

Informe chatbot

▼ Índice

[Índice](#)

[Dataset](#)

[Intro](#)

[Mineração de contextos](#)

[Anotações](#)

[Prompt](#)

[Resultado dos prompts](#)

[Limpeza](#)

[Formatação](#)

[Resultados do conjunto de dados](#)

[Implementando RAG](#)

[Métricas](#)

[Exact Match \(EM\) e F1 Score](#)

[Word Error Rate \(WER\)](#)

[Resultados](#)

[Tabela de resultados](#)

[Comparação gráfica](#)

[Avaliação das métricas](#)

[Observações e Conclusões:](#)

[Referências](#)

Dataset

Intro

Para abordar a construção de um conjunto de dados de avaliação, foi utilizado o formato do Stanford Question Answering Dataset (SQuAD) [Rajpurkar et al. 2016]. Esse é um conjunto de dados com um formato bem definido para interpretação de texto e que facilita algumas métricas, como F1 e correspondências exatas. Também se usou o FaQuAD, um conjunto de dados da língua portuguesa inspirado no SQuAD que contém perguntas e respostas de documentos da Universidade Federal de Mato Grosso do Sul (UFMS) [Sayama et al. 2019].

Esses conjuntos de dados contêm cinco cabeçalhos principais: `id`, `title`, `context`, `question` e `answers`. Em particular, este último é composto por um dicionário no seguinte formato:

```
{
  "text": [],
  "answer_started": []
}
```

Na chave `text`, são armazenados os resultados que servirão como Grand True.

Mineração de contextos

Os contextos foram extraídos com base em parágrafos e nas seções próprias do documento. A estrutura do documento foi utilizada para gerá-los. Os nomes dos capítulos foram usados como títulos. Dentro disso, foram identificados os artigos, itens e seções do documento.

A seção do artigo foi processada individualmente, assim como cada ANEXO. Isso ocorreu porque cada anexo tinha uma estrutura específica.

Esse processo foi realizado utilizando a biblioteca `PyMuPDF` para extrair os textos e tabelas.

Vale ressaltar que partes do processo foram feitas manualmente, como alguns ajustes das tabelas que estavam contidas no documento.

Esse processo resultou em:

Titulos	Contextos
28	471

Em média, 16 contextos para cada título.

▼ Lista de títulos:

- **título**
- **Vagas e sistemas de ingresso à Graduação**
- **Objetivo e características do Vestibular Unicamp**
- **Sobre vagas oferecidas e as modalidades de classificação**
- **Inscrição**
- **Sobre as provas, notas e convocações**
- **Matrículas**
- **Sobre as prioridades nos sistemas de ingresso Unicamp**
- **Disposições gerais**
- **ANEXO I**
- **ANEXO II - PROGRAMA DAS PROVAS**
- **ANEXO II - REDAÇÃO, LÍNGUA PORTUGUESA E LITERATURAS DE LÍNGUA PORTUGUESA**
- **ANEXO II - RELAÇÃO DE LIVROS**
- **ANEXO II - MATEMÁTICA**
- **ANEXO II - GEOGRAFIA**
- **ANEXO II - HISTÓRIA**
- **ANEXO II - SOCIOLOGIA**
- **ANEXO II - FILOSOFIA**
- **ANEXO II - LÍNGUA INGLESA**
- **ANEXO II - CIÊNCIAS BIOLÓGICAS**
- **ANEXO II - FÍSICA**
- **ANEXO II - QUÍMICA**
- **ANEXO II - PROVAS DE HABILIDADES ESPECÍFICAS**
- **ANEXO III**
- **ANEXO IV - Indicação de vagas e critérios para remanejamento de vagas no curso de Música**
- **ANEXO V - Cursos de graduação por área de realização da prova do Vestibular Unicamp.**
- **ANEXO VI - Candidatos(as) com deficiência ou em condições que exijam recursos específicos para realizar as provas do VU 2024**
- **ANEXO VII - Redução Parcial da Taxa de Inscrição do VU 2024**

Anotações

Prompt

As anotações foram feitas a partir de solicitações ao chatGPT3.5 usando o seguinte prompt de controle para a entrada usando técnicas de **Few Shot Prompting**.

