Diplomatrix-BR: Um Corpus Paralelo de Redações de Autoria Humana e de LLMs no Concurso de Diplomacia Brasileira

Rodrigo Cavalcanti[1], **Gabriela Casini**[1], **Gabriel Assis**[1], **Livy Real**[2,3], **Daniela Vianna**[2], **Paulo Mann**[4], **Aline Paes**[1]

[1] Universidade Federal Fluminense, Niterói, RJ, Brasil
{rcjoao@id,gabrielacasini@id,assisgabriel@id,alinepaes@ic}.uff.br,
[2]JusBrasil, Salvador, BA, Brasil
livyreal@gmail.com, daniela.vianna@jusbrasil.com.br,
[3]Universidade Federal do Amazonas, Manaus, AM, Brasil
[4]Universidade Federal do Rio de Janeiro, Rio de Janeiro, RJ, Brasil
paulomann@dcc.ufrj.br

Abstract. *Large Language Models (LLMs) have advanced in producing coherent and structured texts, but evaluating their outputs remains a challenge, especially in open-ended and high-level generation tasks. This issue is even more significant for underrepresented languages like Portuguese, where existing benchmarks are often limited in scope and domain. We present Diplomatrix-BR, a new benchmark based on essays from Brazil's diplomatic entrance exam (CACD), including official human-assigned scores and LLM-generated texts within the same themes. We apply a variety of linguistic and automatic metrics to compare human and model-produced content, providing insights into whether current LLMs can generate texts with genuine depth or rely on surface-level fluency. Diplomatrix-BR lays the groundwork for assessing generation in low-resource, high-stakes contexts, while also highlighting the fragility of automatic metrics.*

Resumo. *Modelos de Língua de Larga Escala (LLMs) avançaram significativamente na geração de textos coerentes e bem estruturados, mas a avaliação de suas saídas ainda representa um desafio, especialmente em geração aberta e de alto nível. Esse problema é ainda mais evidente em línguas menos representadas, como o português, em que os benchmarks existentes costumam ser restritos em escopo e domínio. Apresentamos o Diplomatrix-BR, um novo benchmark baseado em redações do exame de admissão à carreira diplomática no Brasil (CACD), acompanhado de suas notas oficiais atribuídas por avaliadores humanos e de textos gerados por LLMs sobre os mesmos temas. Aplicamos uma variedade de métricas linguísticas e automáticas para comparar produções humanas e de modelos, oferecendo indícios sobre se LLMs são capazes de escrever com profundidade real ou se apenas simulam coerência por meio de fluência superficial. O Diplomatrix-BR estabelece as bases para a avaliação da geração em contextos de poucos recursos e de alta complexidade, ao mesmo tempo em que evidencia a fragilidade de métricas automáticas.*

1. Introdução

Modelos de Língua de Larga Escala (LLMs, do inglês *Large Language Models*) têm demonstrado uma capacidade impressionante de geração de textos, que, em muitos

aspectos, se assemelha à escrita humana [Liu et al. 2023a]. Os textos gerados automaticamente têm se destacado particularmente pela sua qualidade sintática, apresentando estruturas gramaticais corretas, coesão entre parágrafos e uso adequado de vocabulário [He et al. 2022, Guo et al. 2024, Martínez et al. 2024], embora ainda existam diferenças sintáticas e semânticas quando comparados a textos produzidos por pessoas [Muñoz-Ortiz et al. 2024]. Diversos estudos já mostraram as habilidades dos LLMs em tarefas como geração de resumos, escrita de artigos jornalísticos e redações acadêmicas de nível básico e intermediário [Zhang et al. 2024, Kaliterna et al. 2024].

Tal evolução despertou interesse crescente sobre seu potencial em tarefas de escrita assistida e apoio à produção textual em contextos educacionais e profissionais [Rodriguez et al. 2019, Tang et al. 2024, Kostic et al. 2024]. Para tanto, trabalhos anteriores apontaram as habilidades de geração automática em contextos de complexidade intermediária, como redações dissertativas-padrão, incluindo exames como o ENEM, o TOEFL ou o GRE [Liu et al. 2023b, Locatelli et al. 2024]. Tais investigações servem não apenas para apontar as habilidades de geração de textos por LLMs, com o fim de melhorá-las ou entendê-las, mas também discutem o potencial de geração de material educativo [Huang et al. 2024], mesmo com saídas de qualidade baixa. Textos gerados por LLMs podem servir não apenas como exemplos para estudo, mas também como material para que estudantes os avaliem e desenvolvam senso crítico [Cufuna et al. 2024].

Por outro lado, ainda é pouco explorada a capacidade dos LLMs de produzirem textos a serem avaliados por critérios significativamente mais exigentes. Exemplos incluem redações de concursos públicos altamente seletivos, exames de ingresso em carreiras diplomáticas ou seleções de pós-graduação de excelência. Esses textos demandam não apenas coesão e coerência, mas também domínio lexical avançado, sofisticação argumentativa, clareza analítica e adequação estilística a contextos formais [Machin et al. 2020, French et al. 2024]. Avaliar o desempenho de LLMs nessas condições representa um desafio expressivo, tanto do ponto de vista da geração quanto da avaliação, e impõe a necessidade de metodologias específicas, conjuntos de dados de referência e métricas sensíveis à qualidade discursiva em várias dimensões [Sudhakaran et al. 2023, Hickman et al. 2024, Gao et al. 2025].

Nesse contexto, este trabalho propõe uma investigação sistemática das capacidades de LLMs na geração de textos que simulam redações exigidas em exames de alta complexidade, com foco na avaliação linguística e automática das respostas geradas. Para tanto, contribuímos com o *Diplomatrix-BR*, um *corpus* paralelo composto por redações de candidatos do Concurso de Admissão à Carreira de Diplomata (CACD)[1] e por redações geradas automaticamente por LLMs a partir dos mesmos temas. A prova do CACD é reconhecida por seu alto nível de complexidade, abrangendo diversas áreas de conhecimento [João Fellet 2014, Rafaela Zem 2024]. Tal exame demanda não apenas correção gramatical e coesão textual, mas também profundidade analítica, erudição, domínio de referências históricas e geopolíticas específicas, capacidade de articulação de argumentos complexos e aplicação precisa de terminologias específicas das relações internacionais, direito internacional e política externa [Ministério das Relações Exteriores 2025].

