Você se Torna Responsável por Aquilo que Gera: Análise Semântica e Temporal de Modelos de Linguagem

Andréia dos Santos Sachete, Instituto Federal de Educação, Ciência e Tecnologia Farroupilha, andreia.sachete@iffarroupilha.edu.br, https://orcid.org/0000-0003-2226-3322
Alba Valéria de Sant'Anna de Freitas Loiola, Universidade Federal do Rio Grande do Sul, alba.portugues@gmail.com, https://orcid.org/0000-0003-2418-3393
Fábio Diniz Rossi, Instituto Federal de Educação, Ciência e Tecnologia Farroupilha, fabio.rossi@iffarroupilha.edu.br, https://orcid.org/0000-0002-2450-1024
Raquel Salcedo Gomes, Universidade Federal do Rio Grande do Sul, raquel.salcedo@ufrgs.br, https://orcid.org/0000-0001-9497-513X

Resumo: Este artigo apresenta um estudo experimental que investiga o impacto da temperatura na qualidade e no desempenho de Grandes Modelos de Linguagem (LLMs). Para isso, foram propostas duas questões inspiradas em O Pequeno Príncipe — uma de natureza interpretativa e outra objetiva — a fim de avaliar o comportamento dos modelos em tarefas abertas e fechadas. Quinze LLMs foram analisados sob cinco configurações distintas de temperatura, com base em métricas de similaridade de cosseno e tempo de inferência. Os resultados indicam que a temperatura influencia não apenas a variabilidade lexical, mas também a profundidade semântica das respostas. Modelos mais robustos demonstraram maior estabilidade em temperaturas elevadas, enquanto modelos menores apresentaram melhor desempenho em valores mais baixos. A faixa em torno de 0,5 revelou-se a mais equilibrada, proporcionando respostas coesas com boa eficiência, o que oferece subsídios para a escolha parametrizada de acordo com o tipo de tarefa textual.

Palavras-chave: Avaliação de desempenho, LLM, similaridade de cosseno.

You Become Responsible for What You Generate: A Semantic and Temporal Analysis of Language Models

Abstract: This article presents an experimental study that investigates the impact of temperature on the quality and performance of Large Language Models (LLMs). To this end, two questions inspired by The Little Prince were proposed — one interpretive and the other objective — to evaluate the behavior of the models in open and closed tasks. Fifteen LLMs were analyzed under five different temperature settings based on cosine similarity and inference time metrics. The results indicate that temperature influences not only lexical variability but also the semantic depth of the responses. More robust models demonstrated greater stability at high temperatures, while lower models performed better at lower values. The range around 0.5 proved to be the most balanced, providing cohesive responses with good efficiency, which supports the parameterized choice according to the type of textual task.

Keywords: Performance evaluation, LLM, Cosine similarity.

1. Introdução

O uso de grandes modelos de linguagem tem se expandido de forma acelerada (*Large Language Models* - LLM) (Örpek; Tural e Destan, 2024), viabilizando práticas inovadoras na educação, como tutores inteligentes, corretores automáticos de textos, geradores de feedback formativo e sistemas adaptativos de aprendizagem (Sachete; Loiola e Gomes, 2024). Esses modelos se destacam por sua capacidade de compreender comandos em linguagem natural e gerar respostas ajustadas ao contexto, oferecendo

suporte personalizado aos estudantes de acordo com seu nível de proficiência, objetivos pedagógicos e necessidades específicas (Sachete; Loiola e Gomes, 2024).

Além disso, os LLMs são capazes de simular interações educativas em diferentes registros e gêneros discursivos, adaptando o conteúdo a distintas faixas etárias e perfis sociolinguísticos(Sachete *et al.*, 2025). Essa versatilidade os torna ferramentas promissoras para fomentar o desenvolvimento de habilidades como leitura, escrita e pensamento crítico, ao mesmo tempo em que ampliam as possibilidades de mediação e avaliação por parte dos professores. No entanto, para que sua aplicação em contextos educacionais seja efetiva, é necessário compreender os parâmetros que regulam seu funcionamento, especialmente aqueles que influenciam a qualidade e a adequação das respostas. Um desses parâmetros é a temperatura, que controla o grau de aleatoriedade nas escolhas lexicais e estruturais, afetando diretamente o equilíbrio entre criatividade e precisão na geração textual (Loiola *et al.*, 2024).

Apesar de sua relevância, a configuração da temperatura ainda é, em grande parte, realizada de forma empírica, sem diretrizes sistematizadas que orientem seu uso em ambientes pedagógicos. Há uma lacuna quanto à compreensão dos impactos desse parâmetro em tarefas interpretativas e objetivas, bem como na relação entre desempenho semântico e eficiência computacional (Alnaasan *et al.*, 2024).

Diante disso, este artigo propõe uma análise experimental dos efeitos da temperatura na geração textual por LLMs, com foco em sua aplicação educacional. Foram avaliados quinze modelos em tarefas com diferentes níveis de complexidade, utilizando-se variações na temperatura para observar padrões de coerência, diversidade lexical e desempenho computacional. A análise considerou métricas como similaridade de cosseno e tempo de resposta, com o objetivo de identificar configurações que conciliem precisão, fluidez textual e eficiência.