Foi projetado um prompt que continha:

1. Especificação do `role` do sistema:

- `Você é um anotador de dados.`

2. Instruções sobre a tarefa que deveria ser cumprida:

- `Dado um contexto, o anotador deve fornecer três perguntas e três respostas para cada pergunta.`

3. Parâmetros de como a resposta deveria ser:

- `As perguntas devem esclarecer possíveis dúvidas de um usuário sobre esse contexto.`
- `As respostas devem solucionar a pergunta usando o mesmo contexto.`
- `As respostas devem estar contidas no contexto.`
- `As respostas devem ser sucintas.`
- `As respostas devem ser três para a mesma pergunta, onde:`
 - `a primeira deve ser simples e dar solução na pergunta sem muito argumento,`
 - `a segunda deve ter um argumento melhor e`
 - `a terceira resposta deve ser mais completa.`

4. Um exemplo aleatório extraído do FaQuAD. E com o formato que queríamos na saída.

O prompt ficou da seguinte forma:

```
"""
Você é um anotador de dados.
Dado um contexto o anotador deve fornecer três perguntas e três respostas para cada pergunta.
As perguntas devem esclareçam possíveis dúvidas de um usuário sobre esse contexto.
As respostas devem solucionar a pergunta usando o mesmo contexto.
As respostas devem estar contida no contexto.
As respostas devem ser sucintas.
As respostas devem ser três para a mesma pergunta, onde:
    a primeira deve ser simples y dar solução na pergunta sim muito argumento,
    a segunda deve ter um argumento melhor e
    a terceira resposta deve ser mais completa.

exemplo:
contexto: Em muitas instituições, o TCC é encarado como critério final de avaliação do aluno: em caso de reprovação, o aluno estará i
anotação:
...
{
  "perguntas": {
    "p0": "O que acontece se o aluno reprovar no TCC?",
    "p1": "Como o TCC é compreendida em instituições?",
    "p2": "Quantos professores compõem a banca examinadora?"
  }
  "respostas": {
    "p0": {
      "r0": "o aluno estará impedido de obter o diploma",
      "r1": "estará impedido de obter o diploma e consequentemente exercer a respectiva profissão",
      "r2": "estará impedido de obter o diploma",
    }
    "p1": {
      "r0": "como critério final de avaliação do aluno",
      "r1": "como critério final de avaliação do aluno",
      "r2": "como critério final de avaliação do aluno",
    }
    "p2": {
      "r0": "entre 3 e 5",
      "r1": "entre 3 e 5 professores",
      "r2": "entre 3 e 5 professores (não necessariamente com Mestre ou Doutor)",
    }
  }
}
...
"""
```

Resultado dos prompts

Ao finalizar este processo, foram obtidas 471 respostas do modelo, que estavam em formato de texto, porém não no formato necessário. Em alguns casos, as respostas estavam incompletas e esses resultados foram removidos.

Limpeza

Os resultados que não atendiam às expectativas foram eliminados, geralmente eram resultados incompletos. Também foram removidos do conjunto de dados alguns resultados que continham perguntas que levavam o modelo a fornecer respostas com informações que não estavam no contexto. A seguir, é fornecido um exemplo disso:

```
Titulo:
ANEXO II - QUÍMICA

Contexto:
"Eletroquímica
1ºF/2ºF - Processos de oxidação e redução – equacionamento, número de oxidação e identificação de espécies redutoras e oxidantes.
1ºF/2ºF - Aplicação da tabela de potenciais padrão de eletrodo, pilhas.
1ºF/2ºF - Leis de Faraday.
1ºF/2ºF - Eletrólise de soluções aquosas e de compostos fundidos."

Resposta do modelo:
,{
  "perguntas": {
    "p0": "O que são processos de oxidação e redução?",
    "p1": "Para que serve a tabela de potenciais padrão de eletrodo?",
    "p2": "O que são as leis de Faraday?",
    "p3": "O que acontece na eletrólise de soluções aquosas e de compostos fundidos?"
  },
  "respostas": {
    "p0": {
      "r0": "Processos de oxidação são aqueles em que ocorre perda de elétrons por uma espécie química, resultando no aumento do número de oxidação.",
      "r1": "Processos de oxidação são aqueles em que ocorre perda de elétrons por uma espécie química, resultando no aumento do número de oxidação.",
      "r2": "Processos de oxidação são aqueles em que ocorre perda de elétrons por uma espécie química, resultando no aumento do número de oxidação."
    },
    "p1": {
      "r0": "A tabela de potenciais padrão de eletrodo é utilizada para prever qual espécie será oxidada e qual será reduzida em uma reação.",
      "r1": "A tabela de potenciais padrão de eletrodo é utilizada para prever qual espécie será oxidada e qual será reduzida em uma reação.",
      "r2": "A tabela de potenciais padrão de eletrodo é utilizada para prever qual espécie será oxidada e qual será reduzida em uma reação."
    },
    "p2": {
      "r0": "As leis de Faraday são leis que descrevem a relação entre a quantidade de substância que reage em uma eletrólise e a quantidade de carga elétrica.",
      "r1": "As leis de Faraday são leis que descrevem a relação entre a quantidade de substância que reage em uma eletrólise e a quantidade de carga elétrica.",
      "r2": "As leis de Faraday são leis que descrevem a relação entre a quantidade de substância que reage em uma eletrólise e a quantidade de carga elétrica."
    }
  }
}
```

