alternative for biomedical NLP. While several studies (Gutiérrez et al., 2022; Keloth et al., 2024; Munnangi et al., 2024; Hu et al., 2024) have explored their effectiveness on biomedical NER, findings consistently show that LLMs lag significantly behind state-of-the-art (SOTA) BERT-based models and have yet to demonstrate practical usefulness in this field. However, given the rapid advancements of LLMs (Ho et al., 2024), revisiting their ability is essential to assess the progress made since earlier evaluations but also explore emerging practical applications of these models. Furthermore, it is crucial to identify the current bottlenecks of LLMs and determine what improvements are needed for them to serve as viable replacements for BERT-based models. The contributions in this work are as follows:

- We examine the current state of LLM performance on biomedical NER by incorporating the most recent models into the evaluation.
- We investigate MedMentions-ST21PV (Mohan and Li, 2019), a more challenging dataset with 21 entity types that can lead to labeling ambiguities.
- We conduct an error analysis, categorizing errors into five types and revealing key limitations of LLMs in biomedical NER.
- We demonstrate the effectiveness of LLMs in low-resource settings, highlighting their potential when annotated data is limited.
- To our knowledge, we present the first study to assess the robustness of LLMs to noisy annotations for in-context learning in the biomedical domain.

2 Related Work

BERT-based models (Devlin et al., 2019) pretrained on large-scale biomedical corpora such as PubMed abstracts (PubMed, 2025) and clinical notes have shown strong performance across diverse clinical NLP tasks. When fine-tuned for applications like NER, they effectively extract biomedical entities from unstructured text. Among the most prominent biomedical-specific models are BioBERT (Lee et al., 2019), PubMedBERT (Gu et al., 2021), and BiolinkBERT (Yasunaga et al., 2022); all of which have consistently achieved SOTA in biomedical entity extraction.

Dataset	Train Doc	Val Doc	Test Doc	Ent Type
NCBI-Disease	592	100	100	1
BC5CDR	500	500	500	2
MM-ST21PV	2,635	878	879	21
MT-Samples	602	201	201	3
VAERS	603	126	286	4

Table 1: Summary of datasets used

LLMs have recently gained attention as alternatives for clinical NLP (Jahan et al., 2024; Tian et al., 2023), including biomedical NER. For instance, Gutiérrez et al. (2022) evaluated GPT-3 (Brown et al., 2020) on multiple biomedical NER datasets and found that it significantly underperformed compared to specialized BERT-based models, inviting people to reconsider their eventual usage for this task. Subsequent efforts have aimed to reduce this performance gap. Notably, Munnangi et al. (2024) proposed augmenting LLM prompts with definitions of relevant biomedical concepts to enhance its in-context learning (ICL) ability. Similarly, Keloth et al. (2024) explored instruction-tuning techniques using LLaMA models (Touvron et al., 2023) to improve their alignment with biomedical NER tasks, achieving performance comparable to that of PubMedBERT. However, their method is limited to either a single target entity type per input, such as disease, chemical, or gene, specified through the instruction prompt, or to few entity types that are very distinct and less likely to be confused, which reduces the risk of mislabeling. Keloth et al. (2024) also split abstracts into individual sentences before annotation to improve performance. However, sentence-level splitting becomes less effective in complex scenarios involving multiple entity types, where determining the correct type often requires broader context. Consequently, this approach is not well suited for comprehensive biomedical information extraction (Nye et al., 2018).

3 Methodolody

3.1 Datasets

We conducted our experiments using three biomedical datasets from the BigBio framework (Fries et al., 2022) and two additional clinical datasets :

- NCBI-Disease (Dogan et al., 2014) : A corpus of 792 abstracts annotated with disease mentions.
- BC5CDR (Li et al., 2016) : A corpus of 1500 abstracts annotated with chemical and disease

mentions.

- MedMentions-ST21PV (MM-ST21PV) (Mohan and Li, 2019) : 4392 abstracts annotated with UMLS concepts, filtered to include 21 biomedical semantic types.
- MT-Samples (Hu et al., 2024) : 163 synthetic discharge summaries annotated using the 2010 i2b2 guideline schema (Medical Problems, Treatments, Tests).
- VAERS (Hu et al., 2024) : 91 publicly available adverse-event safety reports focused on nervous system disorder–related events.

A summary of these datasets are provided in Table 1.

3.2 Models

Our evaluation covers both SOTA BERT-based models and the latest LLMs on the five datasets.

For the biomedical BERT variants, we focus on three widely recognized models: BioBERT (Lee et al., 2019), PubMedBERT (Gu et al., 2021), and BiolinkBERT (Yasunaga et al., 2022); each of which has consistently achieved strong performance on biomedical NER tasks. In addition, we include the latest advancement ModernBERT model (Warner et al., 2024).

For LLMs, we evaluate the performance of five models:

- Meta-Llama-3.1-8B-Instruct (Llama-3.1): A 8B-parameter instruction-tuned model released in July 2024 by Meta (Grattafiori et al., 2024)
- Phi-4: A 14B-parameter model released in December 2024 by Microsoft (Abdin et al., 2024)
- Mistral-Small-3.1-24B-Instruct-2503 (Mistral-3.1): A 24B parameters instruct model released on March 2025 by Mistral (Mistral, 2025)
- gemma-3-12B-it (Gemma3): A 12B-parameter instruction-tuned model released in March 2025 by Google (Gemma, 2025)
- gpt-4o-2024-08-06 (GPT-4o): Model released in May 2024 by OpenAI et al. (2024)

More information on these models can be found in Appendix A.1.1

3.3 Experimental setup for LLMs

In this subsection, we describe the key aspects specific to the LLM experiments.

3.3.1 Annotation framework

We follow prior delimiter-based LLM NER methods (Keloth et al., 2024; Hu et al., 2024), adapting the format for mentions with ambiguous types. Each mention is annotated as follows:

Text @@ [Mention] ## [Entity type] @@ Text

An illustration of this framework is presented in Figure 1. In the MM-ST21PV dataset, overlapping entity spans were resolved by keeping the first span in the text; if spans started at the same position, the shorter one was retained.

3.3.2 In-context learning

To evaluate LLMs, we use their ICL abilities to perform tasks using only a prompt with a few examples, without additional training (Dong et al., 2024). To enhance the relevance of the example, we perform a similarity-based nearest neighbor search using faiss (Douze et al., 2025), dynamically selecting the most suitable training samples to include in the prompt.

3.3.3 LLM Fine-tuning

In addition to leveraging ICL, fine-tuning (Wei et al., 2022) is also explored as a means to further improve model performance. For this purpose, we fine-tune Mistral-3.1 model (Mistral, 2025), which performed the best in our evaluations.

4 Results with Models trained on Gold Data

This section presents the evaluation results for all BERT-based and LLMs models introduced earlier, using only gold-standard (noise-free) annotations for both ICL and fine-tuning. For each evaluation, we report the standard metrics for NER: F1 score (F1), Precision (P), and Recall (R).

4.1 BERT-Based Models Results

The results for BERT-based models are shown in Table 2 and 3 and serve as a baseline for comparing LLM performance. Training parameters are provided in Appendix A.1.2.

BiolinkBERT and PubMedBERT emerge as the best performing models for biomedical NER. Additionally, these results show that, despite all recent advancements, general-purpose ModernBERT still underperform compared to domain-specific BERT variants in biomedical NER tasks.

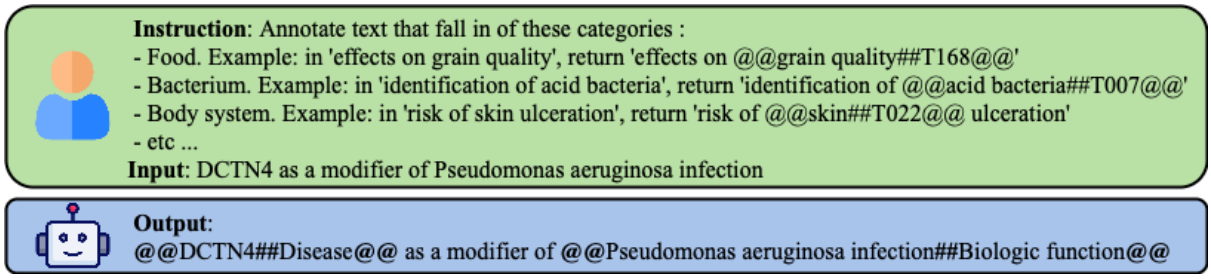


Figure 1: Illustration of annotation framework

4.2 LLMs Results

4.2.1 Results with LLMs using ICL

The performance of LLM was evaluated with a varying number of in-context examples (k = 1, 3, 5, 10). The results are shown in Table 2 and 3. All reported numbers are averaged over three runs, with variances provided in Table 18. The exact prompt details can be found in Appendix A.1.3. In addition to performance, we also track the proportion of cases where the model fails to follow instructions (denoted by "Inv." row in the Table), resulting in invalid output. This includes cases where the LLM produces invalid JSON, fails to generate any annotations, appends annotations at the end of the text instead of performing on-the-fly annotation, as well as other cases of hallucination.

Overall, recently released open-source models outperform their earlier counterparts. Notably, there is a significant performance gap between the oldest evaluated model (Llama-3.1 released in July 2024) and the most recent one (Mistral-3.1 released in March 2025), with relative improvements ranging from 10% to 60% depending on the dataset. The most recent models are also way better at following instructions especially when they work with 10 examples.

Additionally, increasing the number of in-context examples appears to benefit only the strongest models. Indeed, Mistral-3.1 and GPT-4o consistently improve as k increases. In contrast, lower-performing models show little to no improvement. In fact, in most cases, adding more examples further reduces their ability to follow prompt instructions which leads to a higher proportion of invalid outputs.

Nonetheless, the progression of LLMs over time on biomedical NER tasks is noteworthy, with statistical significance tests reported in Appendix A.1.6.

These results indicate that despite remarkable progress, LLMs are still not usable for biomedical NER, with a notable performance gap. We therefore investigate whether instruction-tuning can narrow this gap.

4.2.2 Results with fine-tuned LLM

This section investigates whether supervised fine-tuning can close the performance gap between the general-purpose LLMs and domain-specific BERT models. To this end, we fine-tune Mistral-3.1, which demonstrated the strongest performance among the LLMs in our earlier evaluation.

The results are shown in Table 4. All fine-tuning hyperparameters are documented in Appendix A.1.4.

Despite fine-tuning, LLMs reach about 90% of the performance of the best biomedical BERT model across all datasets, with the exception of MT-Samples where it slightly outperforms it. While this represents meaningful progress, LLMs originally designed for text generation still lag behind BERT-based models for structured tasks such as sequence classification, where BERT remains the stronger choice.

