

Discente: Ian de Brito de Azevedo

Comparação de Modelos de Linguagem

1. Introdução

Este relatório apresenta uma análise comparativa dos modelos de linguagem Gemma, LLaMA e Mistral com base em métricas quantitativas e qualitativas. O objetivo é avaliar o desempenho de cada modelo na geração de resumos, considerando aspectos como previsibilidade, coerência textual, precisão semântica e fidelidade ao conteúdo de referência.

2. Metodologia

2.1. Modelos Avaliados

Gemma (7b): Desenvolvido pelo Google DeepMind, é um modelo de código aberto baseado na arquitetura Transformer, otimizado para tarefas de NLP e disponível sob uma licença permissiva para uso comercial e acadêmico.

LLaMA (3.2:3b): Criado pela Meta, é um modelo de linguagem de código aberto voltado para eficiência e desempenho em tarefas de compreensão e geração de texto, sendo amplamente utilizado na pesquisa e desenvolvimento de IA.

Mistral (v0.3): Desenvolvido pela Mistral AI, é um modelo de código aberto focado em eficiência computacional e geração de texto precisa, projetado para oferecer alta performance com menor custo computacional.

Os modelos foram executados através do **Ollama**, uma plataforma voltada para a execução e otimização de modelos de linguagem em ambiente local. Sendo executado por um container Docker.

2.2. Prompt utilizado no experimento

Artigo selecionado: **INTELIGÊNCIA ARTIFICIAL E A GERAÇÃO DE RESUMOS DE ARTIGOS**

Leia o seguinte texto e gere um resumo de no máximo 100 palavras, destacando os principais conceitos e ideias-chave. O resumo deve ser técnico, sem simplificações excessivas e sem introduções genéricas: A Inteligência Artificial é um tema que está em voga. Reiteradamente ouvimos as pessoas falando acerca de profissionais da área de tecnologia ao meio acadêmico e a sociedade em geral. A Inteligência Artificial (IA) é uma subárea dentro da Ciência da Computação responsável por pesquisar e propor a elaboração de dispositivos computacionais que tentam simular o intelecto humano (Silva, 2013). No contexto das organizações, podem ajudar na melhoria da efetividade e produtividade (Ransbotham et al, 2021). No entanto, como campo interdisciplinar, a IA possui o seu tentáculo em múltiplas áreas do conhecimento, tais como matemática, robótica e ciência da computação (Russell; Norvig, 2013). Neste sentido, esta área é de fundamental relevância, ao passo que facilita no desenvolvimento de trabalho de forma eficiente e eficaz, principalmente nas áreas de conhecimento que hoje estão integradas com a IA. No entanto, Sousa (2023), alerta para algumas desvantagens que a utilização inadequada da IA pode trazer nos campos de ensino em EAD, tais como: alta dependência das infraestruturas e dependência digital em excesso. Diante deste contexto, percebe-se o aumento da utilização da IA em pesquisas acadêmicas e profissionais (Gontijo; Araújo, 2021). Assim, o objetivo do trabalho foi estudar a elaboração de resumos feitos com ferramentas de IA, no intuito de verificar as diferenças substantivas entre os resumos, utilizando como fonte quatro artigos na área da Ética. Além de desenvolver a capacidade analítica no desenvolvimento dos resumos, foram feitas reflexões acerca da capacidade da IA. A seguir, apresentamos a tabela resumo das aplicações. Todos os resumos gerados pela IA tinham qualidade suficiente para o uso na área acadêmica, sendo que os gerados pelo Humata foram os que apresentaram melhor consistência e coesão. A ferramenta Humata oferece apenas um resumo, com os parágrafos contínuos. No tocante aos resumos gerados pelo Resoomer, embora coerentes, foram apresentados por parágrafos e não em texto contínuo, em formato de tópicos para a cada parágrafo ou em passagens mais relevantes. Além disso, eram mais longos e, na maioria das vezes, com duas laudas em média. Com o Tome, os resumos gerados foram coerentes, mas também são gerados em tópicos e em formato de slides. Por final, foi possível verificar e comparar os resumos efetuados pelos bolsistas de iniciação científica. Enquanto dois textos tiveram como ponto forte a coesão, outros dois foram destaque na

coerência. Com este exercício de comparação, foi possível perceber que os resumos gerados por IA e os bolsistas de iniciação científica cumpriram com os requisitos. No entanto, apesar das potencialidades e vantagens com o uso de ferramentas de IA para assistência da escrita acadêmica, é importante ressaltar os limites. Como conclusão, distingue-se que a sabedoria está em utilizar de forma prudente as potencialidades e dirimir os limites.