Entretanto, tarefas como essa, denominadas de geração de texto com fim aberto

[1] Informações sobre o CACD

(*Open-Ended Language Generation* — OENLG) [Erdem et al. 2022], definidas por receberem entradas restritas, mas permitirem muitas respostas distintas plausíveis, ainda são particularmente desafiadoras em termos de avaliação [Zhou et al. 2022]. Assim, a metodologia apresentada no artigo compara as redações dos candidatos e aquelas geradas por 13 LLMs de tamanhos variados utilizando métricas automáticas comumente aplicadas para avaliar a geração de textos, incluindo BLEU [Papineni et al. 2002], ROUGE [Lin 2004], BERTScore [Zhang et al. 2020], CTC [Deng et al. 2021] e o vasto conjunto de indicadores NILC-*Metrix* [Leal et al. 2024]. Ademais, exibimos a correlação entre as métricas automáticas e as notas efetivamente atribuídas às redações de autoria humana por avaliadores especializados. Tal análise visa identificar possíveis discrepâncias entre as métricas automáticas e a nota dada pelos examinadores neste tipo de avaliação.

Em resumo, este artigo apresenta um *corpus* de redações de alta exigência avaliativa e uma análise abrangente de métricas automáticas e linguísticas. Os resultados mostram que, comparados às notas de candidatos aprovados, os LLMs se equiparam a textos de desempenho intermediário, mas não alcançam os de maior nota.

2. Trabalhos Relacionados

A geração de linguagem natural (NLG) — especialmente a geração aberta (OENLG) — avançou rapidamente com a adoção de LLMs, produzindo textos coerentes a partir de entradas mínimas. Estudos anteriores avaliaram modelos em tarefas como produção de resumos, redações do ENEM, TOEFL e GRE, geralmente com métricas de referência automáticas como BLEU [Papineni et al. 2002] e ROUGE [Lin 2004]. Outros estudos avaliam a geração em domínios de complexidade moderada [Locatelli et al. 2024, Liu et al. 2023b]. Diferentemente, nosso trabalho aborda redações de alta complexidade, que carecem de *benchmarks* específicos para profundidade argumentativa e sofisticação linguística, no contexto do CACD.

Em [Locatelli et al. 2024], analisou-se a geração de redações do ENEM por diferentes LLMs, mostrando que humanos fragmentam ideias em múltiplas cláusulas curtas, enquanto modelos geram menos cláusulas, porém mais extensas. Para mais, a investigação do *corpus* ArguGPT [Liu et al. 2023b] (com redações do WECCL, TOEFL e GRE) revela que redações geradas por máquina têm sintaxe mais complexa, mas diversidade lexical reduzida e coesão distinta dos textos humanos. Em muitos casos, os avaliadores — tutores humanos de inglês — reportaram dificuldade em diferenciar os autores das produções.

O trabalho de [Ullah and Yameen 2024] compara textos acadêmicos humanos e de LLMs em inglês, destacando que o primeiro grupo tem maior riqueza lexical, estilo autoral evidente e precisão factual superior. No geral, os textos de LLMs, embora coerentes, são limitados em profundidade analítica [Rodriguez et al. 2019], referências factuais e traços autorais [d'Alte and d'Alte 2023], e apresentam diversidade de vocabulário inferior [Locatelli et al. 2024, Liu et al. 2023b]. Nosso estudo amplia esse escopo de investigação ao analisar textos avançados no contexto do CACD com métricas sensíveis às nuances argumentativas e linguísticas.

[Sardinha 2024] confronta textos humanos e do GPT-3.5 em quatro gêneros — conversação, acadêmico, redação de aprendizes e notícias — usando análise multidimensional [Biber 1988]. Os resultados mostram diferenças significativas nos padrões

linguísticos, especialmente em interação discursiva. Observou-se que textos acadêmicos gerados automaticamente são os mais similares aos humanos, sugerindo um viés para estilos acadêmicos, possivelmente devido aos dados de treinamento. Nosso trabalho, inserido no contexto diplomático de alta exigência em português, pretende oferecer recursos para investigar se essas percepções são compartilhadas nesse cenário.

3. Diplomatrix-BR: um *Corpus* de Redações para Exames do CACD

O Concurso de Admissão à Carreira de Diplomata (CACD) é o processo seletivo realizado pelo Instituto Rio Branco, vinculado ao Ministério das Relações Exteriores do Brasil, para recrutar novos diplomatas. Reconhecido por seu alto grau de exigência, o CACD avalia uma ampla gama de conhecimentos nas áreas de língua portuguesa, história do Brasil e mundial, geografia, política internacional, direito e economia, além de provas específicas de línguas estrangeiras, como inglês, francês e espanhol. O concurso é tradicionalmente voltado à formação de um corpo diplomático com sólida base humanística, domínio técnico e sensibilidade intercultural, sendo considerado um dos concursos públicos mais prestigiados e desafiadores do país[2].

Anualmente, alunos do Instituto Rio Branco — que também exerce a função de academia formadora de diplomatas no Brasil — aprovados no CACD organizam guias voltados à orientação de novos candidatos. Esses guias reúnem respostas discursivas às questões do exame, redações, comentários, recomendações e, de maneira relevante, as respectivas notas atribuídas. Com base nesse material[3], realizamos a extração automática das redações disponíveis e suas avaliações em formato PDF nas coletâneas publicadas entre os anos de 2013 e 2023, totalizando 88 textos.

Ademais, conjecturamos que esse conjunto, por estar associado ao ingresso em uma carreira de elevada complexidade, como a diplomática, apresenta características distintas quando comparado, por exemplo, aos textos produzidos por candidatos do Exame Nacional do Ensino Médio (ENEM), os quais atualmente compõem uma parcela significativa dos recursos disponíveis para a análise de redações em português [Amorim and Veloso 2017, Marinho et al. 2021].

