LLMs are not Zero-Shot Reasoners for Biomedical Information Extraction

Aishik Nagar[1] **Viktor Schlegel**[2,3] **Thanh-Tung Nguyen** **Hao Li**[3]
Yuping Wu[3] **Kuluhan Binici**[4] **Stefan Winkler**[4]

[1]ASUS Intelligent Cloud Services (AICS), Singapore
[2]Imperial College London, Imperial Global Singapore
[3]University of Manchester, United Kingdom
[4]National University of Singapore, Singapore
Correspondence to: `v.schlegel@imperial.ac.uk`

Abstract

Large Language Models (LLMs) are increasingly adopted for applications in healthcare, reaching the performance of domain experts on tasks such as question answering and document summarisation. Despite their success on these tasks, it is unclear how well LLMs perform on tasks that are traditionally pursued in the biomedical domain, such as structured information extraction. To bridge this gap, in this paper, we systematically benchmark LLM performance in Medical Classification and Named Entity Recognition (NER) tasks. We aim to disentangle the contribution of different factors to the performance, particularly the impact of LLMs' task knowledge and reasoning capabilities, their (parametric) domain knowledge, and addition of external knowledge. To this end, we evaluate various open LLMs—including BioMistral and Llama-2 models—on a diverse set of biomedical datasets, using standard prompting, Chain-of-Thought (CoT) and Self-Consistency based reasoning as well as Retrieval-Augmented Generation (RAG) with PubMed and Wikipedia corpora. Counter-intuitively, our results reveal that standard prompting consistently outperforms more complex techniques across both tasks, laying bare the limitations in the current application of CoT, self-consistency and RAG in the biomedical domain. Our findings suggest that advanced prompting methods developed for knowledge- or reasoning-intensive tasks, such as CoT or RAG, are not easily portable to biomedical tasks where precise structured outputs are required. This highlights the need for more effective integration of external knowledge and reasoning mechanisms in LLMs to enhance their performance in real-world biomedical applications.

1 Introduction

The success of Large Language Models (LLMs) is reshaping AI healthcare applications, particularly in Question Answering (Budler et al., 2023; Subramanian et al., 2024), summarization (Van Veen et al., 2024; Schlegel et al., 2023; Nagar et al., 2024), and extracting insights from unstructured patient-generated health data (Li et al., 2023). While advancements in fine-tuning and in-context learning (ICL) have improved LLM performance, these rely on readily available structured training data (Abburi et al., 2023; Zhang et al., 2024; Gutierrez et al., 2022). However, in biomedical contexts, such resources are often unavailable due to domain shifts (Hadi et al., 2023) or ad-hoc requirements —for example when researchers need to process a set of medical records to find patients satisfying inclusion criteria for a clinical trial (Jullien et al., 2023; Hadi et al., 2023) (e.g., whether they're a smoker). This limits the effectiveness of parametric knowledge improvements in LLMs, necessitating strong zero-shot capabilities for structured prediction tasks such as biomedical classification and Named Entity Recognition. Despite this, the literature currently lacks a systematic investigation of other crucial aspects of knowledge utilization in zero-shot performance of LLMs in such tasks.

In order to address this research gap, we first postulate that LLM performance in true zero-shot settings—where only task labels and their meaningful names are provided (Lampert et al., 2014)—hinges on three categories of knowledge: *(a) Parametric Knowledge*: Information embedded in model weights; *(b) Task Knowledge*: Understanding of task-specific labels and context; *(c) External Knowledge*: Additional retrieved context to supplement decision-making.

Existing research evaluating these factors in LLMs for the medical domain focus on knowledge-intensive tasks like Multiple-Choice QA (Nori et al., 2023; Subramanian et al., 2024), but their capabilities in structured prediction tasks, such as medical classification and information extraction, remain underexplored. Additionally, tech-

The Sixth Workshop on Insights from Negative Results in NLP, pages 106–120
May 4, 2025 ©2025 Association for Computational Linguistics

niques like zero-shot Chain-of-Thought (CoT) reasoning (Wei et al., 2022; Wang and Zhou, 2024), self-consistency (Wang et al., 2022), and Retrieval-Augmented Generation (RAG) (Li et al., 2024) require systematic evaluation in these contexts.

Additionally, evaluations often focus on proprietary models like ChatGPT (Biswas, 2023) or GPT-4 (OpenAI, 2023), which pose challenges due to computational cost, privacy concerns, and inaccessibility for techniques like constrained decoding. Despite the growing concerns regarding reliability of LLMs in medical applications, techniques like constrained decoding which have shown promise in mitigating LLM hallucinations have not been systematically applied to medical information extraction or classification.

Thus, four key issues currently hinder progress: *(i)* Reliance on training sets and parametric knowledge for structured prediction, which may be unrealistic; *(ii)* Lack of true zero-shot evaluations for structured tasks beyond surrogate QA; *(iii)* Dependence on large-scale, proprietary LLMs, limiting practical deployment. *(iv)* Lack of a systematic analysis of impact of latest techniques such as Chain-of-thought reasoning, RAG and constrained generation in medical structured prediction tasks.

This paper systematically benchmarks LLMs in medical classification and Named Entity Recognition (NER), assessing task and external knowledge while controlling parametric knowledge. We evaluate CoT reasoning, RAG, and constrained generation, offering insights into their applicability.

First, we present the first comprehensive benchmark of task and external knowledge adaptation for LLMs in medical structured prediction tasks. **Second**, we analyze the impact of knowledge enhancement techniques, including CoT, self-consistency, RAG, and constrained generation. **Third**, we demonstrate that parametric knowledge capacity, i.e., model size, is the primary driver of zero-shot performance, highlighting the limitations and potential of current LLM architectures.

2 Related Work

We survey the existing benchmarking literature for the medical domain in the **appendix section A**, outlining the lack of studies focusing on structured prediction tasks. Furthermore, we cover recent prompting techniques that were proposed to elicit reasoning in LLMs, and augment their domain knowledge, either by better tapping into their parametric knowledge or by explicitly providing them with relevant external context. Notably, we omit approaches that rely on existence of training sets, such as few-shot prompting (Wang et al., 2023) or model fine-tuning, as one of the key challenges in the medical domain is the lack of annotated task data, due to privacy concerns over sharing medical records. Instead, as outlines in the introduction, we focus on “true” zero-shot capabilities of LLMs.

Reasoning- and Knowledge-enhancing approaches: Current work attempts to improve the performance of LLMs from different knowledge utilization perspectives. One of the obvious methods is full parameter domain-specific pretraining (Xie et al., 2024). For example, Chen et al. (2023) propose the largest medical foundation model, trained on both biomedical and clinical data, up to 70B. Bolton et al. (2024), on the other hand, believe larger LLMs are computationally expensive to run, proposing a 2.7B LLM specific for biomedical NLP tasks. When fine-tuned, the relatively small model compete with larger LLMs. In our study, we compare domain-generalist models with those adapted to the medical domain. Since full parameter tuning is costly, many works focus on domain knowledge adaptation by pretraining (Shi et al., 2024; Song et al., 2024) or instruction tuning (Willard and Louf, 2023) with adapters. Training-free approaches encompass chain-of-thought (CoT) (Wei et al., 2022; Jeong et al., 2024), self-consistency (Wang et al., 2022), and, concerned with lack of grounding resulting in hallucination, recent work introduce RAG methods (Li et al., 2024; Wang et al., 2024b; Yu et al., 2023; Munnangi et al., 2024; Wang et al., 2024a; Soong et al., 2023). However, most of these efforts have focused on performance in a particular knowledge paradigm and have lacked a systematic assessment of their performance on structured prediction, which we address in our study.

