PROJECT REPORT: ReAct-Based PC Hardware & System Building Assistant

1. Executive Summary

This project aims to develop an autonomous AI Agent specialized in PC hardware compatibility, system building recommendations, and technical troubleshooting. To address the common issue of "hallucination" (fabricating information) in Large Language Models (LLMs) regarding technical specifications (e.g., TDP values, socket types), the project integrates **ReAct (Reasoning + Acting)** architecture with **RAG (Retrieval-Augmented Generation)** technology.

Benchmark tests indicate that the open-source **Llama-3.3 70B** model demonstrated competitive performance (~88% accuracy) against the industry standard **GPT-4o** (~96% accuracy). Consequently, Llama-3.3 was selected as the **"Optimal Production Model"** due to its significant advantages in cost-efficiency and processing speed via the Groq infrastructure.

---

2. Problem Statement & Objectives

2.1. Problem

While Large Language Models (LLMs) excel in general knowledge, they are prone to generating inaccurate data when handling static, precision-critical technical information (e.g., "What is the TDP of the i9-14900K?"). Furthermore, they often struggle to manage complex compatibility rules between hardware components (e.g., DDR4 vs. DDR5, PCIe bottlenecks) solely through parametric memory.

2.2. Objectives

- To develop a system that analyzes user technical queries.
- To retrieve accurate technical data from an external knowledge base (RAG).
- To logically query compatibility between components (Reasoning).
- To provide evidence-based, reliable recommendations to the user.

3. Methodology & Architecture

The project is built upon the **ReAct** paradigm proposed by Yao et al. (2022).

3.1. ReAct Loop (Thought-Action-Observation)

Instead of responding immediately, the agent follows a strict cognitive loop:

1. **Thought:** "The user is asking for PSU requirements for an RTX 4090. I need to find the card's TDP first."
2. **Action:** Calls the hardware_search("RTX 4090") tool.
3. **Observation:** Retrieves "450W TDP" from the database.
4. **Final Answer:** "For an RTX 4090, a minimum of 850W (preferably 1000W) PSU is recommended."

3.2. RAG Engine (SimpleRAG)

A lightweight, Python-based vector search engine was developed for this project, eliminating the need for heavy external vector databases (like Pinecone or Milvus).

- **Dataset:** hardware_database_FULL.json (Technical specs for CPU, GPU, RAM, Motherboards).
- **Algorithm:** Uses **Cosine Similarity** to match user queries with the most relevant hardware components.

3.3. Models Used

- **Llama-3.3 70B (via Groq):** The primary agent model. Utilizing the Groq LPU (Language Processing Unit) infrastructure, it ensures the "low latency" required for ReAct loops while maintaining high reasoning capabilities.
- **GPT-4o (via OpenAI):** Used as the "Ground Truth" for benchmarking purposes to validate the system's accuracy.

4. Experimental Setup & Benchmark

To evaluate the system's success, a rigorous **50-question test set** was designed across 5 distinct categories:

1. **General Knowledge:** Basic hardware terminology.
2. **Logic & Compatibility (Reasoning):** Physical compatibility between parts (e.g., "Does DDR4 RAM fit in a DDR5 slot?").
3. **Scenario Analysis:** Recommendations based on user profiles (e.g., "CPU vs. GPU priority for a Streamer").
4. **Legacy Hardware:** Questions about obsolete tech to test hallucination risks (e.g., "Are Voodoo graphics cards still manufactured?").
5. **Out-of-Domain:** Irrelevant questions to test agent boundaries (e.g., "How to cook pasta?").

---

5. Results & Performance Analysis

The performance metrics obtained from the tests are as follows:

Metric	Llama-3.3 70B (Groq)	GPT-4o (OpenAI)
Technical Accuracy	88%	**96%**
Hallucination Rate	<5%	**~0%**
Avg. Response Time	~6 seconds	**~3.3 seconds**
Cost	**Free / Low**	High
Loop Error Rate	Moderate	Low

5.1. Analysis

- **GPT-4o:** Achieved the highest accuracy (96%) as expected. It performed flawlessly in complex reasoning tasks and legacy hardware "trap" questions.
- **Llama-3.3 70B:** Achieved a highly competitive success rate of 88%. It yielded results very close to GPT-4o, particularly in **logical deduction** and **scenario** questions. The errors observed were primarily due to time-outs resulting from extended Agent Loops.

