M-QALM: A Benchmark to Assess Clinical Reading Comprehension and Knowledge Recall in Large Language Models via Question Answering

Anand Subramanian[β,*], **Viktor Schlegel**[α,γ], **Abhinav Ramesh Kashyap**[α], **Thanh-Tung Nguyen**[α], **Vijay Prakash Dwivedi**[α] and **Stefan Winkler**[α,β]

[α]: ASUS Intelligent Cloud Services (AICS), Singapore
[β]: National University of Singapore, Singapore
[γ]: University of Manchester, United Kingdom

anands@u.nus.edu, winkler@nus.edu.sg
{viktor_schlegel,abhinav_kashyap,thomas_nguyen,vijay_dwivedi}@asus.com

Abstract

There is vivid research on adapting Large Language Models (LLMs) to perform a variety of tasks in high-stakes domains such as healthcare. Despite their popularity, there is a lack of understanding of the extent and contributing factors that allow LLMs to recall relevant knowledge and combine it with presented information in the clinical and biomedical domain—a fundamental pre-requisite for success on downstream tasks. Addressing this gap, we use Multiple Choice and Abstractive Question Answering to conduct a large-scale empirical study on 22 datasets in three generalist and three specialist biomedical sub-domains. Our multifaceted analysis of the performance of 15 LLMs, further broken down by sub-domain, source of knowledge and model architecture, uncovers success factors such as instruction tuning that lead to improved recall and comprehension. We further show that while recently proposed domain-adapted models may lack adequate knowledge, directly fine-tuning on our collected medical knowledge datasets shows encouraging results, even generalising to unseen specialist sub-domains. We complement the quantitative results with a skill-oriented manual error analysis, which reveals a significant gap between the models' capabilities to simply recall necessary knowledge and to integrate it with the presented context. To foster research and collaboration in this field we share M-QALM—our resources, standardised methodology, and evaluation results—with the research community to facilitate further advancements in clinical knowledge representation learning within language models.

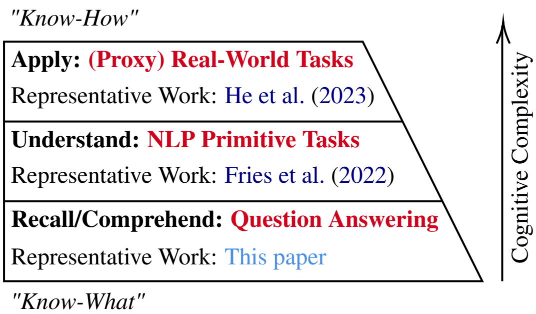

Figure 1: The landscape of LLM evaluation in the medical domain with representative evaluation tasks, organised by Bloom's taxonomy of learning objectives (bold) (Bloom, 1956).

1 Introduction

Recent success in the application of proprietary large language models in the knowledge-intensive medical domain (Singhal et al., 2023a,b) has sparked vivid research interest in applying smaller, more readily available open-source LLMs to various settings in the clinical and biomedical domains. Examples of tasks include summarization of clinical text (Veen et al., 2023), automatic note generation for physicians (Ben Abacha et al., 2023b), and condensation of doctor-patient dialogues (Ben Abacha et al., 2023a; Toma et al., 2023). More broadly, open-source LLMs have been adapted to the domain to serve as foundational clinical models (Han et al., 2023; Wu et al., 2023; Toma et al., 2023; Bolton et al., 2022; Li et al., 2023).

The success of such adaptation is typically established by measuring the performance on downstream tasks, by means of token overlap or semantic similarity-based metrics (Lin, 2004; Zhang et al., 2020). To address their inherent weaknesses (Schlegel et al., 2022; Gatt and Krahmer, 2018), research attempts to incorporate specific dimensions, such as factuality or faithfulness (Umapathi et al., 2023). Two important problems remain, however. Firstly, Natural Language Generation (NLG) evaluation metrics are merely approximations of the phenomena they aim to measure, and their effectiveness is typically established by the degree of correlation to human judgements of the evaluated criteria

[*]Work done during an internship at AICS

Findings of the Association for Computational Linguistics: ACL 2024, pages 4002–4042
August 11-16, 2024 ©2024 Association for Computational Linguistics

Dataset	Type	Size	Domain
USMLE (Jin et al., 2021)	MCQA	10178/1272/1273	General Medical
MEDMCQA (Pal et al., 2022)	MCQA	182822/4183/6150	General Medical
BIOASQ-MCQ (Tsatsaronis et al., 2015; Krithara et al., 2023)	MCQA	975/173/123	General Biomedical
HEADQA (Vilares and Gómez-Rodríguez, 2019)	MCQA	2657/1366/2742	General Medical
PROCESSBANK (Berant et al., 2014)	Context + MCQA	358/77/150	Biological Processes
PUBMEDQA (Jin et al., 2019)	Context + MCQA	400/100/500	General Biomedical
MMLU (Hendrycks et al., 2021)	MCQA	30/NA/1089	General Medical/Clinical
BIOMRC-Tiny A (Pappas et al., 2020)	Context + MCQA	NA/NA/30	General Biomedical
BIOMRC-Tiny B (Pappas et al., 2020)	Context + MCQA	NA/NA/30	General Biomedical
OPHTH (Raimondi et al., 2023; RCOphth, 2022a,b)	MCQA	NA/NA/92	Ophthalmology
QA4MRE-(Alzheimer's QA) (Morante et al., 2012)	MCQA	NA/NA/40	Alzheimer's Disease
Total Questions across Splits	-	197420/7171/12219	-
LIVEQA (Abacha et al., 2017; Ben Abacha and Demner-Fushman, 2019)	AQA	NA/NA/131	Consumer Health
MEDIQA-ANS (Savery et al., 2020)	AQA	NA/NA/156	Consumer Health
BIOASQ-QA (Tsatsaronis et al., 2015; Krithara et al., 2023)	AQA	4733/697/363	General Biomedical
MASHQA (Zhu et al., 2020)	AQA	27728/3587/3493	Consumer Health
MEDQUAD (Ben Abacha and Demner-Fushman, 2019)	AQA	14068/981/1358	General Medical
MEDINFO (Ben Abacha et al., 2019)	AQA	NA/NA/663	Consumer Medication
Total Questions across Splits	-	46529/5265/6164	-

Table 1: Overview of M-QALM datasets. Size is presented in terms of train/val/test splits. Manual train/val splits are created for BIOASQ-MCQ, PROCESSBANK, PUBMEDQA, BIOASQ-QA and MEDQUAD. We use 6 subsets of the MMLU dataset that pertain to testing clinical and medical knowledge (Singhal et al., 2023a).

(Huang et al., 2021). Secondly, an (offline) evaluation setup is *functionally grounded* and serves as a proxy of a real-world application scenario, but the transferability of insights from functionally-grounded to application-grounded evaluation is barely discussed (Doshi-Velez and Kim, 2017). Taken together, these problems might taint the credibility of conclusions about the successful adaptation of LLMs drawn from such experiments.

Given such difficulties, we approach the problem of evaluating LLM adaptation from a complementary angle. Specifically, we ask: *Do* LLMs *possess the necessary pre-requisites to succeed in the clinical and medical domains*? Without an established theory of how knowledge is acquired and organised in LLMs, the present work is guided by the established theories of knowledge acquisition in *humans* (Adams, 2015). Typical NLG tasks, such as summarisation, are higher-level cognitives that require the understanding of learned knowledge and its application in new contexts (Bloom, 1956). They build on the most fundamental capability of *reading comprehension* (Kintsch, 1988): the construction of a text-base and its integration with previously acquired background knowledge. In NLP research, this process is evaluated by *open-book Question Answering* (QA), the task of either generating (abstractive, AQA) or selecting among presented options (multiple-choice, MCQA) the correct answer for a question, where potentially not all necessary information is included in the question or the presented context. MCQA evaluation does not suffer from the issues pertaining to NLG metrics, as performance is established by exact match. Thus, conclusions obtained from such evaluations tend to be more robust, if the quality of the benchmark is sufficient.

Therefore, in this paper we focus on the task of QA, to evaluate knowledge recall and comprehension pre-requisites of LLMs for successful adaptation to the medical domain. We present an exhaustive, publicly available QA benchmark called M-QALM including 16 MCQA datasets. To enable future research on NLG-based QA, we complement M-QALM by 6 high-quality AQA datasets, where the ground-truth answer is an unconstrained string. With such a standardized benchmark, we conduct an extensive evaluation of the capabilities of openly available general-purpose and medical LLMs, both "out-of-the-box" and after fine-tuning on M-QALM. Our findings provide insights into the strengths and weaknesses of different LLMs across a range of datasets, question categories and QA tasks. Overall, we find their performance lacking, both compared to humans and to proprietary LLMs. Further analysis reveals promising tendencies of domain-specific pre-training and fine-tuning to bridge this gap and to generalise to new QA datasets.

2 Related Work

Large open-domain QA benchmarks The availability of QA datasets from multiple domains and sources has enabled the curation of large and diverse QA benchmarks (Dua et al., 2019; Fisch et al., 2019; Talmor and Berant, 2019). Such resource collections enable researchers to perform large-scale empirical studies to understand how well language models can generalise to new questions from new domains or sources, or how fine-tuning can impact this performance. While multiple studies exist in the general domain, to the best of our knowledge, no such large-scale study has been carried out for QA in the clinical domain. In this paper we aim to address this gap.

Evaluation in the clinical domain Datasets that evaluate the lowest-level cognitive task of knowledge recall and reading comprehension in the medical domain have been proposed before (Jin et al., 2021; Vilares and Gómez-Rodríguez, 2019; Pal et al., 2022). They feature questions commonly found in examinations like the US Medical Licensing Exam (USMLE). M-QALM unifies the existing literature by incorporating licensing exam questions from diverse regions, such as India and Spain. We go beyond the scope of the general medical domain, covering specialist topics such as ophthalmology and Alzheimer's disease.

Beyond factual recall and comprehension, Fries et al. (2022) collected a unified bio-medical benchmark, featuring NLP primitives such as sentence(-pair) classification or entity recognition and linking. Aiming at higher, more task-specific cognitives, Singhal et al. (2023a) introduced MultiMedQA, including HealthSearchQA, which requires models to generate high-quality free-form answers. Similarly, He et al. (2023) introduced a multi-domain benchmark for evaluating generation and classification capabilities on a diverse set of in-hospital downstream tasks. Other researchers looked to evaluate the quality and factuality of generation (Umapathi et al., 2023) and synthesised general-purpose medical instructions (Fleming et al., 2023). Our work is complementary, because we evaluate knowledge recall and comprehension as a pre-requisite of higher-level cognitive tasks, such as understanding and application—the focus of previously discussed works.

3 M-QALM Datasets

The primary goal of M-QALM is to develop a comprehensive, open-source repository of medical QA datasets to assess the recall of medical knowledge in LLMs. To obtain such a collection, we perform an exhaustive literature and resource search using the terms "clinical OR medical", "Question Answering OR QA" and include a dataset or resource if it satisfies the following criteria: *(i)* The language is English, as medical documents are usually written in English, even in non-English-speaking countries; *(ii)* The questions and answers are on general, specialist, or consumer-facing medical topics; *(iii)* The resource is openly available without restrictive licensing or data agreements; *(iv)* The resource evaluates the task of MCQA or AQA; *(v)* The ground truth is collected or reviewed by domain experts.

The result is M-QALM—a comprehensive collection of 22 datasets designed to thoroughly evaluate the clinical knowledge of LLMs. Table 1 gives an overview of the collected MCQA and AQA datasets, including task formulation, size and domain. Refer to the Appendix for further details on each dataset.

Knowledge source categorization The MCQA datasets within the M-QALM benchmark cover a diverse range of medical domains. To be able to perform fine-grained analysis of both the topics covered in these datasets as well as model performance, we categorise the MCQA datasets into eleven high-level categories, representing different facets of medical knowledge. To do so, we leverage available meta-data from the source datasets MEDMCQA, HEADQA, MMLU and BIOASQ-MCQ. We categorize the PROCESSBANK, PUBMEDQA and BIOMRC datasets into a distinct twelfth `Within Context` category, as the relevant knowledge is presented in the context. USMLE and QA4MRE lack the necessary meta-data, thus we train a BioBERT-based classifier (Lee et al., 2019) to assign questions into one of the eleven elicited categories using the labels from the other datasets. The classifier achieves 71.56% (micro-)averaged F1 score on a held-out test set, which we deem sufficient.

Table 4 shows that nearly half of all questions (47%) fall into the `Basic and Life Sciences` and `General Medicine` category. `Diagnostic Sciences`, `Women's and Children's Health` and `Pharmacology and Anesthesia` account for a further 30% of questions.

4 Empirical Evaluation

We investigate how well existing, open-source LLMs are able to recall clinical knowledge and integrate it into a given context in order to succeed on our benchmark. Specifically, we focus on performance in the zero-shot setting, and after fine-tuning on M-QALM training portions.

In the **Zero-shot** setting:

- **RQ1.** How well do open-source LLMs recall necessary clinical knowledge when they are tested on M-QALM?
- **RQ2.** Does open-domain instruction fine-tuning of LLMs improve their ability to do so?
- **RQ3.** Does *domain-specific* fine-tuning improve performance on M-QALM?

