RAG vs LONG CONTEXT LLMs

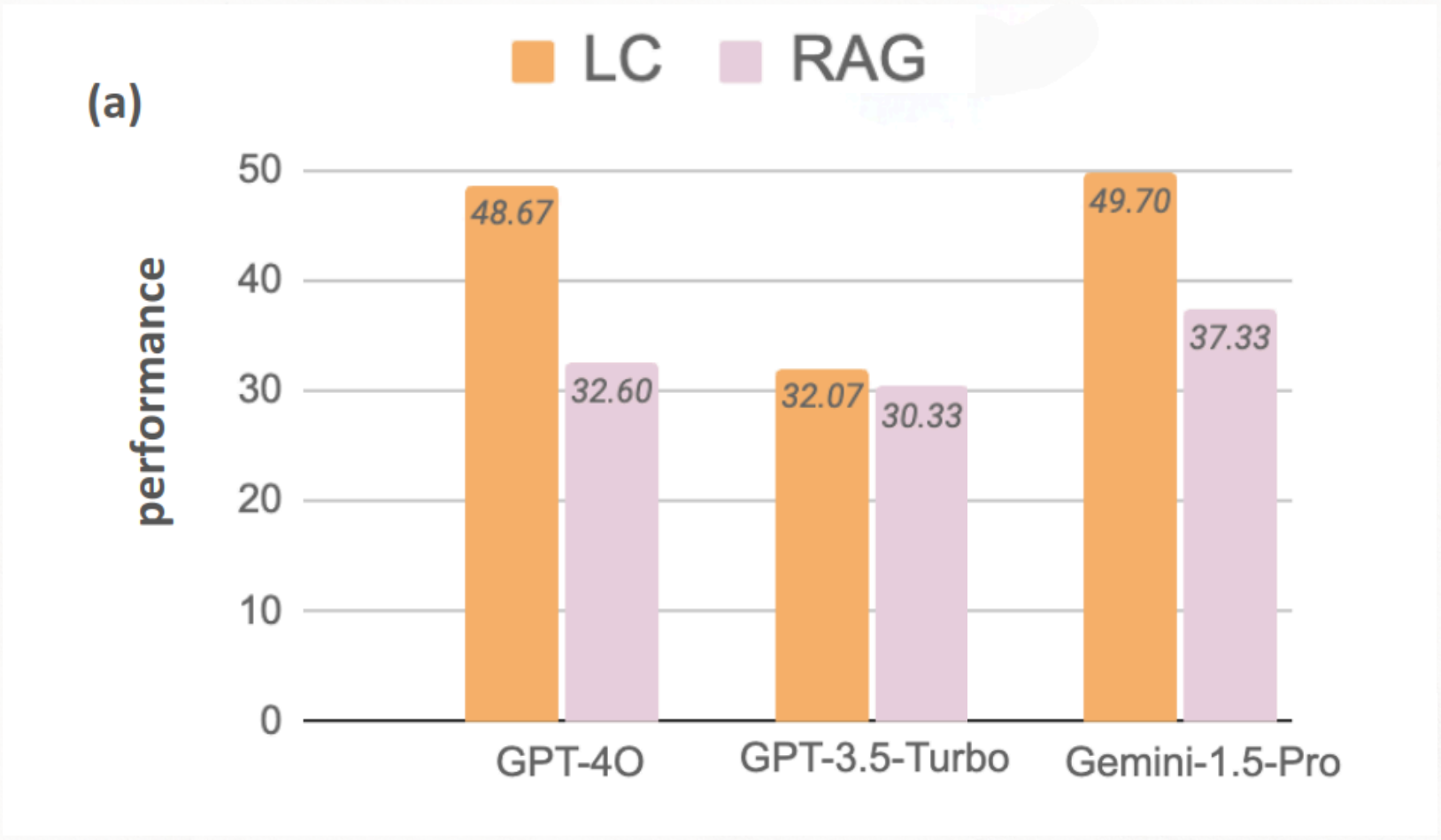
Who is winning the war?

Bhavishya Pandit

ARE RAGs DEAD?

RAG retrieves relevant information based on the query and then prompts an LLM to generate a response in the context of the retrieved information. This approach significantly expands LLM's access to vast amounts of information at a minimal cost.