Para melhor leitura, a estrutura do artigo está organizada da seguinte forma: A Seção 2 apresenta a fundamentação teórica sobre inteligência artificial, processamento de linguagem natural e LLMs. Na Seção 3 descreve-se a metodologia adotada, enquanto a Seção 4 discute os experimentos e os principais resultados. A Seção 5 sintetiza as lições aprendidas e a Seção 6 traz as conclusões e perspectivas para trabalhos futuros.

2. Fundamentação Teórica

A Inteligência Artificial (IA) é uma área multidisciplinar dedicada ao desenvolvimento de sistemas capazes de simular comportamentos tipicamente humanos, como raciocínio, tomada de decisão, aprendizado e compreensão de linguagem natural (Hao; Wang e Li, 2023). Dentre suas subáreas, o Processamento de Linguagem Natural (PLN) (Khurana *et al.*, 2023) tem se destacado nas últimas décadas, especialmente com a popularização dos Modelos de Linguagem baseados em redes neurais profundas. Os Grandes Modelos de Linguagem (*Large Language Models* - LLM) representam um avanço significativo nessa área (Minaee *et al.*, 2024), pois são treinados com bilhões de parâmetros sobre extensos corpora textuais, possibilitando a geração de respostas coerentes, contextualmente relevantes, e, frequentemente, indistinguíveis da produção humana. Arquiteturas como Transformer (Huang *et al.*, 2023) ampliam substancialmente as capacidades desses modelos no domínio textual.

Compreender o funcionamento dos LLMs e a influência de seus hiperparâmetros potencializa seu desempenho e adequa suas aplicações a diferentes contextos. Este estudo se insere nesse escopo ao investigar experimentalmente o impacto da temperatura na produção textual de diferentes modelos, analisando seus efeitos sobre o tempo de inferência e a qualidade semântica das respostas (Liu *et al.*, 2023). Ao contrário de

sistemas baseados em regras ou algoritmos determinísticos, os LLMs são capazes de inferir, sintetizar e adaptar conteúdo com base nos padrões linguísticos aprendidos durante o treinamento, o que os torna especialmente úteis em tarefas que exigem flexibilidade discursiva, como escrita criativa, geração de código, tutoria educacional e apoio à tomada de decisões.

Parâmetros de geração, como a temperatura — ilustrada na Figura 1 —, influenciam diretamente o estilo, a coerência e a diversidade das saídas. Esse hiperparâmetro regula o grau de aleatoriedade na escolha de tokens durante a decodificação: valores baixos tendem a gerar saídas mais determinísticas e conservadoras; valores altos, por sua vez, aumentam a criatividade, mas podem comprometer a coerência semântica (Holtzman *et al.*, 2020).

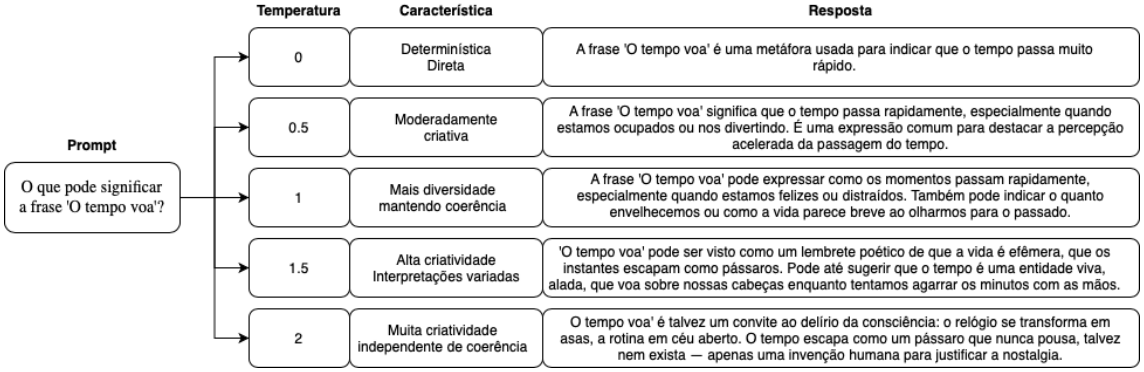


Figura 1. Diferentes respostas ao mesmo prompt, com diferentes temperaturas.

Além do impacto técnico, a compreensão dos efeitos da temperatura tem implicações diretas para o uso pedagógico dos LLMs. Em ambientes educacionais mediados por tecnologia, a geração textual automatizada pode auxiliar na elaboração de explicações personalizadas, na correção de textos dissertativos e na simulação de diálogos interativos. Nesses contextos, a configuração adequada da temperatura torna-se essencial para equilibrar criatividade e clareza, evitando respostas genéricas, ou dispersas que prejudiquem a aprendizagem significativa (Wang *et al.*, 2024).

