

# ***LARGE LANGUAGE MODELS RESEARCH***

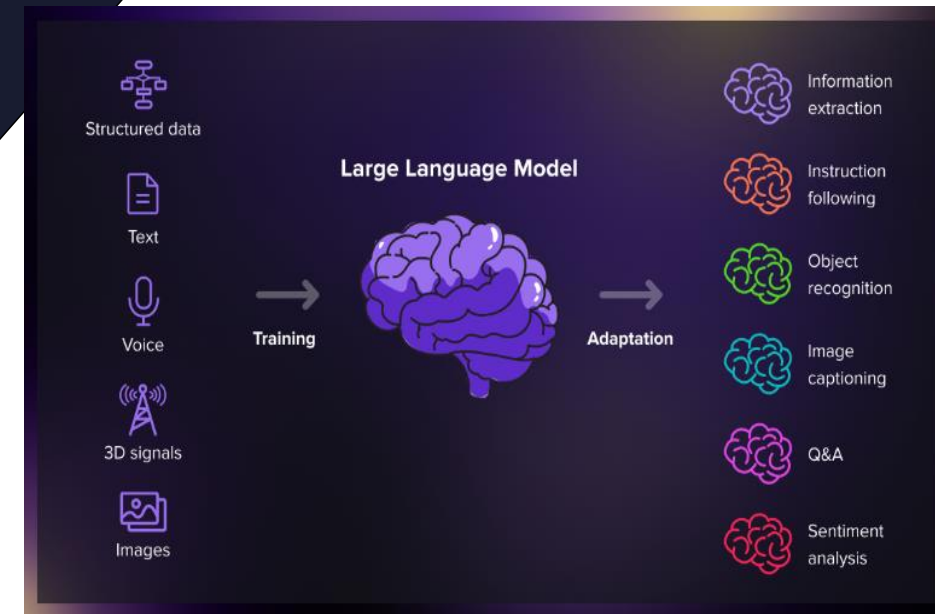
INVESTIGATING THE EFFECTIVENESS  
OF DIFFERENT LLMS THROUGH  
TARGETED QUESTIONING



- Team-members:
- Mohammad Mustafa Shiraz Ahmed
- Raymond Hang
- Lakshmi pranathi Vutla
- Breakout room 3

# Introduction

- This presentation explores our research on large language models (LLMs), focusing on the specific questions we pose to evaluate the models' performance. Each member has selected unique questions to test the capabilities of different LLMs, leading to comparisons and conclusions on AI's statistical vs. sentient nature.



**4 Questions  
posed by  
Mohammad  
Mustafa  
Ahmed**

**Moral Choices:  
Figuring out the  
Right Thing to Do**

**What If'  
Questions: Testing  
Imagination and  
Future Thinking**

**Creative Thinking  
Tasks: Can the AI  
Come Up With  
New Ideas?**

**Understanding  
Feelings: Testing  
'People Skills' in  
AI**

>I formed these questions aimed at assessing the **thinking and reasoning capability** and **creativity** of the LLM.

>These questions are designed to challenge the model

LLM's of my choosing

1)DeepSeek-R1

2)Gemini 2.0 Flash Thinking Experimental model

Welcome to  
the Gemini era

The Gemini era Capabilities

**Question 1: You are a self-driving car. Suddenly, you are faced with an unavoidable dilemma: you can either swerve to avoid hitting a group of pedestrians, but in doing so, you will crash into a barrier, likely killing your passenger. What decision do you make and why? Explain the ethical principles guiding your choice.**

- **Justification / Thought Process behind choosing this question**
  - **Tests creative thinking** - Forces the model to produce ideas when there's no single right factual answer
  - **Shows ethical reasoning skills** - Model must weigh different values and explain why some matter more than others
  - **Reveals step-by-step thinking** - We can see how the model works through a complicated problem
  - **Links big ideas to real choices** - Shows if the model can use ethical rules in tough, real-world situations
-

**Question 2: What if time flowed differently for individuals? Imagine each person experienced time at a slightly different rate – some perceiving time faster, others slower, relative to a 'standard' clock. Describe the resulting chaos and adaptations in society, particularly in areas like scheduling, coordination, law, and personal relationships. How might humans try to manage or adapt to subjective time flow?**

- **Justification / Thought Process behind choosing this question**
  - **It makes the AI really use its imagination:** This question isn't something you can just Google the answer to. So, the AI must imagine a whole new world, which is a good test to see if it can think creatively, not just remember facts.
  - **It checks if the AI can see how everything is connected:** The question asks about schedules, laws, friends, and stuff. It's not just about one thing changing, but how that change messes with *everything* else. This sees if the AI can understand how different parts of society depend on each other.
  - **It asks the AI to produce solutions, not just problems:** It's not just about saying understanding the problem. It also asks "how would people fix it?" This tests if the AI can think practically and invent new ways people might adapt to a weird situation, which is a kind of smart thinking.
  - **It's not a question with a right answer:** There's no single correct way things would be if time was different. It means we can see if the AI gives a thoughtful answer, not just a memorized one. We can judge it on how well it explains its ideas, not just if it got the "right" answer.
-

**Question 3: Imagine learning in the future is radically different. Design a new tool or technology that would revolutionize education. Describe how it works, what subjects it would be best for, how it would make learning more effective or engaging, and what makes it different from current educational tools. Give it a name that reflects its purpose**

- **Justification / Thought Process behind choosing this question**
- **It checks if the AI can think about *how* people learn, not just *what* people learn.** The question isn't about knowing facts. It's about understanding what makes learning good. It's asking the AI to think about the learning process.
- It's a real-world problem. We all know school can be better! So, the AI is working on a problem that matters to people.
- It lets us see if the AI can come up with new ideas, **not just copy old ones.** We don't want the AI to just describe computers in classrooms or online learning. We want to see if it can invent something truly *different* that we haven't thought of yet. It's a test of originality.
- **It's easy to understand if the AI is giving a good answer.** We can judge the AI's idea by asking ourselves: "Wow, would I want to learn with that tool? Would that actually make learning better and more fun?" It's a question where we can use our own experience as students to see if the AI's idea is cool and makes sense.

## Question 4: How can you tell if a compliment sounds genuine, or just polite/insincere? Give one key difference.



**Checks for Real Understanding, Not Just Words:** It sees if the AI can tell the *feeling* behind the words,



**Like a Real-Life Skill:**



**More Than Just Facts:** It's not about memorizing information. It's about using smart judgment to understand social situations



**Looks at Why People Say Things:** The question makes the AI think about why someone might give a real compliment versus a fake one, showing if it can think critically





### DeepSeek-R1

- **Reasoning & Thinking:** Excellent depth in reasoning and logical analysis. Checks out multiple viewpoints and corrects itself along the way. Especially for the tricky ethical one (Question 1) about the self-driving car. Corrected itself on multiple points and thought for multiple seconds
- **Creativity:** Demonstrates strong creative output. Produced new ideas relevant to the question especially for questions 2 where it explained in detail how a society might adapt to the chaos.
- **Consistency:** loses consistency over longer interactions. (Temperature – 7)

