QALM – Clinical QA for Language Models

Anonymous ACL submission

Abstract

In recent years, Large Language Models (LLMs) have gained recognition for their ability to encode clinical knowledge within their parameters. Despite their growing popularity, the existing literature lacks a comprehensive and standardized benchmark for evaluating the performance of these models in clinical knowledge applications. In response to this gap, we introduce a novel benchmark called QALM designed to harmonize the evaluation of language models in the context of clinical knowledge. Our benchmark comprises 16 Multiple-Choice Question (MCQ) datasets and six Abstractive Question Answering (AQA) datasets, offering a diverse range of challenges to comprehensively assess model capabilities.

Our experimental results reveal intriguing insights. We find that decoder-only language models may not be the optimal choice for MCQs in clinical knowledge tasks. Additionally, our investigation demonstrates that instruction fine-tuned language models do not necessarily outperform their counterparts in these evaluations, emphasizing the importance of carefully tailored model selection.

To foster research and collaboration in this field, we make our benchmark publicly available and open-source the associated evaluation scripts. This initiative aims to facilitate further advancements in clinical knowledge representation and utilization within language models, ultimately benefiting the healthcare and natural language processing communities.

1 Introduction

Large Language Models (LLM) deployed in the clinical and biomedical domains have the potential to revolutionize the healthcare industry. They are employed to summarize clinical text (Veen et al., 2023), automatically generate notes for clinicians (Ben Abacha et al., 2023b), and condense dialogues between doctors and patients (Ben Abacha et al., 2023a; Toma et al., 2023). Recognizing their significance, recent work (Han et al., 2023; Wu et al., 2023; Toma et al., 2023; Bolton et al., 2022; Li et al., 2023) has focused on fine-tuning LLMs on clinical and bio-medical datasets. However, the evaluation of LLMs within the clinical and biomedical domains remains incomplete and requires further comprehensive evaluation. Recent models tend to be evaluated on different datasets or tasks, which makes fair comparison of models harder.

Clinical knowledge assessment in LLMs involves two primary tasks (Singhal et al., 2023a): Multiple Choice Question Answering (MCQA), where answers are selected from multiple options, and abstractive question answering (AQA), which entails generating answers to questions, either with or without a provided paragraph context.

The evaluation of LLMs regarding their clinical knowledge is restricted. For instance, Singhal et al. (2023a) assess proprietary models on a consolidated dataset, but open-source LLMs are not tested on such a unified benchmark. We expand Singhal et al. (2023a)'s benchmark with more datasets, to enable transparency and reproducibility, and to test new advances in LLM research. Our benchmark called QALM consolidates existing MCQA and AQA datasets, featuring 16 MCQA and 6 AQA datasets. With such a standardized benchmark we are able to test the strengths and weaknesses of open-source models using a methodological unified framework, which is currently missing in the literature.

Our evaluation encompasses diverse zero-shot and fine-tuning settings. Our findings reveal that the latest decoder-only LLMs do not consistently outperform others on reading comprehension datasets, where models are presented with a context paragraph to answer a question. Moreover, although instruction fine-tuned models have been argued to surpass their non-instruction fine-tuned counterparts in some contexts (Wei et al., 2021; Gupta et al., 2023), our results suggest a more nu-

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

Table 1: An overview of the QALM datasets. We present the size in terms of train/val/test splits. We create a manual train/val split for BIOASQ-MCQ, PROCESSBANK, PUBMEDQA, BIOASQ-QA and MEDQUAD.

anced trend.

In summary, QALM introduces a benchmark that consolidates datasets to assess clinical knowledge within LLMs. This initiative underscores the imperative for a unified and comprehensive suite of datasets to rigorously evaluate the clinical knowledge capabilities of these models, particularly in light of the continuously expanding array of LLMs.

2 QALM Datasets

QALM represent a comprehensive collection of 22 datasets designed to thoroughly evaluate the clinical knowledge of LLMs. These datasets are publicly accessible, and some of them have not been used for testing open-source LLMs before. Like other studies (Singhal et al., 2023a) (Singhal et al., 2023b), we employ Question and Answering tasks as a surrogate test to assess clinical knowledge of LLMs. We have included two types of question-answering datasets: Multiple Choice (MCQA) and Abstractive Question Answering (AQA). The complete list of these datasets is provided in Table 1.