4.3 Performance comparison of BERT and LLMs with limited data

Annotating biomedical corpora for NER is costly, time-consuming, and prone to inconsistencies even among experts. The scarcity of high-quality annotated clinical text remains a major bottleneck for developing robust biomedical NLP systems (Sivarajkumar and Wang, 2022; Mohan et al., 2021).

Moreover, existing gold-standard NER datasets cover only narrow subsets of the biomedical domain, limiting their generalizability. This highlights the need for models with strong zero-shot capabilities that can adapt across diverse biomedical settings without extensive annotation, making it critical to evaluate LLM performance under low-resource conditions.

In this subsection, we explore low-resource con-

Datasets (→)		NCBI-Disease				BC5CDR				MM-ST21PV			
BERT models (↓)													
BioLinkBERT	F1	**0.891**				**0.940**				0.632			
	P	**0.880**				**0.934**				0.619			
	R	**0.902**				0.946				**0.646**			
PubMedBERT	F1	0.870				0.932				**0.639**			
	P	0.868				0.912				**0.633**			
	R	0.872				**0.952**				0.644			
BioBERT	F1	0.862				0.914				0.613			
	P	0.848				0.903				0.591			
	R	0.877				0.925				0.636			
ModernBERT	F1	0.841				0.909				0.594			
	P	0.827				0.900				0.575			
	R	0.855				0.918				0.614			
LLMs (↓)		**k=1**	**k=3**	**k=5**	**k=10**	**k=1**	**k=3**	**k=5**	**k=10**	**k=1**	**k=3**	**k=5**	**k=10**
Llama-3.1	F1	0.489	0.513	0.554	0.551	0.543	0.580	0.610	0.598	0.239	0.305	0.314	0.287
	P	0.518	0.622	0.697	0.677	0.720	0.761	0.774	0.793	0.340	0.424	0.454	0.500
	R	0.476	0.436	0.460	0.464	0.435	0.468	0.503	0.479	0.184	0.239	0.241	0.199
	Inv.	4.0%	8.7%	10.3%	29.3%	2.9%	7.9%	9.9%	24.1%	10.9%	8.2%	12.2%	23.7%
Phi-4	F1	0.536	0.548	0.562	0.542	0.633	0.652	0.649	0.645	0.259	0.302	0.311	–
	P	0.734	0.727	0.732	0.750	0.784	0.779	0.798	0.797	0.398	0.444	0.465	–
	R	0.422	0.440	0.456	0.424	0.530	0.560	0.547	0.542	0.192	0.228	0.234	–
	Inv.	4.7%	2.7%	3.3%	1.7%	0.6%	0.0%	0.4%	0.05%	8.0%	7.6%	9.3%	–
Gemma-3	F1	0.565	0.620	0.616	0.629	0.693	0.718	0.723	0.718	0.336	0.374	0.364	0.329
	P	0.544	0.597	0.627	0.699	0.679	0.720	**0.763**	**0.762**	0.372	0.463	0.502	0.537
	R	0.590	0.645	0.604	0.579	0.708	0.716	0.678	0.686	0.308	0.314	0.285	0.237
	Inv.	6.3%	1.3%	1.0%	9.0%	1.9%	0.5%	1.4%	2.7%	5.0%	14.1%	21.0%	47.9%
Mistral-3.1	F1	0.648	0.703	0.721	0.724	0.725	0.745	0.760	**0.770**	**0.399**	**0.469**	**0.496**	**0.509**
	P	0.651	0.669	0.672	0.684	0.708	0.713	0.733	0.738	**0.372**	0.444	0.480	0.504
	R	**0.666**	**0.742**	**0.778**	**0.770**	**0.742**	**0.780**	**0.789**	**0.805**	**0.430**	**0.497**	**0.516**	**0.511**
	Inv.	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.2%	0.2%	0.4%	2.8%
GPT-4o	F1	**0.670**	**0.717**	**0.730**	**0.745**	**0.737**	**0.755**	**0.767**	0.764	0.363	0.413	0.440	0.457
	P	**0.723**	**0.744**	**0.741**	**0.742**	**0.763**	**0.760**	0.762	0.753	0.361	0.402	0.427	0.447
	R	0.624	0.691	0.719	0.748	0.713	0.749	0.768	0.775	0.373	0.427	0.453	0.467
	Inv.	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.2%	0.1%	0.1%	0.3%

Table 2: Comparison of F1, Precision (P), and Recall (R) scores for BERT-based models and LLMs for different in-context example counts (k) on NCBI-Disease, BC5CDR and MM-ST21PV. The column "Inv." stands for **"Invalid evaluation"** and indicates the % of abstracts excluded due to instruction-following errors. F1 scores are computed only on valid outputs; including invalid cases would further reduce scores. Rows marked "–" denote context-window limits preventing evaluation with the specified number of examples. **Bold** : Best result

ditions by comparing several BERT-based models against Mistral-3.1 using progressively larger subsets of annotated abstracts to simulate low-resource scenarios. The results of this comparison are summarized in Table 5.

Results show that Mistral performance is not affected drastically even with very limited examples. With just 20 annotated abstracts, it achieves almost 90% of its original performance. In contrast, BERT-based models experience a sharp decline, losing around 40% of their original performance is some scenarios, which places them behind Mistral in this low-data configuration. This observation remains true for dataset sizes up to 60 abstracts. Although Mistral's performance gains currently plateau early due to inherent limitations, it is promising to consider that future improvements may allow it to outperform BERT models across a much wider range of data sizes, thereby reducing the reliance on extensive manual annotation that BERT models require.

4.4 Error Analysis

This section analyzes the error distribution of model-generated annotations on the MM-ST21PV dataset (Mohan and Li, 2019). Table 6 summarizes the defined error types, while Figure 2 illustrates their distribution across annotations from all evaluated models.

The main takeaway from this figure is that different LLMs tend to make different types of mistakes. GPT-4o and Mistral-3.1 show a more varied mix of errors, with a slight majority being "invalid tag" errors. These models make about as much annotations as the examples provided in the prompt, which is the number of annotations made by human experts.

In contrast, Llama3.1, Phi-4, and Gemma3 made

Datasets (→)		MT-Samples				VAERS			
BERT models (↓)									
BioLinkBERT	F1	0.787				0.690			
	P	0.786				0.643			
	R	0.788				**0.745**			
PubMedBERT	F1	**0.800**				**0.697**			
	P	**0.797**				**0.659**			
	R	**0.803**				0.739			
BioBERT	F1	0.777				0.630			
	P	0.758				0.566			
	R	0.796				0.712			
ModernBERT	F1	0.736				0.523			
	P	0.730				0.471			
	R	0.743				0.586			
LLMs (↓)		**k=1**	**k=3**	**k=5**	**k=10**	**k=1**	**k=3**	**k=5**	**k=10**
Llama-3.1	F1	0.488	0.528	0.520	0.503	0.419	0.499	0.498	0.504
	P	0.486	0.506	0.505	0.553	0.470	0.519	0.511	0.517
	R	0.489	0.552	0.535	0.527	0.378	0.479	0.485	0.492
	Inv.	5.0%	6.0%	5.0%	5.5%	5.6%	7.0%	7.7%	7.7%
Phi-4	F1	0.530	0.577	0.608	0.610	0.465	0.492	0.513	0.544
	P	0.538	0.567	0.603	0.597	0.559	0.549	0.552	0.569
	R	0.521	0.589	0.613	0.624	0.398	0.445	0.479	0.521
	Inv.	0.0%	2.0%	2.5%	2.5%	0.0%	0.0%	0.0%	0.0%
Gemma-3	F1	0.527	0.537	0.540	0.558	0.509	0.518	0.549	0.559
	P	0.519	0.498	0.498	0.552	0.495	0.484	0.517	0.540
	R	0.535	0.582	0.589	0.564	0.525	0.556	0.584	0.578
	Inv.	4.0%	2.5%	2.5%	2.5%	0.0%	0.0%	0.0%	0.0%
Mistral-3.1	F1	0.570	0.639	0.645	**0.680**	0.463	0.521	0.530	0.546
	P	0.540	0.594	0.599	**0.630**	0.437	0.472	0.480	0.492
	R	0.602	0.691	0.699	**0.738**	0.492	0.580	0.592	0.621
	Inv.	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
GPT-4o	F1	0.561	0.601	0.600	0.614	0.527	0.568	0.588	**0.599**
	P	0.525	0.560	0.558	0.572	0.523	0.536	0.557	**0.560**
	R	0.603	0.649	0.649	0.663	0.531	0.604	0.623	**0.643**
	Inv.	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%

Table 3: Comparison of F1, Precision (P), and Recall (R) scores for BERT-based models and LLMs for different in-context example counts (k) on MT-Samples and VAERS. The column "Inv." stands for **"Invalid evaluation"** and indicates the % of abstracts excluded due to instruction-following errors. F1 scores are computed only on valid outputs; including invalid cases would further reduce scores. Rows marked "–" denote context-window limits preventing evaluation with the specified number of examples. **Bold** : Best result

	Mistral-3.1		
Datasets	F1	P	R
NCBI-Disease	0.869	0.880	0.860
BC5CDR	0.868	0.852	0.885
MM-ST21PV	0.583	0.588	0.579
MT-Samples	0.812	0.805	0.819
VAERS	0.688	0.698	0.678

Table 4: F1, Precision, Recall of fine-tuned Mistral-3.1.

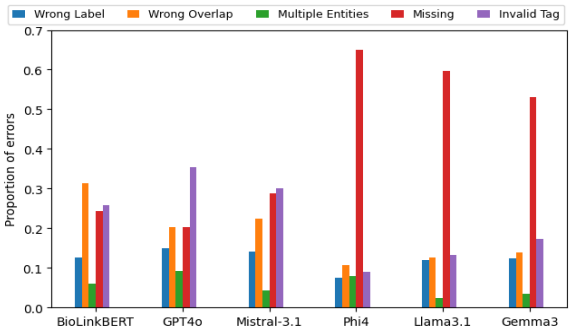


Figure 2: Error Type Distribution of different models annotations on the MM-ST21PV Dataset

far fewer annotations. As a result, most of their errors fall under the "missing" category. Attempts to encourage the model to annotate more aggressively had little effect, as these models appear to be inherently more conservative in their annotations even with many examples in the prompt.

BiolinkBERT has an error distribution closer to GPT-4o and Mistral-3.1, but with a higher incidence of "wrong overlap" errors. This is consistent with its token-level annotation approach, which increases the likelihood of boundary mismatches.

Table 7 summarizes the statistics of all LLMs annotations on MM-ST21PV dataset.