3 Methodology

Our methodology is designed to answer the following two research questions:

1. How well do Large Language Models (LLMs) perform on structured prediction tasks when provided with unstructured inputs?

2. To what extent can approaches that enhance task knowledge and external knowledge improve their performance?

To answer the first research question, we benchmark LLMs on biomedical text classification and NER datasets, focusing on the "true" zero-shot setting to evaluate models' *parametric knowledge*. This reflects real-world scenarios where annotated data is often unavailable due to ad-hoc task requirements, resource limitations and privacy constraints (Giachelle et al., 2021). This leads to what Fries et al. (2022b) describe as "dataset debt", highlighting issues like inconsistent documentation, lack of domain-specific information except generic entities and difficulties adapting datasets to niche domains. Clinicians face significant time constraints, which limit even few-shot annotations (Xia and Yetisgen-Yildiz, 2012; Wac et al., 2024; Farri et al., 2013). These factors make **fine-tuning and few-shot approaches impractical** for structured prediction tasks in the biomedical domain, **positioning zero-shot methods as a scalable solution for real-world biomedical tasks**.

To answer the second question, we compare their zero-shot performance to various methods that aim to enhance *task knowledge* and *external knowledge*, while keeping the *parametric knowledge* static.

Techniques Table 1 lists our methods. We use Vanilla prompting as the baseline, and enhance it with advanced approaches: chain-of-thought (CoT) (Wei et al., 2022) and self-consistency (SC) (Wang et al., 2022), along with retrieval-augmented generation (RAG) (Lewis et al., 2020) that leverages FAISS with PubMed abstracts and Wikipedia articles, embedding documents via `all-MiniLM-L6-v2` (Reimers and Gurevych, 2019). We also apply constrained decoding (Willard and Louf, 2023) to enforce structured outputs. For NER, we adopt a two-stage approach: Stage 1 assigns generic entity labels (e.g., "Body-part"), and Stage 2 refines them to fine-grained labels. Self-consistency is employed in both tasks to aggregate multiple reasoning paths via majority voting.

Complete details of our datasets, techniques and methods are described in Appendix D.

4 Evaluation Results

The complete table of results is provided in table 2. We give an overview of our findings followed by a deeper analysis of the evaluated techniques.

4.1 Overview of results

	Technique	CLS	NER	
		F1	F1-S	F1-L
BioMistral-7B	Vanilla	36.5	3.3	2.2
	CoT	31.3	1.5	1.3
	SC-CoT	20.5	0.8	0.4
	CoT-RAG-P	14.7	1.6	1.2
	CoT-RAG-W	15.5	1.3	1.0
	SC-CoT-RAG-P	19.2	0.5	0.4
	SC-CoT-RAG-W	21.6	0.4	0.3
Llama-2-70B	Vanilla	40.3	8.6	5.8
	CoT	35.9	10.3	7.3
	SC-CoT	28.0	9.1	5.4
	CoT-RAG-P	16.5	9.9	7.1
	CoT-RAG-W	15.7	10.6	7.2
	SC-CoT-RAG-P	27.2	9.0	5.4
	SC-CoT-RAG-W	26.6	9.1	5.3
Llama-2-7B	Vanilla	34.9	6.5	5.2
	CoT	30.6	4.9	2.5
	SC-CoT	24.6	5.1	3.0
	CoT-RAG-P	14.3	4.6	2.3
	CoT-RAG-W	14.5	4.2	1.7
	SC-CoT-RAG-P	25.5	5.7	2.9
	SC-CoT-RAG-W	11.1	5.6	3.2

Table 2: Performance of each model and technique combination across Classification and NER datasets. For classification, we report Micro-F1 and for NER we report both Span-Identification Micro-F1 performance as well as full Micro-F1 performance, including recognizing correct types.

Reasoning and knowledge-enhancing techniques do not improve performance. Figures 1 and 2 compare the best-performing techniques for classification and NER. Surprisingly, Table 2 in the Appendix shows that Standard Prompting consistently achieves the highest average F1 scores across models: BioMistral-7B (36.48%), Llama-2-70B-Chat-AWQ (40.34%), and Llama-2-7b-chat-hf (34.92%). This suggests that for structured prediction tasks, complex reasoning techniques like CoT or RAG do not outperform Standard Prompting.

For NER, Standard Prompting remains effective, but performance varies across models and datasets. Scores are significantly lower than typical F1 scores in biomedical NER benchmarks such as NCBI disease corpus (Doğan et al., 2014;

Technique	Details	Comments
Vanilla	Standard prompting.	Baseline for all tasks.
Chain-of-Thought (CoT) (Wei et al., 2022)	Chain-of-thought reasoning.	Effective for QA and logical reasoning. For NER, adapted into a two-stage approach where generic entity names are first induced (e.g., *Bodypart*), followed by fine-grained labeling.
Self-Consistency (SC) (Wang et al., 2022)	Majority voting across sampled reasoning paths.	Applied in both stages of the two-stage NER approach.
Retrieval-Augmented-Generation (RAG) (Lewis et al., 2020)	Retrieval-augmented generation using FAISS (Douze et al., 2024).	Used PubMed (Sanyal et al., 2021) and Wikipedia as corpora. PubMed improved performance; Wikipedia degraded performance for medical QA (Xiong et al., 2024).
Constrained Decoding (Willard and Louf, 2023)	Restricted outputs to ensure structured extraction.	Avoided hallucinations. Ensured span and label consistency in NER tasks.

Table 1: Techniques Summary with Comments and Details. Complete details can be found in Appendix D.

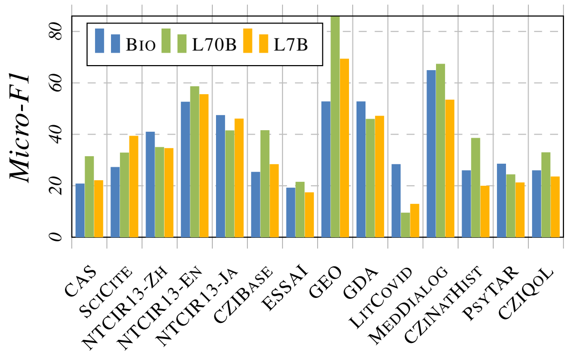

Figure 1: Best-performing *Standard Prompting* method for **Bio**Mistral 7B, **L**lama-**70B** and **L**lama-**7B** for all classification tasks.

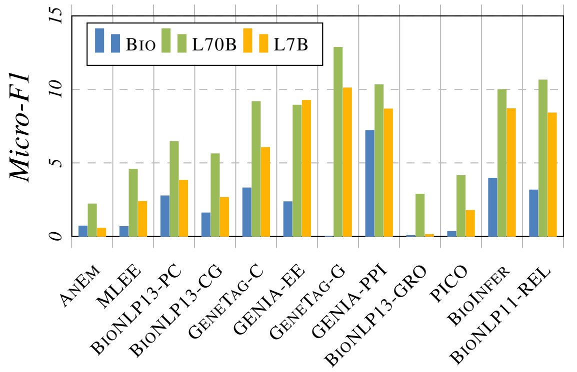

Figure 2: Best-performing *Standard Prompting* method for **Bio**Mistral 7B, **L**lama-**70B** and **L**lama-**7B** for all NER tasks.