However, recent LLMs like Gemini and GPT-4 have demonstrated exceptional capabilities in understanding long contexts directly.

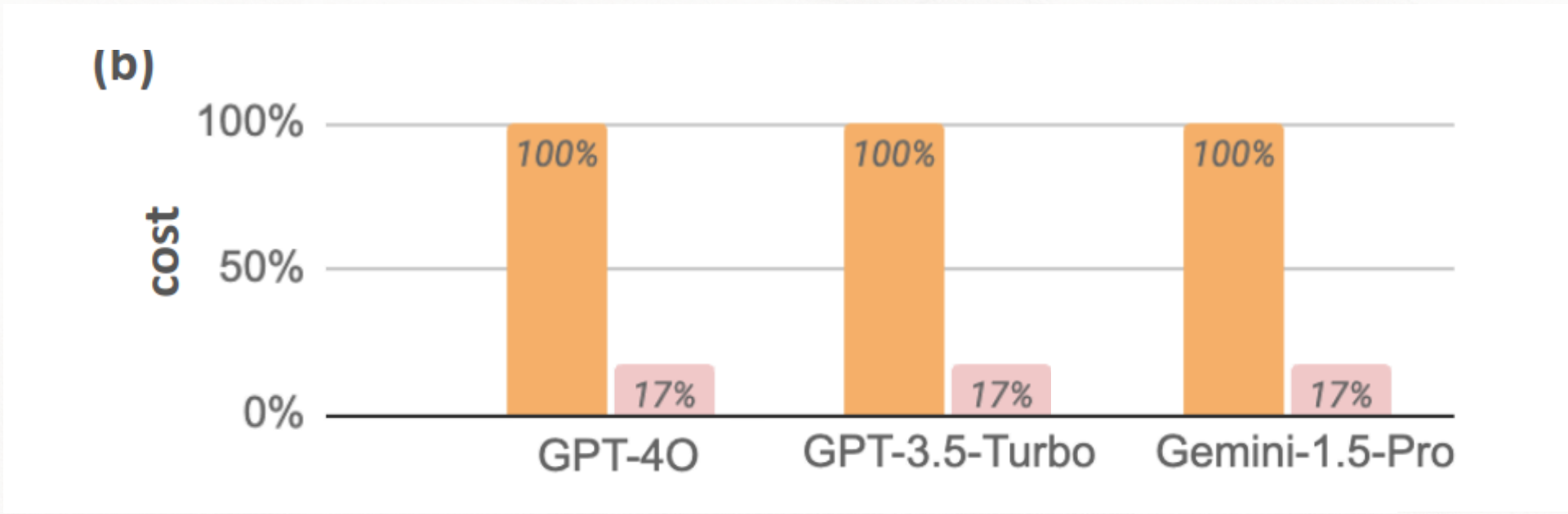


Long Context (LC) LLMs outperform RAGs due to their stronger long-context capabilities producing accurate and reliable results.

Gemini 1.5 can process up to **1 million** tokens. So do we really need RAGs or RAGs have become obselete?

TRADE OFF: COST

Although LLMS with Long Context beat RAGs in performance, the cost of operation seems higher. RAGs can perform the same task with similar proficiency in minimal costs as the pricing depends on the number of tokens used which RAGs minimize significantly.

