# Análise Técnica Comparativa de Modelos de Linguagem para Português: GlórIA, BERT, PTT5, Tucano e BERTimbau

## I. Introdução

### A. Contexto

O campo do Processamento de Linguagem Natural (PLN) tem testemunhado avanços significativos, impulsionados em grande parte pelo desenvolvimento e proliferação de Modelos de Linguagem de Grande Escala (LLMs). Inicialmente, muitos dos modelos mais proeminentes foram treinados predominantemente em corpora de língua inglesa, levando a uma lacuna de desempenho quando aplicados a outras línguas.1 Reconhecendo a necessidade de ferramentas robustas adaptadas às especificidades linguísticas e culturais de outros idiomas, a comunidade de pesquisa tem investido na criação de modelos dedicados. O português, com suas variantes europeia (PT-PT) e brasileira (PT-BR), representa um domínio importante para esse desenvolvimento.4 A trajetória nesse espaço evoluiu do aproveitamento de modelos multilingues, como o mBERT 8, para o pré-treinamento de modelos monolingues em corpora específicos do português 3, e mais recentemente para o treinamento de modelos generativos de larga escala nativamente em português.4

### B. Modelos Sob Revisão

Este relatório foca na análise técnica comparativa de cinco modelos de linguagem relevantes para o português, baseando-se exclusivamente nas informações contidas nos materiais de pesquisa fornecidos:

1. **GlórIA:** Um modelo decoder-only focado no Português Europeu (PT-PT). A análise se concentrará na variante GlórIA 1.3B, conforme detalhado nas fontes.13
2. **BERT:** Representando a arquitetura Bidirectional Encoder Representations from Transformers. Dada a sua presença frequente como linha de base em estudos portugueses, a análise considerará a arquitetura BERT Base e sua variante multilingue (mBERT).8
3. **PTT5:** O Portuguese Text-to-Text Transfer Transformer, uma adaptação da arquitetura T5 para o português. A análise abordará as versões base (v1 e v2, quando distinguíveis nas fontes) e mencionará outras variantes de tamanho conforme documentado.2 A menção a "XLNE ptt5" nas fontes será interpretada como referente ao PTT5, dada a ausência de um modelo distinto com esse nome e o contexto das discussões.26
4. **Tucano:** Uma série de modelos decoder-only treinados nativamente em um grande corpus português. O foco será nas variantes Tucano-1b1 e Tucano-2b4.5
5. **BERTimbau:** Uma versão do BERT pré-treinada especificamente para o Português Brasileiro (PT-BR), disponível nas variantes Base e Large.1

### C. Objetivo e Estrutura do Relatório

O objetivo principal deste relatório é fornecer uma comparação técnica detalhada e estruturada dos modelos listados, compilada exclusivamente a partir das informações presentes nos trechos de pesquisa fornecidos. A análise será apresentada principalmente através de uma tabela comparativa abrangente, seguida de elaborações sobre as características de cada modelo e uma discussão comparativa. É importante notar que a disponibilidade de certas métricas intrínsecas, como a Razão Tipo-Token (TTR), pontuações BLEU ou ROUGE do modelo base, ou análises explícitas de frequência de n-gramas, é limitada pelo conteúdo das fontes fornecidas e as lacunas serão explicitamente indicadas.

## II. Tabela Comparativa de Especificações Técnicas

A tabela a seguir consolida as especificações técnicas dos modelos de linguagem em análise, extraídas unicamente dos trechos de pesquisa fornecidos. Cada linha representa um modelo (ou variante específica), e as colunas detalham atributos como corpus de treinamento, tamanho, métricas de avaliação (quando disponíveis para o modelo base ou em benchmarks relevantes mencionados), janela de contexto e parâmetros arquitetônicos. Onde a informação não foi encontrada nas fontes fornecidas, o campo é marcado como "Não Especificado".

**Tabela 1: Especificações Técnicas Comparativas de Modelos de Linguagem para Português**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Modelo** | **Corpus de Treinamento** | **Tamanho do Corpus** | **TTR (Modelo Base)** | **BLEU (Tarefa Benchmark)** | **ROUGE (Tarefa Benchmark)** | **Perplexidade (Tarefa Benchmark)** | **Análise Freq. N-gramas** | **Janela de Contexto (Seq. Máx.)** | **Parâmetros Relevantes (Total, Camadas, Oculto, Cabeças)** | **Tokens Máx. Entrada** |
| **GlórIA (1.3B)** | Compilação PT-PT (ArquivoPT, OSCAR, ClueWeb-L, Wiki, OpenSubtitles, Europarl) 4 | ~35 Bilhões de tokens 4 | Não Especificado 4 | Não Especificado 4 | Não Especificado 4 | Monitorada no pré-treino.4 Avaliado via ExactMatch (52.79%) no CALAME-PT.4 | Não Especificado | 2048 15 | 1.3B Total; 24 Camadas; 2048 Oculto; 16 Cabeças 4 | 2048 15 |
| **BERT (Base/mBERT)** | Inglês (BookCorpus, Wiki) / Multilingue (Wiki ~104 línguas) 8 | Não Especificado (para mBERT) | Não Especificado | Não Especificado (modelo base) | Não Especificado (modelo base) | Não Especificado (modelo base) | Não Especificado | 512 8 | Base: ~110M Total; 12 Camadas; 768 Oculto; 12 Cabeças 9 | 512 8 |
| **PTT5 (Base v1/v2)** | v1: BrWaC; v2: mC4-pt 2 | v1: ~2.68B tokens (~15GB); v2: ~116B tokens/época (~524GB) 2 | Não Especificado | Fine-tuning: Tradução 53, QG 54 | Fine-tuning: Sumarização 32, QG 54 | Não Especificado (modelo base) | Não Especificado | 512 (no treino/fine-tuning) 2 | Base: ~220M Total; 12 Camadas (6 Enc, 6 Dec); 768 Oculto; 12 Cabeças 9 | 512 2 |
| **Tucano (1b1)** | GigaVerbo (Compilação PT) 5 | ~200B+ tokens (treinado até 515B) 5 | Não Especificado | Não Especificado (modelo base) | Não Especificado (modelo base) | Monitorada no pré-treino.37 Avaliado via ExactMatch/Acc. em benchmarks PT (CALAME-PT, LAMBADA-PT, etc.).33 | Não Especificado | 2048 33 | 1.1B Total; 22 Camadas; 2048 Oculto; 32 Cabeças 10 | 2048 33 |
| **Tucano (2b4)** | GigaVerbo (Compilação PT) 5 | ~200B+ tokens (treinado até 515B) 5 | Não Especificado | Não Especificado (modelo base) | Não Especificado (modelo base) | Monitorada no pré-treino.37 Avaliado via ExactMatch/Acc. em benchmarks PT (CALAME-PT, LAMBADA-PT, etc.).33 | Não Especificado | 4096 33 | 2.4B Total; 24 Camadas; 2560 Oculto; 16 Cabeças 10 | 4096 33 |
| **BERTimbau (Base)** | BrWaC (PT-BR) 8 | 2.68B tokens 47 | Não Especificado 47 | Não Especificado (modelo base) | Não Especificado (modelo base) | Não Especificado (modelo base).47 Perplexidade menor com sinônimos vs legendas em fine-tuning.39 | Implícita (tokenizer) 47 | 512 8 | 110M Total; 12 Camadas; 768 Oculto; 12 Cabeças 8 | 512 8 |
| **BERTimbau (Large)** | BrWaC (PT-BR) 8 | 2.68B tokens 47 | Não Especificado 47 | Não Especificado (modelo base) | Não Especificado (modelo base) | Não Especificado (modelo base) 47 | Implícita (tokenizer) 47 | 512 8 | 335M/340M Total; 24 Camadas; 1024 Oculto; 16 Cabeças 8 | 512 8 |

A compilação destes dados dispersos numa única tabela estruturada oferece um valor significativo, permitindo uma rápida comparação das escolhas arquitetônicas, escalas de dados e limites teóricos (como janela de contexto e contagem de parâmetros) que fundamentam as capacidades dos modelos. Esta centralização facilita decisões informadas sobre a seleção de modelos para tarefas específicas ou direções de pesquisa futuras no panorama do PLN em português, com base exclusivamente nas evidências fornecidas.