Krallinger et al., 2015) and CHEMDNER, where specialized models achieve up to 0.90 Span F1 scores (Kocaman and Talby, 2021; Zhou et al., 2023). However, similar to our findings, zero-shot NER scores tend to be low, even in general domains (Shen et al., 2021) and when providing label descriptions (Picco et al., 2024).

The likely reason for poor performance is that these approaches excel in knowledge- and reasoning-intensive tasks like Question Answering (Nori et al., 2023) or Mathematical Reasoning (Wang and Zhou, 2024; Wang et al., 2022; Li et al., 2024), but structured prediction tasks require understanding task semantics over generic reasoning. These tasks rely less on broad knowledge from biomedical papers or Wikipedia and more on domain-specific application within the given input. Effective models must handle specialized vocabulary, jargon, acronyms, and synonyms varying across subfields (Kim et al., 2007; Zheng et al., 2018; Jiang and Xu, 2024). They must also resolve ambiguity, polysemy, and syntactic nuances in biomedical concepts, which the LLMs to not have been able to capture.

Scale drives improvements. Consistent with prior findings, the 70B model shows notable gains over the 7B model (5.4% for classification, 2.2% for NER Span F1). The largest performance gap appears when using SC with COT and RAG (Wikipedia), where the 70B model surpasses the 7B model by 15.45%. This suggests the larger model excels at leveraging external knowledge when paired with SC and chain-of-thought prompting. The 70B model's greater capacity is particularly beneficial for handling complex reasoning and knowledge integration (Wei et al., 2022). This is further supported by its 10.91% improvement when SC is added to Wikipedia-based RAG, helping mitigate performance drops from irrelevant external information. Unlike classification tasks, where Standard Prompting performed best, NER performance remains stable with advanced prompting techniques, especially in larger models like Llama-2-70B, likely due to the inherent lack of epistemic certainty in NER outputs.

4.2 Detailed Comparison of Prompting Techniques

CoT and SC underperform without sufficient parametric knowledge. For BioMistral-7B, SC-

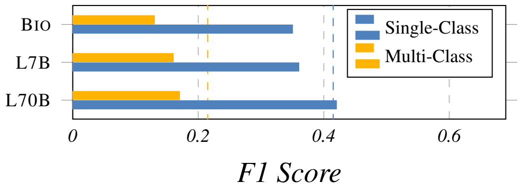

Figure 3: Performance comparison for **Bio**Mistral 7B, **L**lama**7B** and **L**lama**70B** on single- and multi-label datasets, with random guess baselines of 0.415 and 0.215, respectively.

CoT prompts reduce classification performance by about 16%. One reason may be that domain-specific pre-training, while enhancing performance on specialized biomedical tasks, can limit the model's general adaptability(Brokman and Kavuluru, 2024). Similar to RAG, SC does not consistently improve NER. While SC aims to generate multiple reasoning paths and select the most consistent one, it may introduce errors when the model's epistemic certainty in its outputs is low, leading to performance drops. For NER, combining CoT and SC with RAG (Wikipedia) produces the largest performance gap between 70B and 7B models, suggesting that larger models use external knowledge and complex reasoning more effectively when parametric knowledge is limited.

RAG does not help information extraction. Although RAG can improve QA tasks by retrieving relevant facts (Xiong et al., 2024), it seems less useful for classification and information extraction, where irrelevant information misleads the model, creating additional complexity. This results in a drop in classification accuracy, dropping 16.91% with PubMed Corpora and 16.47% with Wikipedia compared to the best classification method.

SC helps filter noise for RAG but does not aid CoT. While SC aims to improve CoT by generating multiple reasoning paths, its efficacy depends on the model's epistemic certainty (Yadkori et al., 2024; Liu et al., 2024). Insufficient parametric knowledge lowers confidence, causing performance declines. BioMistral-7B loses about 16% in classification with SC-CoT prompting. SC also fails to consistently boost NER. However, in the 70B model, combining CoT and SC with RAG (Wikipedia) yields notable gains, indicating that larger models better exploit external knowledge and present higher epistemic certainty owing to their larger parametric capacity.

4.3 Detailed Per-Dataset Analysis

Figure 3 provides the performance comparison of the three models on single and multi-label tasks. Figures 4 and 5 provide a detailed analysis and breakdown of performance of each technique (NER and classification) on each dataset, along with random baselines for each. We discuss their implications below.

Models perform significantly better on public datasets. On public datasets, models average around 30% accuracy, compared to 12% on private datasets, suggesting potential data leakage from publicly available sources used in pre-training or instruction-tuning. Thus, performance on "unseen" tasks may stem from prior exposure rather than true generalization.

Multilingual performance is not scale dependent. As shown in Figure 1, smaller models can match or outperform larger ones on Chinese and Japanese datasets but lag behind in English tasks. This disparity likely results from heavy reliance on English corpora and limited non-English domain exposure, increasing overfitting risks. Factors like language family, data sampling ratios, and sparse representation can also inhibit multilingual models (He et al., 2024; Bagheri Nezhad and Agrawal, 2024).

LLMs struggle on high-complexity tasks. As in Figure 3, LLMs often fail to surpass random baselines for single and multi-class classification, though Figure 4 shows these baselines remain unbeaten in only two of 14 datasets.

5 Conclusion

We provide a comprehensive benchmark and analysis of LLMs in Medical Classification and Named Entity Recognition tasks, revealing several key insights that have significant implications for the field. We carry out a critical investigation of broad claims regarding LLM capabilities by replicating them in various contexts, domains and datasets. We find that models suffer from fundamental drawbacks in generalizability, which hinder their performance in structured information extraction tasks on domain specific problems. This leads to standard prompting outperforming more advanced methods for all models. Our findings underscore that parametric knowledge capacity remains the most important factor in zero-shot settings, with advanced techniques used to augment external knowledge or model reasoning failing to improve performance.

Limitations

While our study provides important insights into LLMs' capabilities for biomedical classification and information extraction, several limitations should be considered when interpreting our results. Our findings are primarily empirical and, although they suggest consistent patterns across models and tasks, further theoretical work is needed to fully understand why advanced prompting techniques fail to improve performance on structured prediction tasks. We deliberately exclude closed-source LLMs from our analysis due to privacy concerns in medical applications and the observed dataset leakage issues, where public dataset contamination is even harder to control for proprietary models. Additionally, our focus on constrained decoding for reliable output parsing necessarily limits us to open-source models where we have access to the generation process.

We also specifically choose not to evaluate in-context learning (ICL) approaches, as our study focuses on "true" zero-shot capabilities where no task-specific examples are available. While techniques like k-NN ICL have shown promise in other domains, they require substantial annotated data to retrieve examples from—which is often unavailable in practical medical settings. Fixed ICL examples could be used, but performance would then largely depend on example selection, essentially reducing the evaluation to the quality of prompt engineering. To balance *(i)* scientific validity and focus on real-world scenarios, where domain experts may not be prompt engineering specialists, with *(ii)* the need to provide useful information to the models, we instead opt for the zero-shot setting—addressing *(i)*—while ensuring semantic clarity through meaningful label names (e.g., using "Control" and "Perturbation" rather than "0" and "1" in the GEO dataset)—addressing *(ii)*.

References

Harika Abburi, Michael Suesserman, Nirmala Pudota, Balaji Veeramani, Edward Bowen, and Sanmitra Bhattacharya. 2023. Generative AI text classification using ensemble LLM approaches. In *IberLEF@SEPLN*, volume 3496 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Sina Bagheri Nezhad and Ameeta Agrawal. 2024. What drives performance in multilingual language models? *arXiv e-prints*, pages arXiv–2404.