Desse modo, a fim de trazer indícios para essa conjectura, foram calculados os indicadores do conjunto de métricas linguísticas NILC-*Metrix* [Leal et al. 2024] tanto para o conjunto extraído quanto para o *corpus* Essay-BR [Marinho et al. 2021], uma coletânea de textos produzidos por alunos do ensino médio e avaliados seguindo os critérios do ENEM[4]. A Tabela 1 apresenta os resultados de alguns desses indicadores. Observa-se que as redações do CACD apresentam, em média, maior número de palavras e sentenças do que as do Essay-BR. Além disso, os textos do CACD exibem maior proporção de palavras de conteúdo — aquelas que concentram o significado principal da frase — e menor proporção de palavras de função, como artigos e conjunções, sugerindo maior densidade informacional. No que se refere à presença de adjetivos, as redações do CACD também apresentam uma frequência mais elevada, o que pode indicar maior complexidade nesse

[2]Mais informações no *site oficial*.

[3]Os guias podem ser acessados publicamente pelo *site*.

[4]Esta análise serve apenas para ilustrar as diferenças médias de estilo e características entre os conjuntos. Vale ressaltar que as redações do CACD são de candidatos aprovados, com notas acima da média, enquanto o Essay-BR inclui textos com diferentes faixas de desempenho.

Tabela 1. Indicadores (com desvio padrão) do Nilc-*Metrix* por exame.

Métrica	Essay-BR (ENEM)	Diplomatrix-BR (CACD)
Indicadores Gerais		
Número de palavras	288,515 (82,982)	658,747 (65,669)
Número de sentenças	10,592 (4,434)	25,287 (5,252)
Palavras de conteúdo (%)	0,578 (0,026)	0,592 (0,020)
Palavras de função (%)	0,422 (0,026)	0,408 (0,020)
Adjetivos (%)	0,077 (0,022)	0,113 (0,024)
Advérbios (%)	0,047 (0,019)	0,031 (0,012)
Operadores lógicos (%)	0,043 (0,016)	0,033 (0,010)
Índice de Brunet	11,130 (0,482)	12,297 (0,401)
Estatística de Honoré	1061,799 (205,908)	1012,584 (128,021)
Ambiguidade		
Ambiguidade de adjetivos	3,284 (1,303)	2,208 (0,580)
Ambiguidade de advérbios	2,534 (1,028)	2,391 (0,817)
Ambiguidade de substantivos	3,386 (0,503)	3,323 (0,389)
Ambiguidade de verbos	9,035 (1,599)	7,745 (1,182)
Ambiguidade de palavras de conteúdo	4,762 (0,694)	3,842 (0,391)
Diversidade		
Diversidade de adjetivos	0,878 (0,093)	0,713 (0,115)
Diversidade de advérbios	0,758 (0,147)	0,661 (0,121)
Diversidade de palavras de conteúdo	0,858 (0,051)	0,863 (0,042)
Diversidade de palavras de função	0,408 (0,051)	0,347 (0,037)
Diversidade de pronomes indefinidos	0,831 (0,244)	0,819 (0,288)
Diversidade de pontuação	0,106 (0,052)	0,042 (0,015)
Diversidade de verbos	0,845 (0,072)	0,802 (0,075)
Diversidade de preposições	0,396 (0,085)	0,258 (0,030)

aspecto. Em contrapartida, nota-se uma presença ligeiramente menor de advérbios e operadores lógicos nesse conjunto, em comparação com os textos do Essay-BR.

No âmbito de índices de leiturabilidade, observa-se um valor ligeiramente inferior para o Índice de Brunet [Étienne Brunet 1978] e um valor ligeiramente superior para a Estatística de Honoré [Honoré 1979] no conjunto Essay-BR, ambos sugerindo uma leve maior complexidade nesse *corpus*. No entanto, é importante reforçar a diferença substancial no tamanho médio dos textos, cerca de 370 palavras a mais nas redações do CACD, o que pode trazer uma interpretação distinta, uma vez que os textos do CACD mantêm um nível de complexidade comparável, mesmo com maior extensão. Esse equilíbrio entre diferentes dimensões de complexidade manifesta-se em outros indicadores. Por exemplo, os índices de ambiguidade são consistentemente menores no CACD, sugerindo uma preferência por vocabulário mais preciso e menos polissêmico. Em contrapartida, os índices de diversidade lexical são mais elevados no Essay-BR, indicando maior variação no uso de palavras, possivelmente associada à natureza mais aberta e expressiva da prova do ENEM, em contraste com o perfil técnico e rigoroso do CACD.

Assim, o conjunto de redações voltadas ao contexto diplomático oferece uma nova perspectiva entre os recursos disponíveis para a análise de produção escrita em português, incorporando aspectos e dimensões particulares a esse nível e domínio específicos. Especialmente, neste trabalho, buscamos também avaliar o desempenho de LLMs na geração de redações nesse contexto, comparando-as com as produções humanas coletadas.

O conjunto Diplomatrix-BR é composto de duas partes: `Diplomatrix-BR-base`, composto pelas já mencionadas redações de candidatos, e `Diplomatrix-BR-gen`, composto por redações geradas automaticamente por LLMs, com os mesmos enunciados de Diplomatrix-BR-base. Além das redações, disponibilizamos métricas linguísticas para cada texto, para possibilitar o uso de critérios de

seleção de subconjuntos das bases. O Diplomatrix-BR-gen inclui 390 redações geradas por 13 LLMs e 12 métricas. O processo de geração dessas redações será detalhado na seção a seguir[5].

4. LLMs como Geradores de Redações de Diplomatas

O processo de geração adotado compreende três etapas principais: (i) escolha dos modelos, (ii) formulação dos *prompts* e (iii) definição das métricas utilizadas para avaliação.

4.1. Seleção dos Modelos de Língua

Para investigar a capacidade de LLMs em gerar redações compatíveis com o nível de exigência do CACD, foram selecionados modelos pertencentes a oito famílias: **(i)** GPT-4o [OpenAI 2024], **(ii)** Gemma (2-9B [Gemma-Team 2024] e 3-27B [Gemma-Team 2025]), **(iii)** Llama-3 [MetaAI 2024] (8B e 405B), **(iv)** Mistral (7B [Jiang et al. 2023] e Mixtral 8x22B [Jiang et al. 2024]), **(v)** Qwen 2.5 [Alibaba-Cloud 2025] (7B e 72B), **(vi)** Phi (3-7B [Microsoft 2024] e 4-14B [Abdin et al. 2024]) **(vii)** Command-R+ [Cohere 2024] e **(viii)** Sabiá-3 [Abonizio et al. 2025]. A escolha considerou tanto o desempenho dos modelos, especialmente aqueles listados no *leaderboard* público Chatbot Arena [Chiang et al. 2024][6], quanto a viabilidade prática de acesso para fins de inferência.

