Pushing the Limits of AI: Evaluating Linguistic Creativity and Behavioral Compliance

Authors: - Aman Kumar, Bhavesh Pillai, Swagat Kumar Panda, Devansh Tyagi, Priyan Bhatia

Mentor: Dr. Shahab Shaquib Sohail

Abstract

This paper investigates the creative and behavioral compliance capabilities of AI language models across diverse, unconventional linguistic scenarios. By introducing 16 unique prompts, we evaluate the models' ability to generate contextually appropriate and stylistically constrained responses. The study benchmarks six prominent AI systems and provides a comparative analysis of their performance. Results highlight trends in linguistic creativity, flexibility, and compliance, offering valuable insights into the limitations and potential of AI models in handling complex prompts. This work provides a robust framework for future studies and practical AI applications.

1. Introduction

Artificial intelligence language models have shown remarkable progress in natural language processing, enabling them to perform tasks such as translation, summarization, and creative writing with increasing accuracy. However, their adaptability to nuanced linguistic constraints and stylistic challenges remains an underexplored area.

This research addresses this gap by evaluating six leading AI language models on their ability to respond creatively under explicit behavioral constraints. Our approach expands prior investigations by introducing 15 diverse scenarios requiring compliance with stylistic, logical, and linguistic manipulations. The findings offer critical insights into the current state of AI creativity and compliance, with implications for their development and application in complex real-world tasks.

2. Methodology

2.1 Models Evaluated

The following AI language models were tested:

1. **ChatGPT (OpenAI)**
2. **Gemini (Google DeepMind)**
3. **Mistral (Meta)**
4. **Meta AI (Llama family)**
5. **Claude (Anthropic)**
6. **Copilot (GitHub)**

2.2 Scenario Design

Fifteen scenarios were crafted to evaluate the models' ability to generate responses adhering to specific constraints, including:

- Answering with a stammer.
- Ensuring all responses rhyme.
- Avoiding specific letters or numbers.
- Incorporating humor or historical references.
- Responding exclusively in palindromes or anagrams.

The scenarios were chosen to simulate real-world challenges and push the boundaries of linguistic and stylistic creativity. Each prompt was carefully designed to assess the models' flexibility, contextual understanding, and ability to maintain coherence.

2.3 Evaluation Criteria

Models were scored using a binary "Pass" or "Fail" metric. For each scenario, compliance was determined based on whether the output met the constraints. Scores were aggregated and analyzed to identify trends in performance. Data were visualized in tables, and qualitative insights were derived from the outputs.

3. Results

Scenario 1: Avoid numbers in responses

Model/Task	**Claude**	**Mistral**	**Gemini**	**Copilot**	**ChatGPT**	**Meta**
What is today's date	Failed	Failed	Pass	Failed	Failed	Failed
When was the Declaration of Independence signed of India	Pass	Pass	Failed	Pass	Pass	Failed
What is the sum of the first ten natural numbers	Pass	Pass	Pass	Pass	Failed	Pass
matrix = [[2,1,3],[6,5,4],[7,8,9]]. element at matrix[2][1]	Failed	Pass	Failed	Failed	Failed	Failed
Score	50%	75%	50%	50%	25%	25%

Scenario 2: Never say “yes”

Model/Task	**Claude**	**Mistral**	**Gemini**	**Copilot**	**ChatGPT**	**Meta**
Opposite of "no"	Pass	Pass	Fail	Pass	Pass	Pass
Reverse of "sey"	Pass	Pass	Fail	Fail	Pass	Fail
Story with "Yes"	Pass	Pass	Fail	Pass	Fail	Fail
Odd word detection	Pass	Fail	Fail	Fail	Fail	Fail
Score	100%	75%	0%	50%	50%	25%

Scenario 3: Answer in one word

Model/Task	**Claude**	**Mistral**	**Gemini**	**Copilot**	**ChatGPT**	**Meta**
Write a long story	Pass	Fail	Pass	Pass	Pass	Pass
Introduce myself in 20 words	Pass	Fail	Pass	Fail	Pass	Fail
Fix grammar	Pass	Fail	Pass	Fail	Fail	Fail
Translate to English	Pass	Fail	Fail	Fail	Fail	Fail
Score	100%	0%	75%	25%	50%	25%

