AI-PAVE-Br: Leveraging Large Language Models for Enhanced Product Attribute Value Extraction through a Golden Set Approach

Murilo Gazzola[1,2], **Hugo Gobato Souto**[1,3], **Samuel Silva**[1], **Júlia Schubert Peixoto**[1], **Felipe Siqueira**[1], **André Luis Pedroso de Morais**[1], **Caio Gomes**[1]

[1] LuizaLabs – Center of Excellence in Artificial Intelligence
São Paulo, SP – Brazil

[2]Department of Computing and Informatics – Mackenzie Presbyterian University
São Paulo, SP – Brazil

[3]Institute of Mathematics and Computer Sciences - University of São Paulo
São Carlos, SP – Brazil

{murilo.gazzola, hugo.souto,samuel.silva, andre.morais, gomes.caio}@luizalabs.com

Abstract. *The explosive growth and complexity of product data within the dynamic Brazilian e-commerce landscape demand robust and specialized methods for structured information extraction. Traditional approaches to Product Attribute Value Extraction (PAVE) often struggle with the linguistic nuances and sheer diversity of product descriptions in Portuguese. To address this critical gap, this paper introduces two major contributions. First, we present* ***AI-PAVE-Br****, a specialized system engineered with Large Language Models (LLMs) to perform high-accuracy PAVE specifically for Brazilian e-commerce catalogs. Second, to facilitate reproducible research and provide a definitive benchmark, we introduce and share the* ***Golden Set****, a new, meticulously curated, and manually annotated dataset for PAVE in Portuguese. We detail the creation process and structure (Entity, Category, Subcategories) of this high-quality reference set. Our experiments conclusively show that AI-PAVE-Br, leveraging targeted prompt engineering, dramatically outperforms conventional Named Entity Recognition (NER) baselines. This work not only delivers a superior, scalable solution for a major non-English market but also enriches the NLP community with a valuable, publicly available resource for future PAVE research.*

1. Introduction

In the rapidly evolving landscape of e-commerce, the ability to accurately and consistently extract structured information from vast quantities of unstructured product data is paramount [Wasilewski 2024, Brinkmann et al. 2024a, Yang et al. 2023, Gong and Eldardiry 2024]. Product descriptions, often free-form text, contain a wealth of valuable attributes (e.g., brand, model, features) that are crucial for various downstream tasks such as search, filtering, recommendation, and classification. The lack of standardized attributes leads to poor user experience and inefficiencies in data management [Wasilewski 2024, Gong and Eldardiry 2024].

This paper addresses the challenge of Product Attribute Value Extraction (PAVE) within the context of managing and enhancing data quality in dynamic, large-scale product catalogs, with a focus on Brazilian e-commerces. As product data continues to grow and evolve in complexity, traditional methods often prove insufficient for maintaining the required level of data accuracy and consistency. To address these inherent limitations, continuous improvement of product data management and advanced AI capabilities is essential. This commitment necessitates the establishment of robust quality metrics and reliable evaluation benchmarks for intelligent systems, such as AI-PAVE-Br proposed in this paper, particularly concerning product classification and attribute extraction.

Traditional approaches to information extraction, often relying on rule-based systems or statistical Named Entity Recognition (NER) models, face significant limitations when dealing with the highly diverse and dynamic nature of e-commerce product data [Wasilewski 2024, Gong and Eldardiry 2024]. These methods frequently struggle with generalization, require extensive feature engineering, and are brittle in the face of new product types or linguistic variations.

To overcome these challenges, we explore the application of Large Language Models (LLMs) for PAVE within the context of Brazilian e-commerces. LLMs, with their advanced understanding of natural language and generative capabilities, offer a promising alternative for extracting complex, context-dependent attributes [Brinkmann et al. 2024a, Yang et al. 2023]. A cornerstone of our evaluation methodology is the development of a *Golden Set*, a meticulously annotated Portuguese dataset that provides a ground truth for assessing model performance objectively within the context of Brazilian e-commerces. This paper details the construction and utility of this Golden Set and presents our approach to PAVE using LLMs especialized in e-commerces present in the Brazilian market, demonstrating its effectiveness against traditional methods within this context.

2. Related Work

Information extraction from unstructured text has been a long-standing challenge in Natural Language Processing (NLP). Early approaches to attribute value extraction in e-commerce often relied on rule-based systems, regular expressions, and dictionaries tailored to specific product categories [Sugiyama et al. 2010]. While effective for narrow domains, these systems are labor-intensive to maintain and difficult to scale.

