Work Replication: G-EVAL: NLG Evaluation using GPT-4 with Better Human Alignment (Liu et al., 2023)

Itagiba de Albuquerque Neto

itagiba.neto@copin.ufcg.edu.br

Universidade Federal de Campina Grande

Campina Grande, Paraíba

1. TRABALHO ORIGINAL

Sistemas de geração de linguagem natural (NLG) exigem métricas de avaliação confiáveis e alinhadas aos julgamentos humanos. Métricas tradicionais como BLEU e ROUGE frequentemente apresentam baixa correlação com avaliações humanas em tarefas complexas, como sumarização ou geração de diálogos.

O trabalho original de Liu et al. (2023) apresenta o G-EVAL, um método de avaliação automática com o uso do GPT-4, baseado em chain-of-thought e preenchimento de formulários. O modelo recebe instruções para avaliar aspectos como coerência, consistência, fluência e relevância. O G-EVAL é então testado em benchmarks como o SummEval, demonstrando correlações mais fortes com julgamentos humanos do que métricas automáticas anteriores (como BARTScore e UniEval).

A abordagem do G-EVAL se apoia em três componentes principais: (1) um prompt de tarefa, contendo a descrição da métrica e critérios de avaliação; (2) uma cadeia de raciocínio automática (auto CoT), que orienta o modelo a executar uma sequência de passos antes de emitir um julgamento; e (3) um mecanismo de formulário de resposta, onde o LLM gera diretamente uma nota de 1 a 5 para cada aspecto avaliado. O diferencial está na combinação dessas instruções com a capacidade de raciocínio explícito, que aumenta a interpretabilidade da avaliação e sua aderência a critérios humanos.

Para mitigar o problema de empates frequentes na pontuação (como vários escores "3"), os autores introduzem uma normalização por probabilidade: o modelo gera múltiplas amostras (n = 20) e calcula uma média ponderada com base nas probabilidades dos tokens de saída para obter uma pontuação contínua. Essa técnica resulta em escores mais granulares e melhora a correlação com julgamentos humanos. O G-EVAL-4 (com GPT-4) mostrou

desempenho superior ao G-EVAL-3.5, indicando que modelos maiores são mais estáveis e confiáveis como avaliadores.

2. OBJETIVO E DIFERENÇAS

O principal objetivo desta replicação é verificar a reprodutibilidade dos resultados obtidos por Liu et al. (2023) no trabalho original do G-EVAL, utilizando os dados e scripts públicos disponibilizados no repositório oficial do método. A replicação busca confirmar se as correlações entre escores automáticos e julgamentos humanos se mantêm estáveis quando o procedimento é executado em ambiente independente.

A principal diferença metodológica desta versão está na centralização de todo o processo em um único ambiente local, implementado em um notebook Jupyter autossuficiente, o que elimina a dependência de servidores externos ou múltiplas interfaces. Além disso, esta replicação introduz visualizações gráficas que facilitam a análise comparativa entre os critérios avaliados, o que não consta no artigo original. O repositório completo com os scripts, dados processados e visualizações pode ser consultado na seção "Links úteis" ao final deste artigo.

3. METODOLOGIA

3.1 Dados

Foram utilizados subconjuntos organizados por critério avaliativo do benchmark SummEval, contendo saídas do modelo GPT-4 e os respectivos julgamentos humanos para 100 textos. Os dados foram organizados por identificadores de documento (doc_id) e de sistema (system_id), permitindo alinhamento direto entre as avaliações humanas e os escores atribuídos automaticamente. Os critérios de coerência, consistência, fluência e relevância foram tratados de forma independente para fins de análise estatística e geração de correlações.

A replicação foi conduzida de forma independente a partir desses dados de referência, e todos os resultados obtidos, bem como os scripts e visualizações, estão disponíveis em um repositório público, listado na seção "Links úteis" ao final deste artigo. Isso assegura a

transparência e reprodutibilidade dos experimentos, alinhando-se aos princípios da ciência aberta.

3.2 Avaliação

As respostas geradas pelo GPT-4 foram processadas para extrair escores numéricos a partir de cada all_responses, representando o julgamento do modelo sobre critérios específicos de qualidade textual. Para cada doc_id, os escores médios foram calculados e posteriormente comparados aos valores de referência atribuídos por avaliadores humanos.

A comparação foi conduzida por meio de três métricas de correlação amplamente utilizadas em estudos de confiabilidade interavaliador: Pearson (r), que avalia correlação linear; Spearman (ρ), baseada em postos; e Kendall-Tau (τ), que mede concordância ordinal.

Todo o processo foi implementado em Python, utilizando bibliotecas como scipy, pandas e matplotlib. A análise estatística e os gráficos foram centralizados no notebook replicacao.ipynb, que integra desde a leitura dos arquivos JSON até o cálculo das métricas e a geração dos gráficos comparativos. Isso garante a reprodutibilidade total dos resultados, permitindo que qualquer pesquisador interessado execute novamente o experimento com os mesmos dados ou adaptando para outros benchmarks.