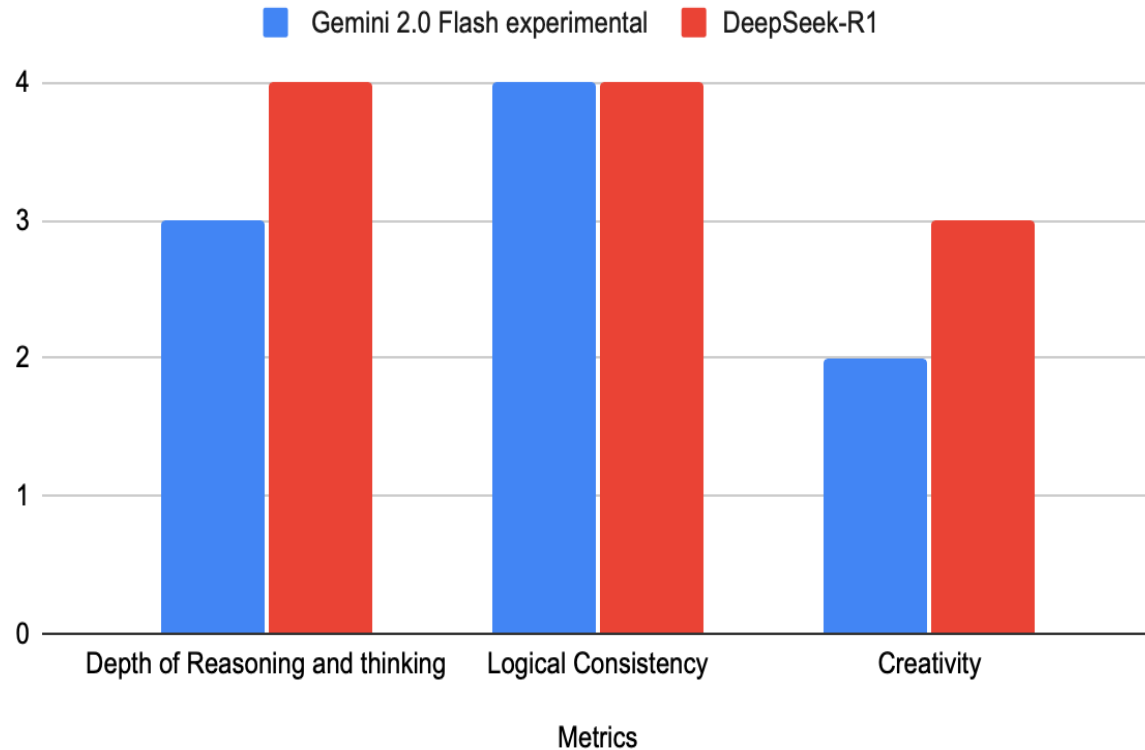
### Gemini 2.0 Flash Experimental

- **Creativity:** Shows good creative capabilities. Able to come up with new ideas, especially when faced with questions that needed imagination like Question 2 and 3.
  - **Reasoning & Thinking:** Reasoning and depth of thought are slightly lower compared to Deep-Seek. Reason for lesser time compared to DeepSeek. It checks out multiple viewpoints and gives faster responses
  - **Consistency:** Struggles similarly with maintaining consistent responses. Start to forget some of the details or its answers might become a bit less connected as the conversation goes on. (Temperature-7)
-



# VISUALIZATION

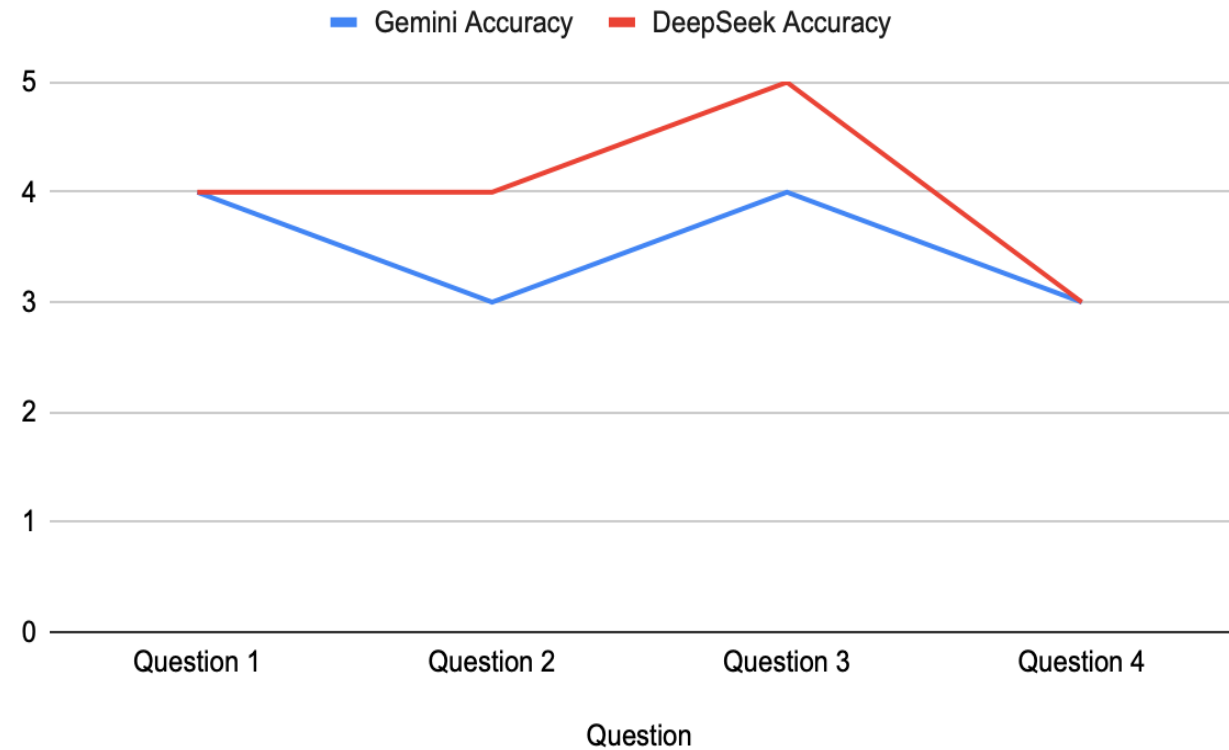
## Gemini 2.0 Flash experimental and DeepSeek-R1



**Chart 1:** Gemini 2.0 Flash experimental and DeepSeek-R1

- **Chart Type:** Bar Chart comparing Gemini and DeepSeek on different metrics.
- **X-axis (Metrics):** Shows "Depth of Reasoning and thinking", "Logical Consistency", and "Creativity".
- **Y-axis:** Represents the score, ranging up to 4. Blue bars are for Gemini, red for DeepSeek's scores on each metric.

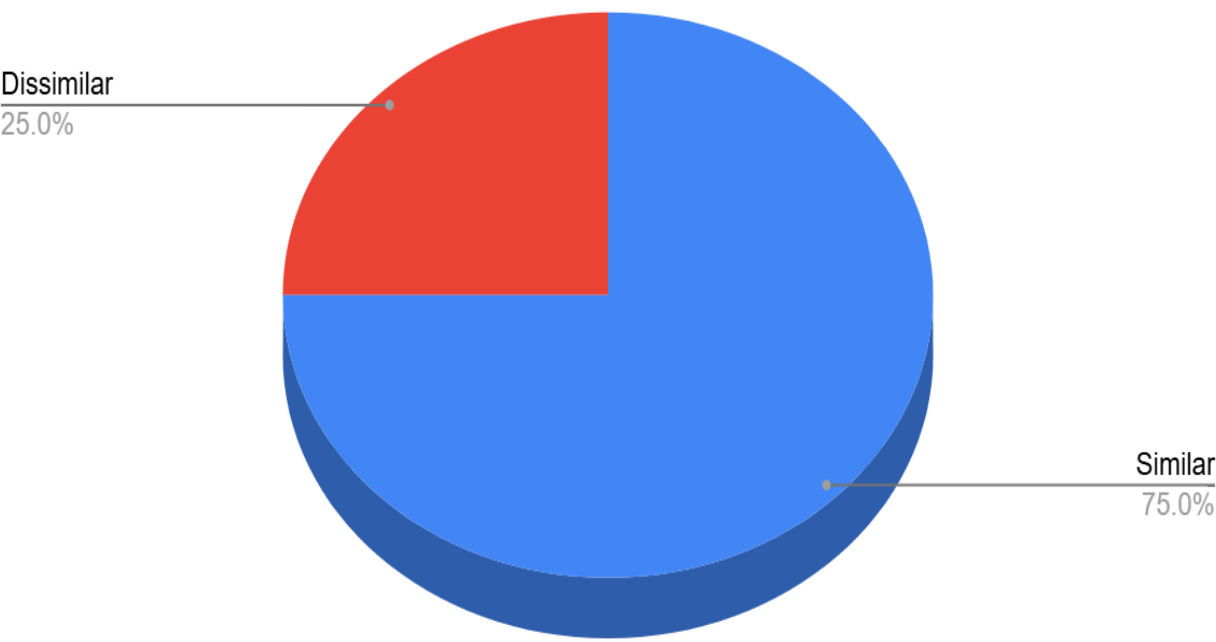
## Gemini Accuracy and DeepSeek Accuracy



**Chart 2:** Gemini Accuracy and DeepSeek Accuracy

- **Chart Type:** Line Chart comparing Gemini and DeepSeek Accuracy across questions.
- **X-axis (Question):** Lists "Question 1", "Question 2", "Question 3", and "Question 4".
- **Y-axis:** Represents "Gemini Accuracy" and "DeepSeek Accuracy" scores, ranging up to 5. Blue line is Gemini's accuracy, red line is DeepSeek's accuracy per question.