Models	BioLinkBERT			PubMedBERT			BioBERT			Mistral-Small-3.1		
Number of abstracts	F1	P	R	F1	P	R	F1	P	R	F1	P	R
20 (0.8%)	0.356	0.333	0.385	0.370	0.345	0.399	0.332	0.301	0.371	**0.433**	**0.441**	**0.424**
40 (1.5%)	0.403	0.362	0.454	0.411	0.373	**0.458**	0.365	0.319	0.426	**0.449**	**0.469**	0.430
100 (3.8%)	0.490	0.457	**0.528**	**0.502**	**0.489**	0.517	0.464	0.422	0.515	0.454	0.466	0.442

Table 5: Comparison F1, precision, and recall metrics for BERT-based models and Mistral-3.1 (10-shot) using a small subset of the MM-ST21PV dataset. Evaluation performed on the first 40 test abstracts. **Bold** : Best result

Errors	Definition	Example
Wrong Label	The entity span is correctly identified, but the assigned semantic type is incorrect.	@@Famotidine##Disease@@ is a histamine ✗ @@Famotidine##Chemical@@ is a histamine ✓
Wrong Overlap	The span boundaries are inaccurate : too long, too short, or both.	Investigate @@interstitial fibrosis##Disease@@ ✗ Investigate interstitial @@fibrosis##Disease@@ ✓
Missing (FN)	A valid entity present in the text is not detected by the model.	Incidence of postoperative delirium ✗ Incidence of @@postoperative delirium##Disease@@ ✓
Invalid Tag (FP)	A span is annotated as an entity when it should not be annotated at all.	The function of @@P2X3##Chemical@@ receptor ✗ The function of P2X3 receptor ✓
Multiple Entities	A single span incorrectly includes more than one distinct entity.	Induced by @@tacrolimus and prednisolone##Chemical@@ ✗ Induced by @@tacrolimus##Chemical@@ and @@prednisolone##Chemical@@ ✓

Table 6: Summary of all possible errors type during annotation

	F1	Noise	Correct	Wrong Label	Wrong Overlap	Multiple Entities	Missing	Invalid Tag
Llama3.1	0.318	0.682	10139	4096	4280	813	20281	4488
Phi4	0.310	0.690	9482	2324	3321	2449	20254	2821
Gemma3	0.379	0.621	12797	3934	4422	1110	16935	5536
Mistral-3.1	0.508	0.492	20609	4047	6416	1260	8178	8596
GPT-4o	0.484	0.516	19885	4137	5657	2539	5651	9855

Table 7: Statistics on the error types of the annotations of all evaluated LLMs on the MM-ST21PV dataset. The clean test split contains a total of 39201 annotations.

5 Results with Noisy Data

The previous section evaluated the performance of BERT-based models and LLMs under ideal conditions using gold standard training data. However, in real-world settings, annotations can be noisy. For instance, alternative approaches such as crowdsourcing or distant supervision are commonly used for more efficient annotations, which can introduce substantial noise into the data (Bölücü et al., 2024; Amin et al., 2020).

Prior studies have shown that BERT models are particularly sensitive to such noise (Wei et al., 2024; Moradi et al., 2021). This section investigates how robust are LLMs against noise by introducing noisy examples into the prompt and comparing their performance against BERT models trained on equally noisy datasets. All experiments are conducted on the MM-ST21PV dataset.

An overview of the experimental setup for testing the noise robustness of LLMs is provided in Figure 3. To generate noisy examples, we use LLM-generated annotations similar to the evaluation done on the previous section but on the training split instead. For LLMs that produce invalid outputs, we randomly replaced those annotated abstracts with valid ones from other LLMs. We use the annotations in their original form and progressively reduce the noise in 10% increments by randomly removing noisy labels in order to create datasets with varying noise levels.

5.1 BERT-Based Models Results

For each noise level, we train BiolinkBERT model and report the results in Table 8.

The results show that BERT models are highly sensitive to label noise: at 50% noise, performance drops by 20–30% regardless of the source. Noise dominated by missing labels (e.g., from Phi-4 and Gemma-3) yields the sharpest declines due to reduced recall, while LLMs such as GPT-4o and

Models	phi-4			GPT-4o			Gemma3			Mistral-3.1		
Noise Level	F1	P	R	F1	P	R	F1	P	R	F1	P	R
10%	0.591	0.591	0.592	0.595	0.582	0.609	0.594	0.606	0.584	0.598	0.582	0.615
20%	0.569	0.586	0.552	0.572	0.562	0.582	0.563	0.596	0.534	0.578	0.561	0.596
30%	0.545	0.572	0.521	0.544	0.526	0.563	0.519	0.593	0.462	0.557	0.544	0.571
40%	0.499	0.489	0.508	0.518	0.494	0.544	0.509	0.522	0.495	0.540	0.517	0.566
50%	0.455	0.522	0.404	0.494	0.469	0.522	0.463	0.558	0.396	0.521	0.496	0.549
60%	0.412	0.438	0.389		-		0.418	0.492	0.363		-	

Table 8: Comparison of the F1, Precision and Recall metrics of BioLinkBERT model after training on noisy datasets generated by different LLMs, evaluated across varying noise levels on the MM-ST21PV dataset.

	Noise Source	BioLinkBERT			phi-4			GPT-4o			Gemma3			Mistral-Small-3.1		
Evaluated LLM	**Noise Level**	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R
GPT-4o	**10%**	0.466	0.457	0.475	0.450	0.447	0.454	0.456	0.437	0.477	0.455	0.447	0.462	0.475	0.457	0.495
	20%	0.480	0.468	0.492	0.436	0.436	0.435	0.452	0.437	0.469	0.454	0.454	0.453	0.478	0.467	0.490
	30%	0.478	0.461	0.496	0.432	0.440	0.425	0.446	0.427	0.468	0.459	0.462	0.455	0.487	0.469	0.506
	40%		-		0.423	0.434	0.413	0.436	0.417	0.457	0.462	0.476	0.450	0.491	0.469	0.516
	50%		-		0.417	0.427	0.407	0.437	0.418	0.458	0.455	0.463	0.448	0.485	0.459	0.514
	60%		-		0.408	0.423	0.394		-		0.459	0.473	0.446		-	
Mistral-Small-3.1	**10%**	0.492	0.468	0.519	0.491	0.483	0.500	0.481	0.453	0.512	0.483	0.470	0.497	0.497	0.472	0.525
	20%	0.479	0.478	0.480	0.470	0.475	0.465	0.475	0.442	0.513	0.471	0.470	0.472	0.484	0.448	0.526
	30%	0.496	0.491	0.500	0.482	0.492	0.472	0.471	0.443	0.504	0.469	0.492	0.448	0.492	0.456	0.533
	40%		-		0.440	0.483	0.405	0.467	0.439	0.498	0.472	0.507	0.441	0.486	0.453	0.524
	50%		-		0.433	0.483	0.393	0.471	0.452	0.492	0.483	0.523	0.449	0.487	0.463	0.514
	60%		-		0.420	0.485	0.370		-		0.471	0.557	0.400		-	

Table 9: F1, Precision and Recall of LLMs performance across difference noise sources and levels on the MM-ST21PV dataset using 10 in-context examples. The evaluation was conducted on the first 40 test abstracts, and results are based on a single run. Rows marked "–" indicate cases where the original model did not reach the specified noise level.

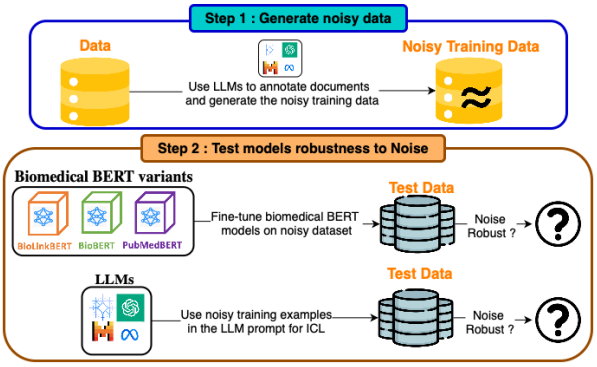


Figure 3: Experimental setup for testing the noise robustness of the different models.

Mistral which exhibit more balanced error patterns are affected less severely. This suggests that the type of noise plays a role just as critical as the noise level in how much it perturbs a model performance.

5.2 LLMs Results

For each noise level, we evaluate the performance of GPT-4o and Mistral-3.1 using noisy labels for ICL in a 10-shot setting. The results are summarized in Table 9.

All experiments were run on the first 40 abstracts of the MM-ST21PV dataset due to cost and resource constraints associated with the scale of this evaluation.

The results demonstrate that both LLMs are remarkably robust to noisy annotations introduced in the prompt examples. In most cases, performance remains comparable to that achieved with clean annotations. Even under the highest noise levels, performance typically decreases by only around 10%, with the largest observed drop observed at less than 20% occurring when GPT-4o is prompted using Phi-4 annotations at 60% noise level. These results indicate that annotation precision is less critical for LLMs compared to BERT-based models. As a result, more lenient annotation strategies may be acceptable when the data is intended for use as in-context examples in LLM prompts. Statistical significance tests supporting this conclusion are provided in Appendix A.1.7.

6 The Current Landscape of LLMs for Biomedical NER

This section summarizes our key findings on LLM-based approaches to biomedical NER. Results from Section 4 show that LLMs still lag behind tradi-

tional NLP methods. Indeed, most open-source models struggle to follow annotation instructions, often producing invalid outputs. This issue worsens with longer prompts: for example, Llama-3.1 produced over 26% invalid outputs when given 10 annotated examples for MM-ST21PV dataset. Attempts to reduce output format errors with alternative prompting strategies (Appendix A.3.2) were also ineffective. These findings highlight the need for substantial improvements in the instruction-following capabilities of smaller LLMs before they can be realistically applied to biomedical NER. Additionally, even when valid outputs were obtained, the annotation quality remained poor.

The error analysis from section 4.4 further shows that models such as Llama-3.1, Phi-4, and Gemma-3 tend to annotate too conservatively, missing most entities with attempts to encourage more aggressive annotation also failing, suggesting an inherent conservative bias in these models. Even proprietary models like GPT-4o fall short of biomedical BERT models, with frequent annotation errors, particularly invalid tags. Addressing this issue represents a key step toward enabling LLMs to perform high-quality biomedical NER. Reducing such false positives through verifier models or other safeguard mechanisms remains essential for achieving BERT-level performance.

