Evaluation of large language model performance on the Biomedical Language Understanding and Reasoning Benchmark

Hui Feng, Francesco Ronzano, Jude LaFleur, Matthew Garber, Rodrigo de Oliveira, Kathryn Rough, Katharine Roth, Jay Nanavati, Khaldoun Zine El Abidine, Christina Mack

Affiliation: Real world solutions, IQVIA

Corresponding author: Hui Feng, hui.feng@iqvia.com

Note: Hui Feng, Francesco Ronzano, Jude LaFleur, Matthew Garber, Rodrigo de Oliveira contributed equally to this article.

Disclosure: No AI-assisted technologies were used to assist with the writing of the submitted work.

Abstract

Background: The ability of large language models (LLMs) to interpret and generate human-like text has been accompanied with speculation about their application in medicine and clinical research. There is limited data available to inform evidence-based decisions on the appropriateness for specific use cases.

Methods: We evaluated and compared four general-purpose LLMs (GPT-4, GPT-3.5-turbo, Flan-T5-XXL, and Zephyr-7B-Beta) and a healthcare-specific LLM (MedLLaMA-13B) on a set of 13 datasets – referred to as the Biomedical Language Understanding and Reasoning Benchmark (BLURB) – covering six commonly needed medical natural language processing tasks: named entity recognition (NER); relation extraction; population, interventions, comparators, and outcomes (PICO); sentence similarity; document classification; and question-answering. All models were evaluated without modification. Model performance was assessed according to a range of prompting strategies (formalised as a systematic, reusable prompting framework) and relied on the standard, task-specific evaluation metrics defined by BLURB.

Results: Across all tasks, GPT-4 outperformed other LLMs, followed by Flan-T5-XXL and GPT-3.5-turbo, then Zephyr-7b-Beta and MedLLaMA-13B. The most performant prompts for GPT-4 and Flan-T5-XXL both outperformed the previously-reported best results for the PubMedQA task. The domain-specific MedLLaMA-13B achieved lower scores for most tasks except for question-answering tasks. We observed a substantial impact of strategically editing the prompt describing the task and a consistent improvement in performance when including examples semantically similar to the input text in the prompt.

Conclusion: These results provide evidence of the potential LLMs may have for medical application and highlight the importance of robust evaluation before adopting LLMs for any specific use cases. Continuing to explore how these emerging technologies can be adapted for the healthcare setting, paired with human expertise, and enhanced through quality control measures will be important research to allow responsible innovation with LLMs in the medical area.

Introduction

Recent advances in large language models (LLMs) have generated substantial interest in using them to perform natural language processing (NLP) tasks in the medical domain, including writing clinical notes, summarizing scientific literature, and reasoning about public health topics.[1]

There are also serious concerns about the safety, ethics, and trustworthiness of LLM outputs,[2] with evidence of plausible – but factually incorrect – hallucinations of medical information,[3] perpetuation of racial bias,[4] and omission of important information from LLM summaries.[5] Calls to rigorously evaluate LLM performance when performing specific medical tasks[6] emphasize the need to ensure adequate information is available for evidence-based decision making when these technologies will impact patient care and treatment.

In addition to the need for robust evaluation, there are open questions about how to prompt models to ensure quality outputs specific biomedical tasks and to what extent domain-specific fine-tuning of models is beneficial. Significant performance increases have been observed when the instructions for the task are accompanied by several examples (few-shot learning), compared to only providing instructions (zero-shot learning).[7] Example selection strategies to populate few-shot prompts can also meaningfully impact performance, specifically by selecting semantically similar examples using sentence-level embeddings.[8,9] Another strategy for improving LLM performance is fine-tuning with data from the medical domain; while this approach requires substantial compute resources, specialized technical expertise, and a relatively large amount of high-quality training data, it could be worthwhile to increase trustworthiness of the LLM's output.

We conducted an empirical evaluation to assess the ability of five LLMs – four general-purpose and one domain-specific (i.e., fine-tuned on biomedical texts) – to perform biomedical NLP tasks, all of which are included in the Biomedical Language Understanding and Reasoning Benchmark (BLURB).[10] In addition to benchmarking, we investigated how the quality of LLM responses can be improved through prompting strategies.

Methods

BLURB is a collection of 13 datasets that cover 6 distinct NLP tasks, each with a pre-defined evaluation metric (Figure 1). We compared 5 LLMs against the BLURB datasets and evaluated performance using the standard BLURB metrics. For each LLM and task, several prompting strategies were compared.

Models

Five LLMs were compared: Azure-based GPT-3.5-Turbo (commercial, general purpose), GPT-4 (commercial, general purpose),[11] Flan-T5-XXL (open-source, general purpose),[12] Zephyr-7B-Beta (open-source, general purpose)[13] and MedLLaMA-13B (open-source, fine-tuned).[14] The only fine-tuned model included in the evaluation, MedLLaMA-13B, was fine-tuned on a variety of clinical and medical text data sources.[15]

None of the models were modified or underwent further fine-tuning for this evaluation.

BLURB benchmarking datasets and healthcare-based tasks

BLURB is a collection of 13 publicly available biomedical NLP datasets used to evaluate the following common medical NLP tasks:

- **Named entity recognition** refers to the extraction of specified categories of information from text - such as identifying all mentions of medications, diseases, or cell types. Associated

datasets are all based on PubMed abstracts with specific annotated mentions: *BC5-chem:* drugs and chemical compounds; *BC5-disease:* diseases; *NCBI-disease:* diseases; *BC2GM:* genes; *JNLPBA:* molecular biology concepts (i.e., protein, DNA, RNA, cell line, cell type).

