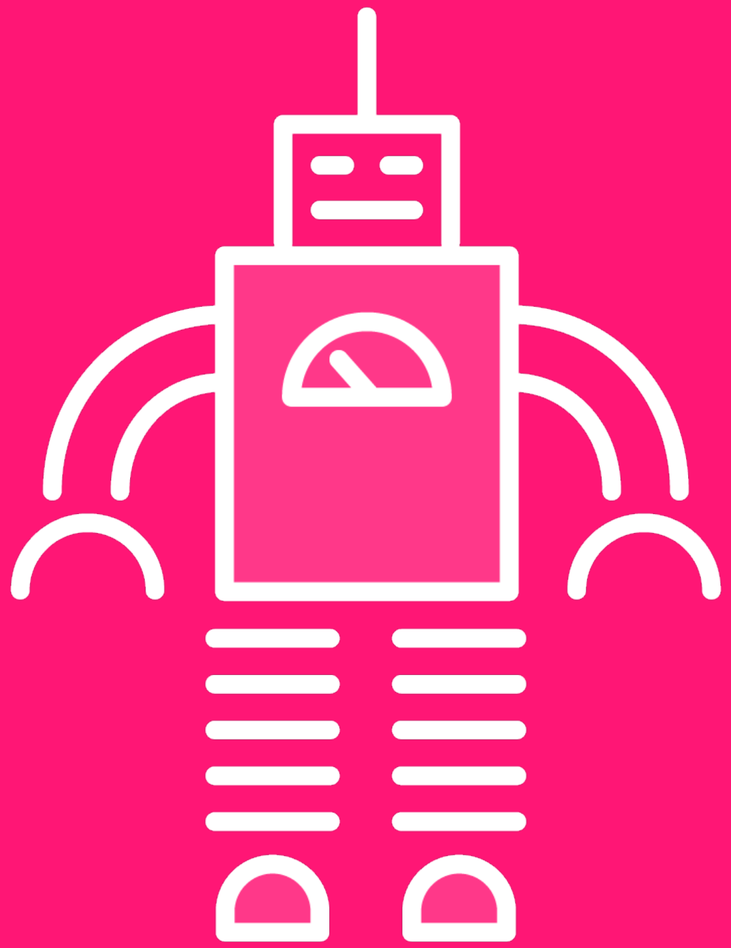
Evaluating Prompt Performance

Amber Israelsen

Trainer | Developer

www.amberisraelsen.com



So, how did I do?

Module Overview

Objective metrics

Subjective metrics

Evaluation techniques

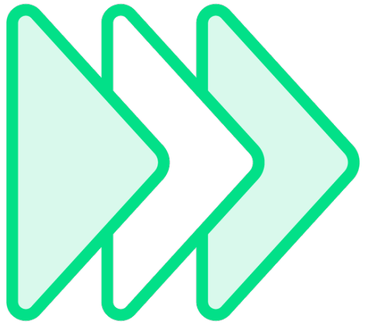
- Surveys and interviews
- A/B testing

Example of adjusting parameters for better results

Objective Metrics

Three Objective Metrics

Accuracy

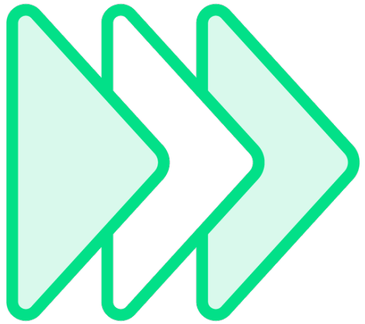


Speed

Relevancy

Three Objective Metrics

Accuracy



Speed

Relevancy

Two Aspects to Accuracy

Factual Correctness

Q: “How long does it take to boil an egg?”

A: “It typically takes about 9-12 minutes to boil an egg.”

Semantic Correctness

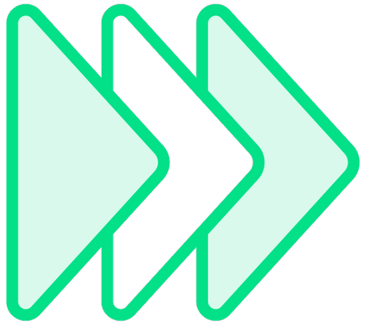
Q: “How long does it take to boil an egg?”

A: “An egg is a food product produced by poultry.”