Som S Biswas. 2023. Role of chat gpt in public health. *Annals of biomedical engineering*, 51(5):868–869.

Elliot Bolton, Abhinav Venigalla, Michihiro Yasunaga, David Hall, Betty Xiong, Tony Lee, Roxana Daneshjou, Jonathan Frankle, Percy Liang, Michael Carbin, et al. 2024. Biomedlm: A 2.7 b parameter language model trained on biomedical text. *arXiv preprint arXiv:2403.18421*.

Àlex Bravo, Janet Piñero, Núria Queralt-Rosinach, Michael Rautschka, and Laura I Furlong. 2015. Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. *BMC Bioinformatics*, 16(1).

Aviv Brokman and Ramakanth Kavuluru. 2024. How important is domain specificity in language models and instruction finetuning for biomedical relation extraction? *arXiv preprint arXiv:2402.13470*.

Leona Cilar Budler, Lucija Gosak, and Gregor Stiglic. 2023. Review of artificial intelligence-based question-answering systems in healthcare. *WIREs Data. Mining. Knowl. Discov.*, 13(2).

Qingyu Chen, Alexis Allot, Robert Leaman, Rezarta Islamaj Doğan, and Zhiyong Lu. 2021. Overview of the biocreative vii litcovid track: multi-label topic classification for covid-19 literature annotation. In *Proceedings of the seventh BioCreative challenge evaluation workshop*.

Shu Chen, Zeqian Ju, Xiangyu Dong, Hongchao Fang, Sicheng Wang, Yue Yang, Jiaqi Zeng, Ruisi Zhang, Ruoyu Zhang, Meng Zhou, Penghui Zhu, and Pengtao Xie. 2020. Meddialog: A large-scale medical dialogue dataset. *CoRR*, abs/2004.03329.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. 2023. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. 2019. Structural scaffolds for citation intent classification in scientific publications. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Rezarta Islamaj Doğan, 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.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. *arXiv preprint*.

Inc. Elucidata. 2022. Geokhoj v1. `https://github.com/ElucidataInc/GEOKhoj-datasets/tree/main/geokhoj_v1`.

Oladimeji Farri, Karen A Monsen, Serguei V Pakhomov, David S Pieczkiewicz, Stuart M Speedie, and Genevieve B Melton. 2013. Effects of time constraints on clinician–computer interaction: a study on information synthesis from ehr clinical notes. *Journal of biomedical informatics*, 46(6):1136–1144.

Hui Feng, Francesco Ronzano, Jude LaFleur, Matthew Garber, Rodrigo de Oliveira, Kathryn Rough, Katharine Roth, Jay Nanavati, Khaldoun Zine El Abidine, and Christina Mack. 2024. Evaluation of large language model performance on the biomedical language understanding and reasoning benchmark. *medRxiv*, pages 2024–05.

Jason Fries, Leon Weber, Natasha Seelam, Gabriel Altay, Debajyoti Datta, Samuele Garda, Sunny Kang, Rosaline Su, Wojciech Kusa, Samuel Cahyawijaya, et al. 2022a. Bigbio: A framework for data-centric biomedical natural language processing. *Advances in Neural Information Processing Systems*, 35:25792–25806.

Jason Alan Fries, Natasha Seelam, Gabriel Altay, Leon Weber, Myungsun Kang, Debajyoti Datta, Ruisi Su, Samuele Garda, Bo Wang, Simon Ott, et al. 2022b. Dataset debt in biomedical language modeling. In *2022 Challenges and Perspectives in Creating Large Language Models, Proceedings of the Workshop*, pages 137–145. Association for Computational Linguistics (ACL).

Fabio Giachelle, Ornella Irrera, and Gianmaria Silvello. 2021. Medtag: a portable and customizable annotation tool for biomedical documents. *BMC Medical Informatics and Decision Making*, 21(1):352.

Natalia Grabar, Vincent Claveau, and Cl'ement Dalloux. 2018. CAS: French corpus with clinical cases. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 122–128, Brussels, Belgium. Association for Computational Linguistics.

Bernal Jimenez Gutierrez, Nikolas McNeal, Clayton Washington, You Chen, Lang Li, Huan Sun, and Yu Su. 2022. Thinking about GPT-3 in-context learning for biomedical ie? think again. In *EMNLP (Findings)*, pages 4497–4512. Association for Computational Linguistics.

Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, et al. 2023. A survey on large language models: Applications, challenges, limitations, and practical usage. *Authorea Preprints*.

Emily Harris. 2023. Large language models answer medical questions accurately, but can't match clinicians' knowledge. *JAMA*.

Yifei He, Alon Benhaim, Barun Patra, Praneetha Vaddamanu, Sanchit Ahuja, Parul Chopra, Vishrav Chaudhary, Han Zhao, and Xia Song. 2024. Scaling laws for multilingual language models. *arXiv preprint arXiv:2410.12883*.

Robert Hoehndorf, Axel cyrille Ngonga Ngomo, Sampo Pyysalo, Tomoko Ohta, Anika Oellrich, and Dietrich Rebholz-schuhmann. 2010. Applying ontology design patterns to the implementation of relations in genia. In *Proceedings of the Fourth International Symposium for Semantic Mining in Biomedicine*.

Hayate Iso, Camille Ruiz, Taichi Murayama, Katsuya Taguchi, Ryo Takeuchi, Hideya Yamamoto, Shoko Wakamiya, and Eiji Aramaki. 2017. Ntcir13 medweb task: multi-label classification of tweets using an ensemble of neural networks. In *NTCIR*.

Minbyul Jeong, Hyeon Hwang, Chanwoong Yoon, Taewhoo Lee, and Jaewoo Kang. 2024. Olaph: Improving factuality in biomedical long-form question answering. *arXiv preprint arXiv:2405.12701*.

Chao Jiang and Wei Xu. 2024. Medreadme: A systematic study for fine-grained sentence readability in medical domain. *arXiv preprint arXiv:2405.02144*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Mael Jullien, Marco Valentino, Hannah Frost, Paul O'Regan, Dónal Landers, and Andre Freitas. 2023. NLI4CT: Multi-Evidence Natural Language Inference for Clinical Trial Reports. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16745–16764, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rajvir Kaur, Jeewani Anupama Ginige, and Oliver Obst. 2023. Ai-based ICD coding and classification approaches using discharge summaries: A systematic literature review. *Expert Syst. Appl.*, 213(Part):118997.

Hyeoneui Kim, Sergey Goryachev, Graciela Rosemblat, Allen Browne, Alla Keselman, and Qing Zeng-Treitler. 2007. Beyond surface characteristics: a new health text-specific readability measurement. In *AMIA Annual Symposium Proceedings*, volume 2007, page 418. American Medical Informatics Association.

Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. 2009. Overview of BioNLP'09 shared task on event extraction. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, pages 1–9, Boulder, Colorado. Association for Computational Linguistics.

Jung-jae Kim, Xu Han, Vivian Lee, and Dietrich Rebholz-Schuhmann. 2013. GRO task: Populating the gene regulation ontology with events and relations. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 50–57, Sofia, Bulgaria. Association for Computational Linguistics.

Veysel Kocaman and David Talby. 2021. Biomedical named entity recognition at scale. In *Pattern Recognition. ICPR International Workshops and Challenges:*

Virtual Event, January 10–15, 2021, Proceedings, Part I, pages 635–646. Springer.

