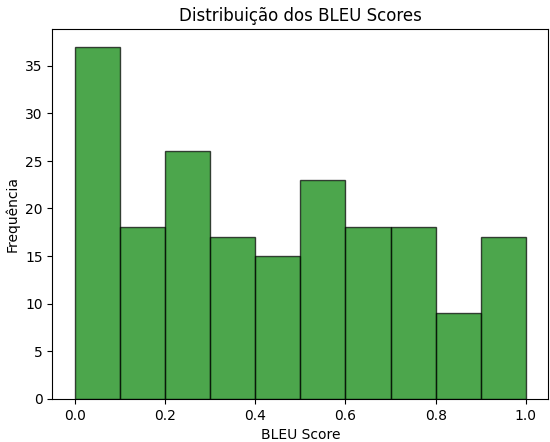
**BLEU Score**

**Análise dos Resultados**

Com base na distribuição dos BLEU Scores calculados entre o código humano e o código gerado pela máquina (GPT + RAG), foram observados os seguintes pontos principais:

****

**Distribuição Variada dos BLEU Scores:**

O histograma mostra uma distribuição dispersa, com valores significativos concentrados em torno de 0.0 e outros espaçados uniformemente até 1.0.

A concentração em BLEU Scores baixos (próximos de 0.0) indica casos em que o código gerado diverge lexicalmente do código humano.

Os valores próximos de 1.0 representam alta similaridade lexical e estrutural entre os códigos.

**Interpretação dos Casos de BLEU Baixo:**

Os baixos valores de BLEU sugerem que o modelo pode ter utilizado construções diferentes ou não conseguiu captar integralmente a intencionalidade do código humano.

Isso pode ocorrer devido à diversidade de soluções possíveis para o mesmo problema ou à qualidade limitada do código humano na base de treinamento.

**Eficácia Geral do Modelo:**

A presença de altos BLEU Scores indica que o modelo é capaz de reproduzir soluções semelhantes ao código humano em muitos casos.

Contudo, a variabilidade na distribuição sugere que a consistência das gerações ainda depende do contexto e do problema apresentado.

**Críticas e Limitações do Experimento**

**Possíveis Viés na Base de Treinamento:**

A qualidade da base de treinamento do modelo GPT + RAG influencia diretamente os resultados. Se o código humano utilizado contém problemas de baixa qualidade, soluções inadequadas podem ter sido aprendidas e replicadas.

Esse viés pode resultar em códigos que, embora lexicamente similares, não sejam funcionalmente corretos ou otimizados.

**Riscos de Introdução de Novos Bugs:**

O código gerado pode introduzir erros devido a interpretações equivocadas do problema original, mesmo que tenha alta similaridade lexical.

A ausência de validação semântica significa que a alta similaridade BLEU pode mascarar falhas lógicas no código.

**Criação de Novas Vulnerabilidades:**

A falta de compreensão contextual pode levar a gerações que expõem o sistema a vulnerabilidades, como ataques de injeção ou uso de práticas inseguras.

O modelo pode replicar padrões inseguros da base de treinamento ou gerar soluções que falham em cenários não previstos.

**Métrica Limitada:**

Embora o BLEU seja uma métrica útil para avaliar similaridade lexical, ela não captura diferenças semânticas ou de funcionalidade. Por isso, não pode ser usada isoladamente para validar a qualidade do código gerado.

**Recomendações Finais**

**Complementar Avaliação com Outras Métricas:**

Combine o BLEU com métricas como Similaridade do Cosseno (CodeBERT), ROUGE ou Levenshtein Distance para capturar aspectos mais amplos da similaridade entre códigos.

**Análise Manual e Testes Automatizados:**

Realize revisões manuais dos códigos gerados, especialmente para casos com BLEU Scores baixos.

Testes unitários e de integração são essenciais para validar a funcionalidade e garantir que o código gerado não introduza falhas.

**Treinamento com Códigos de Alta Qualidade:**

Use bases de treinamento que priorizem soluções seguras, claras e otimizadas para minimizar a propagação de viés negativos.

**Monitoramento Contínuo:**

Utilize ferramentas como SonarQube ou Checkmarx para avaliar vulnerabilidades e a qualidade do código gerado.

**Conclusão**

O experimento revelou que o BLEU Score é uma métrica útil para medir similaridade lexical entre código humano e código gerado por máquina, mas não é suficiente para avaliar a qualidade geral ou funcionalidade do código.

A alta variabilidade nos resultados sugere que o modelo pode gerar códigos de qualidade inconsistentes, dependendo do contexto e da base de treinamento. Para garantir a segurança e a eficiência do código gerado, é fundamental implementar revisões manuais, testes robustos e ferramentas de validação automatizada.