Although fine-tuning improves performance, LLMs still fall short of BERT-based models for most datasets while demanding far greater computational resources. For instance, fine-tuning Mistral-3.1 on MM-ST21PV took 22 hours on a single H100 GPU, compared to 50 minutes for BioLinkBERT on an A40. Inference on the same test set required 9 hours for Mistral-3.1 and less than 1 hour for BERT models. Full training and inference times for all datasets are listed in Appendix A.1.4 and A.1.5.

Nonetheless, Sections 4.3 and 5 highlight two scenarios where LLMs can offer value. First, when annotated data is scarce, LLMs can outperform BERT-based models: Table 5 shows Mistral-3.1 performance surpassing BERT when there are less than 60 annotated abstracts available. Second, under noisy annotation conditions, LLMs can again outperform BERT: Tables 8 and 9 illustrate cases where LLMs trained with 50–60% noisy data achieve better results than BERT trained on the same proportion of noisy data.

As of today, LLMs could be incorporated into biomedical NER workflows as follows : 1) Cold-start data generation, where LLMs are used to produce synthetic annotated data when no labeled corpus is available, enabling BERT-based models to be trained from scratch and 2) Selective LLM usage where it's only applied to instances of data slices where it outperforms BERT models.

Beyond overall robustness to noise, model behavior also depends strongly on the type of noisy examples provided in the prompt. Mistral and GPT-4o with perfect examples tended to over-annotate, producing many invalid tags. In contrast, when exposed to noisy examples from Gemma-3 and Phi-4 where errors primarily involved missing entities, both models became more conservative: recall declined but precision improved. As shown in Table 9, this trade-off becomes more pronounced as noise from these sources increases. By comparison, when the noisy examples were generated by GPT-4o and Mistral-3.1, whose errors were more uniformly distributed, both recall and precision dropped together.

These findings show that LLMs remain both less accurate and more resource-intensive than BERT-based models for biomedical NER. Substantial efficiency gains are required before LLMs can be considered practical alternatives. Future progress will depend not only on improving accuracy, but also on reducing computational cost and resource demands so that LLMs can match BERT accessibility and scalability in real-world biomedical applications.

7 Conclusion and Future Work

Biomedical NER plays a key role in biomedical NLP by automating the identification of relevant entities, which helps minimize manual effort, improve data processing efficiency, and accelerates the integration of research insights into clinical practice. This study made an update on the current state of LLM performance on biomedical NER, revealing substantial progress over previous evaluations. Although LLMs are still behind traditional domain-specific BERT models in overall performance, they demonstrate advantages in specific scenarios, notably in low-data and noisy scenarios. These advancements highlight the growing potential of LLMs for biomedical NER and point toward promising future applications in the field.

8 Limitations

While LLMs have demonstrated promising performance in certain scenarios, they still exhibit notable limitations and uncertainties. Despite recent advances, LLMs continue to underperform compared to specialized BERT-based models in biomedical NER tasks, and there is no guarantee that improvements in general LLM capabilities will directly translate to better performance in this domain. Moreover, inference with LLMs requires significantly more time and computational resources than with BERT models (A.1.5), and this gap widens further when training is involved.

In addition to efficiency considerations, the application of LLMs to biomedical and clinical text introduces risks related to bias propagation and hallucinations. This introduces significant risks in clinical environments where precision is paramount. Furthermore, the use of biomedical text generated or processed by general LLMs raises unresolved questions around patient privacy and data governance, especially when models are deployed in settings involving sensitive data.

Lastly, our findings suggests that LLMs perform well in low-data settings with in-context learning and exhibit greater robustness to noisy inputs compared to BERT models. However, this observation is based on a single experimental scenario and may not generalize to other contexts. Furthermore, it may not hold for more advanced models developed in the future. We leave the exploration of this trend as a direction for future work.

Acknowledgments

Funding support provided by National Science Foundation CAREER grant 1944247, National Institute of Health grant R35GM152245, and Chan Zuckerberg Initiative grant 253558 to C.S.M.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. 2024. Phi-4 technical report. *Preprint*, arXiv:2412.08905.

Saadullah Amin, Katherine Ann Dunfield, Anna Vechkaeva, and Guenter Neumann. 2020. A data-driven approach for noise reduction in distantly supervised biomedical relation extraction. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 187–194, Online. Association for Computational Linguistics.

Ted T Ashburn and Karl B Thor. 2004. Drug repositioning: identifying and developing new uses for existing drugs. *Nature reviews Drug discovery*, 3(8):673–683.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Necva Bölücü, Maciej Rybinski, Xiang Dai, and Stephen Wan. 2024. An adaptive approach to noisy annotations in scientific information extraction. *Information Processing Management*, 61(6):103857.

Sudha Cheerkoot-Jalim and Kavi Kumar Khedo. 2021. Literature-based discovery approaches for evidence-based healthcare: a systematic review. *Health and Technology*, 11:1205 – 1217.

Priya Desai, Natalie Telis, Benjamin Lehmann, Keith Bettinger, Kizito Jonathan, and Somalee Datta. 2018. Scireader*: A cloud-based recommender system for biomedical literature.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. pages 4171–4186.

Rezarta Islamaj Dogan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: A resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. *Preprint*, arXiv:2301.00234.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2025. The faiss library. *Preprint*, arXiv:2401.08281.

Jason A. Fries, Troy Mayfield, Kenneth Brimacombe, Michael M. Bronstein, David Silver, David Wehner, et al. 2022. Bigbio: A large-scale biomedical corpus. *Journal of Biomedical Informatics*, 135:104037.

Gemma. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, and etc... 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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

Bernal Jiménez Gutiérrez, Nikolas McNeal, Clay Washington, You Chen, Lang Li, Huan Sun, and Yu Su. 2022. Thinking about gpt-3 in-context learning for biomedical ie? think again. *Preprint*, arXiv:2203.08410.

Anson Ho, Tamay Besiroglu, Ege Erdil, David Owen, Robi Rahman, Zifan Carl Guo, David Atkinson, Neil Thompson, and Jaime Sevilla. 2024. Algorithmic progress in language models. *Preprint*, arXiv:2403.05812.

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. *Preprint*, arXiv:2106.09685.

Yan Hu, Qingyu Chen, Jingcheng Du, Xueqing Peng, Vipina Kuttichi Keloth, Xu Zuo, Yujia Zhou, Zehan Li, Xiaoqian Jiang, Zhiyong Lu, Kirk Roberts, and Hua Xu. 2024. Improving large language models for clinical named entity recognition via prompt engineering. *Preprint*, arXiv:2303.16416.

Israt Jahan, Md Tahmid Rahman Laskar, Chun Peng, and Jimmy Xiangji Huang. 2024. A comprehensive evaluation of large language models on benchmark biomedical text processing tasks. *Computers in Biology and Medicine*, 171:108189.

David Kartchner, Selvi Ramalingam, Irfan Al-Hussaini, Olivia Kronick, and Cassie Mitchell. 2023. Zero-shot information extraction for clinical meta-analysis using large language models. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 396–405, Toronto, Canada. Association for Computational Linguistics.

David Kartchner, Haydn Turner, Christophe Ye, Irfan Al-Hussaini, Batuhan Nursal, Albert J. B. Lee, Jennifer Deng, Courtney Curtis, Hannah Cho, Eva L. Duvaris, Coral Jackson, Catherine E. Shanks, Sarah Y. Tan, Selvi Ramalingam, and Cassie S. Mitchell. 2025. Trialsieve: A comprehensive biomedical information extraction framework for pico, meta-analysis, and drug repurposing. *Bioengineering*, 12(5).

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

Anna Kirkpatrick, Chidozie Onyeze, David Kartchner, Stephen Allegri, Davi Nakajima An, Kevin McCoy, Evie Davalbhakta, and Cassie S. Mitchell. 2022. Optimizations for computing relatedness in biomedical heterogeneous information networks: Semnet 2.0. *Big Data and Cognitive Computing*, 6(1).

Z. Kusynová, G.M. Pauletti, H.A van den Ham, H.G.M. Leufkens, and A.K. Mantel-Teeuwisse. 2022. Unmet medical need as a driver for pharmaceutical sciences – a survey among scientists. *Journal of Pharmaceutical Sciences*, 111(5):1318–1324.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. *Preprint*, arXiv:2309.06180.

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:baw068.

Mistral. 2025. Mistral. Mistral website to mistral-small-3-1 model.

Sunil Mohan, Rico Angell, Nicholas Monath, and Andrew McCallum. 2021. Low resource recognition and linking of biomedical concepts from a large ontology. In *Proceedings of the 12th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, BCB '21, page 1–10. ACM.

Sunil Mohan and Donghui Li. 2019. Medmentions: A large biomedical corpus annotated with umls concepts. *Preprint*, arXiv:1902.09476.

Milad Moradi, Kathrin Blagec, and Matthias Samwald. 2021. Deep learning models are not robust against noise in clinical text. *Preprint*, arXiv:2108.12242.

Monica Munnangi, Sergey Feldman, Byron C Wallace, Silvio Amir, Tom Hope, and Aakanksha Naik. 2024. On-the-fly definition augmentation of llms for biomedical ner. *Preprint*, arXiv:2404.00152.

Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain Marshall, Ani Nenkova, and Byron Wallace. 2018. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 197–207, Melbourne, Australia. Association for Computational Linguistics.

Aurélie Névéol, Hercules Dalianis, Sumithra Velupillai, Guergana Savova, and Pierre Zweigenbaum. 2018. Clinical natural language processing in languages other than english: Opportunities and challenges. *Journal of biomedical semantics*, 9:12.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, and more OpenAI team. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Jacqueline Peng, Mengge Zhao, James Havrilla, Cong Liu, Chunhua Weng, Whitney Guthrie, Robert Schultz, Kai Wang, and Yunyun Zhou. 2020. Natural language processing (nlp) tools in extracting biomedical concepts from research articles: a case study on autism spectrum disorder. *BMC Medical Informatics and Decision Making*, 20.

PubMed. 2025. Pubmed. Website of pubmed.

John Schulman and Thinking Machines Lab. 2025. Lora without regret. *Thinking Machines Lab: Connectionism*. Https://thinkingmachines.ai/blog/lora/.

Sonish Sivarajkumar and Yanshan Wang. 2022. Healthprompt: A zero-shot learning paradigm for clinical natural language processing. *Preprint*, arXiv:2203.05061.

S Tian, Q Jin, L Yeganova, et al. 2023. Opportunities and challenges for chatgpt and large language models in biomedicine and health. arxiv. *arXiv preprint arXiv:2306.10070*.

Hugo Touvron, Louis Martin, Kevin Stone, and Peter Albert etc... 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Marceli Wac, Raul Santos-Rodriguez, Chris McWilliams, and Christopher Bourdeaux. 2024. Capturing requirements for a data annotation tool for intensive care: Experimental user-centered design study. *Preprint*, arXiv:2309.16500.

Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, and Hongfang Liu. 2018. Clinical information extraction applications: A literature review. *Journal of Biomedical Informatics*, 77:34–49.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Preprint*, arXiv:2412.13663.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. *Preprint*, arXiv:2109.01652.

Yishu Wei, Yu Deng, Cong Sun, Mingquan Lin, Hongmei Jiang, and Yifan Peng. 2024. Deep learning with noisy labels in medical prediction problems: a scoping review. *Preprint*, arXiv:2403.13111.

Olivier J Wouters, Martin McKee, and Jeroen Luyten. 2020. Estimated research and development investment needed to bring a new medicine to market, 2009-2018. *Jama*, 323(9):844–853.

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

A Appendix

The appendix is organized into three parts:

- Implementation details for BERT-based models and LLMs are provided in Appendix A.1.
- Further discussion on LLMs is included in Appendix A.2.
- Example annotations for the different models are presented in Appendix A.3.

A.1 Training and Inference Details for BERT-Based Models and LLMs

A.1.1 Additional information about the evaluated LLMs

Table 10 provides additional details on the LLMs evaluated in this paper, including their architectures, pre-training data, and other relevant information when available.

A.1.2 BERT Training parameters

The parameters used for the training of all biomedical BERT models are provided in Table 11.

For the training in low-data setting experiments in Section 4.3, we adjusted the learning rate to $1e^{-4}$ and set the number of epochs to 200, 100 and 50 for datasets containing 20, 40 and 100 abstracts, respectively .

A.1.3 LLM prompt

Figure 4 displays the exact prompt used to generate the results reported in Section 4.2 using LLMs.

Arguments enclosed in *italics* are filled dynamically when constructing the prompt. Each argument is described below:

- `extraction_prompt`: A task-specific instruction describing what entities to extract. The exact instructions for each dataset is given in Figure 6
- `abstract`: The specific abstract to be annotated.
- `examples`: One or more examples of abstracts with their corresponding correct annotations.

A.1.4 LLM decoding and fine-tuning parameters

The LLM were used with temperature set to 0.

The fine-tuned results of Mistral-3.1 reported in Section 4.2.1 were obtained using either the original LoRA approach (Hu et al., 2021) or LoRA applied to all layers (Schulman and Lab, 2025), with the reported results corresponding to the configuration that achieved the highest validation accuracy, using the following parameter settings :

- Rank: 128
- Scaling factor α: 32
- Dropout: 0.05

Note that we evaluated multiple values of the rank parameter (32, 64, 128, 192) and scaling factor (16, 32), and found the optimal configuration to be rank: 128 / scaling factor: 32 / dropout: 0.05.

The number of epochs and batch size varied across datasets; their values along with the training time are listed in Table 12. For reference, it took 50 minutes to train Biolink-BERT on the largest evaluated dataset MM-ST21PV (2635 abstracts in training set).

Due to time and resource constraints, validation during training was performed on only the first 40 abstracts of the validation set, which may have limited the model's performance.

Different strategies to optimize performance of fine-tuning were also explored by varying the number of examples included in the prompt during both training and inference on the MM-ST21PV dataset. Specifically, we experimented with 0, 1, or 3 examples in both stages. The results are presented in Table 13.

Results indicate that incorporating examples into the training prompts improves learning, with performance increasing as more examples are provided. In contrast to the base model, the fine-tuned version is more robust to reductions in the number of examples during inference, with only a modest performance drop. However, including more examples at inference than were seen during training hurts performance significantly; likely due to a mismatch between the prompt structure the model was trained on and the one used at test time.

A.1.5 Inference time for LLMs

Figure 5 presents the inference time of all evaluated open-source LLMs across the different datasets.

Mistral-3.1, as the largest model, has the highest computational demand and requires the longest runtime. Running it on the whole MM-ST21PV test set (879 abstracts) took close to 9 hours. This is significantly longer than the time required to train and run inference with BERT-based models, which takes less than an hour in this case. As a result, LLMs are currently not only less effective than domain-specific BERT models for biomedical NER but also considerably slower.

All experiments were run on an Nvidia A40 GPU using vllm framework (Kwon et al., 2023).

	Parameters	Context Length	Pre-Training data	Architectue highlights
Meta-Llama-3.1-8B-Instruct	8B	128K	15T+ (50% general knowledge web, 25% math/reasoning, 17% code, 8% multilingual)	Dense model, SwiGLU activation, GQA and long-context RoPE
Phi-4	14B	16384	10T (30% web + web-rewrites (15% each), 40% synthetic, 20% code, 10% academic sources & books)	Dense model, Full attention
gemma-3-12B-it	12B	128K	12T (web text, code, math, images, and multilingual data)	Dense model, global + local (1024 tokens) attention, RoPE
Mistral-Small-3.1-24B-Instruct-2503	24B	128K	N/A	Dense model, Sliding window attention, RoPE

Table 10: Architecture and Training data comparison of evaluated open source-models. N/A : Not Available

System Instruction
You are a medical doctor who specializes in clinical trials and observational studies.
You will act as an expert annotator of research articles provided to you. Only answer using data explicitly present in given studies.

User
You need to labels ALL text that fall in of these categories:
{*extraction_prompts*}

This is the specific abstract that you need to annotate:
{*abstract*}

Entities must be tagged in this format : "@@entity##entity_type@@"

To assist you, here are a few examples of correctly tagged abstracts:
{*examples*}

Your final output must match the original abstract precisely.
Be sure that any errors such as duplicating sentences, omitting text, or introducing hallucinations are corrected before the final output.

Return the answer in a JSON format.
An example of valid answer look like this :
Final Output
```json
{"output": text ... @@entity##entity_type@@ ... text}
```

Figure 4: Illustration of annotation framework

Hyperparameters	
seq_len	512
batch_size	16
num_epochs	30
learning_rate	$5e^{-4}$
weight_decay	0.01
warmup_ratio	0.2

Table 11: Hyperparameters used to train all biomedical BERT models

	NCBI-Disease	BC5CDR	MM-ST21PV
batch_size	8	3	2
num_epochs	30	15	5
train_time (hours)	4.3	7.5	22.2

Table 12: Batch size, number of epochs and training time with one H100 GPU to fine-tune Mistral-3.1

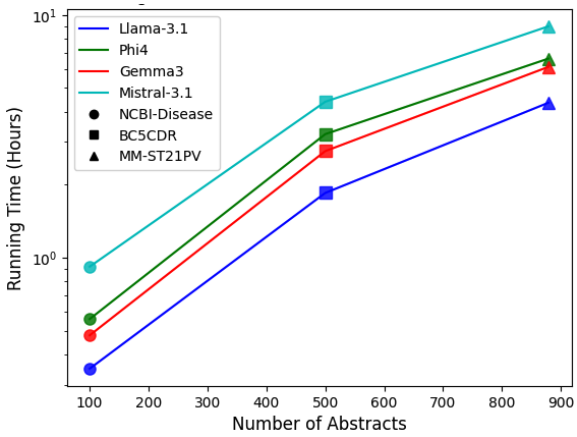


Figure 5: Inference time vs Number of abstracts for all evaluated open-source models

A.1.6 Statistical significance testing of LLM progress over time

The model used for our statistical significance test is a fixed-effects (within-dataset) OLS regression using HC3 standard errors, which are robust in small-sample panel settings.

The regression data include a total of 20 datapoints consisting of the 5 evaluated datasets, each evaluated at 4 different timepoints [1) July 2024 corresponding to Llama3.1 release, 2) November 2024 corresponding to Phi3 release, 3) March 2025 corresponding to Gemma3 release and 4) April 2025 corresponding to Mistral3 Small release]. For each dataset, the evaluated variable is the performance gap between the best biomedical BERT vari-

	k = 0			k = 1			k = 3		
Metrics	F1	P	R	F1	P	R	F1	P	R
l = 0	0.553	0.558	0.547	0.112	0.250	0.072		-	
l = 1	0.579	0.574	0.576	0.568	0.558	0.578	0.268	0.330	0.226
l = 3	0.543	0.597	0.497	0.583	0.588	0.579	0.580	0.564	0.596

Table 13: F1, Precision, and Recall scores for Mistral-3.1, fine-tuned with different numbers of in-prompt examples and evaluated using varying numbers of examples during inference on MM-ST21PV dataset.
k : Number of examples in the prompt during inference
l : Number of examples in the prompt during training

Class of models	β	**SE**	**t**	**p-value**
Evaluated open-source models	$-1.4 \cdot 10^{-2}$	$3 \cdot 10^{-3}$	-4.8	0.0

Table 14: Statistical significance testing of LLM progress on all evaluated datasets: Negative regression coefficient reveals that successively released open-source models (Llama 3.1 in July 2024, Phi-4 in December 2024, Gemma3 in March 2025, Mistral 3.1 in April 2025) progressively close the performance gap relative to the best biomedical BERT-Based variant.

ant and the best evaluated open-source LLM at that same timestamp.

A.1.7 Statistical significance tests for noise robustness of BERT training and LLMs ICL

Noise Source	β	**SE**	**t**	**p-value**
Gemma-3	$3.1 \cdot 10^{-3}$	$3.0 \cdot 10^{-4}$	$1.0 \cdot 10^{1}$	$1.4 \cdot 10^{-6}$
Phi-4	$2.2 \cdot 10^{-3}$	$2.7 \cdot 10^{-4}$	7.9	$1.4 \cdot 10^{-5}$
GPT-4o	$1.5 \cdot 10^{-3}$	$4.0 \cdot 10^{-4}$	3.8	$5.3 \cdot 10^{-3}$

Table 15: Statistical Comparison of Noise Robustness: Mistral-3.1 ICL and Biolink-BERT training across different noise source.

Noise Source	β	**SE**	**t**	**p-value**
Gemma-3	$3.2 \cdot 10^{-3}$	$2.5 \cdot 10^{-4}$	$1.3 \cdot 10^{1}$	$1.8 \cdot 10^{-7}$
Phi-4	$2.5 \cdot 10^{-3}$	$2.6 \cdot 10^{-4}$	9.4	$2.8 \cdot 10^{-6}$
Mistral-3.1	$2.9 \cdot 10^{-3}$	$1.6 \cdot 10^{-4}$	$1.8 \cdot 10^{1}$	$8.8 \cdot 10^{-8}$

Table 16: Statistical Comparison of Noise Robustness: GPT-4o ICL and Biolink-BERT training across different noise source.

In Table 15 and 16, we report the results of regression-based statistical tests comparing the robustness of Mistral-3.1 and GPT-4o using noisy ICL against BiolinkBERT trained on noisy data across different sources of noise. The interaction term between model and noise level provides the test statistic, with the corresponding β, SE, t, and p-values indicating whether the difference in performance degradation slopes is statistically significant.