- **Population, interventions, comparators, and outcomes (PICO)** identifies each of the categories from the PICO research framework from an abstract.[16] The PICO data set, *EMB-PICO,* is a collection of clinical trial abstracts with annotated mentions of these key elements of study design.
- **Relation extraction** assesses the ability of an algorithm to classify the relationship between pairs of entities in a text. There are three datasets: *ChemProt* is a collection of PubMed abstracts that requires the algorithm to classify the relationship described between chemical and protein entities; *DDI* is a collection of texts from PubMed abstracts and DrugBank that requires the algorithm to classify the drug-drug interactions that are described; *GAD* is a collection of sentences from PubMed abstracts that requires the algorithm to identify whether a gene-disease relationship is being described.
- **Sentence similarity** assesses the ability of an algorithm to determine the similarity of the meaning of two sentences. The sentence similarity dataset, *BIOSSES,* is a collection of sentence pairs that have been rated on a scale of 0 (no relation) to 4 (equivalent meanings) by subject matter experts; the algorithm must estimate the annotated score.
- **Document classification** assesses the ability of an algorithm to correctly categorize a document. The document classification dataset, *HoC,* is a collection of PubMed abstracts that have been classified according to whether they discuss specific cancer hallmarks.
- **Question answering** assesses the ability of an algorithm to correctly answer free-text questions. The two question answering datasets are both based on PubMed abstracts and annotated with question-answer pairs based on the provided text: *PubMedQA* captures whether the text contains the answer to a research question (yes/no/maybe); *BioASQ* annotates whether an extract is the correct answer to a research question (yes/no).

Each dataset has clearly defined ground truth labels and was partitioned into training, validation, and test sets as part of the BLURB benchmark; all evaluations reported here were performed on the standard BLURB test set using the provided labels. Further detail on BLURB tasks is available[17] and information is summarized in Figure 1.

Prompting strategies

A series of experiments were conducted to systematically evaluate prompting strategies based on the structure of the prompt and example selection. Each of the five LLMs were provided with prompts constructed from four high-level templates (based on studies from previous work[18,19]) across all datasets to assess the impact of distinct combinations of three key variables:

- **Verbosity** (*short versus long*): Short prompts contain concise instructions to the model. Long prompts provide details (e.g., a description of each label) that could be potentially useful to the model.
- **Number of examples** (*zero-shot versus few-shot)*: Zero-shot prompts provide instructions to the model with no examples. Few-shot prompts provided the model with three examples of inputs and the corresponding correct responses.
- **Selection strategy for examples** (*semantically similar versus random*): Previous studies suggest that when prompting an LLM to analyse a specific text excerpt, performance of LLM response could be improved by including examples that are semantically similar to the text excerpt to be analysed as a part of the prompt.[8,9] Based on these studies, we implemented two approaches for example selection from the training set: semantically

similar selection scheme and random selection scheme. Under the semantically similar selection scheme, the three most semantically similar training examples and their correct answers were provided to the LLM as part of the prompt. For NER tasks, among the three semantically similar examples, we included two positive examples (i.e. ground truth contains target entities) and one negative example (i.e. ground truth contains no entities). Under the random selection scheme, three examples were randomly selected.

For all the few-shot experiments for EBM-PICO dataset and the few-shot experiments for PubMedQA with MedLLaMA-13B and Flan-T5-XXL, we encountered the issue of prompt length exceeding maximum context size of some models, and therefore did not conduct the experiments. We also used separate calls for each category (EBM-PICO) and entity mention pair (ChemProt and DDI) to ensure the consistency of prompts with long-text examples.

The precise prompt texts used in experiments are provided in Appendix A.

Evaluation methodology

For each LLM and each dataset, six experiments were run varying the length of the prompt, number of examples, and the example selection strategy. The same prompts were used to interact with all LLMs; full prompts are available in Appendix A. An overview of the evaluation strategy is provided in Figure 1.

To enforce standardisation and repeatability of evaluations, raw responses from LLMs were formatted according to a standard JSON schema for each BLURB dataset. A detailed explanation of this process can be found in Appendix B.

For all experiments performed, the entire test set was used to evaluate performance for a given dataset using the standard, task-specific metrics from the BLURB benchmark (see Figure 1). For NER tasks, the F1 score (the harmonic mean of sensitivity and positive predictive value [i.e., recall and precision]) was calculated. For the PICO task, we used a generalization of this metric, the macro F1 score (an unweighted mean of the F1 score for each of the four information categories). The micro F1 score – a weighted mean of category-specific F1 scores – was used for relation extraction and document classification. Pearson correlation coefficients were calculated to evaluate performance on the sentence similarity task. For the question answering tasks, accuracy – the overall proportion of correct answers – was used.

All analyses were performed using Python (3.9) packages, with scikit-learn (1.4.0) and scipy (1.12.0) for computing the metrics. The version of GPT models used was gpt-35-turbo, 0613 and gpt-4, 0613. Please reach out to authors for scripts used to perform the evaluations.

Results

Table 1 characterizes the performance of the five LLMs against all data in the test set of all BLURB datasets. For NER, relation extraction and PICO datasets, number of examples contained in each dataset varies from hundreds to thousands or tens of thousands. Document classification and question-answering datasets contain hundreds of examples. Sentence similarity contains twenty examples.

