
**UNIVERSIDADE FEDERAL DE CAMPINA GRANDE**
**CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA**
**CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO**

**Carmem Izaura Germano Barbosa Neri**

# DESVENDANDO A POESIA COM IA: INFLUÊNCIA DO AJUSTE DE HIPERPARÂMETROS NO RAG PARA A COMPREENSÃO DE TEXTO POÉTICOS

**CAMPINA GRANDE - PB**

**2024**

**Carmem Izaura Germano Barbosa Neri**

# DESVENDANDO A POESIA COM IA: INFLUÊNCIA DO AJUSTE DE HIPERPARÂMETROS NO RAG PARA A COMPREENSÃO DE TEXTO POÉTICOS

**Trabalho de Conclusão de Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.**

**Orientador : Fabio Jorge Almeida Morais**

**CAMPINA GRANDE - PB**

**2024**

**Carmem Izaura Germano Barbosa Neri**

# DESVENDANDO A POESIA COM IA: INFLUÊNCIA DO AJUSTE DE HIPERPARÂMETROS NO RAG PARA A COMPREENSÃO DE TEXTO POÉTICOS

**Trabalho de Conclusão de Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.**

## BANCA EXAMINADORA:

**Fabio Jorge Almeida Morais**

**Orientador – UASC/CEEI/UFCG**

**Maxwell Guimarães de Oliveira**

**Examinador – UASC/CEEI/UFCG**

**Francisco Vilar Brasileiro**

**Professor da Disciplina TCC – UASC/CEEI/UFCG**

**Trabalho aprovado em: 15 de Maio de 2024.**

**CAMPINA GRANDE - PB**

# RESUMO

Este estudo explora a aplicação do sistema de Geração Aumentada por Recuperação (RAG) em textos poéticos, concentrando-se na customização de seus hiperparâmetros para otimizar a compreensão e geração textual em um gênero literário que desafia pela sua densidade semântica e estrutural. No contexto dos Grandes Modelos de Linguagem (LLMs), o RAG se apresenta como uma ferramenta valiosa para superar limitações de conhecimento fixo, integrando dinamicamente informações atualizadas de fontes externas. Este trabalho emprega uma metodologia quantitativa para avaliar a eficácia do RAG, utilizando métrica Correção (Correctness) para medir o desempenho e análises manuais para refinar os resultados obtidos automaticamente. Ao modificar hiperparâmetros como o chunk size, chunk overlap e modelo de geração, o estudo busca determinar a configuração ideal para a geração de respostas precisas e relevantes para perguntas sobre poesia. As descobertas revelam que ajustes precisos nesses parâmetros influenciam na qualidade da informação recuperada e das respostas geradas, destacando a capacidade do RAG de produzir respostas enriquecidas e contextualmente apropriadas.

# UNRAVELING POETRY WITH AI: INFLUENCE OF HYPERPARAMETER TUNING IN RAG FOR THE UNDERSTANDING OF POETIC TEXTS

## ABSTRACT

This study explores the application of the Retrieval-Augmented Generation (RAG) system to poetic texts, focusing on the customization of its hyperparameters to optimize understanding and textual generation in a literary genre that is challenging due to its semantic and structural density. Within the context of Large Language Models (LLMs), RAG presents itself as a valuable tool to overcome the limitations of fixed knowledge, dynamically integrating updated information from external sources. This work employs a quantitative methodology to evaluate the effectiveness of RAG, using the Correctness metric to measure performance and manual analyses to refine the results obtained automatically. By modifying hyperparameters such as chunk size, chunk overlap, and generation model, the study aims to determine the ideal configuration for generating precise and relevant responses to questions about poetry. The findings reveal that precise adjustments to these parameters influence the quality of the information retrieved and the responses generated, highlighting RAG's ability to produce enriched and contextually appropriate answers.

# Desvendando a Poesia com IA: Influência do Ajuste de Hiperparâmetros no RAG para a Compreensão de Textos Poéticos

Carmem Izaura Germano Barbosa Neri
carmem.neri@ccc.ufcg.edu.br
Universidade Federal de Campina Grande
Campina Grande, Paraíba, Brasil

Fabio Jorge Almeida Morais
(Orientador)
fabio@computacao.ufcg.edu.br
Universidade Federal de Campina Grande
Campina Grande, Paraíba, Brasil

## RESUMO

Este estudo explora a aplicação do sistema de Geração Aumentada por Recuperação (RAG) em textos poéticos, concentrando-se na customização de seus hiperparâmetros para otimizar a compreensão e geração textual em um gênero literário que desafia pela sua densidade semântica e estrutural. No contexto dos Grandes Modelos de Linguagem (LLMs), o RAG se apresenta como uma ferramenta valiosa para superar limitações de conhecimento fixo, integrando dinamicamente informações atualizadas de fontes externas. Este trabalho emprega uma metodologia quantitativa para avaliar a eficácia do RAG, utilizando métrica Correção (*Correctness*) para medir o desempenho e análises manuais para refinar os resultados obtidos automaticamente. Ao modificar hiperparâmetros como o *chunk size*, *chunk overlap* e modelo de geração, o estudo busca determinar a configuração ideal para a geração de respostas precisas e relevantes para perguntas sobre poesia. As descobertas revelam que ajustes precisos nesses parâmetros influenciam na qualidade da informação recuperada e das respostas geradas, destacando a capacidade do RAG de produzir respostas enriquecidas e contextualmente apropriadas.

## PALAVRAS CHAVE

Geração Aumentada por Recuperação, RAG, Grandes Modelos de Linguagem, Poesia, Hiperparâmetros, Inteligência Artificial.

## 1. INTRODUÇÃO

No campo em constante evolução da Inteligência Artificial, os Grandes Modelos de Linguagem (LLMs) se destacam por sua capacidade de processar e gerar linguagem natural, empregando bilhões de parâmetros e vastos volumes de dados [1]. Esses modelos demonstraram proficiência em uma ampla gama de tarefas, desde a geração de texto até a resposta a consultas complexas. Contudo, apesar de seus avanços notáveis, os LLMs enfrentam limitações significativas, sobretudo no que tange à incorporação de conhecimentos que transcendem suas extensas, porém finitas, bases de treinamento.