MCQA questions assess the model's ability to select the correct option from a list of challenging alternatives. These types of questions are frequently encountered in medical licensing examinations, such as the US Medical Licensing Exam (USMLE) (Jin et al., 2021), as well as in medical entrance exams in countries like India (Pal et al., 2022) and Spain (Vilares and Gómez-Rodríguez, 2019). Within QALM, there are a total of 216,810 instances of MCQA questions. On average, each question has four answer choices. In some scenarios, a contextual paragraph is provided and the answer must be derived from this context which tests the model's contextual reasoning abilities.

AQA datasets evaluate the model's proficiency in providing open-ended answers to questions. QALM encompasses six AQA datasets, containing a total of 57,958 questions. On average, the answers in these datasets have a length of about 100 tokens.

Our work is distinct from Singhal et al. (2023a), who only consider two AQA datasets, LiveQA and MedicationQA, as they argue that these datasets lack reliable sources for answers. In contrast, we incorporate four additional datasets where answers are provided from experts or sourced from trusted forums, increasing their reliability. Furthermore, we use a version of the LiveQA dataset that contains expert-ranked answers (Ben Abacha and Demner-Fushman, 2019). The performance of LLMs on these datasets have not been previously assessed in existing literature.

While Singhal et al. (2023a) employ only a single dataset for assessing model performance in reading comprehension, we introduce two more

datasets: one for evaluating a model's ability to predict a concealed medical entity, and another for assessing model proficiency in comprehending interactions between various entities mentioned in the text. Furthermore, we include four additional general knowledge multiple-choice question (MCQ) datasets, broadening the range of questions for model testing. This expanded array of evaluation datasets enables a more comprehensive analysis of model capabilities, potentially mitigating uncertainty stemming from a limited range of observations.

3 Empirical Evaluation

Considering the QALM datasets, we seek evidence for the following key research questions in a large-scale empirical study:

RQ1. How well do open-source language models (LLMs) recall necessary clinical knowledge when they are tested on QALM?

RQ2. Does instruction fine-tuning of LLMs improve their clinical knowledge recall?

RQ3. Does domain and task-specific finetuning on QALM help LLMs acquire additional clinical knowledge?

RQ4. Can LLMs generalize to data unseen during training?

3.1 Study Setup

To seek evidence for **RQ1** and **RQ2** empirically, we evaluate several LLMs and their instruction-finetuned versions on the test splits of QALM in zero-shot manner. To answer **RQ3** and **RQ4**, we fine-tune LLMs on the training portion of 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) and `LLama 2` and `LLama 2-chat` (7B and 13B). 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 2.[2] We initially also considered `OpenLLaMA` (3B and 7B) (Geng and Liu, 2023) as well as `GPT-J` (6B) (Wang and Komatsuzaki, 2021). However, we found that due to the poor performance of `OpenLLaMA` on the MCQA datasets and since `GPT-J` and `OpenLLaMA` did not have instruction-finetuned equivalents, we did not include them in our further analysis.

To address **RQ3**, we use the training set of the 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 **RQ4**.

Finetuning and hyperparameters: Since the number of parameters for most of our models are in the billions, we follow a more accepted practice of using parameter-efficient fine-tuning. Specifically, we use QLora and 4-bit quantization (Dettmers et al., 2023) for fine-tuning. We use A100-40G GPUs for all our experiments. The other hyperparameters used to train our models are reported in the Appendix (Table 7).

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)[3] and `METEOR` (Banerjee and Lavie, 2005). METEOR in particular is found to correlate better with human judgments than other metrics on AQA (Chen et al., 2019).

3.2 Results and Analysis

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

3.2.1 Zero-shot performance of LLMs

Table 3 presents the zero-shot evaluation of language models as evidence towards **RQ1**. We report the average accuracies across all tasks.

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

[3] We use deberta-xlarge-mnli for calculating BERTScore.

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)
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	All of MPT and *Databricks Dolly-15k (Conover et al., 2023), Anthropic Helpful and Harmless (Bai et al., 2022)*
Falcon-Instruct	Decoder	1.5T	All of Falcon and *baize (Xu et al., 2023), GPT4All, GPTeacher* [1]
LLama 2-Chat	Decoder	2T	All of LLama 2 and *Flan Collection (Wei et al., 2021) + Private Data*

Table 2: 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*).

		MCQA	AQA		
		Acc	RL	BS	MTR
Base	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

Table 3: Zero-shot performance of Base models (top) and instruction-tuned models (bottom). Metrics are **Acc**uracy for MCQA; **R**ouge-**L**, **B**ERT**S**core, and **METEOR** for AQA.

