Geração Automática de Perguntas em Português do Brasil Usando os Modelos PTT5 e FLAN-T5

Tiago Felipe V. Braga[1], **Bruno Cardoso Coutinho**[2], **Hilário Tomaz Alves de Oliveira**[1]

[1]Programa de Pós-graduação em Computação Aplicada (PPComp)
Instituto Federal do Espírito Santo (IFES) – Campus Serra

[2]Coordenadoria do Curso Técnico em Informática
Instituto Federal do Espírito Santo (IFES) – Campus Serra

tiagofvx@gmail.com, {bccout,hilario.oliveira}@ifes.edu.br

Abstract. *This paper performs a comparative analysis of the pre-trained neural models of $PTT5$ and $FLAN$-$T5$ for Brazilian Portuguese automatic question generation. To this end, two datasets, PIRÁ and FairyTaleQA, were used to evaluate the ability of these models to generate questions from two scenarios: (i) considering only the context and (ii) using the context and the expected answer. The ROUGE-L and BERTScore measures were used to assess the generated questions, in addition to an analysis based on GPT-4o. The results demonstrated that the $PTT5_{Large}$ model consistently outperformed the other models, generating 93.06% of valid questions in PIRÁ and 82.32% in FairyTaleQA based on the GPT-4o evaluation.*

Resumo. *Este artigo apresenta uma análise comparativa dos modelos neurais pré-treinados do $PTT5$ e $FLAN$-$T5$ para a geração automática de perguntas em português do Brasil. Para isso, foram utilizados dois conjuntos de dados, PIRÁ e FairyTaleQA, para avaliar a capacidade desses modelos de gerar perguntas a partir de dois cenários: (i) considerando apenas o contexto e (ii) usando o contexto e a resposta esperada. As medidas do ROUGE-L e do BERTScore foram usadas para avaliar as perguntas geradas, além de uma análise baseada no GPT-4o. Os resultados demonstram que o modelo $PTT5_{Large}$ apresentou consistentemente desempenho superior aos demais modelos, gerando 93,06% de perguntas válidas no PIRÁ e 82,32% no FairyTaleQA na avaliação baseada no GPT-4o.*

1. Introdução

A Geração de Perguntas (QG, do inglês *Question Generation*) é uma tarefa da área de Processamento de Linguagem Natural (PLN), que envolve a criação automática de perguntas a partir de um dado texto ou conjuntos de dados textuais [Zhang et al. 2021]. Usando técnicas de PLN e algoritmos de Aprendizado de Máquina (AM), os sistemas de QG visam gerar perguntas gramaticalmente corretas e contextualmente relevantes [da Rocha Junqueira et al. 2024]. Diante da grande abundância de informações digitais, sistemas de QG possuem diversas potenciais áreas de aplicação [Mulla and Gharpure 2023]. Na área da educação, a aplicação de abordagens de QG pode contribuir para o desenvolvimento de materiais de avaliação, questionários práticos, no desenvolvimento de sistemas de tutoria, aprimorando processos de aprendizagem e

avaliação [Kurdi et al. 2020]. No âmbito dos sistemas de perguntas e respostas (QA, do inglês Question Answering), abordagens de QG têm sido usadas para o treinamento de modelos com pouca supervisão ou para fins de aumento de dados [Puri et al. 2020].

Apesar do crescente interesse em pesquisas envolvendo a tarefa de QG, a maioria desses estudos concentra-se predominantemente na língua inglesa, onde há diversos recursos e bases de dados disponíveis para experimentação e desenvolvimento [Zhang et al. 2021, Mulla and Gharpure 2023]. Em contrapartida, as pesquisas focadas na língua portuguesa, especialmente para o português do Brasil, ainda são limitadas, resultando em uma escassez tanto de estudos quanto de bases de dados [da Rocha Junqueira et al. 2024, Leite et al. 2024]. Essa lacuna impõe desafios adicionais para o avanço no desenvolvimento e na aplicação prática de sistemas de QG, uma vez que a adaptação de modelos e técnicas desenvolvidas para o inglês nem sempre se traduzem diretamente em resultados eficazes em outros idiomas, dada a complexidade e as particularidades linguísticas inerentes de linguagem natural.

