INFNET

Desenvolvimento de Data-Driven Apps com Python [24E4_3] - TP2

Aluno: Rodrigo Moreira Avila

Repositório GIT: https://github.com/r-moreira/infnet-data-driven-apps-tp2-p1

Questões teóricas

> Parte 1

Questão 3: Com base na API desenvolvida na Questão 2 (Parte1), explique as principais limitações do modelo de tradução utilizado.

Enumere e discuta:

- Limitações quanto à precisão da tradução.
- Desafios de tempo de resposta e desempenho em grande escala.
- Restrições de custo e escalabilidade.
- Limitações na tradução de gírias, expressões idiomáticas ou linguagem de contexto.

A API desenvolvida na Questão 2 (Parte1) possui limitações.

1. **Limitações quanto à precisão da tradução**: O modelo pode cometer erros assim como qualquer outro modelo de LLM, no caso do modelo em questão, ele pode ter dificuldade ao traduzir frases complexas ou ambíguas.

2. **Desafios de tempo de resposta e desempenho em grande escala**: O modelo pode ter um tempo de resposta maior em grande escala, pois o uso de recursos usado pelo modelo é grande, ao se comparar com uma consulta em banco de dados, por exemplo.

3. **Restrições de custo e escalabilidade**: O modelo pode ter um custo alto para ser utilizado em grande escala, devido a quantidade de recursos usados, e consequentemente, afetando a escalabilidade do modelo.

4. **Limitações na tradução de gírias, expressões idiomáticas ou linguagem de contexto**: O modelo pode ter dificuldade em traduzir gírias, expressões idiomáticas ou linguagem de contexto, pois o modelo nem sempre consegue entender o contexto da frase, e assim, pode traduzir de forma errada.

Questão 4: Com base no modelo GPT-2 utilizado na Questão 1 (Parte 1), explique as principais limitações do modelo no contexto da geração de texto.

Discuta:

- A coerência do texto gerado.
- Possíveis falhas ou incoerências geradas por LLMs.
- Desempenho e questões de latência.
- Limitações na geração de conteúdo apropriado.

A coerência do texto gerado: O modelo GPT-2 se comparado a outros modelos, ele tem menos precisão, e nem sempre consegue gerar um texto coerente, ele acaba alucinando bastante, principalmente em textos mais longos.

Possíveis falhas ou incoerências geradas por LLMs: Falhas ou incoerências são comuns, pois ele não tem acesso a uma base de dados factual, e assim, ele pode gerar informações erradas, tudo depende do treinamento do modelo e da quantidade de dados que ele foi treinado.

Desempenho e questões de latência: O modelo GPT-2 pode ter um desempenho ruim em grande escala, pois o uso de recursos usado pelo modelo é grande, ao se comparar com uma consulta em banco de dados, por exemplo, inclusive ele é muito mais pesado que o modelo de tradução da Questão 2 (Parte1).

Limitações na geração de conteúdo apropriado: O modelo GPT-2 pode acabar gerando conteúdo inadequado ou ofensivo, já que ele não tem uma compreensão real do que está sendo gerado, ele apenas tenta gerar um texto coerente baseado no treinamento que ele recebeu. Além disso, ele pode apresentar viéses e falta de contexto atualizado, o GPT2 é um modelo de 2019.

In [1]:
```
!jupyter nbconvert --to webpdf rodrigo_avila_DR3_TP2_p1.ipynb
```

```
[NbConvertApp] Converting notebook rodrigo_avila_DR3_TP2_p1.ipynb to webpdf
[NbConvertApp] Building PDF
[NbConvertApp] PDF successfully created
[NbConvertApp] Writing 171716 bytes to rodrigo_avila_DR3_TP2_p1.pdf
```