Factually correct, but semantically incorrect because it does not answer the question

Three Objective Metrics

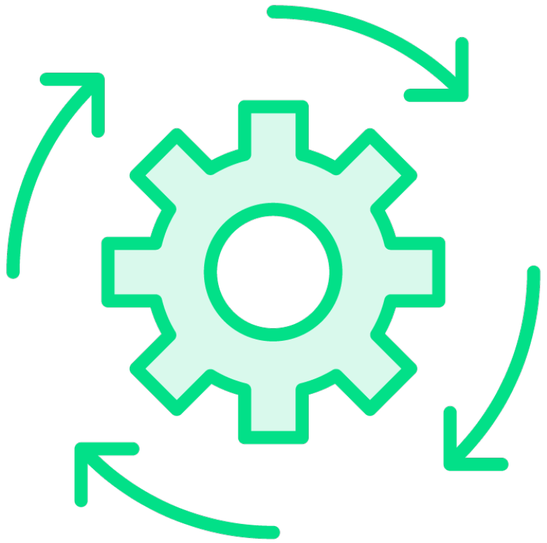
Accuracy



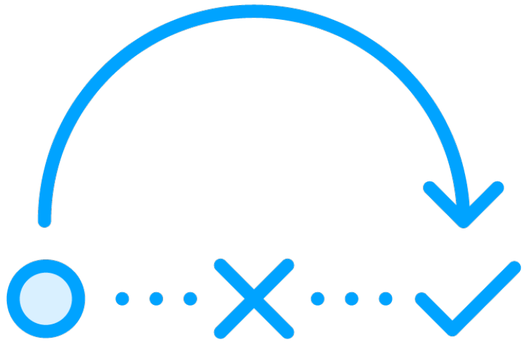
Speed

Relevancy

Measurements of Speed



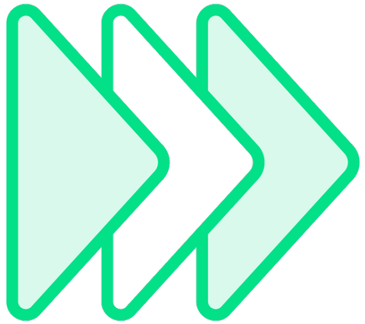
Processing speed
(largely infrastructure)



Response speed
(includes network latency
and other delays)

Three Objective Metrics

Accuracy

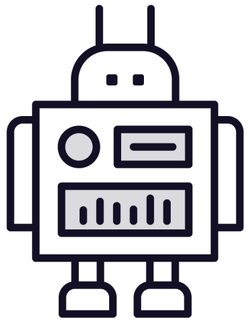


Speed

Relevancy

Give me three book recommendations

Sure, here are three movies you might enjoy...



Subjective Metrics

Subjective Metrics

Human opinions on coherence, tone, and clarity

Coherence

Whether a response makes logical and linguistic sense in its own right

```
Q: What is the capital of France?
A: The capital of France is Paris.
```

Coherence

Whether a response makes logical and linguistic sense in its own right

```
Q: What is the capital of France?
A: The capital of France is Paris.
```

```
Q: Who wrote 'Pride and Prejudice'?
A: Jane Austen wrote 'Pride and Prejudice'.
```

Coherence

Whether a response makes logical and linguistic sense in its own right

```
Q: What is the capital of France?
A: The capital of France is Paris.
```

```
Q: Who wrote 'Pride and Prejudice'?
A: Jane Austen wrote 'Pride and Prejudice'.
```

```
Q: How does photosynthesis work?
A: Green refrigerator fast the elephant.
```

Tone

The perceived emotional quality or “mood” of a message

Tone

Prompt: I just got promoted and I'm really excited!

A: That's fantastic news! Congratulations on your promotion. You must have worked really hard for this achievement. How are you planning to celebrate?

Tone

Prompt: I just got promoted and I'm really excited!

A: That's fantastic news! Congratulations on your promotion. You must have worked really hard for this achievement. How are you planning to celebrate?

A: Promotions involve more responsibilities. It could be more stressful.

Clarity

How understandable or clear responses are

Clarity

Prompt: What's the weather like today?

A: It's sunny and 75 degrees.

Clarity

Prompt: What's the weather like today?

A: It's sunny and 75 degrees.

A: The situation involving the atmospheric conditions in your geographical locale at this current juncture in the temporal continuum is characterized predominantly by the absence of significant cloud cover and a temperature registering at approximately 75 on the Fahrenheit scale.

