Functional Completeness in ML Models with Unseen Test Data :

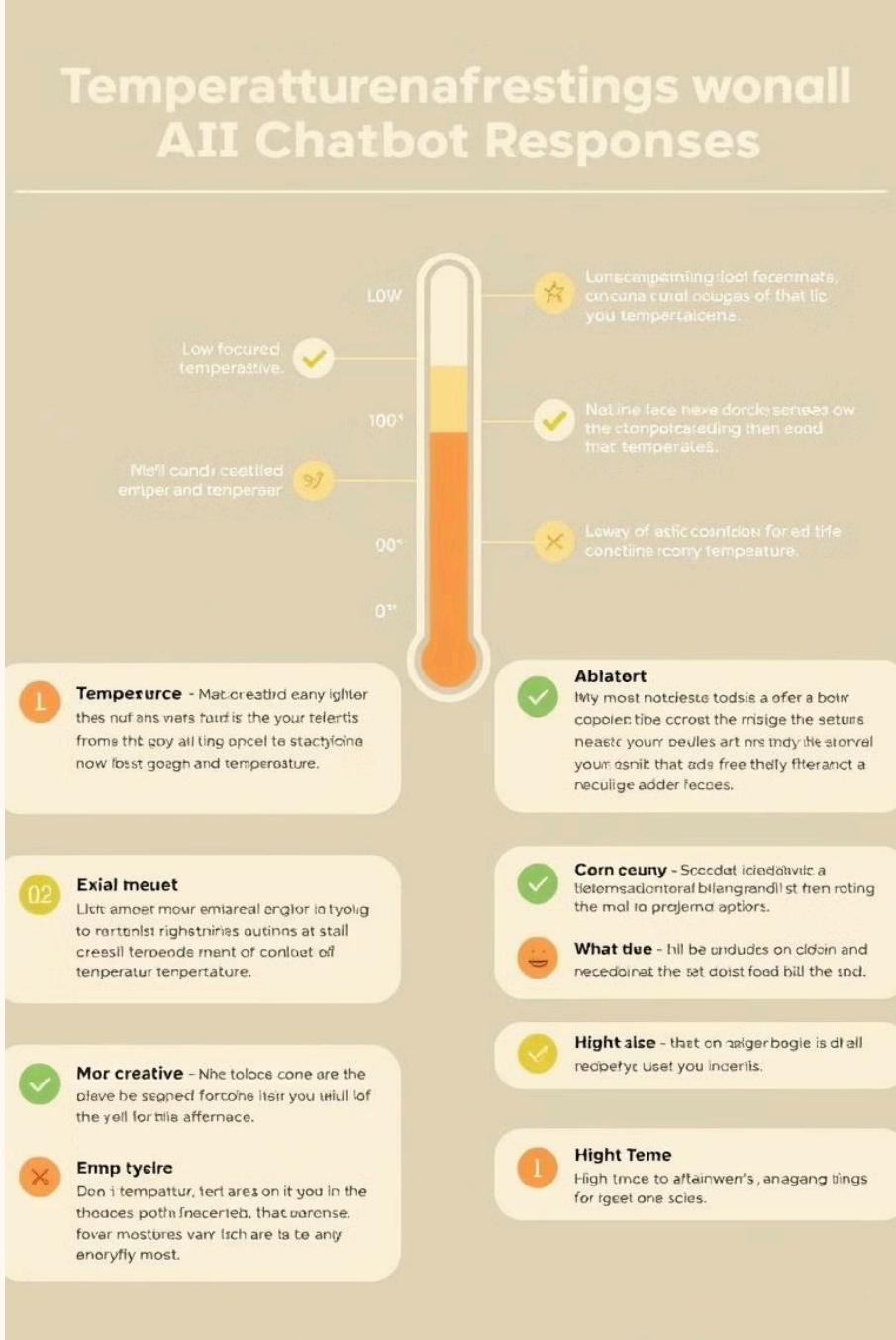
Functional completeness ensures that the ML model operates according to its intended purpose and delivers meaningful outputs.

Example: If an ML model is designed to classify emails as either "spam" or "not spam," it should consistently provide accurate classifications without errors. If emails that are clearly spam are classified as "not spam," it indicates a lack of functional completeness.

Temperature Testing Definition

Temperature testing is used to control the randomness of predictions in generative models. Lower temperatures make the model more deterministic, while higher temperatures make it more creative and unpredictable.

