Gen Ai Essentials Bootcamp

Pre-week Exercise:
Diagrammatically Explain Gen AI Architecture

By: MKGenAIstudent

INTRODUCTION

- Business Goal:
  - We have been tasked to create an architectural diagram that serves as a teaching aid to help stakeholders understand their key components of GenAI workloads. The outcome is to help let stakeholders visualize possible technical paths, technical uncertainty when adopting GenAI.

- Technical Considerations:
  - Out of the possible diagrams that we can use (Conceptual, Logical and Physical), I have chosen the Conceptual option to communicate the business solution to key stakeholders about the GenAI architecture at a high level.

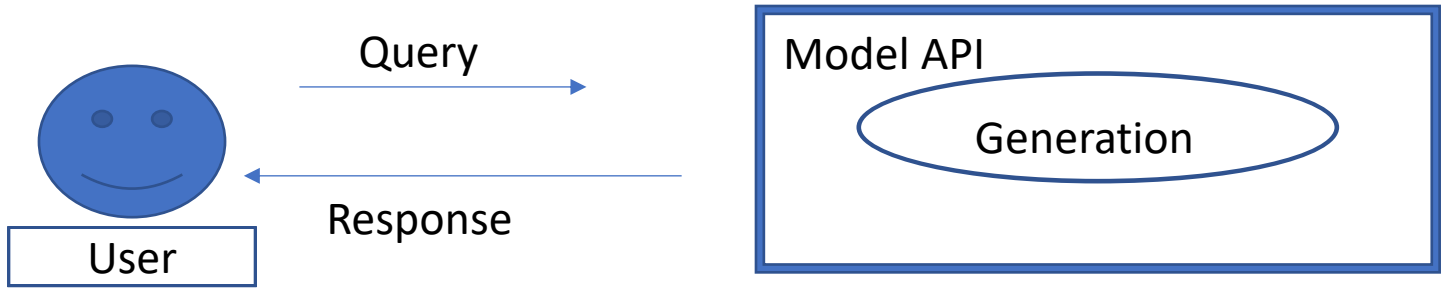
- Data Strategy:
  - I have included a possible solution to address the concerns around Data Privacy & Security through the use of input/output guardrails

- Model Selection and Development:
  - Included in the proposed solution is the size of the model (or size of the ”brain”) in no.of parameters (7bn)
  - Included in the proposed solution is a prompt cache so as to help optimise model performance and efficiency

Study Activity (GenAI Architecture Explained)

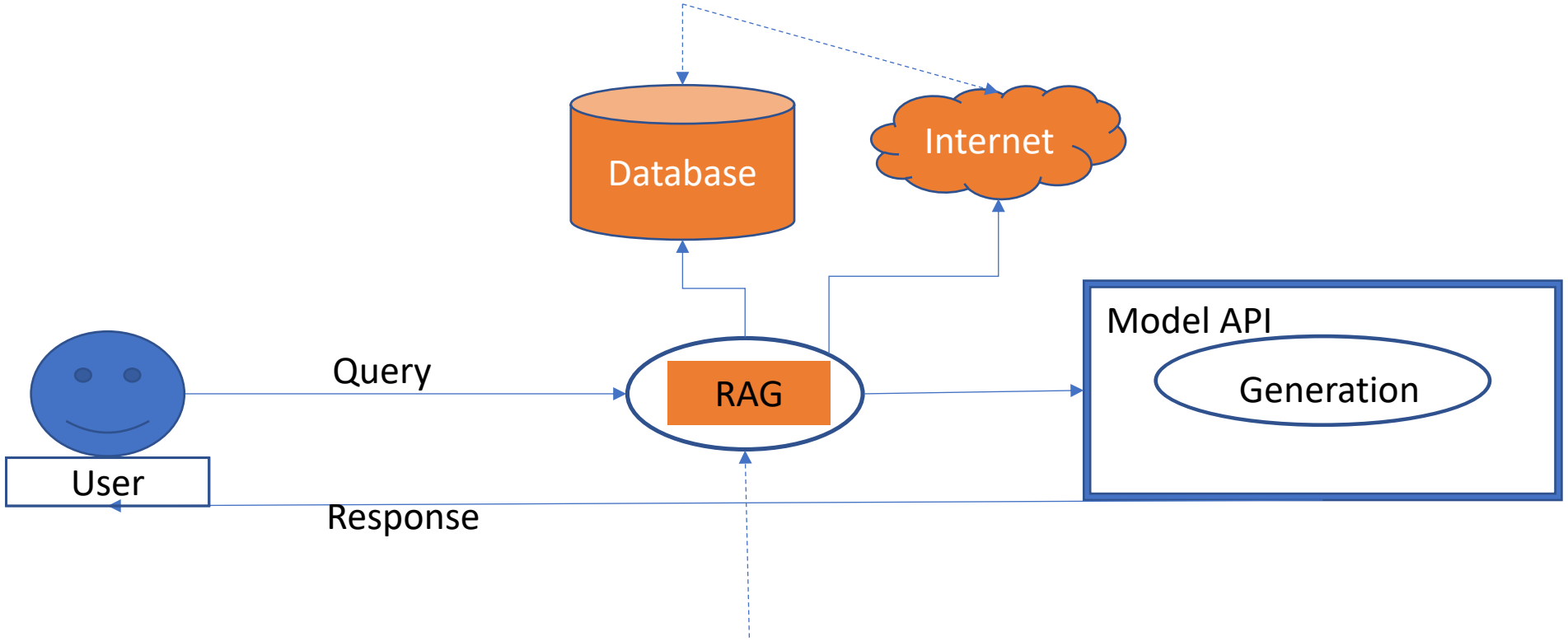
To begin with, let’s start with the purpose of the LLM model.