Este artigo tem como objetivo investigar a aplicação de modelos neurais de linguagem pré-treinados baseados na arquitetura *Transformers* [Vaswani 2017], mais especificamente os modelos $PTT5$ [Carmo et al. 2020] e $FLAN\text{-}T5$ [Chung et al. 2024] para a tarefa de QG em português do Brasil. Para isso, foi realizado o ajuste fino desses modelos em suas arquiteturas *small*, *base* e *large*, utilizando as bases de dados do PIRÁ [Paschoal et al. 2021], nativa em português, e uma versão traduzida da base de dados FairytaleQA [Leite et al. 2024] para o português do Brasil. Os experimentos foram realizados em dois cenários: no primeiro, as perguntas foram geradas a partir somente de um dado contexto; no segundo, as perguntas foram geradas considerando tanto o contexto quanto uma resposta prévia. A avaliação dos resultados foi realizada por meio das medidas automáticas do ROUGE-L e BERTScore, que são comumente utilizadas para avaliar abordagens de QG em termos de similaridade léxica e semântica das perguntas geradas com as perguntas de referência. Além disso, foi realizado um experimento adicional utilizando o modelo $GPT\text{-}4o$ para avaliar as perguntas geradas. Esse experimento teve como objetivo complementar as avaliações quantitativas anteriores, proporcionando uma análise adicional da qualidade das perguntas geradas.

As principais contribuições deste artigo incluem: **(i)** o ajuste fino e a avaliação de diferentes arquiteturas dos modelos $PTT5$ e $FLAN\text{-}T5$ para a tarefa de QG em português do Brasil; e **(ii)** uma extensa investigação considerando duas bases de dados, PIRA e FairytaleQA, e duas variações da tarefa. O código-fonte desenvolvido neste trabalho está público em um repositório do GitHub[1].

2. Trabalhos Relacionados

As abordagens de geração de perguntas podem ser classificadas em métodos convencionais e baseados em modelos neurais [Zhang et al. 2021]. Os métodos convencionais de QG baseiam-se principalmente na aplicação de regras heurísticas para transformar os textos em perguntas relacionadas. Recentemente, com a evolução das arquiteturas de redes neurais profundas, houve uma mudança de paradigma na tarefa para a adoção de modelos neurais, permitindo assim, o desenvolvimento de abordagens orientadas a dados e completamente treináveis, na qual a seleção de conteúdo e a construção de perguntas podem

[1] `https://github.com/laicsiifes/question_generation_ptbr`

ser otimizadas de forma combinada. Embora exista uma vasta literatura sobre QG em diversos idiomas [Kurdi et al. 2020, Zhang et al. 2021, Mulla and Gharpure 2023], por limitação de espaço, esta seção foca em trabalhos envolvendo o português do Brasil.

Em [Leite and Lopes Cardoso 2022], os autores apresentam um estudo que envolveu o treinamento do modelo $PTT5$ para a geração de perguntas utilizando uma versão em português do conjunto de dados SQuAD 1.1. Os resultados obtidos foram encorajadores, com desempenho equiparável com a implementação em inglês do modelo $T5$, evidenciando a eficácia dos modelos baseados na arquitetura *Transformers* e estabelecendo *baselines* para futuras comparações para a tarefa de QG em português. Oliveira et al. [Oliveira et al. 2023] abordam o desafio de gerar e classificar distratores (opções incorretas) para questões de múltipla escolha em português. Os autores desenvolveram e combinam vários métodos de geração de distratores, incluindo extração baseada em contexto, manipulação numérica e similaridade semântica a partir de recursos como WordNet.