2.3. Critérios de Avaliação

2.3.1. Métricas Quantitativas

- **Perplexidade:** Mede a previsibilidade do modelo (valores menores indicam melhor desempenho).
- **Cobertura de Palavras-chave:** Avalia a aderência ao vocabulário esperado.
- **Scores ROUGE:** Mede a similaridade textual com uma referência, incluindo ROUGE-1, ROUGE-2 e ROUGE-L.
- **BERTScore:** Avalia a similaridade semântica utilizando embeddings contextuais.

2.3.2. Métricas Qualitativas

- **Clareza e Coerência:** O resumo é bem estruturado e compreensível?
- **Precisão:** O modelo omitiu informações cruciais ou adicionou algo incorreto?
- **Consistência:** A resposta se mantém coerente quando o mesmo prompt é enviado mais de uma vez?

3. Resultados e Discussão

3.1. Resumos Gerados

Modelo: **gemma:7b**

A Inteligência Artificial (IA) é um campo da Ciência da Computação responsável por dispositivos computacionais que tentam simular o intelecto humano. A IA tem

aplicações em vários domínios, incluindo ensino, robótica e ciência da computação. No entanto, o uso inadequado da IA pode gerar problemas em áreas como ensino em EAD. O estudo analisou a qualidade dos resumos gerados pela IA e concluiu que os melhores resultados foram obtidos com o Humata, que forneceu resumos contínuos e coerentes. Apesar das vantagens da IA, é importante considerar os limites deste campo e usar-o de forma prudente.

Modelo: **llama3.2:3b**

A Inteligência Artificial (IA) é uma subárea da Ciência da Computação que busca simular o intelecto humano através de dispositivos computacionais. Na área acadêmica, a IA pode ajudar na melhoria da efetividade e produtividade. No entanto, existem desvantagens, como alta dependência das infraestruturas digitais e excessiva dependência digital. Além disso, a utilização inadequada da IA pode trazer limites, incluindo a perda de habilidades analíticas humanas. Um estudo foi realizado para avaliar a capacidade da IA em gerar resumos eficientes, comparando os resultados com aqueles produzidos por bolsistas de iniciação científica. Os resultados mostraram que as ferramentas de IA, como o Humata e o Tome, podem gerar resumos de qualidade, mas é importante utilizar essas ferramentas com prudência e sabedoria para evitar a perda de habilidades humanas.

