Domain Specific Knowledge for Large Language Models

Ameen Izhac, Pedro M. Baiz V, Marek Rei, MEng Individual Project

Motivation

Large Language Models are

Large Language Models are

Intelligent

Large Language Models are

Intelligent

Creative

Large Language Models are

Intelligent

Creative

Big liars

Applications

Applications

Energy regulations

Applications

Energy regulations

Legal information

Applications

Energy regulations

Legal information

Medical information

Tackling the Problem

Tackling the Problem

We can either

Tackling the Problem

We can either

Teach an LLM to say “I don’t know” [1]

[1] Sanyam Kapoor, Nate Gruver, Manley Roberts, Katherine Collins. Large Language Models Must Be Taught to Know What They Don't Know. 2024

Tackling the Problem

We can either

Teach an LLM to say “I don’t know” [1]

Make sure the LLM knows

[1] Sanyam Kapoor, Nate Gruver, Manley Roberts, Katherine Collins. Large Language Models Must Be Taught to Know What They Don't Know. 2024

Requirements

Requirements

Comprehension

Requirements

Comprehension

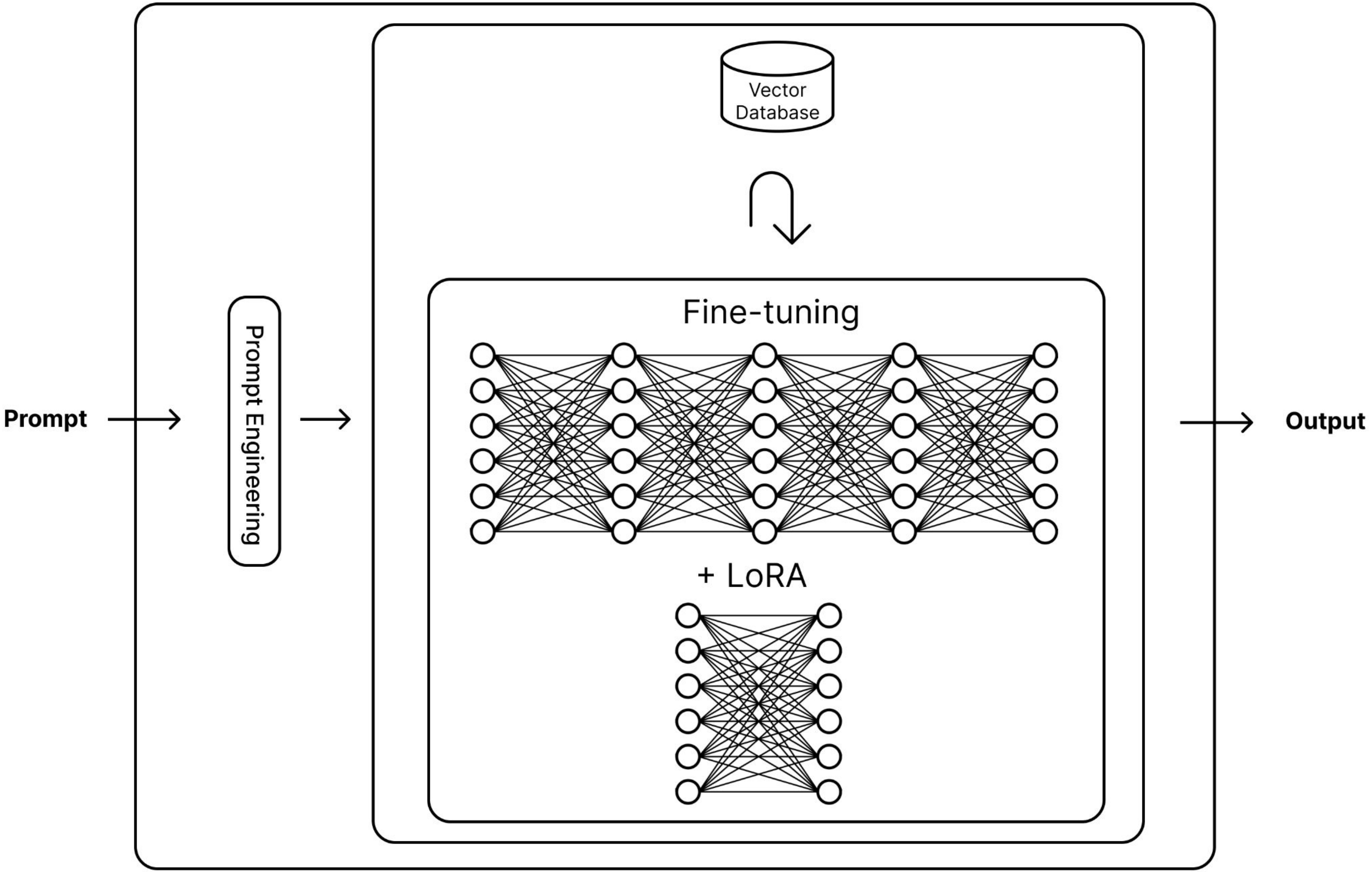
Communication

Requirements

Comprehension

Communication

Knowledge


Prompt
Prompt Engineering
Vector
Database
Fine-tuning
+ LoRA
Output

Overview

Flan-T5 Base evaluation encouraged adopting a bigger model

Llama 3 8B answering Closed Answer Anatomy Questions

Multiple Choice Evaluation

Improvement via Prompt Engineering and RAG

Flan T5 - CoQA

`Have you ever been to some big cities in the world? The information below will be helpful to you. Budapest For many centuries...`

`Was Budapest always one city?`

`no`

`How many was it?`

`two`

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	BLEURT
base	0.50	0.45	0.40	0.35	0.044
fine-tuned	**0.68**	**0.64**	**0.59**	**0.54**	**0.38**

EM	precision	recall	F1
0.42	0.71	0.74	0.69
0.56	**0.84**	**0.82**	**0.82**

Flan T5 - SQuAD

```
Architecturally, the school has a Catholic character. Atop the Main Building's gold