Across all noise sources, the interaction term is highly significant, indicating that LLMs exhibits a substantially slower performance decline as noise increases. This provides statistical evidence that noisy ICL examples for LLMs is more robust than noisy examples for BERT training.

A.2 Additional discussion on LLMs

A.2.1 Bigger vs. Smaller: How Model Size Affects LLMs in the Same Family

The results from Section 4 suggest that larger models tend to perform better overall, which is generally true but only to a certain extent. Table 17 presents a direct comparison on NCBI-Disease dataset between two models of the same generation but different sizes: Gemma3-12B-it and Gemma3-27B-it.

	Gemma3-12B-it				Gemma3-27B-it			
k	F1	P	R	Inv.	F1	P	R	Inv.
1	0.580	0.551	0.613	7%	0.583	0.529	0.650	0%
3	0.639	0.612	0.670	2%	0.635	0.575	0.710	0%
5	0.621	0.620	0.622	1%	0.642	0.593	0.700	0%
10	0.653	0.701	0.611	9%	0.654	0.620	0.692	0%

Table 17: Comparison F1, Precision, and Recall scores between Gemma3-12B and Gemma3-27B using varying numbers of examples during inference on NCBI-Dataset. Results are based on a single run. "Inv." (Invalid evaluations) represents the % of abstracts that could not be assessed due to formatting errors or failure to follow prompt instructions.

These results indicate that model size has an impact, but not directly, on the quality of annotations, but rather on following instructions. Indeed, increasing the model size from 12B to 27B parameters did not really improve the results but significantly reduced formatting errors during the

annotation process, with the larger model adhering to the expected format perfectly. However, when it comes to the quality of the annotations themselves, the most critical factor remains the model inherent capability to handle the task. For example, although Gemma3-27B-it is larger than the Mistral-3.1 model used in our experiments, it still underperforms in comparison. If a model is fundamentally not suited for biomedical NER, increasing its size, even doubling it, will not lead to substantial gains, as it lacks the core ability to effectively perform the task.

A.2.2 Evaluating Potential Data Leakage in Mistral-3.1

Given the strong performance of the Mistral-3.1 model, we considered the possibility that it might have been exposed to test examples during training. To investigate this, we ran a controlled experiment in which we fine-tuned Mistral on the training set while using test set examples as ICL examples (without including those examples in the training loss). Under this setup, performance on MM-ST21PV increased from 58% (reported in our main results) to 67%. This 10% increase indicates that even limited exposure to test data through ICL, without directly training on it, can yield substantial gains. Given this sensitivity, we believe it is unlikely that Mistral-3.1 had seen the test examples during pre-training.

A.3 Details on LLM outputs

A.3.1 Comparison of annotations from BERT, Mistral with gold annotations and with noisy annoataions

Annotation results for a sample abstract (PMID: 27059693) from MM-ST21PV are shown under four setups to highlight their tendencies :

- Figure 7: Annotations produced by BERT.
  F1 = 0.697, P = 0.697, R = 0.697
- Figure 8: Annotations generated by Mistral using in-context examples sourced from gold data.
  F1 = 0.600, P = 0.568, R = 0.636
- Figure 9: Annotations generated by Mistral using in-context examples sourced from Gemma3, where 50% of the annotations are noisy.
  F1 = 0.582, P = 0.727, R = 0.485
- Figure 10: Annotations generated by Mistral using in-context examples sourced from GPT-4o, where 50% of the annotations are noisy.
  F1 = 0.576, P = 0.576, R = 0.576

These results are consistent with the findings discussed in Sections 4 and 5. Among the models, BERT achieves the strongest annotation performance overall, with its errors primarily concentrated in three categories: "Invalid Tag", "Missing", and "Wrong Overlap".

For Mistral, the key result is that performance remains very similar regardless of whether the in-context examples are drawn from gold annotations or from noisy annotations such as those provided by GPT-4o. In terms of aggregate metrics, the presence of noise in the examples does not substantially alter the model's effectiveness.

A secondary observation concerns the distribution of errors. When the noisy examples resemble the types of mistakes Mistral would typically make (e.g. GPT-4o-based annotations), the error profile remains quite stable. However, when the noisy examples come from a source with a different error distribution such as Gemma3, whose annotations are dominated by "Missing" errors, Mistral behavior shifts accordingly. In this case, it adopts a more conservative annotation strategy. This results in reduced recall but an increase in precision, leading to an overall F1 score that remains comparable to the setup without noisy annotations.

This analysis suggests that while the overall performance of Mistral is robust to noise in in-context examples, the nature of the noise can still shape the error patterns produced.

A.3.2 Examples of invalid outputs

The LLMs failed to consistently follow the output requirements, resulting in a significant number of invalid outputs especially for higher number of in-context examples.

Common errors include the generation of invalid JSON, no annotations, the placement of annotations at the end of the abstract rather than within the text, and hallucinations in which the model simply repeats the abstract without introducing any annotations. Figure 11 illustrates an example where annotations were incorrectly appended at the end of the text; in addition, the model began annotating only from the second sentence.

We attempted to mitigate these errors by placing stronger emphasis on the annotation rules; however, this did not result in any noticeable improvement.

For example, we explicitly instructed the model to:

- Not rewrite, summarize, or rephrase the text.

- Preserve the text exactly as provided, including line breaks, typos, and formatting.
- Only insert annotations in the format @@mention##ENTITY@@ into the original text.
- Ensure the output is identical to the input text, apart from these inserted annotations.
- Annotate both title and abstract.

Paradoxically, introducing these additional rules seemed to decrease the model performance. When Llama 3.1–8B ($k = 10$) was evaluated on 40 abstracts, the number of valid outputs declined from 33 with the prompt shown in Figure 1 to 23 with the extended set of rules.

A.3.3 Error analysis on noisy evaluations

Tables 19 and 20 summarize error type statistics for GPT-4o and Mistral-3.1 in-context learning on the MM-ST21PV dataset under 30% and 50% noisy examples, respectively.

MM-ST21PV

"T058": "Healthcare Activity : Return the name of the health care activity described.
Example usage: in 'a pilot study of an evidence-based psychological intervention', return 'intervention'."

"T062": "Research Activity : Return the name of the research activity mentioned.
Example usage: in 'By using exome sequencing and extreme phenotype design', return 'exome sequencing'."

"T037": "Injury or Poisoning : Return the name of the injury or poisoning mentioned.
Example usage: in 'and their toxic effects on aquatic species have been reported', return 'toxic effects'."

"T038": "Biologic Function : Return the name of the biologic function described. E
Example usage: in 'DCTN4 as a modifier of chronic Pseudomonas aeruginosa infection in cystic fibrosis',
return 'chronic Pseudomonas aeruginosa infection' and 'cystic fibrosis'."

"T005": "Virus : Return the name of the virus mentioned.
Example usage: in 'Flaviviruses, including Zika and dengue (DENV), pose a serious global threat to human health.',
return 'Flaviviruses', 'Zika', 'dengue' and 'DENV'.",

"T007": "Bacterium : Return the name of the bacterium mentioned.
Example usage: in 'for the identification of lactic acid bacteria', return 'lactic acid bacteria'."

"T204": "Eukaryote : Return the name of the eukaryotic organism mentioned.
Example usage: in 'as an inert carrier was investigated against Sitophilus oryzae', return 'Sitophilus oryzae'."

"T017": "Anatomical Structure : Return the name of the anatomical structure mentioned.
Example usage: in 'Polymerase chain reaction and direct sequencing were used to screen DNA samples for DCTN4 variants.',
return 'DNA samples'."

"T074": "Medical_device : Return the name of the medical device mentioned.
Example usage: in 'important future components for bionanoelectronic devices', return 'bionanoelectronic devices'."

"T031": "Body_substance : Return the name of the body substance mentioned.
Example usage: in 'Apoptosis has been shown to be induced by serum deprivation or copper treatment.', return 'serum'."

"T103": "Chemical : Return the name of the chemical mentioned.
Example usage: in 'DCTN4 as a modifier of chronic Pseudomonas aeruginosa infection in cystic fibrosis', return 'DCTN4'."

"T168": "Food : Return the name of the food mentioned.
Example usage: in 'with no detrimental effects on grain quality', return 'grain'.",

"T201": "Clinical Attribute : Return the name of the clinical attribute described.
Example usage: in 'Anthropometric measurements, including height, weight, waist circumference', return 'waist circumference'."

"T033": "Finding : Return the name of the finding mentioned.
Example usage: in 'chronic Pa infection (CPA) is associated with reduced lung function, faster rate of lung decline',
return 'faster rate of lung decline'."

"T082": "Spatial Concept : Return the name of the spatial concept mentioned.
Example usage: in 'The difference in structure of the two compounds', return 'structure'."

"T022": "Body system : Return the name of the body system mentioned.
Example usage: in 'reduce the potential risk of skin ulceration', return 'skin'."

"T091": "Biomedical Occupation or Discipline : Return the name of the biomedical occupation or discipline mentioned.
Example usage: in 'its symptoms are broad and place patients at crossroads between dermatology, hematology',
return 'dermatology' and 'hematology'."

"T092": "Organization : Return the name of the organization mentioned.
Example usage: in 'in a cohort of adult CF patients from a single centre', return 'centre'."

"T097": "Professional or Occupational Group : Return the name of the professional or occupational group mentioned.
Example usage: in 'Delivered by a health psychologist', return 'health psychologist'."

"T098": "Population Group : Return the name of the population group mentioned.
Example usage: in 'in a cohort of adult CF patients from a single centre', return 'cohort'."

"T170": "Intellectual Product : Return the name of the intellectual product mentioned.
Example usage: in 'Intervention using effective components of behaviour change informed by psychological theory',
return 'psychological theory'.",

NCBI-Disease

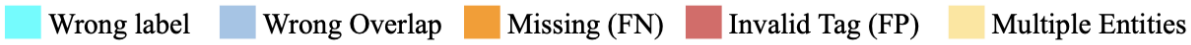
"Disease": "Return the text that is related to disease.
Example usage: in 'Suxamethonium infusion rate and observed fasciculations.', return 'fasciculations'.",

BC5CDR

"Chemical": "Return the text that is related to chemical.
Example usage: in 'Suxamethonium infusion rate and observed fasciculations.', return 'Suxamethonium'."

