UNIVERSIDADE DO ESTADO DO PARÁ – UEPA

BACHARELADO EM ENGENHARIA DE SOFTWARE

Davi Costa Mendes da Silva

Ryan Souza Santana

PROJETO INTEGRADO I

APONTAMENTOS

Trabalho realizado sob orientação do *Prof. Dr. Eng. Ítalo Flexa Di Paolo* como requisito parcial e complementar de avaliação na disciplina Projeto Integrado I.

Ananindeua

2025

O presente documento representa parte da avaliação de Projeto Integrado I ministrada pelo Prof. Dr. Eng. Ítalo Flexa Di Paolo. Nele, estão os apontamentos exigidos com a referência de cada artigo, evento, anais, relatório e revista utilizado na elaboração do projeto de pesquisa. Em cada tópico estão apresentados: o título do documento, a referência do mesmo, um resumo do que se trata e, por último, para que a obra foi utilizada.

Ademais, foram utilizados 10 documentos acadêmicos para a elaboração do projeto. Estes, são fundamentais para a contextualização, definição do estado da arte e auxílio na metodologia aplicada na pesquisa. Pela Inteligência Artificial e seu Modelo Linguagem Larga serem uma realidade recente, a maioria das obras possuem datas de publicação recente, apresentando um tópico com amplo valor e pouco explorado em comparação a outras áreas mais consolidadas.

1. **Fabbri et al. (2021) — “SummEval: Re-evaluating Summarization Evaluation”**

Fabbri, A. R. et al. SummEval: Re-evaluating Summarization Evaluation. **Transactions of the Association for Computational Linguistics**, v. 9, p. 391–409, 26 abr. 2021. Disponivel em: https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00373/100686/SummEval-Re-evaluating-Summarization-Evaluation Acesso em: 19 de maio de 2025.

O artigo propõe uma reavaliação das métricas automáticas de sumarização, utilizando um conjunto diversificado de 23 modelos e 14 métricas. Os autores mostram que métricas tradicionais como ROUGE têm limitações significativas e destacam a necessidade de métodos de avaliação mais alinhados ao julgamento humano. Introduzem também uma toolkit para avaliação padronizada. A pesquisa busca **padronizar e ampliar** a avaliação de sumários gerados por máquinas.

Uso: resenha em dois pontos da justificativa, como citação para embasamento do referencial teórico e uma na metodologia como resenha.

Citação do referencial teórico: “Coherence - the collective quality of all sentences. We align this dimension with the DUC 396 quality question (Dang, 2005) of structure and coherence whereby ‘‘the summary should be well-structured and well-organized. The summary should not just be a heap of related information, but should build from sentence to sentence to a coherent body of information about a topic.’’Consistency - the factual alignment between the summary and the summarized source. A factually consistent summary contains only statements that are entailed by the source document. Annotators were also asked to penalize summaries that contained hallucinated facts. Fluency - the quality of individual sentences. Drawing again from the DUC quality guidelines, sentences in the summary ‘‘should have no formatting problems, capitalization errors or obviously ungrammatical sentences (e.g., fragments,missing components) that make the text difficult to read.’’Relevance - selection of important content from the source. The summary should include only important information from the source document. Annotators were instructed to penalize summaries that contained redundancies and excess information.”

---

2. **Kryscinski et al. (2019) — “Neural Text Summarization: A Critical Evaluation”**

Kryscinski, W. et al. **Neural Text Summarization**: A Critical Evaluation. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Anais...Stroudsburg, PA, USA: Association for Computational Linguistics, 2019. Disponível em: https://www.aclweb.org/anthology/D19-1051 Acesso em: 19 de maio de 2025.

Este trabalho critica o atual estado da pesquisa em sumarização neural. Os autores apontam três problemas principais: (1) datasets ruidosos, (2) métricas automáticas pouco confiáveis e (3) modelos que se aproveitam de **viés de layout** (como o uso das primeiras frases dos textos). O estudo mostra que o progresso nas benchmarks é limitado, e argumenta que há uma **estagnação** na área devido a deficiências metodológicas. Defende uma reestruturação das práticas de avaliação e construção de datasets.

Uso resenha no referencial teórico.

---

3. **Martins (2001) — "Introdução à Sumarização Automática"**

Martins, C. B. et al. **Introdução à sumarização automática**. Relatório Técnico RT-DC, v. 2, 2001. https://sites.icmc.usp.br/taspardo/rtdc00201-cmartinsetal.pdf Acesso em: 29 de abril de 2025.

O relatório explora as bases da sumarização automática de textos. Discute os desafios de replicar a variação e a intenção presentes nos sumários humanos, e apresenta arquiteturas de sistemas e métodos históricos, como uso de palavras-chave. O trabalho é uma referência sólida para compreender os fundamentos e as limitações da sumarização automática.

Uso: resumo na justificativa e referencial teórico.

---

4. **Moon (2025) — "Benchmarking LLMs for Calculus Problem-Solving"**

Moon, I. H. **Benchmarking Large Language Models for Calculus Problem-Solving**: A Comparative Analysis. 30 mar. 2025. Disponível em: <https://arxiv.org/abs/2504.13187> Acesso em: 19 de maio de 2025.

O estudo comparou cinco LLMs em problemas de cálculo diferencial. O trabalho conclui que, embora promissores, os LLMs ainda dependem do ensino humano para fomentar compreensão conceitual profunda, sendo úteis como **ferramentas complementares** em ambientes educacionais.

Uso: resenha na justificativa e no referencial teórico.

---

5. **Nze (2024) — "AI-Powered Chatbots"**

Nze, S. AI-Powered Chatbots. **Global Journal of Human Resource Management**, v. 12, n. 1, p. 10-25, 2024. Disponível em: https://eajournals.org/gjhrm/vol12-issue-6-2024/ai-powered-chatbots/ Acesso em: 13 de maio de 2025.

