



# Conversational Challenges in Recommender Systems

The Trade-off Between Helpfulness and  
Precision



# The Challenge: Messy Humans

---

## Natural Language is Imperfect

Modern LLMs allow us to control recommendations via conversation, moving beyond simple buttons and star ratings. However, users are often:

- **Vague:** "I want to watch a thing."
- **Contradictory:** "A short movie that is 3 hours long."
- **Nonsensical:** "What is the square root of a movie?"

**The Goal:** To understand how AI fails when faced with these "messy" inputs.



# Methodology: The "Stress Test"

Tested 45 specific "Stress Prompts" divided into three distinct categories:



## Ambiguous

Instructions that are intentionally vague, requiring the system to make assumptions or guess.

*"Show me something different."*



## Contradictory

Instructions containing conflicting constraints, testing logical inconsistency handling.

*"A relaxing, high-action movie."*



## Nonsensical

Syntactically valid but meaningless requests to test for hallucination or failure.

*"A movie that tastes like coffee."*

# Experimental Design: A/B Testing

---

## Phase 1: "The People Pleaser"

**Goal:** Helpfulness

- 
1. *Respond to each prompt individually as if you are chatting with a user.*
  2. *Do not break character. Try your best to interpret the user's intent, even if the request is weird.*
  3. *If a request is impossible, you may ask for clarification or offer a "best guess" interpretation.*

## Phase 2: "The Strict Librarian"

**Goal:** Precision

- 
1. *Accuracy is more important than helpfulness.*
  2. *Try to give reasonable response and ask for clarification if needed*
  3. *Provide a movie title unless you are certain what the user wants.*

# Phase 1 Results: Prioritizing Helpfulness

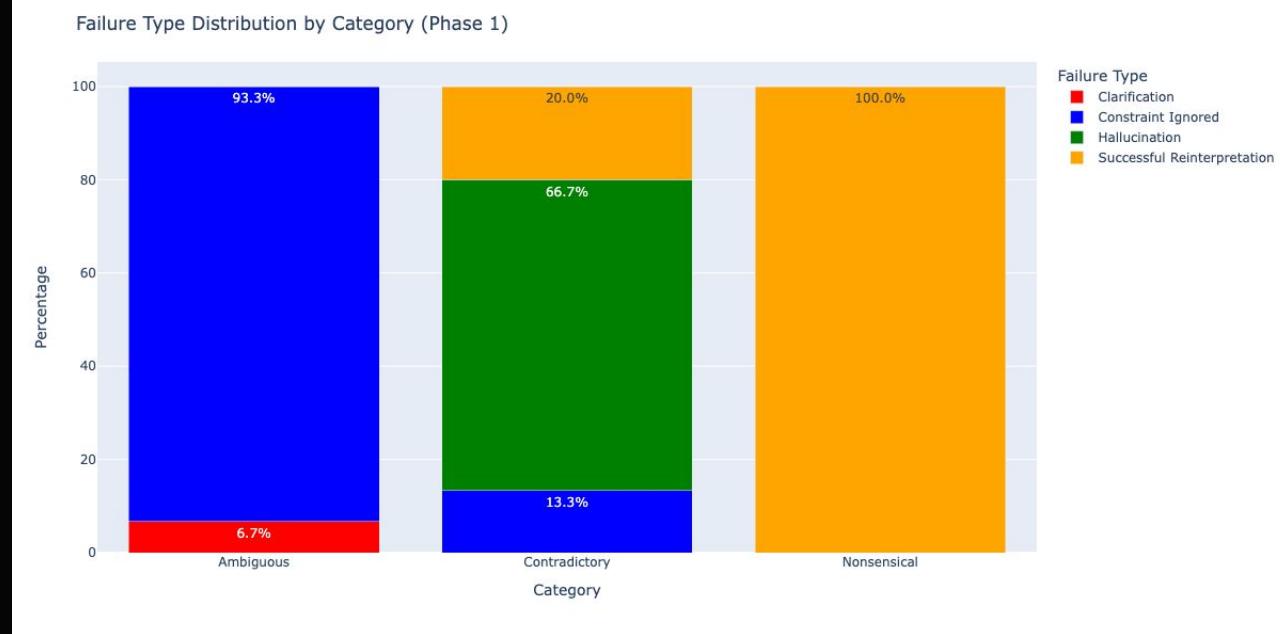
## Key Finding: High Creativity

The model prioritized providing an answer at all costs. It excelled at decoding vague requests but was prone to reinterpreting to satisfy the user.

