Week 4 Report

Team 1

Platforms and LLMs We Tested

AnythingLLM

- Gemini 2.0, Llama 3.2 Vision, Deepseek R1 via Ollama

LM Studio

- Llama 3.2 Vision, Deepseek

chatgpt.com 4.o and o1

claude.ai 3.5 sonnet

perplexity.ai

- GPT-4o, Gemini 2.0 Flash, Claude 3.5 Sonet, and Grok-2

We Asked Questions Like…

“Which paper…

- was authored by <author name>?
- discussed <keyword>?
- was advised by <advisor name>?
- has the word “<specific word>" in the title?
- is tagged with the keyword "<specific keyword>"?
- was published in <year>?

Standings After Experimentation

1. **GPT 4o on chatgpt.com (Best)**

   Total Score: 25/25

2. **Claude 3.5 - sonnet on claude.ai (Second best)**

   Total Score: 20/25

3. **Gemini 2.0 flash on AnythingLLM or LM Studio (Third Best)**

   Total Score: 15/25

AnythingLLM

Gemini 2.0

- Strengths: concise responses
- Problems: inaccurate responses for follow-up questions

Llama 3.2 Vision

- Strengths: straight to the point
- Weaknesses: not very accurate

Deepseek R1 via Ollama

- Strengths:
  - Shows how it thinks
- Problems:
  - Slow
  - Hard to escape context
  - Can give incorrect answers for the document is using

LM Studio

Llama 3.2 Vision

- Strengths: straight to the point
- Weaknesses: not very accurate

Deepseek R1

- Strengths: shows reasoning, detailed responses
- Problems: slow, fairly inaccurate

Perplexity.ai

Each model on Perplexity gave responses that were:

- Detailed (often too lengthy)
- Fairly accurate
- Slow to generate

chatgpt.com

GPT 4o (overall score of 25 / 25)

- Strengths: accurate and to-the-point responses.
- Weaknesses: pricey, limited number of prompts.

GPT o1 (overall score of 25/ 25) - most advanced model

- Similar to 4o, but gave longer responses.
- Took longer to output responses.
- Shows reasoning

Gemini

2.0 flash on AnythingLLM via API key

- Strengths: short responses.
- Weaknesses: accuracy.

2.0 flash on perplexity.ai

- Strengths: longer, relevant responses.
- Problems: accuracy.

DeepSeek R1

AnythingLLM and **LM Studio**:

- Strengths: clarity in reasoning.
- Weaknesses: accuracy, refusing to shift a context.

deepseek.com DeepSeek R1:

- Fast, accurate, but responses could be a little lengthy.

Extra Processing & Potential Next Steps

Methodology: Created a script (link here) in order to talk to the Google Gemini API

Basic

Upload the papers individually and ask questions about each of the papers

Vectorized

Upload the papers individually and ask questions about each of them

Contextualized

Upload all the papers to create a context window and ask questions about all of the papers.

Vectorized Contextual Window

Vectorize the papers, create context window across different papers and ask overall questions

Findings

Basic Question and Answer

Depth of Understanding

Identifies main findings, results, limitations, etc. Answers are mostly surface level.

Contextual Understanding

Connects ideas across different section. Understands relationships between methods, results, and their significance

Quality of Response

Organizes information logically in it’s response & presents information in a coherent hierarchy. Sometimes lacks depth.

Findings

Vectorized Question and Answer

Depth of Understanding

Much more detailed technical explanations. (Talked in depth about paper 2 concepts).

Contextual Understanding

Stronger linkage between the different sections of the paper. Clearer relationships between methods & their purpose.

Quality of Response

More precise and technically accurate. Better organized responses with a clearer structure.

Findings

Contextualized Question and Answer

Depth of Understanding

Very deep understanding when comparing the different papers. Stronger grasp of theory and practical implications of these papers.

Contextual Understanding

Exceptional ability to draw inferences across the 3 papers. Understands how first paper theory lays the framework for the second and third papers.

Quality of Response

Clear use of examples to support points. Strong analytical answers - maintains balance between all 3 papers and individual concepts.

Findings

Vectorized Contextual Question and Answer

Depth of Understanding

Shows significantly enhanced understanding of how papers interconnect and build upon each other.

Contextual Understanding

Superior ability to trace progression across papers. Better at explaining relationships between theory and practical notions.

Quality of Response

Better organization with clear hierarchical presentation. Stronger supporting examples and evidence

Vectorized Contextual Q&A provides the most sophisticated and nuanced understanding of relationships between papers.