Should We Fine-Tune or RAG? Evaluating Different Techniques to Adapt LLMs for Dialogue

Simone Alghisi [†], Massimo Rizzoli [†], Gabriel Roccabruna, Seyed Mahed Mousavi, Giuseppe Riccardi
Signals and Interactive Systems Lab, University of Trento, Italy
{s.alghisi, massimo.rizzoli, giuseppe.riccardi}@unitn.it

https://arxiv.org/abs/2406.06399

San Diego Machine Learning
Ryan Chesler

Overview of Large Language Models

- Huge models trained against text crawls of the internet to guess the next token

- Able to learn structure of language and some factual knowledge from all of the information it is trained against

- Cannot know anything about stuff that happened outside of its training set

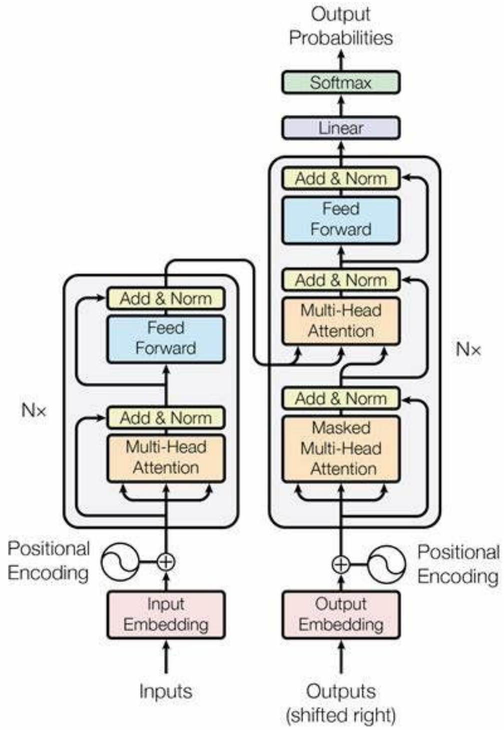
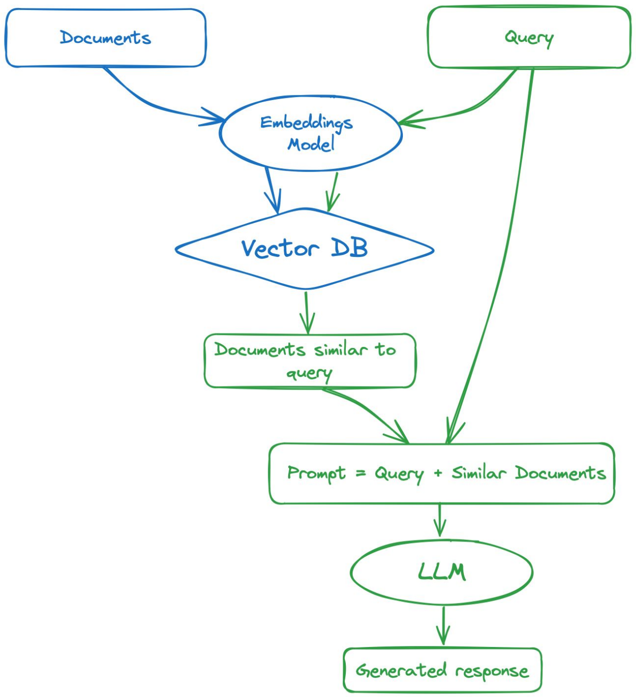


Figure 1: The Transformer - model architecture.