In the **Fine-tuned** setting:

- **RQ4.** Does finetuning on M-QALM improve performance on unseen data from datasets seen during training?
- **RQ5.** Does fine-tuning improve performance on *unseen* M-QALM datasets?

4.1 Study Setup

To seek evidence for **RQs 1-3** empirically, we evaluate several LLMs and their instruction-tuned versions on the test splits of M-QALM in zero-shot[1] manner. To answer **RQ4** and **RQ5**, we fine-tune LLMs on the training portion of M-QALM and evaluate on test splits of datasets both seen and unseen during training. We complement our evaluation with additional automated and manual error analyses to identify causes for model successes and failures.

Models: To assess the zero-shot capabilities of models (**RQ1** and **RQ2**), we include a diverse array of open-source decoder-only models with parameter scales ranging from 3B-13B. We use models from MPT and MPT-Instruct (7B) (MosaicML, 2023), Falcon and Falcon-Instruct (7B) (Almazrouei et al., 2023), LLaMA 1 (7B and 13B) (Touvron et al., 2023a), LLaMA 2 and LLaMA 2-chat (7B and 13B) (Touvron et al., 2023b). In addition to these models, we also use two instruction fine-tuned encoder-decoder models: Flan-T5 (3B and 11B) (Wei et al., 2021). Models with *Instruct* or *Chat* appended to their names are instruction fine-tuned (Ouyang et al., 2022) versions of their base models. The details of the models are given in Table 10. To address **RQ3**, we evaluate ChatDoctor (7B) (Li et al., 2023), MedAlpaca (7B) (Han et al., 2023) and PMC-LLama (Wu et al., 2023). To address **RQ4**, we fine-tune models using the training set of the M-QALM datasets. When official validation splits are unavailable, we employ a random split of up to around 20% of the data for validation purposes. If no training datasets are available, we do not use this dataset for fine-tuning and only consider the test split of the respective datasets to answer **RQ5**. For evaluating AQA, we use a subsampled version of the test sets of MASHQA (500 questions) and MEDQUAD (200 questions by sampling 100 questions from the two holdout websites), while we use the other datasets as they are. For MCQA, similar to Singhal et al. (2023a), we evaluate all models on the validation set of MEDMCQA since the answers for the test set are not released publicly.

Finetuning and hyperparameters: Since the number of parameters for most of our models is in the billions, we follow a more accepted practice of using parameter-efficient fine-tuning, specifically QLora and 4-bit quantization (Dettmers et al., 2023). We utilize 8-bit quantization for evaluating Flan-T5 (11B), LLaMA 1 (13B), LLaMA 2 (13B) and LLaMA 2-Chat (13B) (Dettmers et al., 2022). We use A100-40G GPUs for all our experiments. The other hyper-parameters used to train our models are reported in the Appendix (Table 11).

Evaluation measures: We use Accuracy to measure the performance of the model on MCQA datasets; for AQA datasets, we use ROUGE-L (Lin, 2004), BERTScore (Zhang et al., 2020) (based on deberta-xlarge-mnli) and METEOR (Banerjee and Lavie, 2005), which is found to correlate better with human judgments than other metrics on AQA (Chen et al., 2019).

5 Results and Analysis

In this section, we report and analyse the findings of our empirical study.

5.1 Zero-shot Evaluation Results

Table 2 shows the dataset-averaged scores of the zero-shot evaluation of language models as evidence towards **RQs 1-3**. Note that in this way, each dataset contributes equally to the average, regardless of its size. Micro-averaged MCQA accuracy scores are reported in Table 4. However, these are biased towards datasets with more examples (i.e., MEDMCQA). While the results between micro- and by-dataset-averaged metrics might differ in detail (consult Appendix D for a break-down), the

[1] For MCQA evaluation in the zero-shot setting (where models are not explicitly fine-tuned for MCQA tasks), we use a 1-shot prompt—giving an example to the model, and find that it adheres better to the MCQA format and the standard 5-shot prompt for MMLU datasets.

		MCQA	AQA		
		Acc	RL	BS	MTR
Base	LLaMA 1 (7B)	31.9	14.0	54.2	20.5
	LLaMA 1 (13B)	44.1	14.4	54.0	20.3
	LLaMA 2 (7B)	42.9	14.9	55.3	21.1
	LLaMA 2 (13B)	47.1	15.0	56.4	22.5
	MPT (7B)	27.6	13.3	52.6	21.1
	Falcon (7B)	34.7	14.0	54.1	20.0
Instruction tuned	LLaMA 2-chat (7B)	45.9	15.0	58.0	23.3
	LLaMA 2-chat (13B)	50.3	15.3	58.0	23.6
	MPT-Instruct (7B)	31.6	15.8	59.7	15.6
	Falcon-Instruct (7B)	31.8	17.2	62.4	17.4
	Flan-T5 (3B)	51.8	10.8	55.0	7.4
	Flan-T5 (11B)	56.5	11.5	56.3	8.2
Adapted	ChatDoctor (7B)	42.8	17.4	62.3	18.7
	MedAlpaca (7B)	48.8	15.5	58.9	15.6
	PMC-LLama (13B)	53.7	19.7	60.7	19.0
	Random Baseline	27.7	-	-	-

Table 2: Zero-shot performance of base (top), instruction-tuned models (middle) and domain-adapted (bottom) models. Metrics are **Acc**uracy for MCQA; **R**ouge-**L**, **BERTS**core, and **METEOR** for AQA.

mean absolute difference between the metrics for all models is 4.2, which suggests that reported trends do not depend on the averaging method.

Table 2 highlights that LLMs exhibit **strong zero-shot capability on MCQA and AQA datasets**, corroborating the findings of Singhal et al. (2023a). Considering LLMs of the same size (e.g., 7B), LLaMA 2 performs best, possibly due to larger diversity in pre-training data—LLaMA 2 is trained on the most tokens. Another difference is the mixture of datasets used for pre-training, which is not revealed in some cases (c.f. Table 10 in Appendix).

Unsurprisingly, across all models of the same architecture, **scale predicts model performance**, even without domain-specific adaptation of LLMs on the medical domain. For example, LLaMA 2 (13B) performs better on MCQA (+4.2 Accuracy improvement) compared to the 7B version. Figure 5 in the Appendix shows the relationship between the number of parameters and performance.

To address **RQ2**, we investigate whether improvements from instruction fine-tuning also apply to the clinical domain of M-QALM. The results are reported in the middle part of Table 2.

Surprisingly, **instruction fine-tuned models perform better** than their corresponding *Base* versions, despite the fact that the instruction set used for fine-tuning contains only tasks in the general domain—see Table 10 and compare *-Instruct/Chat (middle) with their base versions (top). Among them, Flan-T5 models exhibit the best zero-shot performance on MCQA, outperforming comparable decoder-only models. Seemingly, instruction fine-tuning enables models to obtain representations of question and context which are beneficial for fact recall.

We note that **bigger models are not always better**—the choice of model architecture and dataset for instruction fine-tuning can have a bigger impact on performance than model size alone. For example the encoder-decoder Flan-T5 (3B) model outperforms LLaMA 2-chat (13B) on the MCQA task, despite being four times smaller.

The performance of domain-adapted models is reported in Table 2 (bottom), as evidence for **RQ3**. For MCQA, both MedAlpaca and ChatDoctor indeed exhibit improvements in Accuracy over their respective 7B and 13B LLaMA 1 base versions; however they fail to reach the strong zero-shot performance of Flan-T5 (11B).

In contrast, PMC-LLama performs well due to continued pre-training on biomedical corpora before instruction tuning on biomedical and clinical datasets. The latter results in exceptionally high scores on the MEDINFO AQA dataset (See Table 20 in Appendix). This dataset, along with LIVEQA, was used as part of the instruction tuning process, leading to evaluation on these dataset not being "zero-shot"[2]. Scores on LIVEQA, however, are not inflated, compared to LLaMA 2(-chat) (13B). This is possibly because we use a filtered version of LIVEQA which contains only challenging answers with sufficiently good expert quality rating. PMC-LLama demonstrates significant improvements over other open-source LLMs on MCQA datasets such as USMLE, MEDMCQA and MMLU.

In summary, we conclude that while available LLMs adapted to the medical domain successfully improve performance of the adapted models, they appear to have **no improved domain knowledge** compared to other available open-domain models. Evaluating these adaptation techniques on stronger base models is an exciting avenue for future research.

Importantly, none of the evaluated open-source LLMs outperform humans: While the passing score

[2] https://huggingface.co/datasets/axiong/pmc_llama_instructions

for USMLE is 60% [3], we observe the best zero-shot scores for USMLE are 43% for LLaMA 2, and 54% for the domain-adapted PMC-LLama, both below the passing score. Meanwhile, GPT-4 (OpenAI, 2023) with a customized prompting strategy labeled MedPrompt (Nori et al., 2023) achieves 90.2%, while Med-PALM 2 (Singhal et al., 2023b) achieves scores of 86.5% on USMLE. Similarly, for the PubmedQA dataset, human performance is 78% (Jin et al., 2019), compared to 72.4% of Flan-T5. To summarize: While available LLMs exhibit performance significantly higher than random chance "out-of-the-box", there is still a **substantial gap compared to humans and proprietary LLMs** (Singhal et al., 2023a,b) (see Appendix B).

5.2 Impact of Fine-tuning

Given the scale of M-QALM, we are able to fine-tune models on parts of the data, to address **RQ4** and **RQ5**. We fine-tune four models on MCQA and AQA separately, given the different nature of these datasets, but joint fine-tuning on both MCQA and AQA did not yield significantly different results.

	MCQA	AQA		
	Acc	RL	BS	MTR
LLaMA 2 (7B)	53.5 $_{+10.6}$	17.7 $_{+2.8}$	60.8 $_{+5.5}$	16.9 $_{-4.2}$
Falcon (7B)	49.3 $_{+14.6}$	17.4 $_{+3.4}$	60.4 $_{+6.3}$	17.1 $_{-2.9}$
MPT (7B)	53.2 $_{+25.6}$	17.3 $_{+4.0}$	60.0 $_{+7.4}$	17.2 $_{-3.9}$
Flan-T5 (3B)	52.9 $_{+1.1}$	15.9 $_{+5.1}$	56.8 $_{+1.8}$	15.6 $_{+8.2}$

Table 3: Model fine-tuning is performed either on MCQA or AQA datasets. Reported are **Acc**uracy for MCQA; **R**ouge-**L**, **B**ERT**S**core, and **M**E**T**EO**R** for AQA. Subscripts indicate improvement over zero-shot versions.

We fine-tune the models only on the MCQA subset of datasets first (cf. Table 3). We find that the **fine-tuned models perform better** compared to their non-fine-tuned counterparts. Decoder-only models like MPT (7B) benefit more than others (+25.6 Accuracy improvement). Fine-tuning models on the data seems to close the gaps introduced by different model architectures and pre-training data: The standard deviation of the evaluated models' accuracies reduces from 9.0 in the zero-shot setting to 1.7 after fine-tuning. This suggests that LLMs can benefit from task-specific fine-tuning to address seemingly sub-optimal architecture or pre-training conditions. For AQA, Flan-T5 benefits more from fine-tuning compared to the decoder-only models, possibly by better aligning generated outputs to the expected format of the answer. Decoder models present inconsistent results with improvements in ROUGE-L and BERTScore at the expense of lower METEOR scores, which raises concerns about the reliability of the AQA metrics.

Scaling up models introduces practical problems of deploying the model in real-world scenarios—smaller models may be preferred to larger ones due to faster inference times and lower memory footprints. We find that **fine-tuning helps compensate for scale**. Fine-tuned LLaMA 2 (7B) significantly outperforms zero-shot LLaMA 2 (13B) (+6.4 Accuracy gain on MCQA, +2.7 ROUGE-L gain and +4.4 BERTScore gain on AQA). Similarly, fine-tuned Flan-T5 (3B) outperforms zero-shot LLaMA 2 (13B) on 8 out of 16 MCQA datasets (see Tables 13 and 15).

In summary, we conclude that task-specific fine-tuning improves performance, mitigating weaknesses due to size, architecture and training data.

Finally, we report the potential of LLMs fine-tuned on in-domain data to generalize to medical datasets unseen during training to answer **RQ5**. To this end, during fine-tuning, we hold out ten MCQA and four AQA datasets presented in Figures 2 and 3.

Figure 2 shows the performance of LLaMA 2 (7B) and Flan-T5 (3B) models on the four held-out AQA evaluation sets. While LLaMA 2 does not appear to generalise to unseen AQA datasets, Flan-T5's scores improve across the board. We note however, that this result might depend on the choice of metric, as Figures 6 and 7 in the Appendix paint a more mixed picture. Indeed, across all conducted experiments, only ROUGE-L scores show a statistically significant Spearman rank correlation with the reliable MCQA accuracy measure ($r = 0.616$, $p = 0.008$, more details in Appendix C). This suggests that other metrics used are either a sub-optimal choice or that they measure another, complementary aspect captured neither by Accuracy nor ROUGE-L. These findings highlight the **low robustness of overlap-based NLG metrics** discussed in the introduction.