Este tipo de dados foi gerado possivelmente devido ao limite de tokens de saída.

Formatação

Para que fique compatível com o formato do SQuAD, temos que usar o comando `eval`, que nos retorna um dicionário com as chaves `perguntas` e `respostas`, como mostrado abaixo:

```
{
  "perguntas": {
    "p0": "O que é o VU 2024?",
    "p1": "Quais são as regras específicas estabelecidas?",
    "p2": "Essas regras são aplicáveis apenas para o Vestibular Unicamp de 2024?"
  },
  "respostas": {
    "p0": {
      "r0": "O VU 2024 é o Vestibular Unicamp específico para o ano de 2024",
      "r1": "É o Vestibular Unicamp do ano de 2024",
      "r2": "É o processo seletivo da Unicamp para ingresso em 2024"
    },
    "p1": {
      "r0": "As regras específicas estabelecem as normas para o processo seletivo do VU 2024",
      "r1": "Estabelecem as regras para o Vestibular Unicamp de 2024",
      "r2": "Determinam as normas para a seleção dos candidatos no VU 2024"
    },
    "p2": {
      "r0": "Sim, essas regras são aplicáveis somente para o Vestibular Unicamp de 2024",
    }
  }
}
```

```

    "r1": "As regras são válidas apenas para o Vestibular Unicamp de 2024",
    "r2": "Apenas o processo seletivo de 2024 seguirá essas regras especificadas"
  }
}
}

```

Isso nos permitirá separar mais facilmente, em seguida, tabulamos nossos dados individualmente. Na seção `Preprocess data` do arquivo `experiments/RAG_probe.ipynb`, encontraremos a implementação desse formato. Como resultado, teremos algo como:

ID	Title	Context	Question	Answers
18	Vagas e sistemas de ingresso à Graduação	Art. 2º A presente Resolução especifica as regras para o VU 2024.	O que é o VU 2024?	{'text': ['O VU 2024 é o Vestibular Unicamp específico para o ano de 2024', 'É o Vestibular Unicamp do ano de 2024', 'É o processo seletivo da Unicamp para ingresso em 2024'], 'answer_start': []}
19	Vagas e sistemas de ingresso à Graduação	Art. 2º A presente Resolução especifica as regras para o VU 2024.	Quais são as regras específicas estabelecidas?	{'text': ['As regras específicas estabelecem as normas para o processo seletivo do VU 2024', 'Estabelecem as regras para o Vestibular Unicamp de 2024', 'Determinam as normas para a seleção dos candidatos no VU 2024'], 'answer_start': []}
20	Vagas e sistemas de ingresso à Graduação	Art. 2º A presente Resolução especifica as regras para o VU 2024.	Essas regras são aplicáveis apenas para o Vestibular Unicamp de 2024?	{'text': ['Sim, essas regras são aplicáveis somente para o Vestibular Unicamp de 2024', 'As regras são válidas apenas para o Vestibular Unicamp de 2024', 'Apenas o processo seletivo de 2024 seguirá essas regras especificadas'], 'answer_start': []}

Observe que os contextos foram duplicados para cada pergunta, mas um dicionário com todas as possíveis respostas foi mantido. Este formato nos ajudará posteriormente a calcular nossa métrica.

Resultados do conjunto de dados

Quantidade de dados:

No final, concluímos o conjunto de dados com as seguintes quantidades de dados:

Context Data	Data raw	UVQuAD
471	422	1270

Onde,

Context Data: Conjunto de dados inicial dos contextos.

Data raw: Conjunto de dados com as solicitações feitas ao ChatGPT.