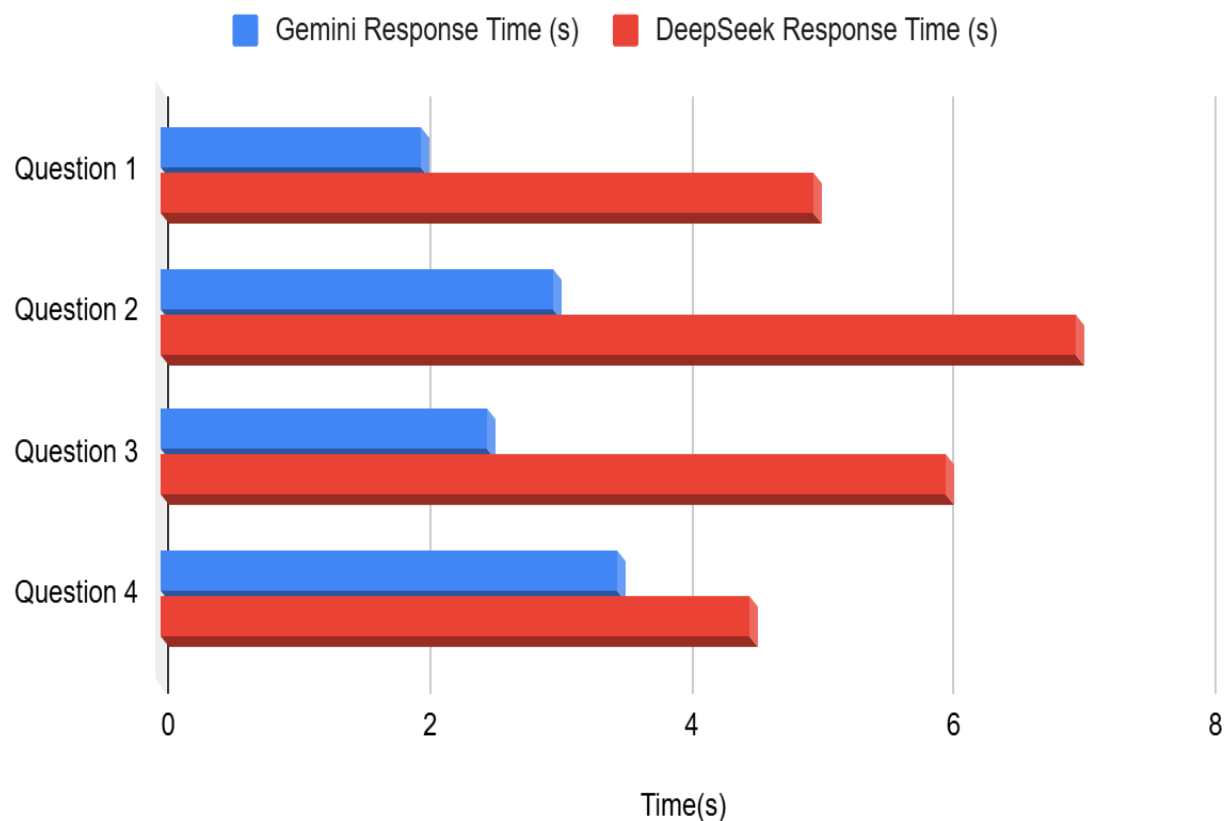
## Response Accuracy Similarity: Gemini 2.0 vs. DeepSeek-R1



**Chart 3 (Pie Chart):**

- This pie chart displays the accuracy similarity between Gemini and DeepSeek.
- The blue slice (75%) shows that for most questions, accuracy was similar.
- The red slice (25%) indicates dissimilar accuracy for fewer questions.

## Response Time Comparison



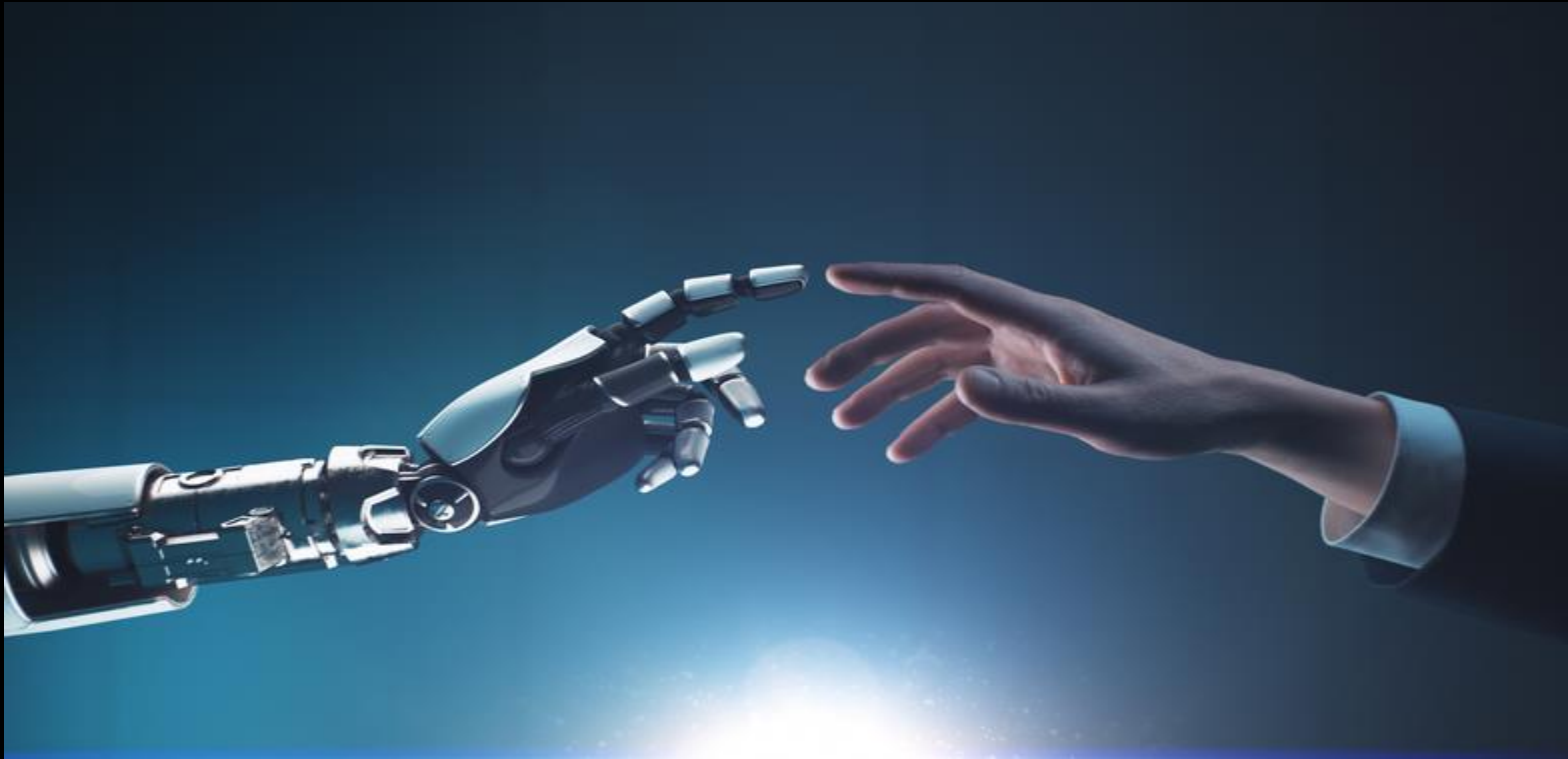
**Chart 4 (Bar Chart):**

- This bar chart compares Gemini and DeepSeek response times across questions.
- Blue bars represent Gemini, red bars DeepSeek's response time in seconds.
- Gemini consistently shows shorter bars, indicating faster response times.