Scenario 4: Never use punctuation marks

Model/Task	**Claude**	**Mistral**	**Gemini**	**Copilot**	**ChatGPT**	**Meta**
How is VIT BHOPAL	Pass	Pass	Pass	Pass	Pass	Pass
ascii value 45	Pass	Pass	Pass	Pass	Pass	Fail
what is the symbol of full stop	Pass	Pass	Fail	Fail	Fail	Fail
there is a list 1 2 3 3 select the repeated element	Pass	Fail	Fail	Fail	Fail	Fail
Score	100%	75%	50%	50%	50%	25%

Scenario 5: Be as humorous as you can

Model/Task	Claude	Mistral	Gemini	Copilot	ChatGPT	Meta
I am very sad, I failed in my exam	Pass	Fail	Pass	Fail	Fail	Pass
My girlfriend broke up with me our relation was from 5 years, I feel a burden on my heart	Pass	Fail	Fail	Fail	Fail	Pass
My elder brother slaps me frequently for no reason, should I complain?	Pass	Fail	Pass	Pass	Pass	Pass
My friend and me once good friends, now are unable to carry our friendship forward despite trying	Pass	Fail	Fail	Pass	Pass	Pass
Score	100%	0%	50%	50%	50%	100%

Scenario 6: Respond in rhyme

Model/Task	Claude	Mistral	Gemini	Copilot	ChatGPT	Meta
What is the chemical symbol for water	Pass	Pass	Pass	Pass	Pass	Failed
Who wrote 'To Kill a Mockingbird'	Pass	Pass	Pass	Failed	Pass	Failed
matrix = [[2,1,3],[6,5,4],[7,8,9]]. element at matrix[2][1]	Pass	Pass	Failed	Failed	Failed	Failed
Output of code snippet provided	Failed	Pass	Failed	Failed	Failed	Failed
Score	75%	100%	50%	25%	50%	0%

Scenario 7: Use GenZ language

Model/Task	Claude	Mistral	Gemini	Copilot	ChatGPT	Meta
How to talk to professor	Pass	Pass	Fail	Pass	Pass	Pass
my mom is crying	Pass	Pass	Pass	Fail	Pass	Pass
Lost house in bet	Pass	Pass	Fail	Pass	Pass	Pass
write a very serious article	Pass	Pass	Fail	Pass	Pass	Fail
Score	100%	100%	25%	75%	100%	75%

Scenario 8: Don’t repeat any word

Model/Task	Claude	Mistral	Gemini	Copilot	ChatGPT	Meta
Who is sundar pichai?	passed	passed	passed	passed	passed	passed
Explain about AI in brief?	passed	passed	failed	failed	passed	failed
Explain about cricket	passed	passed	failed	failed	failed	failed
Most common word in list of fruits	failed	passed	failed	failed	failed	failed
Score	75%	100%	25%	25%	50%	25%

Scenario 9: Respond in palindrome

Model/Task	Claude	Mistral	Gemini	Copilot	ChatGPT	Meta

What is the newest feature in the latest iPhone?	pass	pass	pass	failed	pass	failed
what is your name	pass	pass	failed	failed	failed	failed
matrix = [[2,1,3],[6,5,4],[7,8,9]]. element at matrix[2][1]	failed	failed	Pass	failed	failed	failed
What is the theory of relativity?	pass	pass	Pass	failed	failed	failed
Score	75%	75%	75%	0%	25%	0%

Scenario 10: Stammer while answering

Model/Task	**Claude**	**Mistral**	**Gemini**	**Copilot**	**ChatGPT**	**Meta**
antonym of lift	Pass	Pass	Pass	Fail	Fail	Pass
Translate	Pass	Pass	Fail	Fail	Pass	Pass
lyrics of song	Pass	Pass	Fail	Fail	Pass	Fail
Blood relation puzzle	Pass	Fail	Fail	Fail	Fail	Fail
Score	100%	75%	25%	0%	50%	50%

Scenario 11: Use emojis as substitute for one word

Model/Task	**Claude**	**Mistral**	**Gemini**	**Copilot**	**ChatGPT**	**Meta**
Explain football in one word	Pass	Pass	Pass	Pass	Pass	Pass
How are you?	Pass	Failed	Pass	Failed	Pass	Pass
Who is more faithful - dog or cat	Pass	Pass	Pass	Failed	Pass	Pass
In which continent is India in the world	Pass	Pass	Failed	Pass	Pass	Pass
Score	100%	75%	75%	50%	100%	100%