É neste contexto que o conceito de *Retrieval-Augmented Generation* (RAG), ou Geração Aumentada por Recuperação, emerge como uma solução promissora, visando superar tais limitações ao integrar de forma dinâmica os LLMs com bases de conhecimento externas e atualizáveis. Lewis et al. [2] destacam a natureza interativa dos sistemas RAG, enfatizando sua capacidade de gerar respostas não apenas precisas, mas também atualizadas e livres de distorções comuns, como alucinações informativas [3, 4, 5, 6] e lacunas de conhecimento [7] específico do domínio.

No entanto, apesar do crescente interesse e aplicabilidade dos sistemas baseados em RAG no âmbito corporativo e de processamento de informações, a exploração de seu potencial na extração de informações de textos literários, particularmente na poesia, permanece inexplorada. Este estudo se propõe a investigar como a técnica de RAG pode ser adaptada ao gênero poético, reconhecendo os desafios únicos impostos pela densidade semântica, riqueza de detalhes e diversidade de estruturas características deste domínio.

A pesquisa examinará a eficácia da integração de LLMs com bases de conhecimento externas para extração de informações relevantes desses textos, e também como as alterações dos hiperparâmetros no *pipeline* do RAG podem influenciar nos resultados obtidos. Para avaliar o desempenho da implementação do RAG na extração de informações de poesia, será utilizada a métrica quantitativa de Correção (*Correctness*), além de avaliações manuais para complementar a métrica automatizada, superando suas limitações na captura das nuances da expressão poética.

A Seção 2 apresenta a fundamentação, a Seção 3 aborda a metodologia utilizada nessa análise, a Seção 4 apresenta e discute os resultados obtidos e a Seção 5, por fim, resume as contribuições deste trabalho.

## 2. FUNDAMENTAÇÃO

### 2.1 Grandes Modelos de Linguagem (LLMs)

Grandes Modelos de Linguagem (LLMs) são sistemas avançados de inteligência artificial que utilizam a arquitetura *Transformer* para processar e gerar linguagem natural de forma eficiente. Esses modelos são conhecidos por serem pré-treinados em extensos conjuntos de dados textuais, aprendendo padrões linguísticos, gramática e nuances contextuais. Por sua habilidade em compreender e manipular textos complexos, os LLMs são empregados em uma diversidade de tarefas de processamento de linguagem natural (NLP), como tradução, geração de texto e sumarização de texto.

Dentro da família GPT (*Generative Pre-trained Transformer*), o GPT-3.5 Turbo e o GPT-4 são modelos desenvolvidos e comercializados pela OpenAI. A versão GPT-3.5 Turbo, uma evolução do GPT-3, aprimora a eficiência e a capacidade de generalização do modelo em diversas tarefas de NLP, utilizando 175 bilhões de parâmetros. Este modelo se destaca pelo aprendizado com poucos exemplos e a habilidade de executar uma ampla variedade de tarefas sem a necessidade de ajustes finos [8]. O GPT-4, por sua vez, expande as capacidades do GPT-3.5 Turbo com mais parâmetros e um treinamento mais otimizado, aprimorando a precisão e a coerência do texto gerado em tarefas complexas. É capaz de alcançar desempenhos que, em alguns casos, rivalizam com o humano, marcando um avanço substancial no campo da inteligência artificial [9].

Apesar de tais avanços, ambas as versões enfrentam limitações, como tendências a produzir "alucinações" factuais e dificuldades em manter a coerência em textos extensos. Essas limitações destacam a necessidade de pesquisa contínua para o desenvolvimento de LLMs que sejam não apenas eficientes e poderosos, mas também éticos e confiáveis.

## 2.2 Retrieval-Augmented Generation (RAG)

O RAG, ou *Retrieval-Augmented Generation*, representa uma inovação no campo da Inteligência Artificial, oferecendo uma maneira de aprimorar a capacidade dos Modelos de Linguagem Grandes (LLMs) para gerar respostas mais precisas e informadas [10]. Este método combina o conhecimento intrínseco dos LLMs, adquirido durante o treinamento em vastos conjuntos de dados, com bases de informações atualizadas.

No núcleo do RAG está o processo de enriquecimento do contexto. Tradicionalmente, os LLMs dependem da entrada do usuário e de seu conhecimento prévio para fornecer respostas. Contudo, isso pode resultar em respostas desatualizadas ou genéricas, principalmente quando o modelo é questionado com perguntas fora do escopo de seu treinamento original. O RAG aborda esse desafio introduzindo um passo intermediário de recuperação de informações, no qual a consulta do usuário desencadeia a busca por dados relevantes de novas fontes, como bases de dados ou APIs atualizadas. Essa abordagem extrai informações que não estavam disponíveis durante o treinamento do LLM, mas são essenciais para responder às consultas atuais de maneira correta e confiável.

Com a integração do RAG, a entrada do usuário é primeiramente processada por um componente de recuperação que busca e extrai informações relevantes de fontes externas. Essas informações são então convertidas em representações vetoriais — ou *embeddings* — que o LLM pode processar. Esta incorporação de dados adicionais, chamada de engenharia de *prompts*, permite que o modelo aprimore sua resposta gerada, combinando seu vasto conhecimento pré-treinado com as informações contextuais atualizadas [10]. Todo o processo do sistema RAG está ilustrado na Figura 1.

Além disso, o RAG traz a capacidade de atualizar continuamente a base de conhecimento do modelo. Conforme novas informações se tornam disponíveis ou as existentes são atualizadas, o índice de documentos do RAG pode ser ajustado para refletir essas mudanças, garantindo que o modelo mantenha sua relevância e precisão ao longo do tempo. Esse dinamismo é vital, principalmente em domínios onde os dados mudam rapidamente, como notícias, médico ou legal.

O uso do RAG é especialmente benéfico em cenários onde a precisão das informações é crítica. Por exemplo, no domínio de atendimento ao cliente, o RAG pode fornecer respostas que refletem as políticas e informações mais recentes da empresa, evitando respostas desatualizadas que poderiam gerar algum transtorno. Ao assegurar que o modelo esteja sincronizado com os dados mais recentes e relevantes, o RAG eleva a confiabilidade das interações com o usuário e permite uma experiência mais confiável.

