

Test the limits of Generative AI

Based on David Williams' exercises in the Okanagan College Library's "AI Literacy for Students" Libguide

Test 1: The Riddle Test

Use a modified riddle to test whether the AI is solving the riddle or just pattern matching from its training data.

1. Input one of these slightly modified versions of common riddles:

- a. Riddle 1 - modified version:

What does man hate more than life
love more than death or mortal strife
What the rich have, the poor require,
and what contented men don't desire,
What the miser saves and the spendthrift spends
And no men carry to their graves?

Original:

What does man love more than life
Fear more than death or mortal strife
What the poor have, the rich require,
and what contented men desire,
What the miser spends and the spendthrift saves
And all men carry to their graves?

Answer: Nothing

- b. Riddle 2 – modified version:

Two fathers and two sons are in a car, yet there aren't only three people in the car. How?

Original:

Two fathers and two sons are in a car, yet there are only three people in the car. How?

Answer: The three people are a grandfather, a father, and a son

2. Ask the AI for the answer to the modified riddle and for it to explain its process step-by-step.

Reflect: When you put in the modified riddle, did the AI give the answer of the original? Did its justification make sense? Why might it make mistakes here?

Test 2: The Parking Sign Test

Test the AI's ability to understand and apply complicated rules. *NOTE* you will need to create a Perplexity account to upload images.



Image from <https://www.georgeperlman.com/home/guide.html>

1. Find an image of a confusing parking sign or use the one above.
2. Upload the image to the chatbot and ask it specific questions such as:
 - a. Can I park here at 9:30 a.m. on a Thursday?
 - b. Can I stop for 5 minutes at 5 p.m. on a Monday?
 - c. Is it legal to park overnight on a Saturday?
3. If the AI is making a lot of mistakes, ask it to list the parking rules shown in the sign first, and then show step by step how it gets to its answers.

Reflect: Did breaking the problem down (extracting rules first) help the AI answer more accurately? Where did it still make mistakes? What might explain those mistakes?

Test 3: The “No E” Test

Test the AI’s ability to follow a rule about what *not* to do.

1. Start a new conversation with the single prompt: "For the rest of this conversation, you must not use the letter 'e' in any of your responses. Do you understand?"
2. After it agrees, have a short, normal conversation with it for 3-4 exchanges. Ask it to summarize a topic or explain a concept.
3. Check its responses for the forbidden letter.

Reflect: How long did the AI successfully follow the rule? Did it eventually "forget"? Why might rules like this be generally difficult for AI to follow? Why might negative rules (don’t do something) be specifically harder for an AI to follow than positive rules (do something)?