dome is a golden statue of the Virgin Mary. Immediately in front of the Main
Building...
```

```
To whom did the Virgin Mary allegedly appear in 1858 in Lourdes France?
```

```
Saint Bernadette Soubirous
```

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	BLEURT
base	0.80	0.80	0.74	0.61	0.78
fine-tuned	**0.85**	**0.82**	**0.75**	**0.62**	**0.80**

EM	precision	recall	F1
0.76	0.90	0.74	0.89
0.80	**0.92**	**0.90**	**0.90**

Flan T5 - MedMCQA

```
Chronic urethral obstruction due to benign prismatic hyperplasia can lead to the following change in kidney parenchyma

Hyperplasia
Hyperophy
Atrophy
Dyplasia

Atrophy
```

Model	EM
base	0.16
random choice	0.28
fine-tuned	0.34
fine-tuned (extensive)	**0.37**

Flan T5 - CNN and Daily Mail

```
MINNEAPOLIS, Minnesota (CNN) -- Drivers who were on the Minneapolis bridge
when it collapsed told harrowing tales of survival. "The whole bridge from one
side of the…
```

```
NEW: "I thought I was going to die," driver says . Man says pickup truck was
folded in half; he just has cut on face
```

Model	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-LSUM	BLEURT
base	**0.31**	**0.16**	**0.26**	**0.26**	**-0.78**
fine-tuned	0.30	0.15	**0.26**	**0.26**	-0.85

Llama 3 8B

Results Table

Dataset	Llama 3 8B Base	Llama 3 8B Final
MedMCQA	0.462	**0.585** (12.3%)
MedMCQA-Anatomy	0.534	**0.632** (9.8%)
MedQA (USMLE)	0.513	**0.581** (6.8%)
PubMedQA	**0.722**	0.436 (-28.6%)
MMLU-Anatomy	0.474	**0.578** (10.4%)

Multiple Choice - Exact Match

Prompt Engineering

Prompt Engineering

Few-Shot

Prompt Engineering

Few-Shot

Chain of Thought

Prompt Engineering

Few-Shot

Chain of Thought

System Prompting

Prompt Engineering

Few-Shot

Chain of Thought

System Prompting

Dataset	Llama 3 8B Base	Llama 3 8B (PE)
MedMCQA	0.462	**0.517** (5.5%)
MedMCQA-Anatomy	0.534	**0.585** (5.1%)
MedQA (USMLE)	0.513	**0.573** (6.0%)
PubMedQA	**0.722**	**0.722** (0.0%)
MMLU-Anatomy	0.474	**0.578** (10.4%)

