

Pergunte ao BI: Uma Interface Conversacional para apoio à Tomada de Decisão usando Modelos de Linguagem

Antony Seabra^{1,2}, Claudio Cavalcante^{1,2}, Gabryel Medeiros¹, João Nepomuceno¹, Vitor Sallenave¹, Nicolaas Ruberg¹ and Sérgio Lifschitz² *

¹ BNDES - Área de Tecnologia da Informação, Rio de Janeiro, RJ

²PUC-Rio - Departamento de Informática, Rio de Janeiro, RJ

{gabme, jonep, vsall, nic}@bndes.gov.br
{amedeiros, cfraga, sergio}@inf.puc-rio.br

Resumo. Apresentamos uma ferramenta que integra modelos de linguagem, documentos, bancos de dados relacionais e agentes autônomos para ampliar as capacidades tradicionais de consulta em linguagem natural. Diferente de soluções limitadas a respostas pontuais sobre dados, nossa proposta permite a realização de análises avançadas, geração de insights e previsões automatizadas, apoiando decisões complexas em contextos organizacionais. A ferramenta foi aplicada em um domínio real de gerenciamento de contratos de tecnologia, com foco em análise de riscos e controle orçamentário. O artigo apresenta uma avaliação qualitativa com usuários finais, evidenciando benefícios em termos de acessibilidade, agilidade analítica e suporte estratégico à tomada de decisão. O trabalho contribui com uma proposta prática e extensível para a construção de sistemas de apoio à decisão mediados por inteligência artificial.

Abstract. We present a tool that integrates language models, documents, relational databases, and autonomous agents to extend traditional natural language query capabilities. Unlike solutions limited to specific responses on data, our proposal allows advanced analysis, generation of insights, and automated predictions, supporting complex decisions in organizational contexts. The tool was applied in a real domain of technology contract management, focusing on risk analysis and budget control. The paper presents a qualitative evaluation with end users, evidencing benefits in terms of accessibility, analytical agility, and strategic support for decision-making. The work contributes with a practical and extensible proposal for the construction of decision support systems mediated by artificial intelligence.

1. Introdução

De acordo com [Brynjolfsson et al. 2011], a tomada de decisão baseada em dados tornou-se uma prática fundamental para organizações que buscam eficiência, previsibilidade e vantagem competitiva. No entanto, como destaca [Kimball and Ross 2011], os dados relevantes para embasar decisões estratégicas estão frequentemente dispersos em fontes

*A. Seabra, C. Cavalcante e S. Lifschitz são parcialmente financiados pela CAPES (institucional) e CNPq (auxílio individual).

heterogêneas, como bancos de dados relacionais, documentos não estruturados, planilhas e relatórios históricos. Ferramentas tradicionais de Business Intelligence (BI) geralmente exigem configurações complexas e conhecimento técnico especializado para construção de dashboards ou consultas [Eckerson 2010], dificultando o acesso direto de tomadores de decisão aos insights mais relevantes. Além disso, tais ferramentas tendem a se limitar a consultas pré-definidas e a análises descritivas, deixando lacunas importantes em tarefas mais complexas, como a análise de riscos e a previsão de cenários.

Com os avanços recentes em Modelos de Linguagem de Grande Escala (Large Language Models - LLMs), tornou-se possível interagir com dados organizacionais por meio da linguagem natural [Raiaan et al. 2024]. Entretanto, soluções de perguntas e respostas baseadas apenas em LLMs ainda enfrentam desafios importantes: dificuldade de conectar múltiplas fontes de informação, limitações na geração de consultas estruturadas precisas e baixa capacidade de guiar o usuário ao longo de um processo decisório. Para superar essas limitações, surge uma abordagem mais robusta: a combinação de LLMs com técnicas de recuperação de informações [Chang et al. 2024], como RAG (Retrieval-Augmented Generation) e Text-to-SQL, além da utilização de agentes autônomos de IA capazes de coordenar tarefas complexas, explorar o contexto e oferecer recomendações adaptadas às necessidades do usuário [Wang et al. 2024]. Neste trabalho, propomos uma ferramenta de Business Intelligence desenvolvida para apoiar decisões em contextos reais e complexos. Sendo avaliada no domínio da gestão de contratos de tecnologia, as principais contribuições deste trabalho são:

C1. O design e a implementação de uma ferramenta de BI baseada em uma arquitetura multiagente, que orquestra LLMs, técnicas de RAG e Text-to-SQL para responder a consultas complexas no domínio do gerenciamento de contratos de tecnologia, sem necessidade de fine-tuning dos modelos.

