Preliminary Chatbot Results

Background

On May 1, the web team released the retrieval-augmented generation (RAG) chatbot on book pages that were frequently visited. The books are Deadly Anti Science, Political Determinants of Health, Catland, Dementia Prevention, Fight Heart Disease Like It's Cancer, Black Butterfly, Wrong, Teaching with AI, Math in Drag, The 36 Hour Day.

First Results

Since its release, the chatbot has been used by 20 unique users, who asked total of 26 messages (with 26 responses from the chatbot. The mode usage was a single question and response; however, this was most often the case when a user got a response other than the "I am not equipped to handle questions about the metadata of a book" fallback response. Some users asked an unanswerable question and immediately gave up when the chatbot used its fallback response, though.

Drawbacks

Many people did not ask about the book's content at all. Importantly, this is a RAG chatbot built *to answer* questions about the content of a book. Instead, many people seemed to be interested in details outside the book: ongoing sales/discounts, table of contents and chapter names, or author information. All of these categories of question are unanswerable by the chatbot *in its current incarnation*.

Potential Next Steps

The chatbot was built for a singular purpose: extract the topic of a user's question, find relevant content in the book, use the content to inform the user if the book would be a good fit. Broadly, that is where AI applications excel—doing one specific thing very well instead of being okay at multipurpose tasks.

There may be an interesting "hybrid" option. Using a speed-optimized model (such as GPT4.1-mini), the user query could be "triaged" into a pre-defined category, e.g. "Question about the content of the book," "Question about sales, prices, and discounts," or "Question about the book's metadata," etc. Then, depending on which category the user query is sorted into, a different chunk of information could be fed into the chat response, or a different piece of code could be run. (One immediate connection would be with Andrew's AI Search—a user's query may indicate they actually want to see other books on dementia and could thus be routed right to the search with relevant terms already added, for instance.) Preliminary testing by Ruben has had (anecdotally) good results in the LLM's sorting capabilities, both for speed and accuracy.