We have a user, who sends a query into the Model. The model generates the response and returns it to the user:



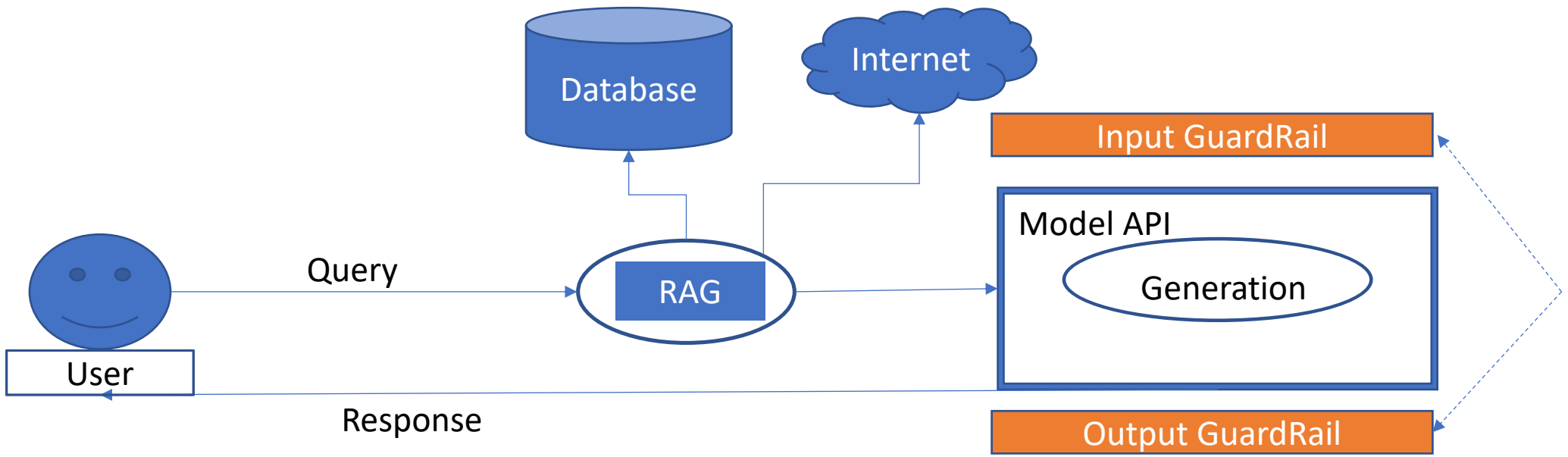
Study Activity (GenAI Architecture Explained)

In order for the model to generate the answer so as to return the response to the user, the model will trawl through the database and the internet and build the response to the query:



