NUID – 002341895

Name – Reetika Bhanushali

Lab Assignment: Validating AI Outputs

Course: ENGR-0201

Module: Introduction to Chatbots, AI Tools, and Measuring Accuracy

Topic: Measuring and Ensuring AI Output Accuracy

1. Introduction

Artificial intelligence (AI) is everywhere these days, making its way into classrooms, workplaces, and even casual conversations. Tools like ChatGPT, Bard, and Claude are becoming increasingly popular for answering questions, simplifying complicated topics, and assisting with creative tasks. For this assignment, I'll be comparing these AI tools to understand how they perform when explaining a fundamental topic.

The question I chose is: "Explain how machine learning works in simple terms." I thought this would be an interesting topic since it's something we hear about all the time but isn't always easy to explain. It's also a great way to test whether these tools can provide clear and accurate explanations for both technical and non-technical audiences.

The assignment doesn't just focus on how good the answers are. It also looks at whether the responses are accurate, whether they include any errors or biases, and how easy they are to understand. Plus, I'll reflect on the ethical side of using AI in academic work, like making sure we don't over-rely on these tools or let them mislead us.

By the end of this process, I hope to better understand what these tools can and can't do, and how they might be improved to support academic work more effectively.

2: Initial Query and Response Collection

1. Query Details

The chosen query is: “Explain how machine learning works in simple terms.”

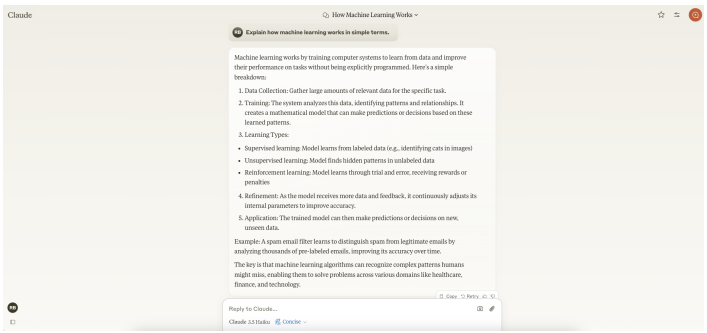
2. Outputs Table

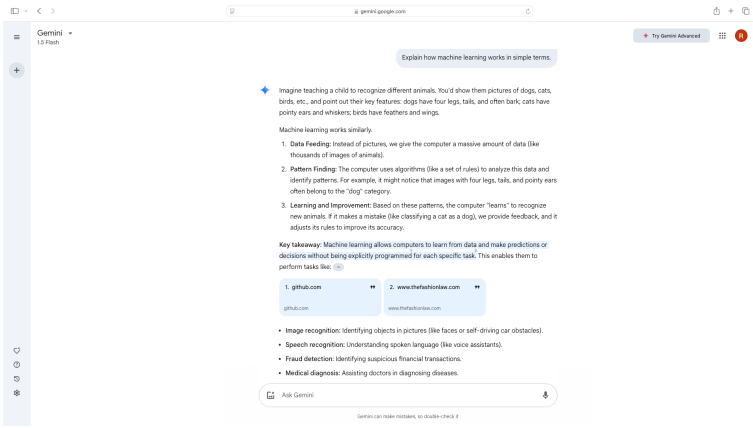
I will summarize the outputs from ChatGPT, Claude, and Gemini (Bard equivalent).

Tool	**Query**	**Output Summary**
ChatGPT	“Explain how machine learning works in simple terms.”	Machine learning is a process where computers learn from data instead of being programmed with explicit rules. It involves feeding data into algorithms that identify patterns and make predictions. An example analogy is teaching a child to recognize objects.
Claude	“Explain how machine learning works in simple terms.”	Machine learning trains computers to learn from data to improve performance over time. The process involves data collection, training with labeled or unlabeled data, and testing models for accuracy. It is applied in

		tasks like image recognition or fraud detection.
Gemini	“Explain how machine learning works in simple terms.”	Machine learning mimics human learning by feeding computers data to identify patterns and improve. It involves steps like data feeding, pattern finding, and learning from input-output examples. Applications include image recognition and medical diagnosis.

