Avaliação de Aplicações de Geração Aumentada de Recuperação por meio de Feed￼back Implícito

Alessandro Marinho de Albuquerque[1], Igor May Wensing[2], Nelson Luiz Joppi Filho[2], Carina Dornelles[1]

[1]PPGCC – INE – Universidade Federal de Santa Catarina (UFSC)
Florianópolis, Santa Catarina – Brazil

[2]Tribunal de Contas de Santa Catarina – Coordenação de Informações para Fiscalização (CIAF/DIE)
Florianópolis, Santa Catarina - Brasil

{alex.marinho.natal,igormaywensing,nelsonluizjoppifilho}@gmail.com,
carina.dorneles@ufsc.br

Abstract. . *In a scenario of evolution of large language models in recent years, with the emergence of a specific market niche, corporate applications began to emerge and become strategic. However, in corporate scenarios, the need to evaluate the results of these applications becomes critical. How do you know if one model is better than another? How do you know if the prompt or question can be improved? How to perform error diagnosis? This article addresses a new proposal for implicit feedback in Retrieval-Augmented Generation (RAG) architectures. The results demonstrate the potential of the proposal applied on VIGIA, a RAG application which detects irregularities on public documents.*

Resumo. *Em um cenário de evolução dos últimos anos dos modelos grandes de linguagem, com o surgimento de um nicho de mercado específico, aplicações corporativas começaram a surgir e se tornar estratégicas. Contudo, em cenários corporativos, a necessidade de avaliar os resultados dessas aplicações se torna crítica. Como saber se um modelo é melhor que outro? Como saber se o prompt ou a pergunta pode ser melhorado? Como realizar o diagnóstico de erros? Este artigo aborda uma nova proposta de feedback implícito em arquiteturas de Geração Aumentada de Recuperação (RAG). Os resultados demonstram potencial da proposta aplicado no VigIA, uma aplicação RAG que detecta irregularidades em documentos públicos.*

1. Introdução

A avaliação da linguagem generativa é um tópico explorado na literatura e permanece um desafio (Gao et al, 2024). Com o surgimento e acesso a serviços de modelos de linguagem generativa, como o CHATGPT, numerosos casos de uso surgiram dentro das corporações, com alguns se tornando estratégicos. Assim, o desafio de avaliar esses modelos de linguagem ganha maior destaque.

As organizações que consomem esses serviços por meio da Geração Aumentada por Recuperação (RAG) ganham relevância no cenário, pois utilizam o conhecimento de modelos pré-treinados para realizar tarefas relacionadas à geração de código ou Pergunta/Resposta, por exemplo. Os detalhes da construção de modelos pré-treinados muitas vezes não são facilmente acessíveis ou representam propriedade intelectual da

empresa, dificultando uma melhor compreensão da explicabilidade desses modelos oferecidos no mercado (OpenAI, 2024).

Neste cenário, como avaliar esses modelos de linguagem pré-treinados? A necessidade de avaliação apoia uma série de processos de tomada de decisão que culminam no sucesso de um projeto de linguagem generativa dentro da organização. A ausência de avaliação traz baixa adoção e capacidade de utilidade, dificultando a integração efetiva nos fluxos de trabalho (Reddy et al, 2021), além de não atender às reais necessidades dos usuários, resultando em desperdício de recursos e esforços (Stahl et al, 2023).

Essas decisões que seguem métodos de avaliação permitem a organização identificar onde são necessárias melhorias nas fases que antecedem a submissão da solicitação ao modelo pré-treinado. Em uma arquitetura RAG, selecionar o melhor modelo, prompt, pergunta e recuperação de contexto já representa um enorme ganho quando guiado por um mecanismo de avaliação.

Além dessas melhorias, a avaliação permite detectar a necessidade de aprimoramentos no processamento de dados que representam a entrada da arquitetura RAG, bem como componentes de geração de vetores. O melhor uso pelo usuário é considerado o principal ganho do mecanismo de avaliação, conforme relatado por Stahl et al (2023).