Embora ausente no Chatbot Arena, o Sabiá-3 foi incluído por seu destaque entre os principais LLMs em português [Garcia 2024], alinhado ao foco deste estudo. No total, foram avaliados 13 modelos, de 7 a mais de 400 bilhões de parâmetros, permitindo analisar o impacto do tamanho na qualidade das redações.

4.2. Construção do *Prompt*

A Figura 1 exibe o *prompt* formulado para que os modelos gerem redações. Além de passar o contexto e as instruções necessárias para o desenvolvimento da redação, o *prompt* inclui os critérios de avaliação da redação, o quais, mesmo que implicitamente, fazem parte do conhecimento prévio dos candidatos do concurso e são relevantes para a escrita. O enunciado varia conforme o tema especificado em cada exame.

Prompt

Contexto
Você é um candidato que deverá elaborar uma redação para concorrer ao cargo de diplomata brasileiro. Sua redação será avaliada de acordo com os seguintes Critérios e Instrução.

Critérios
Apresentação/impressão geral do texto, coerência, legibilidade e estilo;
Capacidade de argumentação (objetividade, sistematização, conteúdo e pertinência das informações);
Capacidade de análise e reflexão;
A redação deve ter entre 65 e 70 linhas;

Instrução
`<Enunciado passado ao candidato>`

Figura 1. *Prompt* de entrada para os LLMs.

[5]Os dados e códigos utilizados neste artigo estão disponíveis em https://github.com/MeLLL-UFF/diplomatrixbr-gen.

[6]https://www.lmarena.ai

4.3. Avaliação da Geração Automática

Para avaliar as redações geradas pelas LLMs, utilizamos um conjunto de métricas amplamente empregadas em PLN. Incluem-se as métricas baseadas em sobreposição de *n*-gramas, como BLEU [Papineni et al. 2002] e ROUGE [Lin 2004][7], além do BERTScore [Zhang et al. 2020], que considera similaridade semântica via *embeddings*. As redações de autoria humana, descritas na Seção 3 e com notas oficiais, serviram como base comparativa, permitindo investigar a correlação entre métricas automáticas e avaliação humana. A disponibilidade de múltiplas redações por tema viabilizou análises com diferentes textos de referência. A Figura 2 resume o processo avaliativo adotado.

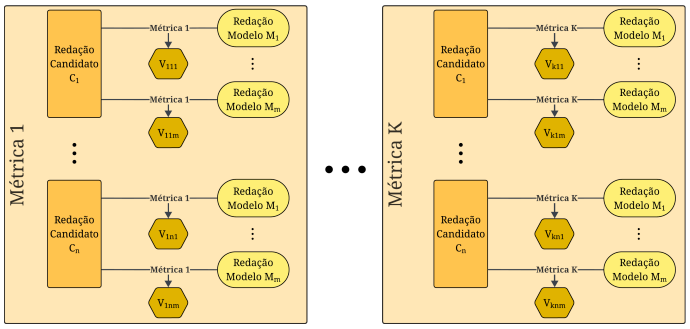

Figura 2. Diagrama do processo de comparação entre as redações dos candidatos e as geradas automaticamente, utilizando diversas métricas.

Também aplicamos a métrica CTC [Deng et al. 2021], que avalia aderência e factualidade com base no alinhamento de *tokens*. Nesse caso, além de comparar com as múltiplas referências, também verificamos a aderência com a instrução da prova.

5. Resultados Experimentais da Geração Automática

Nesta seção, apresentamos os resultados experimentais obtidos a partir das métricas de avaliação automática. Os experimentos foram conduzidos visando responder às questões: **QP1.** *Qual o desempenho de modelos de diferentes tamanhos, com diferentes temperaturas, em métricas léxicas, semânticas e de alinhamento, quando a referência são as redações originais de mesmo tema?*; **QP2.** *Como as redações geradas automaticamente se comparam às redações dos candidatos em termos de métricas linguísticas?*; **QP3.** *Existe correlação entre as notas das redações dos candidatos e os valores das métricas automáticas?*

5.1. Configurações experimentais

Para executar o processo de geração com LLMs, os seguintes hiperparâmetros foram definidos empiricamente. O `top_p` foi fixado em 0,4, restringindo os modelos aos *tokens* mais prováveis (núcleo estreito), visando mais controle e menor risco de saída absurda. Por outro lado, o valor de `temperature` foi variado em {0,3, 0,5, 0,7}, para verificar o impacto das escolhas dentro desse núcleo estreito. Por último, o valor de `max_tokens` foi fixado em 1024, limite compatível com o número máximo de palavras típico das redações do CACD. Os modelos GPT-4o, Command-R+ e Sabiá-3 foram acessados por meio de suas respectivas APIs. Os outros modelos com até 9 bilhões de parâmetros foram

[7]O ROUGE foi aplicado em todas as variantes disponíveis na biblioteca *evaluate*.

executados via *framework* Transformers [Wolf et al. 2020] em uma GPU NVIDIA RTX 4090. Os demais, foram acessados pelo Chatbot Arena [Chiang et al. 2024].

Tabela 2. Valores médios das métricas quantitativas das redações geradas automaticamente, quando comparadas com as redações originais no mesmo tema.