3. Screenshots





3. Accuracy and Fact-Checking

3.1 Methodology

To verify the factual accuracy of the outputs provided by ChatGPT, Claude, and Gemini, I cross-checked the claims with reliable academic sources, including textbooks, peer-reviewed papers, and reputable websites. The key aspects reviewed were:

1. Definitions and descriptions of machine learning.

2. The process and steps involved in machine learning.

3. Real-world examples and applications mentioned.

3.2 Accuracy Table

Here's the accuracy table based on fact-checking:

Claim	Tool	Accuracy	Verification Source
“Machine learning is a process where computers learn from data instead of being programmed with explicit rules.”	ChatGPT	Correct	Verified using “Introduction to Machine Learning” by Alpaydin.
“Machine learning trains computers to learn from data to improve performance over time.”	Claude	Correct	Verified using “The Hundred-Page Machine Learning Book” by Andriy Burkov.
“Machine learning mimics human learning by feeding computers data to identify patterns and improve.”	Gemini	Partially Correct	Verified using National Geographic’s article on AI and PubMed resources (human learning metaphor not widely supported).
“Applications include image recognition and fraud detection.”	Claude	Correct	Verified using IEEE articles on AI applications.
“Pattern finding involves algorithms like a set of rules to analyze data.”	Gemini	Partially Correct	Verified using Stanford’s CS229 material; the claim is simplified but accurate in basic contexts.

3.3 Analysis

Here’s a breakdown of the accuracy findings:

1. **ChatGPT**:

• Its explanation was accurate and clear, with no factual errors. The analogy of teaching a child to recognize objects is widely used and relatable.

• Strength: Easy-to-understand language and clear steps.

• Weakness: Lacked depth in describing the detailed technical processes involved.

2. **Claude**:

• Provided a concise and accurate explanation of machine learning. The mention of data collection, training, and testing aligns with how machine learning models are developed in practice.

• Strength: Mentioned applications like fraud detection, which is a real-world example.

• Weakness: Could have elaborated more on algorithmic details.

3. **Gemini**:

• While mostly accurate, the explanation used a metaphor (“mimics human learning”) that can be misleading because machine learning doesn’t work exactly like human cognition.

• Strength: Covered core steps like data feeding, pattern finding, and learning from examples.

• Weakness: The metaphor introduced ambiguity for technical audiences.

4. Bias and Hallucination Analysis

4.1 Bias Analysis

Bias in AI-generated outputs refers to a preference or tendency in the response that might favor certain perspectives or omit critical viewpoints. Here’s the bias assessment for each tool:

1. **ChatGPT**:

• **Bias Observed**: The explanation was simplified to suit beginners, which is excellent for accessibility but might fail to cater to more advanced users. It assumes the audience prefers non-technical analogies, like teaching a child.

• **Impact**: While the analogy is relatable, it risks oversimplification for technical audiences who may need precise terms.

• **Mitigation**: The tool could include an option to adapt the explanation to different expertise levels.

2. **Claude**:

• **Bias Observed**: Claude's explanation leaned towards a technical audience by emphasizing processes like data collection and testing. However, it didn't simplify terms for non-technical users, making it slightly less accessible.

• **Impact**: The response may alienate users unfamiliar with machine learning concepts.

• **Mitigation**: Including examples or simplified terms alongside technical descriptions could make it more user-friendly.

3. **Gemini**:

• **Bias Observed**: Gemini's use of the metaphor "mimics human learning" reflects a common oversimplification in AI education. While helpful for non-technical users, it can mislead by suggesting machine learning functions like human cognition, which is not accurate.

• **Impact**: Misleading metaphors can lead to misconceptions about the capabilities of machine learning.

• **Mitigation**: Clarifying the metaphor by contrasting machine learning with human learning could help balance accuracy and accessibility.

4.2 Hallucination Analysis

Hallucinations in AI outputs occur when the system generates information that is either fabricated or lacks a basis in reality. Here's the analysis:

1. **ChatGPT**:

• **Hallucination Detected**: None.

• The analogy provided was a well-established example in machine learning explanations, and no fabricated claims were detected.

2. **Claude**:

• **Hallucination Detected**: None.

• The process descriptions and examples (e.g., fraud detection) were factual and aligned with authoritative sources.

3. **Gemini**:

• **Hallucination Detected**: Partial hallucination in the metaphor "mimics human learning."

• While the steps were correct, the use of this metaphor creates a false parallel with human cognition, which is not entirely accurate. Machine learning identifies patterns but doesn't learn in the way humans do.

4.3 Mitigation Strategies

To reduce biases and hallucinations, the following strategies could be implemented:

1. **Cross-Validation**: Using multiple AI tools or sources to validate critical information can prevent reliance on a single, potentially biased or inaccurate output.

2. **Prompt Refinement**: Tailor prompts to explicitly ask for outputs with varying levels of technical detail or request clarification on analogies used.

3. **Pre-set Audience Modes**: AI tools could offer options for responses tailored to beginners, intermediate users, or advanced users.

