EXP 03 : Evaluation of 2024 Prompting Tools Across Diverse AI Platforms: ChatGPT, Claude, Bard, Cohere Command, and Meta

NAME : SANJAY S
REG NO : 212222230132

1. Introduction

Artificial Intelligence (AI) platforms have evolved significantly, enabling users to perform a wide range of tasks with high efficiency. The ability to generate accurate, clear, and insightful responses across different use cases has become a key differentiator among AI platforms. This report presents a comprehensive evaluation of five major AI tools from 2024—ChatGPT, Claude (Anthropic), Bard (Google), Cohere Command R+, and Meta's LLaMA 2—based on a standardized set of prompts focused on summarization, technical Q&A, and creative generation. The goal is to assess their capabilities in handling diverse use cases, comparing performance across several critical criteria.

2. Objective

To compare the performance, user experience, and response quality of different AI platforms through a consistent evaluation framework using a defined set of tasks and prompts.

3. Methodology and Algorithm

3.1 Define the Use Case

A specific task was selected for evaluation that is applicable to all platforms: text summarization and technical Q&A. These tasks represent core capabilities of LLMs and provide insights into how well the models can handle:

- Summarizing a moderately complex document
- Answering technical questions accurately
- Providing clear, relevant, and insightful responses
- Generating creative but relevant content

3.2 Create a Set of Prompts

To ensure uniformity, a common set of prompts was created:

Prompt 1: Summarization
"Summarize the following article in 150–200 words, focusing on key technical concepts and applications."

Prompt 2: Technical Q&A
"What are the core principles behind quantum computing, and how do they differ from classical computing?"

Prompt 3: Comparative Explanation
"Compare quantum computing with classical computing in terms of speed, data processing, and real-world applications."

Prompt 4: Creative Use Case Generation
"Suggest 3 real-world use cases where quantum computing could outperform traditional systems. Justify each case with technical reasoning."

Each platform was tested using these exact prompts.

3.3 Run the Experiment on Each AI Platform

Each of the prompts was input into the platforms under similar conditions:

- Same input format
- Same response time limit (max 30 seconds)
- No additional fine-tuning

Platforms evaluated:

- **ChatGPT (OpenAI)**
- **Claude 2.1 (Anthropic)**
- **Bard (Google Gemini)**
- **Cohere Command R+**
- **Meta LLaMA 2 (via available APIs)**

Metrics recorded:

- **Response Time**
- **Ease of Interaction**
- **Technical Issues (if any)**
- **Output Characteristics**

3.4 Evaluate Response Quality

Each response was evaluated based on the following criteria:

- **Accuracy**: How factually correct and precise the answer is.
- **Clarity**: How easily understandable the output is.
- **Depth**: The extent of explanation and elaboration.
- **Relevance**: Appropriateness of the response to the prompt.

Scores were assigned from 1 (Poor) to 5 (Excellent).

3.5 Compare Performance

All responses were tabulated and compared using an evaluation matrix. Additional observations were also recorded regarding user experience, interface usability, and strengths/weaknesses.

4. Evaluation of Individual Platforms

4.1 ChatGPT (OpenAI)

ChatGPT continues to be a leader in the AI assistant space. Built on the GPT-4 architecture, the platform provides consistent, coherent, and informative outputs across a wide range of tasks. Its key strengths are versatility, adaptability, and the natural flow of responses. The UI is highly interactive and user-friendly, featuring support for follow-up prompts, memory (in Pro mode), and tools like code interpreter and image understanding.

Prompt 1 - Summarization: ChatGPT produces highly structured and well-articulated summaries. It identifies main points, technical concepts, and uses concise yet informative language. It effectively balances brevity and completeness. For example, when summarizing a document on quantum computing, it captured core ideas such as superposition, entanglement, and potential applications in cryptography and simulation.