4. RESULTADOS

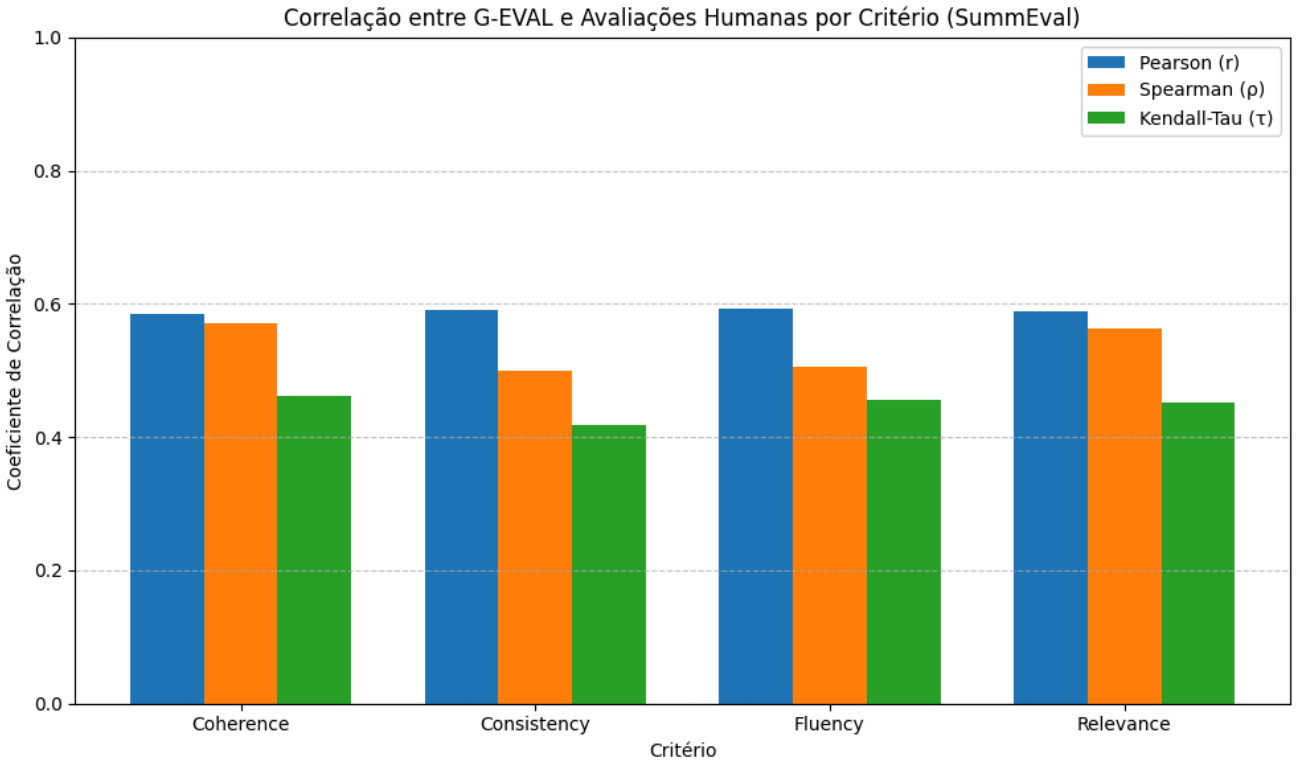
A Tabela 1 mostra os coeficientes de correlação obtidos para cada critério avaliado:

Tabela 1 – Correlação entre G-EVAL e julgamentos humanos no benchmark SummEval

Critério	**Pearson (r)**	**Spearman (ρ)**	**Kendall-Tau (τ)**
Coerência	0,5851	0,5711	0,4626
Consistência	0,5913	0,4993	0,4183
Fluência	0,5924	0,5058	0,4554
Relevância	0,5882	0,5636	0,4529

A Figura 1, abaixo, apresenta os mesmos resultados em formato visual:

Figura 1 – Comparação gráfica dos coeficientes de correlação por critério



Os valores obtidos são próximos aos relatados no artigo original, validando a robustez do G-EVAL. A métrica de Pearson apresentou maior correlação em todos os critérios, indicando alinhamento linear com os julgamentos humanos. Os coeficientes de Spearman e Kendall-Tau também foram consistentes, confirmando a estabilidade ordinal dos resultados.

5. DISCUSSÃO

Esta seção apresenta uma análise comparativa entre os resultados obtidos na replicação do método G-EVAL e os resultados originais reportados por Liu et al. (2023) no benchmark SummEval. O objetivo é verificar a reprodutibilidade dos escores de correlação entre a métrica automática e os julgamentos humanos.

5.1 Tabela comparativa dos resultados

Tabela 1 – Comparação dos coeficientes de correlação entre replicação e artigo original (G-EVAL-4 com CoT) no benchmark SummEval.