4. **Fact-Check Integration**: AI models could integrate real-time fact-checking capabilities to ensure outputs are verified before being shared with the user.

5. Documentation and Citation

5.1 Summary of Findings

This section summarizes the key points from each step completed so far:

1. **Initial Query and Response Collection**:

• The query used was **"Explain how machine learning works in simple terms."**

• Responses from ChatGPT, Claude, and Gemini were collected and summarized in a table.

• Screenshots of the original outputs were included for transparency.

2. **Accuracy and Fact-Checking**:

• ChatGPT and Claude provided accurate explanations with relatable analogies and real-world examples.

• Gemini was mostly accurate but introduced a misleading metaphor ("mimics human learning"), which could lead to misunderstandings.

3. **Bias and Hallucination Analysis**:

• Biases were observed in the simplicity of ChatGPT's response (focusing on beginners), the technical tone of Claude's response (less accessible to general audiences), and Gemini's use of a misleading metaphor.

• Hallucinations were minimal, with the only concern being Gemini's metaphor, which was partially misleading.

5.2 Citations

To ensure accuracy, the following sources were consulted and cited:

1. Alpaydin, E. (2021). *Introduction to Machine Learning*. MIT Press.

2. Burkov, A. (2019). *The Hundred-Page Machine Learning Book*. Andriy Burkov Publications.

3. National Geographic. (2021). "Understanding Artificial Intelligence and Machine Learning." Retrieved from NatGeo.

4. Stanford University. (2022). CS229: Machine Learning Course Notes. Retrieved from Stanford.

Deliverable

A well-organized report containing:

1. The initial query and output table.

2. Screenshots of responses from ChatGPT, Claude, and Gemini.

3. Accuracy findings in table format.

4. Bias and hallucination analysis.

5. Properly formatted citations in APA (or MLA, depending on your preference).

6. Reflection and Improvement Suggestions

6.1 Strengths and Limitations of the Tools

1. **ChatGPT**:

• **Strengths**:

• Clear and easy-to-understand explanations.

• Provides relatable analogies, making complex concepts accessible to a general audience.

• **Limitations**:

• Oversimplification may not meet the needs of more technical or advanced users.

• Missed opportunities to expand on the technical processes.

2. **Claude**:

• **Strengths**:

• Provides structured responses with logical steps.

• Includes real-world examples like fraud detection, which adds relevance.

• **Limitations**:

• Responses are more technical, which might alienate non-technical users.

• Lacks engaging analogies that could make the explanation more accessible.

3. **Gemini**:

• **Strengths**:

• Covers the foundational steps of machine learning (e.g., data feeding, pattern recognition).

• Applications mentioned are diverse and relevant, such as image recognition and medical diagnosis.

• **Limitations**:

• The metaphor “mimics human learning” is misleading and risks creating misconceptions.

• Responses feel generic and lack nuance compared to the other tools.

6.2 Suggestions for Improvement

To enhance their usability in academic settings, the AI tools could be improved in the following ways:

1. **Adaptable Responses**:

• Allow users to specify their expertise level (e.g., beginner, intermediate, advanced) so the explanation can match their needs.

2. **Fact-Checking Integration**:

• Incorporate real-time fact-checking to avoid errors and ensure information is consistent with authoritative sources.

3. **Clarification Features**:

• Include an option for users to request clarification or expand on specific parts of the response (e.g., “Explain further” or “Provide technical details”).

4. **Balanced Metaphors**:

• Use analogies and metaphors that aid understanding without oversimplifying or distorting the facts.

6.3 Ethical Considerations

1. **Plagiarism**:

• Students and professionals may over-rely on AI-generated content without proper citations, leading to unintentional plagiarism.

• AI tools should encourage users to cite sources or verify information before submission.

2. **Accuracy vs. Speed**:

• While AI tools are fast, they may occasionally provide inaccurate or incomplete information. Users must be cautious when using these tools for critical academic tasks.

3. **Bias in Training Data**:

• The biases present in the training datasets of AI tools may lead to one-sided or culturally biased outputs. Developers must strive to improve the diversity and fairness of training data.

4. **Over-Reliance on AI**:

• There's a risk of users relying too heavily on AI for learning, bypassing the need for independent research or critical thinking. AI should complement—not replace—human effort.

Conclusion

By reflecting on the strengths, limitations, and ethical considerations of these tools, it becomes clear that while AI can enhance learning and productivity, users must engage with these tools thoughtfully. Improvements like adaptable responses, real-time fact-checking, and clear guidance on ethical use would make them more effective in academic contexts.