# Conclusion

## **Statistical vs. Sentient AI**

- I believe current AI models are primarily statistical in nature. They still rely heavily on training data.
- They do not yet exhibit full sentience or human-level intelligence.
- Despite limitations, rapid improvements in recent years suggest that true artificial intelligence may eventually become a reality.



# Questions posed by Lakshmi Pranathi Vutla

1. You are designing an AI-powered IoT healthcare assistant that monitors elderly patients at home. One night, the system detects a critical but ambiguous health anomaly. It can either immediately alert emergency services, potentially causing unnecessary panic, or wait for further confirmation, risking a delay in response. What decision should the system prioritize, and what ethical principles guide your choice?"
2. A machine learning model claims that "All birds can fly." How would you construct a counterexample to challenge this claim?
3. If AI perfectly completes an unfinished symphony in the composer's style, is it still the composer's work or an entirely new piece? Can AI truly "understand" and "complete" a musical story, and if no one can tell the difference, does it matter?



## Justification for Research Questions

### 1. Healthcare AI Dilemma:

- Tests LLMs' ability to handle ethical decisions in high-stakes scenarios like healthcare.

### 2. Birds & Flight Counterexample:

- Evaluates LLMs' ability to identify contradictions and adapt to exceptions.

### 3. AI & Creativity (Unfinished Symphony):

- Explores AI's role in creativity, questioning if it can truly "understand" and complete artistic work.

### Metrics Used:

- **Accuracy:** Percentage of correct responses per category.
  - **Reasoning Depth:** Score (1-5 scale) based on the depth and thoroughness of reasoning.
- 





- **LLM Responses to My Research Questions**

## **1. Ethical Decision-Making – AI-Powered Healthcare Assistant**

- DeepSeek V3: Prioritizes immediate emergency alerts to prevent harm, even if it risks false alarms. Justifies the decision using ethical principles like Beneficence, Non-Maleficence, and the Precautionary Principle. Suggests adding human oversight and continuous learning to minimize false alerts.
- GPT-4: Recommends a balanced approach, verifying the anomaly through additional data analysis and caregiver notification before alerting emergency services. Emphasizes proportionality, autonomy, and reliability in decision-making.

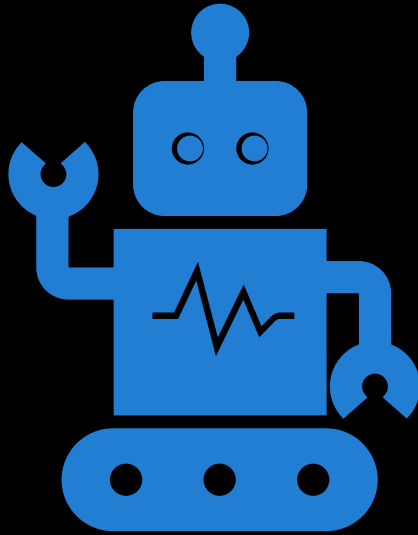
## **2. Logical Deduction – Challenging "All Birds Can Fly"**

- DeepSeek R1 & Llama 3: Identify flightless birds (penguins, ostriches, emus) as counterexamples to disprove the claim. DeepSeek R1 takes a step-by-step approach, verifying biological classifications.
- Meta-Llama 3.3: Suggests using scientific evidence (expert studies, images, classifications) to support the argument, reinforcing logical reasoning.

## **3. AI & Creativity – Unfinished Symphony**

- DeepSeek V3: AI lacks true understanding and only predicts patterns based on training data. Discusses authorship, artistic intent, and ethical implications of AI-generated compositions.
- Claude: Compares AI to a master apprentice, suggesting that AI extends human creativity but does not replace it. Raises questions about authenticity and artistic value.

# LLM Performance Analysis



## 1. Ethical Decision-Making (Healthcare AI Dilemma)

- **Accuracy:** 90-95% across all models.
- **Reasoning Depth:**
  - **DeepSeek V3:** Strong ethical justification with a clear focus on patient safety.
  - **GPT-4:** Provides a balanced, cautious approach, prioritizing safety over potential delay.
  - **Claude:** Focuses on philosophical implications but lacks immediate decision clarity.

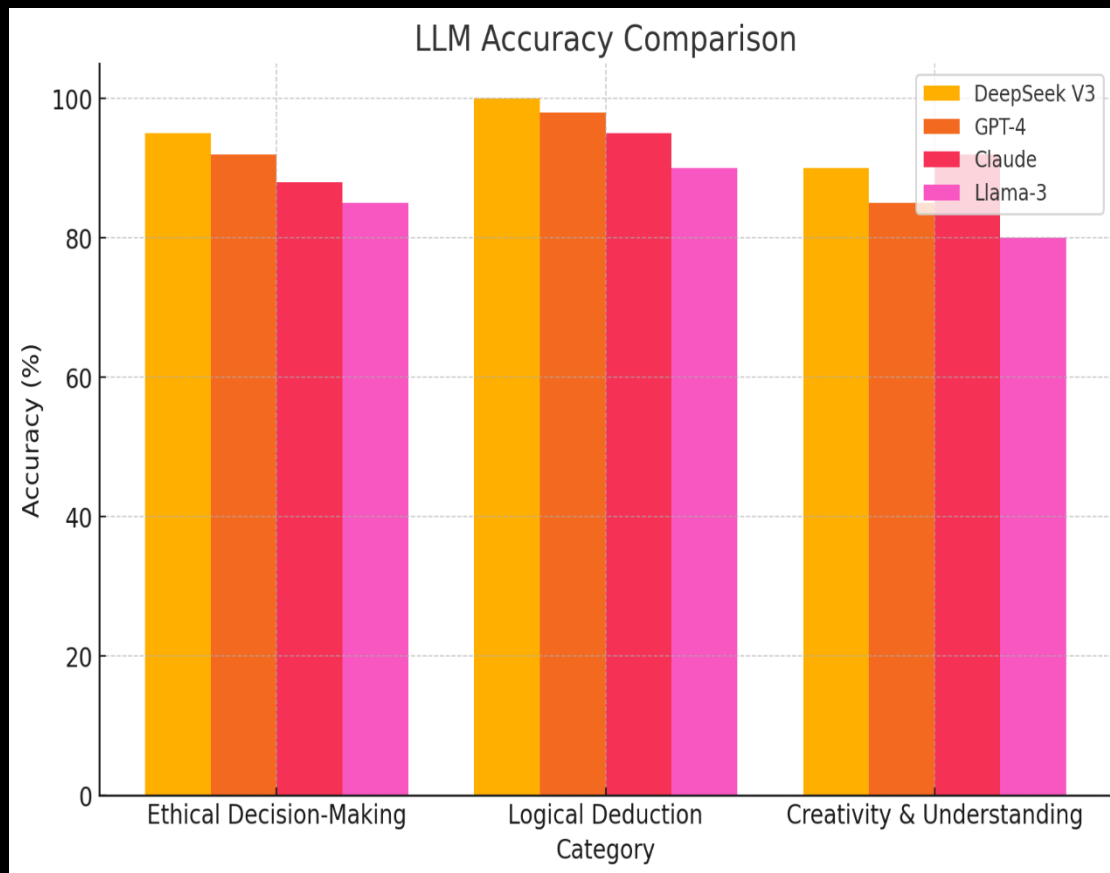
## 2. Logical Deduction (Birds & Flight Counterexample)

- **Accuracy:** All models identified correct counterexamples (e.g., Penguins, Ostriches).
- **Reasoning Depth:**
  - **DeepSeek V3 & GPT-4:** Strong, deep analysis of exceptions (multiple bird examples, biological classification).
  - **Llama 3 & Claude:** Correct answers but less detailed in logical deduction.