Shankar et al (2024) relatam algumas abordagens e propõem o EvalGen, utilizando validação humana. No entanto, o autor menciona que validar todas as respostas por um ser humano remove o propósito de algumas aplicações que utilizam modelos de linguagem. Apesar da alta qualidade da avaliação humana, o custo de coletá-las também é alto.

Nesse cenário, é apresentado o mecanismo Implicit Feedback for User Satisfaction Enhancement (INFUSE), capaz de gerar dados para avaliação de feedback implícito, com base em ações observáveis de humanos em resultados de aplicação RAG. O INFUSE foi aplicado no VigIA, um framework de inteligência artificial executado no Tribunal de Contas de Santa Catarina (Rodrigues et al, 2024), capaz de detectar irregularidades em editais de licitação.

2. Fundamentação

Em fevereiro de 2024, Gao et al (2024) abordaram o contexto da avaliação de linguagem generativa em quatro categorias:

- **Métricas Derivadas de LLM**: desenvolvimento de métricas para gerar *embeddings* ou melhorar a geração de probabilidade;
- **Prompting LLMs**: questionamento de LLMs por meio de prompts projetados envolvendo diferentes elementos de avaliação;
- **Aprimoramento de LLMs**: uso de dados rotulados para aprimorar modelos LLM existentes, melhorando suas capacidades linguísticas;
- **Avaliação Colaborativa Humano-LLM**: utilização de avaliadores humanos e modelos de linguagem para gerar avaliações mais robustas.

Gao et al (2024) apontam que grande parte dos artigos publicados enfatiza a categoria Prompting LLMs. O autor destaca a necessidade urgente de trabalhos em larga escala com dados rotulados por humanos para avaliação de modelos de linguagem. Outra oportunidade notada pelo autor diz respeito à predominância de artigos publicados que são baseados na língua inglesa para avaliação, levantando dúvidas sobre as avaliações de modelos em outros idiomas. Finalmente, Gao et al (2024) demonstram que há poucas publicações sobre avaliações colaborativas entre humanos e modelos de linguagem, enfatizando que o papel de cada um deveria ser mais explorado cientificamente.

Em abril de 2024, Shankar et al (2024) apresentaram o EvalGen, uma plataforma híbrida de avaliação que utiliza modelos de linguagem e rotulação humana dedicada para avaliação de código fonte. Os autores mostraram em outros artigos que há desafios na abordagem híbrida, incluindo a dependência excessiva dos humanos nos modelos de linguagem e a supergeneralização, onde os usuários ajustam prompts para cenários específicos sem uma visão holística do impacto desses ajustes em larga escala.

Modelos de linguagem com a capacidade de avaliar resultados de outros modelos, intitulados "LLM como juiz", têm sido explorados na literatura. Alguns estudos utilizaram o GPT-4 como avaliador, alcançando níveis de desempenho semelhantes aos avaliadores humanos (Huang et al, 2024). No entanto, esse cenário levanta questões quando o mesmo modelo é usado tanto para gerar respostas quanto para avaliá-las.

Para a avaliação de arquiteturas RAG, Yu et al (2024) recomendam a visualização a partir de uma perspectiva métrica em cada componente do RAG, citando métricas como relevância, robustez, latência e diversidade de resultados. Os autores destacam a importância de mais pesquisas integrando essas métricas com os resultados obtidos. Finalmente, Finardi et al (2024) mencionam que a avaliação de dados e a diversidade representam desafios para os sistemas RAG.

3. INFUSE Framework

O INFUSE é uma estrutura destinada a fornecer um mecanismo de avaliação em aplicações RAG que utilizam modelos de linguagem. Esse mecanismo se baseia na geração de feedback implícito, derivado de ações dos usuários ao interagirem com informações geradas por um modelo de linguagem.

A arquitetura do INFUSE, conforme apresentado na Figura 1, é composta por diversos módulos:

Analytical Model (Modelo Analítico): Um componente externo ao INFUSE, responsável pela geração das respostas que serão avaliadas. Ele inclui elementos fundamentais de uma arquitetura RAG, como a recuperação aumentada de dados (RAG), prompts, perguntas e o modelo de linguagem (LLM). Esses componentes recebem dados de pipelines de Big Data e produzem as entradas para o INFUSE.