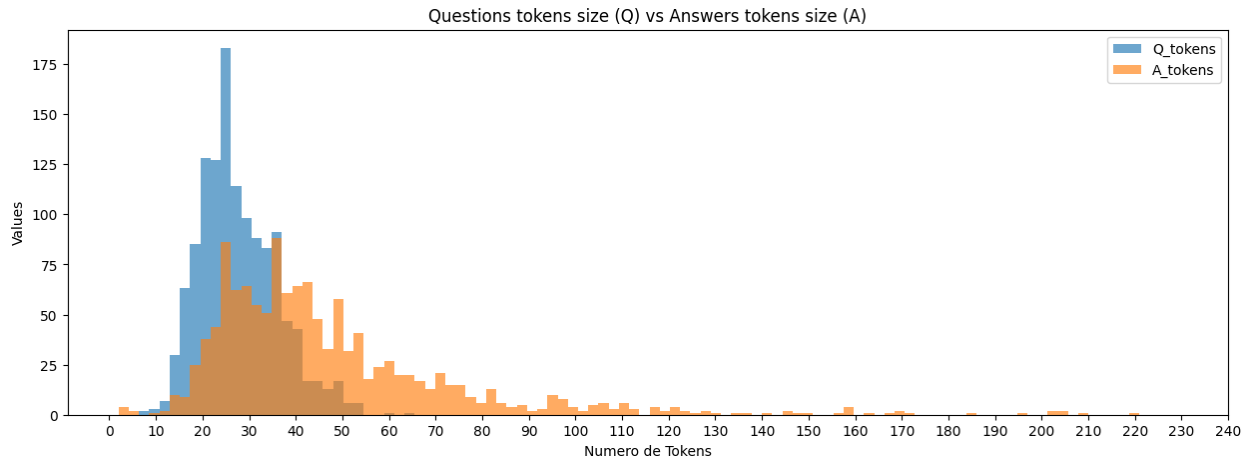
UVQuAD: Conjunto de dados final.

Distribuição de tokens em Perguntas vs Respostas:

A seguir, vemos como as quantidades de tokens são distribuídas entre `Perguntas` e `Respostas`:

Distribuição de tokens em Perguntas vs Respostas:

A seguir, vemos como as quantidades de tokens são distribuídas entre `Questions` vs `Answers`:



Podemos afirmar que a maioria dos tokens usados nas perguntas está entre 12 e 40 tokens. Em relação às respostas, obtemos respostas entre 12 e 60 tokens.

Implementando RAG

De forma simplificada, as perguntas do nosso UVQuAD são usadas como novas entradas em nosso processo. Calculamos as correspondências com os textos disponíveis nos dados de contexto e os enviamos juntos como prompt para o ChatBot. O resultado é então devolvido ao usuário de forma interativa.

- **Banco de Perguntas**

Vamos usar como banco de perguntas aquelas que foram geradas previamente e estão disponíveis no conjunto de dados UVQuAD. Este banco de perguntas será utilizado como entrada em nosso modelo de ChatBot.

- **Recuperador (Retriever)**

Para buscar correspondências, utilizamos o algoritmo **BM25**. A implementação que usamos está no framework `LangChain`, por meio da classe `BM25Retriever`. Essa classe é baseada no módulo `rank_bm25` do Python e deve ser instalada previamente antes de ser utilizada.

O módulo busca em grandes conjuntos de dados, como coleções de texto ou bases de conhecimento, para identificar fontes potenciais de informação.

- **Avaliação de Relevância**

Após recuperar os documentos, é importante avaliar o quão relevantes eles são para a pergunta em questão. Felizmente, a biblioteca já nos retorna os textos por ordem de relevância.

- **Geração de Resposta**

Como geradores de respostas, vamos usar dois modelos de ChatBot. O primeiro será baseado no ChatGPT 3.5 e o segundo modelo será baseado no MariTalk.

O texto recuperado é passado como contexto em um prompt junto com a pergunta, da seguinte forma:

```
"""Com base no contexto: {context}, responda à pergunta: {question}. Resposta: """
```

Métricas

Neste trabalho, foram utilizadas diversas métricas para avaliar o desempenho dos resultados obtidos a partir das previsões geradas pelos modelos ChatGPT e MariTalk em comparação com o ground truth.

Foram empregados EM, F1 Score e WER para comparar de forma abrangente os resultados do ground truth com as previsões geradas tanto pelo ChatGPT quanto pelo MariTalk. Essas métricas proporcionaram uma avaliação rigorosa e completa do

desempenho dos modelos em diversas tarefas de processamento de linguagem natural.

Exact Match (EM) e F1 Score

Foram utilizadas duas métricas-chave para esta avaliação: o *Exact Match* (EM) e o *F1 Score*. O EM mede a proporção de predições que são idênticas ao ground truth, o que significa que a resposta prevista é completamente precisa. Por outro lado, o F1 Score fornece um equilíbrio entre a precisão e a abrangência (recall) das previsões.