C2. A demonstração da eficácia do sistema na resolução de diferentes tipos de perguntas relacionadas à gestão de contratos, como análise de riscos contratuais, comparação de cláusulas entre diferentes contratos e estimativas orçamentárias.

C3. A avaliação qualitativa da precisão do sistema e de sua capacidade transformadora para práticas de gestão de contratos, com usuários finais da solução.

O restante deste artigo está estruturado da seguinte forma: a seção 2 apresenta a arquitetura do sistema e os componentes implementados. A Seção 3 detalha o estudo de caso realizado no domínio do gerenciamento de contratos de tecnologia e discute os resultados da avaliação qualitativa com usuários. Por fim, a Seção 4 apresenta as conclusões e direções futuras para a evolução do trabalho.

2. Arquitetura e Implementação

A figura 1 representa a arquitetura proposta para integrar os diferentes tipos de componentes utilizados. Trata-se de uma combinação de diferentes agentes e técnicas para executar tanto a recuperação dos dados que são a base para as respostas, assim como a geração de análises subsequentes. O processo começa com o usuário inserindo uma consulta ou solicitação no sistema (1). Essa consulta pode variar desde uma pergunta simples sobre um contrato específico até uma solicitação complexa de análise de tendências ou previsão de riscos. A consulta do usuário é inicialmente direcionada ao *Roteador*. Este

componente analisa a consulta para determinar qual agente ou conjunto de agentes é mais adequado para processá-la (2). O roteador é essencial para garantir que a consulta seja encaminhada para o caminho correto, otimizando o processamento.

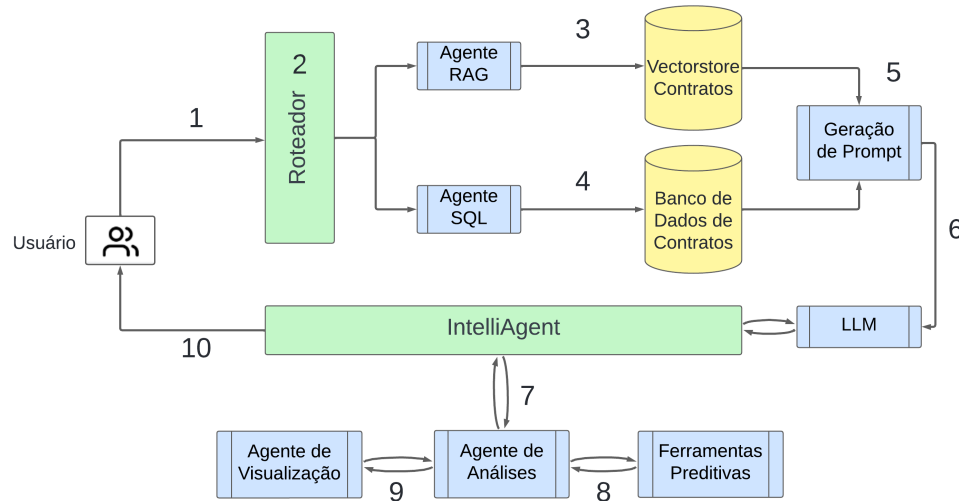


Figure 1. Arquitetura MultiAgente e MultiModal. Fonte: Autores.