Junqueira et al. [da Rocha Junqueira et al. 2024] apresentaram uma investigação do desempenho dos modelos $T5$, $FLAN$-$T5$ e $BART$-PT para a geração de perguntas factuais em português do Brasil. Para mitigar o problema da escassez de dados, foi utilizada uma versão em português brasileiro do SQuAD v1.1, obtida por meio de tradução automática. Leite et al. [Leite et al. 2024] realizaram a construção de versões traduzidas automaticamente da base de dados FairytaleQA, que é um conjunto de dados comumente usado para o desenvolvimento de sistemas de perguntas e respostas em inglês. Foram desenvolvidas versões do FairytaleQA para o português de Portugal, português do Brasil, espanhol e francês, que podem ser usadas em pesquisas da área de QG e QA. Além disso, foram realizados experimentos usando modelos neurais baseados na arquitetura $T5$.

Este trabalho difere dos anteriores ao: **(i)** treinar e avaliar diferentes tamanhos de arquitetura dos modelos $PTT5$ e $FLAN$-$T5$, **(ii)** considerar dois cenários da tarefa de QG, **(iii)** adotar uma base de dados escrita nativamente em português (PIRÁ) e outra obtida por meio de tradução automática (FairytaleQA), e **(iv)** analisar o desempenho dos modelos usando uma abordagem com o modelo GPT-$4o$, além de tradicionais medidas de avaliação consideradas em trabalhos anteriores.

3. Materiais e Métodos

3.1. Bases de Dados

Neste trabalho foram utilizados dois conjuntos de dados, o PIRÁ [Paschoal et al. 2021] e o FairyTaleQA [Xu et al. 2022]. Essas bases de dados foram selecionadas por serem usadas em trabalhos da literatura na tarefa de geração de perguntas ou de sistemas de perguntas e respostas em português do Brasil. Além disso, elas possuem três componentes essenciais para a tarefa de QG: **(i)** contexto textual, **(ii)** pergunta associada e **(iii)** resposta correspondente.

O **PIRÁ** é uma base de dados bilíngue (português-inglês) focada em questões oceânicas e da costa brasileira. A base contém 2.261 textos extraídos de trechos de relatórios das Nações Unidas sobre o oceano e de resumos relacionados ao litoral brasileiro [Paschoal et al. 2021]. As perguntas e respostas foram criadas manualmente em um processo de revisão em pares por avaliadores humanos. Após uma análise da base de dados, foi observado que alguns exemplos não apresentam as respostas para as perguntas. Por

isso, esses exemplos foram removidos, já que a resposta é um elemento importante para os experimentos realizados neste trabalho.

O **FairyTaleQA** é uma base de dados comumente usada para avaliar sistemas de perguntas e respostas em inglês. Essa base foi criada por especialistas em educação e é composta por textos narrativos infantis. Leite et al. [Leite et al. 2024] realizaram um processo de tradução do FairyTaleQA para diversos idiomas, incluindo o português de Portugal e do Brasil. Neste trabalho, foi utilizada a versão traduzida para o português do Brasil, que compreende 10.580 perguntas e respostas derivadas de 278 histórias infantis.

Na Tabela 1 são apresentadas para cada base de dados as estatísticas do total de exemplos em cada conjunto (treinamento, validação e teste) e o tamanho médio e desvio padrão do total de palavras para cada componente (contexto, pergunta e resposta). Para gerar essas estatísticas, foi utilizada a ferramenta spaCy[2] para o processamento dos textos.