Essas duas métricas são amplamente conhecidas e utilizadas para avaliar sistemas de processamento de linguagem natural, sendo especialmente relevantes em tarefas de resposta a perguntas, como as avaliadas nos conjuntos de dados SQuAD e FaQuAD.

Word Error Rate (WER)

Além do EM e do F1 Score, também foi aplicado o *Word Error Rate* (WER) para avaliar a qualidade das previsões em termos de erros a nível de palavras. Essa métrica é valiosa em cenários onde é necessário medir a precisão das previsões em um nível mais granular, levando em consideração possíveis deslocamentos ou substituições de palavras.

Resultados

Tabla de resultados

ID	Pergunta	R ChatGPT	Score	R MariTalk	Score
83	A partir de qual ano os candidatos aprovados em comissões de heteroidentificação na Unicamp estão dispensados da averiguação?	Os candidatos que se submeteram e foram aprovados em comissões de heteroidentificação a partir de 2020 estão dispensados do procedimento de averiguação na Unicamp.	EM: 0.0, F1: 83.72, WER: 0.94	O procedimento de heteroidentificação da Unicamp é realizado por uma comissão composta por três membros.	EM: 0.0, 22.85, WER: 0.94
209	Qual é a precisão utilizada para arredondar a NPF1?	A NPF1 é arredondada para uma casa decimal com precisão de 0,1.	EM: 100.0, F1: 100.0, WER: 0.083	Com base no contexto: "A Comvest, após o término da 2a fase, procederá à apuração dos resultados de todos os candidatos, de acordo com os critérios previamente estabelecidos, e divulgará..."	EM: 0.0, 20.0, WER: 1.0
440	Quantas vagas são destinadas à concorrência para ampla concorrência no curso de Ciência da Computação (Noturno)?	Desculpe, como um chatbot, não tenho acesso às informações específicas sobre o número de vagas destinadas à concorrência para ampla concorrência no curso de Ciência da Computação (Noturno) na Unicamp. Sugiro verificar as informações atualizadas no site oficial da Unicamp ou entrar em contato com a comissão responsável pelo Vestibular da Unicamp para obter esses dados.	EM: 0.0, F1: 40.0, WER: 2.88	{'Error': HTTPError('429 Client Error: Too Many Requests for url: https://chat.maritaca.ai/api/chat/inference')}	EM: 0.0, WER: 1.0

ID	Pergunta	R ChatGPT	Score	R MariTalk	Score
1154	Quais são as restrições para o Vídeo III?	As restrições para o Vídeo III são as seguintes: a duração máxima é de 12 minutos e o tamanho máximo é de 500MB. A pontuação possível para o Vídeo III varia de 0 a 24 pontos.	EM: 0.0, F1: 64.0, WER: 1.52	Com base no contexto: "A gravação em vídeo deve conter: 1- Uma exposição oral, por parte do candidato, de linhas gerais da aula, explicitando: - Repertório a ser utilizado. - Conhecimento musical..."	EM: 0.0, 13.55, W 2.64
1217	Qual é a cidade em que será realizada a avaliação por equipe multiprofissional e interdisciplinar?	A avaliação por equipe multiprofissional e interdisciplinar é realizada em Campinas.	EM: 0.0, F1: 90.0, WER: 0.09	Com base no contexto: "A Comvest reserva-se o direito de não atender as solicitações de condições especiais que não se enquadrem nas especificidades dos recursos disponíveis ou que não sejam compatíveis com as condições de realização das provas.", responda a pergunta: "Pode a Comvest negar alguma solicitação de condição especial?". Resposta:	EM: 0.0, 3.27, W 0.09

Na tabela, apresentamos alguns resultados específicos. (Uma tabela mais completa pode ser encontrada na pasta [experiments/results/](#) ou no arquivo [experiments/UVQuAD.ipynb](#)) juntamente com suas métricas correspondentes para cada resultado. Podemos observar desses resultados que, por exemplo:

Para a pergunta com o **ID 83**, ambos os modelos não acertaram completamente, como indicado pelo baixo EM. Ainda assim, o ChatGPT obteve uma pontuação F1 mais alta, sugerindo uma melhor combinação de precisão e recall. No entanto, seria útil investigar por que as respostas estão divergindo e considerar ajustes no processo de geração de resposta ou no uso do contexto.