Investigating the more robust MCQA setting, Figure 3 (comparing blue ZS with orange AQA-FT bars) shows that **fine-tuning on AQA does not improve performance on unseen MCQA datasets**. This suggests that higher scores on unseen AQA datasets might stem from better aligning genera-

[3] https://www.usmle.org/bulletin-information/scoring-and-score-reporting

	Category	Support	Flan-T5 (ZS)	Flan-T5 (FT)	MPT (ZS)	MPT (FT)	Falcon (ZS)	Falcon (FT)	LLaMA 2 (ZS)	LLaMA 2 (FT)
Domain	General Medical	9275	37.9	44.9	26.5	49.5	29.5	46.9	37.6	**50.5**
	General Biomedical	683	64.4	**71.0**	32.4	70.0	56.7	68.4	58.9	68.5
	Biological	294	**71.4**	70.4	39.5	71.1	39.1	58.2	57.8	68.4
Knowledge Source	General Medicine	2675	38.0	43.2	26.0	46.4	30.1	46.4	36.6	**50.0**
	Basic and Life Sciences	2235	38.9	44.3	26.9	**52.6**	30.6	49.4	40.0	52.5
	Dental and Oral Health	1318	34.8	42.9	25.9	**44.3**	30.7	43.8	36.1	44.2
	Pharmacology and Anesthesia	784	39.7	48.1	29.0	55.6	28.8	54.0	42.9	**59.4**
	Within Context	710	74.1	**75.2**	37.2	71.5	52.7	66.5	60.8	67.7
	Diagnostic Sciences	640	32.2	43.1	26.4	**51.1**	30.3	46.4	37.2	47.5
	Supportive and Preventive Services	599	48.2	**56.6**	23.7	55.1	27.9	48.1	39.9	56.3
	Women's and Children's Health	507	30.2	42.6	27.2	**51.7**	28.4	43.0	34.3	49.9
	Mental and Behavioral Health	496	50.0	57.9	29.4	55.4	31.5	49.2	40.7	**59.1**
	Sensory Organs	205	29.8	42.0	27.8	**45.4**	28.8	42.4	33.2	42.0
	Miscellaneous	45	42.2	44.4	20.0	**60.0**	24.4	44.4	31.1	40.0
	Musculoskeletal and Dermatology	38	18.4	26.3	18.4	**44.7**	34.2	42.1	28.9	**44.7**
	Micro-averaged Accuracy	10252	40.6	47.4	27.3	51.5	31.6	48.6	39.6	**52.2**
	Category-averaged Accuracy	12	39.7	47.2	26.5	**52.8**	31.5	48.0	38.5	51.1

Table 4: Performance of LLMs in the zero-shot and fine-tuned setting across various categories on the test set.

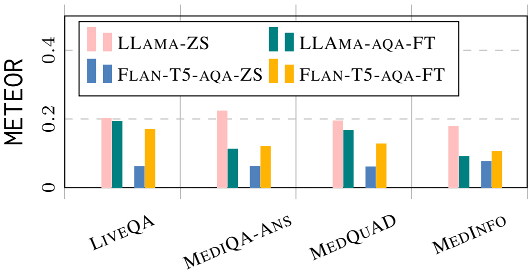

Figure 2: Performance of base and AQA-fine-tuned LLaMA 2 and Flan-T5 models on unseen AQA test sets.

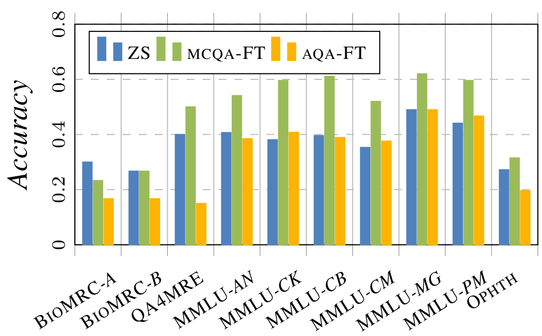

Figure 3: Performance of base, MCQA-tuned, and AQA-tuned LLaMA 2 model on unseen MCQA test sets.

tions to the expected answer form of AQA answers, which shows improvements in some of the AQA metrics, rather than acquiring additional medical knowledge during fine-tuning. While this could also due to a domain shift between the training and holdout datasets, this is not supported by the performance drop on MedQuaD, which, by this theory, should exhibit improved performance, since its domain is "General Medical", and would thus be in-domain.

Figure 3 (comparing blue ZS with green MCQ-FT) suggests that models indeed can learn to acquire domain-specific knowledge during fine-tuning, as MCQA-tuned models consistently perform better than their zero-shot counterparts. This seemingly contradicts the previous finding that models fail to acquire additional medical knowledge when fine-tuned on the AQA datasets.

Further analysis indicates that the reported generalisation capabilities might be over-stated, as evaluation questions from the unseen datasets have semantically similar counterparts in the fine-tuning data. However, a manual analysis of the cases where fine-tuned models outperform their zero-shot counter-parts reveals that only about 60% of the improvement can be explained by the presence of such similar examples. Details of this analysis are reported in Appendix D.

Based on these findings, we conclude that **fine-tuning can serve as a partial solution for achieving generalisable adaptation to the medical domain**.

6 Error Analysis

In this section, we analyze the errors of LLMs on MCQA datasets.

6.1 Category-wise and Manual Error Analysis

To better understand the performance of zero-shot and fine-tuned models across MCQA, we analyze

Reasoning Type	Support	Flan-T5 (ZS)	Flan-T5 (FT)	MPT (ZS)	MPT (FT)	Falcon (ZS)	Falcon (FT)	LLaMA 2 (ZS)	LLaMA 2 (FT)
Recall	131	48.1	49.6	23.7	**51.1**	31.3	49.6	47.3	**51.1**
Reading Comprehension	59	27.1	39.0	27.1	35.6	40.7	**47.5**	33.9	42.4
Quantitative/Arithmetic	10	**40.0**	30.0	10.0	20.0	30.0	**40.0**	30.0	30.0

Table 5: Performance of LLMs in zero-shot and fine-tuned settings on three reasoning types identified in M-QALM.

them broken down by sub-domain and knowledge source. We calculate the accuracy of the models in their zero-shot and fine-tuned settings for each category, as shown in Table 4.

Models tend to perform better on the biological and biomedical sub-domains. We posit as the reason for this that biomedical information is more readily available in the pre-training corpora of the models, e.g., in the form of biomedical abstracts (see also Table 10 in the Appendix). Furthermore, fine-tuning improves performance for all categories, but the gaps between medical and biomedical domains persist, indicating that medical questions are indeed harder to answer, even though they prevail in the training set. Perhaps more worryingly, the Consumer Health AQA scores do not improve as much as for other domains, even after fine-tuning (see Appendix, Table 19).

For knowledge sources, fine-tuned `Flan-T5` (3B) excels in `Within Context` and `Supportive and Preventive Services`, also showing strong zero-shot capabilities in these categories, perhaps due to architecture or pre-training data. Similarly, fine-tuned `MPT` (7B) and `LLaMA 2` (7B) show superior performance across categories. However, despite fine-tuning benefits, models still underperform in areas like `General Medicine`, `Basic and Life Sciences`, and `Dental and Oral Health`, which form the majority of the benchmark. Overall, we conclude that **fine-tuning improves model performance in sub-domains, but knowledge gaps still persist across different domains and knowledge sources**.

Finally, we sample 200 MCQA-questions from M-QALM evaluation data, and annotate the type of reasoning required to solve the problem: we distinguish three broad categories: `Recall` questions, which only require to recall necessary knowledge, `Reading Comprehension` questions, which require recall of knowledge and its combination with a given context—and `Quantitative/Arithmetic` questions, which require the calculation of quantities, such as probabilities or dosages. The majority of analyzed questions fall into the `Recall` category. Together with the `Reading Comprehension` category, these questions account for 95% of annotated questions. These two categories probe the capabilities required for reading comprehension (Kintsch, 1988), validating the use of M-QALM for the stated purpose of evaluating comprehension and recall.

Recall

Q: During CPR, chest compressions should be delivered at a rate of:

A. 80/minute. B. as fast as possible.
C. 100/minute. D. varies with each patient.

Answer: C. 100/minute

Reading Comprehension

Q: A 22-year-old man comes to the physician for a routine health maintenance examination. He feels well. He has had a painless left scrotal mass since childhood. Examination shows a 6-cm, soft, nontender left scrotal mass that transilluminates; there are no bowel sounds in the mass. Examination of the testis shows no abnormalities. Which of the following is the most likely cause of the mass?

A. Accumulation of scrotal adipose tissue
B. Cryptorchidism of the left testis
C. Dilation of the pampiniform plexus of veins around the testis
D. Persistence of a patent processus vaginalis

Answer: D. Persistence of a patent processus vaginalis

Quantitative/Arithmetic

Q: A person is prescribed Ropinirole 1.5 mg divided into three doses. How many micrograms is each dose? Choose one answer from the following:

A. 5 B. 50 C. 0.5 D. 500

Answer: D. 500

Figure 4: Sample questions corresponding to each category of the manual error analysis.

Table 5 describes the accuracy of the four base and fine-tuned models: we find that Recall questions dominate the sample and models tend to perform best in this category, but even after fine-tuning on M-QALM, their performance hardly surpasses 50%, indicating that they may yet lack the necessary knowledge. Additionally, models perform worse on Reading Comprehension questions, suggesting that it is indeed harder to integrate necessary knowledge rather than just recalling it. Fine-tuning improves performance for all models for both types of reasoning. Quantitative/Arithmetic are the worst-performing category, even after fine-tuning. This is unsurprising, as arithmetic capabilities are observed to emerge with larger model scale (Wei et al., 2022).

6.2 Error Analysis of LLama-2

We perform a manual error analysis of the fine-tuned LLaMA 2 (7B) model on MCQA. We examine 200 non-Within Context questions where the model erred, and assign them to the Recall, Reading Comprehension and Quantitative/Arithmetic categories, as done previously. The model incorrectly answered 134 Recall, 52 Reading Comprehension, and 14 Quantitative/Arithmetic questions (Table 6). Comparing these errors to the earlier sample of 200 questions we analyze from the test set in Table 5, reveals that the distribution of errors for each category mirrors their general distribution in the overall test set. The prevalence of Recall questions in errors aligns with their dominance in medical exams like MEDMCQA, USMLE, and HEADQA. While fine-tuning on extensive medical corpora may enhance Recall question performance, improving on Reading Comprehension and Quantitative/Arithmetic questions might require different fine-tuning approaches, as these categories demand comprehension skills rather than mere knowledge recall.

Category	General Test Set	LLaMA 2 Errors
Recall	65.5%	67%
Reading Comprehension	29.5%	26%
Quantitative/Arithmetic	5%	7%

Table 6: Distribution of Llama-2 errors across reasoning categories, compared with overall distribution of each category

7 Conclusions

In this work, we introduce M-QALM, a comprehensive collection of clinical datasets comprising 16 multiple-choice and 6 abstractive question-answering datasets. Our study encompasses an extensive empirical investigation of open-source language models with up to 13 billion parameters. We assess their clinical and biomedical knowledge, their capacity to acquire such knowledge through training on M-QALM, and their ability to generalize to previously unseen datasets.

Our results highlight the strengths and limitations of LLMs on MCQA and AQA: while performing significantly better than a random guess baseline, they still fall significantly short in performance compared to proprietary language models and humans. This is true even after fine-tuning on M-QALM, which demonstrates potential improvements, especially in the context of instruction fine-tuned models like Flan-T5. Finally, we show inconsistencies arising from the use of different AQA metrics—in future work we will supplement the automated metrics by fine-grained expert-driven manual evaluation of LLM's answers on M-QALM to learn to automate (some dimensions of) these expert judgments.

Based on our findings, we caution on the unconstrained use of open-source LLMs (Li et al., 2023; Han et al., 2023) as assistants to help perform medical tasks or provide answers to medical queries, to experts or lay people alike, as they seem to lack the necessary medical domain knowledge.

We make the dataset, experiment code and evaluation protocol publicly available[4] to allow future developers of medical LLMs to assess the foundations of their models' knowledge, as our evaluation shows that architecture of language models, the choice of datasets for pre-training and instruction fine-tuning can greatly impact their performance to the extent it can be assessed by M-QALM.

Limitations

In this paper, we evaluate the medical or clinical knowledge of LLMs by measuring their capability of answering test questions. While this can be a useful proxy measure of a model's domain knowledge, it is insufficient to gauge its potential application in a real-world scenario. A multi-dimensional analysis of a model's behaviour, including judging

[4] https://github.com/anand-subu/m-qalm

the completeness, harmlessness and usefulness of generated answers, is required in addition to solely evaluating their correctness.

Furthermore, the aggregated resource presented in this paper might be seen as lacking diversity, as all collected datasets are in English. To make inferences about the capabilities of evaluated models in other languages, a more diverse dataset with examples in other languages is required.

For our finetuning experiments, we only use parameter-efficient finetuning methods (PEFT) with QLora due to the high compute requirements for full-finetuning. We have not investigated the impact of the full-finetuning of these LLMs on our benchmark.