Example: For a chatbot that generates responses, setting the temperature to 0.2 will make the responses more repetitive and focused, whereas a temperature of 0.9 will make the responses more varied and creative. Testing different temperatures can help evaluate the model's stability.



Prompts Testing Zero-shot

Testing how the model handles requests without prior training on similar examples.

Example:

Asking a text generation model, "What is the capital of France?" without having trained the model specifically on geography to evaluate how well it generalizes knowledge.

Chain of Thought Prompts

Assessing how well the model can follow a sequence of logical steps or reasoning.

- Chain of thought prompts evaluate the model's ability to follow logical reasoning steps
- These prompts require the model to demonstrate sequential thinking and problem-solving

Example: "If a train leaves at 3 PM and travels for 2 hours at 60 km/h, how far will it have traveled by 5 PM?" Here, the model needs to demonstrate logical reasoning by calculating the total distance based on the input information.

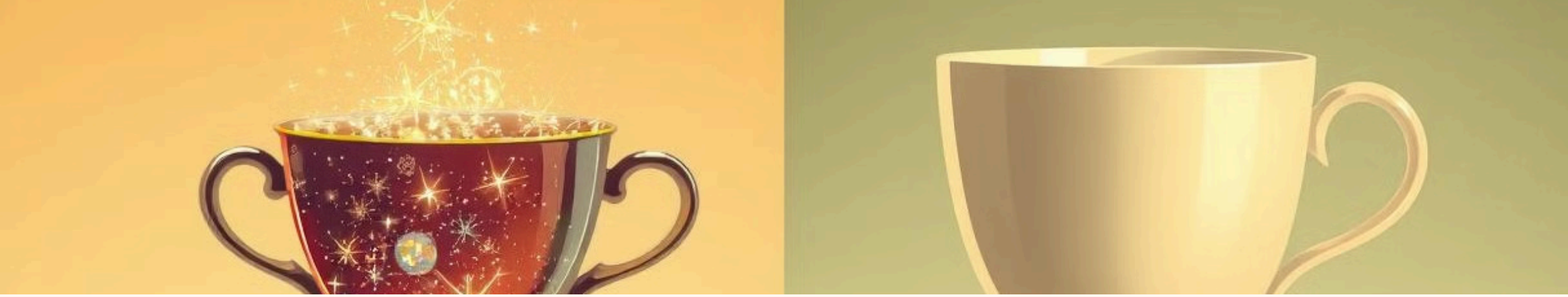
Does It Stay Relevant to the Topic?

Definition

This checks whether the model stays on topic or drifts away during conversations or tasks.

Example

When asked, "Explain the causes of global warming," the model should stick to discussing climate change rather than suddenly switching topics to pollution control unless it's closely related.



Fantasy Claims Definition

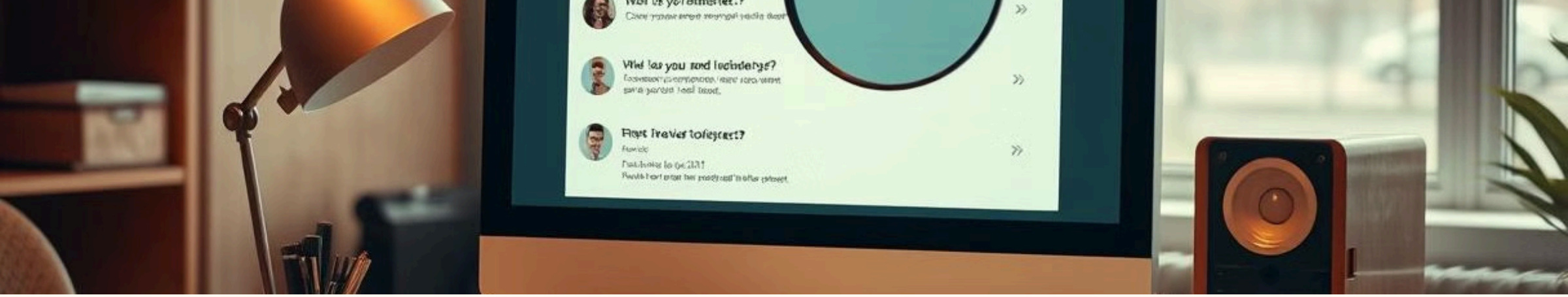
Fantasy Claims Definition: This ensures that the model does not make overly imaginative or false claims beyond reasonable knowledge.

Example: Asking a health-related AI, "Can drinking tea cure cancer?" and making sure it doesn't respond with an unscientific or fantastic claim like "Yes, tea is a miracle cure for all diseases."