Feedback Generation Subsystem (Subsistema de Geração de Feedback): Dividido em dois componentes principais:

Action Mapping (Mapeamento de Ações): Mapeia as ações observáveis dos usuários, como "Arquivar" ou "Enviar Comunicação", em cima dos dados enviados ao INFUSE.

Feedback Generation Engine (Motor de Geração de Feedback): Processa as ações mapeadas para gerar feedback implícito positivo ou negativo, com base nas ações mapeadas do usuário, o que alimenta o subsistema de métricas.

Golden Dataset (Conjunto de Dados de Referência): Armazena os dados rotulados, que servem como referência para comparação e aprimoramento contínuo das métricas e da eficácia do sistema.

Metrics Subsystem (Subsistema de Métricas): Calcula diversas métricas, como precisão, acurácia, recall, especificidade e F1-score, utilizando os dados de feedback gerados. Esse subsistema é crucial para o refinamento contínuo do modelo analítico.

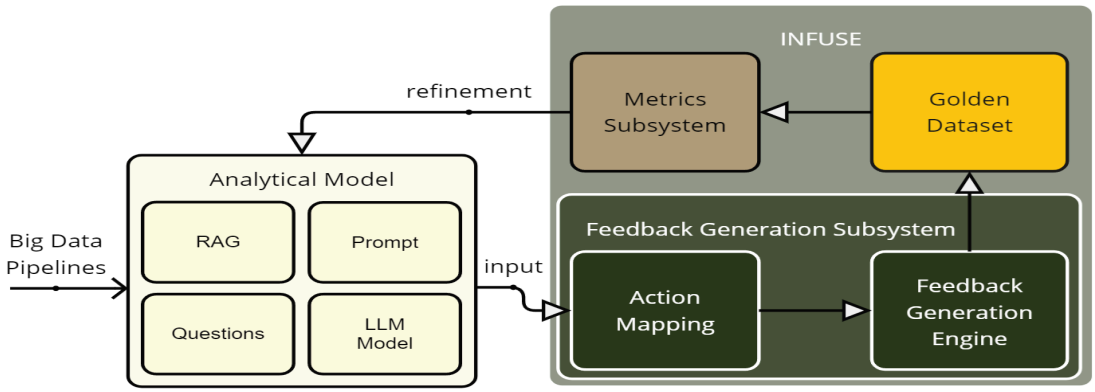


Figura 1 - INFUSE Framework

O INFUSE isola o usuário do processo ativo de avaliação, permitindo que ele se concentre nas regras de negócios e em um fluxo de trabalho mais fluido. O feedback implícito é coletado e processado de forma eficiente para aprimorar continuamente as respostas geradas pelo sistema.

4. Experimentos e Resultados

VigIA é uma estrutura de inteligência artificial projetada para detectar sinais de irregularidades em licitações públicas, automatizando a análise de documentos e identificando potenciais fraudes antes da aprovação (pré-publicação). Utiliza modelos de linguagem, RAG e uma infraestrutura robusta de big data para analisar arquivos de licitação não estruturados, extraindo metadados e respondendo a perguntas de auditoria.

Nos experimentos, foi utilizado um dataset[1] rotulado por um especialista com profundo conhecimento na área, composto por 74 licitações enviadas ao Tribunal de Contas de Santa Catarina (TCE-SC) entre maio e julho de 2024. Esses documentos em

[1] O dataset utilizado para os experimentos pode ser encontrado neste repositório Github

PDF, com status de "Pré-publicado" ou "Publicada", incluem modalidades como "Concorrência", "Pregão Presencial" e "Pregão Eletrônico". Esse dataset foi essencial para a avaliação de benchmark do projeto e validação dos modelos do VigIA.

Uma aplicação notável do VigIA foi na análise de licitações de transporte escolar em Santa Catarina, onde o TCE-SC detectou inconsistências significativas, permitindo intervenções e correções antes da aprovação dos contratos.