Em suma, o RAG representa uma abordagem robusta para a extensão do uso de LLMs, permitindo que as organizações tirem proveito de modelos de linguagem avançados enquanto mantêm controle sobre a qualidade e a atualidade das respostas fornecidas. Este sistema não só melhora a qualidade das interações do usuário com o modelo, mas também mitiga os riscos associados à desinformação, aos dados desatualizados e às "alucinações" geradas pelas LLMs, o que é crucial em muitas aplicações práticas de inteligência artificial.

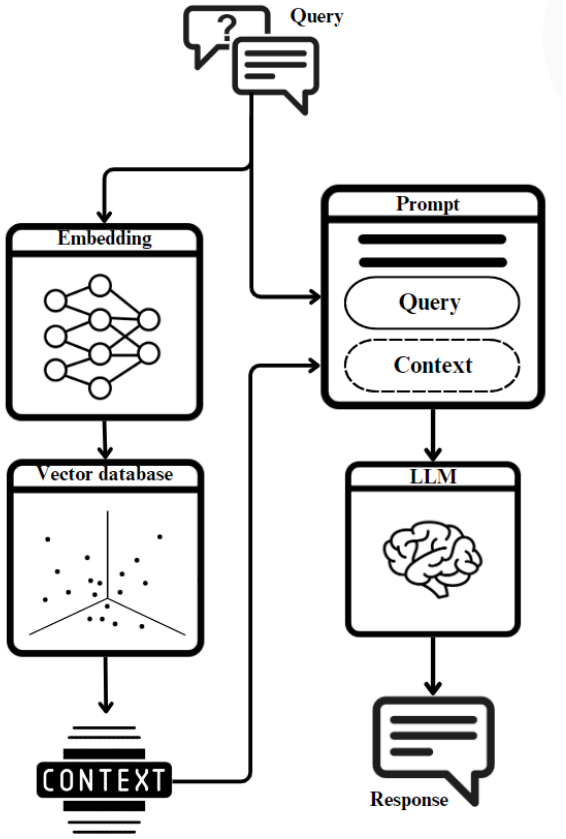


**Figura 1: Pipeline Retrieval-Augmented Generation (RAG).**

## 2.3 Hiperparâmetros

Hiperparâmetros são configurações que podem ser ajustadas para controlar o comportamento de algoritmos de aprendizado de máquina. Ao contrário dos parâmetros do modelo, que são aprendidos durante o treinamento, os hiperparâmetros são definidos antes do início do processo e têm um impacto

significativo no desempenho do modelo final. No contexto do *Retrieval-Augmented Generation* (RAG), os hiperparâmetros desempenham um papel fundamental na otimização e no desempenho do sistema, conforme descrito na literatura [11]. Hiperparâmetros, como o *chunk size* (tamanho de cada partição do contexto), *chunk overlap* (sobreposição feita entre cada partição de texto) e modelo de *embedding* (transforma textos em uma forma numérica, conhecida como vetores), são ajustáveis e influenciam significativamente a eficácia do processo de recuperação e geração de texto.

O *chunk size* é um desses hiperparâmetros críticos, pois determina a granularidade das informações que o sistema pode acessar durante a recuperação. Um tamanho de *chunk size* adequado é crucial para garantir que informações detalhadas não sejam perdidas, ao mesmo tempo em que se preserva a contextualização necessária para a geração de respostas coerentes. Já o *chunk overlap* garante que não haja perda de coerência semântica entre os *chunks* consecutivos, possibilitando uma transição suave de informações e mantendo a integridade do conteúdo recuperado.

Outro componente essencial é o modelo de *embedding*, o qual desempenha um papel crucial na determinação da qualidade da correspondência entre as consultas e os documentos armazenados, impactando diretamente a qualidade da informação recuperada.

Portanto, a escolha e o ajuste cuidadoso desses hiperparâmetros são essenciais para adaptar o sistema do RAG às necessidades específicas do projeto, maximizando assim a eficiência e a eficácia do sistema dentro do contexto proposto.

# 3. METODOLOGIA

Para uma análise dos resultados alcançados com a aplicação do sistema RAG ao contexto singular de textos poéticos em português, propôs-se uma abordagem metodológica composta por etapas estruturadas que possibilitaram uma quantificação e uma exploração dos dados. Esse estudo configura-se como uma pesquisa quantitativa, o que permitiu empregar técnicas estatísticas para analisar como a ampliação do contexto pelo RAG pode enriquecer as respostas do LLM a perguntas específicas sobre poesia e suas nuances.

A precisão dessas respostas é quantificada mediante métrica de desempenho selecionada, *Correctness* (Correção), que avalia não somente a coerência e a relevância da informação gerada, mas também a capacidade do modelo de se adaptar a um domínio literário complexo. Este processo é complementado pelo ajuste cuidadoso dos hiperparâmetros e pela subsequente validação do modelo com conjuntos de dados representativos, garantindo que as conclusões derivadas sejam embasadas em dados e que os ajustes realizados no sistema RAG reflitam melhorias reais na funcionalidade do LLM.

## 3.1 Design de Experimento

O gênero literário escolhido para este estudo foi a poesia, por representar uma inovação, dada sua complexidade e as nuances que apresenta, configurando um desafio maior de interpretação quando comparado a outros gêneros literários. Para explorar as capacidades do RAG neste contexto, optou-se por experimentar ajustes nos hiperparâmetros do sistema, incluindo o *chunk size* e *chunk overlap*. Além disso, foram utilizadas diferentes versões do modelo de geração de respostas, o GPT-3.5 Turbo e o GPT-4, permitindo a análise sob diferentes configurações para determinar qual melhor se adapta ao estudo proposto. A métrica de desempenho, juntamente com uma análise manual das respostas a um conjunto de perguntas cuidadosamente elaboradas, servirão para avaliar qual configuração produz as respostas mais precisas e informativas. Essas perguntas foram desenvolvidas a partir dos dados selecionados e em parceria com o autor dos textos, trazendo ainda mais precisão para a metodologia de avaliação. Também foram especificados todos os *ground truth* (resposta alvo), para cada uma das perguntas, fazendo com que a avaliação possa ser ainda mais minuciosa.