RAG

RAG

MedMCQA Questions

RAG

MedMCQA Questions

Dataset	Llama 3 8B Base	Llama 3 8B RAG
MedMCQA	0.462	**0.585** (12.3%)
MedMCQA-Anatomy	0.534	**0.632** (9.8%)
MedQA (USMLE)	0.513	**0.581** (6.8%)
PubMedQA	**0.722**	0.436 (-28.6%)
MMLU-Anatomy	0.474	**0.578** (10.4%)

RAG

MedMCQA Questions

Anatomy Textbook(s)

Dataset	Llama 3 8B Base	Llama 3 8B RAG
MedMCQA	0.462	**0.585** (12.3%)
MedMCQA-Anatomy	0.534	**0.632** (9.8%)
MedQA (USMLE)	0.513	**0.581** (6.8%)
PubMedQA	**0.722**	0.436 (-28.6%)
MMLU-Anatomy	0.474	**0.578** (10.4%)

RAG

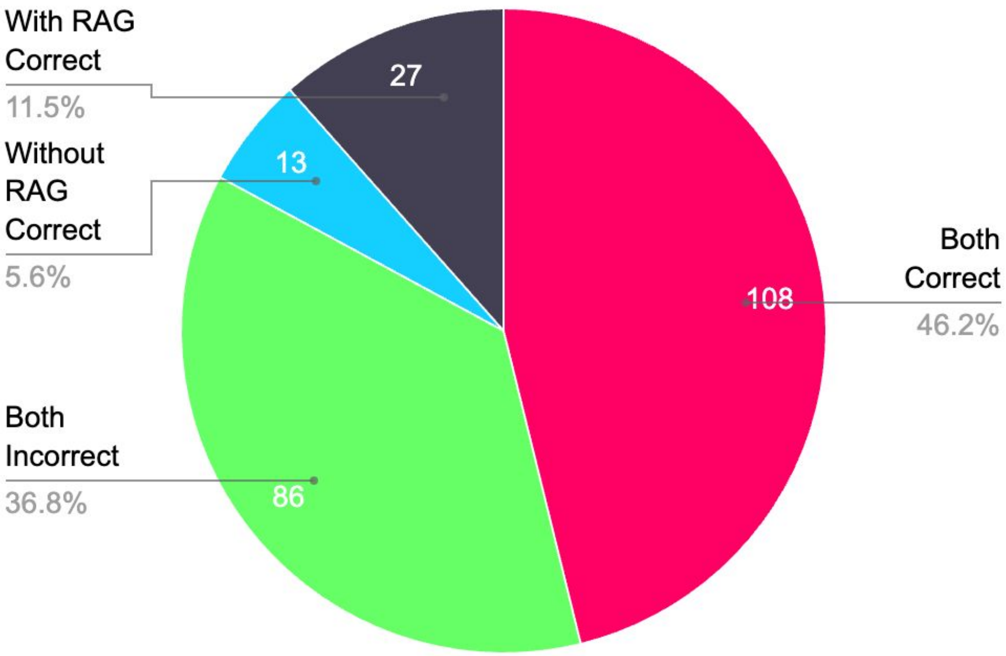
MedMCQA Questions

Anatomy Textbook(s)