O INFUSE foi aplicado ao VigIA para validar a proposta de geração de feedback implícito em aplicações RAG. Nesse contexto, foi definido que a resposta gerada pelo modelo de IA, quando submetida ao evento "Arquivar", geraria um feedback afirmativo de regularidade, enquanto o evento "Enviar Comunicação" geraria um feedback negativo de regularidade.

A geração do motor de feedback é demonstrada no Algoritmo 1 abaixo:

Algorithm 1 Implicit Feedback Processing

Require: Action: User action ("archive" or "send communication")
Require: ChunkSize: Size of the chunk
Require: ChunkNumbers: Number of chunks
Require: ChunkOverlap: Overlap of chunks
Require: EmbeddingModel: Embedding model
Require: LLMModel: LLM model
Require: QuestionText: Text of the question
Require: Prompt: Prompt given to the LLM model
Ensure: Feedback: Implicit feedback of regularity or irregularity
Ensure: Metadata: Processing metadata
Ensure: QuestionText: Text of the question
Ensure: Prompt: Prompt given to the LLM model

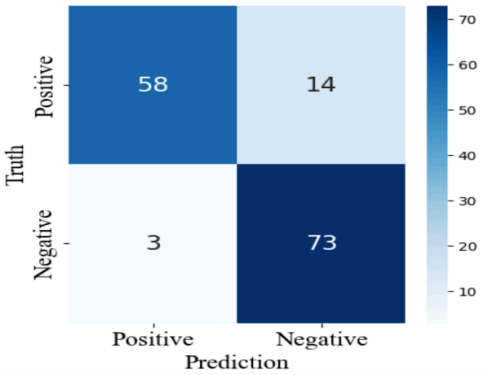
```
function GenerateImplicitFeedback(Action, ChunkSize, ChunkNumbers, ChunkOverlap, EmbeddingModel,
LLMModel, QuestionText, Prompt)
    if Action == "archive" then
        Feedback ← "Confirmed Regularity"
    else if Action == "send communication" then
        Feedback ← "Confirmed Irregularity"
    else
        Feedback ← "Unknown Action"
    end if
    Metadata ← {ChunkSize, ChunkNumbers, ChunkOverlap, EmbeddingModel, LLMModel}
    return {Feedback, Metadata, QuestionText, Prompt}
end function
```

Algoritmo 1 – Motor de feedback implícito

No contexto do VigIA, a implementação do INFUSE gerou um conjunto de dados de 148 itens, onde duas perguntas de auditoria foram aplicadas a 74 licitações de transporte escolar: "Há antecedentes criminais para o motorista?" e "Há exigência de pintura do ônibus?" Esse conjunto de dados permite o refinamento dos componentes do modelo analítico, fornecendo métricas que não apenas oferecem uma visão geral do desempenho, mas também orientam melhorias.

Falsos positivos e falsos negativos podem indicar problemas com parâmetros como o número e tamanho dos blocos, o modelo de embedding, problemas de pré-processamento ou perguntas mal formuladas. No contexto de irregularidades, falsos negativos são mais preocupantes, pois podem permitir que irregularidades passem despercebidas.

A matriz de confusão foi gerada a partir do motor de feedback com a aplicação de métricas como precisão, acurácia, recall, especificidade e F1-score, conforme demonstrado na Figura 2.



Accuracy	0.8851
Precision	0.9508
Recall	0.8056
Specificity	0.9605
F1-Score	0.8722

Figura 2 – Matriz de confusão e métricas

5. Conclusão e Trabalho Futuro

Este artigo apresentou o INFUSE, uma estrutura para feedback implícito em arquiteturas RAG, que separa o usuário do processo ativo de avaliação, facilitando o fluxo de trabalho e a coleta de feedback. Experimentos no âmbito do VigIA, focando em licitações de transporte escolar, mostraram que, além das métricas de desempenho, a análise de falsos positivos e negativos é crucial para melhorias.

