Project Proposal
JSONLLM: Extração Escalável de Atributos de Produtos com Grandes Modelos de Linguagem (LLMs)

João Vítor Fernandes Dias *
Computer Science Postgraduate Program
2024711370
Federal University of Minas Gerais
Belo Horizonte, MG; 31270-901
joaovitorfd2000@gmail.com

Melissa Dias Mattos
Computer Science Postgraduate Program
2024675063
Federal University of Minas Gerais
Belo Horizonte, MG; 31270-901
melissadmattos366@gmail.com

Abstract

Muitos produtos de *e-commerce* possuem múltiplos atributos – como tamanho, cor e material – que são essenciais para auxiliar os clientes em tomadas de decisões de compra informadas. No entanto, a extração e a organização dessas informações a partir de descrições textuais de produtos não estruturadas é uma tarefa desafiadora. Neste trabalho, propomos o JSONLLM, uma abordagem que explora o potencial dos grandes modelos de linguagem (LLMs) para identificar e estruturar atributos de produtos em formato JSON. Nosso método utiliza as capacidades avançadas de compreensão semântica dos LLMs para reconhecer e categorizar com precisão informações provenientes de descrições diversas e complexas. Os objetivos principais deste projeto incluem: integrar LLMs para a extração de atributos de produtos, garantindo a escalabilidade com sistemas reais de *e-commerce*, padronizando e avaliando a precisão do modelo proposto em comparação a técnicas de *prompt engineering*. Espera-se que o sistema resultante melhore significativamente a precisão da extração de atributos de produtos, facilitando a criação de catálogos de produtos mais ricos e detalhados. Vale ressaltar que a estruturação em JSON facilita a integração com sistemas de *e-commerce* existentes, melhorando a experiência do usuário final.

1 Introdução

A extração e estruturação de atributos de produtos em *e-commerce* representa um desafio crucial devido à diversidade e complexidade das descrições de produtos não estruturadas. Muitos produtos possuem múltiplos atributos essenciais, como tamanho, cor e material, que são fundamentais para decisões informadas de compra. Métodos tradicionais de extração de informações frequentemente enfrentam limitações em precisão e adaptabilidade a descrições variadas Brinkmann et al. (2024). Recentemente, o uso de Grandes Modelos de Linguagem (LLMs) tem demonstrado potencial significativo para superar esses desafios, permitindo uma extração mais precisa e estruturada em formatos como JSON, facilitando a integração com sistemas de *e-commerce* existentes Sinha and Gujral (2024). Esta motivação conduz à proposta do sistema JSONLLM, que visa explorar as capacidades avançadas dos LLMs para melhorar a qualidade da extração e apresentação de atributos de produtos.

*ORCID: 0000-0002-8156-9551; GitHub: jvfd3; LinkedIn: jvfd3

Preprint. Under review.

2 Objetivos

Os principais objetivos deste projeto são: 1. Desenvolver um pipeline que integre LLMs para a extração automática e precisa de atributos de produtos a partir de descrições não estruturadas; 2. Criar um esquema padronizado em formato JSON para representar os atributos extraídos de forma clara e interoperável; 3. Avaliar a precisão e eficiência do modelo proposto em comparação com métodos tradicionais de extração, mediante experimentos controlados; 4. Garantir a escalabilidade e aplicabilidade do sistema em ambientes reais de *e-commerce*, com potencial para integração fácil.

3 Metodologia

A metodologia envolve a coleta e preparação de um conjunto de dados representativo de descrições de produtos do *e-commerce*, apoiada por bases públicas como o dataset MAVE (*Multi-source Attribute Value Extraction*), que possui milhões de anotações de pares atributo-valor Yang et al. (2022)[2]. Após o treino com as bases de dados anotadas o modelo será aplicado no *dataset* da *Track Product Search* da TREC (*Text REtrieval Conference*)[3] Campos et al. (2023) para avaliação empírica do método de enriquecimento dos dados.

Será adotada uma arquitetura baseada em LLMs, com experimentações utilizando modelos *open-source*, explorando técnicas de *few-shot* e *zero-shot learning* para extração estruturada. Em etapa posterior, através do *finetuning* um modelo pré-treinado será aprimorado através dos exemplos anotados e sua performance será avaliada.

O *pipeline* incluirá a padronização dos atributos extraídos em JSON, com validações automatizadas e refinamento iterativo via técnicas de autoaperfeiçoamento (*self-refinement*) para aumentar a precisão Brinkmann and Bizer (2025). O desempenho do sistema será avaliado por métricas de precisão, *recall* e f1-*score*, comparado a abordagens baseadas em aprendizado de máquina tradicional e ferramentas existentes.

4 Resultados Esperados e Contribuições

Espera-se que o sistema JSONLLM melhore significativamente a precisão da extração de atributos de produtos, permitindo a geração de catálogos detalhados e padronizados. Assim os catálogos poderão ser dispostos em interfaces agradáveis aos usuários para facilitar a avaliação de opções disponíveis.

A contribuição do projeto inclui o desenvolvimento de um *pipeline* LLM para extração de atributos, o avanço no uso de JSON como padrão para dados estruturados em contexto comercial e a validação empírica da superioridade do método em relação a abordagens tradicionais Brinkmann et al. (2024). Além disso, o projeto pode ser aplicado em múltiplos domínios de comércio eletrônico, bem como em outras bases de dados, potencializando a interoperabilidade entre sistemas diversos.

5 Cronograma

Table 1: Cronograma do Projeto

Prazo	Atividade
02/10	Proposta do projeto
09/10	Revisão de literatura
13/10	Preparação dos dados
14/10	*Mid-Project Check-in* (*Report Draft*)
21/10	Implementação do modelo
28/10	Experimentos e avaliações
18/11	Submissão Final
25/11	Apresentações

[2] `https://github.com/google-research-datasets/MAVE`

[3] `https://huggingface.co/datasets/trec-product-search/product-recommendation-2025`

References

Brinkmann, A. and Bizer, C. (2025). Self-refinement strategies for llm-based product attribute value extraction. *arXiv preprint arXiv:2501.01237*. Accessed October 2025.

Brinkmann, A., Shraga, R., and Bizer, C. (2024). Extractgpt: Exploring the potential of large language models for product attribute value extraction. In Delir Haghighi, P., Greguš, M., Kotsis, G., and Khalil, I., editors, *Information Integration and Web Intelligence*, volume 15342, pages 38–52. Springer Nature Switzerland, Cham.

Campos, D., Kallumadi, S., Rosset, C., Zhai, C. X., and Magnani, A. (2023). Overview of the trec 2023 product product search track.

Sinha, A. and Gujral, E. (2024). Pae: Llm-based product attribute extraction for e-commerce fashion trends. *arXiv preprint arXiv:2405.17533*. Accessed October 2025.

Yang, Y., Kharitonov, E., Ghosh, S., et al. (2022). Mave: A large-scale dataset for product attribute value extraction. Accessed October 2025.

A Desafios e Estratégias de Mitigação

Entre os principais desafios destacam-se a variabilidade e ambiguidade nas descrições dos produtos que dificultam a extração consistente de atributos; a necessidade de adaptação do modelo a múltiplos domínios de produto; e a escalabilidade do *pipeline* para grandes volumes de dados. Para mitigar esses desafios, serão aplicadas estratégias de refinamento iterativo para corrigir erros, adoção de esquemas JSON flexíveis e extensíveis, e uso de conjuntos de dados amplos e diversificados para o treinamento e validação Brinkmann and Bizer (2025).