Para consultas que envolvem análise de dados não estruturados de documentos contratuais, o roteador envia a consulta para o *Agente RAG*. Este agente utiliza a técnica RAG para extrair informações relevantes da *vectorstore* de contratos (3), um banco de dados otimizado para busca de similaridade vetorial em textos, gerado a partir da leitura dos documentos contratuais e transformação dos pedaços de texto (*chunks*) em vetores usando um modelo de *embeddings* adequado. Para consultas que exigem acesso a dados estruturados de bancos de dados, o roteador envia a consulta para o *Agente SQL*. Este agente utiliza a técnica Text-to-SQL para converter a consulta do usuário em uma consulta SQL que pode ser executada pelo banco de dados (4). Uma vez que os agentes são executados, os dados recuperados são utilizados para gerar um prompt dinâmico (5), que carrega uma instrução detalhada e estruturada para guiar o modelo de linguagem na geração da resposta ou análise desejada. O prompt gerado é enviado para o LLM, que então processa as informações e gera uma resposta em linguagem natural (6).

Em seguida, o IntelliAgent recebe as respostas do LLM e coordena a interação entre as diferentes ferramentas através do InsightAgent (7). Essas ferramentas vão tratar as respostas na direção da tomada de decisão, baseada na consulta original do usuário. Ferramentas preditivas podem ser chamadas para realizar análises preditivas com base nos dados recebidos (8). Essas ferramentas podem gerar por exemplo, previsões de vendas ou orçamentos, tendências de custos ou outras informações úteis para a tomada de decisões. Ademais, ferramentas de visualização podem ser chamadas para gerar gráficos (9), caso os dados recebidos tenham um formato tabular que permita sua construção. A saída para o Usuário é gerada pelo IntelliAgent e contém a resposta final (10), que pode incluir texto, visualizações de dados e/ou previsões, dependendo da natureza da consulta original do usuário e dos agentes que foram executados ao longo do processo.

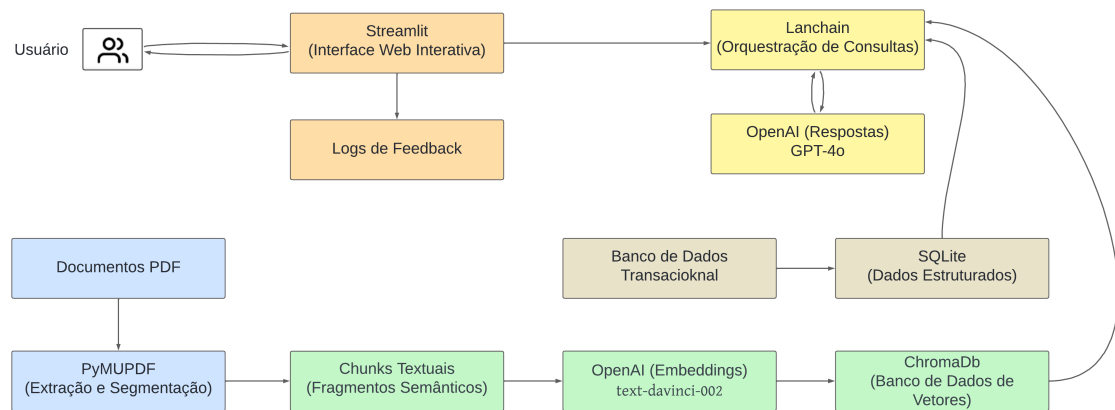



Figure 2. Implementação. Fonte: Autores.


A implementação da solução foi estruturada para garantir eficiência e precisão no processamento de dados contratuais, iniciando pela extração de texto de documentos PDF com a biblioteca PyMuPDF e posterior transformação desses conteúdos em vetores numéricos com o modelo de embeddings text-davinci-002 da OpenAI, armazenados no ChromaDB com metadados relevantes. Os dados estruturados foram gerenciados em SQLite, enquanto a orquestração das interações com o modelo GPT-4o foi realizada via Lanchain. A interface web foi desenvolvida com Streamlit, permitindo a escolha de pessoas, envio de perguntas em linguagem natural e recebimento de respostas detalhadas, com funcionalidades adicionais como exportação de conversas e avaliação da qualidade das respostas.