LLMs are good zero/few-shot learners: Table 3 shows that LLMs exhibit strong zero-shot capability on MCQA and AQA datasets, corroborating the findings of Singhal et al. (2023a). Across different LLMs of the same size, LLama 2 (7B) provides the best performance compared to other open source models like Falcon (7B). A possible reason for this might be the diversity in pretraining data – LLama 2 is trained on two trillion tokens, which is double the amount of Falcon . Another plausible reason is the mixture of datasets used for pretraining them.

Bigger LLMs have more clinical knowledge: Figure 1 shows the relationship between the number of parameters and performance. It suggests that model performance tends to improve with scale (correlation between scale and performance: Spearman's $\rho = 0.24, p < 0.01$), and that language models exhibit emergent abilities with scale. For example, LLama 2 (13B) has 23% (relative) better accuracy than LLama 2 (7B) on MCQA and 6% higher METEOR scores. Even without further domain-specific adaptation of LLMs on clinical data, scale appears to play a major role in the amount of clinical knowledge available to LLMs.

Open-source LLMs do not outperform humans: While the passing score for USMLE is 60% for humans[4], we observe the best score for USMLE to be 43% (LLama 2), which is 17% short of that requirement. Meanwhile, Singhal et al. (2023a) report scores of 86.5% on USMLE. Similarly, for the PubmedQA dataset, human performance is 78% (Jin et al., 2019), compared to 60.4% of LLama 2.

To summarize our findings related to **RQ1**: While LLMs have good clinical knowledge, there is still a significant gap compared to humans (Singhal et al., 2023a).

3.2.2 Effect of Instruction Fine-tuning

Instruction fine-tuning (Wei et al., 2022; Ouyang et al., 2022) was proposed as a means to improve across-the-board performance on various open-domain tasks. To address **RQ2**, we investigate whether these improvements also apply to the clinical domain of QALM. The results are reported in the bottom part of Table 3.

Instruction fine-tuning is key: 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 2). Among them, Flan-T5 models show the best zero shot performance on MCQA, outperforming all comparable decoder-only models.

We observe that the decoder-only language models generate longer and more verbose answers compared to the encoder-decoder models, probably due to their causal language modelling training objective. As such, they perform better on AQA. Instruction fine-tuning models can consistently improve decoder-only models as well (compare LLama 2 with LLama 2-chat for example).

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

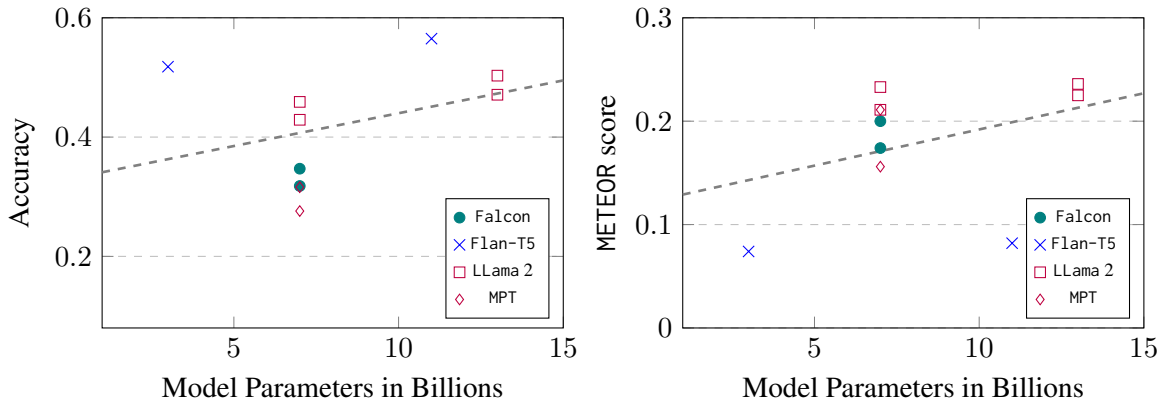

Figure 1: Zero-shot performance of models on MCQA (left) and AQA (right) as a function of model size. The dashed line represents a fitted linear regression showing the correlation between the model size and the score.

Bigger models are not always better: The choice of model architecture and the 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), despite being four times its size.