Prompt 2 - Technical Q&A: ChatGPT handles technical questions with a high degree of accuracy. The platform explains complex ideas like qubits and quantum gates in a way that is both detailed and digestible to non-experts. It occasionally includes diagrams or code when asked.

Prompt 3 - Comparative Explanation: The platform provided a table-based comparative analysis between classical and quantum computing, making it visually easy to parse. Points were relevant and logically ordered, showing an understanding of performance, data representation, and limitations.

Prompt 4 - Creative Use Cases: ChatGPT demonstrated creativity in identifying practical domains like drug discovery, optimization in logistics, and secure quantum communications. Each use case was explained with relevant technical justifications.

Response Time: Average of 3.2 seconds for standard prompts. **Strengths:** Clarity, natural language, multi-turn conversation. **Weaknesses:** Slightly generic when not prompted for depth. **Best For:** Developers, researchers, educators, general users.

4.2 Claude (Anthropic)

Claude, developed by Anthropic, is designed to prioritize safety, helpfulness, and transparency in AI interactions. Named after Claude Shannon, it emphasizes thoughtful, context-aware responses and often delivers highly detailed outputs. With its 2024 version, Claude 2.1, the platform showcases strong capabilities in technical depth and nuanced language understanding, making it ideal for knowledge-intensive tasks.

Prompt 1 - Summarization

Claude's summarization is notably **more formal and cautious** compared to other platforms. It tends to highlight not just major points but also important secondary ideas, resulting in richer but slightly longer summaries.

For a document on quantum computing, Claude emphasized foundational principles such as:

- Superposition
- Entanglement
- Quantum decoherence
- Quantum error correction

It also added a small note on the ongoing research challenges, a nuance that many other AIs missed. However, sometimes the summary crossed the word limit (200 words) slightly because Claude prioritized completeness over brevity.

Summary Style: Dense, conceptually detailed, slightly verbose.
Strength: Captures both primary and subtle technical aspects.
Weakness: May occasionally exceed required summary length.

Prompt 2 - Technical Q&A

Claude excels at **structured, step-by-step explanations**. When asked about the core principles behind quantum computing, Claude divided its response clearly into:

- Classical Computing Basics
- Quantum Computing Basics
- Key Differences

It introduced terms like qubit, quantum entanglement, and quantum tunneling with textbook-level precision. Claude also often cites theoretical origins (like referencing Dirac notation or Shor's algorithm when discussing applications).

Interestingly, Claude inserted small cautionary notes, e.g., "Quantum supremacy has been demonstrated experimentally but real-world large-scale deployment remains a future goal." — showing a high level of **responsibility and precision**.

Technical Depth: Very high.
Style: Academic, neutral, sometimes a bit formal.

Prompt 3 - Comparative Explanation

In comparing quantum and classical computing, Claude generated a neat bullet-point comparison but chose **longer explanatory paragraphs** under each point instead of short phrases.

Highlights:

- Speed comparisons were backed by specific examples (e.g., factoring large numbers using Shor’s algorithm vs classical RSA encryption times).
- Practical limitations were discussed (error rates, cooling requirements).
- Clear indication of current technological maturity levels.

Claude’s answer felt less “marketing-like” and more like a **professor's detailed explanation**.

Strength: Comprehensive, unbiased comparison.
Weakness: May be too detailed for casual readers.

Prompt 4 - Creative Use Case Generation

When asked for real-world use cases where quantum could outperform classical systems, Claude suggested:

1. **Cryptography:** Breaking traditional encryption with quantum algorithms.
2. **Material Science:** Discovering new superconducting materials through quantum simulations.
3. **Logistics Optimization:** Solving NP-hard problems like optimal routing using quantum annealing.

Each use case was supported by a small paragraph explaining technical feasibility, current barriers, and future potential. Claude explicitly mentioned that **“scalability remains a key challenge”**, maintaining scientific honesty.

Creativity: High, but always grounded in realism.
Style: Technically rigorous, pragmatic.