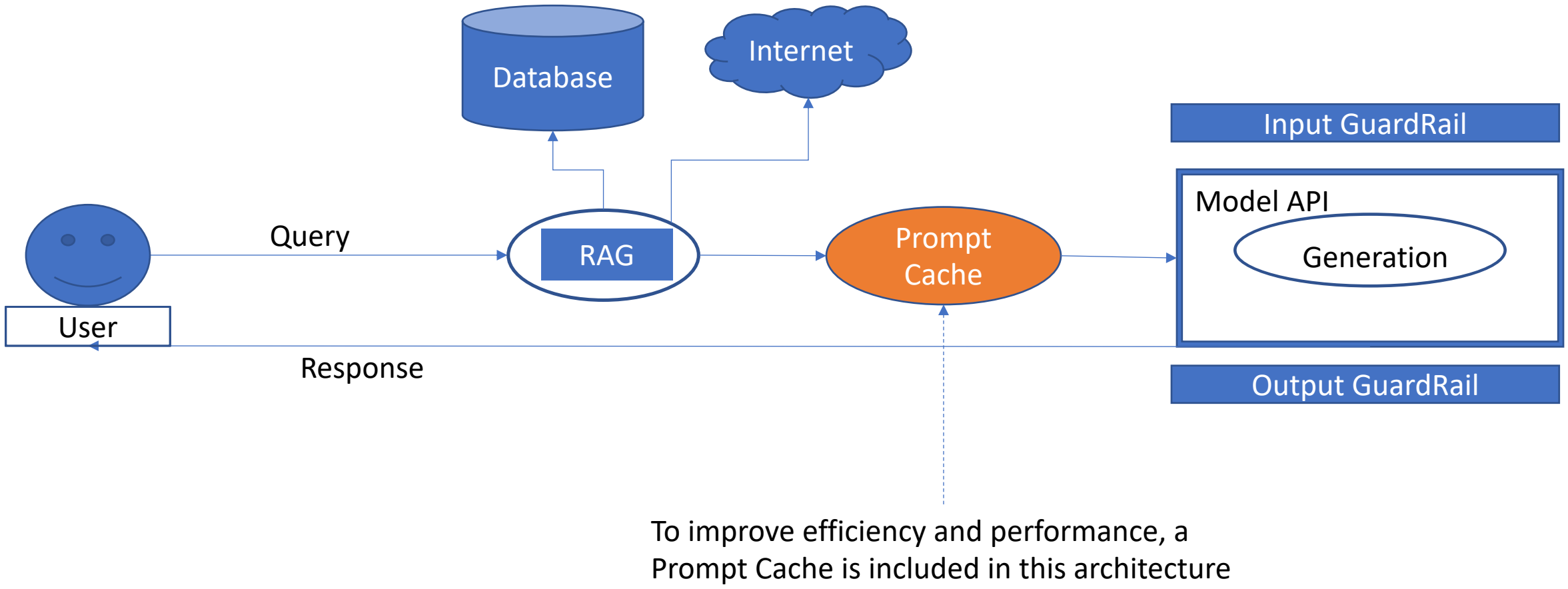
Retrieval Augmentation Generation (RAG) allows us to fetch data from an external source such as a database or the internet.

Study Activity (GenAI Architecture Explained)

