Otimizando a pesquisa e análise de dados semiestruturados com modelos de LLM

Uma aplicação em relatórios de auditoria de empresas listadas na B3

David Marques
david.leal@lcad.inf.ufes.br
Departameto de Informática
Universidade Federal do Espirito Santo
Vitória, Espírito Santo, Brasil

Vagner Antônio Marques
vagner.marques@ufes.br
Departameto de Ciências Contábeis
Universidade Federal do Espirito Santo
Vitória, Espírito Santo, Brasil

Aerty Pinto dos Santos
aerty.santos@edu.ufes.br
Programa de Pós-Graduação em Informática
Universidade Federal do Espirito Santo
Vitória, Espírito Santo, Brasil

Elias Silva de Oliveira
elias@lcad.inf.ufes.br
Programa de Pós-Graduação em Informática
Universidade Federal do Espirito Santo
Vitória, Espírito Santo, Brasil

Abstract

The growth of open data, combined with the rapid development of Large Language Models (LLMs), has enabled significant advances in the scientific and technological fields. In the context of the capital markets, this environment enhances large-scale quantitative and qualitative analysis of disclosed information, which is currently limited due to the semi-structured nature of many source documents. As a labor-intensive and error-prone process, manual data extraction prevents systematic analysis. This extended abstract, as an ongoing work, presents the development of a solution for the automated collection and extraction of data from audit reports, using the CVM Open Data Portal. Through the application of LLMs and prompt engineering, the algorithm systematically interprets and converts textual and numerical information into structured JSON databases, streamlining market analysis and offering strong potential contributions for researchers, investors, regulators, and other stakeholders.

Palavras-chave

Inteligência Artificial, Processamento de Linguagem Natural, Mineração de Texto, Engenharia de Prompt

1 Introdução

Os avanços tecnológicos das últimas décadas têm impulsionado a rápida disponibilização de dados não estruturados, transformando as práticas contábeis e de auditoria [2, 6]. Historicamente, a extração de dados de relatórios de auditoria exige esforço manual oneroso, limitando o escopo de pesquisas e a acurácia das análises do mercado financeiro [11, 15].

O surgimento de Grandes Modelos de Linguagem (LLMs) apresenta uma solução viável para processar grandes volumes textuais com rapidez e eficiência [10, 16]. Modelos avançados e abordagens combinadas garantem respostas atreladas a bases documentais específicas e têm seu desempenho otimizado por meio de Engenharia de Prompt criteriosa [1, 5, 9].

No contexto internacional, estudos recentes evidenciam a eficácia dessa abstração textual no julgamento contábil. Por exemplo, Küster et al. [7] demonstraram que modelos especializados, como o FinBERT, são capazes de identificar sinais preditivos de perdas futuras analisando a semântica dos Principais Assuntos de Auditoria (PAAs). Paralelamente, pesquisas como a de Doi et al. [3] destacam a alta acurácia na classificação automatizada de *Key Audit Matters* (KAMs) quando submetidos a categorizações zero-shot estruturadas, e Lazarev and Sedov [8] validaram a robustez de LLMs como o Gemini na síntese automatizada de tendências financeiras.

Contudo, na literatura nacional, as abordagens ainda carecem de metodologias que absorvam em larga escala as ricas nuances qualitativas legadas do mercado de capitais brasileiro, com análises muitas vezes limitadas à simples quantificação pontual de itens por auditores humanos ou bases amostrais restritas [11, 14].

Diante desse cenário e aproveitando o espaço da cultura de dados abertos (*open data*) [13], este resumo estendido (fruto de um trabalho em andamento) apresenta o protótipo de uma ferramenta baseada em LLM — o *Gemini 2.0 Flash* — estruturada via Engenharia de Prompts para automatizar a coleta e organização de dados não estruturados em relatórios de auditoria (Pareceres e Declarações) de empresas listadas na B3. O objetivo é reduzir a assimetria informacional, mitigando o esforço braçal que coíbe análises sistemáticas, e assim subsidiar pesquisadores, investidores e reguladores com uma extração de dados contábeis muito mais massiva e eficiente.