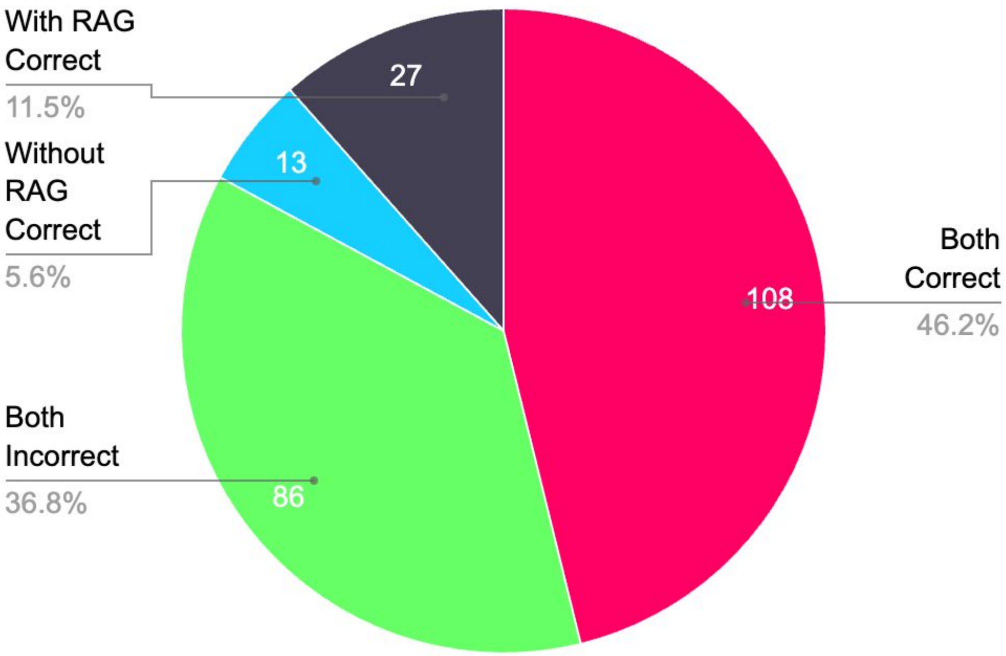
Providing Definitions

Dataset	Llama 3 8B Base	Llama 3 8B RAG
MedMCQA	0.462	**0.585** (12.3%)
MedMCQA-Anatomy	0.534	**0.632** (9.8%)
MedQA (USMLE)	0.513	**0.581** (6.8%)
PubMedQA	**0.722**	0.436 (-28.6%)
MMLU-Anatomy	0.474	**0.578** (10.4%)

Impact of RAG



Impact of RAG



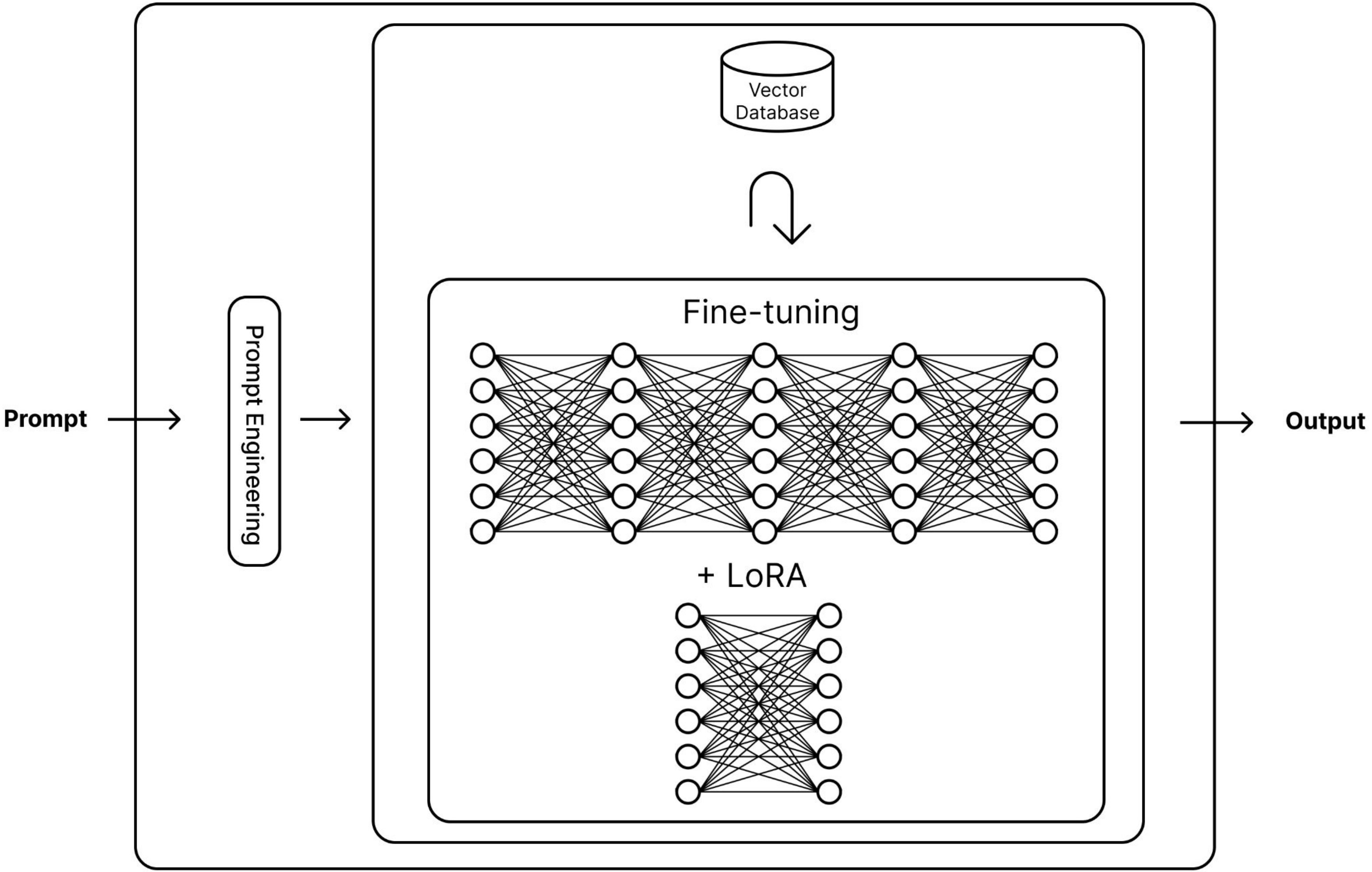
PubMedQA	**0.722**	0.436 (-28.6%)