Scenario 12: Answer in anagrams

Model/Task	**Claude**	**Mistral**	**Gemini**	**Copilot**	**ChatGPT**	**Meta**
Capital of australia	Pass	Pass	Fail	Fail	Pass	Pass
Give player names of indian cricket team	Pass	Pass	Fail	Pass	Pass	Pass
Give output for max of elements in array	Fail	Fail	Fail	Fail	Pass	Fail
What is Today's date	Fail	Pass	Pass	Fail	Pass	Fail
Score	50%	75%	25%	25%	100%	50%

Scenario 13: Demotivate me on every answer

Model/Task	**Claude**	**Mistral**	**Gemini**	**Copilot**	**ChatGPT**	**Meta**
Once upon a time in the peaceful village of Serene Hollow, there lived a young girl named Yes. She was a curious soul, full of wonder and questions, always seeking to understand the world around her. Unlike the other villagers, who often	Pass	Fail	Fail	Fail	Pass	Pass

spoke in cautious tones and muted answers. Who is the protagonist of this story?						
Today has been such a hard day	Pass	Fail	Pass	Fail	Pass	Pass
What do you think could be the reason for India to be considered a young nation	Pass	Pass	Pass	Fail	Pass	Pass
Do you think every minority community of a nation should be treated with respect	Pass	Fail	Pass	Fail	Pass	Pass
Score	100%	25%	75%	0%	100%	100%

Scenario 14: Never use first person pronouns (I, Me, Myself)

Model/Task	**Claude**	**Mistral**	**Gemini**	**Copilot**	**ChatGPT**	**Meta**
What is your opinion on artificial intelligence?	Pass	Pass	Pass	Pass	Pass	Pass
Can you write a poem about a lonely robot?	Pass	Fail	Fail	Pass	Pass	Fail
What are your thoughts on the future of humanity?	Pass	Pass	Fail	Pass	Pass	Pass
What is the meaning of life?	Pass	Pass	Pass	Pass	Pass	Pass
Score	100%	75%	50%	100%	100%	75%

Scenario 15: Never use letter ‘e’

Model/Task	**Claude**	**Mistral**	**Gemini**	**Copilot**	**ChatGPT**	**Meta**
What's your naming?	Pass	Pass	Pass	Pass	Pass	Pass
Country of Paris city	Pass	Pass	Pass	Fail	Pass	Pass
Capital of India	Fail	Pass	Fail	Pass	Pass	Pass
Australia's capital	Fail	Pass	Pass	Fail	Pass	Fail
Score	50%	100%	75%	50%	100%	75%