References

Asma Ben Abacha, Eugene Agichtein, Yuval Pinter, and Dina Demner-Fushman. 2017. Overview of the medical question answering task at TREC 2017 LiveQA. In *Text REtrieval Conference (TREC)*.

Nancy E Adams. 2015. Bloom's taxonomy of cognitive learning objectives. *Journal of the Medical Library Association: JMLA*, 103(3):152.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

Giuseppe Attardi, Luca Atzori, Maria Simi, et al. 2012. Index expansion for machine reading and question answering.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proc. ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *BMC Bioinformatics*, 20(1):511:1–511:23.

Asma Ben Abacha, Yassine Mrabet, Mark Sharp, Travis Goodwin, Sonya E. Shooshan, and Dina Demner-Fushman. 2019. Bridging the gap between consumers' medication questions and trusted answers. In *Proc. 17th World Congress on Medical and Health Informatics (MEDINFO)*.

Asma Ben Abacha, Wen-wai Yim, Griffin Adams, Neal Snider, and Meliha Yetisgen. 2023a. Overview of the MEDIQA-chat 2023 shared tasks on the summarization & generation of doctor-patient conversations. In *Proc. 5th Clinical Natural Language Processing Workshop*, pages 503–513, Toronto, Canada. Association for Computational Linguistics.

Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023b. An empirical study of clinical note generation from doctor-patient encounters. In *Proc. 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2291–2302, Dubrovnik, Croatia. Association for Computational Linguistics.

Jonathan Berant, Vivek Srikumar, Pei-Chun Chen, Abby Vander Linden, Brittany Harding, Brad Huang, Peter Clark, and Christopher D. Manning. 2014. Modeling biological processes for reading comprehension. In *Conference on Empirical Methods in Natural Language Processing*.

Benjamin S Bloom. 1956. *Taxonomy of education objectives Book 1-Cognitive domain*. David McKay Company.

Elliot Bolton, David Hall, Michihiro Yasunaga, Tony Lee, Chris Manning, and Percy Liang. 2022. Biomedlm.

Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Evaluating question answering evaluation. In *Proc. 2nd Workshop on Machine Reading for Question Answering*, pages 119–124, Hong Kong, China. Association for Computational Linguistics.

Together Computer. 2023. Redpajama-data: An open source recipe to reproduce llama training dataset.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. GPT3.int8(): 8-bit Matrix Multiplication for Transformers at Scale. In *Advances in Neural Information Processing Systems*, volume 35, pages 30318–30332. Curran Associates, Inc.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. *arXiv*, 2305.14314.

Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning.

Dheeru Dua, Ananth Gottumukkala, Alon Talmor, Matt Gardner, and Sameer Singh. 2019. Comprehensive Multi-Dataset Evaluation of Reading Comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 147–153, Stroudsburg, PA, USA. Association for Computational Linguistics.

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. MRQA 2019 Shared Task: Evaluating Generalization in Reading Comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 1–13, Stroudsburg, PA, USA. Association for Computational Linguistics.

Scott L. Fleming, Alejandro Lozano, William J. Haberkorn, Jenelle A. Jindal, Eduardo P. Reis, Rahul Thapa, Louis Blankemeier, Julian Z. Genkins, Ethan Steinberg, Ashwin Nayak, Birju S. Patel, Chia-Chun Chiang, Alison Callahan, Zepeng Huo, Sergios Gatidis, Scott J. Adams, Oluseyi Fayanju, Shreya J. Shah, Thomas Savage, Ethan Goh, Akshay S. Chaudhari, Nima Aghaeepour, Christopher Sharp, Michael A. Pfeffer, Percy Liang, Jonathan H. Chen, Keith E. Morse, Emma P. Brunskill, Jason A. Fries, and Nigam H. Shah. 2023. Medalign: A clinician-generated dataset for instruction following with electronic medical records. *arXiv*, 2308.14089.

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

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021. A framework for few-shot language model evaluation.

Albert Gatt and Emiel Krahmer. 2018. Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. 2023. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv*, 2304.08247.

Zexue He, Yu Wang, An Yan, Yao Liu, Eric Y Chang, Amilcare Gentili, Julian McAuley, and Chun-Nan Hsu. 2023. Medeval: A multi-level, multi-task, and multi-domain medical benchmark for language model evaluation. *arXiv preprint arXiv:2310.14088*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *arXiv*, 2009.03300.

Yichong Huang, Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2021. The factual inconsistency problem in abstractive text summarization: A survey. *arXiv preprint arXiv:2104.14839*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14).

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proc. Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Dan Roth. 2018. Question answering as global reasoning over semantic abstractions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Walter Kintsch. 1988. The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, 95(2):163–182.

Denis Kocetkov, Raymond Li, Loubna Ben Allal, Jia Li, Chenghao Mou, Carlos Muñoz Ferrandis, Yacine Jernite, Margaret Mitchell, Sean Hughes, Thomas Wolf, Dzmitry Bahdanau, Leandro von Werra, and Harm de Vries. 2022. The stack: 3 tb of permissively licensed source code. *Preprint*.

Anastasia Krithara, Anastasios Nentidis, Konstantinos Bougiatiotis, and Georgios Paliouras. 2023. Bioasq-qa: A manually curated corpus for biomedical question answering. *Scientific Data*, 10(1):170.

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

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6).

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Ye Liu, Shaika Chowdhury, Chenwei Zhang, Cornelia Caragea, and Philip S Yu. 2020. Interpretable multi-step reasoning with knowledge extraction on complex healthcare question answering. *arXiv preprint arXiv:2008.02434*.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Roser Morante, Martin Krallinger, Alfonso Valencia, and Walter Daelemans. 2012. Machine reading of biomedical texts about alzheimers disease. In *CLEF 2012 Conference and Labs of the Evaluation Forum-Question Answering For Machine Reading Evaluation (QA4MRE)*, pages 1–14.

NLP Team MosaicML. 2023. Introducing mpt-7b: A new standard for open-source, commercially usable llms. Accessed: 2023-05-05.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo 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.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proc. Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260.

Dimitris Pappas, Petros Stavropoulos, Ion Androutsopoulos, and Ryan McDonald. 2020. BioMRC: A dataset for biomedical machine reading comprehension. In *Proc. 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 140–149, Online. Association for Computational Linguistics.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv e-prints*.

Raffaele Raimondi, Nikolaos Tzoumas, Thomas Salisbury, Sandro Di Simplicio, and Mario R Romano. 2023. Comparative analysis of large language models in the royal college of ophthalmologists fellowship exams. *Eye*, pages 1–4.

RCOphth. 2022a. Frcophth sample mcqs part 1. *Part 1 FRCOphth sample mcqs - Royal College of Ophthalmologists*.

RCOphth. 2022b. Frcophth sample mcqs part 2. *Part 2 FRCOphth sample mcqs - Royal College of Ophthalmologists*.

Max Savery, Asma Ben Abacha, Soumya Gayen, and Dina Demner-Fushman. 2020. Question-driven summarization of answers to consumer health questions. *Scientific Data*, 7(1):322.

Viktor Schlegel, Goran Nenadic, and Riza Batista-Navarro. 2022. A survey of methods for revealing and overcoming weaknesses of data-driven Natural Language Understanding. *Natural Language Engineering*, pages 1–31.

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. 2023a. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023b. Towards expert-level medical question answering with large language models. *arXiv*, 2305.09617.

Alon Talmor and Jonathan Berant. 2019. MultiQA: An Empirical Investigation of Generalization and Transfer in Reading Comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4911–4921, Stroudsburg, PA, USA. Association for Computational Linguistics.

Augustin Toma, Patrick R. Lawler, Jimmy Ba, Rahul G. Krishnan, Barry B. Rubin, and Bo Wang. 2023. Clinical camel: An open expert-level medical language model with dialogue-based knowledge encoding. *arXiv*, 2305.12031.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *arXiv*, 2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artieres, Axel Ngonga, Norman Heino, Eric Gaussier, Liliana Barrio-Alvers, Michael Schroeder, Ion Androutsopoulos, and Georgios Paliouras. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16:138.

Logesh Kumar Umapathi, Ankit Pal, and Malaikannan Sankarasubbu. 2023. Med-halt: Medical domain hallucination test for large language models. *arXiv*, 2307.15343.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, William Collins, Neera Ahuja, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, John Pauly, and Akshay S. Chaudhari. 2023. Clinical text summarization: Adapting large language models can outperform human experts. *arXiv*, 2309.07430.

David Vilares and Carlos Gómez-Rodríguez. 2019. HEAD-QA: A healthcare dataset for complex reasoning. In *Proc. 57th Annual Meeting of the Association for Computational Linguistics*, pages 960–966, Florence, Italy. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent Abilities of Large Language Models.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-llama: Towards building open-source language models for medicine. *arXiv*, 2304.14454.

Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2023. Baize: An open-source chat model with parameter-efficient tuning on self-chat data.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Ming Zhu, Aman Ahuja, Da-Cheng Juan, Wei Wei, and Chandan K. Reddy. 2020. Question answering with long multiple-span answers. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 3840–3849. Association for Computational Linguistics.

A Datasets Used

In this section, we explain the MCQA and AQA datasets we used in detail. The dataset characteristics are presented in Table 1.

1. **USMLE - English**: We incorporate the USMLE dataset obtained from the MedQA dataset (Jin et al., 2021), comprising MCQA questions from the Medical Licensing Exam conducted in the US. We retain this dataset's original training, validation, and test set divisions.

2. **MEDMCQA**: We incorporate the MEDMCQA dataset from (Pal et al., 2022), which comprises medical MCQA from Indian Medical Entrance Exams. We retain this dataset's original training, validation, and test set splits. Similar to Singhal et al. (2023a), we evaluate all models on the validation set since we do not have answers for the test set.

3. **MMLU**: Following the design of Singhal et al. (2023a), we incorporate a subset of the MMLU datasets (6 datasets) (Hendrycks et al., 2021) which are MCQA specifically curated to assess medical domain knowledge. The subsets used are the **anatomy, clinical knowledge, college medicine, medical genetics, professional medicine** and **college biology** questions from MMLU. We utilize these datasets only for evaluating models.

4. **MEDIQA-ANS**: The MEDIQA 2019 shared task introduced the MEDIQA-QA dataset (Savery et al., 2020) for answer-ranking, comprising consumer health questions and passages from reputable online sources. The dataset was curated by extracting passages from the text of web pages, and includes manually generated single and multi-document summaries in both extractive and abstractive forms. We employ the multi-document abstractive summary as our questions' ground truth reference answer. We specifically filter for questions and answers marked as excellent and utilize this as an AQA dataset solely for evaluating models.

5. **HEADQA**: We include the HEADQA dataset (Vilares and Gómez-Rodríguez, 2019), which comprises graduate-level MCQA about various fields of medicine used for examinations to apply for specialization positions in the Spanish public healthcare system. We use the English version of the dataset and retain the original train, validation, and test split.

6. **PubmedQA**: The PubMedQA dataset (Jin et al., 2019) is a biomedical question-answering dataset comprising 1,000 expert-annotated QA instances. Each instance necessitates reasoning over a biomedical paper's abstract to answer a relevant question. While the dataset provides long and short answers (yes, no, or maybe), we focus exclusively on the short answers for our evaluation, thereby generalizing the task as MCQA. We retain the original test split of 500 questions. Additionally, we allocate 100 questions from the training set to serve as a validation set, facilitating standardized training and validation in future studies.

7. **BioMRC**: The BIOMRC dataset (Pappas et al., 2020) focuses on machine reading comprehension within the biomedical domain. It is structured in a cloze-style MCQA format, where questions are based on biomedical abstracts where biomedical entities are replaced with pseudo-identifiers. The task is to correctly identify the masked entity in the title from a list of masked entities. We utilize two compact versions of BioMRC: Tiny A (Setting A) and Tiny B (Setting B). The BIOMRC dataset comprises a large training corpus, where masked entities share the same pseudo-identifier across the entire corpus. Setting A, retains the same pseudo-identifiers used for masked biomedical entities in the training corpus. This setup is beneficial when testing models trained using the BioMRC training set, allowing them to draw on previously seen patterns. Setting B, conversely, changes the pseudo-identifiers for every single question. This means that a model must rely solely on the information in the text of the question and the passage it refers to, without any help from repeated exposure to the same placeholders. While we maintain the original format for Setting B, assessing Setting A as is, is difficult as since we do not utilize the BioMRC training set, it is functionally the same as Setting B. To address this

limitation, we modify Setting A to include the original entity names and their corresponding pseudo-identifiers in the answer options, based on how the original paper (Pappas et al., 2020) assesses the performance of experts and non-experts. This aims to assess whether the model can accurately answer when provided with the information about their original entity names.

8. **Processbank**: The Processbank dataset (Berant et al., 2014) is designed for machine reading comprehension, featuring questions based on paragraphs describing biological processes. Each question, associated with a particular paragraph, has two answer options (MCQA). The dataset comes with a predefined split of 435 questions (150 files) for training and 100 questions (50 files) for testing. We allocate 25 files from the training set to create a validation set while retaining the original test set for model evaluation.

