

Assignment #1:

Thinking about LLM Costs in Marketing Applications

Behnam Mo

August 31, 2026

Q1 [20 pt]. BetterBuy offers Bestie, a large-language-model (LLM)-powered artificial-intelligence (AI) assistant that helps customers with their purchase decisions. The Chief Marketing Officer (CMO) has asked you to analyze how customers use Bestie and come up with ways to lower the operating costs of the AI system. You notice that customers often open a Bestie chat and use it more like a search engine than a chatbot. For example, queries typically look like:

- bags for school with side pockets
- pink tumblers that are cute
- tennis shoes, easy to put on, durable

Obviously, this means that users aren't getting the most out of the AI; most of its potential remains unrealized. In one paragraph, explain what you think about this: Is it hurting the firm in any way? In other words, what advantages would the firm have if customers used the AI more like a chatbot? Give at least one advantage.

Q2 [20 pt]. You notice Bestie chats are often quite long; customers keep searching for several products until they find what they prefer and leave the chat. Each turn, they search for a different product, and each time the AI looks at 30 different products in that category before giving its recommendation. The average product description is 6,000 tokens, and the average chat length is 10 turns. You conclude that, on average, Bestie processes tokens just to process product descriptions. If you have 10,000 daily active users, the average annual cost of Bestie is at least (Bestie currently uses OpenAI's Terra model without caching).

Q3 [20 pt]. To save on AI costs, your boss suggests limiting the number of interactions

customers can have with Bestie. He thinks chats shouldn't continue for more than 5 turns. But you think that just makes customers start new chats when the previous one comes to an end, resulting in the same amount of Bestie usage overall. You test this for a month and learn that 70% of your daily users indeed start several chat sessions, while the remaining 30% combine all their queries into one, which Bestie responds to by processing multiple product categories at once. Assuming that the number of active users has remained the same and customers search for the same number of products as before, Bestie's operating cost is now at least

Q4 [18 pt]. What other approaches do you recommend to further lower the costs of Bestie? You do not need to calculate the new costs; just mention 3 approaches that result in less token processing by the AI.

Q5 [22 pt]. Now think about a different scenario: Your marketing team uses AI for campaign ideation. The current AI is GPT-Luna; look up its Application Programming Interface (API) pricing on OpenAI's website. Usually, it takes 6 back-and-forth between the user and Luna to land on an acceptable marketing idea. But when you tried a better model like GPT-Terra for this task, you found that the first shot is always an acceptable idea. Based on what you've learned in this course so far, you think you can save on costs by using GPT-Terra instead of GPT-Luna. Look up GPT-Terra's API pricing and use the right tool in the course material to prove your hunch. The actual numbers do not matter; what matters is your ability to justify the tradeoffs and the counterintuitive result that a model that is more expensive per token can be less expensive per task.