A obra aborda como os relacionamentos cliente/empresa mudaram com a implementação de chatbots e como estas Inteligências Artificiais se tornaram uma alternativa eficaz para abordagens variadas com o usuário final.

Uso: citação na contextualização.

Citação utilizada de referência: "Artificial Intelligence (AI)-powered chatbots have emerged as a transformative technology, fundamentally changing how businesses and organizations engage with their customers by providing real-time, personalized communication. These chatbots, driven by sophisticated algorithms, utilize Natural Language Processing (NLP) and Machine Learning (ML) to understand, interpret, and respond to human language in a manner that is contextually appropriate and relevant."

---

6. **Shen et al. (2023) — "LLMs are Not Yet Human-Level Evaluators for Abstractive Summarization"**

Shen, C. et al. **Large Language Models are Not Yet Human-Level Evaluators for Abstractive Summarization**. Findings of the Association for Computational Linguistics: EMNLP 2023. Anais...Stroudsburg, PA, USA: Association for Computational Linguistics,

2023. Disponível em: https://aclanthology.org/2023.findings-emnlp.278 Acesso em: 19 de maio de 2025.

O artigo questiona o uso de LLMs como substitutos de juízes humanos na avaliação de sumários **abstrativos**. Embora LLMs como ChatGPT e GPT-4 superem métricas automáticas tradicionais, eles apresentam **inconsistência** ao avaliar sistemas com desempenho semelhante e em tarefas de alta qualidade. As avaliações variam conforme o tipo de sumário (fluência, coerência, factualidade).

Uso: resenha no referencial teórico.

---

7. **Son; Won; Lee (2025) — "Optimizing LLMs: A Deep Dive into Prompt Engineering Techniques"**

Son, M.; Won, Y. J.; Lee, S. **Optimizing Large Language Models**: A Deep Dive into Effective Prompt Engineering Techniques. Applied Sciences (Switzerland), v. 15, n. 3, 1 fev. 2025. Disponível em: https://www.mdpi.com/2076-3417/15/3/1430 Acesso em: 19 maio 2025.

Este artigo analisa técnicas avançadas de **Prompt Engineering** para otimizar a performance de LLMs. Técnicas como **In-Context Learning (ICL)**, **Chain of Thought (CoT)**, **Retrieval-Augmented Generation (RAG)**, **Step-by-Step Reasoning (SSR)** e **Tree of Thought (ToT)** foram testadas em diversos benchmarks. A escolha da técnica ideal depende do tipo de tarefa. O trabalho reforça que mesmo LLMs avançados ainda se beneficiam significativamente de engenharia de prompt adequada e personalizada.

Uso: resenha no referencial teórico.

---

8. **Tabosa et al. (2020) — "Avaliação do Desempenho de um Software de Sumarização Automática de Textos"**

Tabosa, H. R. et al. **Avaliação do desempenho de um software de sumarização automática de textos**. Informação & Informação, v. 25, n. 1, p. 189, 1 abr. 2020.

http://www.uel.br/revistas/uel/index.php/informacao/article/view/35928 Acesso em: 19 maio 2025.

Este estudo avalia um software brasileiro de sumarização automática com base em testes cegos comparando seus resumos com os produzidos por humanos. Foram utilizados critérios como **correção gramatical**, **coerência**, **preservação das ideias centrais** e **extensão do resumo**. Os resultados mostraram equivalência em quatro dos cinco critérios. A principal deficiência foi a extensão excessiva dos resumos gerados pela máquina. O estudo sugere que o software é promissor, mas precisa de ajustes e testes mais amplos para consolidação.

Uso resenha na justificativa e resenha no referencial teórico como complementação.

---

9. **Yampolskiy(2016) — "Taxonomy of Pathways to Dangerous Artificial Intelligence"**

Yampolskiy, Roman V. Taxonomy of Pathways to Dangerous Artificial Intelligence. In: **AAAI Workshop: AI, Ethics, and Society**. 2016. p. 143-148. Disponível em: https://cdn.aaai.org/ocs/ws/ws0156/12566-57418-1-PB.pdf Acesso em: 13 maio 2025.

Workshop realizado pela University of Louisville em que o autor destaca a possibilidade de erro que as IAs podem apresentar ao longo do seu ciclo de vida. Nele, são apresentados perigos pré e pós instalação de uma inteligência artificial de fins variados, seja um equipamento militar, de exploração extraterrestre ou um simples radar de velocidade.

Uso: resenha na contextualização

---

10. **Zhang; Liu; Zhang (2023) — "Extractive Summarization via ChatGPT for Faithful Summary Generation"**

Zhang, H.; Liu, X.; Zhang, J. **Extractive Summarization via ChatGPT for Faithful Summary Generation**. Findings of the Association for Computational Linguistics: EMNLP 2023. Anais… Stroudsburg, PA, USA: Association for Computational Linguistics, 2023. Disponível em: https://aclanthology.org/2023.findings-emnlp.214 Acesso em: 19 maio 2025.

Este artigo avalia o desempenho do ChatGPT na tarefa de sumarização **extrativa**. Os resultados mostram que, embora o modelo tenha desempenho inferior em métricas tradicionais como ROUGE, ele se destaca em métricas baseadas em LLMs quanto à **fidelidade** do conteúdo. Além disso, o uso de abordagens como **"extract-then-generate"**, aprendizado por contexto (ICL) e raciocínio encadeado (CoT) melhora a performance. O estudo propõe que a sumarização extrativa via LLMs pode ser promissora para mitigar problemas de alucinação.

Uso: resumo e resenha na justificativa e um resumo no referencial teórico.