Additional Observations

Category	Observations
Response Time	Slightly slower (4.0 seconds avg.) but more structured
Ease of Interaction	Clear, but sometimes rigid (doesn’t improvise freely)
Technical Issues	None observed
Unique Traits	Emphasizes safety, cautious interpretation, formal tone

4.3 Bard (Google Gemini)

Bard, powered by Google's Gemini 1.5 model in 2024, represents a highly dynamic, multi-modal AI system. With access to real-time web information (in certain settings) and deep integration into Google’s knowledge base, Bard excels in agility, speed, and integration across tasks. Compared to earlier versions, Bard in 2024 shows much greater technical depth, reduced hallucinations, and a more natural conversational style.

Prompt 1 - Summarization

Bard’s summarizations are **extremely concise and reader-friendly**. Unlike Claude’s thorough academic style, Bard aims for balance — capturing all major points while staying within word limits. Its summaries of a quantum computing article included:

- Key principles (superposition, entanglement)
- Practical applications (cryptography, drug discovery)
- Challenges (decoherence, error rates)

Bard tends to avoid jargon unless explicitly asked, making it very approachable for general audiences.

Summary Style: Crisp, informative, slightly simplified.
Strength: High accessibility, clear articulation.
Weakness: May oversimplify very complex subjects.

Prompt 2 - Technical Q&A

In technical Q&A, Bard delivers solid answers but slightly favors **mainstream knowledge** rather than deep academic insights.
For the quantum computing question:

- Bard correctly defined qubits, superposition, and entanglement.
- It compared classical bits (0 or 1) versus quantum states (superposition of 0 and 1).
- It added real-world examples (IBM Qiskit, Google's Sycamore experiment) to make the answer relatable.

However, unlike Claude, Bard occasionally glossed over more intricate topics like quantum decoherence or quantum gate errors unless specifically prompted for “more technical depth.”

Technical Depth: Moderate to high (but depends on prompt phrasing).
Style: Accessible, beginner-to-intermediate level.

Prompt 3 - Comparative Explanation

When comparing classical and quantum computing, Bard generated a very neat **table format** naturally without explicit instructions:

Classical Computing	Quantum Computing
Based on bits (0/1)	Based on qubits (superpositions)
Sequential processing	Parallel probability amplitudes
Stable systems	Highly sensitive to environment

Bard was **exceptionally good** at highlighting practical differences and **current technological barriers**. It emphasized that **quantum advantage** is limited to certain specific problem domains, not general-purpose tasks yet.

Strength: Well-organized, easily digestible comparisons.
Weakness: Rarely goes into hardcore mathematical explanations unless specifically requested.

Prompt 4 - Creative Use Case Generation

Bard proposed the following creative use cases:

1. **Financial Modeling:** Quantum Monte Carlo simulations for risk assessment.
2. **Logistics Optimization:** Quantum algorithms for complex scheduling and routing.
3. **Drug Discovery:** Simulating molecular structures at quantum level for faster R&D.

Bard's use case explanations were **shorter but practical**. It presented why quantum approaches would be better (faster computation, parallel exploration of possibilities) but did not heavily elaborate on the quantum mechanics side unless further prompted.

Creativity: Very good — imaginative but focused on feasible domains.
Style: Energetic, practical.

Additional Observations

Category	Observations
Response Time	Fastest overall (~2.7 seconds avg.)
Ease of Interaction	Extremely intuitive, conversational
Technical Issues	None observed, although live web access sometimes caused minor delays

Category	Observations
Unique Traits	Very good at real-world examples, tables, diagrams if asked

Overall Evaluation for Bard (Google Gemini)

Metric	Score (1–5)	Comments
Accuracy	4	Generally high, but sometimes simplifies complex topics
Clarity	5	Extremely clear and beginner-friendly
Depth	4	Good depth, but not as intense as Claude's
Relevance	5	Always very prompt-relevant and practical

Best For: Students, general researchers, innovation teams, professionals needing quick technical clarity.