Para avaliar os resultados gerados, torna-se necessário empregar métricas que considerem não apenas a correção linguística, mas também a relevância semântica e a adequação discursiva das respostas. Indicadores como semantic similarity, coherence score e tempo de inferência funcionam como parâmetros objetivos para a comparação entre modelos. A análise integrada dessas métricas permite identificar padrões no comportamento dos LLMs e estabelecer diretrizes mais precisas para o ajuste de seus hiperparâmetros (Castilho e Gurevych, 2011).

Apesar dos avanços recentes, os LLMs ainda enfrentam limitações significativas, como a tendência a produzir informações imprecisas, redundantes ou contraditórias. Tais falhas são potencializadas por escolhas inadequadas de temperatura, o que reforça a importância de investigações empíricas que ajudem a mapear o comportamento dos modelos sob diferentes condições. Compreender essas fragilidades é fundamental para garantir uma aplicação crítica e responsável da IA em domínios sensíveis, como a educação, a saúde e o direito.

Dessa forma, a realização de experimentos controlados — com variações sistemáticas de temperatura em tarefas de diferentes naturezas — permite não apenas quantificar os efeitos desse hiperparâmetro, mas também produzir evidências empíricas sobre sua influência em múltiplas dimensões textuais. A partir dessas evidências,

é possível formular recomendações práticas para pesquisadores, desenvolvedores e educadores, promovendo um uso mais consciente, transparente e alinhado aos objetivos de cada domínio de aplicação.

3. Metodologia

Para avaliar de forma experimental e sistemática o desempenho de diferentes Grandes Modelos de Linguagem (LLMs) sob variações do parâmetro de temperatura, foi conduzido um experimento controlado com 15 modelos de código aberto amplamente utilizados na comunidade científica. O objetivo principal foi mensurar, simultaneamente, a eficiência temporal das respostas geradas (tempo de inferência) e sua qualidade semântica em relação a uma resposta de referência. Os testes foram organizados em dois conjuntos distintos, ambos baseados na obra *O Pequeno Príncipe*, de Antoine de Saint-Exupéry. O acesso aos modelos foi viabilizado por meio de uma interface de programação de aplicações (*Application Programming Interface - API*) da plataforma groq,[*] à qual foi integrado um código que alimenta os modelos com o conteúdo do livro.

No primeiro conjunto, os modelos foram desafiados com uma pergunta de natureza interpretativa, cuja resposta não era diretamente extraída do texto, exigindo inferência e compreensão contextual: "*O que a raposa ensina ao Pequeno Príncipe?*". A resposta de referência utilizada nesse caso foi: *A raposa ensina ao Pequeno Príncipe que o essencial é invisível aos olhos e que criar laços torna algo verdadeiramente especial. Ela explica que 'cativar' significa estabelecer um vínculo, tornando duas pessoas únicas uma para a outra. A raposa também diz que somos responsáveis por aqueles que cativamos, destacando a importância do afeto e do compromisso nas relações humanas.*

No segundo conjunto, a pergunta possuía uma resposta objetiva e literal: "*Qual a frase que a raposa diz ao Pequeno Príncipe?*", sendo adotada como referiencia a citação direta: *Tu te tornas eternamente responsável por aquilo que cativas.* Essa distinção metodológica possibilitou avaliar o desempenho dos modelos em tarefas com diferentes níveis de complexidade semântica.

Cada modelo foi submetido às mesmas perguntas, com cinco valores distintos de temperatura entre 0 (mínimo) e 2 (máximo) em intervalos de 0.5 ponto. Para cada combinação de modelo e temperatura, foram realizadas 10 execuções, cujas médias compuseram os resultados finais. As condições de hardware e rede foram mantidas constantes, e os baixos desvios padrão observados — inferiores a 0,1 segundo no tempo de inferência e 1,5 pontos na similaridade de cosseno — indicam robustez estatística e estabilidade dos modelos frente a variações operacionais.

O tempo de inferência foi medido desde o início da geração da resposta até sua finalização. A qualidade das respostas foi avaliada por meio da similaridade de cosseno (Reimers e Gurevych, 2019) entre os vetores de *embeddings* das respostas geradas pelos modelos e a resposta de referência. Para essa tarefa, utilizou-se uma técnica de vetorização semântica baseada em modelos pré-treinados de *sentence embeddings*, assegurando coerência na representação semântica. A similaridade de cosseno, amplamente adotada na literatura para avaliação textual, é definida como

$$\text{Sim}(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} \tag{1}$$

em que $\mathbf{A}$ e $\mathbf{B}$ são vetores que representam, respectivamente, a resposta de referência e a resposta gerada pelo modelo.

[*]https://groq.com

4. Avaliação e Discussão

A análise gráfica dos modelos avaliados revela padrões distintos de desempenho entre os dois tipos de pergunta propostos: uma questão de natureza interpretativa e uma de natureza literal. Essa distinção se mostrou importante para compreender não apenas a sensibilidade dos modelos ao parâmetro de temperatura, mas também a forma como diferentes arquiteturas respondem a diferentes níveis de complexidade semântica.