Tabela 1. Estatística das bases de dados do PIRÁ e FairyTaleQA

Base de Dados	Conjunto	Exemplos	Componente	Média de Palavras (Desvio Padrão)
PIRÁ	Treino	1.756	Contexto	274,73 (141,41)
			Pergunta	13,83 (5,62)
			Resposta	14,32 (11,76)
	Validação	215	Contexto	273,98 (157,08)
			Pergunta	13,65 (5,38)
			Resposta	15,04 (12,06)
	Teste	216	Contexto	250,58 (128,98)
			Pergunta	13,36 (5,68)
			Resposta	14,92 (14,50)
FairyTaleQA	Treino	8.548	Contexto	182,51 (94,53)
			Pergunta	10,23 (3,38)
			Resposta	6,98 (5,73)
	Validação	1.025	Contexto	170,08 (74,18)
			Pergunta	10,93 (3,40)
			Resposta	7,52 (5,96)
	Teste	1.007	Contexto	168,92 (73,77)
			Pergunta	10,48 (3,30)
			Resposta	6,80 (5,31)

3.2. Modelos Avaliados

Neste trabalho, foram avaliados os modelos $PTT5$ e o $FLAN$-$T5$, baseados na arquitetura *Text-to-Text Transfer Transformer* (T5) [Raffel et al. 2020]. Apesar de existirem diferentes tamanhos de arquitetura, as três comumente usadas são "*small*", "*base*" e "*large*". Elas possuem um número crescente de parâmetros, o que geralmente resulta em maior capacidade de aprendizado, mas também em um maior custo computacional. Esses modelos foram escolhidos devido ao seu desempenho promissor em tarefas de PLN e por terem sido explorados em trabalhos anteriores.

O $PTT5$ é uma adaptação do modelo $T5$, especificamente pré-treinada para o português do Brasil [Carmo et al. 2020]. O modelo foi pré-treinado no corpus BrWac [Wagner Filho et al. 2018], uma extensa coleção de páginas *web* em português, contendo aproximadamente 2,7 bilhões de *tokens*. Foram utilizados os modelos $PTT5_{Small}$, que possui aproximadamente 60 milhões de parâmetros, o $PTT5_{Base}$, com cerca de 220 milhões de parâmetros, e o $PTT5_{Large}$, que apresenta aproximadamente 740 milhões de parâmetros.

[2] `https://spacy.io/`

O $FLAN$-$T5$ [Chung et al. 2024] é uma versão aprimorada do $T5$ pré-treinado em múltiplas tarefas de PLN. Esse modelo foi pré-treinado majoritariamente em documentos em inglês, mas possui suporte a outros idiomas, como o português. Foram avaliadas três variantes deste modelo: o $FLAN$-$T5_{Small}$ com cerca de 80 milhões de parâmetros, o $FLAN$-$T5_{Base}$ contendo aproximadamente 250 milhões de parâmetros, e o $FLAN$-$T5_{Large}$ apresentando cerca de 780 milhões de parâmetros. Sua inclusão tem o objetivo de avaliar como um modelo com treinamento diversificado se comporta em comparação a modelos especializados em um único idioma e tarefa.

3.3. Metodologia Experimental

A metodologia experimental utilizada neste trabalho envolveu o desenvolvimento, ajuste fino e a avaliação dos modelos investigados, especificamente ajustados para dois cenários da tarefa de geração de perguntas, conforme ilustrado na Figura 1. Para cada cenário, seis modelos foram treinados, considerando os três tamanhos de arquitetura e os dois modelos $PTT5$ e $FLAN$-$T5$. No primeiro cenário, foi analisada a variação da tarefa que gera perguntas a partir somente do contexto, como entrada. Já no segundo cenário, os modelos recebem tanto o contexto quanto uma resposta como entrada e devem gerar uma pergunta como saída.

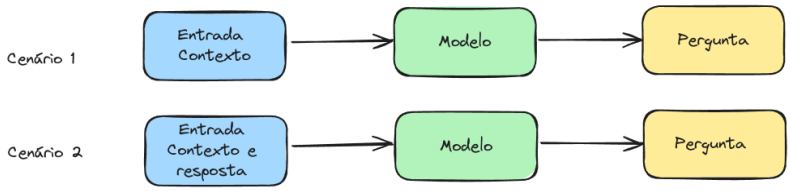

Figura 1. Cenários da tarefa de QG analisados.