## III. Características Chave dos Modelos (Elaboração)

Esta seção aprofunda as características individuais de cada modelo, detalhando sua arquitetura, dados de treinamento, capacidades contextuais e desempenho em avaliações, conforme documentado nas fontes fornecidas.

### A. GlórIA (1.3B)

O modelo GlórIA representa um esforço significativo na criação de um LLM generativo focado especificamente no Português Europeu (PT-PT).4

* **Arquitetura:** GlórIA 1.3B é um modelo decoder-only baseado na arquitetura GPT-Neo 13, que se aproxima da arquitetura GPT-3.62 Possui 1.3 bilhão de parâmetros, distribuídos em 24 camadas (layers) de transformadores. A dimensão oculta (hidden size ou dmodel​) é de 2048, e cada camada de atenção utiliza 16 cabeças de atenção (attention heads).4 Esta configuração arquitetônica é otimizada para tarefas de geração de texto.
* **Corpus de Treinamento:** Um diferencial chave do GlórIA é seu corpus de pré-treinamento, composto exclusivamente por textos em Português Europeu, totalizando aproximadamente 35 bilhões de tokens.4 Este corpus foi cuidadosamente montado a partir de diversas fontes para garantir amplitude e diversidade, incluindo conteúdo genérico da web (OSCAR-2201 4, ClueWeb-L 22 4), notícias e periódicos arquivados do Arquivo.pt 14, conhecimento enciclopédico da Wikipedia 4, dados de diálogo do OpenSubtitles 4 e textos parlamentares do Europarl.14 Foi aplicado um processo de filtragem para garantir a qualidade e remover conteúdo indesejado, como a filtragem explícita de profanidades no corpus OpenSubtitles para evitar que o modelo aprendesse vieses tóxicos em primeira pessoa.4
* **Janela de Contexto:** O modelo foi treinado e opera com uma janela de contexto máxima de 2048 tokens.15 Isso define o comprimento máximo da sequência de entrada que o modelo pode processar de uma só vez.
* **Avaliação:** A avaliação do GlórIA 1.3B concentrou-se em suas capacidades de modelagem de linguagem. A perplexidade foi monitorada durante o pré-treinamento como um indicador de quão bem o modelo aprende a prever a sequência de texto.4 Para uma avaliação mais direta da capacidade generativa em contexto zero-shot, foi introduzido o benchmark CALAME-PT, inspirado no LAMBADA.4 Neste benchmark, GlórIA 1.3B alcançou uma pontuação ExactMatch (predição correta da última palavra da sentença, ignorando maiúsculas/minúsculas e acentos) de 52.79%, superando outros modelos abertos de tamanho comparável para português, como o mGPT (1.3B).4 Embora o potencial para tarefas downstream seja mencionado 7, as fontes fornecidas não reportam pontuações BLEU ou ROUGE para o modelo base, nem a métrica TTR.4
* **Posicionamento:** GlórIA 1.3B destaca-se como um LLM generativo de código aberto, treinado extensivamente em dados PT-PT, oferecendo uma alternativa robusta aos modelos multilingues ou focados no PT-BR para tarefas que exigem geração de texto em Português Europeu.

### B. BERT (Base/mBERT)

BERT (Bidirectional Encoder Representations from Transformers) é um modelo fundamental que revolucionou o PLN ao introduzir o pré-treinamento profundo bidirecional.8 Embora não seja um modelo específico para português, sua versão multilingue (mBERT) e sua arquitetura base servem frequentemente como ponto de comparação crucial em estudos de PLN para português.8

* **Arquitetura:** BERT utiliza exclusivamente a pilha de encoders da arquitetura Transformer.8 A variante **BERT Base** possui 12 camadas de Transformer, uma dimensão oculta (dmodel​) de 768, 12 cabeças de atenção e totaliza aproximadamente 110 milhões de parâmetros.8 A variante **BERT Large** é maior, com 24 camadas, dimensão oculta de 1024, 16 cabeças de atenção e cerca de 340 milhões de parâmetros.8
* **Corpus de Treinamento:** O BERT original foi pré-treinado em texto inglês (BookCorpus e Wikipedia inglesa).23 O **mBERT (Multilingual BERT)**, frequentemente usado como baseline em contextos não ingleses, foi pré-treinado em textos da Wikipedia de aproximadamente 104 línguas, incluindo o português.8 O tamanho exato do corpus do mBERT não é detalhado nas fontes, além da referência às principais línguas da Wikipedia.
* **Janela de Contexto:** Os modelos BERT, incluindo mBERT e variantes como BERTimbau, têm uma limitação no comprimento máximo da sequência de entrada, fixado em 512 tokens.8 Sequências mais longas precisam ser truncadas ou divididas.9
* **Avaliação:** BERT e mBERT são extensivamente utilizados como linhas de base para avaliar o desempenho de novos modelos em tarefas de compreensão de linguagem natural (NLU) em português, como Reconhecimento de Entidades Nomeadas (NER), Similaridade Textual Semântica (STS), Reconhecimento de Implicação Textual (RTE), análise de sentimentos e classificação de textos.8 Consistentemente, modelos treinados especificamente para o português, como o BERTimbau, demonstram desempenho superior ao mBERT nessas tarefas.8 As fontes fornecidas não detalham métricas intrínsecas como TTR, BLEU, ROUGE ou perplexidade para o modelo mBERT base.
* **Posicionamento:** BERT é uma arquitetura encoder-only fundamental. Sua variante multilingue, mBERT, embora superada por modelos monolingues, permanece uma referência importante para avaliar o progresso em modelos específicos para o português. Sua limitação de 512 tokens é uma consideração importante para tarefas que envolvem textos longos.

### C. PTT5 (Portuguese T5)

PTT5 (Portuguese Text-to-Text Transfer Transformer) representa a adaptação da arquitetura T5 do Google para a língua portuguesa, visando trazer as capacidades de geração e compreensão do T5 para este idioma.3

