Processo Seletivo TutorMundi

Pesquisa sobre LLM

Léo Victor Cruz Vasconcelos

Rio de Janeiro
27/02/2024

1. RESUMO

O LLM, ou Large Language Models, exemplificado pelo ChatGPT, destaca-se em processamento de linguagem natural. A questão central é se essa tecnologia pode categorizar perguntas. Dentre as opções, BERT e GPT são notáveis, permitindo duas abordagens: utilizar API pronta ou criar a própria API.

Uma API pronta oferece baixo custo inicial e menos preocupações, com rápida implementação, mas limitado controle. Criar uma própria API demanda alto custo inicial e maior preocupação, mas, proporcionando controle total e possibilidade de fine-tuning.

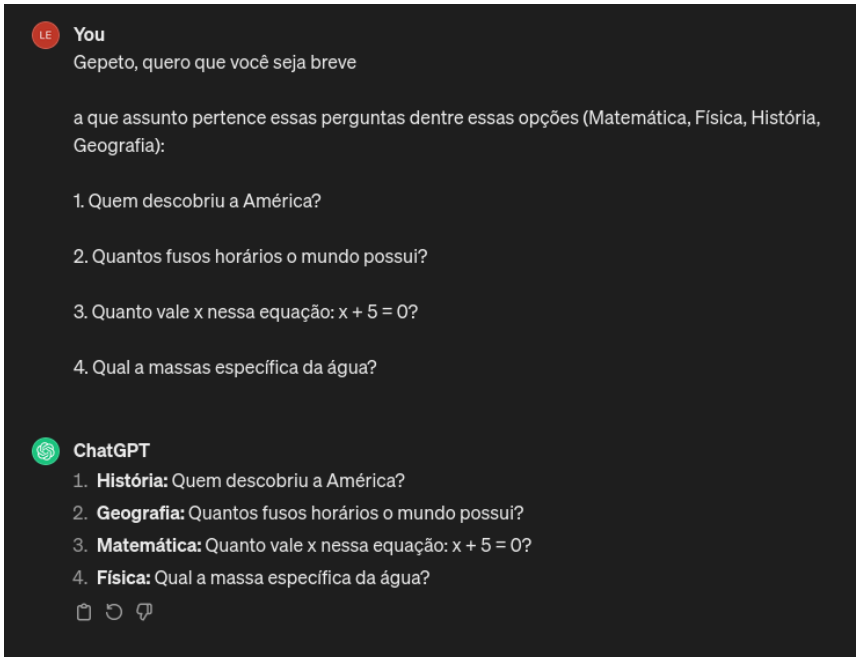
Para API pronta, o custo geral depende da quantidade de acessos ou caracteres, com resposta rápida, custos de recursos humanos e infraestrutura mais baixos. Já construir a própria API é variável, com custo inicial e preocupações mais elevadas, mas, proporcionando controle total e capacidade de fine-tuning.

Uma conclusão mais precisa será possível após implementar duas aplicações mínimas para avaliar o desempenho de cada abordagem.

2. INTRODUÇÃO

LLM deriva do inglês (Large Language Models), significa Grande Modelo de Linguagem e é um algoritmo de aprendizado profundo que pode executar várias funções de processamento de linguagem natural. Com o avanço do poder de processamento, houve o surgimento do ChatGPT, que processa linguagem natural podendo tomar conclusões, responder perguntas, dentre outras possibilidades. Será possível que esse tipo de tecnologia seja capaz de determinar a que assunto pertence uma determinada pergunta?

Figura 1: Teste de perguntas do ChatGPT



Pela figura 1, claramente pode-se ver que essa tecnologia pode ser usada para essa finalidade. Entretanto, como sabemos se essa é a melhor solução? A princípio devemos avaliar 2 critérios: custo e tempo de resposta.

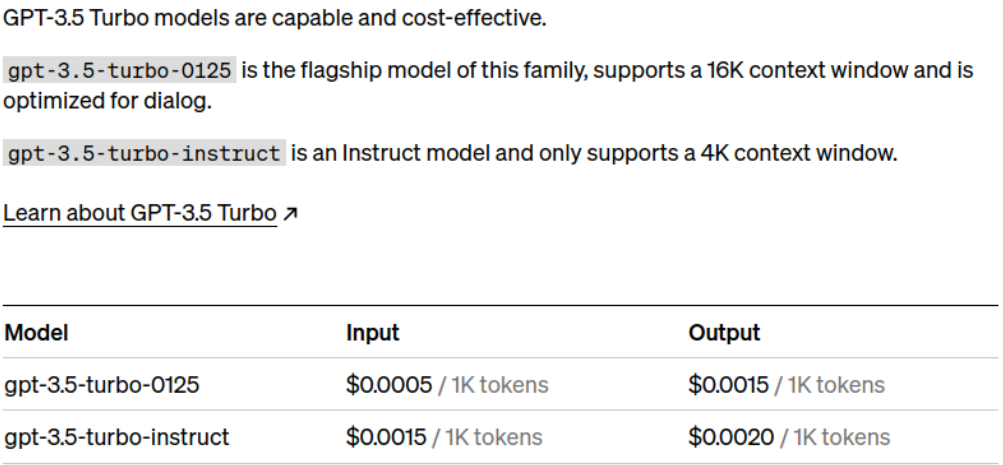
Dentre as soluções possíveis, também precisa avaliar o tempo para realizar a construção da solução.