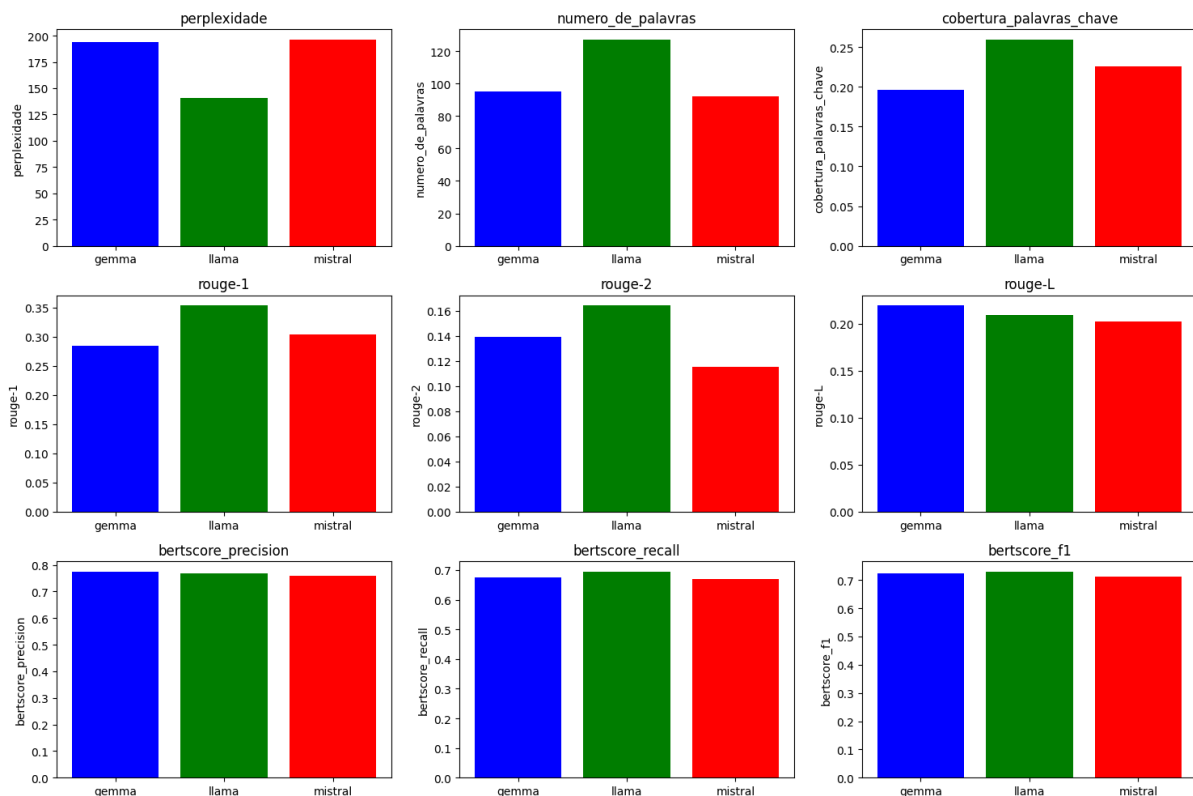
Modelo: **mistral v0.3**

O texto discute a Inteligência Artificial (IA) como área interdisciplinar que simula o intelecto humano e melhora a efetividade e produtividade em organizações. No entorno acadêmico, a IA é usada para gerar resumos de artigos acadêmicos, com qualidade suficiente para uso na área acadêmica. A ferramenta Humata oferece um resumo contínuo e com maior coesão, enquanto os resumos do Resoomer são coerentes mas apresentados em tópicos. O Tome também oferece resumos coerentes mas apresentados em formato de slides. Na conclusão, é importante utilizar a IA de forma prudente e ressaltar seus limites.

3.2. Desempenho Quantitativo

| Métrica | Gemma | LLaMA | Mistral | Melhor Desempenho |
|-----------------------------|--------------|---------------|---------|-------------------|
| Perplexidade | 193.76 | 141.05 | 196.17 | LLaMA |
| Cobertura de Palavras-chave | 0.197 | 0.259 | 0.226 | LLaMA |
| ROUGE-1 | 0.312 | 0.354 | 0.328 | LLaMA |
| ROUGE-2 | 0.145 | 0.164 | 0.152 | LLaMA |
| ROUGE-L | 0.219 | 0.209 | 0.202 | Gemma |
| BERTScore (Precisão) | 0.765 | 0.769 | 0.760 | LLaMA |
| BERTScore (F1) | 0.723 | 0.730 | 0.712 | LLaMA |

O modelo com melhor desempenho em cada métrica está destacado em **negrito**.



O **LLaMA** obteve a menor perplexidade (141.05), indicando maior previsibilidade. A cobertura de palavras-chave foi maior no **LLaMA** (0.259), demonstrando melhor aderência ao vocabulário. O **LLaMA** também apresentou os melhores valores para ROUGE-1 (0.354) e ROUGE-2 (0.164).

3.3. Avaliação Qualitativa

| Modelo | Clareza e Coerência | Precisão | Consistência |
|---------|---------------------|----------|--------------|
| Gemma | Mediana | Mediana | Baixa |
| LLaMA | Alta | Alta | Alta |
| Mistral | Mediana | Alta | Média |

O **LLaMA** gerou os resumos mais claros e coerentes.

O **Mistral** demonstrou alta precisão, mas sofreu com inconsistências.

O **Gemma** apresentou menor consistência quando o mesmo prompt foi repetido.

4. Conclusão

Os resultados indicam que o **LLaMA** tem o melhor desempenho geral, combinando baixa perplexidade, alta coerência textual, melhor cobertura de palavras-chave e maior precisão semântica. O **Mistral** teve um desempenho competitivo em precisão, enquanto o **Gemma** mostrou menor estabilidade e previsibilidade.