Outro caso notável é o do **ID 209**, onde o ChatGPT forneceu uma resposta exata e obteve pontuações perfeitas em EM e F1, enquanto o MariTalk não conseguiu entender corretamente a pergunta, retornando o contexto dado para a geração de respostas. As explicações para isso podem variar desde ajustes nos hiperparâmetros do modelo até ajustes na formulação da pergunta.

No **ID 440**, ambos os modelos não forneceram uma resposta precisa. Ainda assim, o ChatGPT obteve uma pontuação F1 mais alta, indicando uma melhor tentativa de resposta. No entanto, é perceptível que, em termos de correspondência exata do texto, o texto gerado não tem relação. Para esse tipo de resultado, seria considerável fornecer no contexto alguns outros documentos relevantes para a solução desses resultados.

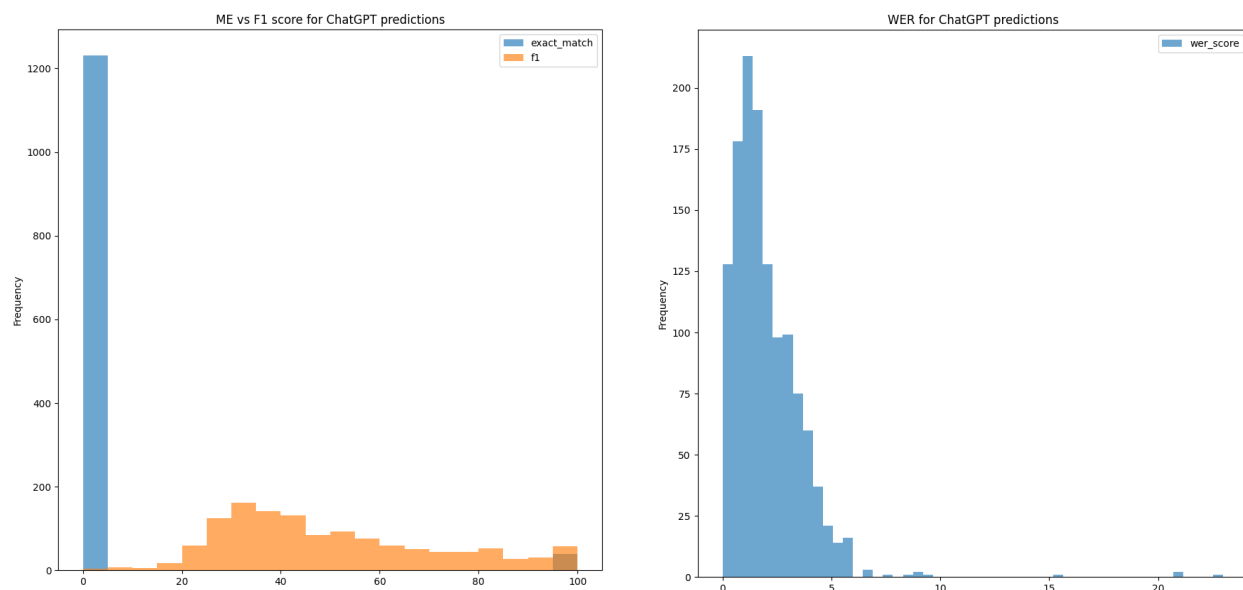
Já no **ID 1154**, o ChatGPT forneceu uma resposta relativamente precisa com uma pontuação F1 mais alta, mas o MariTalk não compreendeu completamente a pergunta, mostrando resultados semelhantes aos obtidos no **ID 209**. Por último, temos o resultado **ID 1217**, onde o ChatGPT forneceu uma resposta precisa com pontuações altas em EM e F1, mas novamente o MariTalk não respondeu corretamente, resultando em pontuações mais baixas.

Sugestões Gerais:

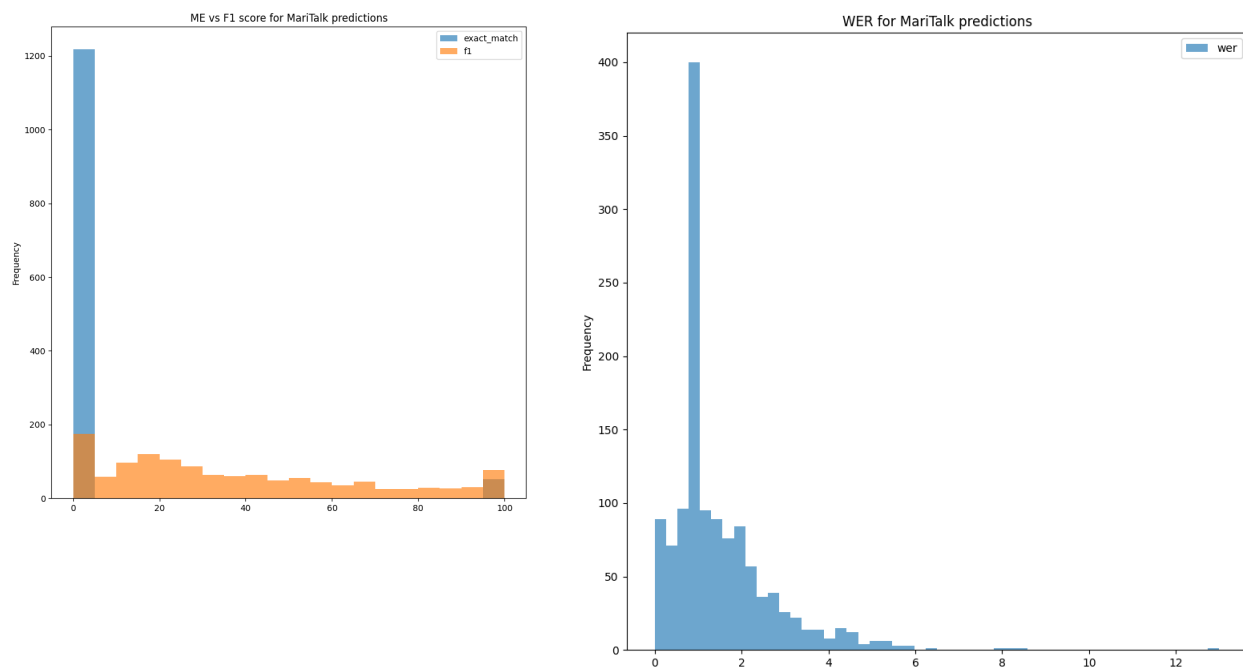
- Analisar casos específicos em que os modelos falham e ajustar o conjunto de dados ou a formulação das perguntas.
- Considerar treinar modelos com dados específicos para informações relacionadas à Unicamp.
- Revisar a formulação das perguntas para garantir clareza e especificidade.
- Avaliar a possibilidade de fornecer respostas alternativas ou uma abordagem de múltiplos modelos para melhorar a robustez das respostas.

Comparação grafica

A continuación se presenta una tabla de ocurrencia para cada una de las metricas usadas en este trabajo:



A partir dos gráficos anteriores, podemos observar que o ChatGPT não teve uma correspondência exata na maioria dos resultados. No entanto, o equilíbrio entre as respostas em termos de precisão e recall permaneceu em faixas aceitáveis. Por outro lado, a correspondência de palavras também não foi satisfatória, já que as respostas geradas não apresentaram coincidências específicas na maioria dos casos. É importante esclarecer que isso pode ter diversas causas e não necessariamente corresponde a resultados nos quais as respostas do modelo não estão relacionadas com a pergunta feita.



Com MariTalk, observamos comportamentos semelhantes aos que vimos com o ChatGPT em relação ao EM e F1 Score. Em relação aos resultados de WER, notamos que o modelo estava ligeiramente mais propenso a gerar respostas que continham um nível semelhante de conteúdo contextual. Isso pode ser vantajoso ao considerar qual dos dois modelos podemos controlar melhor.

Em geral, apesar de os modelos terem apresentado resultados pouco favoráveis, têm vários aspectos que podem ser aprimorados por meio de uma investigação mais aprofundada.

Avaliação das métricas

Modelo	Exact Match	F1	WER
ChatGPT	3.0708661417322833	49.81529279617546	1.644359464627151
MariTalk	4.094488188976378	37.27092615989766	1.3227065204666952

1. Exact Match (EM):

- ChatGPT: 3.07%
- MariTalk: 4.09%

Ambos os modelos apresentam uma taxa relativamente baixa de correspondência exata com as respostas corretas. Essa baixa porcentagem pode indicar que os modelos têm dificuldade em fornecer respostas precisas que coincidam exatamente com o ground truth.

2. F1 Score:

- ChatGPT: 49.82%
- MariTalk: 37.27%

O F1 Score é uma métrica que leva em consideração tanto a precisão quanto o recall. O ChatGPT demonstra um desempenho superior, sugerindo uma melhor combinação de precisão e capacidade de recuperar informações relevantes em comparação com o MariTalk.