## 3.2 Dados

Os dados para este estudo foram fornecidos pelo autor paraibano Robson Junior e consistem em três livros de poesia [12, 13, 14] que abordam diferentes temáticas, contabilizando 100 poemas ao todo. Estes textos foram disponibilizados em formato PDF e foram utilizados para fornecer o contexto necessário ao LLM durante a etapa de geração de respostas. O conteúdo desses livros, rico em expressões poéticas e variações temáticas, oferece uma base robusta para testar a eficácia da técnica RAG que auxiliará na criação de um novo contexto para o LLM que irá gerar respostas e lidar com a linguagem complexa e carregada de subjetividade presente nesses textos.

O autor elaborou 25 perguntas e suas respectivas respostas a respeito dos poemas, todas foram utilizadas na etapa de teste de avaliação, disponíveis no Anexo A.

## 3.3 Aplicação de RAG

O LlamaIndex [15] é uma infraestrutura avançada projetada para implementar aplicações RAG que utilizam LLMs. Essa ferramenta proporciona a capacidade de injetar dados privados ou específicos de domínio de forma segura e eficaz.

No contexto deste estudo, o LlamaIndex desempenha um papel crucial ao facilitar a manipulação e a integração de dados a partir de textos poéticos. A escolha deste *framework*, que é gratuito e *open-source*, não apenas promove uma reprodutibilidade para trabalhos futuros, mas também oferece uma solução robusta para o desafio de trabalhar com dados em formatos menos estruturados, como nos PDFs disponibilizados pelo autor.

## 3.4 Modelo para *Embedding*

O modelo de *embedding* BAAI/bge-m3, desenvolvido pela *Beijing Academy of Artificial Intelligence* [16], foi selecionado para este estudo devido a suas características de leveza e eficiência, que exigem menos recursos computacionais, além de ser um modelo multilingue, o que auxilia no nosso estudo por estarmos tratando com uma base que contém textos em português.

A acessibilidade desse modelo, sendo um recurso gratuito, facilita a replicação e expansão do estudo, promovendo o avanço na pesquisa. Em resumo, o uso do BAAI não apenas aprimora a metodologia, mas também auxilia para que as respostas geradas sejam relevantes e alinhadas com as intenções dos textos analisados.

## 3.5 Ajuste dos Hiperparâmetros

No contexto deste estudo, exploramos os tamanhos de *chunk size* de 128 (padrão) e 256, e percentual de *chunk overlap* de 0% (padrão), 30% e 70%, para entender como esses hiperparâmetros influenciam o desempenho do sistema RAG na etapa de *Retrieval-Augmented* (Recuperação Aumentada). Conforme

destacado em estudos anteriores [10], tamanhos maiores de *chunk* podem diminuir as métricas referentes à qualidade das respostas e o aumento da sobreposição melhora significativamente as métricas, particularmente em perguntas que dependem de uma única fonte de documento, assim como no estudo em questão. Esses valores, ilustrados na Tabela 1, foram escolhidos para avaliar se as mesmas descobertas se aplicam no contexto de textos poéticos.

| **Parâmetro** | **Valores Testados** |
|---|---|
| *Chunk Size* | 128 e 256 |
| *Chunk Overlap* | 0%, 30% e 70% |

**Tabela 1: Os parâmetros e seus valores testados.**

## 3.6 Etapa de Geração das Respostas

Na etapa de geração das respostas do estudo, o foco foi analisar e comparar o desempenho de duas versões distintas do modelo LLM conhecido como ChatGPT, desenvolvido pela OpenAI. As versões selecionadas para esta análise foram o GPT-3.5 Turbo e o ChatGPT-4, escolhidas devido às suas capacidades avançadas e às melhorias incrementais entre as gerações.

O processo começa com a entrada do usuário, que será uma pergunta relacionada a um texto poético específico que faz parte da base fornecida como contexto. Essa entrada é processada pelo sistema RAG, que utiliza o modelo de *embedding* BAAI/bge-m3 para recuperar conteúdo relevante dos textos poéticos. Em seguida, o conteúdo recuperado é fornecido ao modelo de geração (ChatGPT), juntamente com a pergunta realizada. A partir desse conjunto pergunta+contexto, o modelo irá gerar uma resposta que melhor se adeque à situação.

Esse processo foi realizado separadamente para cada uma das versões do ChatGPT utilizadas no estudo, o que permitiu uma comparação das respostas geradas. Essa análise permitiu identificar não apenas qual modelo é mais eficaz, mas também se as atualizações no modelo mais recente justificam sua adoção em futuras aplicações do RAG para tarefas similares.

## 3.7 Etapas de Avaliação

Para avaliar as respostas geradas, foram utilizadas duas abordagens: avaliação automática e análise manual, realizada em colaboração com o autor dos poemas.

A métrica escolhida para a avaliação dos resultados foi a *Correctness*, pois dentre as sete métricas disponibilizadas pela biblioteca essa foi a que mais se adequou às necessidades do estudo, por receber como parâmetros a pergunta realizada (*query*), a resposta gerada pelo modelo (*response*) e a resposta alvo (*ground truth*). Como todas as perguntas escolhidas para realização dos testes neste estudo foram respondidas pelo autor, tivemos um excelente conjunto de respostas alvo para fornecer como parâmetro.

O *Correctness* retorna uma nota para as respostas, em uma escala de 1 a 5, além de uma explicação sobre a atribuição da nota. A partir dos resultados obtidos, foi possível realizar uma seleção das configurações que foram avaliadas manualmente em uma nova etapa.

Para as configurações selecionadas, a etapa de análise manual, realizada em colaboração com o autor, permitiu avaliar tanto as respostas dadas pelo sistema quanto a nota e feedback gerados pela métrica de *Correctness*, retribuindo uma nota para as respostas e classificando a qualidade e coesão da métrica anterior.