2 Solução Proposta

Esta seção descreve, de forma sucinta, o processo empregado na arquitetura e validação do protótipo de extração automatizada de informações contábeis baseada em LLM. O processo foi dividido na aquisição de dados do Portal de Dados Abertos da Comissão de Valores Mobiliários (CVM) e na extração/validação das informações utilizando Engenharia de Prompts.

O *dataset* original é composto por dezenas de milhares de manifestações textuais em Formulários de Demonstrações Financeiras Padronizadas (DFP), especificamente as seções de Pareceres e Declarações. Foi desenvolvido um algoritmo em `Python` [12] para o agrupamento inicial e limpeza básica desse conjunto. A Figura 1 ilustra a arquitetura global do protótipo desenvolvido.

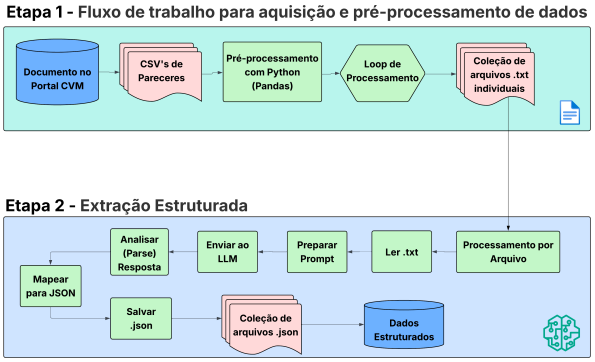

Figura 1: Fluxograma da Arquitetura do Processamento

Para a extração das variáveis de interesse, utilizou-se o modelo `gemini-2.0-flash`. A interação foi guiada por um processo iterativo de Engenharia de Prompt [1], que instruiu a LLM a retornar as saídas já pré-formatadas como um objeto `JSON` estrito contendo duas chaves essenciais: `metadata_origem` e `dados_extraidos_llm` (Figura 2).

```
{
  "metadata_origem": {
    "cnpj": "CNPJ",
    "nome_companhia_original": "Nome"
  },
  "dados_extraidos_llm": {
    "titulos_e_assuntos": [ ... ],
    "datas_mencionadas": [ ... ],
    "contador": { ... }
  }
}
```

Figura 2: Recorte da estrutura JSON desenhada para a saída.

Na fase de validação (analisando os resultados de um subconjunto de amostras), os arquivos foram submetidos a um validador utilizando `jsonschema`. Dos artefatos gerados automatica e aleatoriamente neste teste inicial, uma porção majoritária atendeu à integridade hierárquica esperada. Contudo, em casos específicos apontados na Tabela 1, foram registrados erros comumente ligados à incapacidade pontual da IA lidarem com formatações legadas (ex: quebras de linha não escapadas corretamente dentro do texto jurídico).

Por fim, no que cerne ao comparativo de eficiência computacional e esforço humano, estima-se o tempo necessário para que um auditor realizasse a mesma extração manualmente. Assumindo uma estimativa mínima de 5 minutos por

Tabela 1: Taxonomia dos erros de validação JSON

Tipo de Erro	Descrição Técnica	Qtd.	%
Invalid Control	Quebras de linha não escapadas.	24	85,7
Invalid \escape	Uso incorreto de barras invertidas.	3	10,7
Unterminated String	LLM interrompeu a string subitamente.	1	3,6
Total		**28**	**100**

documento ($T_{unitario}$) para leitura e transcrição estruturada, o tempo total seria calculado como:

$$T_{humano} = N_{docs} \times T_{unitario} = 480 \times \mathbf{5} = 2.400 \text{ min} \approx 40 \text{ horas} \quad (1)$$

Em contraste, a solução automatizada, apesar das restrições de limite de *API* impostas na versão gratuita, exigiu um tempo de supervisão interativa mínima do pesquisador. A Tabela 2 sumariza esse comparativo, demonstrando que a automação via LLMs não apenas mitiga erros de transcrição por cansaço, como tem escalabilidade financeira superior.