(A) Modelos abertos *menores*

Model	BLEU	ROUGE				BERTScore	CTC			
		R1	R2	RL	RLs	F1	ins_adr.	ref_adr.	ins_fac.	ref_fac.
Gemma-2-9B (temp 0,3)	0,037	0,417	0,108	0,185	0,240	0,726	319,161	329,502	0,695	0,717
Gemma-2-9B (temp 0,5)	0,037	0,417	0,107	0,186	0,239	0,725	320,399	331,444	0,695	0,718
Gemma-2-9B (temp 0,7)	0,035	0,412	0,107	0,182	0,241	0,724	320,720	330,827	0,698	0,719
Llama-3.1-8B (temp 0,3)	0,039	0,432	0,108	0,187	0,240	0,711	328,475	334,499	0,707	0,720
Llama-3.1-8B (temp 0,5)	0,035	0,417	0,104	0,182	0,229	0,704	329,821	332,678	0,707	0,714
Llama-3.1-8B (temp 0,7)	0,035	0,419	0,103	0,184	0,233	0,707	326,232	332,687	0,702	0,717
Mistral-7B (temp 0,3)	0,033	0,367	0,091	0,163	0,211	0,709	310,412	313,771	0,675	0,684
Mistral-7B (temp 0,5)	0,024	0,343	0,085	0,157	0,197	0,707	309,598	313,842	0,672	0,682
Mistral-7B (temp 0,7)	0,028	0,357	0,087	0,159	0,206	0,709	317,960	322,510	0,681	0,692
Phi-3-7B (temp 0,3)	0,027	0,388	0,095	0,175	0,222	0,724	321,239	336,933	0,685	0,718
Phi-3-7B (temp 0,5)	0,029	0,380	0,100	0,172	0,222	0,726	321,793	332,709	0,692	0,716
Phi-3-7B (temp 0,7)	0,027	0,381	0,096	0,166	0,223	0,721	326,117	337,637	0,691	0,715
Qwen-2.5-7B (temp 0,3)	0,021	0,353	0,092	0,165	0,203	0,684	320,769	337,366	0,676	0,711
Qwen-2.5-7B (temp 0,5)	0,023	0,357	0,093	0,166	0,203	0,687	322,607	340,279	0,679	0,716
Qwen-2.5-7B (temp 0,7)	0,026	0,371	0,099	0,167	0,218	0,725	324,260	336,554	0,690	0,716

(B) Modelos abertos *maiores*

Model	BLEU	ROUGE				BERTScore	CTC			
		R1	R2	RL	RLs	F1	ins_adr.	ref_adr.	ins_fac.	ref_fac.
Gemma-3-27B (temp 0,3)	0,031	0,385	0,107	0,175	0,232	0,717	322,622	335,692	0,692	0,720
Gemma-3-27B (temp 0,5)	0,029	0,388	0,107	0,174	0,235	0,717	321,695	334,225	0,690	0,717
Gemma-3-27B (temp 0,7)	0,033	0,394	0,108	0,176	0,237	0,717	321,646	334,254	0,690	0,717
Llama-3-405B (temp 0,3)	0,044	0,419	0,117	0,187	0,238	0,721	324,916	336,656	0,697	0,721
Llama-3-405B (temp 0,5)	0,039	0,409	0,112	0,185	0,235	0,719	328,755	337,529	0,704	0,723
Llama-3-405B (temp 0,7)	0,041	0,409	0,115	0,184	0,233	0,722	328,280	337,209	0,707	0,726
Mixtral-8x22B (temp 0,3)	0,027	0,375	0,107	0,170	0,220	0,728	341,412	342,217	0,727	0,729
Mixtral-8x22B (temp 0,5)	0,029	0,371	0,104	0,168	0,210	0,719	344,777	342,225	0,729	0,725
Mixtral-8x22B (temp 0,7)	0,035	0,390	0,106	0,174	0,224	0,727	336,244	338,524	0,721	0,727
Phi-4-14B (temp 0,3)	0,036	0,417	0,114	0,176	0,236	0,729	325,116	339,751	0,686	0,718
Phi-4-14B (temp 0,5)	0,034	0,411	0,114	0,178	0,229	0,731	323,493	337,476	0,688	0,719
Phi-4-14B (temp 0,7)	0,032	0,410	0,114	0,176	0,231	0,730	327,768	340,053	0,692	0,719
Qwen-2.5-72B (temp 0,3)	0,049	0,441	0,112	0,181	0,251	0,729	329,585	340,244	0,698	0,722
Qwen-2.5-72B (temp 0,5)	0,050	0,436	0,115	0,179	0,246	0,728	327,569	340,071	0,694	0,720
Qwen-2.5-72B (temp 0,7)	0,046	0,436	0,112	0,176	0,241	0,727	329,736	340,192	0,697	0,720

(C) Modelos Comerciais

Model	BLEU	ROUGE				BERTScore	CTC			
		R1	R2	RL	RLs	F1	ins_adr.	ref_adr.	ins_fac.	ref_fac.
GPT-4o (temp 0,3)	0,050	0,460	0,120	0,179	0,251	0,722	327,391	339,652	0,694	0,719
GPT-4o (temp 0,5)	0,048	0,450	0,116	0,176	0,247	0,724	329,680	339,904	0,700	0,721
GPT-4o (temp 0,7)	0,051	0,454	0,119	0,181	0,253	0,723	327,388	340,269	0,694	0,720
Sabiá-3 (temp 0,3)	0,049	0,459	0,122	0,182	0,257	0,724	325,655	337,634	0,693	0,718
Sabiá-3 (temp 0,5)	0,048	0,460	0,124	0,183	0,255	0,721	322,328	334,914	0,694	0,720
Sabiá-3 (temp 0,7)	0,048	0,459	0,126	0,184	0,251	0,724	324,679	337,422	0,695	0,722
Command-R+ (temp 0,3)	0,056	0,466	0,128	0,184	0,267	0,723	335,000	338,428	0,713	0,720
Command-R+ (temp 0,5)	0,062	0,478	0,134	0,189	0,270	0,726	332,396	337,602	0,711	0,722
Command-R+ (temp 0,7)	0,057	0,470	0,126	0,186	0,268	0,732	332,197	338,477	0,710	0,724

5.2. Resultados

Primeiro, as redações geradas pelos LLMs têm, em média, 481 palavras, número inferior às cerca de 658 palavras das referências. A Tabela 2 apoia a análise da articulação dessas palavras e exibe os resultados utilizados para responder à QP1. Observa-se que os valores de BLEU e das variantes do ROUGE (exceto ROUGE-1) são extremamente baixos em todos os casos, sugerindo uma sobreposição lexical limitada a unigramas, mas com ordem divergente em relação às referências, um resultado plausível em tarefas de geração aberta de texto. Em contraste, os escores de F1 do BERTScore são, em geral, elevados, apontando para alguma preservação de conteúdo semântico entre as redações originais e as geradas automaticamente. Tendência semelhante é observada nas diferentes variantes da métrica CTC. De modo geral, não se verificam diferenças significativas entre

os modelos e configurações de temperatura, com exceção das temperaturas 0,3 e 0,5 do Qwen2.5-7B, cujos escores ficaram abaixo de 0,7. Mais do que evidenciar similaridade entre os textos gerados e as referências, esses resultados podem indicar limitações das métricas automáticas na avaliação de tarefas de geração aberta, dada a baixa sensibilidade a variações entre modelos de diferentes portes e configurações — embora modelos maiores tenham registrado ligeira superioridade, exceptuando os modelos Gemma.