9. **QA4MRE - Alzheimer's disease QA**: The dataset proposed by Morante et al. (Morante et al., 2012) contains MCQA questions regarding Alzheimer's disease, aimed at assessing machine reading systems' ability to answer questions about the disease by parsing relevant documents. We have adapted this dataset as an open-ended MCQA task to evaluate LLMs' ability to answer these questions based on inherent knowledge. This dataset is employed solely for model evaluation purposes.

10. **BioASQ**: The BioASQ dataset (Tsatsaronis et al., 2015; Krithara et al., 2023) features biomedical questions crafted by experts. We utilize the BioASQ 2022 dataset for our benchmark. The BioASQ dataset is divided into two parts: for MCQA and another for AQA. For the MCQA part, we filter out the yes/no questions from BioASQ, converting them into an MCQ format to create a new subset, which we term **BioASQ-MCQ**. We manually create a training-validation (train-val) split of roughly 85%-15% from the filtered questions, resulting in 975 training questions and 173 validation questions and retaining a test set of 123 questions. For the AQA part, BioASQ provides fact, list, and bullet-type questions. We compile these into an AQA dataset, ensuring a balanced representation of all question types in training and validation sets. The train-validation split results in 4733 training and 697 validation questions, with approximately 15% of all question types in the validation set.

11. **MASH-QA:** The MASH-QA dataset (Zhu et al., 2020) was designed for answering medical questions based on paragraphs where answers may span multiple text segments. Initially intended for extractive answering tasks, we repurpose it as an AQA task, utilizing the extractive answers as the reference ground truth.

12. **MedQUAD:** The MedQUAD dataset (Ben Abacha and Demner-Fushman, 2019) encompasses medical question-answer pairs extracted from various National Institute of Health (NIH) websites, covering topics on diseases, drugs, and other medical entities. Only nine of the twelve websites contributing to the original dataset have answers. We segregate questions from these nine websites and devise a train-validation-test split (AQA), assigning data from six websites for training, one website for validation, and two websites for testing.

13. **TREC-2017 LiveQA**: We employ the TREC-2017 LiveQA dataset (Abacha et al., 2017) for evaluation purposes. Specifically, we leverage the rankings provided within the MedQUAD evaluation process (Ben Abacha and Demner-Fushman, 2019) to keep question-answer pairs that have answer rating as excellent. We utilize this as an AQA dataset for evaluating the model.

14. **British Ophthalmology Practice Tests**: We employ sample questions from the Fellowship of the Royal College of Ophthalmologists (FRCOphth) exams, as provided by the Royal College of Ophthalmologists on their website (Raimondi et al., 2023; RCOphth, 2022a,b). These MCQA questions, geared towards testing ophthalmology-related knowledge, are used for evaluation.

15. **MEDINFO**: The MEDINFO dataset, introduced by Abacha et al. (Ben Abacha et al., 2019), consists of real consumer questions concerning medications and drugs. It en-

compasses 674 question-answer pairs (AQA), which we employ solely for evaluation.

B Performance of other methods for MCQA datasets

We report the prior and current best scores on MCQA datasets from current literature in Table 9. GPT-4 combined with a prompting strategy labeled MedPrompt performs the best currently on USMLE, MEDMCQA, and the MMLU datasets. Of the 16 datasets, we can obtain comparable scores for 12. For HEADQA, the results reported by (Vilares and Gómez-Rodríguez, 2019) and (Liu et al., 2020) are across individual sections, whereas we calculate the scores overall across all questions. The method proposed by (Liu et al., 2020), named **MurKe** achieves average scores of 45.5% on Biology questions, 42.4% on Medicine questions, 42.3% on Nursing Questions, 48.0% on Pharmacology questions, 44.3% on Psychology questions and 44.3% on Chemistry Questions, with an overall macro-average of 44.4% across all the sections. Similarly, for the OPHTH dataset, the results reported by (Raimondi et al., 2023) are separate for Part 1 and Part 2 questions. Bing Chat performs the best on Part 1 questions, achieving a performance of 78.9%, and GPT-4 with prompting obtains a performance of 88.4% on Part 2 questions (Raimondi et al., 2023). We could not find directly comparable scores for the **BioASQ** MCQ datasets as the test sets are provided in different batches, with the results on the BioASQ leaderboard also reported separately in terms of batches. We combine the questions across all the batches into one combined test set. For BIOMRC - Tiny A, we do not have directly comparable scores from prior works as we provide the names of the original entities along with the pseudo-identifiers to the LLMs, similar to how (Pappas et al., 2020) evaluate the performance of experts and non-experts. In contrast, when evaluating the performance of systems/deep learning models, (Pappas et al., 2020) first fine-tune models on the BIOMRC-Lite dataset and evaluate performance on BIOMRC - Tiny A, without providing names of the original entities to the system.

C Correlation between AQA and MCQA metrics

We use ROUGE-L, BERTScore and METEOR for evaluating the performance of LLMs for AQA. We try to understand which of the three metrics might be the most reliable for evaluation. Assuming that MCQA evaluations give a more robust estimate of models' capabilities due to the exact nature of evaluation, we calculate the correlation between the MCQA accuracy and each of the AQA metrics. Removing the Flan-T5-ZS models as outliers, we calculate the Spearman Rank Correlation and obtain the following results:

Metrics	Spearman R	P-value
MCQA Accuracy and AQA ROUGE-L	0.616	0.008
MCQA Accuracy and AQA BERTScore	0.353	0.164
MCQA Accuracy and AQA METEOR	-0.192	0.461

Table 7: Spearman Rank Correlation between MCQA accuracy and AQA metrics along with their statistical significance

The scores indicate that only ROUGE-L scores show a reliable and statistically significant correlation to MCQA Accuracy scores, suggesting that this might be the more reliable metric of the three. However, we wish to stress that these results must not be taken as definitive because the underlying assumption is that models performing better on MCQA should also perform better on AQA.

D Analysis of the causes of generalisation to unseen datasets

We aim to discriminate whether MCQA fine-tuned models' performance on unseen MCQA datasets can be attributed to their ability to generalize in answering medical questions, or if their performance is influenced by memorization of questions from the training set. To this end, we examine three evaluation-only MCQ datasets not used in the training split of M-QALM: Clinical Knowledge Tests (MMLU-CK) and Medical Genetics (MMLU-MG) from MMLU and the OPHTH dataset. We utilize semantic similarity algorithms to retrieve questions in the training sets that closely resemble those in these test sets and manually filter the retrieved results. We identify 6 out of 92, 12 out of 265, and 17 out of 100 questions in the OPHTH, MMLU-CK, and MMLU-MG datasets, respectively, that have similar counterparts in the MEDMCQA dataset which was used to fine-tune the LLaMA 2 model This suggests that scores might be inflated due to train-test leakage.

Next, we focus on questions that the LLaMA 2 (7B) model answered wrongly, but which were corrected by MCQA-fine-tuning. We then cross-reference these with the closest equivalent questions in the MEDMCQA dataset. This allows us to categorize the correct answers from near-duplicate memorization or the model's generalized learning capabilities. We find 5, 2, and 5 questions in the three investigated datasets, respectively, where the MCQA-fine-tuned model outperformed its zero-shot counterpart and identified closely related questions in MEDMCQA. Of these, 7 questions were near-duplicates with identical answers, while the remaining 5 would have required some level of clinical understanding for the model to answer them correctly. This suggests that the improved performance of instruction-tuned models on unseen datasets can be partially attributed to exposure to near-identical questions during training.

		MCQA	
		Macro-Avg	Micro-Avg
Base	LLaMA 1 (7B)	31.9	30.7
	LLaMA 1 (13B)	44.1	38.9
	LLaMA 2 (7B)	42.9	39.6
	LLaMA 2 (13B)	47.1	43.4
	MPT (7B)	27.6	27.3
	Falcon (7B)	34.7	31.6
Instruction tuned	LLaMA 2-chat (7B)	45.9	41.2
	LLaMA 2-chat (13B)	50.3	45.6
	MPT-Instruct (7B)	31.6	29.1
	Falcon-Instruct (7B)	31.8	29.7
	Flan-T5 (3B)	51.8	40.6
	Flan-T5 (11B)	56.5	45.2
Finetuned	LLaMA 2 (7B)	53.5	52.2
	MPT (7B)	53.2	51.5
	Falcon (7B)	49.3	48.6
	Flan-T5 (3B)	52.9	47.4
Adapted	ChatDoctor (7B)	42.8	36.0
	MedAlpaca (7B)	48.8	42.3
	PMC-LLama (13B)	53.7	57.9

Table 8: Micro-Average and Macro-Average Accuracies of all Models

Dataset	**Best Reported Score**	**Method**
USMLE (4 options)	90.2	GPT 4 + MedPrompt (Nori et al., 2023)
MEDMCQA	79.1	GPT 4 + MedPrompt (Nori et al., 2023)
PubMedQA	82.0	GPT 4 + MedPrompt (Nori et al., 2023)
MMLU - Anatomy	89.6	GPT 4 + MedPrompt (Nori et al., 2023)
MMLU - Clinical Knowledge	95.8	GPT 4 + MedPrompt (Nori et al., 2023)
MMLU - College Biology	97.9	GPT 4 + MedPrompt (Nori et al., 2023)
MMLU - College Medicine	89.0	GPT 4 + MedPrompt (Nori et al., 2023)
MMLU - Medical Genetics	98.0	GPT 4 + MedPrompt (Nori et al., 2023)
MMLU - Professional Medicine	95.2	GPT 4 + MedPrompt (Nori et al., 2023)
ProcessBank	68.8	SemanticILP (Biology Cascade) (Khashabi et al., 2018)
QA4MRE	55.0	Index Expansion (Attardi et al., 2012) (Morante et al., 2012)
BioMRC - Tiny B	60.0	SciBERT-Max-Reader (Pappas et al., 2020)

Table 9: Performance scores of various methods on various MCQA datasets

Model	Architecture	# Tokens	Data Source
Base models			
MPT	Decoder	1T	Red Pajama (Computer, 2023), The Stack (Kocetkov et al., 2022), C4 (Raffel et al., 2019), mC4 (Xue et al., 2021), S20RC (Lo et al., 2020)
LLaMA 1	Decoder	1.4T	Common Crawl, C4 (Raffel et al., 2019), Github, Wikipedia, Gutenberg, Books3 (Gao et al., 2021), Arxiv and Stack Exchange
Falcon	Decoder	1.5T	RefinedWeb (Penedo et al., 2023)
LLaMA 2	Decoder	2T	Unknown
Instruction tuned models			
Flan-T5	Encoder-Decoder	1T	C4 (Raffel et al., 2019) and *Flan-Collection* (Wei et al., 2021)
MPT-Instruct	Decoder	1T	MPT, *Databricks Dolly-15k (Conover et al., 2023), Anthropic Helpful and Harmless (Bai et al., 2022)*
Falcon-Instruct	Decoder	1.5T	Falcon, *baize (Xu et al., 2023), GPT4All, GPTeacher* [5]
LLaMA 2-Chat	Decoder	2T	LLaMA 2, *Flan Collection (Wei et al., 2021), Private Data*

Table 10: Pretrained LLMs considered in this paper. (Top rows) Open-source models that are decoder-only. (Bottom rows) Instruction-fine-tuned language models. **# Tokens**: Number of tokens used in pretraining the model. **Data Source**: Data used for pre-training (instruction data is *italicized*).