Martin Krallinger, Florian Leitner, Obdulia Rabal, Miguel Vazquez, Julen Oyarzabal, and Alfonso Valencia. 2015. Chemdner: The drugs and chemical names extraction challenge. *Journal of cheminformatics*, 7:1–11.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*.

Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. 2014. Attribute-based classification for zero-shot visual object categorization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(3):453–465.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Hao Li, Yu-Ping Wu, Viktor Schlegel, Riza Batista-Navarro, Thanh-Tung Nguyen, Abhinav Ramesh Kashyap, Xiao-Jun Zeng, Daniel Beck, Stefan Winkler, and Goran Nenadic. 2023. Team: PULSAR at probsum 2023: PULSAR: pre-training with extracted healthcare terms for summarising patients' problems and data augmentation with black-box large language models. In *BioNLP@ACL*, pages 503–509. Association for Computational Linguistics.

Mingchen Li, Huixue Zhou, Han Yang, and Rui Zhang. 2024. Rt: a retrieving and chain-of-thought framework for few-shot medical named entity recognition. *Journal of the American Medical Informatics Association*, page ocae095.

Linyu Liu, Yu Pan, Xiaocheng Li, and Guanting Chen. 2024. Uncertainty estimation and quantification for llms: A simple supervised approach. *arXiv preprint arXiv:2404.15993*.

Itay Manes, Naama Ronn, David Cohen, Ran Ilan Ber, Zehavi Horowitz-Kugler, and Gabriel Stanovsky. 2024. K-QA: A real-world medical q&a benchmark. *CoRR*, abs/2401.14493.

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. *arXiv preprint arXiv:2404.00152*.

Aishik Nagar, Yutong Liu, Andy T Liu, Viktor Schlegel, Vijay Prakash Dwivedi, Arun-Kumar Kaliya-Perumal, Guna Pratheep Kalanchiam, Yili Tang, and Robby T Tan. 2024. umedsum: A unified framework for advancing medical abstractive summarization. *arXiv preprint arXiv:2408.12095*.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolò Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, Renqian Luo, Scott Mayer McKinney, Robert Osazuwa Ness, Hoifung Poon, Tao Qin, Naoto Usuyama, Chris White, and Eric Horvitz. 2023. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *CoRR*, abs/2311.16452.

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.

Tomoko Ohta, Sampo Pyysalo, Jin-Dong Kim, and Jun'ichi Tsujii. 2010. A reevaluation of biomedical named entity - term relations. *Journal of bioinformatics and computational biology*, 8:917–28.

Tomoko Ohta, Sampo Pyysalo, Rafal Rak, Andrew Rowley, Hong-Woo Chun, Sung-Jae Jung, Sung-Pil Choi, Sophia Ananiadou, and Jun'ichi Tsujii. 2013. Overview of the pathway curation (PC) task of BioNLP shared task 2013. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 67–75, Sofia, Bulgaria. Association for Computational Linguistics.

Tomoko Ohta, Sampo Pyysalo, Jun'ichi Tsujii, and Sophia Ananiadou. 2012. Open-domain anatomical entity mention detection. In *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse*, volume W12-43. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Gabriele Picco, Leopold Fuchs, Marcos Martínez Galindo, Alberto Purpura, Vanessa López, and Hoang Thanh Lam. 2024. Description boosting for zero-shot entity and relation classification. *CoRR*, abs/2406.02245.

Sampo Pyysalo, Filip Ginter, Juho Heimonen, Jari Bj"orne, Jorma Boberg, Jouni J"arvinen, and Tapio Salakoski. 2007. Bioinfer: a corpus for information extraction in the biomedical domain. *BMC bioinformatics*, 8(1):1–24.

Sampo Pyysalo, Tomoko Ohta, and Sophia Ananiadou. 2013. Overview of the cancer genetics (CG) task of BioNLP shared task 2013. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 58–66, Sofia, Bulgaria. Association for Computational Linguistics.

Sampo Pyysalo, Tomoko Ohta, Jin-Dong Kim, and Jun'ichi Tsujii. 2009. Static relations: a piece in the biomedical information extraction puzzle. In *Proceedings of the BioNLP 2009 Workshop*, pages 1–9, Boulder, Colorado. Association for Computational Linguistics.

Sampo Pyysalo, Tomoko Ohta, Makoto Miwa, Han-Cheol Cho, Jun'ichi Tsujii, and Sophia Ananiadou. 2012. Event extraction across multiple levels of biological organization. *Bioinformatics*, 28(18):i575–i581.

Sampo Pyysalo, Tomoko Ohta, and Jun'ichi Tsujii. 2011. Overview of the entity relations (rel) supporting task of bionlp shared task 2011. In *Proceedings of the BioNLP Shared Task 2011 Workshop*, BioNLP Shared Task '11, page 83–88, USA. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP/IJCNLP (1)*, pages 3980–3990. Association for Computational Linguistics.

Debarshi Kumar Sanyal, Plaban Kumar Bhowmick, and Partha Pratim Das. 2021. A review of author name disambiguation techniques for the pubmed bibliographic database. *J. Inf. Sci.*, 47(2).

Viktor Schlegel, Hao Li, Yu-Ping Wu, Anand Subramanian, Thanh-Tung Nguyen, Abhinav Ramesh Kashyap, Daniel Beck, Xiao-Jun Zeng, Riza Theresa Batista-Navarro, Stefan Winkler, and Goran Nenadic. 2023. PULSAR at mediqa-sum 2023: Large language models augmented by synthetic dialogue convert patient dialogues to medical records. In *CLEF (Working Notes)*, volume 3497 of *CEUR Workshop Proceedings*, pages 1668–1679. CEUR-WS.org.

Yongliang Shen, Xinyin Ma, Zeqi Tan, Shuai Zhang, Wen Wang, and Weiming Lu. 2021. Locate and label: A two-stage identifier for nested named entity recognition. *arXiv preprint arXiv:2105.06804*.

Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, Hang Wu, Carl Yang, and May D. Wang. 2024. Medadapter: Efficient test-time adaptation of large language models towards medical reasoning. *Preprint*, arXiv:2405.03000.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Yiping Song, Juhua Zhang, Zhiliang Tian, Yuxin Yang, Minlie Huang, and Dongsheng Li. 2024. Llm-based privacy data augmentation guided by knowledge distillation with a distribution tutor for medical text classification. *arXiv preprint arXiv:2402.16515*.

David Soong, Sriram Sridhar, Han Si, Jan-Samuel Wagner, Ana Caroline Costa Sá, Christina Y Yu, Kubra Karagoz, Meijian Guan, Hisham Hamadeh, and Brandon W Higgs. 2023. Improving accuracy of gpt-3/4 results on biomedical data using a retrieval-augmented language model. *arXiv preprint arXiv:2305.17116*.

Saurabh Srivastava, Anto PV, Shashank Menon, Ajay Sukumar, Alan Philipose, Stevin Prince, Sooraj Thomas, et al. 2024. Functional benchmarks for robust evaluation of reasoning performance, and the reasoning gap. *arXiv preprint arXiv:2402.19450*.

Anand Subramanian, Viktor Schlegel, Abhinav Ramesh Kashyap, Thanh-Tung Nguyen, Vijay Prakash Dwivedi, and Stefan Winkler. 2024. M-QALM: A Benchmark to Assess Clinical Reading Comprehension and Knowledge Recall in Large Language Models via Question Answering. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 4002–4042, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Lorraine Tanabe, Natalie Xie, Lynne H Thom, Wayne Matten, and W John Wilbur. 2005. GENETAG: a tagged corpus for gene/protein named entity recognition. *BMC Bioinformatics*, 6.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, et al. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature medicine*, 30(4):1134–1142.