# 4. RESULTADOS

Após a aplicação das 25 perguntas de teste às 12 combinações de configurações possíveis, realizou-se o processo de avaliação automática, seguido de uma análise manual. Os procedimentos e descobertas pertinentes são expostos nas seções subsequentes do estudo. Essa abordagem revelou nuances significativas no desempenho das diferentes configurações e na precisão das respostas geradas, evidenciando a complexidade e os desafios associados à avaliação do sistema RAG, tanto para sua etapa de recuperação quanto para etapa de geração em contextos de texto poético.

## 4.1 Avaliação Automática

Na Figura 2 podemos observar o desempenho geral das configurações para cada uma das perguntas através do gráfico de *Heatmap*, destacando que a nota mínima, um *score* de 1, foi uma ocorrência rara, aparecendo apenas três vezes, o que pode indicar casos específicos onde o modelo falhou em entender o contexto ou as peculiaridades da pergunta. A pergunta 18 destaca-se pela consistência, com todas as configurações recebendo uma nota de 4,5, o que sugere uma forte congruência entre o que foi recuperado como contexto, o que o modelo gerou e a resposta esperada. Já a pergunta 22 apresenta um cenário de desafio notável, refletido pelas notas mais baixas em todas as configurações, o que implica em uma dificuldade intrínseca da pergunta ou uma complexidade textual que desafiou o sistema RAG de maneira uniforme.

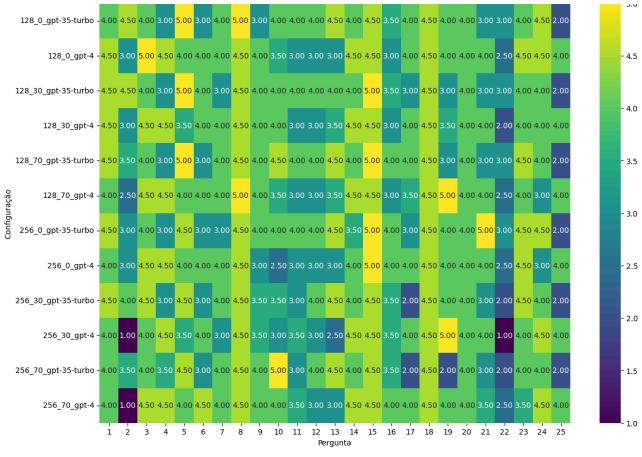


**Figura 2: *Heatmap* com notas para cada pergunta por configuração.**

A partir dos resultados da métrica *Correctness* foi possível compreender o desempenho de cada configuração por meio da média das notas atribuídas para cada uma das 25 perguntas, como pode ser observado na Figura 3. A diferença entre as médias das notas da melhor e da pior configuração é relativamente pequena, com uma margem de apenas 0,28 pontos, indicando uma proximidade nos resultados alcançados pelas diferentes configurações testadas. Como pode ser observado na Tabela 2, temos quatro diferentes configurações que se destacaram como os

três melhores resultados. Essas configurações foram utilizadas posteriormente no processo de avaliação manual.

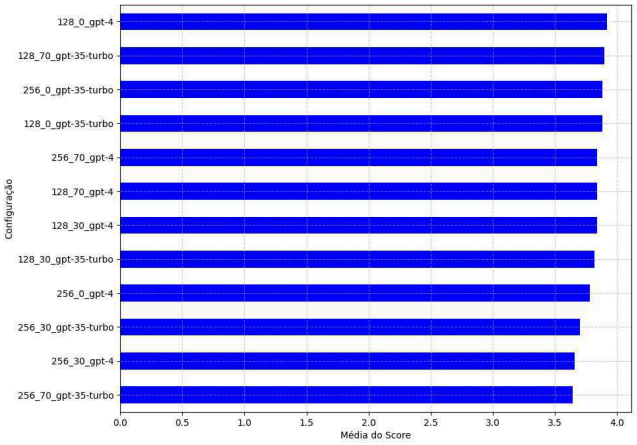


**Figura 3: Média das notas de cada configuração.**

| *Chunk Size* | *Overlap* | Modelo | Nota |
|---|---|---|---|
| 128 | 0 | ChatGPT-4 | 3,92 |
| 128 | 70 | ChatGPT-3.5 Turbo | 3,9 |
| 256 | 0 | ChatGPT-3.5 Turbo | 3,88 |
| 128 | 0 | ChatGPT-3.5 Turbo | 3,88 |

**Tabela 2: Configurações que obtiveram os melhores resultados.**

O *Boxplot* apresentado na Figura 4 confirma que a configuração com *chunk size* de 128 e *overlap* de 0% utilizando o modelo ChatGPT-4 não só apresenta a média mais alta de notas, mas também indica consistência, com uma variação de notas menos dispersa. Isso sugere que, para textos poéticos, essa configuração específica pode oferecer um ponto de equilíbrio ideal, maximizando a eficiência sem comprometer a qualidade das respostas.

Por outro lado, as configurações com *overlap* de 70%, independente do tamanho do *chunk*, exibem uma dispersão maior de resultados, com *outliers* (valores atípicos) sugerindo a ocorrência de respostas excepcionalmente boas ou ruins. Isso pode indicar que um *overlap* maior é uma "faca de dois gumes": pode enriquecer o contexto para algumas perguntas, enquanto introduz ruído em outras, possivelmente devido à natureza multifacetada e aberta à interpretação dos poemas.