Across 11 of the 13 BLURB datasets, GPT-4 had the highest score (Table 1). For all NER datasets, GPT-4 exceeded the comparators across all prompts (higher 100% of the time, with F1 scores ranging from 47.6-78.2). For BC5-chemical, BC5-disease and NCBI-disease, the gap was large (for example, difference between best performing GPT-4 prompting strategy and the next best performing model

for NCBI-disease was 14.49 points); for others, it was considerably smaller (for example, difference between best performing GPT-4 prompting strategy and next best performing model for BC2GM was 3.99 points). For GAD, prompting Zephyr-7B-Beta with semantically similar examples performed better than any GPT-4 prompting strategy (3 points higher than the best of GPT-4). For PubMedQA, all prompting strategies for Flan-T5-XXL outperformed GPT-4 (the best Flan-T5-XXL prompting strategy achieved 1.40 points higher than the best of GPT-4).

For all models and datasets, performance varied widely based on the prompting strategy used (Table 1). Notably, prompts with semantically similar examples had the highest score for 7 of 11 datasets. For GPT-4, the differences between the highest and lowest scores were less than 5 points for NER and PICO datasets, but it were much larger for the three relation extraction tasks (range: 9.55 to 26.34 points). The differences were about 12 points for sentence similarity and document classification datasets and less than 8 points for question-answering. However, this pattern is not generalizable across models; for example, the differences between highest and lowest scores for Flan-T5-XXL were for 10 to 20 points for the NER datasets and less than 8 points for the relation extraction datasets. Although there seems to be no clear best prompting strategies across models or tasks, performance improvements clearly could be made by choosing the optimal prompting strategy. For instance, the accuracy for Zephyr-7B-Beta on PubMedQA dataset was enhanced from 18.40% to 59.40% by changing from a short, zero-shot prompt to a long, few-shot prompt with random examples.

Figure 2 summarizes the average best performance of each model by task. We consider each model, prompt, and dataset as one combination and report the average per-model score across datasets of the same task. Overall, GPT-4 showed the highest average best performance for all tasks, followed by Flan-T5-XXL and GPT-3.5-turbo, then Zephyr-7B-Beta and MedLLaMA-13B. For PICO, only zero-shot experiments were performed and the performance was not ideal – no combination achieved a score higher than 34%. Across tasks, MedLLaMA-13B showed lower performance than other LLMs except for question-answering tasks.

Prompting strategies: impact of short versus long instructions

Figure 3 compares the mean per-task performance of long versus short instructions by model when no examples are included in the prompt. There is no clear best strategy across tasks and models. In some cases, providing longer prompts dramatically improved the performance of a model. For example, for the sentence similarity task, considering the zero-shot scenarios, long instructions improved the MedLLaMA-13B model performance by 16.36 points and the Zephyr-7B-Beta model by 52.27 points; yet for the documentation classification task, more verbose prompts improved MedLLaMA-13B performance by 15.63 points but decreased performance of Zephyr-7B-Beta by 26.00 points. For other models, like Flan-T5-XXL and GPT-4, more verbose instructions tended to decrease performance, as observed in four out of six tasks. For the relation extraction task, all models saw increased performance from the longer instructions, though the magnitude of the improvement varied (Figure 3). For all other tasks, changes in performance were mixed across models.

The mean per-task performance of long versus short instructions by model when using few-shot prompting strategies are displayed in Appendix Figure A.1 (random example selection) and Appendix Figure A.2 (semantically similar example selection). The effect of longer versus shorter prompts was highly dependent on the model and the task. Like zero-shot prompts, performance increased for all LLMs when using long prompts for the relation extraction task, regardless of example selection method.

Prompting strategies: impact of zero-shot versus random few-shot versus semantically similar few-shot

Figure 4 compares the mean per-task performance of providing no examples (i.e., zero-shot), three randomly selected examples (i.e., random few-shot), and three semantically similar examples (i.e., semantically similar few-shot) in the prompt when short instructions are given. For most tasks and models, the introduction of three randomly selected examples resulted in modest changes in performance comparing to the zero-shot scenario. Occasionally the differences were more substantial; for example, there were large score increases in the question answering task for Zephyr-7B-Beta and large score decreases in the sentence similarity task for MedLLaMA-13B.

Providing three semantically similar examples to the short prompt resulted in a more consistent pattern of improvement across tasks and models. Only Flan-T5-XXL's performance was negatively affected, and this occurred in 3 of 5 tasks. For all other models, the addition of semantically similar examples resulted in improvements that ranged from modest (e.g., the NER task for GPT-3.5-Turbo) to substantial (e.g., the sentence similarity task for Zephyr-7B-Beta).

We observed similar patterns for long prompts scenarios; the mean per-task performance across different example selection strategies when using long prompts are compared in Appendix Figure A.3. Like the short prompts scenarios, providing random examples in the prompts had a modest and occasionally negative impact on the performance, while providing semantically similar examples resulted in consistent improvement across all tasks and models.