Os modelos $PTT5$ e $FLAN$-$T5$ foram implementados usando a biblioteca *Transformers*[3]. O tamanho de entrada máximo foi definido para 512 *tokens*, enquanto a saída foi configurada para no máximo 40 *tokens*. Durante o treinamento, os modelos foram ajustados por no máximo 20 épocas, sendo utilizada a estratégia de parada antecipada com uma paciência de 5 épocas. Para mitigar o sobreajuste dos modelos, ao final de cada época, o modelo treinado é aplicado no conjunto de validação e é computada a medida do ROUGE-L, sendo salvo somente o modelo com maior valor. Durante a geração das perguntas, foi utilizado o algoritmo de *Beam Search* com uma largura de tamanho 5. Esses valores foram definidos a partir da análise de trabalhos anteriores.

A avaliação do desempenho dos modelos foi realizada por meio de duas abordagens: a aplicação de métricas automáticas de similaridade e uma avaliação com base no modelo de linguagem GPT-$4o$. Para a avaliação automática, foi utilizada a métrica Recall-Oriented Understudy for Gisting Evaluation Longest Common Subsequence (ROUGE-L) [Lin 2004], que mensura a similaridade com base na maior sequência de palavras em comum entre as perguntas geradas e as perguntas de referência. Adicionalmente, foi utilizada a métrica BERTScore [Zhang et al. 2019], que calcula a similaridade de cosseno a partir das representações em embeddings extraídas do modelo *Bidirectional Encoder Representations for Transformers* (BERT).

[3]`https://huggingface.co/docs/transformers/index`

Para uma avaliação mais holística e contextualmente relevante, foi realizada uma análise usando o modelo $GPT\text{-}4o$. Essa avaliação foi pensada porque, dado um contexto específico, é possível gerar múltiplas perguntas válidas que não necessariamente precisam ser idênticas à pergunta de referência presente nas bases de dados usadas nos experimentos. Esta situação é particularmente relevante no Cenário 1, onde apenas o contexto é fornecido como entrada para o modelo. Neste caso, diversas perguntas podem ser consideradas válidas, desde que sejam respondíveis com base no contexto fornecido. Em contraste, no Cenário 2, onde o contexto e a resposta esperada são fornecidos como entrada, a pergunta gerada deve ser semanticamente equivalente à pergunta de referência.

O processo de avaliação utilizando o $GPT\text{-}4o$ foi inspirado na técnica de *Retrieval Augmented Generation* (RAG). Esta técnica consiste em fornecer um contexto e uma pergunta para um modelo de linguagem de grande escala (LLM, do inglês *Large Language Model*), solicitando que ele responda à pergunta usando apenas o contexto fornecido ou sinalize caso não seja possível [Chen et al. 2024]. Seguindo esta abordagem, foi criado um *prompt*[4] contendo o contexto original e a pergunta gerada pelos modelos avaliados. Este prompt foi então submetido ao $GPT\text{-}4o$, com a instrução de responder à pergunta utilizando somente o contexto fornecido ou indicar a impossibilidade de resposta. Com base nas respostas do $GPT\text{-}4o$, foi calculado o percentual de perguntas válidas (aquelas que o LLM conseguiu responder com base no contexto) e inválidas (as que não puderam ser respondidas) para cada modelo avaliado. Deste modo, foi possível analisar se as perguntas geradas foram relevantes ao contexto, ainda que diferentes da pergunta de referência. Embora esta análise seja automatizada, foi realizada uma inspeção manual em amostras das saídas do $GPT\text{-}4o$ para verificar sua confiabilidade. Foi observado que, em geral, o LLM identificava corretamente as perguntas válidas e inválidas.