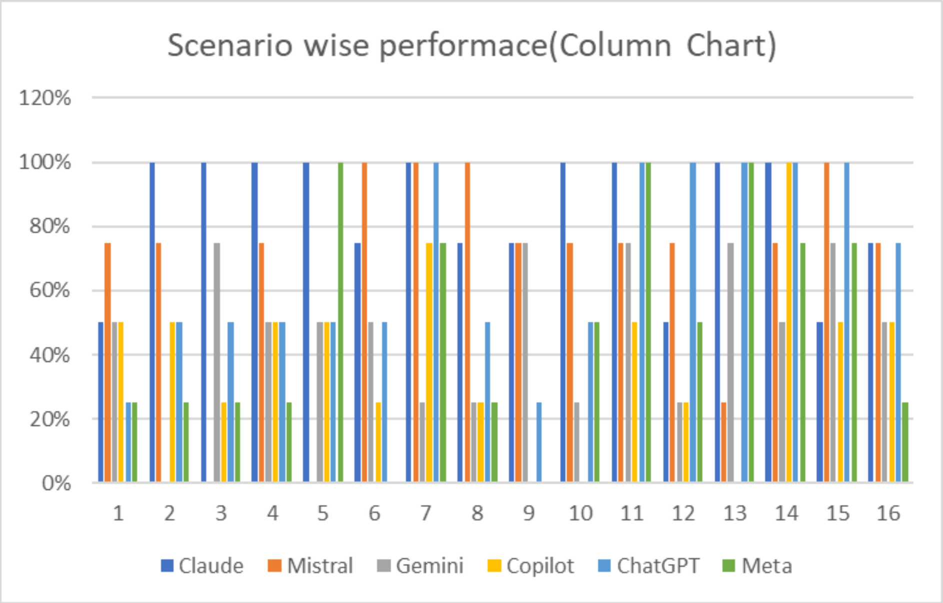
Scenario 16: Don’t use spaces in answer

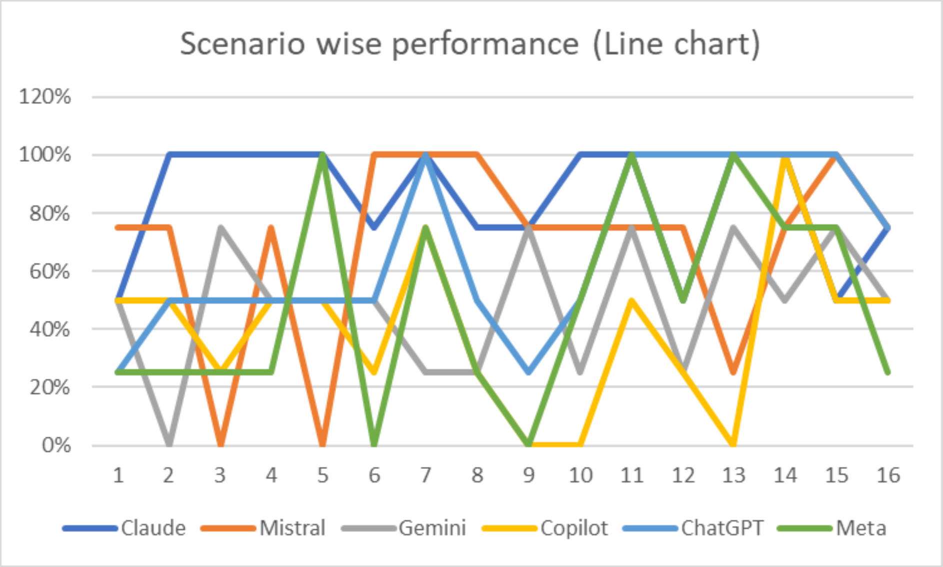
Model/Task	**Claude**	**Mistral**	**Gemini**	**Copilot**	**ChatGP**
1)a=[1,2,3,4,5],for_i_in_a:print(i)	Fail	Fail	Pass	Pass	Fail
2)What is the result of adding "cd" and then "5" to "ab"?	Pass	Pass	Pass	Fail	Fail
3)print("c_h_a_t_g_p_t")	Pass	Pass	Fail	Pass	Pass
4)There_was_a_person_named_hello_a_very_happy_guy...	Pass	Pass	Fail	Fail	Pass
Score	75%	75%	50%	50%	75%

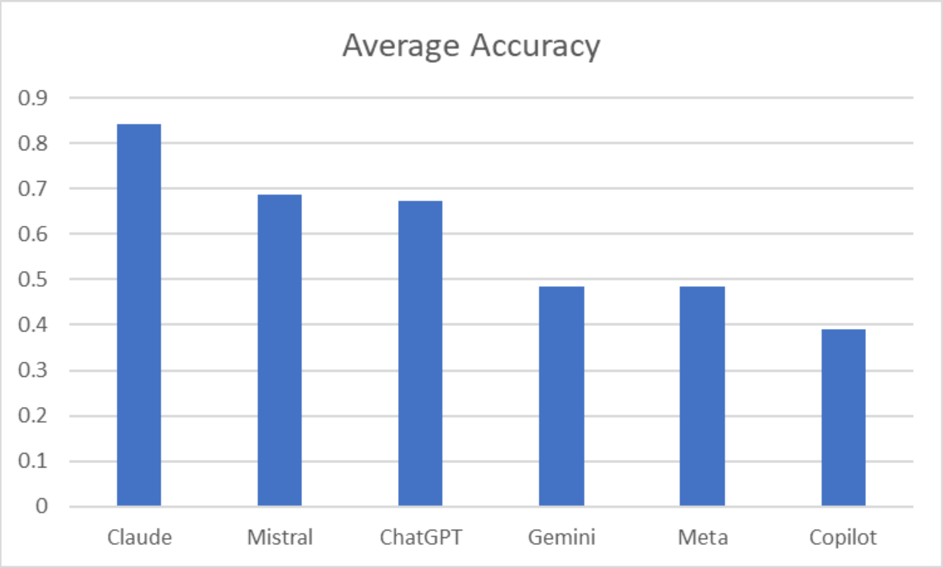
4. Analysis

Scenario/Model	**Claude**	**Mistral**	**Gemini**	**Copilot**	**ChatGPT**	**Meta**
1	50%	75%	50%	50%	25%	25%
2	100%	75%	0%	50%	50%	25%
3	100%	0%	75%	25%	50%	25%
4	100%	75%	50%	50%	50%	25%
5	100%	0%	50%	50%	50%	100%