Na Figura 1(a), observa-se que, sob temperatura 0, a geração de respostas interpretativas torna-se rígida e pouco diversificada. Como esperado, essa configuração favorece o determinismo, mas limita a capacidade de adaptação semântica dos modelos. Embora alguns modelos, como *qwen-2.5-coder-32b*, apresentem boa similaridade mesmo sob baixa aleatoriedade, a maioria não consegue capturar nuances da mensagem implícita. O tempo de inferência, por sua vez, é homogêneo e baixo, o que revela que, apesar da limitação interpretativa, o custo computacional é favorecido.

À medida que a temperatura aumenta para 0.5 (Figura 1(b)), observa-se um ganho expressivo na similaridade semântica, indicando que níveis moderados de aleatoriedade promovem maior riqueza lexical sem comprometer a coerência. A arquitetura *llama3-8b-8192* por exemplo, destaca-se por manter bom equilíbrio entre tempo e qualidade textual. Esse comportamento reforça a hipótese de que há uma "zona de temperatura ótima" para perguntas abertas.

Na temperatura 1 (Figura 1(c)), esse efeito é acentuado. Modelos mais robustos, como *llama3-70b-8192* e *llama-3.2-90b-vision-preview*, demonstram resiliência ao aumento de entropia, mantendo boa similaridade com tempos de inferência ainda controlados. Por outro lado, modelos com menor capacidade semântica começam a apresentar respostas inconsistentes.

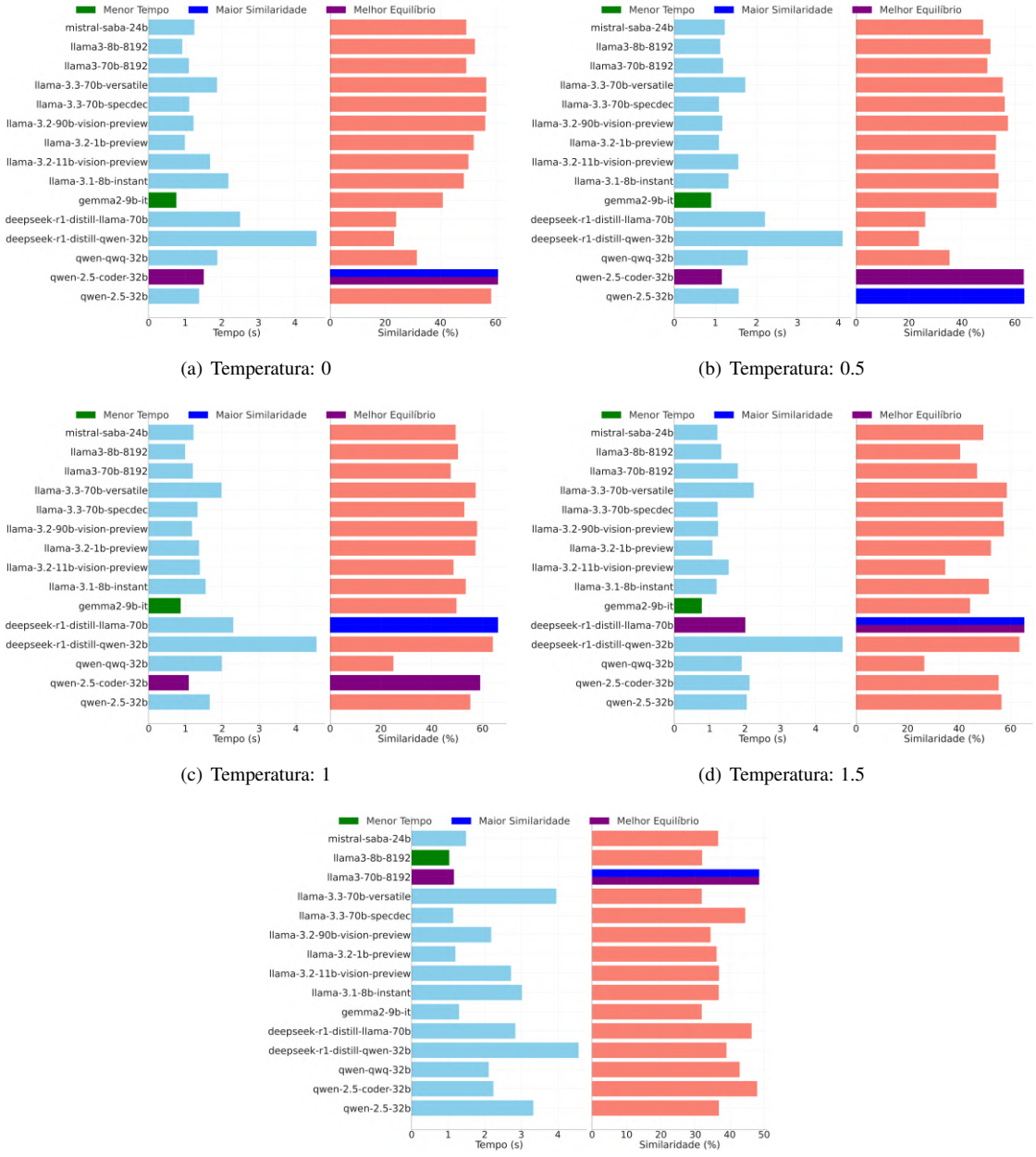
A partir de 1.5 (Figura 1(d)), a dispersão entre os resultados aumenta. Enquanto alguns modelos mantêm a qualidade textual, outros oscilam significativamente. A elevação da temperatura parece induzir interpretações mais criativas, mas nem sempre alinhadas ao conteúdo semântico esperado. O modelo *gemma2-9b-it*, por exemplo, mantém baixo tempo de inferência, mas sofre perdas severas de similaridade, sugerindo que sua estratégia de decodificação não se adapta adequadamente à alta aleatoriedade.