3. Avaliação

A avaliação da eficácia da solução proposta foi realizada no domínio específico da Gestão de Contratos, e nos concentramos em demonstrar como a ferramenta facilita e aprimora a tomada de decisões estratégicas através da análise de dados complexos. Para este propósito, selecionamos duas áreas críticas para a gestão contratual: a análise de riscos e a previsão orçamentária. Estas áreas foram escolhidas por seu impacto direto na saúde financeira e operacional das organizações, onde insights precisos e oportunos podem resultar em economias significativas e na mitigação de potenciais problemas.

A figura 3 ilustra um exemplo de como o Agente RAG foi utilizado para identificar riscos associados a contratos de banco de dados, respondendo à pergunta do usuário sobre os riscos de gestão desses contratos. O Agente RAG, ao processar os documentos contratuais, identificou seções relevantes que abordavam a "Matriz de Riscos" e outras cláusulas relacionadas ao gerenciamento de riscos. Essas informações foram então passadas para o IntelliAgent, que as resumizou, destacando os pontos mais importantes para o gestor do contrato. O resumo gerado pelo IntelliAgent inclui a alocação de riscos entre as partes, a identificação conjunta de riscos, a importância da Matriz de Riscos, a cláusula de equilíbrio econômico-financeiro e a proibição de aditivos em determinadas situações. Este processo demonstra a capacidade do sistema em extrair informações complexas de documentos contratuais e apresentá-las de forma concisa para a tomada de decisões.

 Em relação aos contratos de banco de dados, quais são os riscos associados à sua gestão?



Gerenciamento de Riscos em Contratos de Banco de Dados

Os contratos do BNDES possuem uma abordagem consistente para o gerenciamento de riscos, conforme identificado na seção "Matriz de Riscos" dos contratos. Aqui estão alguns pontos importantes sobre como os riscos podem ser mitigados pelo gestor do contrato:

- **Alocação de Riscos:** O BNDES e o contratado estabelecem que o risco deve ser alocado à parte com a maior capacidade de gerenciá-lo e absorvê-lo, visando obter o melhor custo contratual.
- **Identificação dos Riscos:** Ambas as partes identificam os riscos decorrentes da relação contratual.
- **Matriz de Riscos:** Os respectivos responsáveis pelos riscos identificados são estabelecidos na Matriz de Riscos, que é parte do anexo do contrato.
- **Cláusula de Equilíbrio Econômico-Financeiro:** Qualquer reajuste de preço mencionado na Matriz de Riscos deve respeitar a cláusula de equilíbrio econômico-financeiro do contrato.
- **Proibição de Aditivos:** É vedada a celebração de aditivos contratuais decorrentes de eventos supervenientes que sejam de responsabilidade do contratado, conforme alocado na Matriz de Riscos.

Esses pontos são fundamentais para garantir que os riscos sejam gerenciados de forma eficaz e que o gestor do contrato tenha diretrizes claras para mitigá-los.

Figure 3. Análise de Riscos de Contratos de Bancos de Dados. Fonte: Autores.

Em relação à previsão orçamentária, o sistema extrai informações sobre os valores dos contratos e seus respectivos parâmetros do banco de dados transacional. Com essas informações em mãos, o IntelliAgent orquestra a execução de uma análise preditiva, neste caso utilizando uma série temporal para todos os contratos iniciados a partir de 2020, e calculando os valores

 Considerando que todos os contratos sofrerão um reajuste de 14,75% ao seu término, projete o orçamento necessários para os próximos 5 anos.