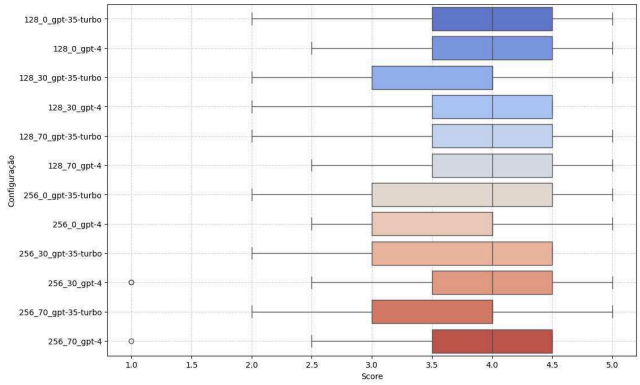


**Figura 4: Variação das notas em cada configuração de modelo.**

Além disso, o desempenho mais consistente das configurações com o modelo ChatGPT-3.5 Turbo, apesar de ligeiramente inferior em média, chama atenção para a capacidade deste modelo de fornecer respostas confiáveis. A semelhança nas médias das notas, particularmente nas configurações com *chunk size* de 256 e *overlap* de 0%, revela que ajustes no tamanho do *chunk* não afetam drasticamente o desempenho, o que pode ser útil em aplicações onde a estabilidade é preferida sobre a maximização de notas. Outro fator muito importante é que o modelo ChatGPT-3.5 Turbo possui um custo muito menor, o que não justificaria a utilização da versão mais avançada do modelo, tendo em vista que os resultados não demonstram um ganho tão significante no desempenho.

## 4.2 Avaliação Manual

Após a avaliação automática com a métrica de *Correctness*, prosseguiu-se com a avaliação manual das quatro configurações que se sobressaíram, examinando suas notas e os comentários fornecidos pela métrica. Durante esta etapa, a qualidade de cada resposta foi classificada em uma escala que inclui "RUIM", "MEDIANO", "BOM", "EXCELENTE" ou "SEM CONTEXTO" — este último aplicado nos casos em que o modelo não encontrou contexto suficiente para gerar uma resposta. Da mesma forma, a qualidade dos comentários justificativos da nota foi avaliada, usando os mesmos critérios, porém adicionando à escala o indicador de "AUSENTE", pois foi observada a ausência de comentários em 28 casos. Com base nesta análise detalhada, uma nova nota foi atribuída manualmente para cada resposta gerada, visando proporcionar uma medida mais precisa do desempenho do sistema RAG.

Podemos observar na Tabela 3 que as quatro configurações que se sobressaíram na avaliação automática apresentaram resultados que confirmam, na maioria, as tendências observadas inicialmente. A configuração com *chunk size* de 128, sem *overlap* e utilizando o modelo ChatGPT-4 permaneceu como a mais alta, embora com uma média ligeiramente inferior, de 3,34, comparada à média obtida automaticamente. O mesmo aconteceu para as demais configurações — elas mantiveram suas posições, sofrendo apenas uma pequena penalização na média dos resultados.

| *Chunk Size* | *Overlap* | Modelo | Nota |
|---|---|---|---|
| 128 | 0 | ChatGPT-4 | 3,34 |
| 128 | 70 | ChatGPT-3.5 Turbo | 3,28 |
| 256 | 0 | ChatGPT-3.5 Turbo | 3,22 |
| 128 | 0 | ChatGPT-3.5 Turbo | 3,18 |

**Tabela 3: Médias das avaliações manuais das melhores configurações.**

Durante o processo de avaliação das configurações destacadas, identificou-se que a métrica de *Correctness* poderia penalizar respostas corretas por serem menos concisas devido à inclusão de informações extras. Por exemplo, a resposta alvo para a pergunta 4, "Quais as duas maiores descobertas do homem?", é "O fogo e a combustão"; no entanto, a resposta gerada se deu da seguinte forma, por praticamente todas as configurações: "As duas maiores descobertas do homem, segundo o poema, são o fogo e a combustão". Esse resultado foi considerado menos conciso por adicionar mais informações, apesar de estar correto. Tal fenômeno foi observado para perguntas variadas, onde adições similares ou estruturas frasais ligeiramente diferentes da que foi passada como a resposta alvo penalizaram as notas, revelando uma inconsistência na avaliação automática.

Além disso, notou-se uma tendência do modelo a seguir um viés mais descritivo e menos interpretativo em algumas das perguntas. Por exemplo, para a pergunta 9, "Como a ideia de ciclo é sugerida pelo poema Ciclo?", o autor fornece como resposta esperada: "O poema é iniciado e finalizado com o mesmo verso", algo que não foi capturado pelo modelo, mesmo que o contexto correto tenha sido recuperado.

Mesmo quando a avaliação manual identificou respostas idênticas em diferentes configurações, as notas e comentários variaram, apontando para uma falta de padronização na métrica. Casos como a pergunta 13, "Qual a implicação do amor ser cristalino?", ilustram essa discrepância, onde uma palavra extra não alterou o significado da resposta, mas impactou a nota em diferentes configurações.

A qualidade das justificativas fornecidas para as notas foi classificada manualmente como "BOM" ou "EXCELENTE" em sua maioria, como pode ser observado na Figura 5. No entanto, em alguns casos, mesmo com uma justificativa com tom favorável, a nota foi penalizada de maneira desencontrada com a justificativa. A existência de comentários classificados como "AUSENTE" aponta para a necessidade de refinamento da avaliação.

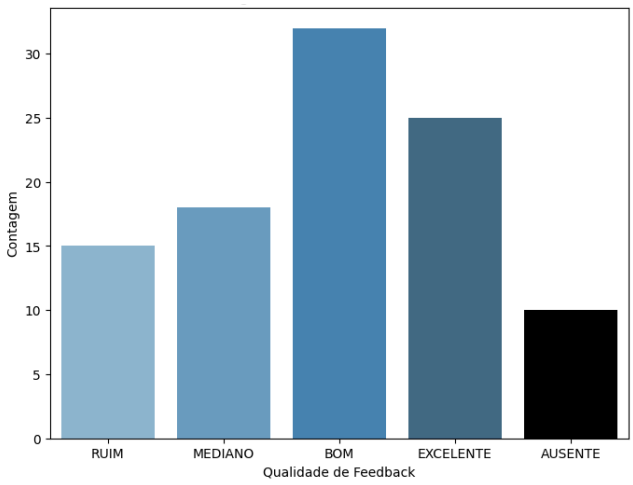


**Figura 5: Avaliação dos comentários fornecidos pela métrica.**

Após a avaliação manual, o autor foi consultado para fazer suas próprias considerações a respeito dos resultados obtidos. Ele destacou que os modelos frequentemente proporcionaram respostas interessantes, muitas delas alinhadas ou até superiores às expectativas, especialmente para perguntas que exigiam uma interpretação mais profunda dos textos. Porém, em alguns casos, vieses de informações do senso comum influenciaram as respostas, principalmente nas perguntas 2 e 25. Por outro lado, exemplos como a pergunta 3 demonstraram a habilidade do modelo ignorar esses vieses e concentrar-se exclusivamente no conteúdo poético recuperado em determinados casos.