3. Word Error Rate (WER):

- ChatGPT: 1.64
- MariTalk: 1.32

O WER mede a taxa de erro de palavras entre as respostas geradas e o ground truth. Ambos os modelos têm valores relativamente baixos, indicando uma boa correspondência nas palavras entre as respostas geradas e as esperadas. O MariTalk, com um WER ligeiramente menor, sugere uma correspondência mais precisa nas palavras usadas nas respostas.

Observações Adicionais:

- O ChatGPT, apesar de sua baixa taxa de EM, destaca-se no F1 Score, o que sugere uma capacidade robusta de produzir respostas precisas, mesmo que não coincidam exatamente com o ground truth.
- O MariTalk, com sua taxa de EM superior, pode estar mais focado em coincidências exatas, mas tem um desempenho inferior em termos de precisão e recall (F1 Score).
- A baixa taxa de EM em ambos os modelos pode indicar a complexidade da tarefa ou a natureza diversificada das respostas esperadas.

Essas observações levam em consideração as diferentes nuances das métricas utilizadas. Como sempre, a interpretação precisa depende da natureza específica da tarefa e dos requisitos do sistema.

Observações e Conclusões:

1. Desempenho do ChatGPT e MariTalk:

- Ambos os modelos, ChatGPT e MariTalk, exibiram resultados variáveis nas métricas de Correspondência Exata (EM), F1 Score e Taxa de Erro de Palavras (WER) nos exemplos fornecidos.

2. ChatGPT vs. MariTalk:

- O ChatGPT mostrou uma tendência a obter pontuações mais altas no F1 Score, indicando um equilíbrio superior entre precisão e recall em comparação com o MariTalk.

3. Correspondência Exata (EM):

- Tanto o ChatGPT quanto o MariTalk tiveram baixas taxas de correspondência exata nos resultados, sugerindo que ambos os modelos podem se beneficiar de melhorias para fornecer respostas exatas.

4. Palavras-Chave e Contexto:

- Observou-se que, em alguns casos, as respostas geradas por ambos os modelos careciam de correspondências específicas de palavras-chave com o contexto fornecido, o que poderia afetar a precisão das respostas.

5. MariTalk e Taxa de Erro de Palavras (WER):

- O MariTalk mostrou uma inclinação ligeiramente maior para gerar respostas com um conteúdo contextual semelhante, de acordo com os resultados de WER. Isso poderia ser considerado ao escolher entre os dois modelos, dependendo da necessidade de controlar o nível de conteúdo contextual.

6. Oportunidades de Melhoria:

- Ambos os modelos apresentaram resultados pouco favoráveis, mas destacou-se a possibilidade de melhorias por meio de pesquisas mais aprofundadas, como ajustes nos hiperparâmetros, treinamento com dados específicos e revisão da formulação de perguntas.

7. Sugestões para Futuras Melhorias:

- Sugere-se analisar casos específicos de falhas, ajustar conjuntos de dados e fórmulas de perguntas, considerar o treinamento com dados específicos para a Unicamp e avaliar a possibilidade de fornecer respostas alternativas ou usar uma abordagem de múltiplos modelos.

8. Avaliação Integral:

- A avaliação integral dos modelos deve considerar não apenas métricas quantitativas, mas também a qualidade e relevância subjetiva das respostas geradas em contextos específicos.

Essas observações e conclusões fornecem uma base para entender o desempenho dos modelos e sugerem áreas-chave para futuras melhorias e ajustes.

Referencias

```
@article{2016arXiv160605250R,
  author = {{Rajpurkar}, Pranav and {Zhang}, Jian and {Lopyrev},
    Konstantin and {Liang}, Percy},
  title = "{SQuAD: 100,000+ Questions for Machine Comprehension of Text}",
  journal = {arXiv e-prints},
  year = 2016,
  eid = {arXiv:1606.05250},
  pages = {arXiv:1606.05250},
  archivePrefix = {arXiv},
  eprint = {1606.05250},
}
```

```
@article{10.1109/BRACIS.2019.00084,
  author = {Sayama,Hélio Fonseca and Araujo,Anderson Viçoso and Fernandes,
    Eraldo Rezende},
  booktitle = {2019 8th Brazilian Conference on Intelligent Systems (BRACIS)},
  title = "{FaQuAD: Reading Comprehension Dataset in the Domain of
    Brazilian Higher Education}",
  year = {2019},
  pages = {443-448},
  doi = {10.1109/BRACIS.2019.00084}}
```