Evaluation Techniques

Surveys and Interviews

Did the AI understand your prompt correctly?

Was the response AI gave relevant to your prompt?

Was the AI's response easy to understand?

Did the AI provide a complete answer to your prompt, or did it miss anything?

Surveys and Interviews

Did you find the AI’s response helpful?

Was the AI’s response delivered in an appropriate tone?

How satisfied are you with the speed of the AI’s response?

Did you feel the conversation with the AI flowed naturally?

A/B Testing for a Customer Service Bot

Group A	VS	Group B
Uses the prompt: “I’m sorry to hear you’re not satisfied with your purchase. Could you please tell me the order number and the reason for the return?”		Uses the prompt: “I understand you want to return an item. Can I have the order number and the specific issue with the product?”

- Which prompt led to more successful return transactions?
- Which prompt resulted in shorter conversation lengths (indicating possibly smoother interactions)?
- Which prompt received higher customer satisfaction scores in post-interaction surveys?

Fine Tuning with OpenAI APIs

Fine tuning guide

platform.openai.com/docs/guides/fine-tuning

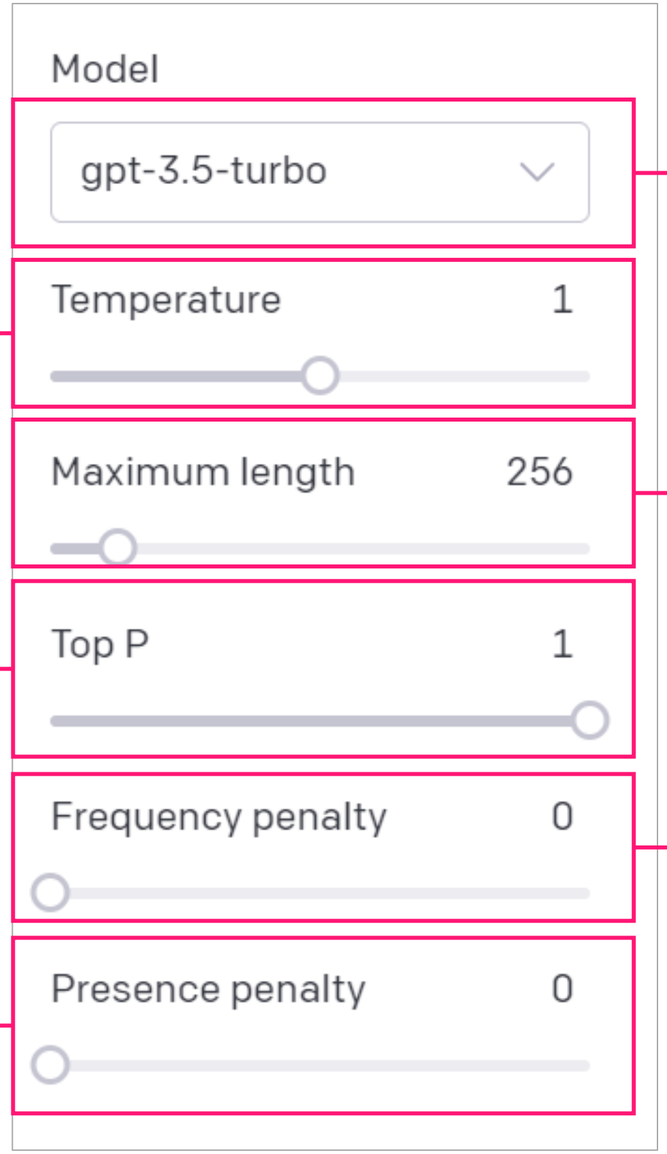
Example: Getting Better Results by Adjusting Parameters

About Parameters

The higher the temperature, the more variable/creative the responses

Controls diversity; 0.5 means half of all likelihood-weighted options are considered

Controls likelihood of talking about new topics



The model selected will determine the parameters available

Controls the length of the response (and cost)

Controls likelihood of repeating the same line verbatim

Example

Prompt:
The inventor of the iPhone is

Module Summary

Module Summary

There are a variety of metrics that can be used to evaluate prompt performance

- Objective metrics
- Subjective metrics
- Evaluation techniques

Adjust parameters to achieve better results

Up Next:

Using Advanced Prompting Techniques