Marceli Wac, Raul Santos-Rodriguez, Chris McWilliams, and Christopher Bourdeaux. 2024. Evaluation of a data annotation platform for large, time-series datasets in intensive care: Mixed methods study. *arXiv preprint arXiv:2410.16959*.

Junda Wang, Zhichao Yang, Zonghai Yao, and Hong Yu. 2024a. Jmlr: Joint medical llm and retrieval training for enhancing reasoning and professional question answering capability. *arXiv preprint arXiv:2402.17887*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Xuezhi Wang and Denny Zhou. 2024. Chain-of-thought reasoning without prompting. *arXiv preprint arXiv:2402.10200*.

Yang Wang, Xueping Peng, Tao Shen, Allison Clarke, Clement Schlegel, Paul Martin, and Guodong Long. 2023. Soft prompt transfer for zero-shot and few-shot learning in EHR understanding. In *ADMA (3)*, volume 14178 of *Lecture Notes in Computer Science*, pages 18–32. Springer.

Zihao Wang, Anji Liu, Haowei Lin, Jiaqi Li, Xiaojian Ma, and Yitao Liang. 2024b. Rat: Retrieval augmented thoughts elicit context-aware reasoning in long-horizon generation. *arXiv preprint arXiv:2403.05313*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Brandon T Willard and Rémi Louf. 2023. Efficient guided generation for large language models. *arXiv e-prints*, pages arXiv–2307.

Fei Xia and Meliha Yetisgen-Yildiz. 2012. Clinical corpus annotation: challenges and strategies. In *Proceedings of the third workshop on building and evaluating resources for biomedical text mining (BioTxtM'2012) in conjunction with the international conference on language resources and evaluation (LREC), Istanbul, Turkey*, pages 21–27.

Qianqian Xie, Qingyu Chen, Aokun Chen, Cheng Peng, Yan Hu, Fongci Lin, Xueqing Peng, Jimin Huang, Jeffrey Zhang, Vipina Keloth, et al. 2024. Me llama: Foundation large language models for medical applications. *arXiv preprint arXiv:2402.12749*.

Tingyu Xie, Qi Li, Jian Zhang, Yan Zhang, Zuozhu Liu, and Hongwei Wang. 2023. Empirical study of zero-shot ner with chatgpt. *arXiv preprint arXiv:2310.10035*.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. *arXiv preprint arXiv:2402.13178*.

Yasin Abbasi Yadkori, Ilja Kuzborskij, András György, and Csaba Szepesvári. 2024. To believe or not to believe your llm. *arXiv preprint arXiv:2406.02543*.

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. ACI-BENCH: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *CoRR*, abs/2306.02022.

Guoxin Yu, Lemao Liu, Haiyun Jiang, Shuming Shi, and Xiang Ao. 2023. Retrieval-augmented few-shot text classification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6721–6735.

Meishan Zhang, Bin Wang, Hao Fei, and Min Zhang. 2024. In-context learning for few-shot nested named entity recognition. In *ICASSP*, pages 10026–10030. IEEE.

Jiaping Zheng, Hong Yu, et al. 2018. Assessing the readability of medical documents: a ranking approach. *JMIR medical informatics*, 6(1):e8611.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2023. Universalner: Targeted distillation from large language models for open named entity recognition. *arXiv preprint arXiv:2308.03279*.

Maryam Zolnoori, Kin Wah Fung, Timothy B. Patrick, Paul Fontelo, Hadi Kharrazi, Anthony Faiola, Yi Shuan Shirley Wu, Christina E. Eldredge, Jake Luo, Mike Conway, Jiaxi Zhu, Soo Kyung Park, Kelly Xu, Hamideh Moayyed, and Somaieh Goudarzvand. 2019. A systematic approach for developing a corpus of patient reported adverse drug events: A case study for SSRI and SNRI medications. *Journal of Biomedical Informatics*, 90.

A Related Work

Existing LLMs Benchmarks: With the rising popularity of LLMs, many works evaluated their performance in the biomedical and clinical domains. These works typically focus on evaluating domain-knowledge by means of Question Answering (Singhal et al., 2023; Harris, 2023; Subramanian et al., 2024), or focus directly on possible application scenarios, such as summarisation (Li et al., 2023; Yim et al., 2023) or clinical coding (Kaur et al., 2023). Many works combine these two directions in an effort to provide more comprehensive benchmarks (Srivastava et al., 2024; Xiong et al., 2024; Feng et al., 2024; Chen et al., 2020; Manes et al., 2024). However, many of these works overlook the wealth of existing literature and plethora of available resources for traditional structured prediction tasks in the biomedical domain, such as document classification, entity recognition and linking and event and relation extraction (e.g., Pyysalo et al. (2007; 2012) to name a few). Fries et al. (2022a) have provided a comprehensive and unified collection of these resources, however their work prioritises reportage of the resource collection over benchmarking results. Their preliminary evaluations suggest that their evaluated pre-LLM era models barely surpass the random guess baseline in the zero-shot setting. We build upon their work by providing a detailed analysis to what extent approaches to enhance reasoning and knowledge in LLMs help to challenge this status quo.

B Datasets

Table 3 and 4 list the huggingface dataset cards and citations for each classification and ner dataset used in the paper respectively.

For datasets considered private, we assume that models have not been trained on these datasets due to their restricted access, which requires Data Use Agreements (DUAs) and other permissions. Consequently, the likelihood of these datasets being included in common web crawls is low.

We have signed all the relevant Data Use Agreements (DUAs) and strictly adhere to their provisions. We do not redistribute the data and advise those wishing to reproduce experiments involving private datasets to consult the corresponding Hugging Face dataset cards for guidance on obtaining the necessary data.

C Compute Details

1. Hardware used (GPU/CPU): We used a mix of different shared computational facilities with nVidia A100-SXM4-80GB, RTX6000 with 24GB and L40S with 48GB. Debian OS was used for all the compute servers.

2. Memory: The machines used had between 256 GB and 1TB of memory

3. Software and libraries used: The environment can be reproduced from the texttenvironment.yaml file in the supplementary material

4. Model details: The models used have been described in detail in the main paper submission under the Models subsection of the Methodology section.

5. Random seed of 42 was used for all random sampling purposes

D Methodology

Datasets Since we evaluate different prompting techniques, we restrict the choice of tasks to those where the number of possible labels is small enough to fit in the evaluated LLMs' context window. We restrict the number of labels to ten and the mean length of the input documents to at most 2048 tokens. This leaves us with 14 different classification datasets from the BigBio collection[1]. For the NER task, we sample 12 datasets from the pool of those that satisfy the criteria. The resulting dataset sample features four non-English datasets and six non-public classification datasets, which allows us to investigate whether LLMs perform better on minority languages or on data that is less likely to be found in public pre-training corpora. We run the evaluation on the official test-set split where available, otherwise we consider the full dataset. For datasets with more than 500 instances, we sample 500 random but fixed instances to speed up the experiments. Overall, our selection spans English and non-english source data, publicly available and private datasets, and various domains such as scientific papers, medical notes and social media.