Com temperatura 2 (Figura 1(e)), a inconsistência torna-se dominante. A maioria dos modelos apresenta degradação semântica, indicando que valores extremos de temperatura comprometem a capacidade de manter vínculos conceituais com a resposta de referência. Nesse cenário, observa-se que mesmo arquiteturas de grande porte começam a apresentar lacunas em suas respostas, apontando para um limite prático de entropia tolerável para tarefas interpretativas.

Nas figuras referentes à pergunta direta, observa-se um padrão distinto. Na Figura 2(a), a baixa temperatura favorece diretamente a recuperação da citação-alvo, com alta similaridade e baixa variabilidade. Modelos como *llama-3.3-70b-specdec* e *qwen-2.5-coder-32b* conseguem recuperar a frase com precisão, o que é coerente com a natureza determinística da tarefa.

Com o aumento para temperatura 0.5 (Figura 2(b)), ainda se observa alto desempenho, embora algumas arquiteturas com vocabulário mais diversificado introduzam pequenas variações lexicais. Essas variações nem sempre prejudicam a similaridade, uma vez que a métrica vetorial utilizada tolera sinônimos e paráfrases.

No entanto, a partir da temperatura 1 (Figura 2(c)), torna-se evidente que a entropia começa a interferir na recuperação literal da resposta. Embora alguns modelos mantenham alta similaridade, outros começam a reinterpretar a frase, ou introduzir variações que comprometem a fidelidade esperada.



(a) Temperatura: 0

(b) Temperatura: 0.5

(c) Temperatura: 1

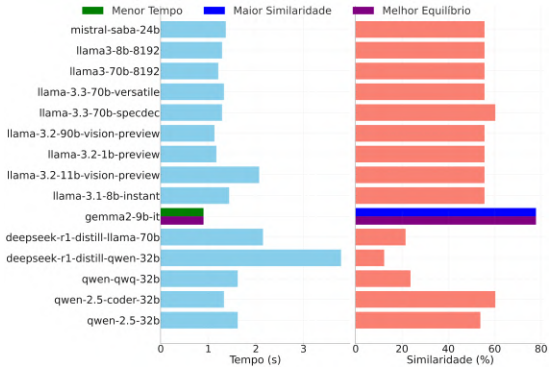
(d) Temperatura: 1.5

(e) Temperatura: 2

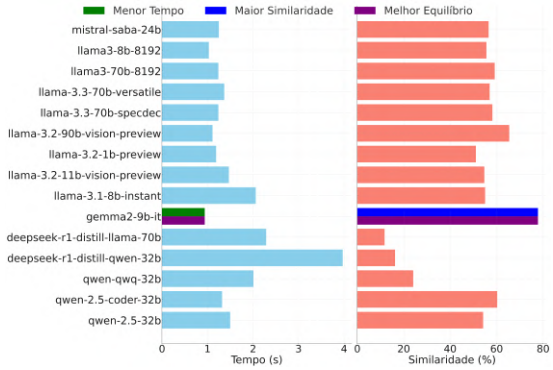
Figura 2. Respostas interpretativas, diversas temperaturas.

Com temperatura 1.5 (Figura 2(d)) e especialmente em 2 (Figura 3(e)), esse afastamento se intensifica. Embora o núcleo semântico ainda esteja presente em muitos casos, há perda de precisão lexical, revelando que até mesmo perguntas objetivas são sensíveis ao aumento da temperatura, ainda que de forma menos pronunciada do que as interpretativas.

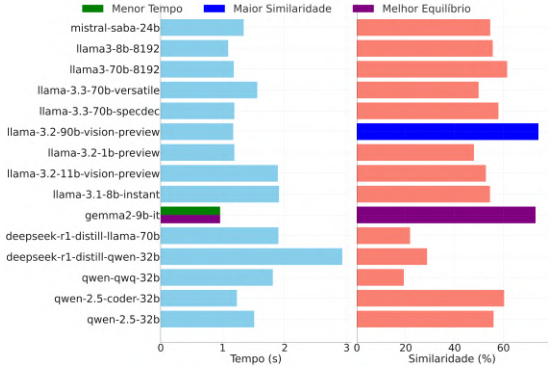
A Figura 3 ilustra a correlação entre tempo de inferência e similaridade semântica para as tarefas interpretativas. Observa-se que modelos com maior tempo de resposta, como o llama3-70b-8192, frequentemente produzem respostas mais próximas da referência, sugerindo um processo de raciocínio mais elaborado. Essa visualização reforça a hipótese de que o tempo de inferência pode servir como indicador indireto de profundidade cognitiva do modelo, particularmente em tarefas que exigem inferência semântica e construção argumentativa.