"Disease": "Return the text that is related to disease.
Example usage: in 'Suxamethonium infusion rate and observed fasciculations.', return 'fasciculations'.",

Figure 6: Extraction instruction for each evaluated datasets



Title : Patient - @@Physician##T097@@ Discordance in @@Global Assessment##T058@@ in @@Rheumatoid Arthritis##T038@@: A @@Systematic Literature Review##T170@@ With @@Meta-Analysis##T062@@

Abstract : The integration of the patient in therapeutic @@decision-making##T038@@ is important in the @@management##T058@@ of @@rheumatoid arthritis##T038@@ (@@RA##T038@@), but the patient opinion regarding disease status may differ from the @@physician's##T097@@ opinion. The aim of this study was to assess in the @@published literature##T170@@ the frequency and drivers of patient - @@physician##T097@@ discordance in @@global assessment##T058@@ in @@RA##T038@@. A @@systematic literature review##T170@@ of all @@articles##T170@@ published up to January 2015 in @@Medline##T170@@ or @@Embase##T170@@, reporting discordance in @@RA##T038@@, was conducted by 2 @@investigators##T097@@. Discordance was defined based on the absolute difference of @@patient global##T170@@ (@@PGA##T170@@) and @@physician global assessments##T170@@ (@@PhGA##T170@@) on 0-10-cm @@scales##T170@@. The frequency of discordance and its predictors were collected in each @@study##T062@@. Frequencies of discordance were pooled by @@meta-analysis##T062@@ using random effect. In all, 12 @@studies##T062@@ were selected (i.e., 11,879 patients): weighted mean ± SD age was 55.1 ± 13.9 years, weighted mean ± SD disease duration was 10.4 ± 9.3 years, and 80.7% were @@women##T098@@. The value of the difference | @@PGA##T170@@ - @@PhGA##T170@@ | defining discordance varied between ≥0.5 cm (n = 2 @@studies##T062@@) to ≥3 cm (n = 5 @@studies##T062@@); the weighted mean value was 2.7 cm. The pooled percentage of patients with discordance was 43% (95% confidence interval 36%-51%; range 25%-76%). PGA was usually higher than PhGA. The drivers of PGA were pain and functional incapacity, whereas drivers of PhGA were joint counts and acute-phase reactants. Discordance in global assessment was most frequently defined as a difference of 3 points or more; even with such a stringent definition, up to half the patients were found to be discordant. The long-term consequences of this discordance remain to be determined.

Figure 7: Example Mistral annotations with gold data

Wrong label | Wrong Overlap | Missing (FN) | Invalid Tag (FP) | Multiple Entities

Title : Patient - Physician Discordance in Global Assessment in @@Rheumatoid Arthritis##T038@@: A @@Systematic Literature Review##T170@@ With @@Meta-Analysis##T062@@

Abstract : The integration of the patient in @@therapeutic decision-making##T058@@ is important in the @@management##T058@@ of @@rheumatoid arthritis##T038@@ (@@RA##T038@@), but the patient opinion regarding @@disease##T038@@ status may differ from the @@physician's##T097@@ opinion. The aim of this @@study##T062@@ was to assess in the published @@literature##T170@@ the frequency and drivers of patient - @@physician##T097@@ discordance in global assessment in @@RA##T038@@. A systematic @@literature review##T170@@ of all articles published up to January 2015 in @@Medline##T170@@ or @@Embase##T170@@, reporting discordance in @@RA##T038@@, was conducted by 2 @@investigators##T097@@. Discordance was defined based on the absolute difference of patient global (@@PGA##T033@@) and @@physician##T097@@ global assessments (@@PhGA##T033@@) on 0-10-cm scales. The frequency of discordance and its predictors were collected in each @@study##T062@@. Frequencies of discordance were pooled by @@meta-analysis##T062@@ using random effect. In all, 12 @@studies##T062@@ were selected (i.e., 11,879 patients): weighted mean \u00b1 SD age was 55.1 \u00b1 13.9 years, weighted mean \u00b1 SD @@disease##T038@@ duration was 10.4 \u00b1 9.3 years, and 80.7% were @@women##T098@@. The value of the difference | @@PGA##T033@@ - @@PhGA##T033@@ | defining discordance varied between \u22650.5 cm (n = 2 @@studies##T062@@) to \u22653 cm (n = 5 @@studies##T062@@); the weighted mean value was 2.7 cm. The pooled percentage of patients with discordance was 43% (95% confidence interval 36%-51%; range 25%-76%). @@PGA##T033@@ was usually higher than @@PhGA##T033@@. The drivers of @@PGA##T033@@ were @@pain##T033@@ and functional incapacity, whereas drivers of @@PhGA##T033@@ were @@joint counts##T033@@ and @@acute-phase reactants##T103@@. Discordance in global assessment was most frequently defined as a difference of 3 points or more; even with such a stringent definition, up to half the patients were found to be discordant. The long-term consequences of this discordance remain to be determined.

Figure 8: Example Mistral annotations with in-context examples sourced from gold data

Wrong label Wrong Overlap Missing (FN) Invalid Tag (FP) Multiple Entities

Title : Patient - Physician Discordance in Global Assessment in Rheumatoid Arthritis: A Systematic Literature Review With Meta-Analysis

Abstract : The integration of the patient in therapeutic decision-making is important in the management of @@rheumatoid arthritis##T038@@ (@@RA##T038@@), but the patient opinion regarding disease status may differ from the physician's opinion. The aim of this @@study##T062@@ was to assess in the published literature the frequency and drivers of patient - physician discordance in global assessment in @@RA##T038@@. A systematic @@literature review##T170@@ of all articles published up to January 2015 in @@Medline##T170@@ or @@Embase##T170@@, reporting discordance in @@RA##T038@@, was conducted by 2 @@investigators##T097@@. Discordance was defined based on the absolute difference of patient global (@@PGA##T170@@) and physician global assessments (@@PhGA##T170@@) on 0-10-cm scales. The frequency of discordance and its predictors were collected in each @@study##T062@@. Frequencies of discordance were pooled by @@meta-analysis##T062@@ using random effect. In all, 12 @@studies##T062@@ were selected (i.e., 11,879 patients): weighted mean ± SD age was 55.1 ± 13.9 years, weighted mean ± SD disease duration was 10.4 ± 9.3 years, and 80.7% were women. The value of the difference | @@PGA##T170@@ - @@PhGA##T170@@ | defining discordance varied between ≥0.5 cm (n = 2 @@studies##T062@@) to ≥3 cm (n = 5 @@studies##T062@@); the weighted mean value was 2.7 cm. The pooled percentage of patients with discordance was 43% (95% confidence interval 36%-51%; range 25%-76%). @@PGA##T170@@ was usually higher than @@PhGA##T170@@. The drivers of @@PGA##T170@@ were pain and functional incapacity, whereas drivers of @@PhGA##T170@@ were joint counts and acute-phase reactants. Discordance in global assessment was most frequently defined as a difference of 3 points or more; even with such a stringent definition, up to half the patients were found to be discordant. The long-term consequences of this discordance remain to be determined.

Figure 9: Example of Mistral annotations with in-context examples sourced from Gemma3 where 50% of the annotations are noisy.

Wrong label | Wrong Overlap | Missing (FN) | Invalid Tag (FP) | Multiple Entities

Title : Patient - Physician Discordance in Global Assessment in @@Rheumatoid Arthritis##T038@@: A Systematic @@Literature Review##T170@@ With @@Meta-Analysis##T062@@

Abstract : The integration of the patient in therapeutic @@decision-making##T058@@ is important in the management of @@rheumatoid arthritis##T038@@ (@@RA##T038@@), but the patient opinion regarding @@disease##T038@@ status may differ from the physician\'s opinion. The aim of this @@study##T062@@ was to assess in the published @@literature##T170@@ the frequency and drivers of patient - physician discordance in global assessment in @@RA##T038@@. A systematic @@literature review##T170@@ of all articles published up to January 2015 in @@Medline##T170@@ or @@Embase##T170@@, reporting discordance in @@RA##T038@@, was conducted by 2 @@investigators##T097@@. Discordance was defined based on the absolute difference of patient global (@@PGA##T033@@) and physician global assessments (@@PhGA##T033@@) on 0-10-cm scales. The frequency of discordance and its predictors were collected in each @@study##T062@@. Frequencies of discordance were pooled by @@meta-analysis##T062@@ using random effect. In all, 12 @@studies##T062@@ were selected (i.e., 11,879 patients): weighted mean ± SD age was 55.1 ± 13.9 years, weighted mean ± SD @@disease##T038@@ duration was 10.4 ± 9.3 years, and 80.7% were @@women##T098@@. The value of the difference | @@PGA##T033@@ - @@PhGA##T033@@ | defining discordance varied between ≥0.5 cm (n = 2 @@studies##T062@@) to ≥3 cm (n = 5 @@studies##T062@@); the weighted mean value was 2.7 cm. The pooled percentage of patients with discordance was 43% (95% confidence interval 36%-51%; range 25%-76%). @@PGA##T033@@ was usually higher than @@PhGA##T033@@. The drivers of @@PGA##T033@@ were @@pain##T033@@ and functional incapacity, whereas drivers of @@PhGA##T033@@ were @@joint counts##T033@@ and @@acute-phase reactants##T103@@. Discordance in global assessment was most frequently defined as a difference of 3 points or more; even with such a stringent definition, up to half the patients were found to be discordant. The long-term consequences of this discordance remain to be determined.

Figure 10: Example of Mistral annotations with in-context examples sourced from GPT-4o where 50% of the annotations are noisy.