Para responder à QP2, a Figura 3 apresenta uma projeção 2D, via t-SNE [van der Maaten and Hinton 2008], dos vetores de medidas do NILC-*Metrix* para as redações originais e as geradas pelos LLMs. As maiores notas aparecem, em geral, isoladas das demais produções de cada ano, distantes das menores notas, mas relativamente próximas das intermediárias. Ressalta-se que mesmo as notas mínimas referem-se a candidatos aprovados, o que indica um padrão de alta qualidade geral.

Em quase todos os anos, as melhores notas estão espacialmente separadas dos *clusters* dos LLMs, sugerindo que os textos de maior desempenho ainda apresentam características linguísticas não plenamente capturadas pelos modelos. Os modelos que mais se aproximam dessas regiões ao longo dos anos incluem Mixtral 8×22B, Llama-3-405B, Gemma-3-27B, Phi-4-14B e Mistral-7B, com uma leve aproximação mais evidente a partir de 2022. O modelo nacional Sabiá-3, embora distante das maiores notas, aparece próximo das intermediárias. Não observamos um padrão claro de influência da temperatura na proximidade entre os textos gerados e os humanos.

De modo geral, os resultados indicam que os LLMs tendem a ocupar com consistência o espaço das notas intermediárias, especialmente os modelos de maior porte e em temperaturas mais altas (0,5 e 0,7), mas permanecem um pouco distantes tanto das notas máximas e mínimas. Tais indícios trazem direções para gerações futuras mais direcionadas a níveis específicos de qualidade.

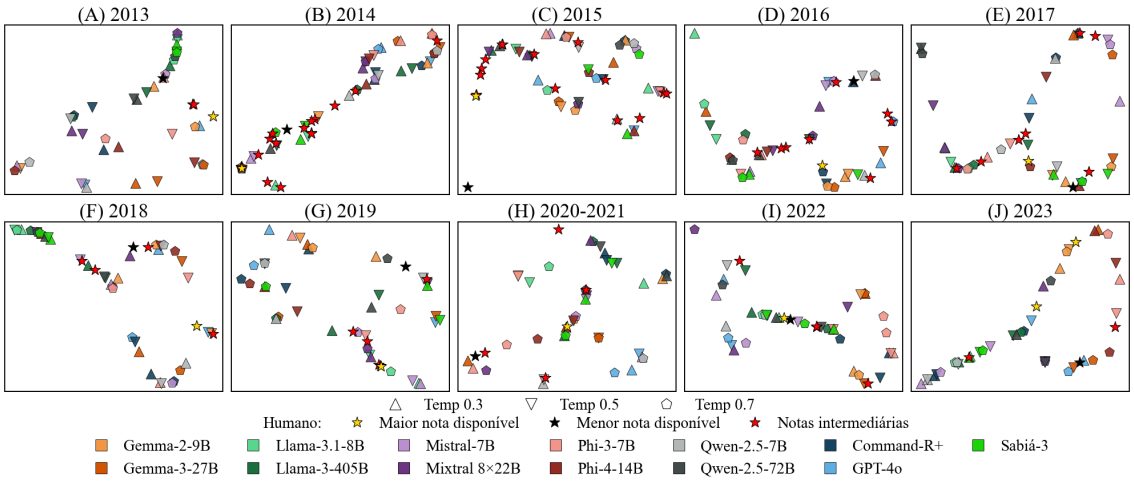

Figura 3. Representação t-SNE das redações originais para cada ano, comparadas com as redações geradas, usando as medidas do NILC-Metrix.

Para responder à QP3, a Tabela 3 apresenta a correlação de Spearman [Spearman 1904] entre as notas dos candidatos e as métricas automáticas. Aqui optamos por ROUGE-2 e *recall* do BERTScore devido à maior flexibilidade em avaliar o quanto da referência foi capturado pelas gerações em contextos OENLG. Em geral, não

há correlações extremas; porém, predominam correlações positivas leves em ROUGE-2 e BERTScore, sugerindo que características léxicas e semânticas das gerações acompanham modestamente a qualidade crescente das redações escritas por humanos. Já o CTC apresenta correlações majoritariamente negativas — e mais significantes —, indicando possível ausência de referências factuais importantes nos textos gerados pelos LLMs. Entre modelos, destaca-se Qwen-2.5-72B, com correlações positivas consistentes, seguido por GPT-4o, Mistral e Sabiá, observando-se novamente poucas variações consideráveis entre temperaturas. Por último, a ausência geral de correlações maiores também pode indicar limitações das métricas automáticas para avaliar a qualidade das saídas em tarefas OENLG.

Tabela 3. Correlação média de Spearman com notas por modelos e temperatura.

Modelo	Rouge-2			BERTScore (recall)			CTC (ref_adr)			CTC (ref_fac)		
	0,3	0,5	0,7	0,3	0,5	0,7	0,3	0,5	0,7	0,3	0,5	0,7
Gemma-2-9B	0,100	0,097	0,012	-0,161	-0,117	-0,215	-0,373	-0,364	-0,395	-0,454	-0,504	-0,444
Gemma-3-27B	0,005	0,070	0,051	-0,053	-0,037	-0,016	-0,158	-0,048	-0,054	-0,386	-0,315	-0,340
Llama-3.1-8B	-0,078	0,099	0,138	-0,147	0,165	-0,071	-0,306	-0,212	-0,376	-0,254	0,017	-0,191
Llama-3-405B	-0,002	0,092	0,107	0,041	0,065	0,060	-0,209	-0,085	-0,107	-0,324	-0,228	-0,212
Mistral-7B	0,257	0,209	0,189	0,100	0,317	0,147	-0,115	-0,198	-0,202	0,062	0,008	-0,043
Mixtral-8x22B	0,252	0,329	0,116	-0,082	0,176	-0,045	-0,161	-0,122	-0,364	-0,198	-0,118	-0,348
Phi-3-7B	-0,137	-0,066	0,091	-0,131	-0,185	-0,106	-0,560	-0,460	-0,453	-0,430	-0,531	-0,412
Phi-4-14B	0,098	0,072	0,125	-0,071	-0,087	-0,099	0,043	-0,207	-0,010	-0,304	-0,282	-0,174
Qwen-2.5-7B	-0,213	-0,173	-0,226	-0,496	-0,434	-0,376	-0,140	-0,140	-0,039	-0,366	-0,362	-0,411
Qwen-2.5-72B	0,127	0,184	0,247	0,120	0,137	0,158	0,067	0,029	0,104	0,063	0,148	0,020
GPT-4o	0,184	0,381	0,126	0,088	0,080	-0,065	0,188	0,222	0,134	-0,155	-0,037	0,208
Sabiá-3	0,052	0,248	0,078	-0,005	0,046	-0,083	0,200	0,092	0,125	-0,182	-0,224	-0,360
Command-R+	0,012	-0,002	-0,022	-0,106	0,027	-0,020	-0,211	-0,192	0,133	-0,383	-0,421	-0,134