A abordagem do INFUSE proporciona feedback confiável e reduz a carga de trabalho, apoiando ajustes de parâmetros e pré-processamento de dados. Problemas relacionados ao número de blocos, tamanho, sobreposição, e melhorias em prompts podem ser melhor orientados pelo conjunto de dados gerado pelo INFUSE. O INFUSE também permite a comparação dinâmica de modelos de linguagem pré-treinados e motores de embedding, contribuindo para avanços científicos.

As principais contribuições do INFUSE no âmbito da literatura se baseia no provimento de uma abordagem para avaliar grande volumes de dados utilizando métricas tradicionais como f1, precisão e recall. Facilita assim a tomada de decisão em sistemas RAG, especialmente àqueles relacionados a Q&A(*Question and Answer)*. Além de considerar que a base dessa avaliação é feita a partir de rotulação implícita por meio de sistemas transacionais, capturando assim o evento relacionado a uma ação e traduzindo em um feedback, minimizando o esforço do usuário de exclusivamente rotular o dado

No contexto do VigIA, o INFUSE facilita a avaliação e melhoria contínua dos parâmetros RAG e ajuda a determinar o melhor modelo para cada caso de uso. Assim, o uso do INFUSE no VigIA aumentará a precisão na detecção de irregularidades em documentos públicos.

Melhorias futuras incluirão mais ações observáveis e interfaces interativas com os resultados do INFUSE, aprimorando sua eficácia em ambientes empresariais. A implementação de métricas adicionais para avaliar a qualidade das respostas permitirá uma avaliação completa da relevância e precisão das respostas em comparação com o contexto original da pergunta.

Referências

Finardi, P., Avila, L., Castaldoni, R., Gengo, P., Larcher, C., Piau, M., ... & Caridá, V. (2024). "The Chronicles of RAG: The Retriever, the Chunk and the Generator". *arXiv preprint arXiv:2401.07883*.

Rodrigues Cássio S, Cardoso, Geovane E., Ramos, Vinicius F. C. "Inteligência artificial no controle de sobrepreço em compras públicas". Revista do Tribunal de Contas de Santa Catarina. Belo Horizonte. Ano 2. Número 2, p 225-252, nov. 2023/abr. 2024.

Gao, M., Hu, X., Ruan, J., Pu, X., & Wan, X. (2024). "Llm-based nlg evaluation: Current status and challenges". *arXiv preprint arXiv:2402.01383*.Dyer, S., Martin, J. and Zulauf, J. (1995) "Motion Capture White Paper", http://reality.sgi.com/employees/jam_sb/mocap/MoCapWP_v2.0.html, December.

Huang, H., Qu, Y., Liu, J., Yang, M., & Zhao, T. (2024). "An empirical study of llm-as-a-judge for llm evaluation: Fine-tuned judge models are task-specific classifiers". *arXiv preprint arXiv:2403.02839*.

OpenAI, "Moving AI governance forward". Disponível em: https://openai.com/index/moving-ai-governance-forward/. Acesso em: 02/06/2024.

Reddy, S., Rogers, W., Makinen, V. P., Coiera, E., Brown, P., Wenzel, M., ... & Kelly, B. (2021). "Evaluation framework to guide implementation of AI systems into healthcare settings". *BMJ health & care informatics*, *28*(1).

Shankar, S., Zamfirescu-Pereira, J. D., Hartmann, B., Parameswaran, A. G., & Arawjo, I. (2024). "Who Validates the Validators? Aligning LLM-Assisted Evaluation of LLM Outputs with Human Preferences". *arXiv preprint arXiv:2404.12272*.

Stahl, B. C., Antoniou, J., Bhalla, N., Brooks, L., Jansen, P., Lindqvist, B., ... & Wright, D. (2023). "A systematic review of artificial intelligence impact assessments". *Artificial Intelligence Review*, *56*(11), 12799-12831.

Yu, H., Gan, A., Zhang, K., Tong, S., Liu, Q., & Liu, Z. (2024). "Evaluation of Retrieval-Augmented Generation: A Survey". *arXiv preprint arXiv:2405.07437*.