Discussion

This paper evaluates multiple state-of-the-art LLMs using a standard set of benchmarking tasks and systematically compares the performance of different prompting strategies. GPT-4 generally had the best performance on most tasks and datasets; however, for some tasks, one or more smaller open-source models performed similarly. Across tasks and models, LLMs made a sizable number or errors, indicating that they likely require some form of human oversight and correction to meet adequate quality standards for use in clinical research and medical applications. Our results also underscore the impact of prompt design on model performance, with the inclusion of semantically similar examples generally improving scores; however, there is no clear 'optimal' prompting strategy that generalizes across tasks and models.

MedLLaMA-13B, the only model that has been fine-tuned for the biomedical domain, had consistently lower performance than nearly all the general-purpose comparator LLMs. One possible explanation is that the generalizability of the model might have been reduced after being fine-tuned mainly for question-answering task[20,21]. Enhancing the fine-tuning process for better performance on these tasks could be an interesting area to explore but is out of scope for this work.

For some of the BLURB tasks, even the best general LLM results had substantially lower performance than other published models, including PubMedBERT, the baseline model introduced along with BLURB benchmark.[10] GPT-4's F1-score trailed PubMedBERT by 15.54 to 31.55 points for the named entity recognition datasets and by 24.78 to 41.46 points for relation extraction datasets. However, LLMs showed a more competitive performance on question answering datasets, especially PubMedQA. Here, both GPT-4 and Flan-T5-XXL outperformed BioLinkBERT-Large, the leading model on the BLURB leader board.[22]

Our findings suggest that more verbose, detailed prompts may not always be an effective strategy for improving LLM performance. When the additional information provided is specific to the task

and is not routinely encountered during training, adding information regarding entity definitions for NER or how to answer questions for QA was rarely effective, whereas explaining complex relations for RE had an appreciable impact. Overall, adding semantically similar examples to prompts has been the most consistent way to increase performance across tasks and models.

Our study has several limitations worth noting. First, the datasets of the BLURB benchmark are primarily constructed using publicly available data from abstracts catalogued in PubMed. It is unclear how model performance on specific tasks, such as NER, would generalize to excerpts from other medical data with substantially differences in format or content. Second, heuristics (detailed in Appendix B) were used to normalize model output and enable an automated evaluation of performance. This may not account for tasks that models are able to perform in a more conversational manner. Finally, there are more sophisticated prompting strategies that may further improve model performance but were beyond the scope of this work. For example, exploring the utility of chain-of-thoughts[23] and automating prompt tuning strategies for biomedical NLP tasks is an area of promising future research.[24-26]

These findings provide a comprehensive evaluation of the performance of LLMs on a variety of biomedical natural language processing tasks. The heterogeneity in results across prompting strategies, models, and datasets underscore the importance of evaluating the performance of a given model and prompt work for specific tasks. Our result has suggested great potential in adopting LLMs to execute biomedical tasks; yet it also showed the gap of only using the current stage LLMs for these tasks. Continuing to explore how to enhance the performance of these models in medical settings, paired with human expertise and quality control measures, will be important to allow responsible innovations with LLMs in the medical field.

Ethics Statement

All results are based on publicly available data from non-human subjects; no ethics approvals were required.

Acknowledgements and Disclosure

All authors are employees of IQVIA. This study is funded by IQVIA.

FR had received research fundings from Torres Quevedo R&D Contractor, Spanish Ministry of Science, Innovation and Universities (up to 11/2021). HF, KRough, JN, CM, KZ have stock in IQVIA. RO has stock in Arria NLG. KRough has stock in Google. JN has stock in Microsoft, AZ, Nvidia, Meta. CM has stock in AZ, J&J, and MindMed. FR was previously employed by Medbioinformatics Solutions SL. RO was previously employed by Arria NLG. JN was previously employed by AZ. KRough was previously employed by Google.

Author Contributions

Conceptualization: Hui Feng, Francesco Ronzano
Methodology: Jude LaFleur, Matthew Garber, Rodrigo de Oliveira, Francesco Ronzano
Supervision: Hui Feng, Francesco Ronzano, Jay Nanavati
Validation: Matthew Garber, Jude LaFleur, Rodrigo de Oliveira, Katharine Roth
Writing – original draft: Kathryn Rough, Hui Feng, Rodrigo de Oliveira, Matthew Garber, Jude LaFleur, Francesco Ronzano
Writing – review & editing: Jay Nanavati, Khaldoun Zine El Abidine, Christina Mack

Tables

Table 1. Performance of five large language models on each of the Biomedical Language Understanding and Reasoning Benchmark (BLURB) tasks according to varying prompting strategies *(best performing result appears in bold).*