Tabela 2: Comparativo preliminar de Eficiência: Humano vs. LLM

Métrica	Processo Manual	Automatizado (LLM)
Esforço Humano Ativo	≈ 40 horas	≈ 1 hora (Configuração)*
Tempo Decorrido	≈ 5 Dias	≈ 48 horas (*rate limit*)
Custo Financeiro Variável	Alto (H/h do avaliador)	Zero (*API* Gratuita)
Disponibilidade	Limitada as horas úteis	24/7 (Ininterrupto)
Escalabilidade	Baixa (Linear)	Alta (Limitada por *API*)

*Tempo focado na criação inicial dos scripts.

Tais resultados preliminares evidenciam a notável capacidade do processo para massificar e escalar estudos investigativos no cenário contábil e de auditoria [15].

3 Considerações Finais

Este resumo estendido delineou a proposta de um protótipo fundamentado em Grandes Modelos de Linguagem (LLMs) e Engenharia de Prompt para extrair atributos-chave de relatórios de auditoria da Comissão de Valores Mobiliários (CVM). O volume massivo de informações textuais contido anualmente nos Formulários de Referência e nas Demonstrações Financeiras Padronizadas tem se mostrado um obstáculo empírico considerável para pesquisadores e analistas, os quais rotineiramente necessitam despender recursos exaustivos na coleta manual para a construção de bancos de dados consistentes.

Embora se trate de um trabalho em andamento que ainda demanda extensa validação humana sobre o conteúdo semântico, os resultados parciais revelam um expressivo acerto estrutural e logístico na geração automatizada de relatórios em formato `JSON`. O modelo, ancorado pela API do *Gemini 2.0 Flash*, provou-se capaz de sintetizar e hierarquizar dezenas de diferentes facetas desses documentos — desde qualificações de opinião e datas-chave até complexamente estruturados Assuntos Principais de Auditoria (PAAs). Essa abordagem algorítmica abstrai dezenas de horas de *esforço humano ativo*

em meras frações de tempo empregadas na supervisão passiva dos artefatos produzidos [6].

Contudo, as limitações inerentes de o modelo lidar estritamente com dados ruidosos legados e as ocorrências do fenômeno de alucinação algorítmica reiteram a importância premente de camadas adicionais de *compliance* normativo e validação cruzada, semelhantes aos testes descritos no nosso fluxo JSON. Entendemos que ferramentas preditivas e gerativas não substituem o crivo e a ética fundamentais do auditor, não obstante, demonstram-se valiosíssimos assistentes técnicos de alto rendimento no preâmbulo investigativo contábil [4].

Como desdobramentos futuros que nortearão a etapa completa deste trabalho, pretendemos não apenas aprimorar a capacidade da LLM em identificar contextos oblíquos e jargões técnico-jurídicos de alta ambiguidade, mas também envolver a consolidação contínua dessa arquitetura por meio da criação de uma *API* ou repositório público. Nosso objetivo basilar reside em democratizar o acesso tempestivo a este acervo quantificado, alavancando novos horizontes de governança e pesquisas no mercado acadêmico e financeiro [2, 13].

Agradecimentos

Os autores agradecem à Fundação de Amparo à Pesquisa e Inovação do Espírito Santo (Fapes) pelo apoio financeiro na realização desta pesquisa.

REFERÊNCIAS

[1] Xavier Amatriain. 2024. PROMPT DESIGN AND ENGINEERING: INTRODUCTION AND ADVANCED METHODS. arXiv:2401.14423 [cs.SE]

[2] Jeremy Bertomeu. 2020. Machine learning improves accounting: discussion, implementation and research opportunities. *Rev Account Stud* 25, 3 (Sept. 2020), 1135–1155. https://doi.org/10.1007/s11142-020-09554-9

[3] Nobushige Doi, Yusuke Nobuta, and Takeshi Mizuno. 2025. Zero-Shot Text Classification Using Large Language Models for Key Audit Matters in Japanese Audit Reports. *International Journal of Smart Computing and Artificial Intelligence* 9, 1 (2025), 1–14. IJSCAI865.

