

WebGPT: Browser-assisted question-answering with human feedback

Monique Monteiro –
moniquelouise@gmail.com



Conceitos importantes

- Coleta de referências à medida que usuários humanos navegam pela interface
- Desafio crescente em NLP:
 - Long-form question answering (LFQA): resposta do tamanho de um parágrafo para uma pergunta aberta
 - IR (ex.: Microsoft Bing Search API) + síntese (fine-tuning do GPT-3)

Contribuição

- Ambiente de navegação web baseado em texto com um qual um modelo de linguagem pode interagir
 - Melhoria do ambiente de recuperação e síntese fim-a-fim
 - *Imitation learning, reinforcement learning*
- Geração de respostas com referências
 - Crucial para auxiliar os anotadores





Datasets

- Modelos treinados para responder a perguntas do ELI5
- Dados adicionais:
 - Demonstrações de humanos usando o ambiente de navegação Web para responder a perguntas: *behavior cloning* (*fine-tuning* supervisionado)
 - Comparações entre duas respostas geradas pelo modelo para a mesma pergunta com suas próprias referências: modelagem de recompensa e RL/*rejection sampling*

Treinamentos

Behavior cloning (BC): fine-tuning em demonstrações usando aprendizagem supervisionada a partir dos comandos gerados pelos usuários (*labels*)

Reward modeling (RM): entropia cruzada a partir das comparações

Reinforcement learning (RL): PPO

Rejection sampling (best-of-n): amostragem de um número fixo de respostas e seleção daquela com maior recompensa (em tempo de inferência)



Resultados interessantes

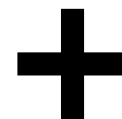
- As respostas do modelo são preferidas por humanos:
 - 56% do tempo em relação às respostas fornecidas por humanos
 - 69% do tempo em relação à resposta mais votada do Reddit
 - Atrás da performance humana no TruthfulQA (respostas curtas)
- Melhor modelo:
 - *Behavior cloning + rejection sampling*

Dúvida básica

- Esse resultado é usado para retreinar o modelo?

Our results are shown in Figure 2. Our best model, the 175B best-of-64 model, produces answers that are preferred to those written by our human demonstrators 56% of the time. This suggests that the use of human feedback is essential, since one would not expect to exceed 50% preference by imitating demonstrations alone (although it may still be possible, by producing a less noisy policy).

- Equações do apêndice I

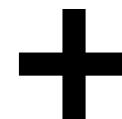
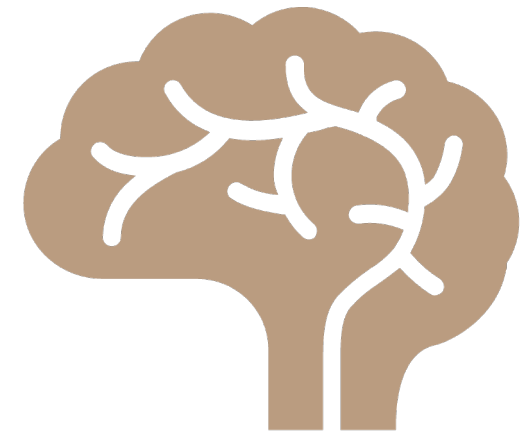




Tópico avançado


- Certamente o modelo GPT 3.5/4 que estamos usando recebeu todo o pré-treino na geração de referências descrito no paper

Visconde: Multi-document QA with GPT-3 and Neural Reranking





Conceitos importantes

-
- Capacidade few-shot de LLMs pode reduzir os custos para resolução de tarefas da *question answering* (QA)
 - Implementação de sistemas de QA para diferentes domínios sem a necessidade de datasets específicos anotados
 - Melhor desempenho quando modelos são induzidos a mostrar evidências
 - *CoT (Chain-of-Thought)*
- 

Contribuição



Um sistema de *question-answering* (QA) que pode responder perguntas cujas evidências de suporte estão espalhadas por múltiplos documentos (potencialmente longos)