Dataset	Prompting strategy	Flan-T5-XXL	GPT-3.5-Turbo	GPT-4	MedLLaMA-13B	Zephyr-7B-Beta
Named entity recognition tasks *(F1 score)*						
BC2GM (6,325 examples)	short, zero-shot	39.11	45.75	51.21	19.31	34.32
	short, few-shot (random)	41.44	47.69	54.63	22.39	38.36
	short, few-shot (semantically similar)	43.15	50.71	54.70	33.24	42.86
	long, zero-shot	34.63	45.56	50.30	29.04	32.84
	long, few-shot (random)	38.17	46.14	53.18	31.03	38.54
	long, few-shot (semantically similar)	41.11	49.21	53.99	41.68	43.79
BC5-chemical (5,385 examples)	short, zero-shot	65.61	60.64	75.06	27.07	49.04
	short, few-shot (random)	66.98	66.41	77.35	24.55	53.16
	short, few-shot (semantically similar)	65.18	62.55	76.20	27.00	57.57
	long, zero-shot	49.74	65.08	76.47	43.57	53.40
	long, few -shot (random)	63.70	62.42	78.23	55.15	54.35
	long, few-shot (semantically similar)	64.31	59.09	77.79	53.88	59.34
BC5-disease (4,424 examples)	short, zero-shot	52.16	44.61	60.63	21.84	30.84
	short, few-shot (random)	54.05	48.98	62.15	11.23	36.03
	short, few-shot (semantically similar)	54.67	47.00	56.28	13.88	37.81
	long, zero-shot	34.29	41.50	55.52	27.16	35.52
	long, few-shot (random)	50.93	47.59	63.93	33.94	36.60
	long, few-shot (semantically similar)	52.77	45.89	56.84	27.45	38.26
JNLPBA (8,662 examples)	short, zero-shot	35.84	39.12	44.94	15.33	22.74
	short, few-shot (random)	37.99	40.25	45.43	23.85	33.13
	short, few-shot (semantically similar)	40.34	42.01	45.99	27.83	38.76
	long, zero-shot	25.24	38.95	43.55	12.75	24.14
	long, few-shot (random)	33.05	40.69	45.51	32.28	34.49
	long, few-shot (semantically similar)	37.38	42.18	47.55	36.49	39.68
NCBI-disease (960 examples)	short, zero-shot	51.63	47.46	64.67	22.85	30.64
	short, few-shot (random)	51.78	47.96	65.18	13.44	33.57
	short, few-shot (semantically similar)	56.10	49.39	68.97	26.07	42.56
	long, zero-shot	27.58	55.72	58.95	31.41	37.51
	long, few-shot (random)	44.10	47.19	65.98	35.76	37.91
	long, few-shot (semantically similar)	50.87	50.39	70.59	45.88	42.14
Populations, interventions, comparators, outcomes task *(macro F1 score)*						
EBM-PICO	short, zero-shot	28.42	23.78	33.49	10.80	14.32

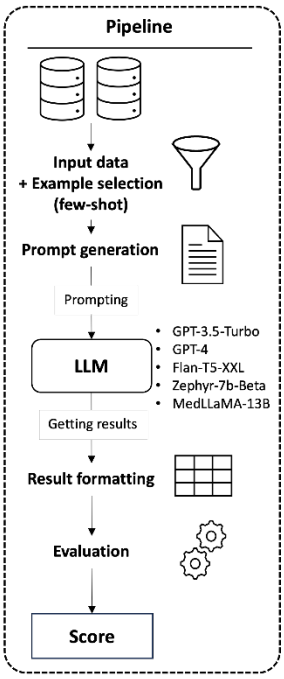
(16,364 examples)	long, zero-shot	24.56	20.46	31.11	10.29	13.36
Relation extraction tasks *(micro F1 score)*						
ChemProt (15,745 examples)	short, zero-shot	14.97	4.17	11.08	7.49	4.24
	short, few-shot (random)	16.08	6.81	16.62	5.40	5.33
	short, few-shot (semantically similar)	17.63	7.53	31.61	12.40	13.31
	long, zero-shot	20.33	31.51	38.25	7.91	19.22
	long, few-shot (random)	19.94	26.46	37.59	8.86	13.07
	long, few-shot (semantically similar)	22.39	21.64	47.42	14.27	19.57
DDI (5,716 examples)	short, zero-shot	15.19	35.18	37.70	10.27	14.91
	short, few-shot (random)	16.01	18.26	27.98	8.79	16.06
	short, few-shot (semantically similar)	16.90	34.69	44.66	7.91	18.85
	long, zero-shot	18.96	40.97	34.95	12.48	18.90
	long, few-shot (random)	19.46	20.76	29.12	9.86	18.97
	long, few-shot (semantically similar)	19.75	36.53	40.90	14.27	19.62
GAD (534 examples)	short, zero-shot	51.12	51.31	50.00	46.82	49.25
	short, few-shot (random)	50.94	49.06	54.68	48.50	47.00
	short, few-shot (semantically similar)	56.18	51.12	59.55	49.81	62.55
	long, zero-shot	50.19	48.88	51.50	51.69	47.00
	long, few-shot (random)	49.81	47.75	52.81	50.94	47.00
	long, few-shot (semantically similar)	57.49	51.50	59.18	52.25	61.61
Sentence similarity task *(Pearson correlation coefficient)*						
BIOSSES (20 examples)	short, zero-shot	90.88	48.84	89.27	-2.65	15.15
	short, few-shot (random)	65.82	79.80	84.65	-15.00	41.78
	short, few-shot (semantically similar)	75.61	82.70	89.03	28.08	72.31
	long, zero-shot	89.86	72.69	80.53	13.71	67.42
	long, few-shot (random)	90.27	93.02	87.08	-30.86	56.79
	long, few-shot (semantically similar)	91.20	92.20	93.18	10.45	77.04
Document classification task *(micro F1 score)*						
HoC (371 examples)	short, zero-shot	49.81	54.10	62.52	0.79	44.11
	short, few-shot (random)	50.73	55.09	62.78	24.16	42.32
	short, few-shot (semantically similar)	47.69	57.57	66.81	42.18	51.74
	long, zero-shot	43.33	43.18	54.45	16.42	18.11
	long, few-shot (random)	39.19	45.44	56.24	21.69	43.29
	long, few-shot (semantically similar)	51.36	44.79	60.88	49.65	47.97
Question answering tasks *(accuracy)*						
BioASQ (140 examples)	short, zero-shot	60.00	77.14	83.57	67.14	60.71
	short, few-shot (random)	60.00	80.71	82.86	66.43	61.43
	short, few-shot (semantically similar)	61.43	81.43	81.43	69.29	64.29
	long, zero-shot	62.86	70.00	85.71	67.14	59.29