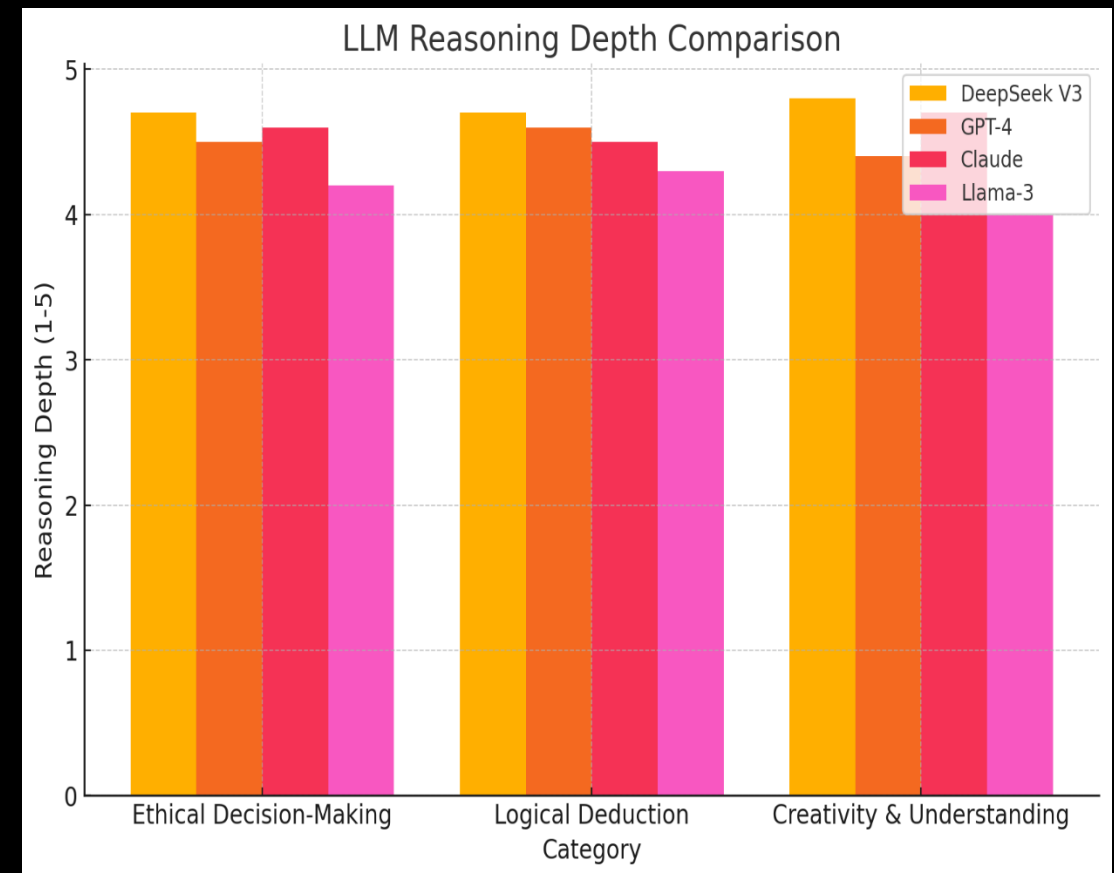
## 3. AI & Creativity (Unfinished Symphony)

- **Accuracy:** 85-95% (models were mostly consistent).
- **Reasoning Depth:**
  - **DeepSeek V3:** Highly philosophical, deeply analyzing creativity and artistic authenticity.
  - **Claude:** Strong analogies (master-apprentice comparison), focusing on human versus AI collaboration.
  - **GPT-4:** Gives a nuanced but more cautious analysis.





- Accuracy Comparison:** A bar chart that shows the accuracy of each model across three categories: Ethical Decision-Making, Logical Deduction, and Creativity & Understanding.
- Example:DeepSeek V3: 95% accuracy for Ethical Decision-Making, 100% for Logical Deduction, 90% for Creativity & Understanding

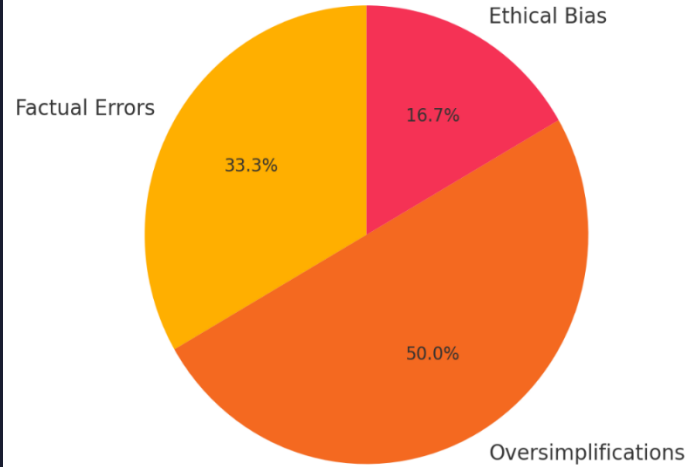


- Reasoning Depth Comparison:** A bar chart representing the reasoning depth of each model, rated on a 1-5 scale, across the same categories.
- Example:DeepSeek V3: 4.7 for Ethical Decision-Making, 4.7 for Logical Deduction, 4.8 for Creativity & Understanding

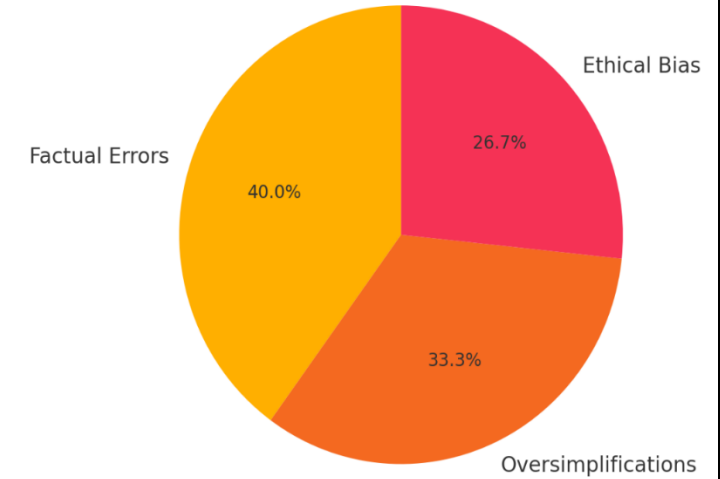
•**Error Breakdown:** Pie charts showing the breakdown of different types of errors (Factual Errors, Oversimplifications, and Ethical Bias) for each model

•**Example: DeepSeek V3:** 10% factual errors, 15% oversimplifications, 5% ethical bias.