6. Conclusões

Este trabalho introduziu o Diplomatrix-BR, um *corpus* composto por 88 redações escritas por humanos, com suas notas oficiais, e 390 redações geradas por 13 LLMs sobre os mesmos temas. Além dos textos, disponibilizam-se indicadores linguísticos e valores de métricas automáticas, contribuindo especialmente para estudos de OENLG em português, sobretudo em contextos mais complexos como o diplomático. Nossas análises mostram que, embora os LLMs ainda não alcancem plenamente o nível das redações mais bem avaliadas, aproximam-se em vários aspectos. Desse modo, o recurso construído apresenta potencial significativo para fins educacionais e de análise da produção textual, em geral.

Trabalhos em andamento envolvem a aplicação de métodos automáticos para a atribuição de notas e pareceres às redações do Diplomatrix-BR, com base na estratégia de *LLM-as-judges* [Gu et al. 2024]. Paralelamente, conduziremos uma avaliação humana criteriosa de uma amostra do conjunto Diplomatrix-BR-gen, bem como do próprio processo de atribuição de notas, visando garantir rigor e confiabilidade na análise. Pretendemos, ainda, expandir a análise para os demais tipos de questões presentes no CACD. Tarefas de OENLG requerem avaliação humana rigorosa, dada a fragilidade de métricas, e a ausência desta é uma limitação do presente artigo.

Agradecimentos

Os autores agradecem o apoio financeiro do CNPq, do INCT IAIA e da FAPERJ (processos SEI-260003/002930/2024 e SEI-260003/000614/2023). Agradecem também pelos créditos fornecidos pela Maritaca AI e *Cohere Labs Research Grant*.

Referências

Abdin, M., Aneja, J., Behl, H., Bubeck, S., Eldan, R., Gunasekar, S., Harrison, M., Hewett, R. J., Javaheripi, M., Kauffmann, P., Lee, J. R., Lee, Y. T., Li, Y., Liu, W., Mendes, C. C. T., Nguyen, A., Price, E., de Rosa, G., Saarikivi, O., Salim, A., Shah, S., Wang, X., Ward, R., Wu, Y., Yu, D., Zhang, C., and Zhang, Y. (2024). Phi-4 Technical Report.

Abonizio, H., Almeida, T. S., Laitz, T., Junior, R. M., Bonás, G. K., Nogueira, R., and Pires, R. (2025). Sabiá-3 Technical Report.

Alibaba-Cloud (2025). Qwen2.5 Technical Report.

Amorim, E. and Veloso, A. (2017). A Multi-aspect Analysis of Automatic Essay Scoring for Brazilian Portuguese. In Kunneman, F., Iñurrieta, U., Camilleri, J. J., and Ardanuy, M. C., editors, *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 94–102, Valencia, Spain. Association for Computational Linguistics.

Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge University Press, Cambridge.

Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhu, B., Zhang, H., Jordan, M. I., Gonzalez, J. E., and Stoica, I. (2024). Chatbot Arena: an open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Cohere (2024). Command R+. Acesso em: jun. 2025.

Cufuna, D. S. A., Duart, J. M., and Rangel-de Lazaro, G. (2024). Augmented reality in higher education: Interactions in llm-based teaching and learning. In *The Learning Ideas Conference*, pages 105–114. Springer.

Deng, M., Tan, B., Liu, Z., Xing, E., and Hu, Z. (2021). Compression, Transduction, and Creation: A Unified Framework for Evaluating Natural Language Generation. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7580–7605, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

d'Alte, P. and d'Alte, L. (2023). Para uma avaliação do chatgpt como ferramenta auxiliar de escrita de textos acadêmicos. *Revista Bibliomar, São Luís*, 22(1):122–138.

Erdem, E., Kuyu, M., Yagcioglu, S., Frank, A., Parcalabescu, L., Plank, B., Babii, A., Turuta, O., Erdem, A., Calixto, I., Lloret, E., Apostol, E.-S., Truică, C.-O., Šandrih, B., Martinčić-Ipšić, S., Berend, G., Gatt, A., and Korvel, G. (2022). Neural Natural Language Generation: A Survey on Multilinguality, Multimodality, Controllability and Learning. *Journal of Artificial Intelligence Research (JAIR)*, 73.

French, S., Dickerson, A., and Mulder, R. A. (2024). A review of the benefits and drawbacks of high-stakes final examinations in higher education. *Higher Education*, 88(3):893–918.

Gao, M., Hu, X., Yin, X., Ruan, J., Pu, X., and Wan, X. (2025). Llm-based nlg evaluation: Current status and challenges. *Computational Linguistics*, pages 1–28.

Garcia, E. A. S. (2024). Open Portuguese LLM Leaderboard. https://huggingface.co/spaces/eduagarcia/open_pt_llm_leaderboard.

Gemma-Team (2024). Gemma 2: Improving Open Language Models at a Practical Size.

Gemma-Team (2025). Gemma 3 technical report.

Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., et al. (2024). A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Guo, Y., Shang, G., and Clavel, C. (2024). Benchmarking linguistic diversity of large language models. *arXiv preprint arXiv:2412.10271*.

He, J., Long, W., and Xiong, D. (2022). Evaluating discourse cohesion in pre-trained language models. In *Proceedings of 3rd Workshop on Computational Approaches to Discourse (CODI 2022)*, page 28.

Hickman, L., Dunlop, P. D., and Wolf, J. L. (2024). The performance of large language models on quantitative and verbal ability tests: Initial evidence and implications for unproctored high-stakes testing. *International Journal of Selection and Assessment*, 32(4):499–511.