 Para projetar o orçamento necessário para os próximos 5 anos, considerando o reajuste de 14,75% ao término dos contratos, segue a lista dos contratos com seus valores globais acumulados atuais e os valores projetados com o reajuste:

OCS Número/Ano	Valor Global Acumulado Atual (R\$)	Valor Projetado com Reajuste (R\$)
0206/2020	36,000.00	41,310.00
0185/2020	2,190,000.00	2,512,175.00
0189/2020	3,281,446.00	3,767,660.93
0190/2020	1,980,000.00	2,273,050.00
...
0118/2022	262,095.23	300,754.39

Figure 4. Orçamento. Fonte: Autores.

contrato a contrato com base no índice de reajuste nos parâmetros de datas de cada contrato. A figura 4 mostra os resultados e a figura 5 ilustra a capacidade do IntelliAgent de gerar visualizações gráficas a partir dos dados tabulares de previsão orçamentária. As avaliações realizadas por gestores de contratos revelaram diferentes níveis de desempenho do sistema frente às questões de competência. Na categoria de Avaliação de Riscos, o desempenho foi limitado, principalmente quando relacionado a questões legais. Já na categoria Projeções Orçamentárias, os resultados indicaram um bom potencial de uso preditivo, especialmente em relação às despesas futuras, ainda que haja espaço para melhorar a modelagem do impacto inflacionário.

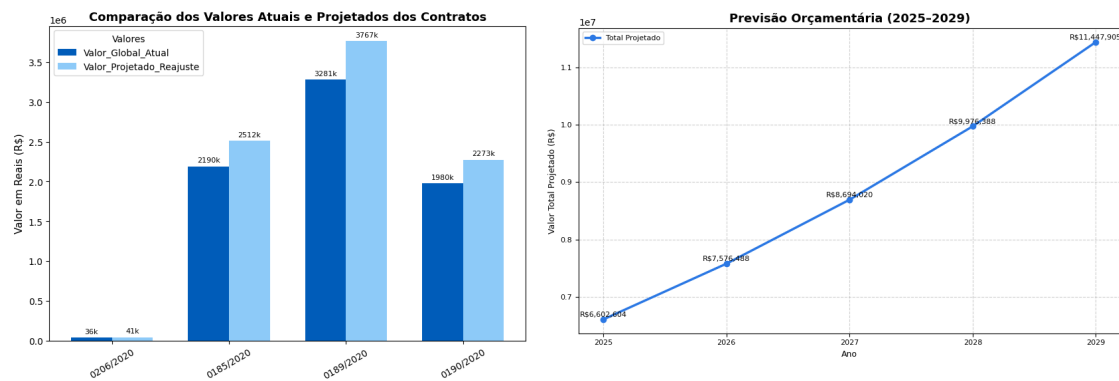


Figure 5. Visualização da previsão orçamentária. Fonte: Autores.

4. Conclusões e Trabalhos Futuros

A implementação da solução demonstra o forte potencial da integração entre LLMs e funcionalidades de BI, aliada a agentes de IA, RAG, Text-to-SQL e engenharia de prompts dinâmicos, para aprimorar a gestão de contratos por meio da extração de insights de documentos e bases de dados. A avaliação no domínio evidenciou sua eficácia em análises como identificação de riscos, comparação de SLAs e previsão orçamentária, destacando-se pela flexibilidade e personalização proporcionadas pela orquestração do IntelliAgent. Para trabalhos futuros, pretende-se ampliar a solução com identificação automática de padrões complexos, modelos de recomendação personalizados, visualizações interativas e sua aplicação em outros domínios, como gestão de projetos, e finanças.

Disclaimer: Este artigo representa a opinião dos autores e é produto de sua pesquisa profissional. Não foi elaborado para representar a posição ou as opiniões do BNDES ou de seus membros, nem tampouco reflete qualquer posição oficial da instituição.

References

- Brynjolfsson, E., Hitt, L. M., and Kim, H. H. (2011). Strength in numbers: How does data-driven decisionmaking affect firm performance? *Available at SSRN 1819486*.
- Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., et al. (2024). A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.
- Eckerson, W. W. (2010). *Performance dashboards: measuring, monitoring, and managing your business*. John Wiley & Sons.
- Kimball, R. and Ross, M. (2011). *The data warehouse toolkit: the definitive guide to dimensional modeling*. John Wiley & Sons.
- Raiaan, M. A. K., Mukta, M. S. H., Fatema, K., Fahad, N. M., Sakib, S., Mim, M. M. J., Ahmad, J., Ali, M. E., and Azam, S. (2024). A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE access*, 12:26839–26874.
- Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., et al. (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.