---

[1] for the GAD dataset, we only select 1 fold out of the 10 available, as the folds feature the same task for different data, unlike other datasets. We also skipped the Chinese subset of meddialog as we had difficulties loading the dataset

Dataset Name	HuggingFace Card	Citation
GAD	bigbio/gad	(Bravo et al., 2015)
GEO	bigbio/geokhoj_v1	(Elucidata, 2022)
MEDDIALOG	bigbio/meddialog	(Chen et al., 2020)
CZIBASE	bigbio/czi_drsm	
CZIQOL	bigbio/czi_drsm	
CZINATHIST	bigbio/czi_drsm	
LITCOVID	bigbio/bc7_litcovid	(Chen et al., 2021)
CAS	bigbio/cas	(Grabar et al., 2018)
ESSAI	bigbio/essai	(Grabar et al., 2018)
NTCIR13-JA	bigbio/ntcir_13 _medweb	(Iso et al., 2017)
NTCIR13-EN	bigbio/ntcir_13 _medweb	(Iso et al., 2017)
NTCIR13-ZH	bigbio/ntcir_13 _medweb	(Iso et al., 2017)
PSYTAR	bigbio/psytar	(Zolnoori et al., 2019)
SCICITE	bigbio/scicite	(Cohan et al., 2019)

Table 3: Datasets used for classification tasks.

Classification: The datasets used for classification tasks include both single-label and multi-label datasets, covering a wide range of biomedical and clinical domains. For single-label classification, the GAD dataset focuses on identifying associations between genes and diseases (Bravo et al., 2015), while the GEO dataset is concerned with classifying microarray, transcriptomics, and single-cell experiments from the Gene Expression Omnibus (GEO) database (Elucidata, 2022). The MEDDIALOG dataset aims to classify dialogue snippets as either being said by a doctor or a patient (Chen et al., 2020). Furthermore, the CZIDRSM dataset has several subsets, including one for classifying research articles based on aspects of disease research (CZIBASE), and others for identifying whether a paper describes substantive research into Quality of Life (CZIQOL) or is a natural history study (CZINATHIST).

In multi-label classification, the LITCOVID dataset is used for the classification of COVID-19-related articles (Chen et al., 2021). The CAS and ESSAI datasets are utilized for identify negation and uncertainty clinical cases from French-speaking countries (Grabar et al., 2018). The NTCIR13 datasets include subsets for disease classification of tweets in Japanese (*-JA), English (*-EN), and Chinese (*-ZH) (Iso et al., 2017). Additionally, the PSYTAR dataset is used for sentence classification of various drug-related effects, such as Adverse Drug Reactions (ADR) and Withdrawal Symptoms (WDs) (Zolnoori et al., 2019), while the SCICITE dataset is used for citation intent classification based on the context within computer science and biomedical domains (Cohan et al., 2019).

NER: The datasets for Named Entity Recognition (NER) tasks are similarly divided into entity recognition (single entity type) and classification (multiple entity types). In the single-type category, the GENETAG dataset is used for gene/protein NER, with two annotation versions: the original GENETAG-G and the corrected GENETAG-C (Tanabe et al., 2005). Additionally, the GENIA-PPI dataset focuses on protein-protein interactions or gene regulatory relations within the GENIA corpus, capturing primarily static relations (Pyysalo et al., 2009; Hoehndorf et al., 2010; Ohta et al., 2010).

The multiple-type NER datasets encompass various complex biomedical tasks. The ANEM dataset targets anatomical entity recognition (Ohta et al., 2012), while the BIOINFER dataset focuses on recognizing proteins, genes, and RNA entities (Pyysalo et al., 2007). The GENIA-EE dataset is used for the GENIA Event corpus (Kim et al., 2009), and the BIONLP11-REL dataset is employed for extracting part-of relations between genes/proteins and associated entities (Pyysalo et al., 2011). Furthermore, the BIONLP-13-CG dataset is used for Cancer Genetics (CG) information extraction, focusing on recognizing events

Dataset Name	HuggingFace Card	Citation
GeneTag-G	bigbio/genetag	(Tanabe et al., 2005)
GeneTag-C	bigbio/genetag	(Tanabe et al., 2005)
GENIA-PPI	bigbio/genia _relation_corpus	(Pyysalo et al., 2009; Hoehndorf et al., 2010; Ohta et al., 2010)
AnEm	bigbio/an_em	(Ohta et al., 2012)
BioInfer	bigbio/bioinfer	(Pyysalo et al., 2007)
Genia-EE	bigbio/bionlp _shared_task_2009	(Kim et al., 2009)
BioNLP11-REL	bigbio/bionlp_st _2011_rel	(Pyysalo et al., 2011)
BioNLP-13-CG	bigbio/bionlp_st _2013_cg	(Pyysalo et al., 2013)
BioNLP-13-GRO	bigbio/bionlp_st _2013_gro	(Kim et al., 2013)
BioNLP-13-PC	bigbio/bionlp_st _2013_pc	(Ohta et al., 2013)
PICO	bigbio/ebm_pico	(Nye et al., 2018)
MLEE	bigbio/mlee	(Pyysalo et al., 2012)

Table 4: Datasets used for NER tasks.

represented as structured n-ary associations of given physical entities (Pyysalo et al., 2013). The BioNLP-13-GRO dataset aims to populate the Gene Regulation Ontology with events and relations (Kim et al., 2013), and the BioNLP-13-PC dataset is used for the automatic extraction of biomolecular reactions from text (Ohta et al., 2013). Lastly, the PICO dataset deals with recognizing (P)articipants, (I)nterventions, and (O)utcomes (Nye et al., 2018), and the MLEE dataset is used for event extraction related to angiogenesis (Pyysalo et al., 2012).

Models For our experiments, we employed two instruction-tuned variants of the Llama-2 model—7B and 70B—both (Touvron et al., 2023), alongside the BioMistral-7B model (Labrak et al., 2024) which was further pre-trained on the biomedical domain. Since we make use of constrained generation to generate model outputs and guide the models decoding process, we retrict the evaluation to open source models since this process is not possible for proprietary models such as GPT-4.

Techniques Table 1 summarizes the techniques used in this study and highlights relevant nuances and comments. These techniques include Vanilla (standard prompting), CoT (chain-of-thought reasoning) (Wei et al., 2022), and SC (self-consistency) (Wang et al., 2022), as well as RAG (retrieval-augmented generation) (Lewis et al., 2020). For RAG, we used FAISS (Douze et al., 2024; Johnson et al., 2019) with PubMed abstracts (Sanyal et al., 2021) and Wikipedia articles as corpora, embedding documents with `all-MiniLM-L6-v2` (Reimers and Gurevych, 2019). We also implemented constrained decoding for structured output generation (Willard and Louf, 2023), crucial for ensuring reliable outputs in NER and classification tasks. A novel two-stage approach for NER was adopted, inspired by (Shen et al., 2021), where general entities were assigned in Stage 1 and refined in Stage 2.

Standard prompting was used as a baseline for both the Classification as well as the NER tasks. *Chain-of-thought reasoning* (Wei et al., 2022) has been shown to improve performance, particularly in QA and logical reasoning tasks. Thus, we also ran experiments with *chain-of-thought* reasoning to measure its impact on model performance. For the NER task, we adapted a more guided, *two-stage approach* (Shen et al., 2021) to implement a novel chain-of-thought reasoning approach. Here, The first stage involves inducing a generic entity name from a datasets' known entity labels—e.g., "Bodypart" for the NER labels describing different bodyparts—and then labelling the input document with that generic entity type. In the second stage all entities labelled in this way are further disambiguated with their respective fine-grained dataset NER labels. *Retrieval Augmented Generation* (Lewis et al., 2020) has been established as an effective technique to improve model performance by introducing relevant non-parameteric knowledge to models and thus grounding the generated outputs to factual information. Xiong et al. (2024) conducted a systematic study of RAG on medical QA, and we incorporate their findings into our study. We used PubMed abstracts (Sanyal et al., 2021) and Wikipedia articles as knowledge corpora, because Xiong et al.'s (2024) experiments found that using PubMed improved performance over non RAG techniques, while using Wikipedia reduced performance in medical QA tasks. Our goal was to evaluate whether the same holds true for structured prediction tasks as well. For the RAG module, we made use of FAISS (Douze et al., 2024; Johnson et al., 2019), which allows retrieval of most similar documents based on semantic similarity, where we used the `all-MiniLM-L6-v2` sentence transformers (Reimers and Gurevych, 2019) model for embedding input documents and corpora. For each experiment, the number of retrieved documents was computed based on the maximum possible documents which could be used without exceeding the token limit of the model.