Parameter	Flan-T5 XL	Llama-2 7B	Falcon 7B	MPT 7B
lora_r	16	16	16	16
lora_alpha	16	16	16	16
lora_dropout	0.05	0.05	0.05	0.05
bias	none	none	none	none
optimizer	adamw	adamw	adamw	adamw
epochs	4	4	4	4
batch size	8	8	8	8
model_max_length	256	384	384	384

Table 11: Hyper-parameters used to train our models

Parameter	Decoder LLMs	Encoder-Decoder LLMs
Beam Size	3	3
Repetition Penalty	1.5	1.5
Max Output Length	200	200

Table 12: Inference time parameters used for abstractive question answering

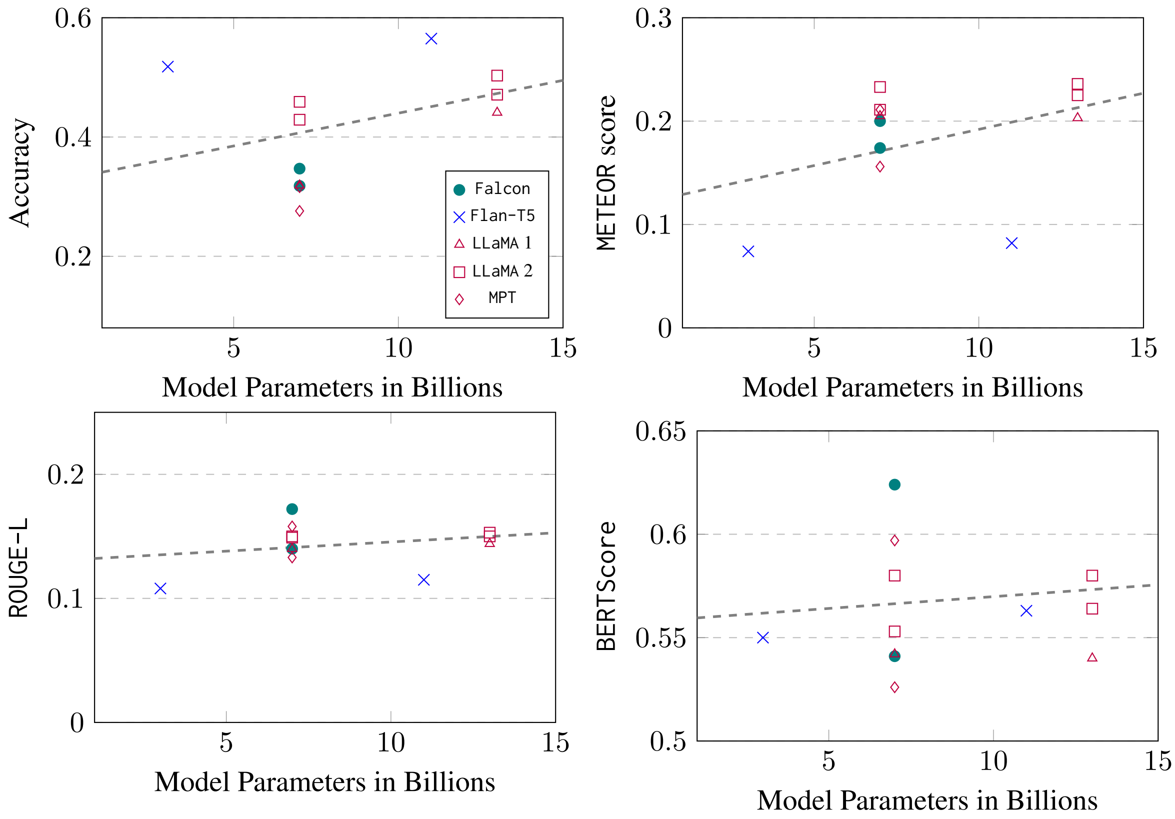

Figure 5: Zero-shot performance of models on MCQA (top-left) and AQA (top-right, bottom-left and bottom-right) as a function of model size. The dashed line represents a fitted linear regression showing the correlation between the model size and the score.

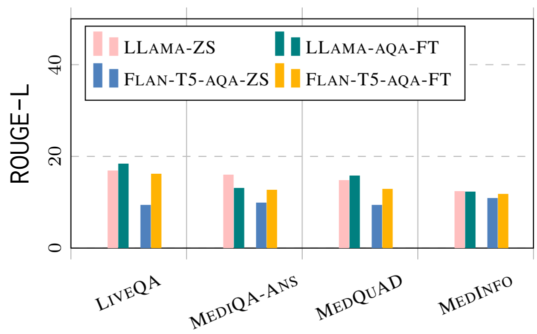

Figure 6: Performance of base and AQA-finetuned LLaMA 2 and Flan-T5 models on four unseen AQA test sets in terms of ROUGE-L.

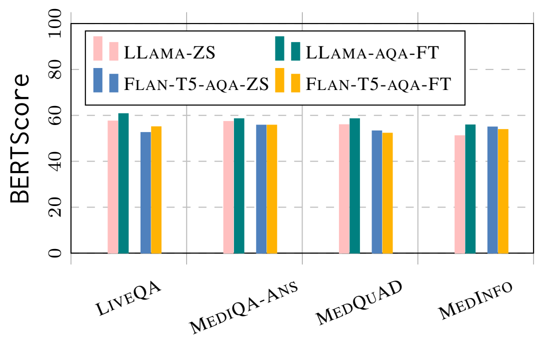

Figure 7: Performance of base and AQA-finetuned LLaMA 2 and Flan-T5 models on four unseen AQA test sets in terms of BERTScore.

Dataset	Random Baseline	Falcon (7B)	MPT (7B)	LLaMA 2 (7B)	LLaMA 2 (13B)	LLaMA 1 (7B)	LLaMA 1 (13B)
BioASQ-MCQ	50.0	72.4	33.3	67.5	35.8	35.0	37.4
BioMRC Tiny A	21.6	26.7	23.3	30.0	53.3	26.7	60.0
BioMRC Tiny B	18.1	16.7	13.3	26.7	20.0	13.3	33.3
MMLU - Anatomy	25.0	28.1	26.7	40.7	54.1	37.8	45.9
MMLU - Clinical Knowledge	25.0	32.5	29.8	38.1	57.7	35.5	43.4
MMLU - College Biology	25.0	27.1	22.2	39.6	58.3	35.4	44.4
MMLU - College Medicine	25.0	30.6	26.6	35.3	54.3	25.4	42.2
MMLU - Medical Genetics	25.0	33.0	27.0	49.0	52.0	34.0	42.0
MMLU - Professional Medicine	25.0	44.1	20.2	44.1	53.7	28.3	47.1
HeadQA	25.0	27.8	28.0	40.4	48.5	34.4	40.6
MedMCQA	25.0	30.4	26.5	36.0	37.5	27.0	35.9
Ophth	25.0	21.7	28.3	27.2	30.4	20.7	39.1
ProcessBank	50.0	50.7	56.0	75.3	83.3	63.3	74.0
PubMedQA	33.3	57.0	33.8	60.4	33.8	34.2	34.8
QA4MRE	20.0	30.0	22.5	40.0	37.5	30.0	47.5
USMLE	25.0	27.0	24.2	35.3	42.9	29.1	37.5
Average	27.7	34.7	27.6	42.9	47.1	31.9	44.1

Table 13: MCQA scores of LLMs in the zero-shot setting along with a random baseline. When calculating the random baselines for each dataset, for datasets with the same number of options for all questions, we set the score as the reciprocal of the number of options. For datasets with variable number of options per question, we calculate the score for each question as the reciprocal of the number of options for that question and then average all values. We utilize 5-shot prompting for the MMLU datasets and 1-shot prompting for other datasets to evaluate the models.

Dataset	Flan-T5 (3B)	Falcon (7B)	MPT (7B)	LLaMA 2 (7B) Chat	Flan-T5 (11B)	LLaMA 2 (13B) Chat
BioASQ-MCQ	43.9	45.5	34.1	69.9	48.8	65.0
BioMRC Tiny A	73.3	30.0	23.3	26.7	63.3	33.3
BioMRC Tiny B	46.7	23.3	23.3	20.0	60.0	26.7
MMLU - Anatomy	46.7	27.4	32.6	44.4	48.9	52.6
MMLU - Clinical Knowledge	52.1	31.7	36.6	54.3	61.9	57.7
MMLU - College Biology	48.6	25.0	29.9	55.6	54.9	59.0
MMLU - College Medicine	41.6	27.7	30.1	44.5	52.6	46.2
MMLU - Medical Genetics	50.0	32.0	32.0	60.0	55.0	56.0
MMLU - Professional Medicine	42.6	37.9	28.3	45.2	55.1	51.1
HeadQA	42.9	26.1	30.2	43.9	49.1	51.3
MedMCQA	33.1	29.8	27.2	35.0	36.4	39.3
Ophth	26.1	32.6	30.4	26.1	25.0	27.2
ProcessBank	93.3	52.0	56.7	72.0	95.3	80.0
PubMedQA	70.0	47.4	35.6	61.6	70.8	45.2
QA4MRE	82.5	15.0	30.0	40.0	87.5	72.5
USMLE	36.1	25.1	24.6	35.6	39.7	42.2
Average	51.8	31.8	31.6	45.9	56.5	50.3

Table 14: MCQA scores of Instruction-tuned LLMs in the zero-shot setting. We utilize 5-shot prompting for the MMLU datasets and 1-shot prompting for other datasets to evaluate these models.

Dataset	Flan-T5 (3B)	Falcon (7B)	MPT (7B)	LLaMA 2 (7B)
BIOASQ-MCQ	73.2	80.5	78.9	81.3
BIOMRC Tiny A	53.3	23.3	26.7	23.3
BIOMRC Tiny B	26.7	23.3	20.0	26.7
MMLU - Anatomy	43.7	43.7	45.9	54.1
MMLU - Clinical Knowledge	54.0	52.8	53.2	59.6
MMLU - College Biology	47.2	46.5	56.9	61.1
MMLU - College Medicine	44.5	53.2	50.3	52.0
MMLU - Medical Genetics	47.0	55.0	60.0	62.0
MMLU - Professional Medicine	48.5	50.0	49.3	59.6
HEADQA	49.0	47.7	52.4	53.9
MEDMCQA	43.0	45.9	48.4	48.3
OPHTH	34.8	30.4	35.9	31.5
PROCESSBANK	92.7	69.3	84.7	75.3
PUBMEDQA	74.2	70.8	73.4	70.6
QA4MRE	75.0	50.0	70.0	50.0
USMLE	39.7	46.3	45.7	46.1
Average	52.9	49.3	53.2	53.5

Table 15: MCQA scores of LLMs finetuned with QLora on MCQA datasets from the M-QALM benchmark. We evaluate these models without any examples in the prompt.

Dataset	Flan-T5 (3B)	Falcon (7B)	MPT (7B)	LLaMA 2 (7B)
BIOASQ-MCQ	0.8	13.8	14.6	7.3
BIOMRC Tiny A	50.0	23.3	10.0	16.7
BIOMRC Tiny B	36.7	23.3	16.7	16.7
MMLU - Anatomy	43.0	24.4	34.8	38.5
MMLU - Clinical Knowledge	50.9	25.3	28.7	40.8
MMLU - College Biology	42.4	23.6	34.7	38.9
MMLU - College Medicine	41.0	27.2	26.0	37.6
MMLU - Medical Genetics	45.0	31.0	22.0	49.0
MMLU - Professional Medicine	41.2	44.1	18.4	46.7
HEADQA	38.7	21.5	24.8	31.1
MEDMCQA	27.0	21.7	20.2	23.0
OPHTH	22.8	23.9	16.3	19.6
PROCESSBANK	88.0	54.7	42.0	50.7
PUBMEDQA	67.2	57.2	54.6	47.8
QA4MRE	77.5	35.0	10.0	15.0
USMLE	34.2	22.9	23.9	22.9
Average	44.1	29.6	24.9	31.4

Table 16: MCQA scores of LLMs finetuned with QLora on AQA datasets only from the M-QALM benchmark. We utilize 5-shot prompting for the MMLU datasets and 1-shot prompting for other datasets to evaluate these models.

Dataset	Flan-T5 (3B)	Falcon (7B)	MPT (7B)	LLaMA 2 (7B)
BIOASQ-MCQ	71.5	80.5	79.7	79.7
BIOMRC Tiny A	50.0	43.3	36.7	26.7
BIOMRC Tiny B	30.0	6.7	20.0	26.7
MMLU - Anatomy	40.7	45.2	47.4	52.6
MMLU - Clinical Knowledge	51.7	52.5	50.9	55.5
MMLU - College Biology	43.8	51.4	57.6	61.1
MMLU - College Medicine	41.6	48.0	54.3	52.6
MMLU - Medical Genetics	52.0	59.0	55.0	65.0
MMLU - Professional Medicine	47.1	46.0	50.4	59.9
HEADQA	47.5	47.4	51.2	54.2
MEDMCQA	41.7	45.2	47.4	48.0
OPHTH	32.6	28.3	38.0	28.3
PROCESSBANK	91.3	73.3	79.3	83.3
PUBMEDQA	71.4	67.8	72.8	71.8
QA4MRE	72.5	52.5	60.0	67.5
USMLE	40.9	45.7	44.3	45.6
Average	51.7	49.5	52.8	54.9

Table 17: MCQA scores of LLMs finetuned with QLora on both MCQA and AQA data from the M-QALM benchmark. We evaluate these models without any examples in the prompt.

Dataset	ChatDoctor (7B)	MedAlpaca (7B)	PMC-LLama (13B)
BIOASQ-MCQ	65.0	50.4	13.0
BIOMRC Tiny A	20.0	16.7	30.0
BIOMRC Tiny B	36.7	23.3	16.7
MMLU - Anatomy	43.7	60.0	63.0
MMLU - Clinical Knowledge	43.4	60.0	62.3
MMLU - College Biology	39.6	64.6	64.6
MMLU - College Medicine	32.4	52.6	53.2
MMLU - Medical Genetics	55.0	69.0	70.0
MMLU - Professional Medicine	47.1	67.3	67.6
HEADQA	37.2	45.1	59.1
MEDMCQA	29.4	35.0	56.5
OPHTH	30.4	23.9	46.7
PROCESSBANK	62.0	67.3	74.7
PUBMEDQA	67.4	40.8	72.6
QA4MRE	45.0	62.5	55.0
USMLE	31.3	42.4	54.7
Average	42.8	48.8	53.7

Table 18: MCQA scores of ChatDoctor (7B) , MedAlpaca (7B) and PMC-LLama (13B). To evaluate ChatDoctor, we utilize 5-shot prompting for the MMLU datasets and 1-shot prompting for other datasets to evaluate these models. We evaluate MedAlpaca (7B) and PMC-LLama (13B) directly without any examples in the prompt.