3. TECNOLOGIAS POSSÍVEIS

Dentre as tecnologias possíveis de LLM, 2 soluções possíveis são BERT e GPT, podendo ser usadas para classificar mas para as duas, haveria 2 metodologias que são: usar API pronta ou desenvolver a próprio GPT ou BERT.

Usar API pronta significa que pode-se implementar rapidamente um programa que realiza um consulta na API, usando um Prompt específico, para a API devolver apenas o resultado desejado. Com isso, pouco tempo foi empregado na construção dessa solução, além disso, ela não requer um custo inicial de compra de infraestrutura. Geralmente, as APIs são cobradas pela quantidade de acessos. No caso do ChatGPT, a API é cobradas por quantidades de caracteres como pode ser visto na figura 2: (tokens equivalem a 4 caracteres)

Figura 2: Custos da API do ChatGPT3.5



GPT-3.5 Turbo models are capable and cost-effective.

`gpt-3.5-turbo-0125` is the flagship model of this family, supports a 16K context window and is optimized for dialog.

`gpt-3.5-turbo-instruct` is an Instruct model and only supports a 4K context window.

Learn about GPT-3.5 Turbo ↗

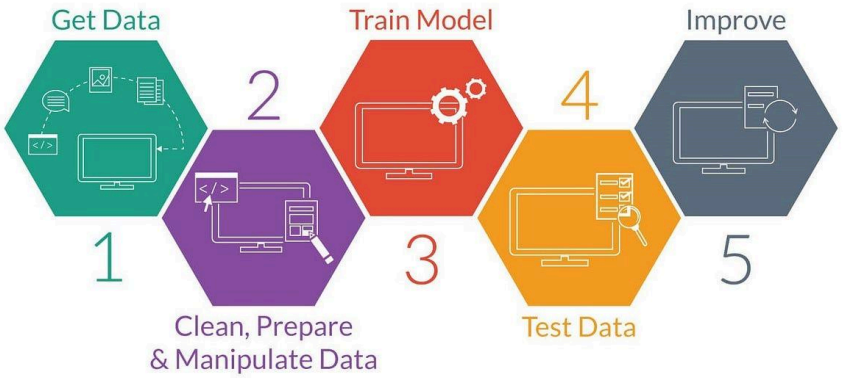
Model	Input	Output
gpt-3.5-turbo-0125	$0.0005 / 1K tokens	$0.0015 / 1K tokens
gpt-3.5-turbo-instruct	$0.0015 / 1K tokens	$0.0020 / 1K tokens

Esse modelo de pagamento é chamado de *pay-as-you-go pricing* e é usado por diversas plataformas pelo mundo.

Desenvolver a própria API pode ser custosa mas a longo prazo pode ser compensadora, pois os custos por consulta podem ser menores que usar uma API pronta.

Esse é um exemplo de Aprendizado Supervisionado e um problema é de classificação e para construir esse modelo, precisa-se adquirir os dados, preparar os dados e treinar o modelo, avaliar o modelo, melhorar o modelo, e quando esse modelo adquirir a acurácia desejada, a API é colocada em produção.

Figura 3: Como Criar um Modelo de Inteligência Artificial



Há vantagens em usar as duas formas, também. Se as 2 forem usadas, há possibilidade de usar a API pronta para avaliar e treinar e API nova, por exemplo. Usar as 2 significa também obter essa *Feature* para que o consumidor usufrua disso mais rápido, à medida que uma solução nova é desenvolvida.

4. CONCLUSÃO

Pode ser que criar uma API seja mais barata do que usar uma pronta mas deve-se avaliar o custo e o tempo de resposta. Para isso, poderia-se fazer uma estimativa do custo de API ao analisar quantos caracteres, por mês, são enviados e recebidos, e para criar a própria API (que é bem mais difícil de prever), quantos novos funcionários eu deveria contratar para criar esse modelo? Por quanto tempo tenho que pagar esses funcionários? Quanto de infraestrutura seria gasto?

Critérios	API Pronta	Criar Própria API
Custo Inicial	Baixíssimo custo inicial	Alto Custo Inicial
Preocupação *	Baixa Preocupação	Alta Preocupação
Tempo para ficar pronto	Tempo menor	Tempo maior
Controle	Pouco Controle	Controle 100%
Fine Tuning?	É possível, mas aumenta o custo	Sim

* Preocupação é uma métrica que está relacionado ao maior número de fatores que influenciam no sucesso da operação, nesse caso podemos citar:

- Recursos Humanos (Tenho recursos humanos capacitados?)
- Dados (Tenho os dados necessários?)
- Tempo (Quanto tempo vai durar? Como posso garantir que vai estar pronto em uma determinada data?)

Pode-se concluir melhor após uma construção de 2 aplicações mínimas seguindo os critérios abaixo.

API Pronta	Construir a Própria API
Custo da API em Geral? Tempo de Resposta?	Custo dos Recursos Humanos? Custo da Infraestrutura? Custo Inicial? Tempo de Resposta?

5. FONTES
    - https://www.elastic.co/pt/what-is/large-language-models
    - https://openai.com/pricing
    - https://medium.com/@mudasserch1/5-core-steps-to-understand-machine-learning-workflow-a-guide-for-beginners-737040850d9b
    - https://www.cedrotech.com/blog/api-ou-desenvolvimento-pr%C3%B3prio/