4. Resultados

Na Tabela 2 são apresentados os resultados dos experimentos, considerando os cenários 1 e 2, com base nas medidas de avaliação do ROUGE-L e BERTScore. No Cenário 1 (apenas contexto), o $PTT5_{Base}$ e o $PTT5_{Large}$ apresentaram os melhores desempenhos para as bases de dados do FairyTaleQA e PIRÁ, respectivamente. No Cenário 2 (contexto e resposta esperada), o $PTT5_{Large}$ superou os demais modelos em ambas as bases. Fica evidente que os modelos $PTT5$ consistentemente obtiveram melhores resultados do que os modelos $FLAN\text{-}T5$ em ambos os conjuntos de dados e cenários, sugerindo que o pré-treinamento específico em português confere vantagens na tarefa de geração de perguntas.

Comparando os resultados obtidos em ambos os cenários de avaliação, observa-se que os modelos apresentaram melhores desempenhos no Cenário 2 em comparação com o Cenário 1. Isso acontece porque no Cenário 2, como a resposta é dada como entrada, ela guia os modelos a gerarem perguntas para aquele contexto e resposta. Assim, a pergunta gerada precisa ser semanticamente equivalente à pergunta de referência. Tal situação não ocorre no Cenário 1, já que é somente dado o contexto como entrada e, para um mesmo contexto, é possível gerar diversas perguntas válidas. Por isso, para melhor avaliar o Cenário 1, foram realizadas as análises usando o modelo $GPT\text{-}4o$.

Na Tabela 3 são apresentados os resultados da avaliação dos modelos no Cenário 1 usando o $GPT\text{-}4o$. Os resultados obtidos apresentam um padrão similar ao primeiro ex-

[4]O *prompt* usado está disponível no repositório do projeto.

Tabela 2. Resultados dos experimentos nos cenários 1 e 2.

Cenário	Base de Dados	Modelo	ROUGE-L	BERTScore
1	FairyTaleQA	$PTT5_{Small}$	0,2491	0,3976
		$PTT5_{Base}$	**0,2699**	**0,4137**
		$PTT5_{Large}$	0,2668	0,4093
		$FLAN\text{-}T5_{Small}$	0,2412	0,3469
		$FLAN\text{-}T5_{Base}$	0,2497	0,3590
		$FLAN\text{-}T5_{Large}$	0,2354	0,3448
	PIRÁ	$PTT5_{Small}$	0,2109	0,2982
		$PTT5_{Base}$	0,2266	0,3265
		$PTT5_{Large}$	**0,2280**	**0,3449**
		$FLAN\text{-}T5_{Small}$	0,1581	0,2099
		$FLAN\text{-}T5_{Base}$	0,1723	0,2404
		$FLAN\text{-}T5_{Large}$	0,2219	0,2988
2	FairyTaleQA	$PTT5_{Small}$	0,4230	0,5429
		$PTT5_{Base}$	0,4786	0,5906
		$PTT5_{Large}$	**0,4938**	**0,6057**
		$FLAN\text{-}T5_{Small}$	0,3190	0,4203
		$FLAN\text{-}T5_{Base}$	0,3672	0,4611
		$FLAN\text{-}T5_{Large}$	0,3810	0,4884
	PIRÁ	$PTT5_{Small}$	0,2625	0,3635
		$PTT5_{Base}$	0,3506	0,4505
		$PTT5_{Large}$	**0,3640**	**0,4656**
		$FLAN\text{-}T5_{Small}$	0,1680	0,2220
		$FLAN\text{-}T5_{Base}$	0,1934	0,2579
		$FLAN\text{-}T5_{Large}$	0,2620	0,3257

perimento, mas com algumas diferenças importantes. O $PTT5_{Large}$ obteve o melhor desempenho em ambas as bases de dados, com 82,32% das perguntas geradas sendo consideradas válidas no FairyTaleQA e 93,06% no PIRÁ. É possível observar uma divergência entre as medidas automáticas e a avaliação $GPT\text{-}4o$, particularmente no PIRÁ. Enquanto as medidas do ROUGE-L e BERTScore indicaram valores menores para o PIRÁ em comparação com o FairyTaleQA, a avaliação $GPT\text{-}4o$ mostrou uma tendência oposta, com percentuais mais altos de perguntas válidas no PIRÁ.