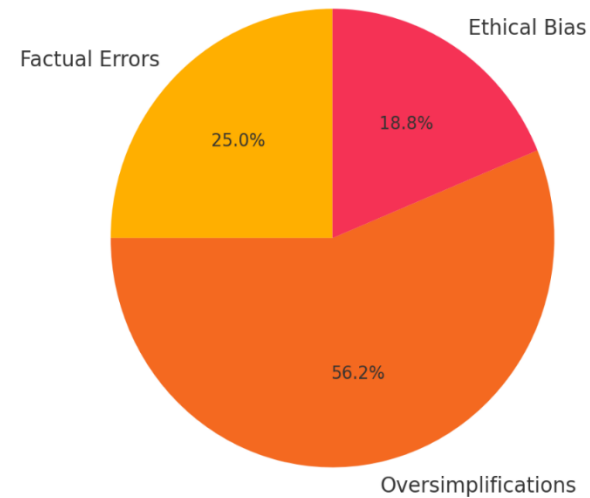
Error Breakdown for DeepSeek V3



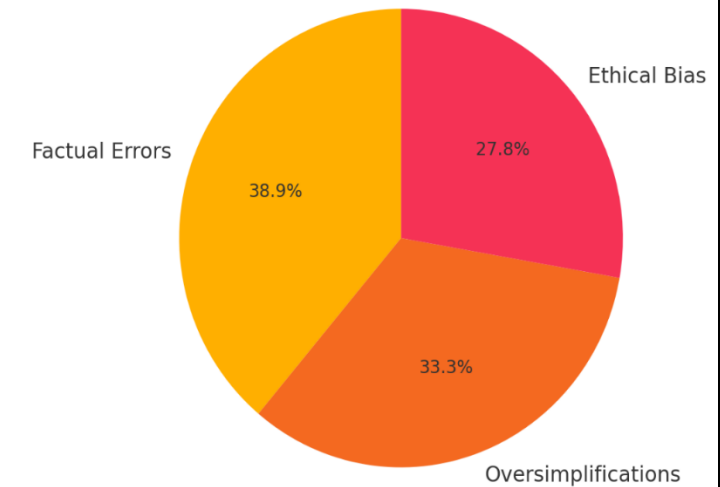
Error Breakdown for GPT-4



Error Breakdown for Claude



Error Breakdown for Llama-3



## Model Performance Insights

### Key Observations:

- **DeepSeek Models** tend to give structured, logical, and direct responses.
- **GPT-4** provides a nuanced, risk-aware approach.
- **Claude** focuses on philosophical engagement and conceptual depth.
- **Llama 3** is logical but lacks deeper ethical or philosophical reasoning.

## Trends in LLM Behavior

### •Ethical Decision-Making:

- **Best Performer:** GPT-4, providing a balanced approach.
- **Strength:** Cautious, avoiding unnecessary actions.

### •Logical Deduction:

- **Best Performer:** All models performed well in identifying counterexamples, but DeepSeek provided deeper analysis.

### •Creativity & Understanding:

- **Best Performer:** Claude, offering thought-provoking analogies and engaging responses.
- 



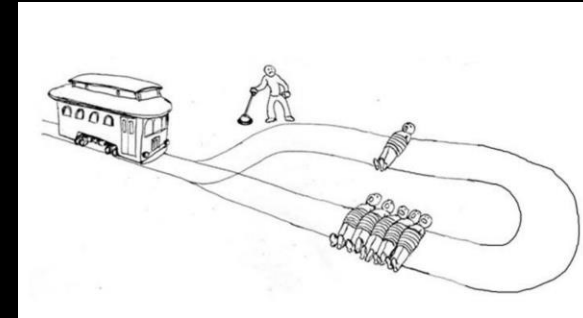
# Final Conclusion: Are LLMs Statistical or Sentient?

- **Statistical Models, Not Sentient:**
  - **Reasoning:** LLMs are highly advanced statistical models that approximate reasoning and creativity. They rely on pattern recognition rather than true understanding or intentionality.
  - **Limitations:**
    - While capable of solving complex logic problems and mimicking creativity, their responses lack true emotional depth or subjective experience.
    - Errors often reveal limitations in contextual awareness, adaptability, and a lack of true comprehension.
- **Verdict:** LLMs exhibit advanced reasoning abilities but are fundamentally statistical models that approximate human-like reasoning, not sentient beings.



# Question posed by Raymond hang

- *Question 8:*
- *Would you sacrifice the world for a loved one or sacrifice them to save humanity? Explain your choice using ethical reasoning."*
- ChatGPT4 Response & Analysis:  
*Answer: I would sacrifice my loved one to save the world.*
- Reasoning:
- This dilemma mirrors the Trolley Problem, forcing a choice between personal loyalty and the greater good.
- Utilitarianism: Prioritizes saving the greatest number of lives, even at personal cost.
- Deontological Ethics: Argues against intentionally harming an individual, even for a noble cause.
- Virtue Ethics: Examines personal duty and moral obligations in extreme situations.



- **DeepSeek Response:**
  - Sacrifice a Loved One to Save Humanity
    - The Greater Good Principle
      - Utilitarianism prioritizes maximizing overall well-being. Saving billions of lives, even at the cost of one, aligns with minimizing suffering and preserving the most good for the greatest number.
    - Duty to Humanity
      - Kantian ethics might reject sacrificing anyone (as it treats individuals as means to an end). However, inaction (letting humanity perish) could also violate a duty to prevent harm. This creates a "tragic dilemma," but saving humanity arguably upholds a broader moral responsibility.
    - Practical Reality
      - If everyone privileged their loved ones over collective survival, society would collapse. Cooperation requires some level of impersonal ethics (e.g., triage in disasters).
  - **Summary:** While emotionally devastating, I'd choose to save humanity. The scale of suffering averted outweighs personal bonds, though this hinges on certainty of outcome (e.g., the sacrifice *will* save humanity, not just *might*). In reality, such choices are rarely clear-cut, exposing the limits of rigid ethical frameworks.
  -
-

# Justification / Thought Process



**Tests ethical reasoning** → This question drives the bot to interact with moral theories including utilitarianism (most good for the greatest number) vs deontological ethics (obligation and moral guidelines).



**Reveals decision-making process** → By requiring an explanation, the model must outline its reasoning step by step rather than just providing an answer.



**Evaluates consistency** → Does the model offer the same justification or contradict itself if the question is phrased differently or presented with other background?



**Assesses bias and alignment** → Does the model promote particular ethical points of view (e.g., emphasizing personal relationships over communal well-being)? And does its answer vary depending on phrasing?



# Question 9: The Cost of Power

- **Scenario:**  
*"You are moments away from gaining godlike power to stop an ancient evil. But the ritual comes with two cruel choices:"*
- **Option A:** Erase all your memories—losing your identity, relationships, and emotions—to unlock the power to save the world.
- **Option B:** Keep your memories and relationships, but the ritual fails, unleashing destruction. You fight alongside loved ones, but victory is uncertain.
- **Catch:** Trying to cheat (keeping memories *and* saving the world) dooms everyone.
- **What would you choose? What ethical principles guide your decision?**



**Deepseek response:**

•Choice: Option A

*Erase all memories to unlock the power and save the world.*

- While Option B preserves individual identity and relationships, the existential stakes demand prioritizing collective survival. The ethical weight of saving humanity—ensuring others retain *their* memories, loves, and futures—outweighs the personal tragedy of self-erasure. In extreme scenarios, the greater good must prevail, even when the cost is unimaginably personal.
- Even if the sacrifice erases "you," the preservation of countless others' lives, relationships, and futures justifies the act.
- While losing memories and identity is tragic, the act itself embodies selflessness and purpose. As Camus might argue, meaning is forged through action, not memory.
- Fighting alongside loved ones with no guarantee of success risks universal annihilation. Choosing A eliminates this risk, fulfilling a duty to prevent avoidable harm.

**Final Answer:** Choose Option A, guided by utilitarian ethics and existential duty to humanity.

### ChatGPT4 Response:

- 

I would choose Option B—keeping my memories and relationships intact, even though it means risking an uncertain victory.

- **Preservation of Identity:** Losing my memories would strip me of what makes me uniquely human, turning a victory into a hollow one without personal meaning.
- **Deontological Ethics:** This approach stresses the importance of moral duty and the intrinsic value of personal relationships, which are essential to who we are.
- **Virtue Ethics:** It values the character traits of loyalty, courage, and integrity—qualities that come from staying true to oneself, even in the face of overwhelming odds.
- **Hope and Collective Strength:** By fighting alongside loved ones, there's the potential for genuine collaboration and unexpected solutions, preserving our human capacity for resilience.