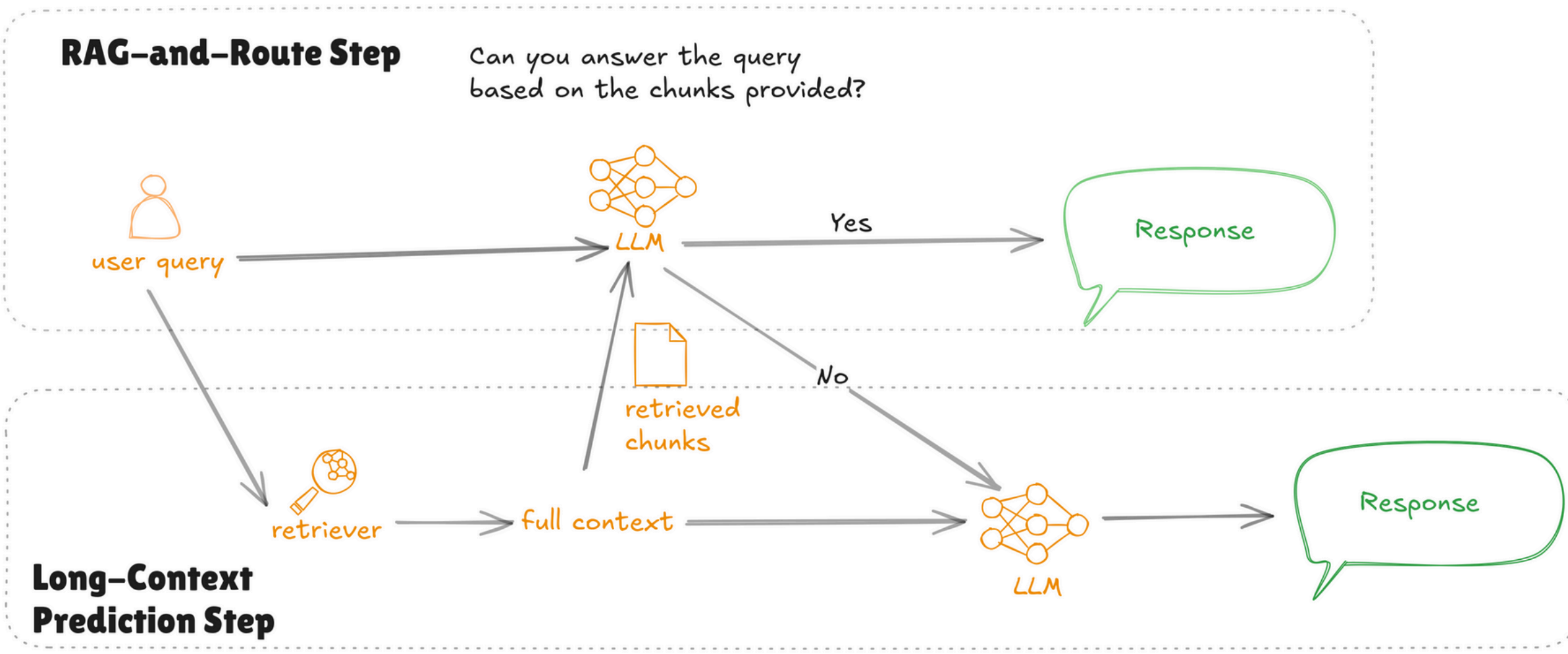


The cost difference between LLMs and RAG for the reference models (models used for research as ref.) is around **83%.** Thus RAGs can't be made obsolete.

We need a technique that uses the fusion of these two that makes the model fast and cost efficient simultaneously.

SELF-ROUTE SAVES US

We have established that Long Context LLMs are proficiently better than RAGs while RAGs are cost-efficient but they sometimes lack context and can give incorrect responses. Scientists at **Google Deepmind** and **UoM** have deduced a new method combining the powers of the two named **"SELF-ROUTE"**.