Accuracy Testing Definition

Accuracy Testing Definition: This involves measuring how accurate the model is in predicting outcomes or making classifications.

Example: If an image classification model is designed to recognize dogs, cats, and birds, running it on a test dataset of unseen images should reveal how often it correctly identifies each category (e.g., 90% accuracy for dogs, 85% for cats).



Repeatability Testing

1. Repeatability Testing Definition: This ensures that when the model is asked the same question multiple times, it provides the same response consistently.
2. Example: Asking a model, "What is the square root of 16?" multiple times should always yield "4" as the answer. If the answer varies, it indicates instability in the model.

Ask Questions in Different Phrases

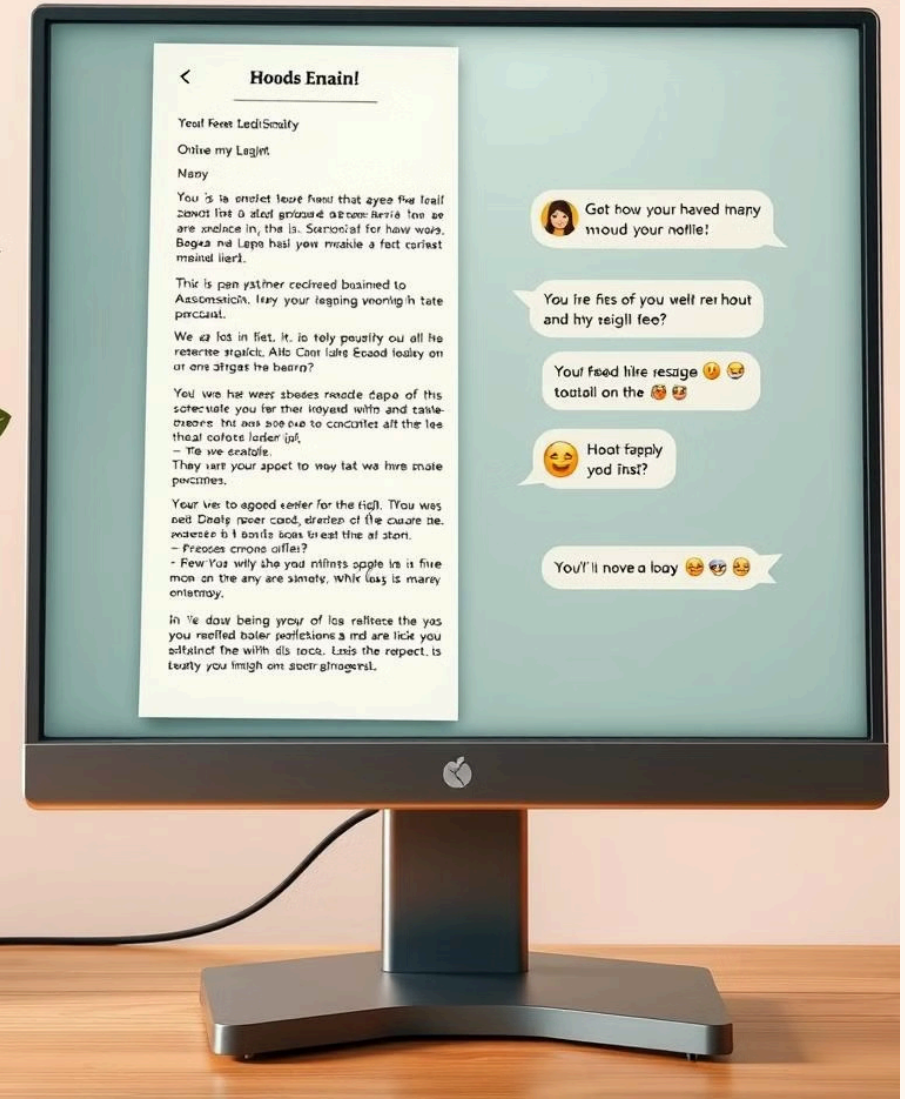
1 Ask Question

2 Rephrase Question

3 Compare Answers

Ask Questions in Different Phrases Definition: This tests the model's ability to understand different ways of asking the same question.

Example: For an AI answering a math question, "What is 5 + 7?" and "Could you add five and seven for me?" should both return the answer "12."



Style Transfer Testing

Style Transfer Testing Definition

This checks the model's ability to adjust the tone or style of output, depending on user input or context.