	long, few-shot (random)	64.29	81.43	82.14	67.86	57.14
	long, few-shot (semantically similar)	61.43	78.57	84.29	65.71	60.00
PubMedQA (500 examples)	short, zero-shot	76.40	63.40	67.40	55.40	18.40
	short, few-shot (random)	76.60	58.40	72.60	N/A[a]	56.80
	short, few-shot (semantically similar)	N/A[a]	63.40	72.20	N/A[a]	56.40
	long, zero-shot	76.80	63.00	70.60	44.20	21.00
	long, few-shot (random)	76.40	56.80	74.20	N/A[a]	59.40
	long, few-shot (semantically similar)	N/A[a]	60.40	75.40	N/A[a]	58.00

[a] ***Note:*** *Prompt length exceeded the maximum context size of models for the set-ups of all the few-shot experiments for EBM-PICO dataset and the few-shot experiments for PubMedQA with MedLLaMA-13B and Flan-T5-XXL; therefore we did not conduct the experiments for these set-ups, resulting in the N/A values in the table.*

Figures

Figure 1. An illustration of the methodology used for this evaluation of the Biomedical Language Understanding and Reasoning Benchmark (BLURB) tasks

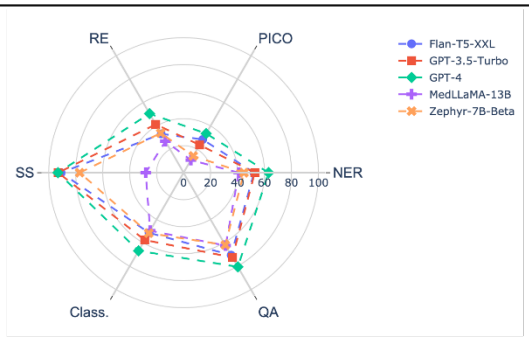


Task	Metrics	Dataset
Named-Entity Recognition (NER) Locating and classifying named entities from input text to pre-defined categories.	**F1 entity-level:** F1 score[a] calculated at named entity level.	BioCreative II Gene Mention Recognition (**BC2GM**)
		BC5CDR Drug/Chemical (**BC5-chem**)
		BC5CDR Disease (**BC5-disease**)
		JNLPBA: chemical recognition
		NCBI Disease: disease recognition
Populations, Interventions, Comparators, and Outcomes (PICO) Locating and classifying the above four elements from text.	**Macro F1 word-level:** Unweighted average of F1 score[a] across the four classes.	**EBM-PICO**: extract PICO from medical abstracts describing clinical trials
Relation Extraction (RE) Extracting semantic relationships from text.	**Micro F1:** Global average of F1 score[a] for all mentions of relationships.	**ChemProt**: chemical-protein interactions
		Drug-Disease Interaction (**DDI**)
		Gene-Disease Associations (**GAD**)
Sentence Similarity (SS) A measure of how similar two pieces of text are, or to what degree they express the same meaning.	**Pearson:** The covariance of the two variables divided by the product of their standard deviations.	**BIOSSES**: biomedical sentence similarity
Document Classification (Class.) Assigning a document one or more classes or categories.	**Average Micro F1:** Unweighted average of micro F1 score[a] for all classes.	Hallmarks of Cancer (**HoC**)
Question Answering (QA) Answering questions, sometimes for a specific domain, and often based on a collection of natural language texts or documents.	**Accuracy[b]:** Number of correct predictions divided by total number of predictions.	**BioASQ**: biomedical QA
		PubMedQA: biomedical QA based on PubMed abstracts

[a] $F_1 = TP/(TP + \frac{1}{2}(FP + FN))$, *TP: number of true positives, FP: number of false positives, FN: number of false negatives, TN: number of true negatives.*
[b] *Accuracy* $= (TP + TN)/(TP + TN + FP + FN)$

Figure 2. Normalized mean of task-specific scores for the best-performing prompt for each large language model (LLM) for the Biomedical Language Understanding and Reasoning Benchmark (BLURB)

TASK / MODEL	Flan-T5-XXL	GPT-3.5-Turbo	GPT-4	MedLLaMA-13B	Zephyr-7B-Beta
NER	52.25	52.80	63.00	40.94	44.73
PICO	28.42	23.78	33.49	10.80	14.32
RE	33.21	41.34	50.54	26.93	33.91
SS	90.88	93.02	93.18	28.08	77.04
Class.	51.36	57.57	66.81	49.65	51.74
QA	70.55	72.42	80.56	62.35	61.85



Note: 1) NER: Named-Entity Recognition; PICO: Populations, Interventions, Comparators, and Outcomes; RE: Relation Extraction; SS: Sentence Similarity; Class.: Document Classification; QA: Question Answering. 2) Score for each task and model combination displayed in this table is the unweighted average from the best scores we obtained from each model for all datasets included in the task, calculated as SUM(model best score for each dataset in the task) / (number of datasets in the task).

