Ex.No:03

Date:18.09.24

Comparative Evaluation of Prompting Tools Across AI Platforms in a Specific Use Case

Aim :

To compare the performance, user experience, and response quality of prompting tools on various AI platforms within a specific use case, such as text summarization or technical question-answering. This experiment will assess each platform's ability to generate accurate, high-quality, and user-friendly responses to understand which platform best suits the selected use case.

Introduction :

Generative AI platforms, such as OpenAI's GPT, Google's Bard, and others, offer prompting tools that users can leverage for tasks like summarization, question-answering, and more. These platforms vary in their architecture, response style, user interface, and overall performance, making it essential to compare them within specific use cases. By evaluating how each platform handles a selected task, this experiment aims to determine which tool delivers the best user experience and response quality for that use case.

This experiment will compare platforms on three primary criteria:

• **Performance:** How effectively the platform completes the task in terms of accuracy, coherence, and depth.

• **User Experience (UX):** The intuitiveness, ease of use, and feedback mechanisms of the platform's interface.

• **Response Quality:**

The relevance, clarity, and completeness of the platform's generated response.

Theory 1.

AI Platform Comparison and Use Case Specificity:

o **Prompting Tools:** These platforms have various built-in tools to help users interact with the model, including interface elements, feedback options, and response customization features

o **Use Case Specificity:** Different tasks require varying levels of model understanding and contextualization.

For example, summarization requires cohesive reduction of information, while technical Q&A demands high accuracy and precision.

Platforms may vary in strengths depending on the use case.

2. **Evaluation Metrics:** o Performance: Measures how well each platform completes the task based on the clarity, accuracy, and depth of the response. o User Experience (UX):

Examines ease of use, clarity of interface, user control over prompts, and responsiveness of the platform.

o **Response Quality:** Evaluates the relevance and usefulness of responses, including coherence, fluency, and level of detail.

3. **Common Platforms for Comparison:**

o **OpenAI's GPT (ChatGPT):**

Known for high-quality natural language understanding and generation capabilities.

o **Google Bard:** Provides powerful search-integrated responses, helpful for general information and exploratory tasks. o Microsoft Bing Chat (powered by GPT-4): Merges search functionality with generative AI to produce contextually rich responses

o **Anthropic's Claude:** Focused on safety and interpretability, particularly in knowledge-based or constrained applications.

Procedure

1. **Define the Use Case:**

o Choose a specific use case where generative AI models are commonly applied, such as:

▪ **Technical Question-Answering:**

Answering domain-specific questions accurately. ▪ Summarization: Reducing lengthy texts into concise summaries.

2. Select Prompts and Tasks: o Design two or three prompts specific to the selected use case.
Examples:

▪ **For Technical Q&A:** "Explain the concept of quantum entanglement in simple terms."

▪ **For Summarization:** "Summarize the following article on climate change impact."

3. **Engage Each Platform Using the Prompts:**

o Test each platform (e.g., GPT, Bard, Bing Chat, Claude) with identical prompts and record the responses.

Repeat each prompt to evaluate consistency in responses.

4. **Evaluate Responses Based on Performance, UX, and Quality:**

o Use predefined criteria to assess each response:

▪ **Performance:** Assess accuracy, relevance, and coherence.

▪ **User Experience:** Document ease of interaction, customization options, and interface features.

▪ **Response Quality:** Evaluate the clarity, depth, and fluency of the response.

5. **Document Observations:**

o Record findings in a comparison table for clarity.

Platform	Use Case	Performance	User Experience	Response Quality
OpenAI GPT	Summarization	High accuracy, clear summary	Intuitive interface, responsive	Clear, relevant, retains key details
	Technical Q&A	Detailed, often nuanced	Simple and user-friendly	Highly accurate, detailed with examples
Google Bard	Summarization	Informative but less concise	Flexible search integration	Adequate relevance, slightly verbose
	Technical Q&A	Correct but generalized	Integration with web search	Accurate but less in-depth
Microsoft Bing Chat	Summarization	High accuracy, concise	Search-enabled interface	Relevant, well-balanced summary
	Technical Q&A	Detailed, context-aware	Enhanced by web search	Accurate, useful for latest information
Claude	Summarization	Balanced, slightly verbose	Safe, clear guidance	Clear but can lack depth in complex topics
	Technical Q&A	Safe and well-reasoned	Simple but thorough	Consistently relevant, conservative in depth

Discussion:

1. **Performance and Response Quality:**
   o **OpenAI GPT**: produced the highest quality responses for both technical Q&A and summarization, displaying a nuanced understanding and providing detailed answers in technical contexts and precise, context-aware summaries.
   o **Google Bard**: excelled in exploratory and general information tasks, benefiting from its web search integration, although its responses tended to be more generalized. It performed better in summarization when compared to technical Q&A, where it occasionally lacked in-depth detail.
   o **Microsoft Bing Chat:** demonstrated a balanced performance, using web search integration effectively to provide context-rich responses. It was particularly strong in technical Q&A tasks, where real-time information was a factor, offering accurate, relevant answers with up-to-date details.
   o **Claude**: produced safe, clear responses but showed conservative depth in technical contexts, likely prioritizing interpretability and cautiousness over extensive detail. It was consistent in output, though occasionally verbose in summarization.
2. **User Experience (UX):** o OpenAI GPT and Claude both offered intuitive, user-friendly interfaces that were easy to navigate for all prompt types. OpenAI GPT's simplicity and responsiveness were especially advantageous for task-oriented use cases.
   o **Google Bard and Microsoft Bing Chat:** provided search-enhanced experiences, making them ideal for exploratory tasks where immediate access to real-time data is valuable. Microsoft Bing Chat's integration with real-time search results gave it an edge in the

technical Q&A use case, where up-to-date accuracy is often critical. 3. Implications for Use Cases:

o For technical Q&A tasks, OpenAI GPT and Microsoft Bing Chat outperformed other platforms, especially when depth, accuracy, and up-to-date relevance were critical.

o In summarization tasks, OpenAI GPT excelled due to its coherence and ability to distill key points concisely, though Claude also performed well when clarity and safety were prioritized.

The findings suggest that when use cases require real-time data (e.g., live event summaries or breaking news questions), platforms with search integration like Microsoft Bing Chat and Google Bard are advantageous.

Conclusion

This experiment demonstrates the importance of platform selection based on specific use cases, as each AI platform exhibits unique strengths.

OpenAI GPT consistently provided high-quality, detailed responses across all tested scenarios, making it highly suited for tasks requiring nuanced and comprehensive answers. Microsoft Bing Chat proved ideal for use cases where up-to-date information is critical due to its web search integration, while Google Bard excelled in exploratory tasks but showed limitations in providing depth for technical questions.

Claude offered safe, interpretable responses and performed well in general summarization but was less detailed in technical contexts.

These results emphasize the importance of matching the platform's capabilities with the demands of specific tasks, particularly in specialized applications like technical Q&A.

This study suggests that by aligning platform choice with task requirements, users can maximize response quality and user experience, achieving optimal results in professional or information-sensitive applications