Lab 13 Part G

1. Stateless API
Since LLMs don’t retain memory between requests, they can’t remember previous conversations unless I include the entire context every time. To deal with this, I can manage state on my side by keeping a running log of the conversation or important variables in my app and resending that context in each prompt. It’s a bit of extra work, but it keeps the interaction coherent.

2. Not trained on your data
Out of the box, an LLM doesn’t know anything specific about my business, project, or internal systems. To solve this, I can use techniques like retrieval-augmented generation (RAG), where I pass relevant documents or knowledge snippets along with my prompt. This lets the model "look up" what it needs to know in real time without retraining it.

3. Limited size of data you can send
There’s a cap on how much information I can include in a single request. So, I try to be strategic summarizing long documents, chunking information, or ranking the most relevant pieces to send. Tools like embedding-based search can help choose the right info dynamically.

4. Prone to hallucinations
LLMs can sometimes make up facts confidently, which is risky. To reduce this, I try to always verify model responses against trusted sources or only allow it to generate answers based on a fixed set of inputs or documents I provide. Adding citations or references can also help track where the information is coming from.

5. Not aware of my APIs
LLMs don’t automatically know how my custom APIs work, which means I need to guide them. One way to fix this is by providing clear documentation or even examples of API calls in the prompt. Better yet, with function calling support in newer models, I can define structured functions and let the LLM decide when to call them making the integration smoother.

6. Not aware of real-time data
LLMs can’t access live or up-to-the-minute data. If I need real-time info (like stock prices or weather), I’d build a system where the LLM asks a function or external service to fetch that info when needed. It becomes more of a smart coordinator than a source of truth.