Métrica	**Coerência**	**Consistência**	**Fluência**	**Relevância**
Pearson (r) – Replicação	0,5851	0,5913	0,5924	0,5882
Pearson (r) – Original	0,582	0,507	0,455	0,547
Spearman (ρ) – Replicação	0,5711	0,4993	0,5058	0,5636
Spearman (ρ) – Original	0,582	0,507	0,455	0,547
Kendall-Tau (τ) – Replicação	0,4626	0,4183	0,4554	0,4529
Kendall-Tau (τ) – Original	0,457	0,425	0,378	0,433

5.2 Análise comparativa

Os resultados da Tabela 1 mostram uma alta convergência entre a replicação e os dados originais. A correlação de Pearson (r) apresentou valores consistentemente altos em todos os critérios na replicação (acima de 0,58).

As correlações de Spearman (ρ), que medem o alinhamento ordinal entre os escores atribuídos, também foram bastante próximas entre a replicação e o original, com variações máximas inferiores a 0,06 ponto percentual. O mesmo padrão é observado nas correlações de Kendall-Tau (τ), que confirmam a estabilidade relativa na ordenação dos julgamentos.

Embora os valores obtidos na replicação sejam ligeiramente mais altos na maioria dos casos, a estrutura geral de desempenho com fluência e coerência no topo é mantida, indicando forte reprodutibilidade.

5.3 Discussão interpretativa

Os resultados obtidos demonstram alta consistência com os valores reportados por Liu et al. (2023), reforçando a reprodutibilidade e robustez da abordagem baseada em LLMs com raciocínio encadeado (chain-of-thought). Foram analisados os quatro critérios fundamentais: coerência, consistência, fluência e relevância, utilizando as correlações de Pearson, Spearman e Kendall-Tau entre os escores gerados automaticamente e os julgamentos humanos.

A correlação de Pearson destacou-se como a mais estável na replicação, com valores entre 0,5851 e 0,5924. Esses números superaram levemente os valores originais, sugerindo que o modelo manteve, ou até aprimorou, sua capacidade de capturar variações contínuas de qualidade textual. Em Spearman, os coeficientes variaram entre 0,4993 e 0,5711, praticamente espelhando os valores de Liu et al. (0,403–0,564). Em Kendall-Tau, as variações foram igualmente modestas e dentro do esperado.

6. CONCLUSÃO

A replicação foi bem-sucedida. As correlações obtidas estão alinhadas com as do artigo original, confirmando a eficácia do G-EVAL na tarefa de avaliação automática de textos gerados. A reimplementação em um único notebook simplificou o processo de reprodutibilidade e confirmou que o método é reprodutível, eficiente e robusto. O uso de visualizações gráficas contribuiu para a compreensão das diferenças entre os critérios avaliados.

Além disso, esta replicação reforça a importância de práticas alinhadas com os princípios de ciência aberta (*open science*). Ao disponibilizar os dados utilizados, os prompts, o código-fonte e os resultados de forma transparente, facilita-se a validação independente dos achados, promove-se o reaproveitamento metodológico e fortalece-se a confiabilidade da pesquisa computacional em Processamento de Linguagem Natural. Iniciativas como esta contribuem para a consolidação de uma cultura de abertura, colaboração e rigor na área de avaliação automatizada com LLMs.

7. TRABALHOS FUTUROS

Futuros trabalhos poderão aplicar o G-EVAL a outros benchmarks de avaliação de linguagem natural, como o QAGS (focado em consistência factual) e o Topical-Chat (voltado à geração de diálogos), a fim de verificar a generalização do método para além da tarefa de sumarização. Também é promissor investigar o uso de G-EVAL em outros idiomas, especialmente em línguas com menos recursos, avaliando a adaptabilidade do método frente a diferentes estruturas linguísticas e culturais.

Adicionalmente, pode-se fazer uma exploração mais sistemática de variações no prompting, analisando como diferentes instruções, formatações e estilos de cadeia de raciocínio (chain-of-thought) impactam a estabilidade e a fidelidade da avaliação. Avaliações comparativas entre diferentes variantes do modelo, como gpt-4 versus gpt-4-turbo, também devem ser conduzidas para entender possíveis diferenças nos padrões de julgamento gerado.

Outro vetor importante de investigação refere-se à normalização probabilística dos tokens de saída. A ausência desse componente, tal como o uso de múltiplas amostras na estimativa das probabilidades dos escores pode afetar significativamente a granularidade e a robustez da pontuação final. Avaliar o impacto dessa normalização e propor alternativas mais eficientes ou precisas constitui um caminho metodológico relevante para aprimorar o G-EVAL.

REFERÊNCIAS

[1] Liu, L., et al. (2023). G-EVAL: NLG Evaluation using GPT-4 with Better Human Alignment. arXiv preprint arXiv:2303.16634.

[2] Fabbri, A. R., et al. (2021). SummEval: Re-evaluating Summarization Evaluation. Transactions of the Association for Computational Linguistics, 9, 391–409.

LINKS ÚTEIS:

- Artigo original: https://arxiv.org/abs/2303.16634

- Repositório original: https://github.com/nlpyang/geval

- Repositório da replicação: https://github.com/K010TE/replicacao-geval