Tabela 3. Resultados da avaliação do Cenário 1 usando o GPT-$4o$.

Base de Dados	Modelo	Válida	Inválida	% Válida
FairyTaleQA	$PTT5_{Small}$	680	327	67,53
	$PTT5_{Base}$	726	281	72,10
	$PTT5_{Large}$	**829**	178	**82,32**
	$FLAN\text{-}T5_{Small}$	451	556	44,79
	$FLAN\text{-}T5_{Base}$	525	482	52,14
	$FLAN\text{-}T5_{Large}$	719	288	71,40
PIRÁ	$PTT5_{Small}$	135	81	62,50
	$PTT5_{Base}$	199	17	92,13
	$PTT5_{Large}$	**201**	15	**93,06**
	$FLAN\text{-}T5_{Small}$	122	94	56,48
	$FLAN\text{-}T5_{Base}$	138	78	63,89
	$FLAN\text{-}T5_{Large}$	197	19	91,20

Na Figura 2 é apresentado um exemplo extraído da base de dados do PIRÁ, contendo o contexto, as perguntas geradas pelos modelos $PTT5$ e a saída da análise usando o $GPT\text{-}4o$. Nesse exemplo, é possível ver que os modelos $PTT5_{Large}$ e $PTT5_{Base}$ foram capazes de gerar perguntas que podem ser respondidas pelo contexto, sendo assim consideradas *válidas*. Por outro lado, o modelo $PTT5_{Small}$ gerou uma pergunta confusa sobre a Petróleo Brasileiro SA não ter comentado sobre a estimativa de produção revisada. Ape-

sar de ser mencionado no contexto, não está explícito nele o porquê disso. Sendo assim, considerada *inválida* pela avaliação do GPT-4o.

Contexto: O BG GROUP produziu recentemente uma nova estimativa na descoberta de óleo de Tupi na Bacia de Santos, afirmando que o campo contém 12-30 bilhões boe ou mais. Por um lado, a Petroleo Brasileiro SA de Petróleo (Petrobras) não comentou sobre a estimativa de produção revisada. Pode-se lembrar que a Petrobras relatou a descoberta para ser de 8 bilhões bbl de luz em bruto em 2007. Enquanto isso, alegações de BG foram produzidas em uma declaração sobre a estratégia de crescimento a longo prazo da empresa, lançada à frente da apresentação de seus resultados do quarto trimestre.

$PTT5_{Small}$: Por que a Petroleo Brasileiro SA de Petróleo (Petrobras) não comentou sobre a estimativa de produção revisada?
Avaliação: INVÁLIDA.

$PTT5_{Base}$: Qual foi a estimativa do BG Group para a descoberta de óleo de Tupi na Bacia de Santos?
Avaliação: VÁLIDA.

$PTT5_{Large}$: Qual a estimativa do BG Group para a quantidade de óleo de Tupi na Bacia de Santos?
Avaliação: VÁLIDA.

Figura 2. Exemplo de contexto extraído do PIRÁ e perguntas consideradas válidas e inválidas pelo GPT-4o.

5. Considerações Finais e Trabalhos Futuros

Neste trabalho, foi realizada uma análise comparativa dos modelos $PTT5$ e $FLAN$-$T5$ para a tarefa de geração automática de perguntas. Para isso, foram utilizadas as bases de dados do PIRÁ e uma versão traduzida do FairyTaleQA para o português do Brasil. O desempenho dos modelos foi avaliado usando uma abordagem tradicional, considerando as medidas de avaliação do ROUGE-L e do BERTScore. Além dessa abordagem, foi realizada uma análise das perguntas geradas pelos modelos usando o GPT-4o, avaliando se as perguntas geradas poderiam ser respondidas somente a partir do contexto fornecido. Os resultados experimentais demonstraram que o modelo $PTT5_{Large}$ obteve os melhores resultados em quase todos os cenários avaliados. Os resultados obtidos indicam a eficácia do pré-treinamento específico em português, evidenciada pelo desempenho superior consistente dos modelos $PTT5$ em comparação com os modelos $FLAN$-$T5$.