Overview of Retrieval Augmented Generation

- Motivation: Give the large language model the relevant context to respond correctly
- Done by creating a knowledge store and then using a retrieval system to extract information related to the users query and passing to the LLM
- Heavily reliant on the strength of being able to retrieve the correct documents



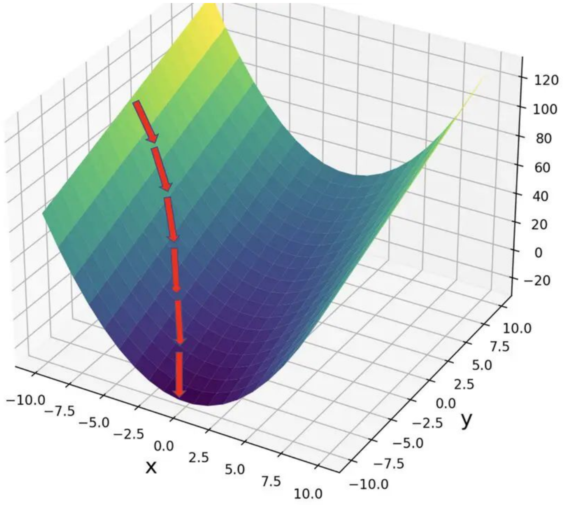
Phases of LLMs

PRE-TRAINING	FINETUNING	VECTOR DATABASE	ICL
All Weights and Params Modified	All Weights and Params Modified	Document Embedding	In-Context Learning
1,000 x H100 (80GB) GPUs 10,000 hours of training Trained using Internet Data USD ~$10M	A lot less GPU memory needed A lot less hours of training Trained using Tasked-Based Data A lot less costs involved	Complements LLMs for More Precise Responses	Inference Advanced Prompt Engineering
These are the Foundation or Base Models	**Instruct-Tuning (CoT)** Using tasked-based public data **(Task-Specific) Finetuning** Using tasked-based private data **Distillation** Using big LLMs to create responses, then use that to train smaller models	Uses Similarity Search on Private Documents Uses LLMs for Coherence and Stylistic Responses	Uses Few Shot Learning
Analogy 12 Years of Primary and Secondary Schooling	Analogy 6 Years of University Bachelors and Masters (Bootcamp)	Analogy Private Notes, References, And Training Materials	Analogy At Work, the “Job”, Applied Skills

Fine-tuning

- Training a model to do something more specific than just guess the next token
- Training it to follow instructions or teach it new information
- Huge downside that if your data is more narrow than the first phase it might lose its ability to do other language tasks

  https://huggingface.co/datasets/OpenAssistant/oasst1



Should We Fine-Tune or RAG?

- "Our analysis shows that there is no universal best-technique for adapting large language models as the efficacy of each technique depends on both the base LLM and the specific type of dialogue"

- Evaluated across four domains
  - Open-Domain Dialogue
  - Knowledge-Grounded Dialogue
  - Task-Oriented Dialogue
  - Question Answering

Open Domain Dialogue

- Daily Dialog - https://arxiv.org/abs/1710.03957

A: I'm worried about something.
B: What's that?
A: Well, I have to drive to school for a meeting this morning, and I'm going to end up getting stuck in rush-hour traffic.
B: That's annoying, but nothing to worry about. *Just breathe deeply when you feel yourself getting upset.*
A: Ok, I'll try that.
B: Is there anything else bothering you?
A: Just one more thing. A school called me this morning to see if I could teach a few classes this weekend and I don't know what to do.
B: Do you have any other plans this weekend?
A: I'm supposed to work on a paper that'd due on Monday.
B: *Try not to take on more than you can handle.*
A: You're right. I probably should just work on my paper. Thanks!

Figure 1: An example in **DailyDialog** dataset. Some text is shortened for space. Best viewed in color.

Knowledge-Grounded Dialogue

- Wizard of Wikipedia - https://arxiv.org/abs/1811.01241

Topic:	Lifeguard
Apprentice:	So I am a lifeguard. Know anything about saving lives in water?
Wizard:	I'm impressed! It's a big responsibility to supervise other people's safety in the water! Tell me more.
Apprentice:	Well, I help make sure people do not drown or get injured while in or near the water!
Knowledge:	A lifeguard is a rescuer who supervises the safety and rescue of swimmers, surfers, . . . Lifeguards are strong swimmers and trained in CPR/AED first aid, certified in water . . . . . . In some areas, the lifeguard service also carries out mountain rescues, or may function as the primary EMS provider.
Wizard:	I've heard that in some places, lifeguards also help with other sorts of emergencies, like mountain rescues! Is that part of your job too?
Apprentice:	I have! I feel like you know much about this! What brings you to know so much?
Wizard:	Oh, that's about the extent of my knowledge. I've just been around beaches and I've always admired lifeguards. I'm not a super strong swimmer myself.