Gargalos ainda estão na etapa de recuperação

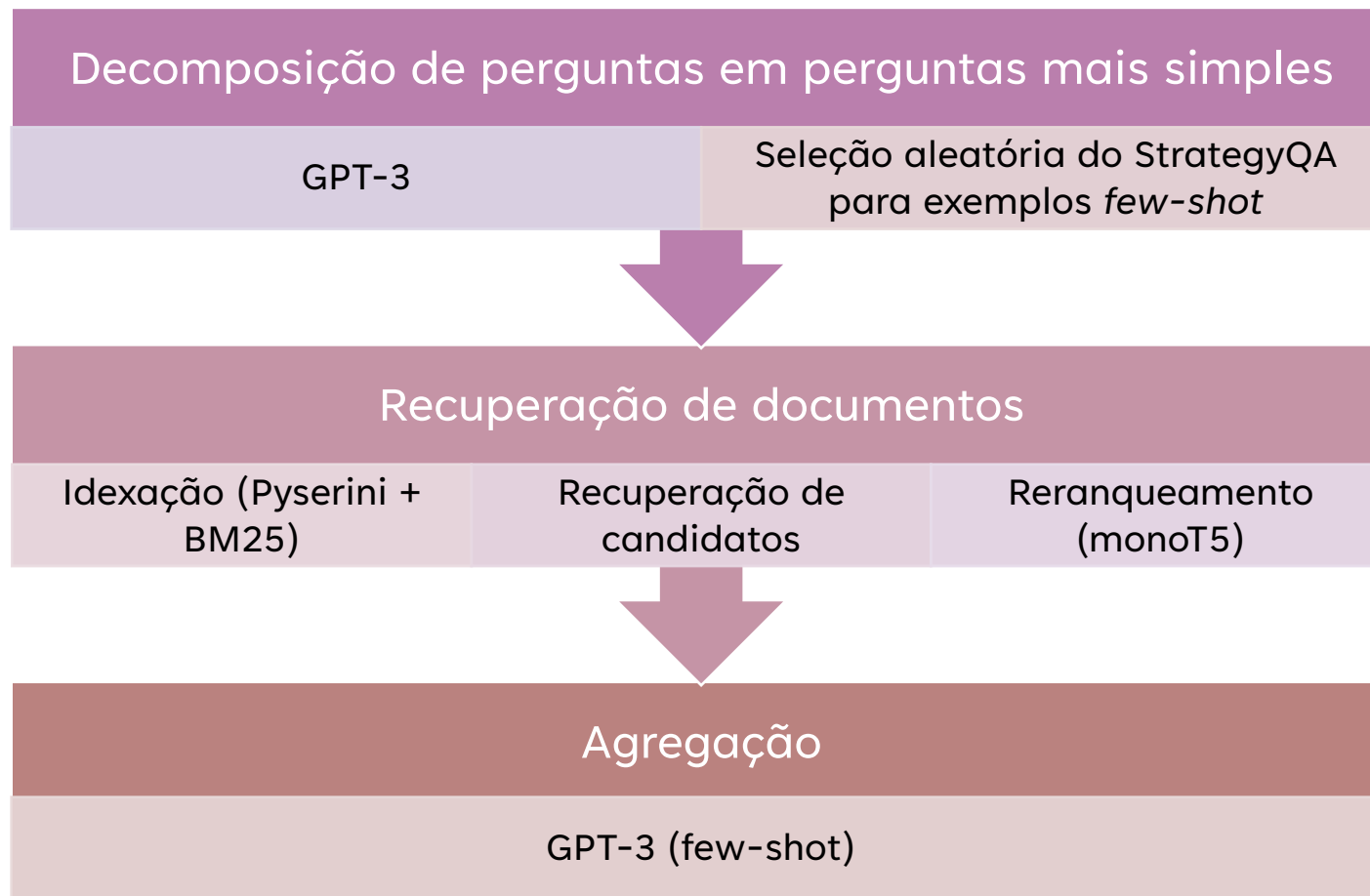


Leitores com performance próxima à humana



Pipeline: BM25 + monoT5 + GPT-3

Etapas



Prompts

Estáticos

- Lista pré-definida

Dinâmicos

- K exemplos mais próximos (KNN + SentenceTransformers)

Resultados interessantes



Prompts dinâmicos apresentam melhor desempenho



Uso de contextos “gold” melhor do que busca na base inteira ou apenas links



Melhores resultados com CoT

Sistemas responde até a perguntas marcadas como sem resposta.





Dúvida básica

- Quando se fala nas diferentes configurações (1) GoldCtx, 2) Linked pages e 3) dataset inteiro, a terceira está excluída das duas primeiras?