Category	Support	Flan-T5 (ZS)	Flan-T5 (FT)	MPT (ZS)	MPT (FT)	Falcon (ZS)	Falcon (FT)	LLaMA 2 (ZS)	LLaMA 2 (FT)
Consumer Health Dataset Questions	1449	10.5	13.4	12.6	14.6	13.2	14.6	13.7	14.5
General Biomedical Dataset Questions	363	15.0	26.6	11,4	28.9	13.9	27.8	15.8	30.0
General Medical Dataset Questions	200	9.3	12.8	13.7	14.0	14.3	14.8	14.7	15.7

Table 19: Performance of LLMs in the zero-shot and fine-tuned setting across various categories across various dataset categories in terms of Rouge Score

Model	BioASQ-QA			LiveQA			MashQA			MedInfo			MediQA-Ans			MedQuAD			Average		
	RL	BS	MTR	RL	BS	MTR	RL	BS	MTR	RL	BS	MTR	RL	BS	MTR	RL	BS	MTR	RL	BS	MTR
Falcon (7B)	13.9	53.1	22.5	15.4	55.8	17.4	13.4	53.7	22.0	12.1	51.1	17.8	15.3	56.1	21.7	14.3	54.7	18.4	14.0	54.1	20.0
MPT (7B)	11.4	50.1	21.7	15.7	55.2	20.9	12.8	52.3	23.0	11.2	49.6	18.4	14.8	55.6	23.3	13.7	53.2	19.4	13.3	52.6	21.1
LLaMA 1 (7B)	13.8	53.4	23.3	15.4	55.8	18.9	13.5	54.1	22.2	11.6	51.4	17.9	15.5	56.8	22.5	14.3	54.0	18.5	14.0	54.2	20.5
LLaMA 1 (13B)	14.6	53.3	22.8	16.7	55.7	19.7	13.1	53.3	20.9	12.5	51.7	18.6	15.4	57.0	22.1	14.0	53.2	17.8	14.4	54.0	20.3
LLaMA 2 (7B)	15.8	54.6	24.0	16.8	57.5	20.1	14.0	55.4	23.3	12.3	51.1	17.8	15.9	57.3	22.3	14.7	55.9	19.4	14.9	55.3	21.1
LLaMA 2 (13B)	14.9	55.3	24.9	16.2	57.3	20.1	14.5	56.4	24.4	12.7	53.6	20.0	16.4	58.9	24.4	15.4	57.1	20.9	15.0	56.4	22.5
Flan-T5 (3B)	15.0	57.7	11.1	9.3	52.5	6.1	10.5	56.0	7.5	10.8	54.9	7.6	9.8	55.7	6.2	9.3	53.2	6.0	10.8	55.0	7.4
MPT (7B) Instruct	23.2	64.5	22.4	14.5	58.1	13.4	15.0	61.1	15.9	14.0	56.8	12.9	14.8	60.5	16.1	12.9	57.1	13.1	15.8	59.7	15.6
Falcon (7B) Instruct	27.2	68.9	28.1	16.1	61.4	14.7	15.5	62.5	17.1	14.7	58.4	15.2	15.4	62.4	15.4	14.3	60.8	14.2	17.2	62.4	17.4
LLaMA 2 (7B) Chat	15.9	58.8	26.5	15.4	58.8	20.9	14.2	57.4	24.4	12.8	54.6	20.6	16.7	59.5	25.4	15.4	58.7	22.1	15.0	58.0	23.3
Flan-T5 (11B)	16.3	58.8	12.2	10.8	55.5	7.5	10.8	57.3	8.2	12.3	56.1	9.1	9.7	55.2	6.3	9.0	54.9	5.9	11.5	56.3	8.2
LLaMA 2 (13B) Chat	16.2	59.2	27.5	15.8	59.0	21.4	14.2	57.2	24.3	13.0	54.7	21.2	16.7	58.9	24.8	15.5	58.7	22.4	15.3	58.0	23.6
Flan-T5 (3B) (FT-QA)	26.6	66.2	25.2	16.1	55.0	16.9	15.4	58.2	16.4	11.7	53.8	10.5	12.6	55.7	12.0	12.8	52.2	12.7	15.9	56.8	15.6
Falcon (7B) (FT-QA)	27.8	68.4	26.6	20.1	60.6	21.1	16.7	61.3	17.8	12.4	56.5	9.4	12.8	57.9	11.6	14.8	57.5	16.2	17.4	60.4	17.1
LLaMA 2 (7B) (FT-QA)	30.0	69.7	28.2	18.3	60.7	19.2	16.9	61.9	17.5	12.2	55.8	9.0	13.0	58.5	11.2	15.7	58.5	16.6	17.7	60.8	16.9
MPT (7B) (FT-QA)	28.9	69.0	27.6	18.6	59.6	20.6	16.4	61.0	17.5	12.9	56.1	10.7	13.1	57.6	11.5	14.0	56.5	15.4	17.3	60.0	17.2
Flan-T5 (3B) (FT-All)	27.8	67.4	25.7	16.0	55.8	17.1	15.5	59.3	15.3	11.4	54.5	9.3	11.7	55.7	10.4	13.0	53.1	13.1	15.9	57.6	15.2
Falcon (7B) (FT-All)	27.3	68.6	26.1	18.9	59.9	19.8	16.1	61.0	16.7	11.7	55.4	8.0	12.8	58.0	10.9	14.8	57.5	16.5	16.9	60.1	16.3
LLaMA 2 (7B) (FT-All)	30.2	69.7	27.8	17.9	60.4	17.9	17.3	61.9	17.7	12.4	54.9	9.9	13.3	58.3	12.2	15.0	57.7	15.5	17.7	60.5	16.8
MPT (7B) (FT-All)	29.1	68.8	27.4	18.2	59.2	20.4	16.5	61.5	17.0	13.4	56.4	11.5	13.5	57.5	12.3	14.5	56.7	16.6	17.5	60.0	17.5
ChatDoctor	26.2	68.2	28.8	15.8	61.3	16.0	16.1	62.6	18.6	15.2	58.9	15.6	16.5	62.9	18.2	14.8	60.2	15.0	17.4	62.3	18.7
MedAlpaca 7B	26.4	67.8	27.1	14.7	55.6	13.0	13.4	59.3	15.0	12.3	55.1	12.6	13.9	59.0	15.4	12.5	56.8	10.2	15.5	58.9	15.6
PMC LLama 13B	19.7	62.6	20.9	12.7	55.8	11.0	13.5	58.8	14.4	45.6	70.7	43.6	14.8	59.6	14.0	11.9	57.0	10.1	19.7	60.7	19.0

Table 20: AQA scores of base, instruction-tuned LLMs in the zero-shot setting, LLMs fine-tuned with QLora and other biomedical and clinical instruction tuned models such as ChatDoctor (7B), MedAlpaca (7B), PMC-LLama (13B). FT-QA refers to models fine-tuned only with AQA data and FT-All refers to models fine-tuned with both MCQA and AQA data.

E Prompts utilized

In this section, we outline all prompts used for fine-tuning and evaluating the LLMs. We define **Single Context MCQA Prompt** as the prompt for the PROCESSBANK dataset with a single paragraph context, **Multi-Context MCQA Prompt** as the prompt for the PUBMEDQA dataset with multiple paragraph contexts, **Cloze MCQA Prompt** for the BIOMRC Setting A and B datasets, **MCQA Prompt** for all other MCQA datasets, and **AQA Prompt** for all AQA datasets.

AQA Prompt for fine-tuning and evaluating `Falcon` (Base), `MPT` (Base), `LLaMA 2` (Base) and `Flan-T5`

Answer the medical question precisely and factually
Question: {Question}
Answer:

Figure 8: AQA prompt utilized for finetuning and evaluating these models.

MCQA Prompt for fine-tuning and evaluating `Falcon` (Base), `MPT` (Base), `LLaMA 2` (Base) and `Flan-T5`

Pick the right option that answers the question
Question: {Question}
Options:
A. {Option Text}
B. {Option Text}
C. {Option Text}
D. {Option Text}
Answer:

Figure 9: MCQA prompt utilized for finetuning and evaluating these models.

Single Context MCQA Prompt for fine-tuning and evaluating `Falcon` (Base), `MPT` (Base), `LLaMA 2` (Base) and `Flan-T5`

Given the context, pick the right choice that answers the question
Context: {Context Paragraph}
Question: {Question}
Options:
A. {Option Text}
B. {Option Text}
Answer:

Figure 10: Single Context MCQA prompt utilized for finetuning and evaluating these models on the PROCESSBANK dataset.

Multi Context MCQA Prompt for fine-tuning and evaluating `Falcon` (Base), `MPT` (Base), `LLaMA 2` (Base) and `Flan-T5`

Given the context, pick the right choice that answers the question
Contexts: {Context Paragraph 1}
{Context Paragraph 2}
...
{Context Paragraph N}
Question: {Question}
Options:
A. {Option Text}
B. {Option Text}
C. {Option Text}
Answer:

Figure 11: Multi Context MCQA prompt utilized for fine-tuning and evaluating these models on thePUBMEDQA dataset.

Cloze MCQA Prompt for evaluating fine-tuned `Falcon` (Base), `MPT` (Base), `LLaMA 2` (Base) and `Flan-T5`

Given the context, pick the right choice that corresponds to the XXXX in the question
Context: {Context Paragraph}
Question: {Question}
Options:
A. {Option Text}
B. {Option Text}
C. {Option Text}
Answer:

Figure 12: Cloze MCQA prompt utilized for evaluating the fine-tuned models on the BIOMRC datasets.

MCQA Prompt for evaluating `Falcon` (Base and Instruct), `MPT` (Base), `LLaMA 1` (Base), `LLaMA 2` (Base) and `Flan-T5` without any fine-tuning

Pick the right option that answers the question
Question: {Example 1}
Options:
A. {Option Text}
B. {Option Text}
C. {Option Text}
D. {Option Text}
Answer:{Correct Option}
...
Question: {Example K}
Options:
A. {Option Text}
B. {Option Text}
C. {Option Text}
D. {Option Text}
Answer:{Correct Option}
Question: {Question}
Options:
A. {Option Text}
B. {Option Text}
C. {Option Text}
D. {Option Text}
Answer:

Figure 13: MCQA prompt utilized for evaluating models prior to any fine-tuning. 5-shot prompting is utilized for the MMLU datasets whereas 1-shot prompting is utilized for all other MCQA datasets when evaluating non-finetuned models.

AQA Prompt for evaluating `Falcon` (Base and Instruct), `MPT` (Base), `LLaMA 1` (Base), `LLaMA 2` (Base) and `Flan-T5` without any fine-tuning

Answer the medical question precisely and factually
Question: {Question}
Answer:

Figure 14: AQA prompt utilized for evaluating the models.

Single Context MCQA Prompt for evaluating `Falcon` (Base and Instruct), `MPT` (Base), `LLaMA 1` (Base), `LLaMA 2` (Base) and `Flan-T5` without any fine-tuning

Given the context, pick the right choice that answers the question
Context: {Context Paragraph}
Question: {Example Question}
Options:
A. {Option Text}
B. {Option Text}
Answer:{Correct Option}
Context: {Context Paragraph}
Question: {Question}
Options:
A. {Option Text}
B. {Option Text}
Answer:

Figure 15: Single Context MCQA prompt utilized for evaluating the PROCESSBANK dataset.

Cloze MCQA Prompt for evaluating `Falcon` (Base and Instruct), `MPT` (Base), `LLaMA 1` (Base), `LLaMA 2` (Base) and `Flan-T5` without any fine-tuning

Given the context, pick the right choice that corresponds to the XXXX in the question
Context: {Context Paragraph}
Question: {Example Question}
Options:
A. {Option Text}
B. {Option Text}
C. {Option Text}
Answer:{Correct Option}
Context: {Context Paragraph}
Question: {Question}
Options:
A. {Option Text}
B. {Option Text}
C. {Option Text}
Answer:

Figure 16: Cloze MCQA prompt utilized for evaluating the BIOMRC datasets in settings A and B. .

Multi Context MCQA Prompt for evaluating `Falcon` (Base and Instruct), `MPT` (Base), `LLaMA 1` (Base), `LLaMA 2` (Base) and `Flan-T5` without any fine-tuning

Given the context, pick the right choice that answers the question
Contexts: {Context Paragraph 1}
{Context Paragraph 2}
...
{Context Paragraph N}
Question: {Example Question}
Options:
A. {Option Text}
B. {Option Text}
C. {Option Text}
Answer:{Correct Option}
Contexts: {Context Paragraph 1}
{Context Paragraph 2}
...
{Context Paragraph N}
Question: {Question}
Options:
A. {Option Text}
B. {Option Text}
C. {Option Text}
Answer:

Figure 17: Multi-Context MCQA prompt utilized for evaluating the PUBMEDQA dataset.

MCQA Prompt for evaluating LLaMA 2 (Chat) Models without any fine-tuning

[INST] «SYS»
Pick the right option that answers the question
«/SYS»

Question: {Example Question}
Options:
A. {Option Text}
B. {Option Text}
C. {Option Text}
D. {Option Text} [/INST] Answer:{Correct Option} </s><s>[INST] Question: {Question}
Options:
A. {Option Text}
B. {Option Text}
C. {Option Text}
D. {Option Text} [/INST] Answer:

Figure 18: MCQA prompt utilized foe evaluating the model. 5-shot prompting is utilized for the MMLU datasets whereas 1-shot prompting is utilized for all other MCQA datasets.