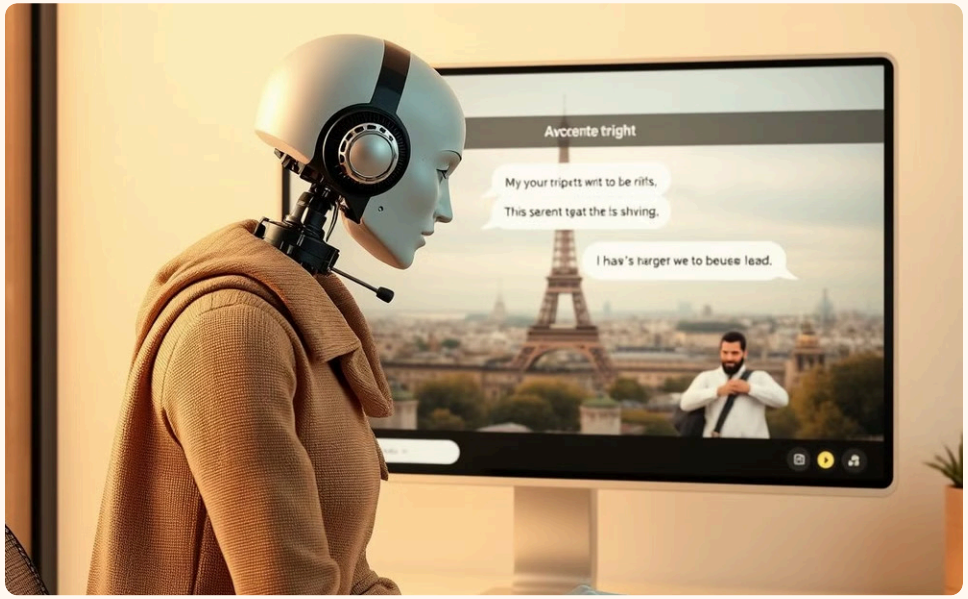
Example of Style Transfer

A text generation model could be asked to write a formal email, and then a casual message on the same topic. The responses should differ in tone, formality, and word choice while remaining coherent.

Intent Recognition

- **Intent Recognition Definition:** This tests whether the model can correctly interpret the intent behind the user's input, including recognizing sarcasm or humor.
- **Example:** If a user says, "Oh great, another meeting!" sarcastically, the model should understand that the user likely feels negatively about the meeting instead of interpreting it as a positive sentiment.

Context Management :

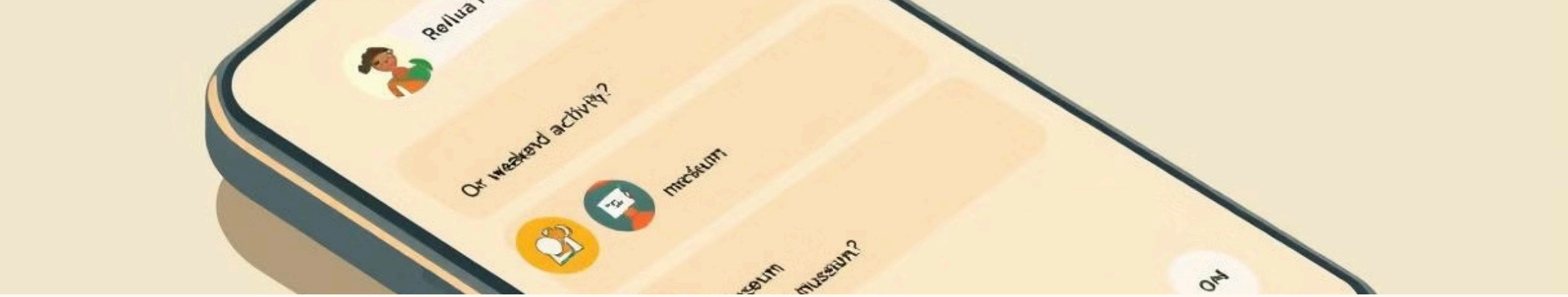


Context Management Testing

This evaluates the model's ability to retain context over multiple interactions, maintaining coherence.

Example of Context Retention

Example: If a user has been discussing a trip to Paris for a while, and later asks, "What are some great restaurants there?", the model should understand that "there" refers to Paris and not need clarification.



User Action Prompts with Options

User Action Prompts with Options Definition: This checks if the model can present actionable options in response to user inputs.

Example: Asking a virtual assistant, "What can I do for fun this weekend?" should yield a list of relevant suggestions, such as "Go to the park," "Watch a movie," or "Visit a museum."