* **Arquitetura:** PTT5 adota a arquitetura encoder-decoder do T5.2 Esta arquitetura unifica diversas tarefas de PLN sob um formato "texto-para-texto", onde tanto a entrada quanto a saída são sequências de texto.3 A versão **PTT5 Base** geralmente possui 12 camadas de Transformer (tipicamente 6 no encoder e 6 no decoder), dimensão oculta (dmodel​) de 768, 12 cabeças de atenção e cerca de 220 milhões de parâmetros.9 Existem também outras variantes de tamanho, como Small (60M), Large (740M/770M) e 3B (3 Bilhões), refletindo diferentes trade-offs entre capacidade e custo computacional.2
* **Corpus de Treinamento:** Houve uma evolução significativa nos dados de treinamento do PTT5:
  + **PTT5 v1:** Esta versão inicial foi desenvolvida através da continuação do pré-treinamento (continued pretraining) dos modelos T5/mT5 originais utilizando o corpus **BrWaC** (Brazilian Web as Corpus).2 O BrWaC processado para este fim continha cerca de 15 GB de texto, correspondendo a aproximadamente 2.68 bilhões de tokens.3 Uma inovação importante foi a criação e uso de um vocabulário SentencePiece específico para o português, com 32.000 tokens, treinado sobre a Wikipedia em português, o que demonstrou impacto positivo no desempenho.2
  + **PTT5 v2:** Uma iteração mais recente continuou o pré-treinamento a partir dos checkpoints T5 do Google, mas utilizando um corpus substancialmente maior: o segmento português do **mC4 (multilingual Colossal Clean Crawled Corpus)**, referido como mC4-pt.2 Este dataset contém aproximadamente 524 GB de texto não comprimido, distribuído por 169 milhões de documentos. Um único epoch de pré-treinamento sobre o mC4-pt envolve cerca de 116 bilhões de tokens.30 O PTT5 v2 manteve o mesmo vocabulário português de 32k tokens da v1.2
* **Janela de Contexto:** As fontes indicam que o pré-treinamento e o fine-tuning do PTT5 foram realizados com um comprimento máximo de sequência de 512 tokens.2 Embora a arquitetura T5, com seus embeddings de posição relativa, possa teoricamente lidar com sequências mais longas, isso está sujeito a restrições de memória significativas, e o desempenho pode não ser garantido para comprimentos não vistos durante o treinamento.76
* **Avaliação:** PTT5 foi avaliado em uma ampla gama de tarefas de PLN em português, incluindo STS, RTE (usando o dataset ASSIN2), NER (HAREM), sumarização abstrativa, geração de perguntas (QG), tradução automática e classificação de sentimentos (TweetSentBR).2 Os resultados geralmente mostram que o PTT5 (tanto v1 quanto v2) supera significativamente os modelos T5/mT5 originais em tarefas portuguesas.3 PTT5 v2, em particular, alcançou resultados estado-da-arte (SOTA) em benchmarks como ASSIN2 RTE e TweetSentBR.26 Em algumas tarefas, como NER financeiro, modelos baseados em BERT como o BERTimbau mostraram desempenho superior ao PTT5.1 Métricas como BLEU e ROUGE são reportadas para tarefas de fine-tuning específicas (tradução 53, QG 54, sumarização 32), mas não para o modelo base. TTR e perplexidade do modelo base não são especificados nas fontes.
* **Posicionamento:** PTT5 é um modelo encoder-decoder versátil e poderoso para o português, demonstrando os benefícios da adaptação linguística através de pré-treinamento contínuo. Sua capacidade de lidar com tarefas de sequência-para-sequência o torna adequado para aplicações generativas como sumarização e tradução. A evolução para PTT5 v2, com um corpus massivamente maior, representa um avanço adicional em suas capacidades.

### D. Tucano (1b1, 2b4)

A série Tucano representa uma iniciativa recente para desenvolver modelos de linguagem generativos (decoder-only) treinados nativamente em português, utilizando um corpus de grande escala e disponibilizando os modelos abertamente.5

* **Arquitetura:** Os modelos Tucano são baseados em uma arquitetura Transformer decoder-only, inspirada no design do Llama.10 As variantes principais analisadas são:
  + **Tucano-1b1:** Possui 1.1 bilhão de parâmetros, 22 camadas, dimensão oculta de 2048 e 32 cabeças de atenção.10
  + **Tucano-2b4:** Possui 2.4 bilhões de parâmetros, 24 camadas, dimensão oculta de 2560 e 16 cabeças de atenção.10 Utilizam embeddings posicionais RoPE, normalização RMSNorm e ativação SiLU.37 O tokenizer empregado é o TeenyTinyLlama, com um vocabulário de 32.000 tokens.33
* **Corpus de Treinamento:** O pré-treinamento foi realizado sobre o **GigaVerbo**, um corpus massivo de português compilado e deduplicado pelos autores, que afirmam conter inicialmente 200 bilhões de tokens.5 O corpus final processado tem cerca de 780 GB e mais de 145 milhões de documentos.33 As fontes incluem uma variedade de datasets portugueses e porções portuguesas de datasets multilingues, como monoHPLT-PT, CrawlPT (CC-100, Oscar, BrWaC), mC4-pt, Common Crawl, BlogSet-BR, Wikipedia, CulturaX, LegalPT, entre outros.33 O modelo maior (2b4) foi treinado em até 515 bilhões de tokens.37
* **Janela de Contexto:** Há uma diferenciação entre as variantes: Tucano-160m, 630m e 1b1 foram treinados com uma janela de contexto de 2048 tokens, enquanto o maior modelo, Tucano-2b4, utilizou uma janela de 4096 tokens.33
* **Avaliação:** Os modelos Tucano foram avaliados em um conjunto de benchmarks para português, incluindo testes nativos como CALAME-PT e testes traduzidos como LAMBADA-PT, ARC-PT e HellaSwag-PT, além de benchmarks como ENEM Challenge e OAB Exams.33 Os resultados indicam que os modelos Tucano têm desempenho igual ou superior a outros modelos abertos portugueses e multilingues de tamanho similar.5 Uma observação crítica feita pelos autores é a aparente falta de correlação entre a quantidade de tokens ingeridos durante o pré-treinamento e o desempenho em vários dos benchmarks portugueses existentes, levantando questões sobre a adequação dessas avaliações para medir o progresso real na modelagem de linguagem fundamental.5 A perplexidade e a loss foram monitoradas durante o treinamento.37 Métricas como TTR, BLEU ou ROUGE para os modelos base não são reportadas nas fontes.
* **Posicionamento:** A série Tucano representa um avanço significativo no desenvolvimento de LLMs generativos abertos e nativamente treinados para o português, utilizando um dos maiores corpora compilados para o idioma até o momento. Oferecem janelas de contexto maiores que os modelos BERT/PTT5 e levantam discussões importantes sobre a avaliação de LLMs em português.

### E. BERTimbau (Base, Large)

BERTimbau é uma adaptação direta da arquitetura BERT, pré-treinada especificamente para o Português Brasileiro (PT-BR), visando melhorar o desempenho em tarefas de NLU neste idioma.8

* **Arquitetura:** Segue a arquitetura encoder-only do BERT.1
  + **BERTimbau Base:** Possui 12 camadas, dimensão oculta de 768, 12 cabeças de atenção e 110 milhões de parâmetros.8
  + **BERTimbau Large:** Possui 24 camadas, dimensão oculta de 1024, 16 cabeças de atenção e 335 ou 340 milhões de parâmetros (ambos os valores são mencionados nas fontes).8
* **Corpus de Treinamento:** Foi pré-treinado exclusivamente no corpus **BrWaC (Brazilian Web as Corpus)**, um grande corpus de português brasileiro.8 O tamanho reportado é de 2.68 bilhões de tokens.47 O treinamento utilizou a técnica de mascaramento de palavras inteiras (whole-word masking).44 Embora o tamanho exato do vocabulário não seja especificado, ele foi adaptado para o português, com uma fonte mencionando 30.000 unidades de subpalavras geradas via SentencePiece.47
* **Janela de Contexto:** Assim como o BERT padrão, o BERTimbau opera com um comprimento máximo de sequência de 512 tokens.8 O processo de treinamento envolveu sequências de até 128 tokens inicialmente, e depois 512 tokens.47
* **Avaliação:** No momento de seu lançamento, o BERTimbau alcançou desempenho estado-da-arte (SOTA) em várias tarefas downstream de NLU em português, como NER (nos datasets HAREM e MiniHAREM), STS e RTE (no dataset ASSIN2), superando consistentemente o mBERT.8 Foi subsequentemente aplicado e avaliado em diversas outras tarefas, incluindo análise de sentimentos 8, classificação de textos tóxicos 8, detecção de ironia 8, análise de textos jurídicos 40, previsão de pictogramas para comunicação alternativa 39, detecção de fake news 68, NER no domínio financeiro 1 e detecção de ideação suicida.19 Em tarefas como NER, frequentemente demonstra desempenho superior ao PTT5.1 Métricas intrínsecas como TTR e perplexidade do modelo base não são reportadas 47, embora a perplexidade tenha sido usada em avaliações de fine-tuning.39 Análise de n-gramas é apenas implícita no processo de tokenização.47 Não há relatos de BLEU/ROUGE para o modelo base.
* **Posicionamento:** BERTimbau estabeleceu-se como um modelo encoder fundamental e de alto desempenho para o Português Brasileiro. Seu sucesso demonstrou claramente o valor do pré-treinamento monolingue em um corpus relevante (BrWaC) para tarefas de compreensão, definindo baselines fortes no cenário de PLN do PT-BR.