Self-consistency, proposed by Wang et al. (2022), improves chain-of-thought reasoning of LLMs by sampling reasoning paths for a given problem, followed by a majority vote for the final answer. We also conduct a set of experiments employing *self-consistency* to investigate whether such improvements can be observed on structured prediction tasks in the medical domain as well. For classification tasks, self consistency was employed to generate multiple reasoning chains for the given problem, followed by answer extraction from each reasoning chain and majority voting to select the final answer. For NER tasks, since we follow the two-stage approach, self-consistency was employed in both stages. Multiple general entity labels were generated in the first stage, and entities were extracted for each such label. In the second stage, self consistency was again used for the entity selection phase as well as the entity label determination step. Majority voting was utilised in final label or class selection in each case (Xie et al., 2023).

Constrained decoding in LLMs (Willard and Louf, 2023) was used to ensure structured information extraction and text generation. This allowed us to evaluate the LLMs for the task at hand without the added variability due to the aleatoric uncertainties brought about by the probabilistic language generation fundamental to the architectures of the models. More specifically, for classification tasks, we ensured the presense of at least one label in the generated outputs. For NER we restricted the generation of spans occurring in text in the first step, and in the second step, for each of the spans we restricted the generation to any of the possible labels. This is also one of the reasons why we opted against evaluating API-based closed-source LLMs[2], as in our initial experiments the hallucinations in generated outputs created problems with reliably parsing the structured outputs.

We refer to chain of thought as CoT, Self-consistency as SC, RAG as RAG-{P|W} for PubMed and Wikipedia corpora, respectively, and to standard prompting as VANILLA.

E Analysis and Performance Breakdown

Figures 4 and 5 provide a detailed analysis and breakdown of performance of each technique (NER and classification) on each dataset, along with random baselines for each. Figure 3 provides the performance comparison of the three models on single and multi-label tasks. A complete discussion for these figures and their implications can be found in section 4.3.

As discussed in section 4.3, LLMs struggle on high-complexity tasks. Even the best performing model, Llama2 70B performs well on only relatively low-complexity tasks (CZIBASE, NTCIR13-EN) and moderate tasks (GEO), but struggles with higher-complexity datasets (BIONLP13-CG, GENIA-EE). In tasks requiring nuanced interpretation (PICO, BIONLP13-GRO), performance remains low. Although RAG (Retrieval-Augmented Generation) sometimes boosts results, it does not universally enhance biomedical information extraction or classification. These findings indicate that even the most advanced general-purpose and domain-specific LLMs are not good zero-shot reasoners for structured prediction tasks such as biomedical information extraction, especially for complex task settings.

F Results Analysis

---

[2] The other reason being their intransparancy with regard to training data, which violates our "true" zero-shot setting.

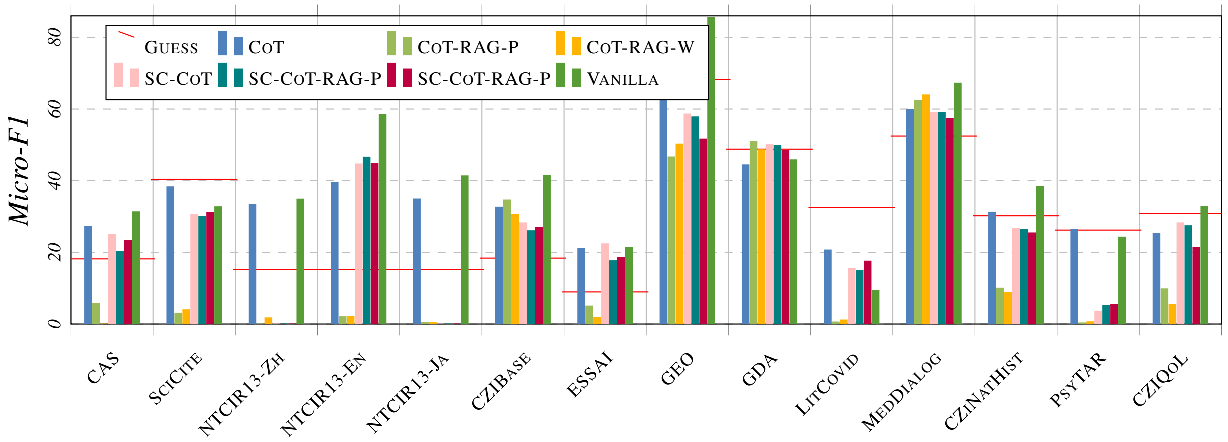

Figure 4: Breakdown of the Micro-F1 performance of each technique and the random guess baseline for all classification datasets, compared against the random guess baseline.

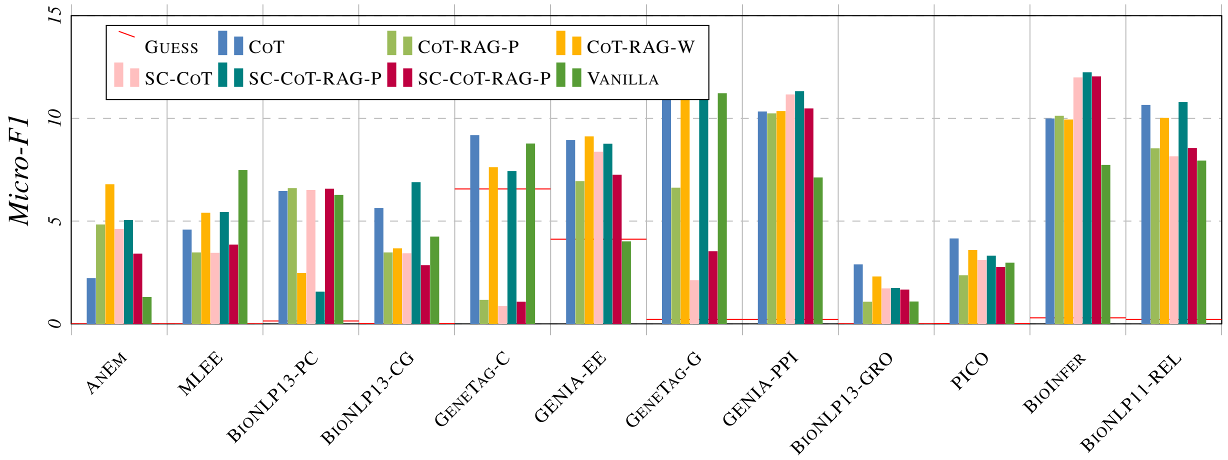

Figure 5: Breakdown of each technique and the random guess baseline on all NER datasets as measured by the Micro-F1 scores. A prediction is counted as correct when both the span and its assigned label are found in the ground truth