[4] Lazarus Elad Fotoh and Tatenda Mugwira. 2025. Exploring Large Language Models in external audits: Implications and ethical considerations. *International Journal of Accounting Information Systems* 56 (2025), 100748. https://doi.org/10.1016/j.accinf.2025.100748

[5] Xinyu Hu, Yihong Tang, Zhi Jin, et al. 2024. Large language models for software engineering: Survey and open problems. *Comput. Surveys* 56, 7 (2024), 1–33.

[6] Hussein Issa, Ting Sun, and Miklos A. Vasarhelyi. 2016. Research Ideas for Artificial Intelligence in Auditing: The Formalization of Audit and Workforce Supplementation. *Journal of Emerging Technologies in Accounting* 13, 2 (Sept. 2016), 1–20. https://doi.org/10.2308/jeta-10511

[7] Stephan Küster, Tobias Steindl, and Max Göttsche. 2025. The informational content of key audit matters: Evidence from using artificial intelligence in textual analysis. *Contemporary Accounting Research* —, — (2025), 1–32. https://doi.org/10.1111/1911-3846.13070

[8] Andrei Lazarev and Dmitrii Sedov. 2024. Utilizing Modern Large Language Models (LLM) for Financial Trend Analysis and Digest Creation. In *2024 6th International Conference on Control Systems, Mathematical Modeling, Automation and Energy Efficiency (SUMMA)*. 317–321. https://doi.org/10.1109/SUMMA64428.2024.10803746

[9] Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Sebastian Borgeaud, Antonia Creswell, et al. 2022. Solving Quantitative Reasoning Problems with Language Models. In *Advances in Neural Information Processing Systems*, Vol. 35. 3843–3857.

[10] Evangelos Liaras, Michail Nerantzidis, and Antonios Alexandridis. 2024. Machine learning in accounting and finance research: a literature review. *Rev Quant Finan Acc* 63, 4 (Nov. 2024), 1431–1471. https://doi.org/10.1007/s11156-024-01306-z

[11] Vagner Antônio Marques, Lanna Nogueira Pereira, Idamo Favalessa De Aquino, and Viviane Da Costa Freitag. 2021. Has it become more readable? Empirical evidence of key matters in independent audit reports,. *Rev. contab. finanç.* 32, 87 (Dec. 2021), 444–460. https://doi.org/10.1590/1808-057x202112990

[12] Wes McKinney. 2012. *Python for data analysis: Data Wrangling with Pandas, NumPy, and IPython*. O'Reilly Media, Inc.

[13] Peter Murray-Rust. 2008. Open Data in Science. *Nat Prec* (Jan. 2008). https://doi.org/10.1038/npre.2008.1526.1

[14] Aerty Pinto dos Santos, Eduardo Almeida Santos Oliveira, Juliana Pinheiro Campos Pirovani, and Elias de Oliveira. 2025. Extraindo o significado de uma palavra: uma abordagem de Inteligência Artificial. *Transinformação* 37 (nov 2025). https://periodicos.puc-campinas.edu.br/transinfo/article/view/14829

[15] Kleyverson Leonardo Dos Santos, Renan Bittencourt Guerra, Vagner Antônio Marques, and Elizeu Maria Júnior. 2020. Os Principais Assuntos de Auditoria Importam? Uma análise de sua associação com o Gerenciamento de Resultados. *REPeC* 14, 1 (March 2020). https://doi.org/10.17524/repec.v14i1.2432

[16] Lina Zhou, Shimei Pan, Jianwu Wang, and Athanasios V. Vasilakos. 2017. Machine learning on big data: Opportunities and challenges. *Neurocomputing* 237 (May 2017), 350–361. https://doi.org/10.1016/j.neucom.2017.01.026