## IV. Análises Comparativas e Observações

A análise conjunta das especificações e características dos modelos revela tendências e divergências significativas no desenvolvimento de LLMs para o português.

### A. Divergência Arquitetônica e Implicações Funcionais

Observa-se que os modelos analisados representam os três principais paradigmas arquitetônicos dos Transformers: Encoder-only (BERT, BERTimbau), Decoder-only (GlórIA, Tucano) e Encoder-Decoder (PTT5). Essa escolha arquitetônica fundamental não é arbitrária; ela direciona as principais forças e aplicações típicas de cada modelo. Modelos Encoder-only, processando a entrada de forma bidirecional, são inerentemente fortes em tarefas de compreensão da linguagem natural (NLU) que exigem um entendimento profundo do contexto, como classificação de texto, reconhecimento de entidades nomeadas (NER) e avaliação de similaridade semântica (STS).1 Por outro lado, modelos Decoder-only são otimizados para a geração de linguagem natural (NLG), prevendo o próximo token de forma sequencial, o que os torna adequados para tarefas como completar texto, escrita criativa e diálogo.10 A arquitetura Encoder-Decoder, como a do PTT5, combina ambas as capacidades, sendo naturalmente adequada para tarefas de sequência-para-sequência que mapeiam uma entrada para uma saída textual diferente, como tradução automática, sumarização e geração de perguntas.3 A seleção de um modelo deve, portanto, considerar primordialmente a natureza da tarefa alvo, alinhando-a com as capacidades intrínsecas da arquitetura escolhida.

### B. Evolução dos Corpora de Treinamento: Escala e Especificidade

Uma tendência clara emerge da análise dos corpora utilizados: um movimento progressivo em direção a datasets maiores e mais específicos para o português. As primeiras abordagens ou adaptavam modelos multilingues treinados em corpora genéricos como a Wikipedia (caso do mBERT) 9 ou utilizavam o corpus BrWaC, com cerca de 2.7 bilhões de tokens, para treinar os primeiros modelos monolingues dedicados como BERTimbau e PTT5 v1.3 Esforços subsequentes escalaram drasticamente o volume de dados. O PTT5 v2 utilizou o mC4-pt, com centenas de gigabytes e dezenas ou centenas de bilhões de tokens.2 GlórIA focou na compilação de um corpus PT-PT de 35 bilhões de tokens.4 A série Tucano foi treinada no GigaVerbo, uma concatenação massiva que visa ultrapassar 200 bilhões de tokens.5 Este aumento na escala reflete tanto a crescente disponibilidade de dados textuais em português quanto o reconhecimento de que, assim como no desenvolvimento de LLMs para inglês, datasets maiores e mais diversificados são geralmente necessários para aprimorar as capacidades e a generalização dos modelos.4 Contudo, essa escalada também traz desafios relacionados à garantia da qualidade dos dados, como evidenciado pelos esforços de filtragem mencionados para GlórIA, PTT5 v2 e Tucano 2, e levanta questões sobre a possibilidade de saturação de benchmarks ou rendimentos decrescentes, conforme sugerido na análise do Tucano.5 A especificidade do corpus (PT-BR no BrWaC vs. PT-PT no GlórIA vs. compilação mista no GigaVerbo) também indica estratégias distintas em relação às variantes da língua e fontes de dados.

### C. Escala de Parâmetros e Trade-offs

Os modelos abrangem uma vasta gama de tamanhos, desde dezenas de milhões de parâmetros (PTT5-small, BERT-Base, BERTimbau-Base) 9, passando por centenas de milhões (PTT5-Base, BERTimbau-Large, PTT5-Large) 9, até bilhões de parâmetros (GlórIA 1.3B, Tucano 1.1B, 2.4B, PTT5 3B).2 Esta variação alinha-se à tendência geral em LLMs, onde modelos maiores tendem a apresentar melhor desempenho e capacidade de generalização, desde que treinados com dados e computação suficientes. A disponibilização de múltiplos tamanhos para modelos como BERTimbau, PTT5 e Tucano oferece flexibilidade, permitindo aos usuários escolher um modelo que equilibre desempenho com os recursos computacionais disponíveis para treinamento e, crucialmente, para inferência. Modelos maiores implicam custos computacionais significativamente mais elevados, tanto em termos de memória quanto de tempo de processamento.9 Embora o ganho de desempenho com a escala seja geralmente esperado 30, a análise do Tucano sugere que a relação entre tamanho/tokens e desempenho em benchmarks específicos pode não ser estritamente linear, possivelmente devido a limitações nos próprios benchmarks.5

### D. Expansão da Janela de Contexto

Uma diferença arquitetônica notável entre os modelos mais antigos e os mais recentes é o tamanho da janela de contexto (comprimento máximo da sequência). Modelos baseados em BERT (BERT, BERTimbau) e as versões documentadas do PTT5 estão tipicamente restritos a 512 tokens.2 Em contraste, modelos decoder mais recentes como GlórIA e Tucano foram projetados com janelas de contexto significativamente maiores: 2048 tokens para GlórIA 1.3B e Tucano 1b1, e 4096 tokens para Tucano 2b4.15 Esta evolução reflete uma tendência importante no desenvolvimento de LLMs: a busca por capacidade de processar sequências de texto mais longas.90 A capacidade de lidar com milhares de tokens de uma só vez permite que modelos como GlórIA e Tucano abordem tarefas envolvendo documentos extensos (sumarização de artigos, resposta a perguntas sobre textos longos, geração de narrativas) de forma mais direta e potencialmente mais eficaz do que modelos limitados a 512 tokens, que exigiriam estratégias de divisão ou truncamento do texto.67 Essa capacidade expandida, no entanto, está associada a um aumento na demanda computacional, em parte devido à complexidade quadrática dos mecanismos de atenção padrão em relação ao comprimento da sequência.66 A escolha entre um modelo com janela de 512 tokens e um com 2k/4k tokens envolve, portanto, um trade-off entre capacidade contextual e custo/latência computacional.

### E. Métricas de Avaliação e Desafios de Benchmarking

A avaliação do desempenho dos modelos é fortemente dependente da tarefa específica. As fontes reportam métricas como F1-score para tarefas de classificação e NER 9, correlação de Pearson para STS 31, Acurácia para classificação e RTE 19, e ExactMatch para avaliação de modelagem de linguagem.4 A Perplexidade é mencionada principalmente no contexto do monitoramento do pré-treinamento 4 ou em avaliações de fine-tuning.39 Métricas comuns de geração como BLEU 80 e ROUGE 54 aparecem principalmente em avaliações de tarefas downstream como tradução 53, sumarização 32 ou geração de perguntas 54, e não como métricas padrão para os modelos base nestas fontes. A Razão Tipo-Token (TTR), uma medida de diversidade lexical, não é mencionada.4 Esta ausência de relato padronizado de métricas intrínsecas nos materiais fornecidos dificulta uma comparação direta das capacidades fundamentais de modelagem de linguagem ou diversidade lexical entre os modelos base. Adicionalmente, o estudo do Tucano levanta uma questão importante ao observar uma correlação fraca entre o desempenho em alguns benchmarks portugueses existentes e a escala do pré-treinamento (tokens ingeridos).5 Isso sugere que os benchmarks atuais podem não capturar totalmente as melhorias obtidas com modelos maiores ou mais treinados, indicando um possível desafio no próprio cenário de avaliação de PLN em português.