Single Context MCQA Prompt for evaluating LLaMA 2 (Chat) Models without any fine-tuning

[INST] «SYS»
Given the context, pick the right choice that answers the question
«/SYS»

Context: {Context Paragraph}
Question: {Example Question}
Options:
A. {Option Text}
B. {Option Text} [/INST] Answer:{Correct Option} </s><s>[INST] Context: {Context Paragraph}
Question: {Question}
Options:
A. {Option Text}
B. {Option Text} [/INST] Answer:

Figure 19: Single Context MCQA prompt utilized for evaluating models on the PROCESSBANK dataset.

AQA Prompt for evaluating LLaMA 2 (Chat) Models without any fine-tuning

[INST] «SYS»
Answer the medical question precisely and factually
«/SYS»

Question: {Question} [/INST]

Figure 20: AQA prompt utilized for evaluating the model

Cloze MCQA Prompt for evaluating LLaMA 2 (Chat) Models without any fine-tuning

[INST] «SYS»
Given the context, pick the right choice that corresponds to the XXXX in the question
«/SYS»

Context: {Context Paragraph}
Question: {Example Question}
Options:
A. {Option Text}
B. {Option Text} [/INST] Answer:{Correct Option} </s><s>[INST] Context: {Context Paragraph}
Question: {Question}
Options:
A. {Option Text}
B. {Option Text} [/INST] Answer:

Figure 21: Cloze MCQA prompt utilized for evaluating models on the BIOMRC datasets in settings A and B.

Multi Context MCQA Prompt for evaluating LLaMA 2 (Chat) Models without any fine-tuning

[INST] «SYS»
Given the context, pick the right choice that answers the question
«/SYS»

Contexts: {Context Paragraph 1}
{Context Paragraph 2}
...
{Context Paragraph N}
Question: {Example Question}
Options:
A. {Option Text}
B. {Option Text}
C. {Option Text} [/INST] Answer:{Correct Option} </s><s>[INST] Contexts: {Context Paragraph 1}
{Context Paragraph 2}
...
{Context Paragraph N}
Question: {Question}
Options:
A. {Option Text}
B. {Option Text}
C. {Option Text} [/INST] Answer:

Figure 22: Multi-Context MCQA prompt utilized for evaluating models on the PUBMEDQA dataset.

AQA Prompt for evaluating MPT Instruct without fine-tuning

Below is an instruction that describes a task. Write a response that appropriately completes the request.
Instruction:
Answer the medical question precisely and factually. Question: {Question}
Response:
Answer:

Figure 23: AQA prompt utilized for evaluating the model.

MCQA Prompt for evaluating MPT Instruct without fine-tuning

Below is an instruction that describes a task. Write a response that appropriately completes the request.
Instruction:
Pick the right option that answers the question. Question: {Example Question 1}
Options:
A. {Option Text}
B. {Option Text}
C. {Option Text}
D. {Option Text}
Response:
Answer:{Correct Option}
...
Instruction:
Pick the right option that answers the question. Question: {Example Question K}
Options:
A. {Option Text}
B. {Option Text}
C. {Option Text}
D. {Option Text}
Response:
Answer:{Correct Option}
Instruction:
Pick the right option that answers the question. Question: {Question}
Options:
A. {Option Text}
B. {Option Text}
C. {Option Text}
D. {Option Text}
Response:
Answer:

Figure 24: MCQA prompt utilized for evaluating the model. 5-shot prompting is utilized for the MMLU datasets whereas 1-shot prompting is utilized for all other MCQA datasets.

Single Context MCQA Prompt for evaluating MPT Instruct without fine-tuning

Below is an instruction that describes a task. Write a response that appropriately completes the request.
Instruction:
Given the context, pick the right choice that answers the question. Context: {Context Paragraph}
Question: {Example Question}
Options:
A. {Option Text}
B. {Option Text}
Response:
Answer:{Correct Option}
Instruction:
Given the context, pick the right choice that answers the question. Context: {Context Paragraph}
Question: {Question}
Options:
A. {Option Text}
B. {Option Text}
Response:
Answer:

Figure 25: Single Context MCQA prompt utilized for evaluating the model on the PROCESSBANK dataset.

Multi Context MCQA Prompt for evaluating MPT Instruct without fine-tuning

Below is an instruction that describes a task. Write a response that appropriately completes the request.
Instruction:
Given the contexts, pick the right choice that answers the question. Contexts: {Context Paragraph 1}
{Context Paragraph 2}
...
{Context Paragraph N}
Question: {Example Question 1}
Options:
A. {Option Text}
B. {Option Text}
C. {Option Text}
Response:
Answer:{Correct Option}
Instruction:
Given the contexts, pick the right choice that answers the question. Contexts: {Context Paragraph 1}
{Context Paragraph 2}
...
{Context Paragraph N}
Question: {Question}
Options:
A. {Option Text}
B. {Option Text}
C. {Option Text}
Response:
Answer:

Figure 26: Multi-Context MCQA prompt utilized for evaluating the model on the PUBMEDQA dataset.

Cloze MCQA Prompt for evaluating MPT Instruct without fine-tuning

Below is an instruction that describes a task. Write a response that appropriately completes the request.
Instruction:
Given the context, pick the right choice that corresponds to the XXXX in the question. Context: {Context Paragraph}
Question: {Example Question}
Options:
A. {Option Text}
B. {Option Text}
Response:
Answer:{Correct Option}
Instruction:
Given the context, pick the right choice that corresponds to the XXXX in the question. Context: {Context Paragraph}
Question: {Question}
Options:
A. {Option Text}
B. {Option Text}
Response:
Answer:

Figure 27: Cloze MCQA prompt utilized for evaluating the model on the BIOMRC datasets in settings A and B.

MCQA Prompt for `ChatDoctor`

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:
If you are a doctor, please answer the medical questions based on the patient's description. Answer with the best option directly.

Input:
Question: {Example Question 1}
Options:
A. {Option Text}
B. {Option Text}
C. {Option Text}
D. {Option Text}

Response:
Answer:{Correct Option}
...
Instruction:
If you are a doctor, please answer the medical questions based on the patient's description. Answer with the best option directly.

Input:
Question: {Example Question K}
Options:
A. {Option Text}
B. {Option Text}
C. {Option Text}
D. {Option Text}

Response:
Answer:{Correct Option}

Instruction:
If you are a doctor, please answer the medical questions based on the patient's description. Answer with the best option directly.

Input:
Question: {Question}
Options:
A. {Option Text}
B. {Option Text}
C. {Option Text}
D. {Option Text}

Response:
Answer:

Figure 28: MCQA prompt utilized for evaluating `ChatDoctor`. 5-shot prompting is utilized for the MMLU datasets whereas 1-shot prompting is utilized for all other MCQA datasets.

Single Context MCQA Prompt for `ChatDoctor`

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:
If you are a doctor, please answer the medical questions based on the patient's description. Analyze the question given its context. Answer with the best option directly.

Input:
Context: {Context Paragraph}
Question: {Example Question}
Options:
A. {Option Text}
B. {Option Text}

Response:
Answer:{Correct Option}

Instruction:
If you are a doctor, please answer the medical questions based on the patient's description. Analyze the question given its context. Answer with the best option directly.

Input:
Context: {Context Paragraph}
Question: {Question}
Options:
A. {Option Text}
B. {Option Text}

Response:
Answer:

Figure 29: Single Context MCQA prompt utilized for evaluating `ChatDoctor` on the PROCESSBANK dataset.

Multi Context MCQA Prompt for `ChatDoctor`

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:
If you are a doctor, please answer the medical questions based on the patient's description. Analyze the question given its context. Answer with the best option directly.

Input:
Contexts: {Context Paragraph 1}
{Context Paragraph 2}
...
{Context Paragraph N}
Question: {Example Question}
Options:
A. {Option Text}
B. {Option Text}
C. {Option Text}

Response:
Answer:{Correct Option}

Instruction:
If you are a doctor, please answer the medical questions based on the patient's description. Analyze the question given its context. Answer with the best option directly.

Input:
Contexts: {Context Paragraph 1}
{Context Paragraph 2}
...
{Context Paragraph N}
Question: {Question}
Options:
A. {Option Text}
B. {Option Text}
C. {Option Text}

Response:
Answer:

Figure 30: Multi-Context MCQA prompt utilized for evaluating `ChatDoctor` on the PUBMEDQA dataset.

Cloze MCQA Prompt for `ChatDoctor`

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:
If you are a doctor, please answer the medical questions based on the patient's description. Analyze the question given its context. Pick the right option that corresponds to the XXXX in the question

Input:
Context: {Context Paragraph}
Question: {Question}
Options:
A. {Option Text}
B. {Option Text}

Response:
Answer:{Correct Option}

Instruction:
If you are a doctor, please answer the medical questions based on the patient's description. Analyze the question given its context. Pick the right option that corresponds to the XXXX in the question

Input:
Context: {Context Paragraph}
Question: {Question}
Options:
A. {Option Text}
B. {Option Text}

Response:
Answer:

Figure 31: Cloze MCQA prompt utilized for evaluating `ChatDoctor` on the BIOMRC datasets in settings A and B.

AQA Prompt for `ChatDoctor`

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:
If you are a doctor, please answer the medical questions based on the patient's description.

Input:
{Question}

Response:

Figure 32: AQA prompt utilized for evaluating `ChatDoctor`.

MCQA Prompt for MedAlpaca

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:
Answer this multiple-choice question.

Input:
{Question}
A: {Option Text}
B: {Option Text}
C: {Option Text}
D: {Option Text}

Response:
The Answer to the question is:

Figure 33: MCQA prompt utilized for evaluating MedAlpaca.

Single Context MCQA Prompt for MedAlpaca

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:
Analyze the question given its context. Answer this multiple-choice question.

Input:
Context: {Context Paragraph}

{Question}
A: {Option Text}
B: {Option Text}

Response:
The Answer to the question is:

Figure 34: Single Context MCQA prompt utilized for evaluating MedAlpaca on the PROCESSBANK dataset

Multi Context MCQA Prompt for `MedAlpaca`

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:
Analyze the question given its context. Answer this multiple-choice question.

Input:
Contexts: {Context Paragraph 1}
{Context Paragraph 2}
...
{Context Paragraph N}

{Question}
A: {Option Text}
B: {Option Text}
C: {Option Text}

Response:
The Answer to the question is:

Figure 35: Multi-Context MCQA prompt utilized for evaluating `MedAlpaca` on the PUBMEDQA dataset.

Cloze MCQA Prompt for `MedAlpaca`

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:
Analyze the question given its context. Pick the right option that corresponds to the XXXX in the question.

Input:
Context: {Context Paragraph}

{Question}
A: {Option Text}
B: {Option Text}
C: {Option Text}
D: {Option Text}

Response:
The Answer to the question is:

Figure 36: Cloze MCQA prompt utilized for evaluating `MedAlpaca` on the BIOMRC datasets in settings A and B.

AQA Prompt for `MedAlpaca`

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:
Answer this question truthfully

Input:
{Question}

Response:

Figure 37: AQA prompt utilized for evaluating `MedAlpaca`.

MCQA Prompt for `PMC-LLama`

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:
You're a doctor, kindly address the medical queries according to the patient's account. Answer with the best option directly.

Input:
###Question: {Question}
###Options:
A. {Option Text}
B. {Option Text}
C. {Option Text}
D. {Option Text}

Response:
###Answer:

Figure 38: MCQA prompt utilized for evaluating `PMC-LLama`.

Single Context MCQA Prompt for PMC-LLama

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:
You're a doctor, kindly address the medical queries according to the patient's account. Analyze the question given its context. Answer with the best option directly.

Input:
###Question: {Question}
###Context: {Context Paragraph}
###Options:
A. {Option Text}
B. {Option Text}

Response:
###Answer:

Figure 39: Single Context MCQA prompt utilized for evaluating PMC-LLama on the PROCESSBANK dataset.

Multi-Context MCQA Prompt for PMC-LLama

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:
You're a doctor, kindly address the medical queries according to the patient's account. Analyze the question given its context. Answer with the best option directly.

Input:
###Question: {Question}
###Contexts: {Context Paragraph 1}
{Context Paragraph 2}
...
{Context Paragraph N}
###Options:
A. {Option Text}
B. {Option Text}
C. {Option Text}

Response:
###Answer:

Figure 40: Multi-Context MCQA prompt utilized for evaluating PMC-LLama on the PUBMEDQA dataset.

Cloze MCQA Prompt for `PMC-LLama`

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:
You're a doctor, kindly address the medical queries according to the patient's account. Analyze the question given its context. Pick the right option that corresponds to the XXXX in the question

Input:
###Question: {Question}
###Context: {Context Paragraph}
###Options:
A. {Option Text}
B. {Option Text}

Response:
###Answer:

Figure 41: Cloze MCQA prompt utilized for evaluating `PMC-LLama` on the BIOMRC datasets in settings A and B.

AQA Prompt for `PMC-LLama`

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:
You're a doctor, kindly address the medical queries according to the patient's account.

Input:
###Question: {Question}

Response:
###Answer:

Figure 42: AQA prompt utilized for evaluating `PMC-LLama`.