Figure 3. Mean performance of each large language model (LLM) for each task in the Biomedical Language Understanding and Reasoning Benchmark (BLURB) using short style versus long style zero-shot (i.e., no example provided to the model) prompts (see Appendix A for more details on prompt styles and templates).

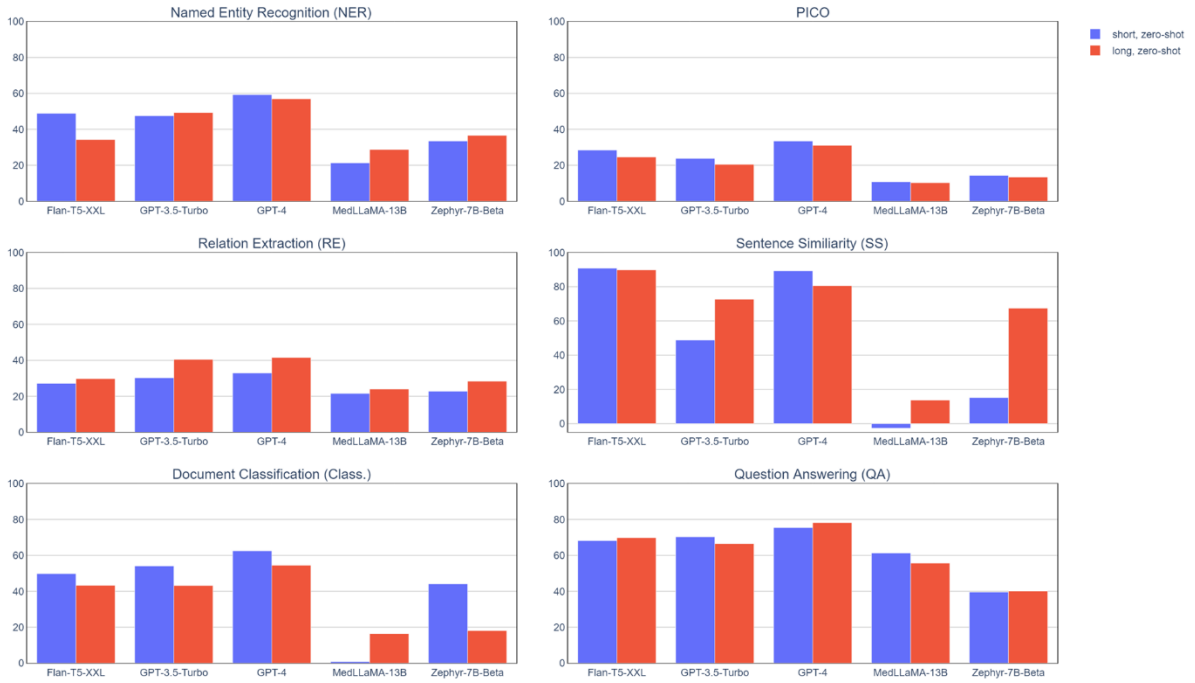
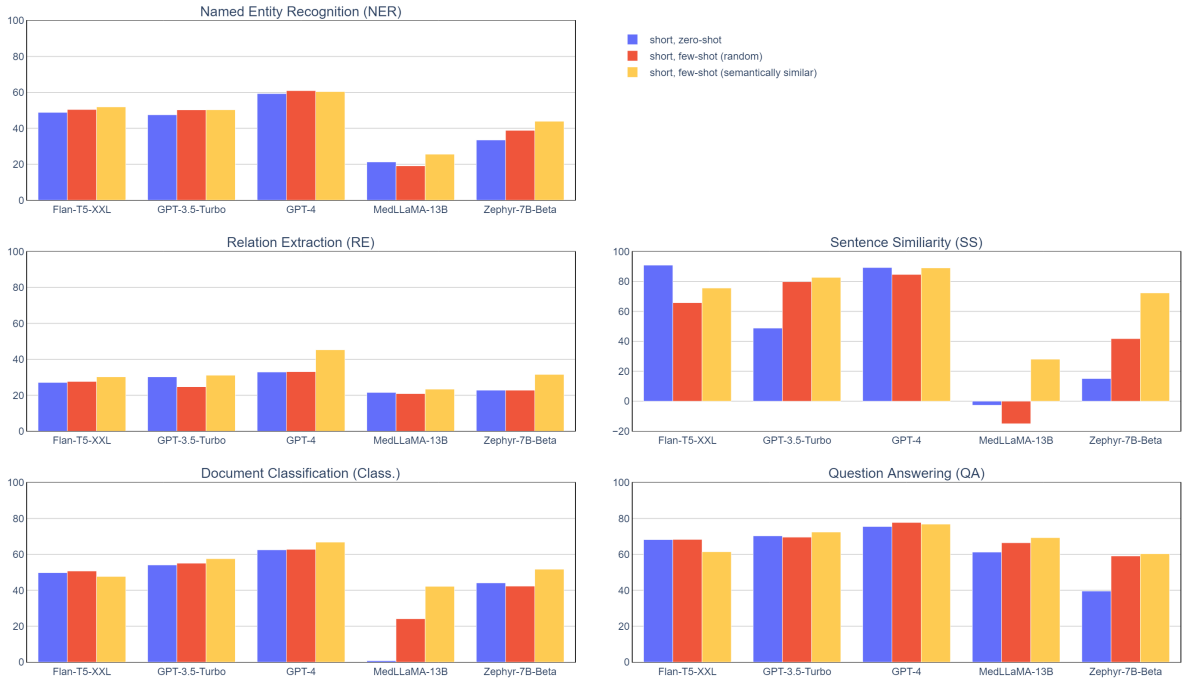


Figure 4. Mean performance of each large language model (LLM) for each task in the Biomedical Language Understanding and Reasoning Benchmark (BLURB) using different example selection methods: short, zero-shot: short prompt style, no example provided to the model; short, few-shot (random): short prompt style, three randomly selected examples provided to the model; short, few-shot (semantically similar): short prompt style, three examples selected based on semantic similarity provided to the model (see Appendix A for more details on prompt styles and templates).