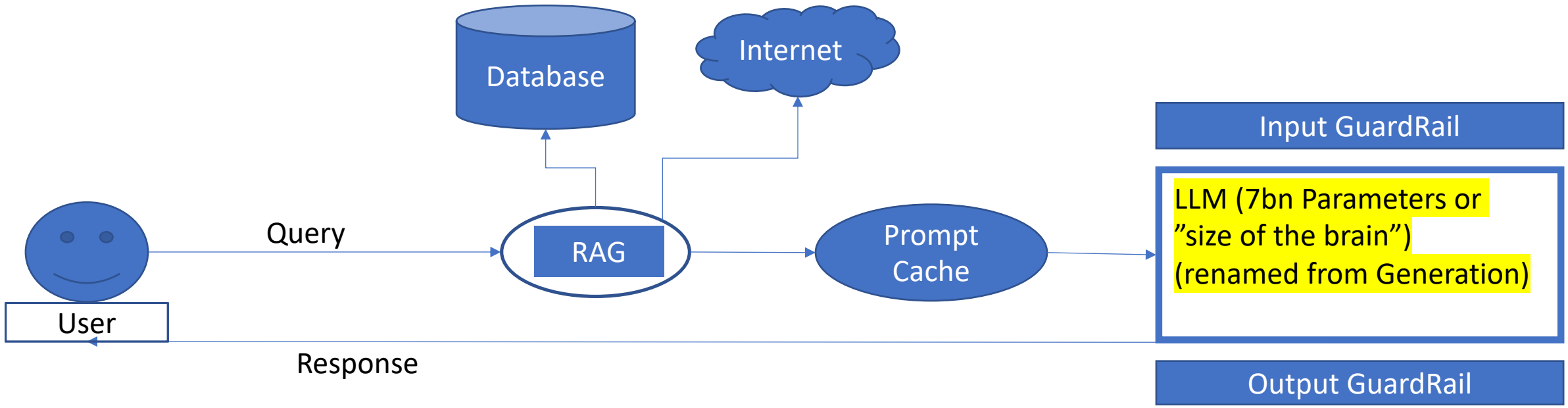


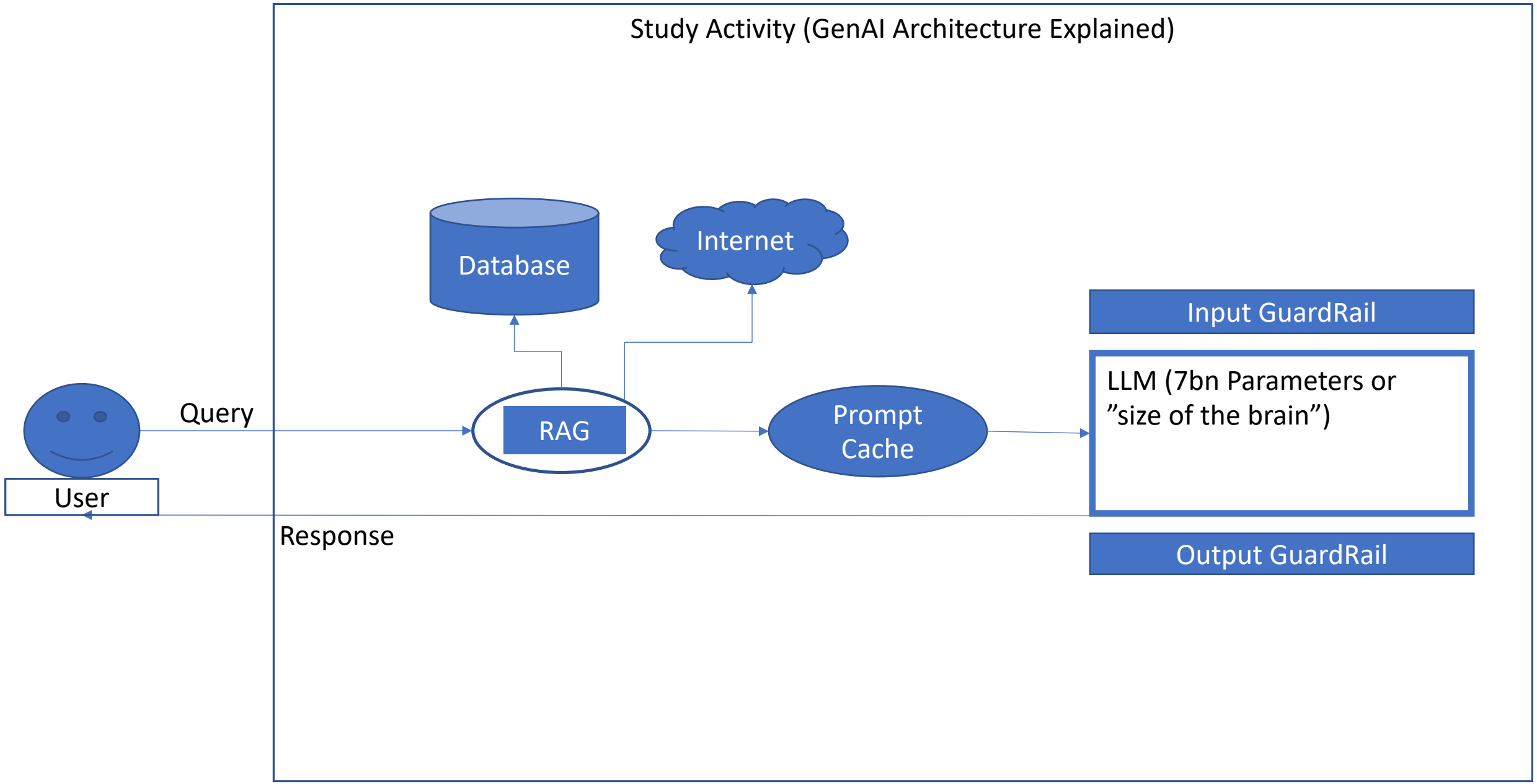
For security, we've included GuardRails, to help protect things like sensitive data or Personally Identifiable Information

Study Activity (GenAI Architecture Explained)



Study Activity (GenAI Architecture Explained)




Study Activity (GenAI Architecture Explained)
Database
Internet
Input GuardRail
LLM (7bn Parameters or ”size of the brain”)
Query
RAG
Prompt Cache
User
Response
Output GuardRail