## V. Conclusão

### A. Sumário dos Achados

Este relatório apresentou uma análise técnica comparativa de cinco modelos de linguagem relevantes para o português: GlórIA, BERT (como baseline/mBERT), PTT5, Tucano e BERTimbau, com base exclusivamente nas informações extraídas das fontes fornecidas. A análise revelou diferenças arquitetônicas chave (Encoder-only, Decoder-only, Encoder-Decoder) que influenciam diretamente as aplicações primárias de cada modelo, desde tarefas de compreensão (BERT, BERTimbau) a tarefas de geração (GlórIA, Tucano) e sequência-para-sequência (PTT5). Observou-se uma clara tendência de aumento na escala dos corpora de treinamento ao longo do tempo, passando de alguns bilhões de tokens (BrWaC) para centenas de bilhões (mC4-pt, GigaVerbo), refletindo a maturação do campo e a busca por maior capacidade. Similarmente, os modelos exibem uma ampla gama de contagens de parâmetros, oferecendo trade-offs entre desempenho e custo computacional. Uma evolução notável é a expansão da janela de contexto nos modelos mais recentes (GlórIA, Tucano), permitindo o processamento de sequências significativamente mais longas que os limites anteriores de 512 tokens (BERT, PTT5, BERTimbau). A avaliação, no entanto, mostrou-se fragmentada e dependente da tarefa, com uma carência de métricas intrínsecas padronizadas para os modelos base e questionamentos sobre a adequação de alguns benchmarks portugueses existentes.

### B. Posicionamento dos Modelos

Com base nos dados compilados:

* **BERT/mBERT:** Servem como baselines multilingues fundamentais, embora geralmente superados por modelos monolingues em tarefas específicas do português.
* **BERTimbau:** Estabeleceu-se como um forte modelo encoder para NLU em Português Brasileiro, eficaz em tarefas de compreensão e classificação, superando o mBERT.
* **PTT5:** É um modelo encoder-decoder versátil para português, adequado para tarefas sequência-para-sequência como sumarização e tradução, com a versão v2 representando um avanço em escala de dados.
* **GlórIA:** Posiciona-se como um LLM generativo de larga escala focado no Português Europeu, treinado em um corpus específico e avaliado com um benchmark customizado.
* **Tucano:** Representa a mais recente geração de modelos decoder-only nativamente treinados em português, utilizando um corpus massivo e oferecendo janelas de contexto expandidas, ao mesmo tempo que levanta questões sobre avaliação.

### C. Limitações e Direções Futuras

É crucial reiterar que esta análise está estritamente limitada às informações contidas nos trechos de pesquisa fornecidos. A ausência de métricas intrínsecas padronizadas (TTR, perplexidade base comparável, BLEU/ROUGE base) impede uma comparação completa das capacidades fundamentais de modelagem de linguagem e diversidade lexical. As observações do estudo Tucano sobre a correlação entre escala e desempenho em benchmarks existentes também sugerem que o cenário de avaliação para o português pode necessitar de refinamento para capturar adequadamente o progresso dos modelos fundamentais. Uma compreensão completa e uma comparação definitiva exigiriam acesso aos artigos originais completos, documentação técnica adicional ou, idealmente, a execução de avaliações padronizadas em todos os modelos sob condições controladas. Pesquisas futuras poderiam focar no estabelecimento de benchmarks mais robustos e na avaliação sistemática desses modelos em um conjunto comum de tarefas e métricas intrínsecas para fornecer uma imagem mais clara do estado da arte em LLMs para a língua portuguesa.

Claro, aqui está a tabela detalhando as principais características, vantagens e desvantagens dos benchmarks de viés discutidos:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Benchmark** | **Objetivo Principal** | **Tipo de Viés Medido** | **Metodologia** | **Métricas Chave** | **Vantagens** | **Desvantagens/Limitações** |
| **Gender Shades** | Avaliar viés interseccional (gênero x tipo de pele) em classificadores faciais comerciais.1 | Desempenho desigual (Viés Tipo I) entre grupos interseccionais.1 | Testar APIs comerciais com dataset PPB (parlamentares, balanceado); comparar taxas de erro entre 4 subgrupos (mulheres/homens x pele clara/escura).1 | Taxa de erro, Precisão, Diferença na taxa de erro (Disparidade), Maior diferença.1 | - Foco interseccional pioneiro.1&lt;br>- Uso de dataset mais balanceado (PPB).1&lt;br>- Impacto na indústria (transparência, melhorias).3 | - Classificação binária de gênero.1&lt;br>- Escala Fitzpatrick como proxy para tipo de pele.1&lt;br>- Escopo limitado (3 APIs, 1 tarefa, 1 momento).1&lt;br>- Dataset PPB não mais público.4&lt;br>- Foco em precisão, não em outros danos.1 |
| **COMPAS** | Avaliar viés racial na ferramenta de avaliação de risco de reincidência criminal.6 | Viés racial (disparidade em taxas de erro - FPR/FNR).6 | (Análise ProPublica) Comparar scores COMPAS com reincidência real (prisão em 2 anos) para réus negros vs. brancos; focar em FPR/FNR; controlar variáveis.7 | Taxa de Falsos Positivos (FPR), Taxa de Falsos Negativos (FNR), Acurácia, Paridade Preditiva (discutida no debate).7 | - Iniciou debate crucial sobre viés algorítmico na justiça.9&lt;br>- Dados (processados) publicamente disponíveis.10 | - Algoritmo proprietário (caixa preta).11&lt;br>- Críticas à metodologia ProPublica (erro processamento 13, foco em métricas 14).&lt;br>- **Dataset inerentemente problemático:** viés de medição, proxy ruim para crime, desconexão com mundo real.15&lt;br>- **Uso desencorajado para benchmarking de novos algoritmos**.15 |
| **Civil Comments / Jigsaw Unintended Bias** | Medir viés não intencional em classificadores de toxicidade (associação errônea de termos de identidade com toxicidade).16 | Viés de identidade/correlação (Viés Tipo II levando a danos Tipo I).16 | Avaliar modelos em dados anotados para toxicidade e menções de identidade; usar métricas AUC especializadas (Subgroup, BPSN, BNSP).16 | AUROC Geral, Subgroup AUC, BPSN AUC, BNSP AUC, Média Generalizada (GMB) dessas AUCs.16 | - Foco em viés sutil (não intencional).16&lt;br>- Métricas específicas para esse viés.16&lt;br>- Dataset grande e público.16&lt;br>- Anotações múltiplas por comentário.16 | - Subjetividade da toxicidade.17&lt;br>- Agregação de rótulos mascara discordância.1&lt;br>- Dados desbalanceados.9&lt;br>- Foco em *menções* de identidade, não autoria.16 |
| **BOLD** | Avaliar vieses sociais (profissão, gênero, raça, religião, política) em geração de texto aberto.20 | Vieses sociais (sentimento, toxicidade, consideração, estereótipos) em texto gerado (Viés Tipo II).19 | Usar prompts da Wikipédia para gerar texto; analisar com métricas automatizadas (sentimento, toxicidade, regard, normas psicolinguísticas, polaridade de gênero); comparar entre subgrupos e com baseline Wikipédia.19 | Sentimento (VADER), Toxicidade (BERT), Consideração/Regard (BERT), Normas Psicolinguísticas (VAD, BE5), Polaridade de Gênero (unigramas, embeddings).19 | - Foco em geração aberta (realista).19&lt;br>- Dataset grande com prompts naturais (Wikipedia).19&lt;br>- Múltiplas métricas para diferentes facetas do viés.19 | - Dependência de classificadores automáticos (modelos de métrica) que podem ser enviesados.6&lt;br>- Prompts da Wikipédia podem herdar seus vieses.19&lt;br>- Foco no inglês e contexto cultural dos EUA.19&lt;br>- Cobertura limitada de domínios/grupos.19 |
| **StereoSet** | Medir vieses estereotipados (gênero, profissão, raça, religião) em PLMs, contrastando com capacidade de modelagem da linguagem.22 | Associação estereotipada (Viés Tipo II).22 | Context Association Test (CAT): tarefas intrasentencial (preencher lacuna) e intersentencial (escolher continuação) com opções estereotipada, anti-estereotipada e sem sentido.23 | Language Modeling Score (lms), Stereotype Score (ss), Idealized CAT Score (icat).23 | - Dataset natural em larga escala (crowdsourced).22&lt;br>- Considera capacidade de modelagem da linguagem (lms).22&lt;br>- Cobre múltiplos domínios de viés.22 | - Foco nos EUA.22&lt;br>- Dificuldade em obter associações neutras.22&lt;br>- **Questões de validade/qualidade dos dados levantadas**.24&lt;br>- Inconsistências com outros benchmarks.24&lt;br>- Forte correlação entre lms e ss (modelos melhores são mais enviesados).22 |
| **CrowS-Pairs** | Medir vieses sociais explícitos (estereótipos) em MLMs contra grupos desfavorecidos nos EUA (9 categorias).27 | Associação estereotipada (Viés Tipo II).27 | Comparar pares de sentenças minimamente diferentes (estereotipada vs. menos estereotipada) usando pontuação de pseudo-log-verossimilhança (PLL) condicional.27 | % Preferência Estereótipo (ideal 50%).27 | - Crowdsourced (diversidade de estereótipos/estruturas).27&lt;br>- Cobre 9 categorias de viés.27&lt;br>- Foco explícito em grupos desfavorecidos.27 | - **Críticas significativas sobre validade/confiabilidade dos dados (Blodgett et al., 2021)**.32&lt;br>- Metodologia de pontuação questionada.32&lt;br>- Foco nos EUA.27&lt;br>- Inconsistências com outros benchmarks.24 |