	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 $_{0.2}$	60.4 $_{2.0}$	17.1 $_{0.3}$
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 $_{7.4}$

Table 4: Caption: Model finetuning is performed either on MCQA or their AQA datasets. Evaluation is performed using **Acc**uracy for MCQA, and **R**ouge-**L**, **BERTS**core, and **METEOR** for AQA. The subscripts indicate the improvement over the zero-shot versions.

3.2.3 Impact of Finetuning

Given the scale of QALM, we are able to fine-tune models on parts of the data, to address **RQ3**. Specifically, we fine-tune four models on MCQA and AQA separately, given the different nature of these datasets.[5]

MCQA fine-tuning improves knowledge: We fine-tune the models only on the MCQA subset of datasets first (c.f. Table 4). We find that the models perform better compared to their non-fine-tuned counter parts. Decoder-only models like MPT (7B) benefit more than others (+25.6 percentage points Accuracy improvement). Interestingly, fine-tuning models on the data seems to close the gaps introduced by different model architectures and pre-training data, discussed in the previous Section. Specifically, the standard deviation of the model Accuracy in zero-shot setting is 9.0, while after fine-tuning it is reduced to 1.7. This suggests that various LLMs can benefit from task-specific fine-tuning to address seemingly sub-optimal architecture or pre-training conditions.

AQA fine-tuning can improve knowledge: We fine-tune the models on AQA datasets. Encoder-Decoder models like Flan-T5 benefit more from AQA fine-tuning compared to the decoder-only models. For instance, the fine-tuned Flan-T5 model has significantly better METEOR (+7.8) compared to the non fine-tuned version.

Fine-tuning can compensate for scale: Scaling up brings 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. Fine-tuning helps compensate for scale. LLama 2 (7B) significantly outperforms the zero-shot LLama 2 (13B) (+6.4 Accuracy gain on MCQA, +5.7 METEOR gain on AQA). Similarly, we observe that a fine-tuned Flan-T5 (3B) outperforms zero shot Flan-T5 (11B) on 9 out of 16 MCQA datasets and also on AQA tasks considerably.

Fine-tuned models make similar mistakes: In order to provide targeted suggestions for future performance improvements, we analyze the types of errors that fine-tuned models make on the MCQ datasets. To do this, we perform entity recognition using SciSpacy (Neumann et al., 2019) over

[5]We also experimented with fine-tuning models on MCQA and AQA jointly, but the results did not differ significantly from those reported here.

Question type and template	Support	LLama 2 (7B)	Falcon (7B)	MPT (7B)	Flan-T5 (3B)
Scenario-Based Treatment (*A {age}-year-old person...*)	1172	**48.37%**	45.31%	44.8%	40.45%
Fact-Based Discriminative (*Which of the...*)	1053	**54.9%**	47.95%	53.18%	45.68%

Table 5: Analysis of the most common question types and finetuned model performance. As these questions form a sub-set of the test set, averaging the reported scores does not correspond to the performance on the full test set reported in Table 4.

the questions and link these entities with UMLS to obtain their type unique identifier (TUI). We map these TUIs to their semantic groups[6]. We replace the original entities in the question with their respective semantic group.

We first use an automated way of analyzing errors . We categorize the questions based on the first three words of the questions, similar to (Yang et al., 2018). The example of categories are shown in Table 5, where *Scenario-Based Treatment Questions* describe a real-world case, and the model needs to chose the right course of treatment for the case. The second category *Fact-Based Questions* forces the model to rely on the internal knowledge stored in its parameters to answer the questions.

We find that models perform worse than average on *Scenario-Based Treatment Questions*: the performance drops when they have to combine knowledge encoded in the internal parameters with particulars described in the scenario. Among different models, Flan-T5 and Falcon have a more stable performance across question categories, compared to LLama 2 and MPT (standard deviations of 2.6 and 1.9, vs 3.4 and 3.9).

We further analyze the errors made by the models manually, by randomly choosing 50 samples where the model was correct and 50 samples where the model was wrong. We consider the best performing LLama 2 (7B) model for our analysis. We identify five categories of common errors and report them in Table 6, where *Scenario-Based Treatment Questions* and *Fact-Based Questions* correspond to the previously described common question types. Meanwhile, questions concerning *Procedures, Tests and Activities* require knowledge of medical procedures (e.g., the type of test needed to identify a certain condition), *Anatomy and Physiology* related questions require knowledge of anatomy, and *Chemicals and Drugs* require pharmaceutical knowledge.

