Pipeline Demo

Below are some simple visualizations based on a sample dataset of 1,000 LLM prompts. Insights here are pretty simplistic, but that's mostly because we're making visualizations in the absence of business context.

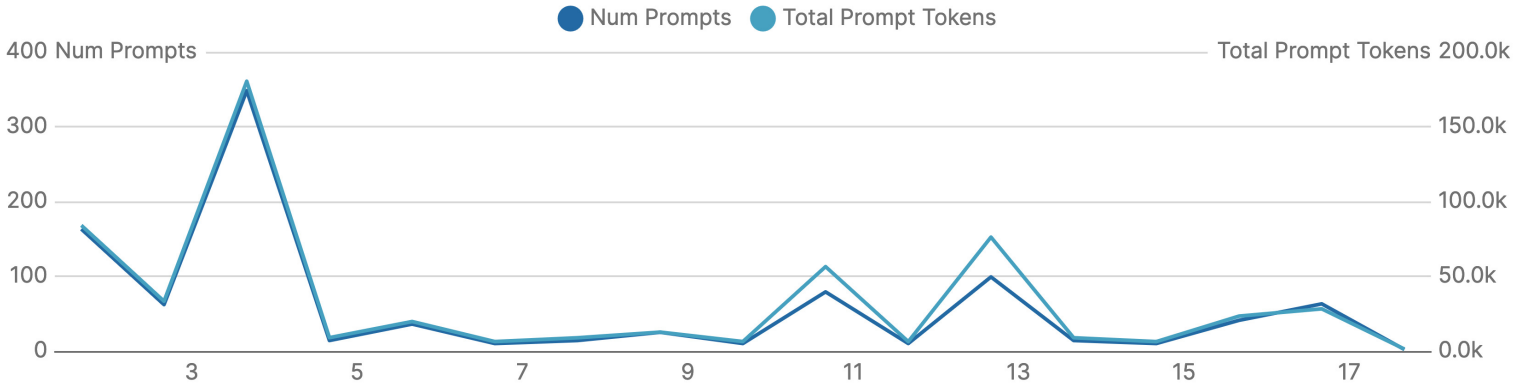
▼ Why Evidence?

We chose Evidence as our data visualization tool here, to choose something where the reader could see how this could be extended to a broader platform (instead of one-shot analyses like Python notebooks), but also remain self-contained (as opposed to requiring users of this demo to sign up for Looker or Mode or something, or to have every user of this demo hitting my personal Google Looker Studio account).

Overall Prompt Volume

With the exception of a couple of days (the 11th and the 13th), prompt volume looks pretty tightly correlated with token count.