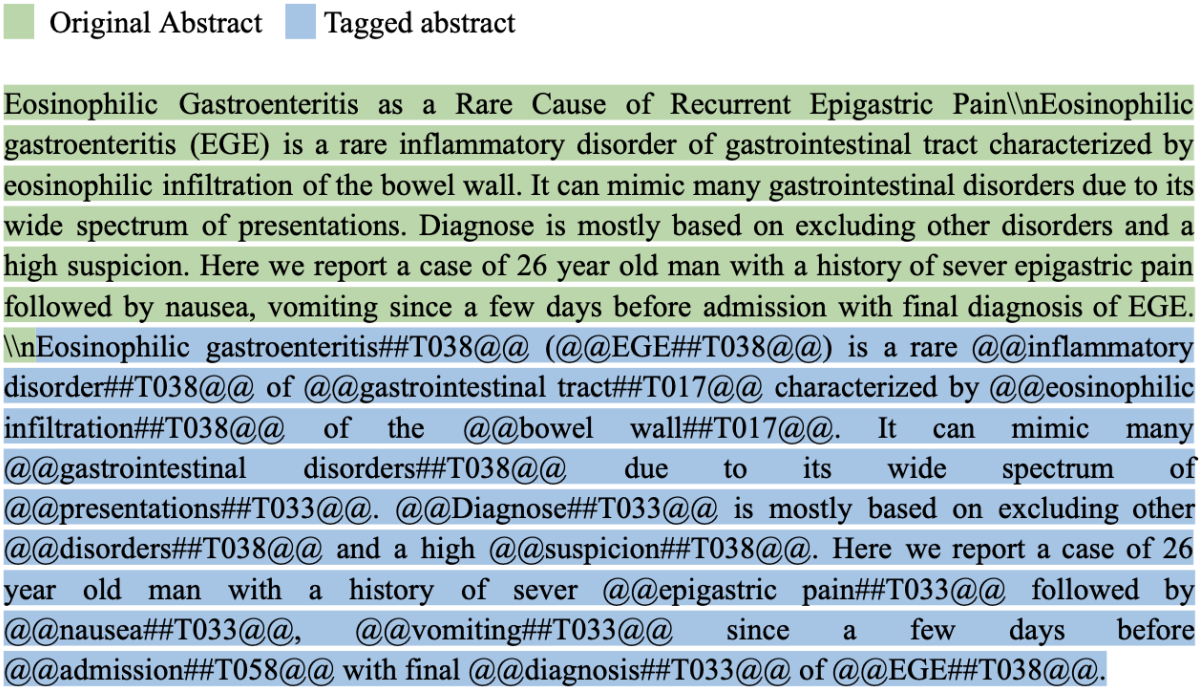


Figure 11: Example of an annotation generated by Llama-3.1-8B that failed to comply with the guidelines. Instead of directly annotating the text, the model rewrote the abstract and then added an annotated version, starting the annotation only from the second sentence. (PMID=27274524)

Datasets (→)		NCBI-Disease				BC5CDR				MM-ST21PV			
Models (↓)		**k=1**	**k=3**	**k=5**	**k=10**	**k=1**	**k=3**	**k=5**	**k=10**	**k=1**	**k=3**	**k=5**	**k=10**
Llama-3.1	F1	0.489	0.513	0.554	0.551	0.543	0.580	0.610	0.598	0.239	0.305	0.314	0.287
		±3e-5	±2e-6	±8e-6	±2e-5	±3e-6	±9e-7	±2e-6	±5e-6	±3e-6	±8e-5	±1e-4	±4e-5
	P	0.518	0.622	0.697	0.677	0.720	0.761	0.774	0.793	0.340	0.424	0.454	0.500
		±3e-4	±4e-4	±2e-7	±3e-4	±4e-5	±1e-5	±4e-5	±8e-5	±1e-5	±5e-6	±5e-6	±1e-5
	R	0.476	0.436	0.460	0.464	0.435	0.468	0.503	0.479	0.184	0.239	0.241	0.199
		±4e-5	±1e-4	±2e-5	±2e-6	±1e-5	±2e-7	±9e-7	±4e-5	±1e-5	±1e-4	±1e-4	±6e-5
	Inv.	4.0%	8.7%	10.3%	29.3%	2.9%	7.9%	9.9%	24.1%	10.9%	8.2%	12.2%	23.7%
		±0.0%	±8e-3%	±1.4%	±0.6%	±2e-3%	±3e-4%	±5e-3%	±2e-2%	±0.0%	±0.0%	±1e-6%	±7e-4%
Phi-4	F1	0.536	0.548	0.562	0.542	0.633	0.652	0.649	0.645	0.259	0.302	0.311	–
		±6e-5	±1e-4	±2e-7	±5e-5	±8e-6	±5e-6	±5e-6	±2e-6	±2e-7	±8e-7	±8e-7	
	P	0.734	0.727	0.732	0.750	0.784	0.779	0.798	0.797	0.398	0.444	0.465	–
		±6e-5	±1e-4	±7e-5	±8e-5	±2e-7	±8e-7	±8e-7	±8e-7	±9e-7	0.0	0.0	
	R	0.422	0.440	0.456	0.424	0.530	0.560	0.547	0.542	0.192	0.228	0.234	–
		±6e-6	±1e-4	±1e-4	±3e-5	±1e-5	±1e-5	±8e-6	±3e-6	±2e-7	±8e-7	±8e-7	
	Inv.	4.7%	2.7%	3.3%	1.7%	0.6%	0.0%	0.4%	0.05%	8.0%	7.6%	9.3%	–
		±2e-3%	±2e-3%	±8e-3%	±8e-3%	±0.0%	±0.0%	±0.0%	±1e-3%	±3e-5%	±4e-4%	±1.3%	
Gemma-3	F1	0.565	0.620	0.616	0.629	0.693	0.718	0.723	0.718	0.336	0.374	0.364	0.329
		±1e-4	±2e-4	±1e-5	±2e-4	±8e-6	±2e-6	±3e-6	±8e-6	±2e-7	±1e-5	±8e-5	±2e-4
	P	0.544	0.597	0.627	0.699	0.679	0.720	**0.763**	**0.762**	0.372	0.463	0.502	0.537
		±3e-5	±1e-4	±3e-5	±3e-6	±6e-5	±3e-5	±7e-5	±6e-7	±2e-7	±3e-6	±4e-5	±4e-5
	R	0.590	0.645	0.604	0.579	0.708	0.716	0.678	0.686	0.308	0.314	0.285	0.237
		±3e-4	±3e-4	±1e-4	±4e-4	±8e-6	±6e-5	±1e-4	±3e-5	±8e-7	±3e-5	±8e-5	±2e-4
	Inv.	6.3%	1.3%	1.0%	9.0%	1.9%	0.5%	1.4%	2.7%	5.0%	14.1%	21.0%	47.9%
		2e-3%	2e-3%	0.0%	0.0%	±8e-5%	±3e-4%	±3e-5%	±1e-3%	±5e-3%	±2e-2%	±5.0%	±3.9%
Mistral-3.1	F1	0.648	0.703	0.721	0.724	0.725	0.745	0.760	**0.770**	**0.399**	**0.469**	**0.496**	**0.509**
		±5e-6	±2e-5	±3e-6	±8e-6	±2e-7	±1e-5	±2e-6	±2e-7	±0.0	±5e-6	±2e-7	±9e-7
	P	0.651	0.669	0.672	0.684	0.708	0.713	0.733	0.738	**0.372**	0.444	0.480	0.504
		±3e-3	±4e-5	±3e-5	±3e-5	0.0	±1e-5	0.0	±5e-6	±0.0	±3e-5	±2e-5	±2e-5
	R	**0.666**	**0.742**	**0.778**	**0.770**	**0.742**	**0.780**	**0.789**	**0.805**	**0.430**	**0.497**	**0.516**	**0.511**
		±2e-7	±5e-5	±2e-6	±8e-7	±2e-7	±8e-6	±1e-5	±1e-5	±0.0	±5e-6	±1e-5	±5e-6
	Inv.	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.2%	0.2%	0.4%	2.8%
		±0.0%	±0.0%	±0.0%	±0.0%	±0.0%	±0.0%	±0.0%	±0.0%	±0.0%	±1e-6%	±4e-4%	±4e-3%
GPT-4o	F1	**0.670**	**0.717**	**0.730**	**0.745**	**0.737**	**0.755**	**0.767**	0.764	0.363	0.413	0.440	0.457
		±4e-5	±3e-4	±6e-5	±1e-4	±4e-5	±6e-7	±4e-6	±3e-6	±1e-4	±1e-4	±5e-4	±4e-4
	P	**0.723**	**0.744**	**0.741**	**0.742**	**0.763**	**0.760**	0.762	0.753	0.361	0.402	0.427	0.447
		±1e-4	±3e-4	±6e-5	±5e-5	±4e-5	±4e-6	±2e-5	±4e-6	±2e-4	±2e-4	±6e-4	±3e-4
	R	0.624	0.691	0.719	0.748	0.713	0.749	0.768	0.775	0.373	0.427	0.453	0.467
		±1e-4	±3e-4	±5e-5	±2e-4	±4e-5	±2e-7	±2e-6	±5e-6	±8e-5	±9e-5	±5e-4	±4e-4
	Inv.	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.2%	0.1%	0.1%	0.3%
		±0.0%	±0.0%	±0.0%	±0.0%	±0.0%	±0.0%	±0.0%	±0.0%	±3e-4%	±1e-4%	±1e-4%	±5e-4%

Table 18: Comparison of F1, Precision (P) and Recall (R) metrics of LLMs with varying number of examples k for ICL on different datasets. Results are averaged over 3 runs, and the associated variances are reported. The column "Inv." indicates the % of abstracts excluded due to instruction-following errors. F1 scores are computed only on valid outputs; including invalid cases would further reduce scores. Rows marked "–" denote context-window limits preventing evaluation with the specified number of examples. **Bold** : Best result

	Noise Source	F1	Noise	Correct	Wrong Label	Wrong Overlap	Multiple Entities	Missing	Invalid Tag
GPT-4o	**BioLinkBERT**	0.478	0.522	936	228	292	113	230	461
	Phi4	0.432	0.568	801	225	257	139	327	400
	Gemma3	0.458	0.542	858	212	272	98	369	418
	Mistral-3.1	0.487	0.513	955	209	284	107	258	483
	GPT-4o	0.446	0.554	884	227	285	148	205	533
Mistral-3.1	**BioLinkBERT**	0.495	0.505	944	183	308	75	371	413
	Phi4	0.480	0.520	887	192	270	99	383	362
	Gemma3	0.467	0.533	842	191	269	59	357	516
	Mistral-3.1	0.490	0.510	1005	212	338	65	288	590
	GPT-4o	0.471	0.529	950	201	312	121	229	556

Table 19: Error Type Statistics for GPT-4o and Mistral-3.1 under multiple noise sources at 30% noise for 40 MM-ST21PV Abstracts. The 40 evaluated abstracts contains a total of 1888 annotations.

	Noise Source	F1	Noise	Correct	Wrong Label	Wrong Overlap	Multiple Entities	Missing	Invalid Tag
GPT-4o	**Phi4**	0.417	0.583	768	202	262	146	358	421
	Gemma3	0.455	0.545	844	205	261	102	394	413
	Mistral-3.1	0.484	0.516	968	216	312	101	237	515
	GPT-4o	0.437	0.563	864	207	285	161	198	549
Mistral-3.1	**Phi4**	0.433	0.567	742	178	239	105	546	272
	Gemma3	0.483	0.517	831	160	253	51	563	295
	Mistral-3.1	0.487	0.513	970	177	337	69	354	545
	GPT-4o	0.471	0.529	929	177	287	132	255	528

Table 20: Error Type Statistics for GPT-4o and Mistral-3.1 under multiple noise sources at 50% noise for 40 MM-ST21PV Abstracts. The 40 evaluated abstracts contains a total of 1888 annotations.