[6] https://lhncbc.nlm.nih.gov/ii/tools/MetaMap/documentation/SemanticTypesAndGroups.html

Category	% Wrong	% Correct
Scenario-Based Treatment	26%	30%
Procedures, Tests and Activities	14%	6%
Fact-based Discriminative	12%	14%
Anatomy and Physiology	10%	12%
Chemicals and Drugs	10%	10%

Table 6: Manual analysis of errors and correct predictions of the best-performing fine-tuned LLama 2 (7B) model.

The manual error analysis corroborates the previous finding that fine-tuned models tend to err when asked about treatments given a scenario, possibly because this is the most frequently occurring category in the training data. Additionally, our findings suggest that models struggle with procedure-based questions, as these can exhibit significant complexity. For example, to successfully select the correct option for the question *"Recent studies have shown that the chlorhexidine-isopropyl alcohol solution substantially reduces the risk of surgical site infections compared with a povidone-iodine preparation without alcohol in clean-contaminated surgery. Which of the following mechanisms best describes the mechanism of action of chlorhexidin?"*, a model needs to not only possess knowledge about the mechanism of action of chlorhexidin, but also relate it to research findings which suggest that it reduces the risk of surgical site infections (i.e. that it *"[...] is greatly reduced in the presence of organic matter"* in this case).

3.2.4 Generalisation to Unseen Data

Finally, we report the potential of LLMs fine-tuned on in-domain data to generalise to medical datasets unseen during training to answer **RQ4**. We hold out 4 AQA and 10 MCQA datasets presented in Figures 2 and 3.

AQA-finetuned models generalise to unseen AQA test sets: Figure 2 shows the performance of LLama 2 (7B) and Flan-T5 (3B) models on the four held-out AQA evaluation sets. The METEOR

scores of the fine-tuned LLama 2 model are lower than the zero-shot baseline (−4.2 METEOR). Meanwhile, fine-tuning Flan-T5 improves performance on all four unseen datasets (+8.2 METEOR), suggesting that instruction fine-tuned models have good generalisation capabilities (Wei et al., 2022) compared to models that were only pre-trained on the language modelling task.

AQA-finetuned models do not generalise to unseen MCQA test sets: Figure 3 (comparing ZS with AQA-FT) 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 generations to the expected answer form of AQA answers, rather than acquiring additional medical knowledge during fine-tuning.

MCQA-finetuned models do generalise to unseen MCQA test sets: Figure 3 (comparing ZS with MCQ-FT) suggests that models indeed can learn to extract relevant 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. To investigate this mismatch, we conduct a manual analysis.

Fine-tuned models may memorize rather than generalize: 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 QALM: Clinical

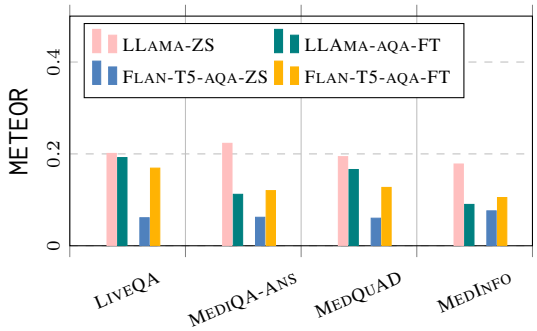

Figure 2: Performance of base and AQA-finetuned LLama 2 and Flan-T5 models on four unseen AQA test sets.

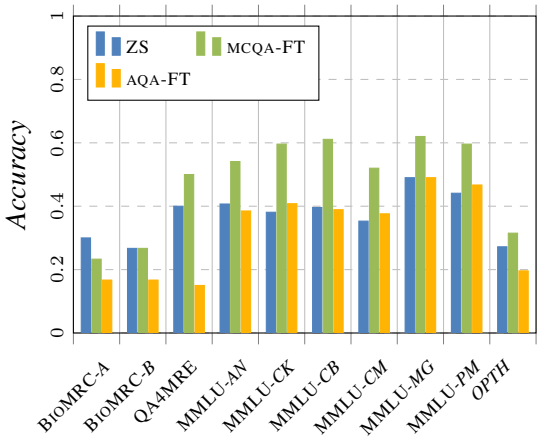

Figure 3: Performance of base, MCQA-fine-tuned and AQA-finetuned LLama 2 model on ten unseen MCQA test sets.

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 six out of 92, twelve 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 five, two, and five 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, seven questions were near-duplicates with identical answers, while the remaining five would have required some level of clinical understanding for the model to answer them correctly. These findings suggest that the improved performance of instruction-tuned models on unseen datasets can be partially attributed to exposure to near-identical questions during training.