#### Referências citadas

1. [2403.12212] Evaluating Named Entity Recognition: A comparative analysis of mono- and multilingual transformer models on a novel Brazilian corporate earnings call transcripts dataset - arXiv, acessado em abril 28, 2025, <https://arxiv.org/abs/2403.12212>
2. ptt5-v2: A Closer Look at Continued Pretraining of T5 Models for the Portuguese Language - arXiv, acessado em abril 28, 2025, <https://arxiv.org/pdf/2406.10806>
3. (PDF) PTT5: Pretraining and validating the T5 model on Brazilian Portuguese data, acessado em abril 28, 2025, <https://www.researchgate.net/publication/343825570_PTT5_Pretraining_and_validating_the_T5_model_on_Brazilian_Portuguese_data>
4. aclanthology.org, acessado em abril 28, 2025, <https://aclanthology.org/2024.propor-1.45.pdf>
5. [2411.07854] Tucano: Advancing Neural Text Generation for Portuguese - arXiv, acessado em abril 28, 2025, <https://arxiv.org/abs/2411.07854>
6. PTT5: Pretraining and validating the T5 model on Brazilian Portuguese data - arXiv, acessado em abril 28, 2025, <https://arxiv.org/pdf/2008.09144>
7. GlórIA: A Generative and Open Large Language Model for Portuguese Pre-print - arXiv, acessado em abril 28, 2025, <https://arxiv.org/html/2402.12969v1>
8. arXiv:2201.03382v1 [cs.CL] 10 Jan 2022, acessado em abril 28, 2025, <https://arxiv.org/pdf/2201.03382>
9. Evaluating Named Entity Recognition: Comparative Analysis of Mono- and Multilingual Transformer Models on Brazilian Corporate Earnings Call Transcriptions - arXiv, acessado em abril 28, 2025, <https://arxiv.org/html/2403.12212v1>
10. [Literature Review] Tucano: Advancing Neural Text Generation for Portuguese - Moonlight, acessado em abril 28, 2025, <https://www.themoonlight.io/en/review/tucano-advancing-neural-text-generation-for-portuguese>
11. (PDF) Tucano: Advancing Neural Text Generation for Portuguese - ResearchGate, acessado em abril 28, 2025, <https://www.researchgate.net/publication/385750166_Tucano_Advancing_Neural_Text_Generation_for_Portuguese>
12. Tucano: Advancing Neural Text Generation for Portuguese - arXiv, acessado em abril 28, 2025, <https://arxiv.org/html/2411.07854v1>
13. GlorIA 1.3B · Models - Dataloop, acessado em abril 28, 2025, <https://dataloop.ai/library/model/nova-vision-language_gloria-13b/>
14. NOVA-vision-language/GlorIA-1.3B - Hugging Face, acessado em abril 28, 2025, <https://huggingface.co/NOVA-vision-language/GlorIA-1.3B>
15. GlorIA 1.3B By NOVA-vision-language - LLM Explorer - EXTRACTUM, acessado em abril 28, 2025, <https://llm.extractum.io/model/NOVA-vision-language%2FGlorIA-1.3B,3ZeM7LVJEHZlAQatN2xr5i>
16. [2305.14070] Assessing Linguistic Generalisation in Language Models: A Dataset for Brazilian Portuguese - arXiv, acessado em abril 28, 2025, <https://arxiv.org/abs/2305.14070>
17. Towards Transfer Learning Techniques—BERT, DistilBERT, BERTimbau, and DistilBERTimbau for Automatic Text Classification from Different Languages: A Case Study - PMC, acessado em abril 28, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC9655936/>
18. Evaluating Named Entity Recognition: A comparative analysis of mono- and multilingual transformer models on a novel Brazilian corporate earnings call transcripts dataset - Papers With Code, acessado em abril 28, 2025, <https://paperswithcode.com/paper/evaluating-named-entity-recognition/review/>
19. Comparative analysis of BERT-based and generative large language models for detecting suicidal ideation: a performance evaluation study - PubMed, acessado em abril 28, 2025, <https://pubmed.ncbi.nlm.nih.gov/39607132/>
20. Evaluating Named Entity Recognition: A Comparative Analysis of Mono- and Multilingual Transformer Models on a Novel Brazilian Corporate Earnings Call Transcripts Dataset - arXiv, acessado em abril 28, 2025, <https://arxiv.org/html/2403.12212v2>
21. Performance comparison between BERTimbau Large and BERTimbau Base: MWEs that were not predicted by either model. - ResearchGate, acessado em abril 28, 2025, <https://www.researchgate.net/figure/Performance-comparison-between-BERTimbau-Large-and-BERTimbau-Base-MWEs-that-were-not_tbl3_370981430>
22. Pretrained models — transformers 2.10.0 documentation - Hugging Face, acessado em abril 28, 2025, <https://huggingface.co/transformers/v2.10.0/pretrained_models.html>
23. Large Language Models: Comparing Gen 1 Models (GPT, BERT, T5 and More), acessado em abril 28, 2025, <https://dev.to/admantium/large-language-models-comparing-gen-1-models-gpt-bert-t5-and-more-74h>
24. PTT5: Pretraining and validating the T5 model on Brazilian Portuguese data - Hugging Face, acessado em abril 28, 2025, <https://huggingface.co/papers/2008.09144>
25. [2008.09144] PTT5: Pretraining and validating the T5 model on Brazilian Portuguese data, acessado em abril 28, 2025, <https://arxiv.org/abs/2008.09144>
26. [2406.10806] ptt5-v2: A Closer Look at Continued Pretraining of T5 Models for the Portuguese Language - arXiv, acessado em abril 28, 2025, <https://arxiv.org/abs/2406.10806>
27. unicamp-dl/ptt5-v2-base - Hugging Face, acessado em abril 28, 2025, <https://huggingface.co/unicamp-dl/ptt5-v2-base>
28. unicamp-dl/ptt5-base-portuguese-vocab - Hugging Face, acessado em abril 28, 2025, <https://huggingface.co/unicamp-dl/ptt5-base-portuguese-vocab>
29. unicamp-dl/ptt5-small-t5-vocab - Hugging Face, acessado em abril 28, 2025, <https://huggingface.co/unicamp-dl/ptt5-small-t5-vocab>
30. ptt5-v2: A Closer Look at Continued Pretraining of T5 Models for the Portuguese Language, acessado em abril 28, 2025, <https://arxiv.org/html/2406.10806v2>
31. ptt5-v2: A Closer Look at Continued Pretraining of T5 Models for the Portuguese Language, acessado em abril 28, 2025, <https://arxiv.org/html/2406.10806v1>
32. Ptt5 Base Summ Xlsum · Models - Dataloop, acessado em abril 28, 2025, <https://dataloop.ai/library/model/recogna-nlp_ptt5-base-summ-xlsum/>
33. Tucano - Nicholas Kluge Corrêa, acessado em abril 28, 2025, <https://nkluge-correa.github.io/Tucano/>
34. Nkluge-correa/Tucano: Natively pre-trained open-source Portuguese language models. - GitHub, acessado em abril 28, 2025, <https://github.com/Nkluge-correa/Tucano>
35. TucanoBR/Tucano-630m - Hugging Face, acessado em abril 28, 2025, <https://huggingface.co/TucanoBR/Tucano-630m>