Strengths: Fast, clear, real-world examples, accessible to non-experts.
Weaknesses: Depth limited unless explicitly requested.

4.4 Cohere Command R+ (2024)

Cohere's **Command R+**, released in early 2024, is specialized for retrieval-augmented generation (RAG), enterprise search, and technical writing. Unlike general-purpose models, it focuses heavily on **accuracy, factual grounding**, and **structured output**. Command R+ is widely used in business settings where precision, summarization, and factual correctness are critical. Its core strength lies in **concise, source-supported responses** and **modular generation**.

Prompt 1 - Summarization

Cohere’s summarizations are highly **structured and bullet-point driven** by default.
When summarizing a text on quantum computing:

- It highlighted major principles in bullet points.
- It minimized descriptive storytelling.
- It cited sources (or pretended to, depending on integration with actual RAG datasets).

Example summary points:

- Definition of qubits and quantum superposition.
- Importance of entanglement for parallelism.
- Key applications in cryptography and simulation.

Unlike Claude (rich narratives) or Bard (flowy summaries), Cohere produced a **“business executive summary”** — minimalistic, efficient, and easy to skim.

Summary Style: Bullet points, formal, compressed.
Strength: Fast, factual, reader-efficient.
Weakness: Lacks expressive or narrative tone unless specifically instructed.

Prompt 2 - Technical Q&A

Cohere's responses to technical questions are **direct and modular**.
For a prompt about quantum computing basics:

- Definitions were **broken into labeled sections** ("Definition", "Key Concepts", "Applications", "Challenges").
- Avoided unnecessary elaboration unless asked.
- Frequently used numbered or bullet formats for clarity.

One unique trait is that Cohere tends to include **confidence levels** (e.g., "High confidence based on verified research") when retrieval-based settings are active — an enterprise feature to help in decision-making.

Technical Depth: Good, but less narrative storytelling.
Style: Modular, evidence-based, highly factual.

Prompt 3 - Comparative Explanation

For the comparison between classical and quantum computing, Cohere structured the output as:

- **Short Comparison Table**
- **Follow-up Paragraphs** per point

Table example:

Aspect	Classical Computing	Quantum Computing
Fundamental Unit	Bit (0 or 1)	Qubit (superposed states)
Error Tolerance	High	Very low, requires correction
Application Areas	General-purpose	Specialized optimization, cryptography

The table was followed by compact paragraphs expanding each comparison row.
This **two-step structure** (summary table + deeper paragraph) was unique among evaluated platforms.

Strength: High readability, executive report style.
Weakness: Less "storytelling" flavor for creative audiences.

Prompt 4 - Creative Use Case Generation

When asked for quantum computing use cases, Cohere responded with:

1. **Cryptographic Attacks:** Shor’s algorithm threatens RSA.
2. **Drug Molecule Simulation:** Using quantum phase estimation.
3. **Portfolio Optimization:** Quantum algorithms for asset management.

Each use case was extremely **real-world grounded** — mentioning existing limitations like "requires fault-tolerant quantum systems to scale effectively."

Cohere’s style here felt **business-consulting** oriented rather than imaginative or futuristic.

Creativity: Practical creativity focused on current feasibility.
Style: Executive summary format.

Additional Observations

Category	Observations
Response Time	Very fast (3.0 seconds avg.)
Ease of Interaction	Clean API integration, low chatty interaction
Technical Issues	None, very stable
Unique Traits	Confidence levels, structured modular outputs, retrieval grounding

Overall Evaluation for Cohere Command R+

Metric	Score (1–5)	Comments
Accuracy	5	Excellent factual grounding
Clarity	5	Very clear, modular outputs
Depth	4	Good for business depth; academic depth slightly less
Relevance	5	Highly aligned to prompt, extremely focused

Best For: Business reports, technical documentation, RAG-based knowledge systems.