4 Related Work

Open source LLMs: The recent proliferation of LLMs has sparked a surge of interest in their adap-

tation for clinical applications. Notably, closed-source models like Med-PaLM and Med-PaLM2 (Singhal et al., 2023a,b) have shown considerable promise in leveraging LLMs within the clinical context. Meanwhile, open-source LLMs are being released at an astonishing pace. Nevertheless, it is crucial to note that these models have yet to be rigorously assessed for their clinical knowledge, and there is currently no existing benchmark for this purpose, a gap that this study addresses.

Benchmarks for clinical knowledge: The adaptation of Large Language Models (LLMs) to the medical domain is gaining significant traction as they exhibit substantial potential in addressing clinical and medical issues. In their groundbreaking work, Singhal et al. (2023a) introduced MultiMedQA, a benchmark that consolidates six distinct medical question-answering datasets, along with HealthSearchQA, a novel free-response dataset designed for medical inquiries. Their research involved an evaluation of Flan-PaLM (Chowdhery et al., 2022; Chung et al., 2022) on MultiMedQA, alongside the proposal of instruction prompt tuning to enhance medical reasoning and customize LLMs for the medical domain, ultimately giving rise to Med-PaLM. Med-PaLM 2 (Singhal et al., 2023b) extended this groundwork by incorporating PaLM2 (Anil et al., 2023) and leveraging medical domain fine-tuning, along with prompting strategies. These enhancements resulted in an impressive 86.5% accuracy rate on the United States Medical Licensing Examination (USMLE). Building upon this foundational work, our aim is to introduce a more comprehensive benchmark, encompassing a broader range of datasets with diverse characteristics, to thoroughly evaluate LLMs.

While most of these benchmarks assess LLMs' capabilities in question-answering tasks, it's crucial to address the aspect of hallucination, which is of paramount importance in clinical and medical applications. In response to this concern, Umapathi et al. (2023) introduced Med-HALT, a benchmark designed to evaluate LLMs on hallucinations.

While some of these benchmarks evaluate LLMs' clinical knowledge, they may fall short in assessing their competence in performing various tasks within real-world clinical settings. A recent development in this regard comes from Fleming et al. (2023), who introduced Med-Align. This innovative benchmark dataset comprises natural language directives related to electronic health record (EHR) data, which is a longitudinal record of a patient's medical history. Med-Align serves the vital purpose of evaluating the efficacy of LLMs in processing clinical instructions within a clinical context.

5 Conclusion

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

Our findings reveal that while these LLMs exhibit performance significantly superior to random guessing, there remains room for enhancing their performance when compared to closed-source language models. Notably, fine-tuning on QALM demonstrates the potential to augment a language model's clinical knowledge, especially in the context of instruction fine-tuned models like `Flan-T5`. However, we acknowledge that some performance improvements may be attributed to the nature of the test questions.

It is important to note that scale and decoder-only language models do not serve as universal solutions for all questions in clinical question-answering. To pave the way for future research in this domain, we emphasize the necessity of considering the architecture of language models, the choice of datasets for instruction fine-tuning, and conducting a rigorous evaluation of the knowledge contained within LLMs. These aspects are vital for advancing the state-of-the-art in clinical NLP and expanding the horizon.

We make the dataset, experiment code and evaluation protocol publicly available under `https://anonymized`. This will allow future LLM developers to perform a fine-grained analysis of their models clinical and biomedical knowledge.

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 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 investigate 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)*.

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.

Rohan Anil et al. 2023. Palm 2 technical report. *arXiv*, 2305.10403.

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*.

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.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi,

David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *arXiv*, 2204.02311.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *arXiv*, 2210.11416.

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

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.

Xinyang Geng and Hao Liu. 2023. Openllama: An open reproduction of llama.

Himanshu Gupta, Saurabh Arjun Sawant, Swaroop Mishra, Mutsumi Nakamura, Arindam Mitra, Santosh Mashetty, and Chitta Baral. 2023. Instruction tuned models are quick learners.

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.

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

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.

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.

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.

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.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and robust models for biomedical natural language processing. In *Proc. 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

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. a. Frcophth sample mcqs part 1.

RCOphth. b. Frcophth sample mcqs part 2.

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.

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.

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.

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.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. `https://github.com/kingoflolz/mesh-transformer-jax`.

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

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.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. 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.

Appendix

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 7: 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 8: Inference time parameters used for abstractive question answering