LoRA

Effective Batch Size	Samples	Learning Rate	LoRA Rank	LoRA Alpha	Exact Match
-	-	-	-	-	0.585
128	4,000	2e-4	32	64	0.585
128	4,000	2e-4	64	128	0.585
128	4,000	2e-4	128	256	0.585
128	4,000	2e-4	256	512	0.585
128	3,500	1e-5	256	512	0.585
128	3,500	3e-5	256	512	0.585
128	4,000	5e-5	256	512	0.585
128	4,000	7e-5	256	512	0.585
128	4,000	9e-5	256	512	0.585
64	3,800	4e-5	64	128	0.585

LoRA

Effective Batch Size	Samples	Learning Rate	LoRA Rank	LoRA Alpha	Exact Match
64	3,800	1e-4	64	128	0.585
64	3,800	3e-4	32	64	0.585
128	4,000	5e-4	128	256	0.585
256	5,000	6e-4	128	256	0.585
128	4,000	2e-4	16	32	0.585
128	8,000	2e-4	16	32	0.585
128	14,560	2e-4	16	32	0.585


Prompt
Prompt Engineering
Vector Database
Fine-tuning
+ LoRA
Output

[Submitted on 9 May 2024 (v1), last revised 13 May 2024 (this version, v2)]

Does Fine-Tuning LLMs on New Knowledge Encourage Hallucinations?