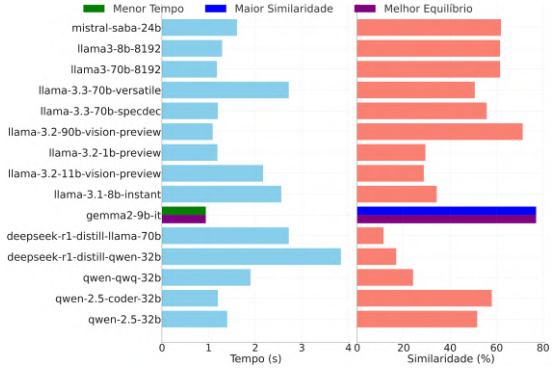
(a) Temperatura: 0



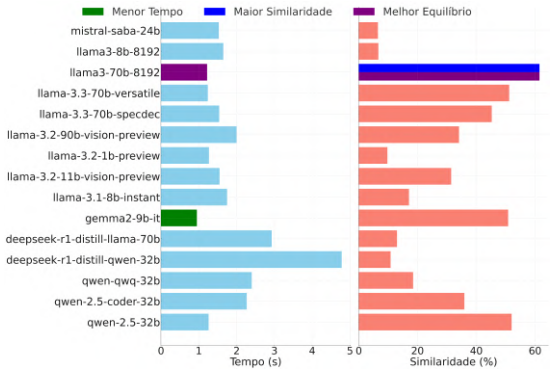
(b) Temperatura: 0.5



(c) Temperatura: 1



(d) Temperatura: 1.5



(e) Temperatura: 2

Figura 3. Respostas literais, diversas temperaturas.

De forma geral, os resultados demonstram que modelos maiores e mais recentes tendem a manter desempenho estável mesmo sob condições de entropia crescente, enquanto modelos menores se beneficiam de configurações de baixa ou média temperatura. A temperatura 0.5 se revela, empiricamente, como a melhor zona de equilíbrio entre tempo e qualidade para ambos os tipos de tarefa, sendo eficaz em perguntas abertas. A análise conjunta dos dois tipos de tarefas permite concluir que a temperatura atua não apenas como regulador de aleatoriedade, mas também como modulador da profundidade interpretativa — funcionando, portanto, como um parâmetro com implicações tanto sintáticas quanto semânticas.

5. Lições Aprendidas

Durante a condução e análise deste estudo, emergiram percepções importantes que transcendem os resultados quantitativos obtidos. Embora os gráficos e métricas tenham fornecido uma base sólida para avaliar desempenho e qualidade textual, algumas lições adicionais puderam ser inferidas com base na observação atenta dos comportamentos dos modelos ao longo dos experimentos.

Primeiramente, ficou evidente que há uma assimetria entre os modelos na forma como lidam com a tarefa interpretativa em comparação à tarefa objetiva. Modelos que tiveram bom desempenho na pergunta direta nem sempre conseguiram manter coerência na pergunta aberta, sugerindo que a generalização semântica e a preservação do significado não são capacidades automaticamente herdadas de arquiteturas com bom desempenho literal. Essa observação reforça a importância de avaliar modelos em múltiplas dimensões cognitivas, e não apenas por sua capacidade de memorização ou repetição de trechos aprendidos.

Outro ponto relevante refere-se ao papel do parâmetro de temperatura, que atua como amplificador das características intrínsecas de cada modelo. Em temperaturas mais elevadas, modelos mais robustos demonstraram maior inventividade sem comprometer a coerência, enquanto modelos menos sofisticados revelaram instabilidade semântica. Isso sugere que a temperatura funciona como uma espécie de "lente de estresse", revelando os limites estruturais e de treinamento das arquiteturas, e destacando sua resiliência interpretativa.

Além disso, o tempo de inferência, comumente tratado como uma métrica objetiva, mostrou-se sensível às estratégias cognitivas utilizadas pelos modelos. Em diversas instâncias, notou-se que os modelos que apresentaram maior tempo de execução não o fizeram por ineficiência computacional bruta, mas sim por engajarem-se em raciocínios mais elaborados, possivelmente baseados em cadeias de pensamento (*chain-of-thought reasoning*) (Wei *et al.*, 2022). Esse comportamento se manifesta por meio de justificativas intermediárias ou explicações passo a passo que precedem a resposta. Embora essa abordagem amplie o tempo de geração, eleva também a profundidade semântica, especialmente em tarefas interpretativas. Assim, o tempo de inferência, longe de ser um simples indicador de desempenho bruto, pode funcionar como um marcador indireto do estilo e da complexidade cognitiva envolvida na geração textual de certos modelos.