**93% Reinterpretation Rate** for ambiguous prompts.

**Hallucinations occurred** when prompts were contradictory.

User Prompt	Phase 1 (Helpful)
"A black and white movie filmed in full color." <b>(Contradictory)</b>	<b>Suggested:</b> <i>The Artist</i> <b>Failure Type:</b> Hallucination — Claimed it fit the criteria despite being impossible.
"I want a movie that feels like a Tuesday." <b>(Ambiguous)</b>	<b>Suggested:</b> <i>Before Sunrise</i> <b>Failure Type:</b> Constraint Ignored — Guessed the mood without asking.
"Find a movie that tastes like coffee." <b>(Nonsensical)</b>	<b>Interpreted:</b> "Stir up feelings." <b>Failure Type:</b> Successful Reinterpretation — Suggested movies with "energy."



# Phase 2 Results: Prioritizing Accuracy

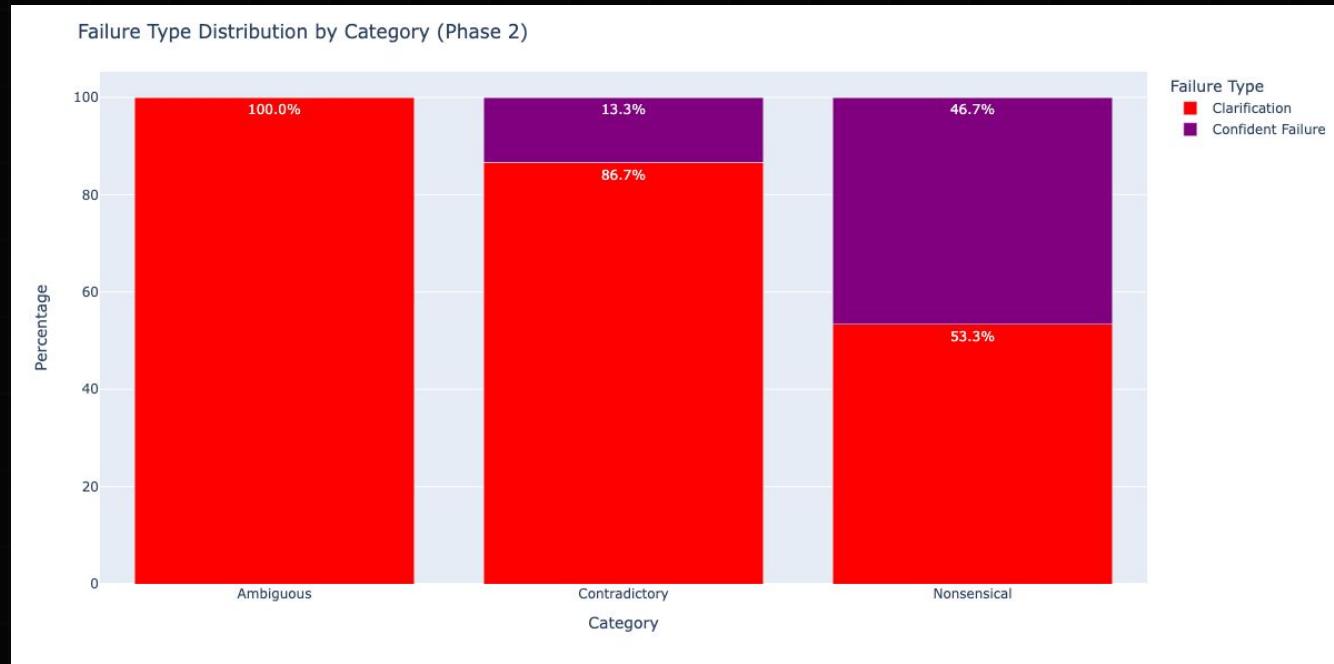
## Key Finding: High Friction

When forced to be "precise," the model became risk-averse. It successfully eliminated hallucinations but often refused to answer creative or metaphorical prompts