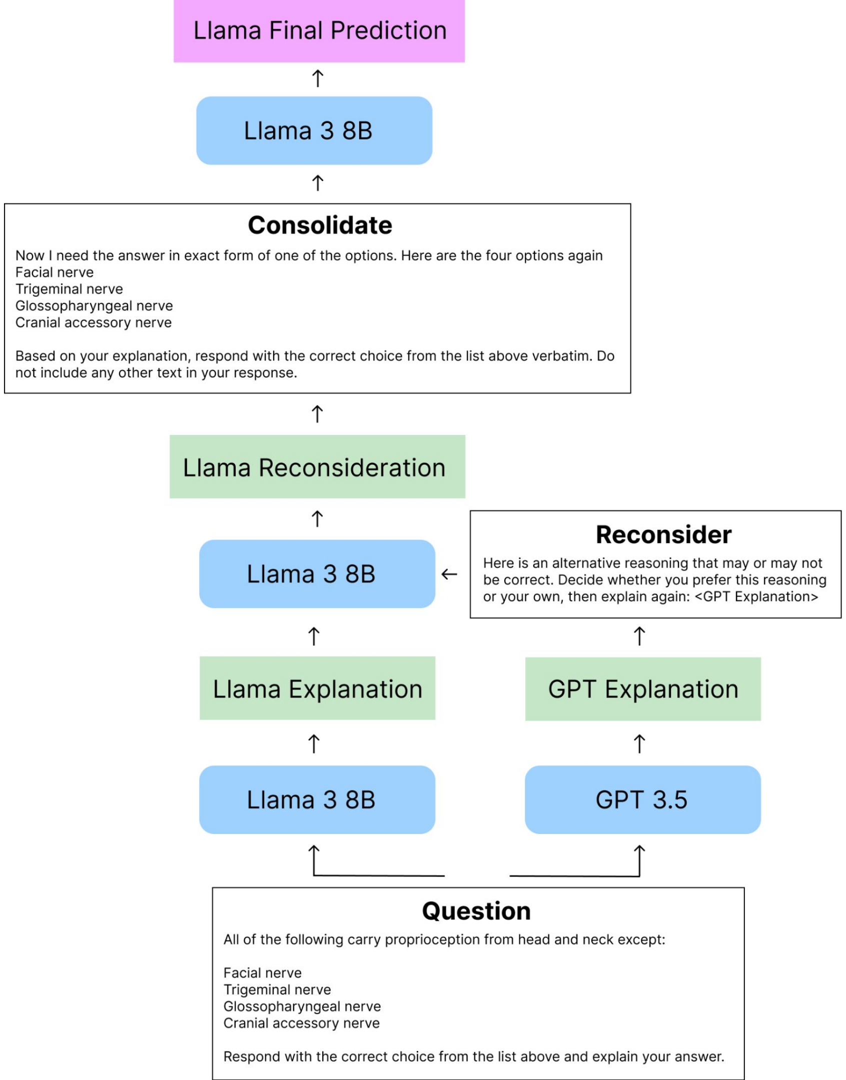
Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, Jonathan Herzig

When large language models are aligned via supervised fine-tuning, they may encounter new factual information that was not acquired through pre-training. It is often conjectured that this can teach the model the behavior of hallucinating factually incorrect responses, as the model is trained to generate facts that are not grounded in its pre-existing knowledge. In this work, we study the impact of such exposure to new knowledge on the capability of the fine-tuned model to utilize its pre-existing knowledge. To this end, we design a controlled setup, focused on closed-book QA, where we vary the proportion of the fine-tuning examples that introduce new knowledge. We demonstrate that large language models struggle to acquire new factual knowledge through fine-tuning, as fine-tuning examples that introduce new knowledge are learned significantly slower than those consistent with the model's knowledge. However, we also find that as the examples with new knowledge are eventually learned, they linearly increase the model's tendency to hallucinate. Taken together, our results highlight the risk in introducing new factual knowledge through fine-tuning, and support the view that large language models mostly acquire factual knowledge through pre-training, whereas fine-tuning teaches them to use it more efficiently.

Second Opinion

	Meta Llama 3 70B	Gemini Pro 1.5 Published	Claude 3 Sonnet Published
MMLU 5-shot	**82.0**	81.9	79.0
GPQA 0-shot	39.5	**41.5** CoT	38.5 CoT
HumanEval 0-shot	**81.7**	71.9	73.0
GSM-8K 8-shot, CoT	**93.0**	91.7 11-shot	92.3 0-shot
MATH 4-shot, CoT	50.4	**58.5** Minerva prompt	40.5

Second Opinion



Second Opinion

Dataset	GPT 3.5	GPT 3.5 (SO)	Llama 3 8B	Llama 3 8B (SO)
MedMCQA	**0.500**	0.449	0.449	0.487
MedMCQA-Anatomy	0.521	0.551	0.543	**0.603**
MedQA (USMLE)	**0.645**	0.564	0.526	0.598
PubMedQA	0.449	0.808	**0.821**	0.701
MMLU-Anatomy	0.644	0.563	0.570	**0.667**

Dataset	GPT 3.5 SO Gain	Llama 3 8B SO Gain	Disagreement
MedMCQA	-0.1%	3.8%	29.1%
MedMCQA-Anatomy	3.0%	6.0%	34.2%
MedQA (USMLE)	-8.1%	7.2%	29.1%
PubMedQA	35.9%	-12.0%	44.9%
MMLU-Anatomy	-8.1%	9.7%	17.1%
Mean	4.5%	2.9%	30.9%
Standard Deviation	7.3%	7.7%	4.0%

Conclusion and Future Work

Custom Knowledge Base

Dynamic Knowledge Base

Long Answer Subdomain