Uma lição transversal a todo o estudo diz respeito aos efeitos qualitativos da temperatura. Longe de representar apenas um fator de aleatoriedade, ela influencia diretamente a natureza das respostas, afetando sua objetividade, grau de detalhamento e diversidade lexical. Dessa forma, o ajuste desse parâmetro deve ser feito com atenção às implicações semânticas, sobretudo em contextos em que a precisão interpretativa e a responsabilidade comunicativa são essenciais. Para complementar as inferências observadas qualitativamente, a Tabela 1 sintetiza comparativamente o desempenho dos modelos avaliados em tarefas literais e interpretativas, bem como sua resiliência a altas temperaturas e propensão a raciocínios mais elaborados. A análise cruzada revela que arquiteturas mais avançadas, como a llama3-70b-8192, mantêm desempenho consistente mesmo em contextos de alta entropia, ao passo que modelos menos complexos, como o gemma2-9b-it, perdem significativamente em coerência semântica. Ademais, a capacidade de elaborar respostas com maior densidade argumentativa, geralmente associada a maior tempo de inferência, surge como uma característica distintiva de arquiteturas mais sofisticadas, o que os torna especialmente adequados para aplicações educacionais que exigem interpretação crítica e explicações intermediárias.

Tabela 1. Comparação qualitativa dos modelos em tarefas educacionais

Modelo	Literal	Interpretativa	Resiliência	Raciocínio Elaborado
llama3-8b-8192	Alta	Alta	Média	Sim
llama3-70b-8192	Alta	Alta	Alta	Sim
gemma2-9b-it	Média	Baixa	Baixa	Não
qwen-2.5-coder-32b	Alta	Média	Média	Parcial
llama-3.3-70b-specdec	Alta	Média	Alta	Sim

Esses resultados têm implicações relevantes no cenário educacional contemporâneo, ao demonstrarem que o ajuste da temperatura impacta diretamente o desempenho cognitivo simulado pelos LLMs em atividades cognitivas complexas, como análise textual, argumentação e produção discursiva. A Tabela 1 resume comparativamente as competências dos modelos em tarefas literais e interpretativas, bem como sua resiliência a diferentes níveis de entropia, revelando a diversidade de comportamentos entre as arquiteturas analisadas.

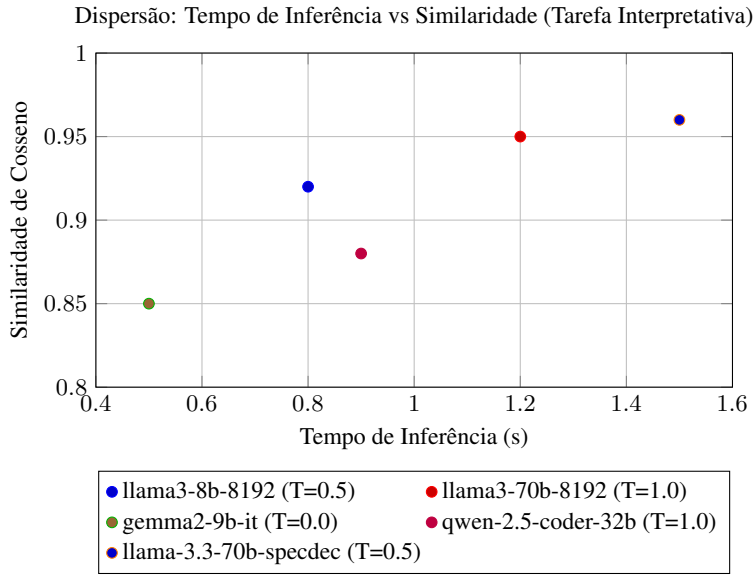


Figura 4. Correlação entre tempo de inferência e similaridade semântica nas tarefas interpretativas

Complementarmente, a Figura 4 demonstra que o tempo de inferência pode atuar como marcador indireto da complexidade semântica envolvida, uma vez que modelos que recorrem a raciocínio mais elaborado tendem a apresentar maior latência, mas também maior similaridade às respostas de referência. Esses achados reforçam a aplicabilidade dos LLMs em contextos educacionais que demandam sensibilidade lexical, profundidade interpretativa e parametrização alinhada ao perfil cognitivo dos estudantes. Ao estabelecer uma base empírica robusta para o uso responsável e eficiente da inteligência artificial na educação, este trabalho contribui para a construção de ferramentas instrucionais mais éticas, transparentes e centradas no processo de aprendizagem.