References

1. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. Journal of Medical Systems: Springer; 2023. p. 33.
2. WHO. https://www.who.int/news/item/16-05-2023-who-calls-for-safe-and-ethical-ai-for-health.
3. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, Scales N, Tanwani A, Cole-Lewis H, Pfohl S. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172-180.
4. Omiye JA, Lester JC, Spichak S, Rotemberg V, Daneshjou R. Large language models propagate race-based medicine. *NPJ Digital Medicine*. 2023;6(1):195.
5. Tang L, Sun Z, Idnay B, Nestor JG, Soroush A, Elias PA, Xu Z, Ding Y, Durrett G, Rousseau JF. Evaluating large language models on medical evidence summarization. *npj Digital Medicine*. 2023;6(1):158.
6. Shah NH, Entwistle D, Pfeffer MA. Creation and adoption of large language models in medicine. *Jama*. 2023;330(9):866-869.
7. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, others. Language models are few-shot learners. Advances in neural information processing systems2020. p. 1877-1901.
8. Liu J, Shen D, Zhang Y, Dolan WB, Carin L, Chen W. What Makes Good In-Context Examples for GPT-3? 2022:100–114.
9. Wang S, Sun X, Li X, Ouyang R, Wu F, Zhang T, Li J, Wang G. Gpt-ner: Named entity recognition via large language models. arXiv preprint arXiv:2304104282023.
10. Yu G, Robert T, Hao C, Lucas M, Naoto U, Xiadong L, Tristan N, Jianfeng G, Hoifung P. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Transactions on Computing for Healthcare*. 2021:1-23.
11. Azure OpenAI Service models, 2024; https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/models.
12. Chung HW, Hou L, Longpre S, Zoph B, Tay Y, Fedus W, Li Y, Wang X, Dehghani M, Brahma S, Webson A, Gu SS, Dai Z, Suzgun M, Chen X, Chowdhery A, Castro-Ros A, Pellat M, Robinson K, Valter D, Narang S, Mishra G, Yu A, Zhao V, Huang Y, Dai A, Yu H, Petrov S, Chi EH, Dean J, Devlin J, Roberts A, Zhou D, Le QV, Wei J. Scaling Instruction-Finetuned Language Models. 2022.
13. Tunstall L, Beeching E, Lambert N, Rajani N, Rasul K, Belkada Y, Huang S, von Werra L, Fourrier C, Habib N, others. Zephyr: Direct distillation of lm alignment. arXiv preprint arXiv:2310169442023.
14. Wu C, Lin W, Zhang X, Zhang Y, Wang Y, Xie W. PMC-LLaMA: Towards Building Open-source Language Models for Medicine. 2023.
15. Wu C, Lin W, Zhang X, Zhang Y, Xie W, Wang Y. PMC-LLaMA: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*. 2024:ocae045.
16. Richardson WS, Wilson MC, Nishikawa J, Hayward RS. The well-built clinical question: a key to evidence-based decisions. *ACP J Club*. Nov-Dec 1995;123(3):A12-3.
17. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, Naumann T, Gao J, Poon H. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*. 2021;3(1):1-23.
18. Chen Q, Du J, Hu Y, Keloth VK, Peng X, Raja K, Zhang R, Lu Z, Xu H. Large language models in biomedical natural language processing: benchmarks, baselines, and recommendations. 2023.
19. Jahan I, Laskar MTR, Peng C, Huang J. Evaluation of ChatGPT on Biomedical Tasks: A Zero-Shot Comparison with Fine-Tuned Generative Transformers. 2023:326–336.
20. Wang Y, Si S, Li D, Lukasik M, Yu F, Hsieh C-J, Dhillon IS, Kumar S. Two-stage LLM fine-tuning with less specialization and more generalization. 2024 https://arxiv.org/pdf/2211.00635v3.

21. Yang H, Zhang Y, Xu J, Lu H, Heng PA, Lam W. Unveiling the Generalization Power of Fine-Tuned Large Language Models, 2024. *arXiv preprint arXiv:240309162*.
22. Yasunaga M, Leskovec J, Liang P. Linkbert: Pretraining language models with document links. arXiv preprint arXiv:2203158272022.
23. Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E, Le QV, Zhou D. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*. 2022;35:24824-24837.
24. Lester B, Al-Rfou R, Constant N. The power of scale for parameter-efficient prompt tuning, 2021. *arXiv preprint arXiv:210408691*.
25. Li L, Zhang Y, Chen L. Prompt distillation for efficient llm-based recommendation. 2023:1348-1357.
26. Peng C, Yang X, Smith KE, Yu Z, Chen A, Bian J, Wu Y. Model Tuning or Prompt Tuning? A Study of Large Language Models for Clinical Concept and Relation Extraction, 2023. *arXiv preprint arXiv:231006239*.