Task-Oriented Dialogue

- Ninth Dialog System Technology Challenge: DSTC9 - https://arxiv.org/abs/2011.06486

Task #1	**Knowledge-seeking Turn Detection**
Goal	To decide whether to continue existing flow or trigger the knowledge access branch for a given utterance and dialog history
Input	Current user utterance, dialog context, and domain API and knowledge sources
Output	Binary class (requires knowledge access or not)
Task #2	**Knowledge Selection**
Goal	To select proper knowledge sources from the domain knowledge-base given dialog context at each turn with knowledge access
Input	Current user utterance, dialog context, and the entire set of knowledge candidates
Output	Ranking of top-k knowledge candidates
Task #3	**Knowledge-grounded Response Generation**
Goal	To generate a system response for a given triple of input utterance, dialog context, and the selected knowledge sources
Input	Current user utterance, dialog context, and selected knowledge sources
Output	Generated system response

Question Answering

- NarrativeQA - https://huggingface.co/datasets/deepmind/narrativeqa


```
{
    "document": {
        "id": "23jncj2n3534563110",
        "kind": "movie",
        "url": "https://www.imsdb.com/Movie%20Scripts/Name%20of%20Movie.html",
        "file_size": 80473,
        "word_count": 41000,
        "start": "MOVIE screenplay by",
        "end": ". THE END",
        "summary": {
            "text": "Joe Bloggs begins his journey exploring...",
            "tokens": ["Joe", "Bloggs", "begins", "his", "journey", "exploring",...],
            "url": "http://en.wikipedia.org/wiki/Name_of_Movie",
            "title": "Name of Movie (film)"
        },
        "text": "MOVIE screenplay by John Doe\nSCENE 1..."
    },
    "question": {
        "text": "Where does Joe Bloggs live?",
        "tokens": ["Where", "does", "Joe", "Bloggs", "live", "?"],
    },
    "answers": [
        {"text": "At home", "tokens": ["At", "home"]},
        {"text": "His house", "tokens": ["His", "house"]}
    ]
}
```

Methods and Results

- In-context learning vs fine-tuning
- Evaluated for retrieved knowledge vs gold knowledge
- Automatic evaluation

Model	Technique	External Knowledge	Perplexity			
			ODD	KGD	TOD	QA
Llama2$_C$	*In-Context Learning*	No Know.	64.13	35.17	25.15	1442.26
		Retrieved Know.		33.10	24.72	625.08
		Gold Know.		24.40	23.81	298.16
	Fine-Tuning	No Know.	**5.67** $\pm$ **0.01**	7.63 $\pm$ 0.01	**3.06** $\pm$ **0.01**	12.03 $\pm$ 0.06
		Retrieved Know.		6.95 $\pm$ 0.01	3.97 $\pm$ 0.01	5.47 $\pm$ 0.02
		Gold Know.		**4.38** $\pm$ **0.01**	3.12 $\pm$ 0.01	**4.98** $\pm$ **0.01**
Mistral$_I$	*In-Context Learning*	No Know.	14.19	15.31	9.82	91.42
		Retrieved Know.		14.75	9.76	42.58
		Gold Know.		9.81	9.37	16.74
	Fine-Tuning	No Know.	**6.41** $\pm$ **0.01**	8.67 $\pm$ 0.01	**3.56** $\pm$ **0.01**	14.11 $\pm$ 0.01
		Retrieved Know.		7.78 $\pm$ 0.01	3.61 $\pm$ 0.01	5.97 $\pm$ 0.01
		Gold Know.		**5.17** $\pm$ **0.01**	3.58 $\pm$ 0.01	**4.88** $\pm$ **0.01**

