Part G: LLM Limitations

While LLMs (Large Language Models) are extremely powerful, they do have several limitations that developers and users need to understand to maximize their effectiveness.

1. Stateless API

LLMs are stateless, meaning they do not remember previous interactions. Each prompt is independent, which can lead to the loss of context over multiple exchanges.

Solution:

- Use external state management or memory systems to retain context across interactions.
- Split conversations or sessions into smaller, self-contained prompts.

2. Not Trained on Your Data

LLMs are not trained on your specific data, so they may lack detailed understanding of your domain, system, or application.

Solution:

- Provide context-specific examples in prompts.
- Use fine-tuning or domain-specific models if possible to cater to particular needs.

3. Limited Size of Data You Can Send

LLMs generally have a maximum token limit for each prompt, which restricts the amount of data that can be processed.

Solution:

- Pre-process and chunk the data into smaller sections.
- Use summarization tools to condense the data before sending it to the LLM.

4. Prone to Hallucinations

LLMs can sometimes generate outputs that seem plausible but are actually incorrect or fabricated.

Solution:

- Implement a post-processing or verification step to ensure the accuracy of the model's responses.

- Use domain-specific tools that can cross-check facts or corroborate information.

5. Not Aware of Your APIs

LLMs are unaware of your APIs or any custom tools unless explicitly mentioned.

Solution:

- Use prompts to provide context about your APIs.

- Integrate external systems that can help bridge the gap between LLMs and real-time data, such as API wrappers or data pipelines.

6. Not Aware of Real-Time Data

LLMs do not have access to real-time information, limiting their usefulness for live updates or real-time data analysis.

Solution:

- Provide real-time data as input in the prompt or integrate with real-time data sources to keep outputs up-to-date.

- Use hybrid models that combine LLMs with real-time systems for live data processing.