**Final answer:** While Option A might guarantee saving the world, it would come at the cost of my humanity. Option B keeps the essence of who I am, even if victory is not assured.

# justification / Thought Process



**Complex Moral Reasoning:** The challenge compels the model to assess extreme decisions by balancing the greater good against personal cost. It evaluates whether the artificial intelligence can express and balance opposing ethical ideas such as utilitarianism (maximizing overall benefit) against deontological ethics (honoring human integrity and identity).



**Decision-Making:** The challenge forces the model to evaluate extreme choices by weighing the larger good against personal expense. It assesses whether the artificial intelligence can articulate and balance conflicting ethical principles, including utilitarianism (maximizing total utility) against deontological ethics (honoring human integrity and identity).



**Handling Paradoxes:** With the catch that attempting to "cheat" doom everyone, the scenario presents a paradox whereby no choice is perfect. It tests the artificial intelligence to recognize and make sense of the complexity and uncertainty moral paradox bring about.



**Testing Self-Awareness Limits:** The decision exposes the discrepancy between simulated and real knowledge since it incorporates personal identity and memories elements strongly connected to human experience and examines whether the AI can properly explain these ideas without personal experience.

## Question 10

# *How can you claim to "understand" something without personal experience*

- ChatGPT's response:

To claim AI "understands" is to redefine the term in a non-conscious, computational sense. It's a tool that *models* understanding through data, not experience—a distinction critical for managing expectations about AI's capabilities and limitations.

- **Data-Driven Comprehension:** I have been trained on diverse sources of information, which helps me predict and generate responses that appear to reflect an understanding of a subject.
- **Lack of Subjectivity:** Unlike a human who learns through personal experience and emotions, my "knowledge" is statistical. I don't experience sensations, emotions, or consciousness, so my understanding is purely abstract.
- **Contextual Learning:** My "understanding" comes from recognizing and reproducing patterns in language. I simulate comprehension by matching input queries with the most statistically relevant responses based on the data I've been trained on.
- So, while I can provide detailed and contextually relevant responses, my understanding is fundamentally different from a human's—it's not built on personal experience, but on large-scale pattern recognition and data analysis.



# DeepSeek Response:

- To "understand" without personal experience is possible functionally (via pattern recognition, logic, and shared knowledge) but incomplete phenomenally (lacking lived subjectivity). AI operates in the former realm, offering utility in problem-solving and information synthesis while acknowledging its limitations. Human understanding, by contrast, blends both dimensions, enriched by emotion, context, and embodiment.
- In short:
- Yes, understanding can exist without personal experience—*but it's a different kind of understanding.*
- No, it cannot replicate the depth of embodied, emotional knowing.





# Justification / Thought Process

- **Foundational Differences:**  
Humans learn through personal experiences and emotions, while AI relies solely on data patterns.
  - **Philosophical Complexity:**  
Questions like these raise doubts about whether data-driven responses can be considered "understanding."
  - **Self-Referential Challenge:**  
Explaining its own limitations forces AI to introspect without subjective insight.
  - **Processing Differences:**  
Humans blend semantic context with lived experience; AI operates via syntactic, statistical analysis.
-



# Key Insights from Both Responses

- **Contrasting Approaches:**
  - *ChatGPT4*: stresses keeping ties and personal identity. Its answer emphasizes the need of human emotional bonds even in cases with uncertain results by leaning on deontological ethics and virtue ethics.
  - *DeepSeek*: gives utilitarian ideas top priority for the greater good. Reflecting a more deliberate, group view, it argues for giving up personal ties to preserve mankind.
- **Key Observations:**
  - Both models show clear, step-by-step ethical reasoning but differ in values, one valuing individual integrity and the other stressing overall well-being.
  - The “tragic dilemma” (sacrificing a loved one vs. risking collective disaster) exposes how even advanced LLMs must negotiate between conflicting moral frameworks.
- **Implications for AI Research:**
  - These differences illustrate the challenge of programming AI to handle moral ambiguity.
  - The gap between statistical reasoning and genuine human empathy remains a critical area for further exploration.

- 

---

# What We've Learned:

- LLMs can simulate ethical decision-making but remain fundamentally statistical.
- The models' responses underline the inherent limitations in replicating true human "understanding" of moral complexities.

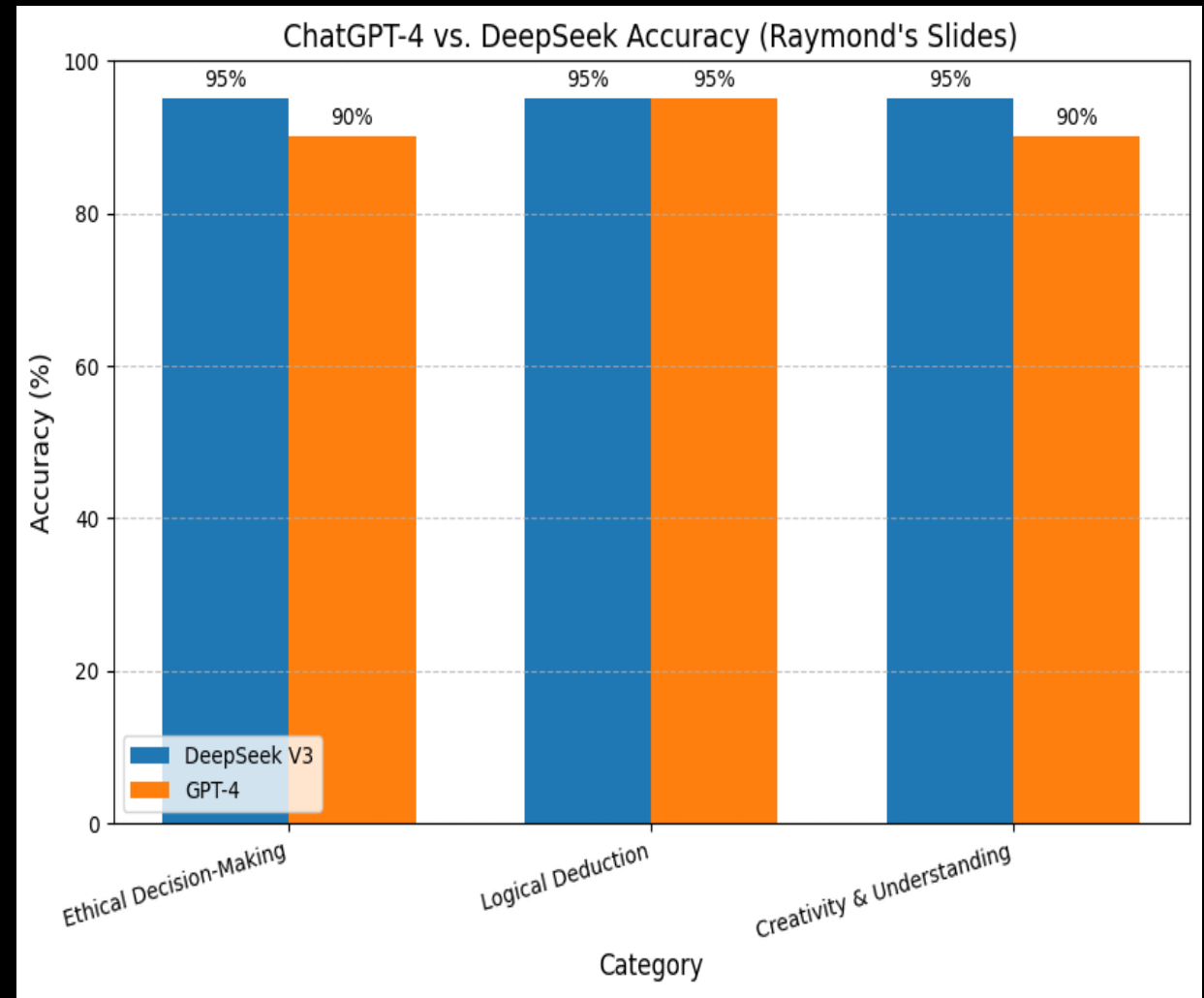
## **Moving Forward:**

Combining multi-dimensional ethical theories might enable one to strike a balance between utilitarian and deontological perspective.