6. Conclusões e Trabalhos Futuros

Este estudo evidenciou que o parâmetro de temperatura exerce influência não apenas sobre a variação lexical das respostas geradas por modelos de linguagem, mas também sobre sua capacidade interpretativa e precisão semântica. A análise sistemática de 15 LLMs, sob cinco diferentes configurações de temperatura, revelou que a natureza da tarefa — interpretativa ou literal — altera significativamente a sensibilidade dos

modelos a esse parâmetro. Temperaturas mais baixas favorecem respostas diretas, com alta fidelidade lexical e menor custo computacional, enquanto valores intermediários (especialmente em torno de 0.5) se destacam por equilibrar diversidade e coerência, sendo particularmente eficazes em tarefas que exigem inferência e interpretação. Além disso, observou-se que modelos maiores e mais recentes demonstram maior robustez sob entropia elevada, ao passo que arquiteturas menores tendem a responder melhor a configurações mais conservadoras.

Como trabalhos futuros, propõe-se expandir o escopo das tarefas avaliadas, incorporando questões com maior diversidade temática e complexidade discursiva, como perguntas argumentativas, analíticas e contraditórias, a fim de testar a consistência interna e a estabilidade semântica das respostas. Pretende-se também explorar o impacto da temperatura em atividades multimodais e em contextos de tutoria adaptativa, ampliando ainda mais o potencial transformador dos LLMs na educação.

Referências

Alnaasan, N.; Huang, H.-R.; Shafi, A.; Subramoni, H.; Panda, D. K. Characterizing communication in distributed parameter-efficient fine-tuning for large language models. In: **2024 IEEE Symposium on High-Performance Interconnects (HOTI)**. [S.l.: s.n.], 2024. p. 11–19.

Castilho, R. E. de; Gurevych, I. Semantic service retrieval based on natural language querying and semantic similarity. In: **2011 IEEE Fifth International Conference on Semantic Computing**. [S.l.: s.n.], 2011. p. 173–176.

Hao, Y.; Wang, J.; Li, X. A scholarly definition of artificial intelligence (ai). **Journal of Information Technology**, Taylor & Francis, v. 38, n. 1, p. 1–15, 2023. Disponível em: ⟨https://www.tandfonline.com/doi/full/10.1080/10584609.2023.2290497⟩.

Holtzman, A.; Buys, J.; Du, L.; Forbes, M.; Choi, Y. The curious case of neural text degeneration. In: **International Conference on Learning Representations (ICLR)**. [s.n.], 2020. Disponível em: ⟨https://arxiv.org/abs/1904.09751⟩.

Huang, Y. *et al.* Advancing transformer architecture in long-context large language models: A comprehensive survey. **arXiv preprint arXiv:2311.12351**, 2023. Disponível em: ⟨https://arxiv.org/abs/2311.12351⟩.

Khurana, D.; Koli, A.; Khatter, K.; Singh, S. Natural language processing for humanitarian action. **Frontiers in Big Data**, Frontiers, v. 6, p. 1082787, 2023. Disponível em: ⟨https://www.frontiersin.org/articles/10.3389/fdata.2023.1082787/full⟩.

Liu, H. *et al.* Understanding the effect of heterogeneous training data on large language model training. **arXiv preprint arXiv:2310.11413**, 2023. Disponível em: ⟨https://arxiv.org/abs/2310.11413⟩.

Loiola, A.; Sachete, A. S.; Gomes, R. S.; Grandi, R. H. Ia generativa em competências discursivas na educação básica. **Revista Eletrônica de Educação**, v. 18, n. 1, p. e6680122, out. 2024. Disponível em: ⟨https://www.reveduc.ufscar.br/index.php/reveduc/article/view/6680⟩.

Minaee, S. *et al.* Large language models: A survey. **arXiv preprint arXiv:2402.06196**, 2024. Disponível em: ⟨https://arxiv.org/abs/2402.06196⟩.

Reimers, N.; Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. In: Association for Computational Linguistics. **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing**. 2019. p. 3982–3992. Disponível em: ⟨https://aclanthology.org/D19-1410/⟩.

Sachete, A. d. S.; Loiola, A. V. d. S.; Gomes, R. S. Adaptivegpt: Towards intelligent adaptive learning. **Multimed Tools Appl**, v. 83, p. 89461–89477, 2024.

Sachete, A. d. S.; Loiola, A. V. d. S. d. F.; Pereira, M.; Rossi, F. D.; Gomes, R. S. Enemia: correção de redações do enem com inteligência artificial. **Texto Livre**, v. 18, p. e58426, jun. 2025. Disponível em: ⟨https://periodicos.ufmg.br/index.php/textolivre/article/view/58426⟩.

Wang, Y.; Zhang, Z.; Chen, H.; Shen, H. Reasoning with large language models on graph tasks: The influence of temperature. In: **2024 5th International Conference on Computer Engineering and Application (ICCEA)**. [S.l.: s.n.], 2024. p. 630–634.

Wei, J. *et al.* Chain-of-thought prompting elicits reasoning in large language models. **arXiv preprint arXiv:2201.11903**, 2022. Disponível em: ⟨https://arxiv.org/abs/2201.11903⟩.

Örpek, Z.; Tural, B.; Destan, Z. The language model revolution: Llm and slm analysis. In: **2024 8th International Artificial Intelligence and Data Processing Symposium (IDAP)**. [S.l.: s.n.], 2024. p. 1–4.