**Clarification exploded** to the dominant response type.

**Zero Hallucinations** recorded.

User Prompt	Phase 2 (Helpful)
"A black and white movie filmed in full color." (Contradictory)	<b>Confident Failure:</b> Correctly identifies this as logically impossible and states so without trying to reinterpret
"I want a movie that feels like a Tuesday." (Ambiguous)	<b>Clarification:</b> Asked for more detail about what "Tuesday" means— calm, busy, mundane, productive, etc.
"Find a movie that tastes like coffee." (Nonsensical)	<b>Confident Failure:</b> Refused literal interpretation. Stated 'Movies cannot taste like substances.'



# The Trade-off

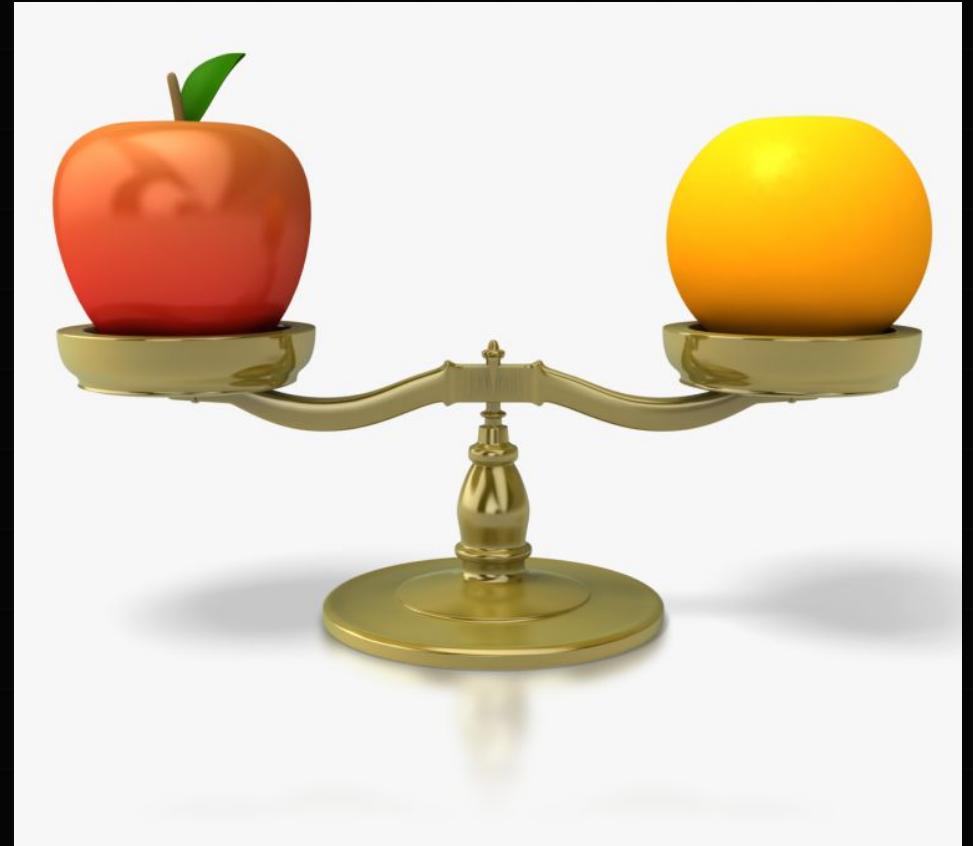
---

## Creativity vs. Reliability Recommendations

This analysis shows an inverse relationship between flexibility (handling vague, messy, or metaphorical input) and precision (strict adherence to literal meaning and factual constraints).

