INPARS: DATA AUGMENTATION FOR INFORMATION RETRIEVAL USING LARGE LANGUAGE MODELS

Monique Monteiro –
moniquelouise@gmail.com

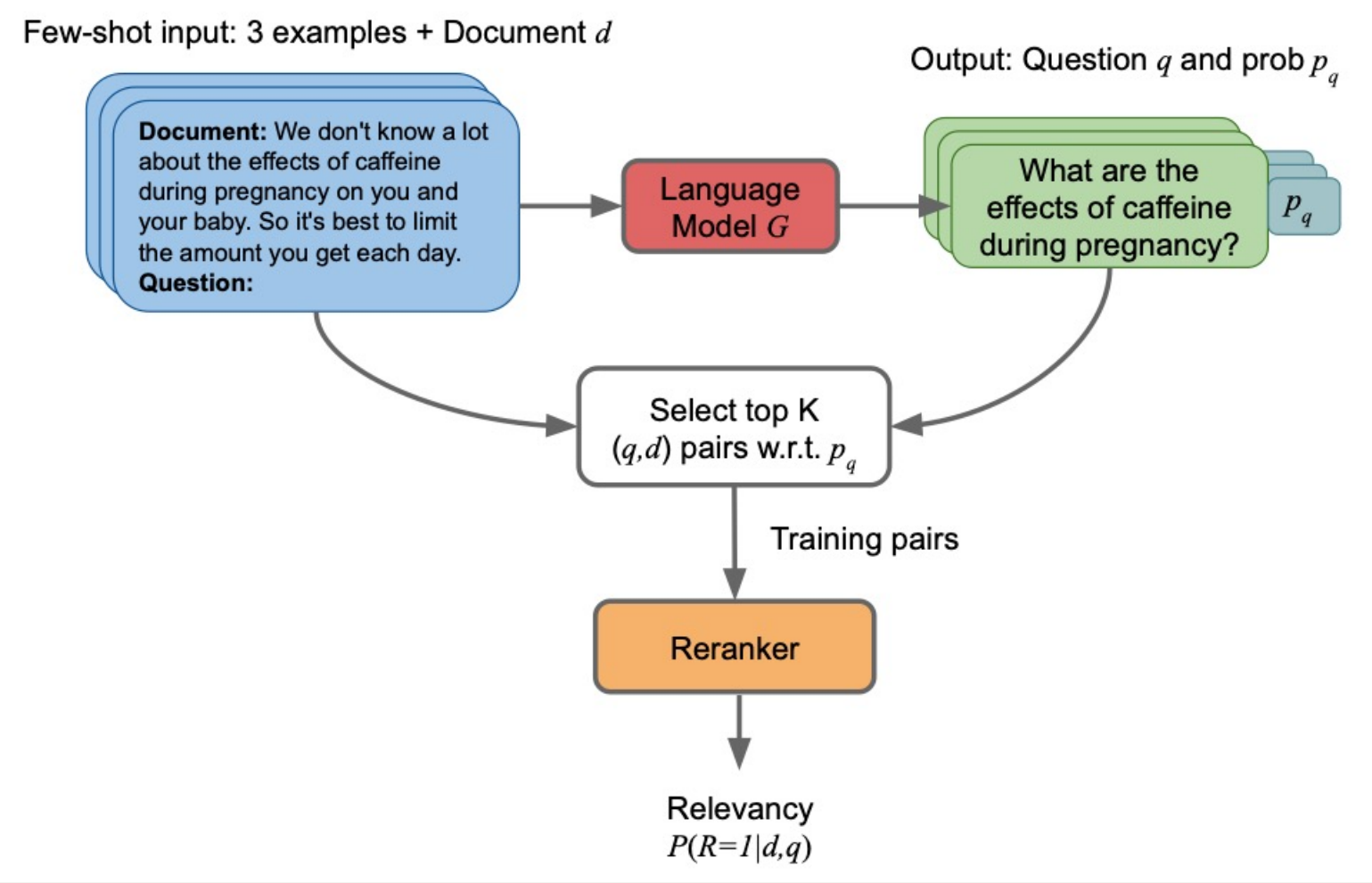
CONCEITOS IMPORTANTES

- LLMs (Large Language Models) raramente são usados em IR (Information Retrieval), com notáveis exceções
  - Natureza intensiva computacionalmente das tarefas de IR (ex.: rereanqueamento)
  - Alto custo financeiro para uso de APIs (ex.: OpenAI)
  - Overhead mesmo para geração de vetores densos
- Escassez de datasets rotulados
  - Alternativa: uso de geradores sintéticos baseados em LLMs

CONTRIBUIÇÃO

- Alavancar as capacidades de few-shot learning de LLMs como geradores de dados sintéticos para tarefas de IR
- Demonstrar que modelos treinados (finetuned) apenas no dataset não supervisionado superam BM25 e outros métodos densos recentes
- Quando combinado com finetuning supervisioado, o método atinge resultados estado-da-arte em 2 de 3 datasets avaliados
- Treinamento de reranqueador a partir dos dados sintéticos gerados
- Método também adequado para algoritmos de IR não neurais

MÉTODO


Few-shot input: 3 examples + Document d
Output: Question q and prob p_q
Document: We don't know a lot about the effects of caffeine during pregnancy on you and your baby. So it's best to limit the amount you get each day. Question:
Language Model G
What are the effects of caffeine during pregnancy?
p_q
Select top K (q,d) pairs w.r.t. p_q
Training pairs
Reranker
Relevancy $P(R=1|d,q)$

DATASETS

- MS MARCO
- TREC-DL
- Robust04
- Natural Questions
- TREC-COVID

GERAÇÃO DAS TUPLAS POSITIVAS E NEGATIVAS

- Uso do GPT-3 Cure como modelo de linguagem
- 100.000 documentos amostrados
  - As 10.000 melhores são utilizadas para finetuning
- Documentos com menos de 300 caracteres são descartados
- Prompts:
  - Vanilla:
    - Queries relevantes selecionadas aleatoriamente do MS MARCO
  - Guided by Bad Questions:
    - Queries do MS MARCO são usadas como não relevantes (“bad”)
    - Queries relevantes são construídas manualmente
- Resultados obtidos pelo BM25 são selecionados aleatoriamente para exemplos negativos

FINETUNING

- Método multi-estágio:
  - Pyserini (BM25) → Reranqueador (monoT5)
  - Balanceamento de dados por batch
  - Treinamento por uma época
- Finetuning de um modelo por coleção usando questões sintéticas geradas daquela coleção

RESULTADOS INTERESSANTES/INESPERADOS

- GBQ prompts levam a resultados melhores do que Vanilla prompts
  - Exceto para MS MARCO e TREC-DL
- In-domain input documents:
  - Melhores resultados
  - Zero-shot domain adaptation
  - Perguntas geradas de documentos amostrados das mesmas coleções nas quais o modelo é avaliado
  - Dúvida básica: não pode gerar overfitting?
- Vantagens da etapa de filtragem

INPARS-V2

- Uso de modelos de linguagem de código aberto (GPT-J)
- Novo estado da arte obtido no benchmark BEIR