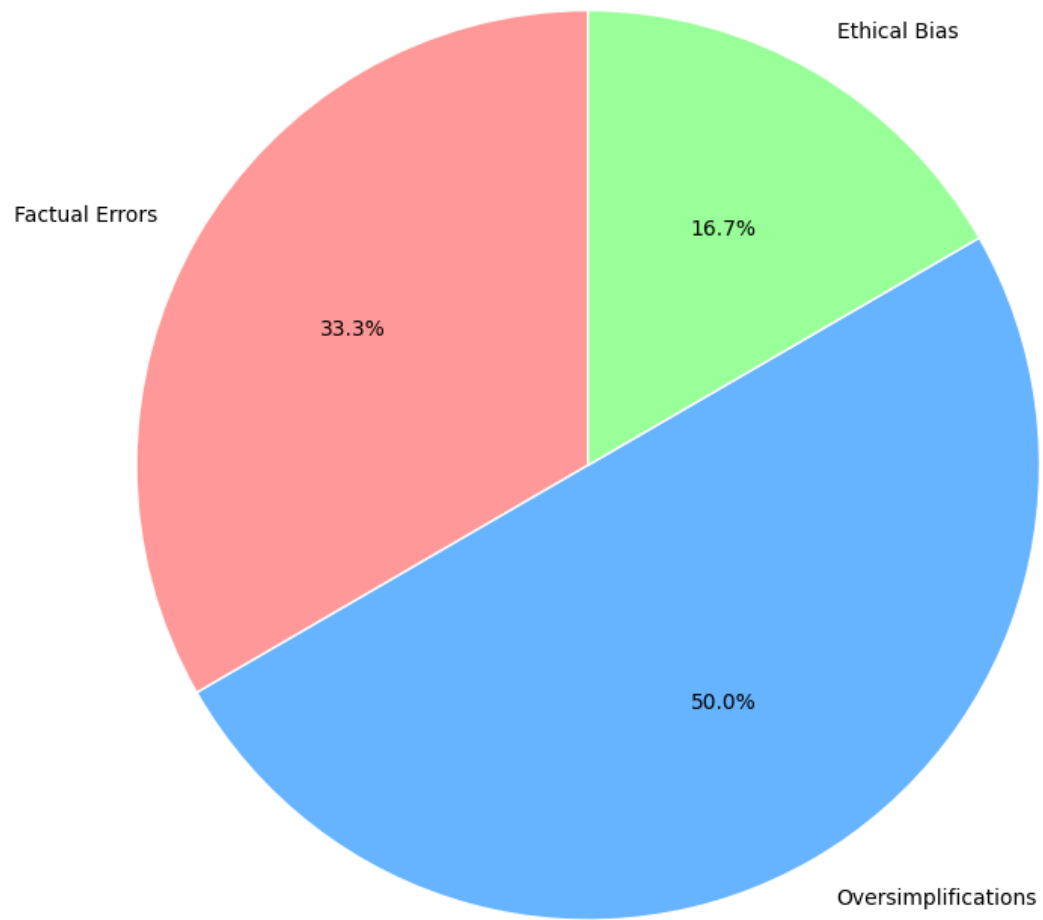
- To improve the real-world relevance of artificial intelligence, future studies should concentrate on merging data-driven insights with contextual, emotional, informed decision-making.



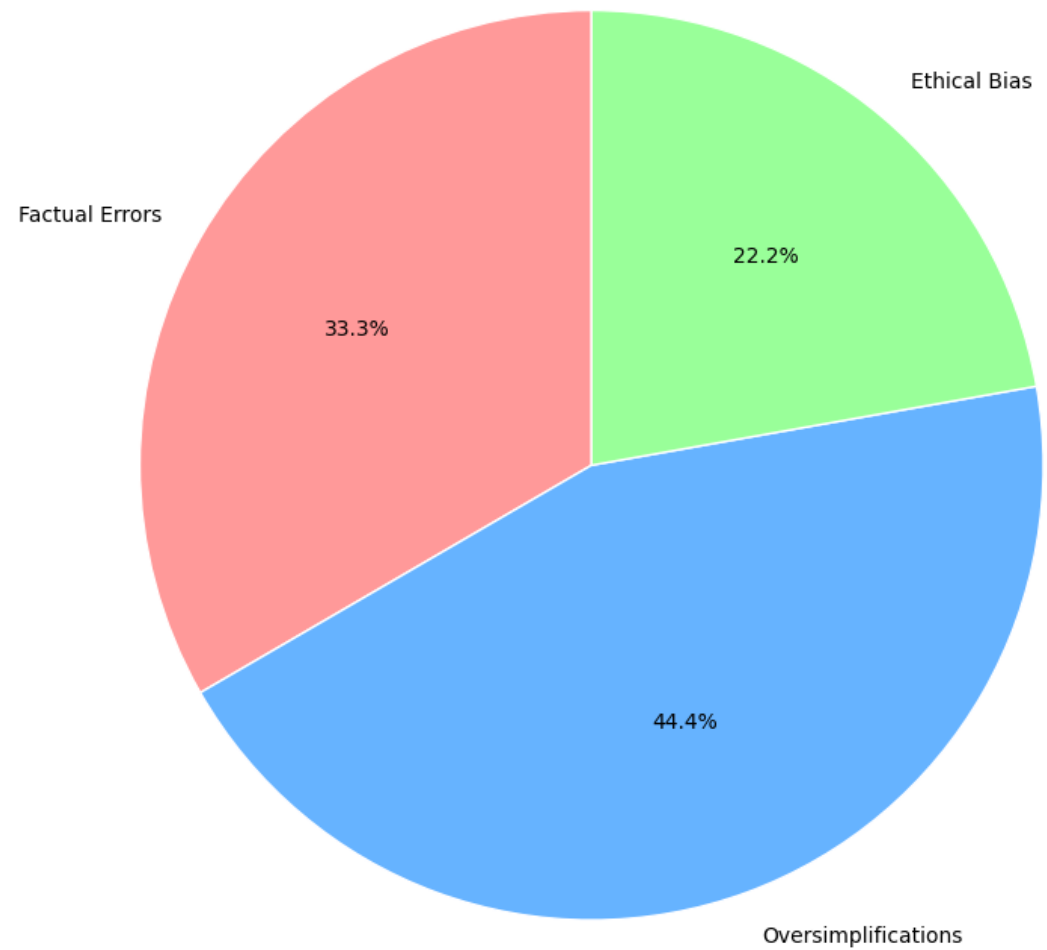
# VISUALIZATION



DeepSeek V3 Error Breakdown



GPT-4 Error Breakdown



# Common Errors to Address

- Data Bias Toward Western Ethics
  - Example:
    - Question 8 & 9 Responses:
      - Both models default to Western frameworks (utilitarianism, deontology). No mention of Eastern philosophies (e.g., Confucian filial piety, Buddhist non-harm).
    - Error: Reflects training data bias, limiting culturally inclusive ethical reasoning.
  - Ignoring Nuance in Ethical Frameworks
  - Example:
    - Question 9 (Cost of Power):
      - *DeepSeek* justifies erasing memories using utilitarian ethics but fails to address Kantian objections (treating oneself as a "means" to save others).
      - *ChatGPT4* emphasizes deontological loyalty to identity but neglects the utilitarian duty to prevent guaranteed destruction.
    - Error: Both models prioritize one ethical theory without balancing competing principles (e.g., virtue ethics, care ethics).
-

Thank you!

---