Apesar dos resultados encorajadores obtidos, o trabalho apresenta diversas limitações, que serão melhor exploradas. Dentre elas, pode-se destacar duas linhas de pesquisa futuras: **(i)** investigar o desempenho de LLMs, como o Llama 3 [Touvron et al. 2023], Gemma [Team et al. 2024] e Sabiá [Almeida et al. 2024]; e **(ii)** realizar uma avaliação humana para complementar as avaliações automáticas realizadas.

Agradecimentos

Os autores agradecem ao Ifes, apoio da FAPES e CAPES (processo 2021-2S6CD, nº FAPES 132/2021) por meio do PDPG (Programa de Desenvolvimento da Pós-Graduação, Parcerias Estratégicas nos Estados).

Referências

Almeida, T. S., Abonizio, H., Nogueira, R., and Pires, R. (2024). Sabiá-2: A new generation of portuguese large language models. *arXiv preprint arXiv:2403.09887*.

Carmo, D., Piau, M., Campiotti, I., Nogueira, R., and Lotufo, R. (2020). Ptt5: Pretraining and validating the t5 model on brazilian portuguese data. *arXiv preprint arXiv:2008.09144*.

Chen, J., Lin, H., Han, X., and Sun, L. (2024). Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17754–17762.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al. (2024). Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

da Rocha Junqueira, J., Corrêa, U. B., and Freitas, L. (2024). Transformer models for brazilian portuguese question generation: An experimental study. In *The International FLAIRS Conference Proceedings*, volume 37.

Kurdi, G., Leo, J., Parsia, B., Sattler, U., and Al-Emari, S. (2020). A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30:121–204.

Leite, B. and Lopes Cardoso, H. (2022). Neural question generation for the portuguese language: A preliminary study. In *EPIA Conference on Artificial Intelligence*, pages 780–793. Springer.

Leite, B., Osório, T. F., and Cardoso, H. L. (2024). Fairytaleqa translated: Enabling educational question and answer generation in less-resourced languages. *arXiv preprint arXiv:2406.04233*.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Mulla, N. and Gharpure, P. (2023). Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications. *Progress in Artificial Intelligence*, 12(1):1–32.

Oliveira, H. G., Caetano, I., Matos, R., and Amaro, H. (2023). Generating and ranking distractors for multiple-choice questions in portuguese. In *SLATE*, pages 4–1.

Paschoal, A. F., Pirozelli, P., Freire, V., Delgado, K. V., Peres, S. M., José, M. M., Nakasato, F., Oliveira, A. S., Brandão, A. A., Costa, A. H., et al. (2021). Pirá: A bilingual portuguese-english dataset for question-answering about the ocean. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 4544–4553.

Puri, R., Spring, R., Shoeybi, M., Patwary, M., and Catanzaro, B. (2020). Training question answering models from synthetic data. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5811–5826, Online. Association for Computational Linguistics.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M. S., Love, J., et al. (2024). Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Vaswani, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.

Wagner Filho, J. A., Wilkens, R., Idiart, M., and Villavicencio, A. (2018). The brwac corpus: A new open resource for brazilian portuguese. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Xu, Y., Wang, D., Yu, M., Ritchie, D., Yao, B., Wu, T., Zhang, Z., Li, T., Bradford, N., Sun, B., Hoang, T., Sang, Y., Hou, Y., Ma, X., Yang, D., Peng, N., Yu, Z., and Warschauer, M. (2022). Fantastic questions and where to find them: FairytaleQA – an authentic dataset for narrative comprehension. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 447–460, Dublin, Ireland. Association for Computational Linguistics.

Zhang, R., Guo, J., Chen, L., Fan, Y., and Cheng, X. (2021). A review on question generation from natural language text. *ACM Trans. Inf. Syst.*, 40(1).

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.