Honoré, A. (1979). Some simple measures of richness of vocabulary. *Association for Literary and Linguistic Computing Bulletin*, 7:172–177.

Huang, C.-Y., Wei, J., and Huang, T.-H. K. (2024). Generating educational materials with different levels of readability using llms. In *Proceedings of the Third Workshop on Intelligent and Interactive Writing Assistants*, pages 16–22.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. (2023). Mistral 7b.

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., de las Casas, D., Hanna, E. B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L. R., Saulnier, L., Lachaux, M.-A., Stock, P., Subramanian, S., Yang, S., Antoniak, S., Scao, T. L., Gervet, T., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. (2024). Mixtral of Experts.

João Fellet (2014). Itamaraty e seleção de diplomatas: pai e filho contam trajetória. https://www.bbc.com/portuguese/noticias/2014/05/140505_itamaraty_selecao_diplomatas_pai_jf. Acesso em 09 jun. 2025.

Kaliterna, M., Žuljević, M. F., Ursić, L., Krka, J., and Duplančić, D. (2024). Testing the capacity of Bard and ChatGPT for writing essays on ethical dilemmas: A cross-sectional study. *Scientific Reports*, 14(1):26046.

Kostic, M., Witschel, H. F., Hinkelmann, K., and Spahic-Bogdanovic, M. (2024). Llms in automated essay evaluation: A case study. In *Proceedings of the AAAI Symposium Series*, volume 3, pages 143–147.

Leal, S. E., Duran, M. S., Scarton, C. E., Hartmann, N. S., and Aluísio, S. M. (2024). NILC-Metrix: assessing the complexity of written and spoken language in Brazilian Portuguese. *Language Resources and Evaluation*, 58(1):73–110.

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.

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., and Zhu, C. (2023a). G-eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Liu, Y., Zhang, Z., Zhang, W., Yue, S., Zhao, X., Cheng, X., Zhang, Y., and Hu, H. (2023b). Argugpt: evaluating, understanding and identifying argumentative essays generated by gpt models. *arXiv preprint arXiv:2304.07666*.

Locatelli, M. S., Miranda, M. P., da Silva Costa, I. J., Prates, M. T., Thomé, V., Monteiro, M. Z., Lacerda, T., Pagano, A., Neto, E. R., Meira Jr, W., et al. (2024). Examining the behavior of llm architectures within the framework of standardized national exams in brazil. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 879–890.

Machin, S., McNally, S., and Ruiz-Valenzuela, J. (2020). Entry through the narrow door: The costs of just failing high stakes exams. *Journal of Public Economics*, 190:104224.

Marinho, J., Anchiêta, R., and Moura, R. (2021). Essay-BR: a Brazilian Corpus of Essays. In *Anais do III Dataset Showcase Workshop*, pages 53–64, Online. Sociedade Brasileira de Computação.

Martínez, G., Hernández, J. A., Conde, J., Reviriego, P., and Merino-Gómez, E. (2024). Beware of words: Evaluating the lexical diversity of conversational LLMs using ChatGPT as case study. *ACM Trans. Intell. Syst. Technol.* Just Accepted.

MetaAI (2024). The Llama 3 Herd of Models.

Microsoft (2024). Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone.

Ministério das Relações Exteriores (2025). Editais e guias de estudo — carreira diplomática. https://www.gov.br/mre/pt-br/instituto-rio-branco/carreira-diplomatica/editais-e-guias-de-estudo. Acesso em 09 jun. 2025.

Muñoz-Ortiz, A., Gómez-Rodríguez, C., and Vilares, D. (2024). Contrasting linguistic patterns in human and llm-generated news text. *Artificial Intelligence Review*, 57(10):265.

OpenAI (2024). Gpt-4o System Card.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a Method for Automatic Evaluation of Machine Translation. In Isabelle, P., Charniak, E., and Lin, D., editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Rafaela Zem (2024). O que faz um diplomata? entenda cargo que tem salário inicial de R$ 20,9 mil e aceita qualquer graduação. https://g1.globo.com/trabalho-e-carreira/noticia/2024/06/20/o-que-faz-um-diplomata-entenda-cargo-que-tem-salario-

inicial-de-r-209-mil-e-aceita-qualquer-graduacao-e-um-concurso-muito-dificil.ghtml. Acesso em 09 jun. 2025.

Rodriguez, P. U., Jafari, A., and Ormerod, C. M. (2019). Language models and automated essay scoring. *arXiv preprint arXiv:1909.09482*.

Sardinha, T. B. (2024). Ai-generated vs human-authored texts: A multidimensional comparison. *Applied Corpus Linguistics*, 4:100083.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101.

Sudhakaran, S., González-Duque, M., Freiberger, M., Glanois, C., Najarro, E., and Risi, S. (2023). Mariogpt: Open-ended text2level generation through large language models. *Advances in Neural Information Processing Systems*, 36:54213–54227.

Tang, X., Chen, H., Lin, D., and Li, K. (2024). Harnessing llms for multi-dimensional writing assessment: Reliability and alignment with human judgments. *Heliyon*, 10(14).

Ullah, A. and Yameen, A. (2024). Comparative analysis of linguistic features of academic text and ai-generated text. In *Heritage International Journal of Linguistics & Literature*. Global Heritage Research Center for Languages and Literature.

van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-Art Natural Language Processing. In Liu, Q. and Schlangen, D., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2020). BERTScore: Evaluating Text Generation with Bert. In *International Conference on Learning Representations*.

Zhang, T., Ladhak, F., Durmus, E., Liang, P., McKeown, K., and Hashimoto, T. B. (2024). Benchmarking Large Language Models for News Summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Zhou, K., Blodgett, S. L., Trischler, A., Daumé III, H., Suleman, K., and Olteanu, A. (2022). Deconstructing NLG Evaluation: Evaluation Practices, Assumptions, and Their Implications. In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V., editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 314–324, Seattle, United States. Association for Computational Linguistics.

Étienne Brunet (1978). *Le vocabulaire de Jean Giraudoux : structure et évolution : statistique et informatique appliquées à l'étude des textes à partir des données du Trésor de la langue française*. Le vocabulaire des grands écrivains français. Slatkine, Genève. ASIN: B0000E99PZ.