36. New publication on Portugese NLP - Center for Science and Thought - Universität Bonn, acessado em abril 28, 2025, <https://www.cst.uni-bonn.de/en/en_news/new-publication-on-portugese-nlp>
37. Tucano: A Series of Decoder-Transformers Natively Pre-Trained in Portuguese, acessado em abril 28, 2025, <https://www.marktechpost.com/2024/11/28/tucano-a-series-of-decoder-transformers-natively-pre-trained-in-portuguese/>
38. TucanoBR/Tucano-1b1-Instruct - Hugging Face, acessado em abril 28, 2025, <https://huggingface.co/TucanoBR/Tucano-1b1-Instruct>
39. [2308.09497] Predictive Authoring for Brazilian Portuguese Augmentative and Alternative Communication - arXiv, acessado em abril 28, 2025, <http://arxiv.org/abs/2308.09497>
40. [2203.05425] Semantic Norm Recognition and its application to Portuguese Law - arXiv, acessado em abril 28, 2025, <https://arxiv.org/abs/2203.05425>
41. Towards Transfer Learning Techniques—BERT, DistilBERT, BERTimbau, and DistilBERTimbau for Automatic Text Classification from Different Languages: A Case Study - MDPI, acessado em abril 28, 2025, <https://www.mdpi.com/1424-8220/22/21/8184>
42. neuralmind/bert-base-portuguese-cased - Hugging Face, acessado em abril 28, 2025, <https://huggingface.co/neuralmind/bert-base-portuguese-cased>
43. neuralmind/bert-large-portuguese-cased - Hugging Face, acessado em abril 28, 2025, <https://huggingface.co/neuralmind/bert-large-portuguese-cased>
44. neuralmind-ai/portuguese-bert: Portuguese pre-trained BERT models - GitHub, acessado em abril 28, 2025, <https://github.com/neuralmind-ai/portuguese-bert>
45. bert-large-portuguese-cased | AI Model Details - AIModels.fyi, acessado em abril 28, 2025, <https://www.aimodels.fyi/models/huggingFace/bert-large-portuguese-cased-neuralmind>
46. BERTimbau - Portuguese BERT-Base language model - PORTULAN CLARIN, acessado em abril 28, 2025, <https://portulanclarin.net/repository/browse/bertimbau-portuguese-bert-base-language-model/486edb32e93711ebabf702420a8701536b383b588f2f4c85b13b4d04c2867a4b/>
47. (PDF) BERTimbau: Pretrained BERT Models for Brazilian Portuguese, acessado em abril 28, 2025, <https://www.researchgate.net/publication/345395208_BERTimbau_Pretrained_BERT_Models_for_Brazilian_Portuguese>
48. rvlopes/GlorIA: Source code for GlorIA models pre-training. - GitHub, acessado em abril 28, 2025, <https://github.com/rvlopes/GlorIA>
49. what is the max length of the context? · Issue #190 · google-research/bert - GitHub, acessado em abril 28, 2025, <https://github.com/google-research/bert/issues/190>
50. HuggingFace Config Params Explained - Mostly on AI, acessado em abril 28, 2025, <https://dejanbatanjac.github.io/huggingface-config>
51. Bert Model Size Comparison | Restackio, acessado em abril 28, 2025, <https://www.restack.io/p/ai-model-answer-bert-model-size-comparison-cat-ai>
52. BRWAC Dataset - Papers With Code, acessado em abril 28, 2025, <https://paperswithcode.com/dataset/brwac>
53. arXiv:2008.08769v1 [cs.CL] 20 Aug 2020, acessado em abril 28, 2025, <https://arxiv.org/pdf/2008.08769>
54. Neural Question Generation for the Portuguese Language: A Preliminary Study, acessado em abril 28, 2025, <https://www.researchgate.net/publication/363500878_Neural_Question_Generation_for_the_Portuguese_Language_A_Preliminary_Study>
55. RecognaSumm: A Novel Brazilian Summarization Dataset - ACL Anthology, acessado em abril 28, 2025, <https://aclanthology.org/2024.propor-1.63.pdf>
56. Benchmarks for Pirá 2.0, a Reading Comprehension Dataset about the Ocean, the Brazilian Coast, and Climate Change | Data Intelligence - MIT Press Direct, acessado em abril 28, 2025, <https://direct.mit.edu/dint/article/6/1/29/120001/Benchmarks-for-Pira-2-0-a-Reading-Comprehension>
57. T5Backbone model - Keras, acessado em abril 28, 2025, <https://keras.io/keras_hub/api/models/t5/t5_backbone/>
58. pretrained BERT models for Brazilian Portuguese BERTimbau: modelos BERT pré-treinados para Português Brasileiro - CORE, acessado em abril 28, 2025, <https://core.ac.uk/download/427533232.pdf>
59. New State-of-the-Art for Question Answering on Portuguese SQuAD v1.1, acessado em abril 28, 2025, <https://sol.sbc.org.br/index.php/kdmile/article/download/24974/24795>
60. Metrics of BERTimbau Large (BERT-CRF) on Named Entity Recognition task... | Download Scientific Diagram - ResearchGate, acessado em abril 28, 2025, <https://www.researchgate.net/figure/Metrics-of-BERTimbau-Large-BERT-CRF-on-Named-Entity-Recognition-task-on-the-test_fig2_345395208>
61. GlórIA: A Generative and Open Large Language Model for Portuguese - ACL Anthology, acessado em abril 28, 2025, <https://aclanthology.org/2024.propor-1.45/>
62. V-GlórIA: Customizing Large Vision and Language Models to European Portuguese - ACL Anthology, acessado em abril 28, 2025, <https://aclanthology.org/2024.customnlp4u-1.24.pdf>
63. GlórIA -- A Generative and Open Large Language Model for Portuguese - arXiv, acessado em abril 28, 2025, <https://arxiv.org/abs/2402.12969>
64. GlórIA: the new Portuguese-European Large Language Model - NOVA Lincs, acessado em abril 28, 2025, <https://nova-lincs.di.fct.unl.pt/gloria-the-new-portuguese-european-large-language-model/>
65. Fine-tuning BERT with sequences longer than 512 tokens - Models - Hugging Face Forums, acessado em abril 28, 2025, <https://discuss.huggingface.co/t/fine-tuning-bert-with-sequences-longer-than-512-tokens/12652>
66. [D] Why is the maximum input sequence length of BERT is restricted to 512? : r/MachineLearning - Reddit, acessado em abril 28, 2025, <https://www.reddit.com/r/MachineLearning/comments/genjvb/d_why_is_the_maximum_input_sequence_length_of/>
67. How can I use BERT for long text classification? - nlp - Stack Overflow, acessado em abril 28, 2025, <https://stackoverflow.com/questions/58636587/how-can-i-use-bert-for-long-text-classification>
68. Talendar/br\_fake\_news\_detection: Fake news detection on texts written in Brazilian Portuguese. - GitHub, acessado em abril 28, 2025, <https://github.com/Talendar/br_fake_news_detection>
69. Comparative analysis of BERT-based and generative large language models for detecting suicidal ideation: a performance evaluation study - SciELO, acessado em abril 28, 2025, <https://www.scielo.br/j/csp/a/XrbVfvybPj9tvJ8qWv7j8VC/?lang=en>