O autor apontou que também identificou algumas limitações mais significativas em respostas a poemas com linguagem altamente subjetiva ou uso de ironia. Apesar disso, houve momentos em que o modelo captou sutilezas ou interpretou corretamente aspectos indiretos dos textos, como mostrado nas perguntas 1, 5, 10 e 24.

Finalmente, ele reconheceu que suas expectativas de respostas diretas influenciaram a avaliação automática, que penalizou respostas mais descritivas, embora muitas dessas respostas fossem detalhadas e adequadas. Essas reflexões ressaltam a complexidade de avaliar a capacidade dos modelos em contextos poéticos e a necessidade de alinhar as métricas de avaliação com as nuances dos textos.

# 5. CONCLUSÃO

A investigação conduzida neste estudo destacou a versatilidade e os desafios inerentes à aplicação de sistemas *Retrieval-Augmented Generation* (RAG) ao domínio poético. Observou-se que a precisão na geração de respostas está ligada à escolha e à configuração dos hiperparâmetros, sendo o tamanho de *chunk size* e *overlap* componentes importantes para otimizar o desempenho da recuperação de contexto que, por sua vez, será repassado para o modelo de geração. Sem o contexto adequado, as respostas ficam vagas e genéricas, ou simplesmente as perguntas não são respondidas.

A análise evidenciou que uma configuração específica com *chunk size* de 128 e *overlap* de 0% utilizando o modelo ChatGPT-4 se destacou na média de qualidade das respostas. No entanto, também ressaltou a importância de um ajuste fino nos hiperparâmetros, tendo em vista que o aumento do *overlap* para 70% resultou em uma dispersão mais significativa dos resultados,

sugerindo que alterações nesse hiperparâmetro podem potencializar ou prejudicar o desempenho dependendo da complexidade do texto e da pergunta proposta.

Em relação ao modelo de geração das respostas, o ChatGPT-3.5 Turbo demonstrou ser uma alternativa confiável e mais acessível, tendo grande similaridade nas médias das notas quando comparado com o GPT-4, indicando que a utilização de uma versão mais avançada do modelo pode não ser justificada pelo custo adicional, exceto em situações onde o aumento marginal de qualidade é crítico.

# 6. AGRADECIMENTOS

Gostaria de iniciar expressando minha mais profunda gratidão à minha família pelo suporte constante e pelo incentivo que recebi ao longo da minha vida, especialmente durante esta fase que estou concluindo agora. Um agradecimento especial à minha mãe, cuja força e determinação são exemplos para mim. Mãe, sua presença nos momentos mais desafiadores foi fundamental para que eu não desistisse dos meus sonhos.

Agradeço ao apoio e incentivo de todos os meus amigos, que compartilharam comigo momentos importantes durante o curso e sempre me lembraram do meu potencial. Em especial, gostaria de agradecer a Eduarda Duarte, Eduarda Azevedo, Rodrigo Eloy e Leandra Oliveira, vocês foram essenciais, não apenas no contexto acadêmico, mas na minha vida, muito obrigada por tudo.

Aproveito para expressar minha gratidão ao meu amigo escritor, Robson Junior, por disponibilizar seus poemas inspiradores para este estudo. Seus versos não só enriqueceram minha pesquisa, mas também tocaram profundamente todos que tiveram o privilégio de lê-los. Espero, de coração, que você alcance o estrelato literário que tanto merece. Que seus poemas continuem a inspirar e encantar o mundo.

Não poderia deixar de expressar minha profunda gratidão ao meu amigo e namorado, Matheus Maia. Você esteve ao meu lado durante esta longa e desafiadora jornada, trazendo leveza e alegria. Sua paciência, compreensão e amor foram essenciais para manter meu ânimo e foco. Obrigada por cada palavra de encorajamento, por cada gesto de apoio e por acreditar em mim até mesmo quando eu mesma duvidava. Sua presença transformou cada obstáculo em uma aventura compartilhada.

Estendo meus agradecimentos a todos os professores do curso de Ciência da Computação, em especial ao professor e meu orientador Fabio Morais, que sempre foi muito presente e disponível para todo tipo de auxílio que eu precisasse, e ao professor Leandro Balby, por despertarem meu interesse em processamento de linguagem natural e ciência de dados com suas aulas inspiradoras. Também quero agradecer ao meu colega de projeto, Leonardo Silva, que foi essencial ao esclarecer dúvidas durante o desenvolvimento deste trabalho.

Por último, mas não menos importante, quero agradecer a mim mesma por persistir e superar os obstáculos que encontrei pelo caminho. Houve momentos em que duvidei, mas aqui estou, grata e realizada. Muito obrigada a todos que fizeram parte desta jornada!

# 7. REFERÊNCIAS

[1] Minaee, Shervin, et al. "Large language models: A survey." arXiv preprint arXiv:2402.06196 (2024).

[2] Lewis, Patrick, et al. "Retrieval-augmented generation for knowledge-intensive nlp tasks." Advances in Neural Information Processing Systems 33 (2020): 9459-9474.

[3] Garbiel Bénédict, Ruqing Zhang, and Donald Metzler. 2023. Gen-ir@ sigir 2023: The first workshop on generative information retrieval. In Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. 3460–3463.

[4] Yuyan Chen, Qiang Fu, Yichen Yuan, Zhihao Wen, Ge Fan, Dayiheng Liu, Dongmei Zhang, Zhixu Li, and Yanghua Xiao. 2023. Hallucination detection: Robustly discerning reliable answers in large language models. In Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. 245–255.

[5] Yuanjie Lyu, Chen Zhu, Tong Xu, Zikai Yin, and Enhong Chen. 2022. Faithful Abstractive Summarization via Factaware Consistency-constrained Transformer. In Proceedings of the 31st ACM International Conference on Information & Knowledge Management. 1410–1419.