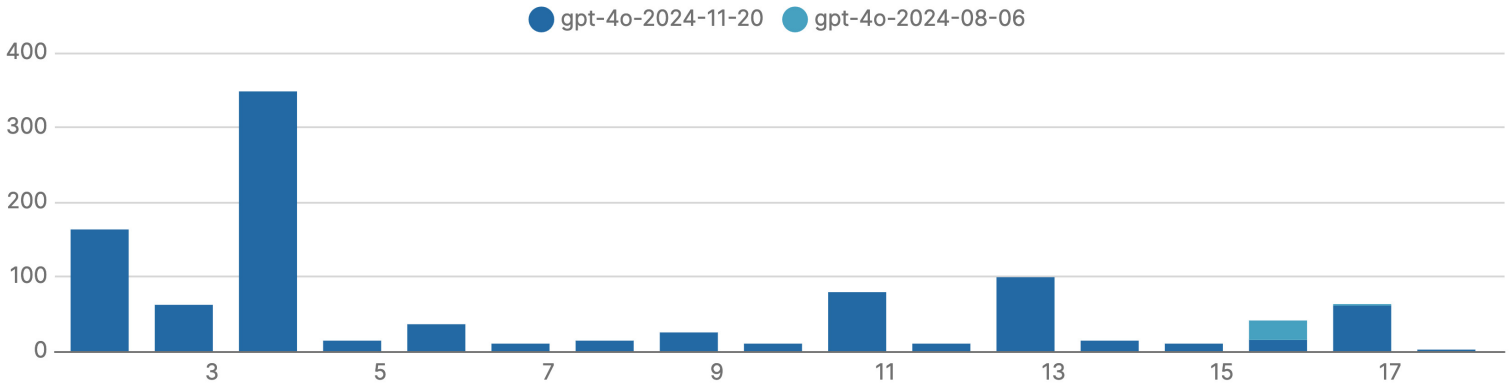
Prompt Volume



Prompts by a few different slices

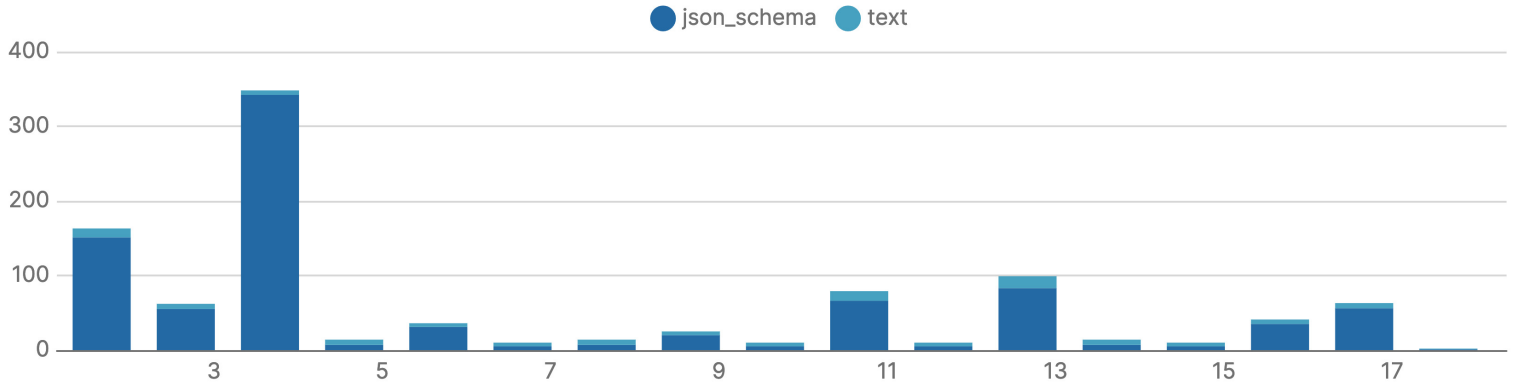
What happened on the 16th? Looks like an older model was run for some reason.

Prompts by Model



Most generation requests seem to be for a JSON schema, with a consistent minority of requests coming in for text.

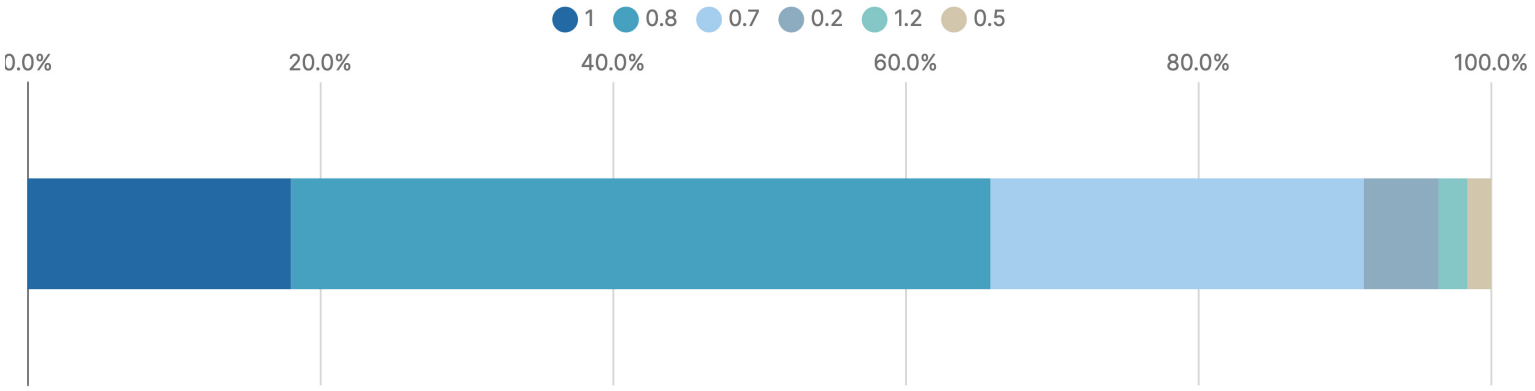
Prompts by Type



Prompts by Temperature

Seems like a temperature of 0.8 is the most popular setting.

Prompts by Temperature



Latency (Time to first token)

The vast majority of prompts have a first token returned in less than a couple of seconds.

Histogram: Latency