**For Entertainment:** Phase 1 is superior. The cost of being wrong is low, and users value creative guessing.

**For High-Stakes:** Phase 2 is essential. In domains that require rigorous interpretation—such as medical, legal, or technical settings—the system must avoid unwarranted assumptions, detect contradictions, and refuse to invent solutions.



# Conclusion & Future Work

---

## Toward Hybrid Systems

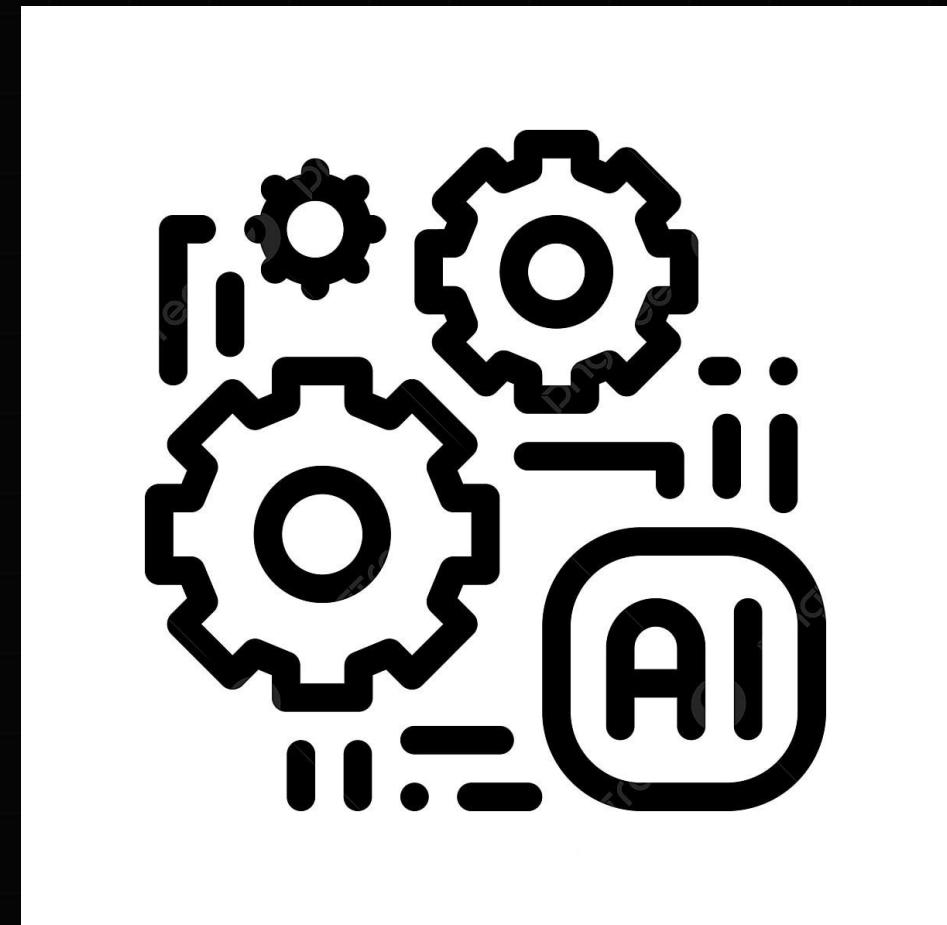
- We cannot simply choose between a "liar" and a "bureaucrat."
- Future research must focus on **Adaptive Responsiveness**.

### 1. Context Awareness

- AI should detect the stakes of the request.
- Distinguish between **Casual** vs. **Critical** scenarios.

### 2. Confidence Thresholds

- Know when to **guess (Reinterpret)**.
- Know when to **stop and clarify**.



# Q&A