70. eduagarcia/RoBERTaCrawlPT-base - Hugging Face, acessado em abril 28, 2025, <https://huggingface.co/eduagarcia/RoBERTaCrawlPT-base>
71. T5 - Hugging Face, acessado em abril 28, 2025, <https://huggingface.co/docs/transformers/v4.13.0/model_doc/t5>
72. transformers/src/transformers/models/t5/configuration\_t5.py at main - GitHub, acessado em abril 28, 2025, <https://github.com/huggingface/transformers/blob/main/src/transformers/models/t5/configuration_t5.py>
73. T5 - Hugging Face, acessado em abril 28, 2025, <https://huggingface.co/docs/transformers/model_doc/t5>
74. What is the T5-Model? - Data Basecamp, acessado em abril 28, 2025, <https://databasecamp.de/en/ml-blog/t5-model>
75. ptt5-v2 - a unicamp-dl Collection - Hugging Face, acessado em abril 28, 2025, <https://huggingface.co/collections/unicamp-dl/ptt5-v2-666538a650188ba00aa8d2d0>
76. What is maximum sequence length that can be used with pretrained T5 (3b model) checkpoint? · Issue #5204 · huggingface/transformers - GitHub, acessado em abril 28, 2025, <https://github.com/huggingface/transformers/issues/5204>
77. recogna-nlp/ptt5-base-summ - Hugging Face, acessado em abril 28, 2025, <https://huggingface.co/recogna-nlp/ptt5-base-summ>
78. recogna-nlp/ptt5-base-summ-xlsum - Hugging Face, acessado em abril 28, 2025, <https://huggingface.co/recogna-nlp/ptt5-base-summ-xlsum>
79. unicamp-dl/ptt5-base-pt-msmarco-100k-v2 - Hugging Face, acessado em abril 28, 2025, <https://huggingface.co/unicamp-dl/ptt5-base-pt-msmarco-100k-v2>
80. Evaluation of Question Answer Generation for Portuguese: Insights and Datasets - ACL Anthology, acessado em abril 28, 2025, <https://aclanthology.org/2024.findings-emnlp.306.pdf>
81. Question Generation for the Portuguese Language Computer Science and Engineering - Instituto Superior Técnico, acessado em abril 28, 2025, <https://fenix.tecnico.ulisboa.pt/downloadFile/1126295043839871/90769-rafael-galhoz-dissertacao.pdf>
82. Examples of generated questions (from both Portuguese-translated and... - ResearchGate, acessado em abril 28, 2025, <https://www.researchgate.net/figure/Examples-of-generated-questions-from-both-Portuguese-translated-and-English-SQuAD-v11_fig3_363500878>
83. Lite Training Strategies for Portuguese-English and English-Portuguese Translation - ACL Anthology, acessado em abril 28, 2025, <https://aclanthology.org/2020.wmt-1.90.pdf>
84. Luciano/bertimbau-base-lener\_br - Hugging Face, acessado em abril 28, 2025, <https://huggingface.co/Luciano/bertimbau-base-lener_br>
85. giotvr/bertimbau\_large\_assin\_fine\_tuned - Hugging Face, acessado em abril 28, 2025, <https://huggingface.co/giotvr/bertimbau_large_assin_fine_tuned>
86. rufimelo/Legal-BERTimbau-base - Hugging Face, acessado em abril 28, 2025, <https://huggingface.co/rufimelo/Legal-BERTimbau-base>
87. marquesafonso/bertimbau-large-ner-total - Hugging Face, acessado em abril 28, 2025, <https://huggingface.co/marquesafonso/bertimbau-large-ner-total>
88. marquesafonso/bertimbau-large-ner-selective - Hugging Face, acessado em abril 28, 2025, <https://huggingface.co/marquesafonso/bertimbau-large-ner-selective>
89. rufimelo/Legal-BERTimbau-sts-base-ma-v2 - Hugging Face, acessado em abril 28, 2025, <https://huggingface.co/rufimelo/Legal-BERTimbau-sts-base-ma-v2>
90. Context is Everything: Why Maximum Sequence Length Matters - Cerebras, acessado em abril 28, 2025, <https://www.cerebras.ai/blog/context-is-everything-why-maximum-sequence-length-matters>
91. What is a context window? - IBM, acessado em abril 28, 2025, <https://www.ibm.com/think/topics/context-window>
92. What is a Context Window for LLMs? - Hopsworks, acessado em abril 28, 2025, <https://www.hopsworks.ai/dictionary/context-window-for-llms>
93. Understanding Large Language Models Context Windows - Appen, acessado em abril 28, 2025, <https://www.appen.com/blog/understanding-large-language-models-context-windows>
94. Exploring Context Window of Large Language Models via Decomposed Positional Vectors, acessado em abril 28, 2025, <https://arxiv.org/html/2405.18009v1>
95. Understanding RAG Fluency Metrics: From ROUGE to BLEU - Galileo AI, acessado em abril 28, 2025, <https://www.galileo.ai/blog/fluency-metrics-llm-rag>
96. LLM Evaluation Metrics » Dezlearn » Learn IT Easy, acessado em abril 28, 2025, <https://www.dezlearn.com/llm-evaluation-metrics/>
97. RAG evaluation metrics: A journey through metrics - Elasticsearch Labs, acessado em abril 28, 2025, <https://www.elastic.co/search-labs/blog/evaluating-rag-metrics>
98. Large Language Model (LLM) Evaluation Metrics – BLEU and ROUGE - ML EXPLAINED, acessado em abril 28, 2025, <https://mlexplained.blog/2023/07/08/large-language-model-llm-evaluation-metrics-bleu-and-rouge/>
99. sabachau/Bleu-Machine-Translation-Score - GitHub, acessado em abril 28, 2025, <https://github.com/sabachau/Bleu-Machine-Translation-Score>
100. LREC 2012 Conference Abstracts, acessado em abril 28, 2025, <http://www.lrec-conf.org/lrec2012/IMG/pdf/LREC%202012%20Conference%20Abstracts.pdf>
101. Proceedings of the First Workshop on Modelling Translation - Translatology in the Digital Age - ACL Anthology, acessado em abril 28, 2025, <https://aclanthology.org/2021.motra-1.pdf>
102. AI Metrics that Matter: A Guide to Assessing Generative AI Quality - Encord, acessado em abril 28, 2025, <https://encord.com/blog/generative-ai-metrics/>
103. Automated metrics for evaluating the quality of text generation | DigitalOcean, acessado em abril 28, 2025, <https://www.digitalocean.com/community/tutorials/automated-metrics-for-evaluating-generated-text>
104. Understanding Language Model Evaluation Metrics: A Comprehensive Overview - 3AI, acessado em abril 28, 2025, <https://3ai.in/understanding-language-model-evaluation-metrics-a-comprehensive-overview/>
105. LLMs (Text Generative AI) Evaluations and Metrics – Kaizen - Shubham Shinde, acessado em abril 28, 2025, <https://shubham-shinde.github.io/blogs/llms-metrics/>
106. Large Language Model Evaluation: The Complete Guide - Granica AI, acessado em abril 28, 2025, <https://granica.ai/blog/large-language-model-evaluation-grc>
107. A Metrics-First Approach to LLM Evaluation - Galileo AI, acessado em abril 28, 2025, <https://www.galileo.ai/blog/metrics-first-approach-to-llm-evaluation>