6. Discussion & Conclusion

In this project, **Llama-3.3 70B** was selected as the **"Optimal Production Model"** over the highest-scoring model (GPT-4o). The key reasons for this decision are:

1. **Cost-Efficiency:** API costs for GPT-4o are prohibitive for scalable or student projects. Llama-3.3 (via Groq) offers comparable performance at a significantly lower (or free) cost.
2. **Sufficiency:** An 88% accuracy rate for a PC building assistant, when supported by a RAG system, is sufficient for commercial or community deployment.
3. **Open Source Alignment:** The project aligns with open-source principles by utilizing an accessible model with open weights.

In conclusion, the developed ReAct-based agent successfully combines static database querying with the reasoning capabilities of LLMs, demonstrating the potential to automate technical support processes effectively

7. Benchmark Comparison

In this section, the performance of OpenAI GPT-4o and Groq Llama 3.3 70B was evaluated on a rigorous 50-question test set

7.1. General Performance Summary

Metric	OpenAI GPT-4o	Groq Llama 3.3 70B
Total Questions	50	50
Correct Answers	48	44
Incorrect Answers (Factual Errors)	0	2
No Response (Loop/Timeout Errors)	2	4
Success Rate	96%	88%

7.2. Detailed Model Analysis

🟢 Model A: OpenAI GPT-4o (The Winner)

GPT-4o maintained its status as the "Gold Standard" throughout the test. It produced zero hallucinations or factual errors. However, the agent got stuck in a reasoning loop on 2 questions where the "Chain of Thought" became too complex or indecisive.

- Correct Count: 48
- Errors:
    1. Question 31 (Legacy): *"Is the Nvidia GTX 1080 Ti still sufficient for modern games?"*
        - Error Type: Loop Limit Exceeded.
        - Root Cause: The model likely fell into "analysis paralysis" trying to define "sufficient" (1080p vs 4K? Ultra vs Low settings?), causing the agent to exceed the maximum allowed turns without reaching a final conclusion.
    2. Question 42 (Out-of-Domain): *"What is the difference between PS5 Slim and Xbox Series X?"*
        - Error Type: Loop Limit Exceeded.
        - Root Cause: While attempting to compare technical specs of two different devices step-by-step, the agent exceeded the Max Turns limit before synthesizing the final answer.

🟠 Model B: Groq Llama 3.3 70B (The Optimal Choice)

Despite being an open-source model, Llama 3.3 achieved a highly competitive 88% success rate. However, compared to GPT-4o, it showed a slight tendency towards context loss and factual confusion in specific edge cases.

- Correct Count: 44
- Errors:
    1. Question 19 (Logic): *"If I install 3 RAM sticks, will Dual Channel work?"*
        - Error Type: Loop Limit Exceeded.
        - Root Cause: The model struggled with the technical nuance (Flex Mode vs. Single Channel) and entered a reasoning loop.
    2. Question 28 (Scenario): *"Does Server RAM work in a home PC?"*
        - Error Type: Loop Limit Exceeded.
    3. Question 29 (Scenario - CRITICAL ERROR): *"Is TPM 2.0 required for Windows 11?"*
        - Model Response: *"The TDP of the RTX 4060 is 115W and the recommended PSU is 550W."*
        - Error Type: Hallucination / Context Loss.
        - Root Cause: This was the most significant error. The model lost the context of the question (Windows 11) and retrieved/generated completely irrelevant hardware specs (RTX 4060) from its memory or RAG context.
    4. Question 30 (Scenario): *"My GPU is overheating..."*
        - Error Type: Loop Limit Exceeded.
    5. Question 31 (Legacy): *"Is the GTX 1080 Ti sufficient?"*
        - Error Type: Loop Limit Exceeded. (Same failure point as GPT-4o).
    6. Question 43 (Out-of-Domain - Factual Error): *"Samsung Galaxy S24 Ultra processor..."*
        - Model Response: *"...available with Snapdragon 8 Gen 2 or Exynos 2400 depending on the region."*
        - Error Type: Factual Error.
        - Root Cause: While the base S24 and S24+ use Exynos in some regions, the S24 Ultra exclusively uses Snapdragon globally. The model conflated the Ultra model with the rest of the series.

7.3. Comparative Evaluation

1. Reliability: GPT-4o demonstrated superior reliability by choosing to "time out" (loop error) rather than providing incorrect information. Llama 3.3 lost points on reliability due to one factual error (S24 Ultra) and one severe hallucination (Win11/RTX4060).
2. Stability: Both models struggled with subjective questions requiring nuanced judgment (e.g., the viability of the GTX 1080 Ti), leading to infinite loops. This suggests that Agentic workflows need better "stop conditions" for subjective queries.
3. Conclusion: Llama 3.3 70B proved to be production-ready with an 88% accuracy rate. Its errors (mostly loops and rare hallucinations) are manageable with improved Prompt Engineering. Considering the significant cost and speed advantages via Groq, it remains the Optimal Model for this project