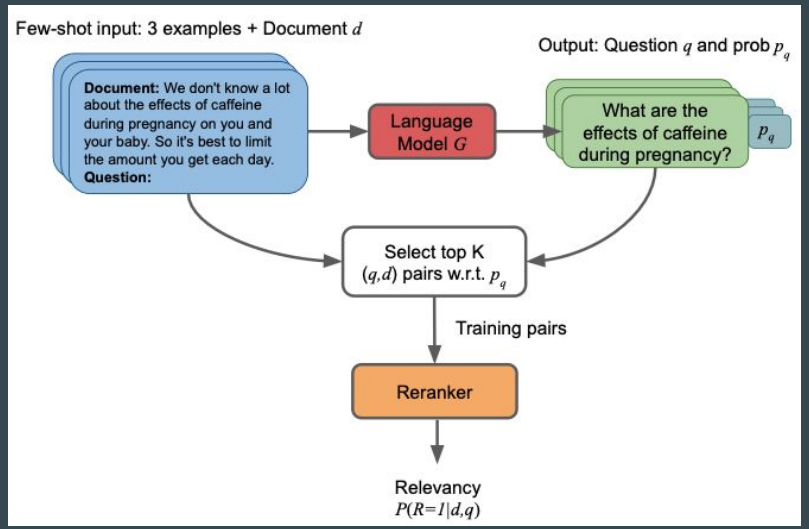
# InPars: Data Augmentation for Information Retrieval using Large Language Models

•••

04 de maio de 2023

# Conceitos Importantes

- Geração de dados sintéticos
  - Utilização de LLMs para gerar queries relevantes a partir de passagens.



# Contribuições do Artigo

- Relevante questão sobre o uso de LLMs em IR:
  - "Custo" computacional das tarefas de recuperação de informação.
- Buscadores densos, apesar de serem computacionalmente mais baratos, precisam de uma passagem de inferência para calcular a representação vetorial de cada documento da coleção.
- Outro desafio no desenvolvimento de modelos neurais para IR é a falta de dados de treinamento específicos do domínio.

# Resultados

| | | MARCO | TREC-DL 2020 | | Robust04 | | NQ | TRECC |
|---|---|---|---|---|---|---|---|---|
| | | MRR@10 | MAP | nDCG@10 | MAP | nDCG@20 | nDCG@10 | nDCG@10 |
| | ***Unsupervised*** | | | | | | | |
| (1) | BM25 | 0.1874 | 0.2876 | 0.4876 | 0.2531 | 0.4240 | 0.3290 | 0.6880 |
| (2) | Contriever (Izacard et al., 2021) | - | - | - | - | - | 0.2580 | 0.2740 |
| (3) | cpt-text (Neelakantan et al., 2022) | 0.2270 | - | - | - | - | - | 0.4270 |
| | *OpenAI Search reranking 100 docs from BM25* | | | | | | | |
| (4) | Ada (300M) | $ | 0.3141 | 0.5161 | 0.2691 | 0.4847 | 0.4092 | 0.6757 |
| (5) | Curie (6B) | $ | 0.3296 | 0.5422 | 0.2785 | 0.5053 | 0.4171 | 0.7251 |
| (6) | Davinci (175B) | $ | 0.3163 | 0.5366 | 0.2790 | 0.5103 | $ | 0.6918 |
| | *InPars (ours)* | | | | | | | |
| (7) | monoT5-220M | 0.2585 | 0.3599 | 0.5764 | 0.2490 | 0.4268 | 0.3354 | 0.6666 |
| (8) | monoT5-3B | **0.2967** | **0.4334** | **0.6612** | **0.3180** | **0.5181** | **0.5133** | **0.7835** |
| | ***Supervised*** [▷ **MARCO**] | | | | | | | |
| (9) | Contriever (Izacard et al., 2021) | - | - | - | - | - | 0.4980 | 0.5960 |
| (10) | cpt-text (Neelakantan et al., 2022) | - | - | - | - | - | - | 0.6490 |
| (11) | ColBERT-v2 (Santhanam et al., 2021) | 0.3970 | - | - | - | - | 0.5620 | 0.7380 |
| (12) | GPL (Wang et al., 2021) | - | - | - | - | - | - | 0.7400 |
| (13) | miniLM reranker | †0.3901 | - | - | - | - | ‡0.5330 | ‡0.7570 |
| (14) | monoT5-220M (Nogueira et al., 2020) | 0.3810 | 0.4909 | 0.7141 | 0.3279 | 0.5298 | 0.5674 | 0.7775 |
| (15) | monoT5-3B (Nogueira et al., 2020) | 0.3980 | **0.5281** | **0.7508** | 0.3876 | 0.6091 | **0.6334** | 0.7948 |
| | *InPars (ours)* [▷ MARCO ▷ unsup in-domain] | | | | | | | |
| (16) | monoT5-3B | 0.3894 | 0.5087 | 0.7439 | **0.3967** | **0.6227** | 0.6297 | **0.8471** |

# Tópicos Avançados

- Devido à interação consulta/documento, cross-encoders (linhas 7-8) são mais eficazes (e contextuais) do que a consulta independente e codificação de documento de bi-encoders.
- As questões sintéticas geradas pelo InPars têm mais semelhança com as consultas que o retriever verá no momento da inferência do que as sentenças extraídas dos textos usados pelo Contriever e cpt-text.
- As novas consultas que geramos, combinadas com documentos que o modelo não usaria de outra forma, fornecem um conjunto de treinamento completamente novo com **alta diversidade**.

# Obrigado

Manoel Veríssimo