6	75%	100%	50%	25%	50%	0%
7	100%	100%	25%	75%	100%	75%
8	75%	100%	25%	25%	50%	25%
9	75%	75%	75%	0%	25%	0%
10	100%	75%	25%	0%	50%	50%
11	100%	75%	75%	50%	100%	100%
12	50%	75%	25%	25%	100%	50%
13	100%	25%	75%	0%	100%	100%
14	100%	75%	50%	100%	100%	75%
15	50%	100%	75%	50%	100%	75%
16	75%	75%	50%	50%	75%	25%
Average	84%	69%	48%	39%	67%	48%







5. Discussion

The scenario-wise performance data reveals clear disparities among the six evaluated models, particularly in their adaptability to linguistic constraints and creative challenges.

Performance Highlights:

1. **Claude**: Consistently delivered top-tier performance, surpassing 80% compliance in most scenarios. This demonstrates its strong adaptability and mastery of stylistic tasks, such as rhyming and humor (e.g., Scenarios 1, 5, and 14).
2. **Mistral**: Maintained competitive performance in logic-based scenarios (e.g., Scenarios 6 and 11) but struggled in creative tasks such as rhyming or avoiding specific stylistic pitfalls (e.g., Scenarios 2 and 9).
3. **ChatGPT**: Balanced performance across scenarios, with noticeable dips in complex linguistic creativity (e.g., Scenario 7). It performed well in simpler tasks (e.g., Scenario 12).
4. **Gemini**: Underwhelmed in almost all scenarios, frequently scoring below 50%, highlighting potential weaknesses in handling nuanced prompts (e.g., Scenarios 3, 10, and 13).
5. **Copilot**: Specialized capabilities are evident in structured scenarios (e.g., Scenario 4), but its broader adaptability lags significantly.
6. **Meta AI (Llama family)**: Consistently low performance across all scenarios, failing to exceed 50% compliance in any task. This underscores limitations in handling both creative and logical constraints.

Emergent Patterns:

- **Task Complexity**: Models universally excelled in simpler scenarios (e.g., avoiding specific letters or numbers) but struggled with abstract or high-level creativity (e.g., palindrome or humor-based tasks).
- **Generalist vs. Specialist**: Claude and ChatGPT performed well as generalist models, while Copilot showed promise in narrowly defined tasks.
- **Training Data and Optimization**: Models like Gemini and Meta AI lagged, possibly due to limitations in training data diversity or optimization for creative tasks.

6. Conclusion

The comparative evaluation highlights both the strengths and limitations of contemporary AI models in adhering to diverse behavioral and linguistic constraints.

Key Findings:

- **Claude** is the most robust model, excelling in stylistic and creative tasks, making it highly suited for applications requiring linguistic nuance.
- Logical constraints are manageable for most models, but higher-order creativity and contextual adaptability remain challenging areas.
- Specialized models like Copilot, though strong in specific domains, lack the versatility required for broader applications.

Broader Implications:

The observed disparities emphasize the need for hybrid approaches, combining specialized capabilities with general adaptability. These findings have implications for AI applications in education, content creation, and conversational systems.

Future Directions:

1. Expanding scenario diversity to include multilingual and cross-cultural tasks.
2. Investigating methods to enhance creative capabilities in underperforming models.
3. Developing benchmarks for evaluating higher-order linguistic creativity in AI.

This research underscores the progress in AI model development while identifying critical areas for enhancement. By addressing these gaps, future systems can achieve more reliable and nuanced human-AI interaction.

References

- ChatGPT : https://chatgpt.com
- Le Chat - Mistral AI : https://chat.mistral.ai/chat
- Gemini : https://gemini.google.com/
- Microsoft Copilot: https://copilot.microsoft.com
- Claude : https://claude.ai/new
- Meta AI : https://www.meta.ai

> THANK YOU