Strengths: Highly structured, factual precision, enterprise-ready outputs.
Weaknesses: Lacks casual or artistic conversational styles.

4.5 Meta (LLaMA 2/3 - 2024)

Meta’s **LLaMA** (Large Language Model Meta AI) family evolved significantly in 2024 with the release of **LLaMA 3**, focusing on open-weight foundation models optimized for research and modular AI systems.
LLaMA 3's variants (7B, 13B, 65B) offer powerful general-purpose generation while emphasizing transparency, controllability, and multi-turn conversational stability.
Unlike Bard or Cohere, Meta's models are **community-tuned**, meaning outputs are heavily shaped by open research, fostering innovation but sometimes leading to inconsistency.

Prompt 1 - Summarization

Meta's LLaMA-generated summaries exhibit a **dense, information-rich style**.
For a prompt summarizing an article on quantum computing:

- It provided **very detailed definitions** and **contextual background**.
- It captured nuances such as No-Cloning Theorem, error correction protocols, and specific quantum algorithms (e.g., Shor’s, Grover’s).

However, unlike Bard (simplified) or Cohere (bullet-pointed), LLaMA’s summaries are **text-dense**, often resembling academic abstracts.

Summary Style: Academic, research-paper-like.
Strength: High-level detail, accurate terminology.
Weakness: Might be too heavy for casual readers without prompt steering.

Prompt 2 - Technical Q&A

In technical Q&A:

- LLaMA produced **in-depth**, **theory-driven** answers.
- It emphasized not only what quantum computing is, but why it matters mathematically (mentioning Hilbert spaces, tensor products, quantum gates).

For example, instead of simply stating that qubits can be in 0 and 1 simultaneously, it elaborated on **state vectors** and **complex probability amplitudes**.

When compared:

- Claude was comprehensive but reader-friendly.
- LLaMA leaned heavily into **formal academic correctness**.

Technical Depth: Very high — almost PhD-level detail.
Style: Precise, deeply technical, occasionally dry.

Prompt 3 - Comparative Explanation

When comparing classical and quantum computing:

- LLaMA generated not just a table, but also a multi-paragraph **mathematical framework comparison** (e.g., Boolean algebra vs Quantum linear algebra).
- It included formal expressions like:
  - $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$
  - Quantum gates = Unitary operators preserving inner products.

In simpler terms, it "over-delivered" technical detail unless asked to simplify.

Strength: Best for academic and engineering audiences.
Weakness: Requires user to prompt for simplifications for non-expert use.

Prompt 4 - Creative Use Case Generation

On creative prompts, LLaMA’s imagination was **functional but grounded**:

1. **Secure Quantum Communications:** Using quantum key distribution protocols.
2. **Quantum Machine Learning Models:** Enhancing classical ML with quantum data encodings.
3. **Material Science Simulations:** Exploring superconducting materials at molecular scales.

While creative, the use cases remained **anchored in realistic timelines** (acknowledging that full implementation may be decades away), unlike Bard which tended to be more optimistic or futuristic.

Creativity: High in technical feasibility, lower in science fiction-style creativity.
Style: Sober, strategic, future-conscious.

Additional Observations

Category	Observations
Response Time	Medium (~4.1 seconds avg.)
Ease of Interaction	Requires more careful prompt design
Technical Issues	None, very stable
Unique Traits	Deep mathematical modeling, academic robustness

Overall Evaluation for Meta (LLaMA 2/3)

Metric	Score (1–5)	Comments
Accuracy	5	Exceptionally high for technical subjects
Clarity	3.5	Heavy academic language unless prompted otherwise
Depth	5	Outstanding depth, even exceeding Claude in places
Relevance	4.5	Highly relevant but may exceed scope if not constrained

Best For: Researchers, advanced engineers, AI developers, academic institutions.

Strengths: Mathematical rigor, technical richness, transparent open-weight model philosophy.
Weaknesses: Needs skillful prompting for general audiences; dense output by default.