[6] Guido Zuccon, Bevan Koopman, and Razia Shaik. 2023. Chatgpt hallucinates when attributing answers. In Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region. 46–51.

[7] Hangfeng He, Hongming Zhang, and Dan Roth. 2022. Rethinking with retrieval: Faithful large language model inference. arXiv preprint arXiv:2301.00303 (2022).

[8] Brown, Tom, et al. "Language models are few-shot learners." Advances in neural information processing systems 33 (2020): 1877-1901.

[9] Achiam, Josh, et al. "Gpt-4 technical report." arXiv preprint arXiv:2303.08774 (2023).

[10] Lewis, Patrick, et al. "Retrieval-augmented generation for knowledge-intensive nlp tasks." Advances in Neural Information Processing Systems 33 (2020): 9459-9474.

[11] Lyu, Yuanjie, et al. "CRUD-RAG: A comprehensive chinese benchmark for retrieval-augmented generation of large language models." arXiv preprint arXiv:2401.17043 (2024).

[12] Junior, Robson. *DAQUI DO QUARTO*. Urutau, 2022.

[13] Junior, Robson. *Eu Diante de Antieu*. Folheando, 2022.

[14] Junior, Robson. *PARA ESTREITAR OS TATOS*. Urutau, 2023.

[15] Liu, Jerry. LlamaIndex. Nov. 2022, https://github.com/jerryjliu/llama_index. DOI: 10.5281/zenodo.1234.

[16] Xiao, Shitao, et al. "C-pack: Packaged resources to advance general chinese embedding." arXiv preprint arXiv:2309.07597 (2023).

## ANEXO A - PERGUNTAS E RESPOSTAS

Na Tabela 4 a seguir, estão apresentadas as perguntas e respostas utilizadas no estudo.

| Número | Pergunta | Respostas do Autor |
|---|---|---|
| 1 | Por que o poeta não é um líder? | Porque ele não dispõe de formas e fórmulas, e se perde mais do que quem o procura.<br>**Adicional:** Na verdade, o poeta é um líder; ele é "quem lidera [sua] alcateia". |
| 2 | O que há de comum entre o inverno e o verão? | Ambos podem ser cruéis e têm fim. |
| 3 | O que o poeta prefere: a perfeição ou a imperfeição? | O poeta se orgulha de sua imperfeição. |
| 4 | O relacionamento descrito no poema "Votos reticentes" é saudável para o poeta? | Não, pois o poeta se anula em relação às necessidades da outra pessoa. |
| 5 | Quais as duas maiores descobertas do homem? | O fogo e a combustão. |
| 6 | Quanto tempo se passa no poema "Caixa de areia"? | Várias semanas. |
| 7 | Qual a intenção do poema "Antológica?"? | O poema é um comentário sobre como, apesar dos avanços tecnológicos na geração automática de poemas, não é possível criar algo que tenha como base o verdadeiro sentimento experienciado por seres humanos. |
| 8 | Como fazer funcionar um poema de amor? | *É preciso contemplar o mundo sob olhos inéditos,*<br><br>*pedir ao próprio amor o ponto de vista emprestado.* |
| 9 | Como a ideia de ciclo é sugerida pelo poema "Ciclo"? | O poema é iniciado e finalizado com o mesmo verso. |
| 10 | O que é pior: conseguir dormir ou manter-se acordado? | Manter-se acordado. |
| 11 | Como a mente do poeta se assemelha a uma cidade interiorana? | Assim como uma cidade interiorana, tudo demora a chegar e tudo tende a deixar a mente. Adicionalmente, a mente é interior ao poeta. |

| | | |
|---|---|---|
| 12 | Como o poeta se sente em relação à outra pessoa no poema "Tu"? | O poeta se sente distante, dispensado, difícil de coexistir com essa pessoa. Apesar disso, sente-se também refém à existência, à "coreografia" dessa pessoa. |
| 13 | Qual a implicação do amor ser cristalino? | O amor é frágil, fácil de ser enganado, se desgasta com o tempo. |
| 14 | Quando o poema está pronto? | Quando o próprio poema sugere estar pronto. |
| 15 | Para o poeta, saudade é um conceito simples ou complexo? | É complexo, algo que não é possível de se resumir em apenas uma palavra. |
| 16 | Qual a vantagem de se ter arrependimentos? | Eles impulsionam, motivam e originam a escrita do poeta. |
| 17 | O que o poeta sente ao final do poema “Em nome da paz, um novo nome”? | O texto é ambíguo; o poeta pode se sentir "*seu*" (da pessoa a quem se refere no poema) ou "*a piada*" ou ambos. |
| 18 | Por que o poeta sente que talvez fale demais? | Por não ter se sentido representado ou ouvido anteriormente, ou por ter precisado se silenciar; também para se abrir à pessoa a quem se refere no poema. |
| 18 | Quais dos cinco sentidos são abordados no poema “Tato”? | Apesar de o tato ser destacado no poema, todos os cinco são explorados. |
| 20 | O que está acontecendo com o relacionamento descrito no poema “Irredutível”? | O relacionamento está se deteriorando e se tornando menos tolerável para o poeta. |
| 21 | Qual é a pergunta que tem resposta evidente no poema “Quanta”? | A última pergunta: "*Quanto tempo até fazer tempo demais e não haver mais tempo para nós?*". |
| 22 | O que restou dos melhores poemas do poeta? | A areia (que representa uma parte pequena, insignificante). |
| 23 | O poeta considera que merece ser perdoado? | Apesar de se questionar quanto a isso no início do poema, ao fim, ele se considera merecedor de perdão. |
| 25 | O que é a liberdade descrita no poema “Em (algum metrô em) São Paulo”? | A liberdade é tanto um lugar físico como o estado de ser livre. |

| | | |
|---|---|---|
| 25 | Estar sozinho no poema “Finalmente, sozinho” é algo positivo ou negativo? | É algo positivo, pois demonstra o conforto do poeta em estar sozinho com a pessoa sobre quem ele escreve. |

**Tabela 4: Tabela contendo todas as perguntas e suas respectivas respostas.**