Human evaluation

- Checked for contextualization, appropriateness, correctness, validity
- 75 manual annotators using Prolific

Model	Technique	External Knowledge	Contextualization				Appropriateness			Validity
			ODD	KGD	TOD	QA	ODD	KGD	TOD	QA
Llama2$_C$	*In-Context Learning*	No Know.	**85**	70	70	50	**80**	70	60	10
		Retrieved Know.		75	65	70		75	45	35
		Gold Know.		**90**	40	**90**		**85**	45	**80**
	Fine-Tuning	No Know.	45	60	70	15	50	65	60	15
		Retrieved Know.		65	**90**	45		80	80	45
		Gold Know.		80	85	85		65	**85**	75
Mistral$_I$	*In-Context Learning*	No Know.	**90**	80	70	20	**85**	**85**	65	20
		Retrieved Know.		75	65	40		65	60	25
		Gold Know.		90	55	**75**		70	55	**80**
	Fine-Tuning	No Know.	55	90	**85**	25	55	80	80	20
		Retrieved Know.		**95**	**85**	30		**85**	**90**	40
		Gold Know.		80	75	70		65	70	70
Ground-Truth			95	80	95	90	100	85	95	90

Hallucinations

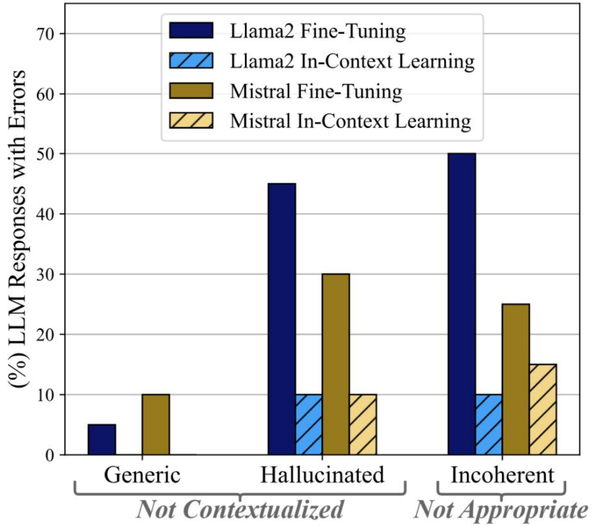


Figure 1: Percentage of LLM responses (y-axis) for each error type (*Not Contextualized* and *Not Appropriate*) and their explanation (Generic, Hallucinated, and Incoherent) (x-axis), for Llama2$_C$ and Mistral$_I$, adapted with In-Context Learning and Fine-Tuning in Open-Domain Dialogues (ODDs).

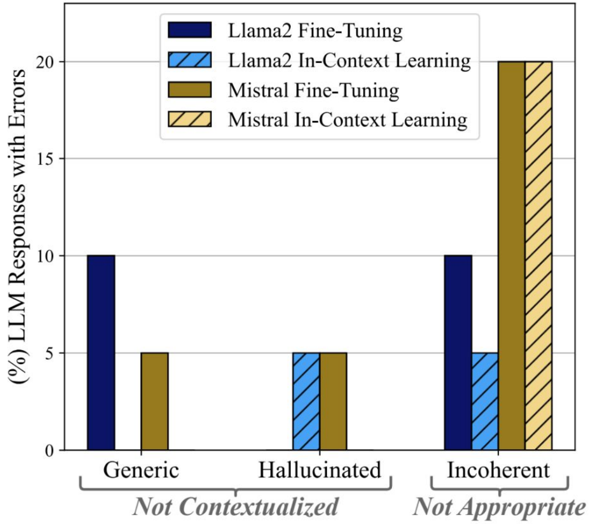


Figure 2: Percentage of LLM responses (y-axis) for each error type (*Not Contextualized* and *Not Appropriate*) and their explanation (Generic, Hallucinated, and Incoherent) (x-axis), for Llama2$_C$ and Mistral$_I$, adapted with In-Context Learning and Fine-Tuning in Knowledge-Grounded Dialogues (KGDs).