SELF-ROUTE is a two-step process-

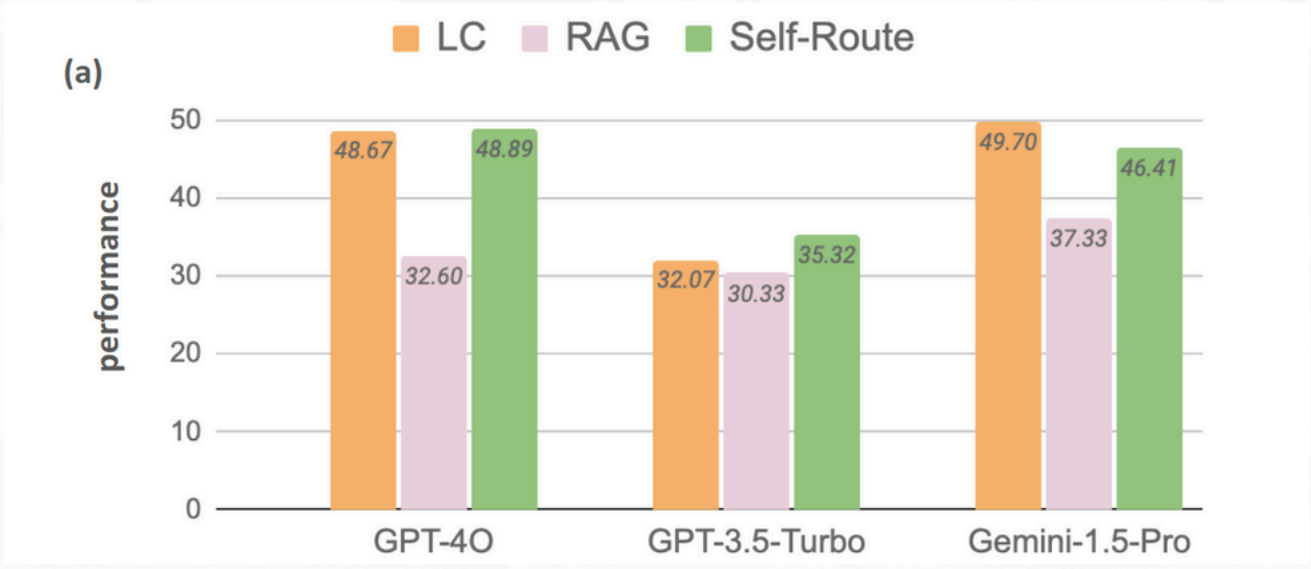
1. **RAG-and-Route Step** – The model processes the query along with the retrieved chunks and assesses whether the query can be addressed using the retrieved information. If it can, the model generates a response. If not, it moves to the next step.

2. **Long-Context Prediction Step** – The entire context is then provided to the model, which generates the final response based on this extended input.

KEY TAKEAWAYS

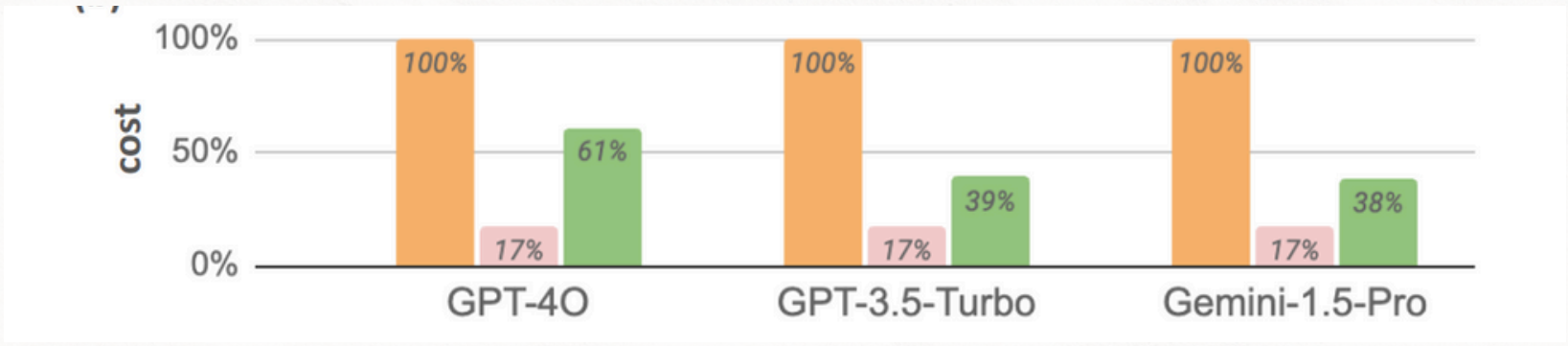
- **Use RAG (Retrieval-Augmented Generation) when:**

1. There is a need for lower computational costs.
2. The input significantly exceeds the model's context window size, making RAG more effective for handling large amounts of data.
3. Performance is important, but it is not the top priority in long-context understanding.



- **Use LC (Long-Context LLMs) when:**

1. Superior performance in handling long-context input is required.
2. A task demands accuracy and deeper understanding of extended sequences.
3. Sufficient resources are available to support higher computational costs.



- **Use the proposed "SELF-ROUTE" method when:**

1. The method dynamically inquires whether a user query can be answered using retrieved chunks; if not, it resorts to using an LLM.
2. There is a goal to achieve comparable performance to LC while benefiting from reduced computational overhead.

Follow to stay updated on AI/ML

SAVE

LIKE

SHARE