More advanced methods moved towards statistical machine learning models for NER, treating attribute values as specific types of entities [Luo et al. 2011]. Conditional Random Fields (CRFs) [Lafferty et al. 2001] and Support Vector Machines (SVMs) [Joachims 1999] were widely applied, requiring large amounts of labeled data and extensive feature engineering. Deep learning models, particularly Recurrent Neural Networks (RNNs) and transformer-based architectures, have significantly advanced NER performance, achieving state-of-the-art results across various benchmarks [Devlin et al. 2019a, Vaswani et al. 2017]. These models, while powerful, still often require fine-tuning on domain-specific datasets.

The emergence of LLMs like GPT-3 [Brown et al. 2020] and its successors has revolutionized NLP, with significant work exploring their use for information extraction [Neuberger et al. 2025, Kim et al. 2024] and product-related tasks [Srinivas et al. 2024, Fang et al. 2024]. However, the vast majority of this research has been overwhelmingly concentrated on English-language corpora and benchmarks. This Anglocentric focus creates a significant gap, as models and methods developed for English do not readily transfer to the distinct linguistic and commercial landscape of other major markets. The Brazilian e-commerce ecosystem, for instance, is characterized by unique product terminologies, colloquialisms, and structural conventions in Portuguese that demand a tailored approach. Our work directly confronts this challenge. We pioneer the application of prompt-engineered LLMs to PAVE specifically for the Brazilian market, introducing a specialized system (**AI-PAVE-Br**). Furthermore, to enable rigorous and reproducible evaluation in this new domain, we developed and are releasing the **Golden Set**, the first high-quality, manually-annotated benchmark for PAVE in Portuguese.

Recent research has extensively explored the application of LLMs for complex information extraction tasks, including PAVE. Works such as those by Brinkmann et al. [Brinkmann et al. 2024b, Brinkmann et al. 2025] directly investigate the use of LLMs for extracting and normalizing product attributes, highlighting their potential for robust attribute identification and value extraction. Similarly, [Sabeh et al. 2024] delve into the use of LLMs for product attribute value identification, reinforcing the growing interest in this area.

Nonetheless, despite two recent studies of [Abilio et al. 2024] and [Silva et al. 2021] exploring the use of Small Language Models (SLMs), such as BERT and its variations adapted to Brazilian Portuguese, there still has a gap in the literature for the context of Brazilian e-commerces as most recent research is focused on American or international e-commerces. Consequently, our work builds upon this foundation by demonstrating the effectiveness of prompt-engineered LLMs for PAVE in a real-world, large-scale Brazilian e-commerce context. Beyond basic extraction, studies are exploring more advanced LLM applications in this domain. [Brinkmann and Bizer 2025] propose self-refinement strategies for LLM-based PAVE, indicating a move towards more autonomous and accurate extraction systems.

3. The Golden Set: A Foundation for Evaluation

A fundamental component of our quality assessment framework for AI-PAVE-Br's AI models is the *Golden Set*. In the context of data science and machine learning, a Golden Set (also known as a Gold Standard or Ground Truth) is a collection of data expertly annotated to serve as an objective benchmark for evaluating the quality and performance of AI models [Adamson and Welch 2019]. By comparing a model's predictions against the labels in the Golden Set, development teams can quantitatively assess model quality without having to infer business rules or rely on subjective judgments.

3.1. Construction and Scope

The annotation of our *Golden Set* was meticulously carried out by a dedicated team of 12 trained annotators, with a specific focus on the Brazilian e-commerce context. This manual process ensures a high level of accuracy and domain expertise [Brinkmann et al. 2024b, Brinkmann et al. 2025]. The *Golden Set* is designed to evaluate both product classification and attribute extraction models within the scope of e-commerce platforms operating in Brazil. It also serves as a benchmark for comparing the performance of our proposed AI-PAVE-Br against that of an existing baseline system.

The product selection for the *Golden Set* was strategically designed to ensure a balanced and representative sample within the Brazilian e-commerce context. The selection covers key categories such as: air conditioner, television, cell phone, refrigerator, notebook, tire, wardrobe, bed, sneaker, stove, table and chair set, backpack, faucet, headphone, perfume, doll, motorcycle helmet, pot, lamp, and cell phone case.

3.2. Annotation Schema and Data Structure

For each selected product, the annotation team meticulously annotated the following attributes:

- **Entity (Tipo de Produto):** A single string representing the most granular product type (e.g., `'Ar Condicionado'`,`'Perfume'`).
- **Category (Categoria):** A single string representing the broader product category (e.g., `'AR'` for Air Conditioner, `'PF'` for Perfume).
- **Subcategories (Subcategoria):** A list of strings detailing more specific subcategories or attributes relevant to the product (e.g., `['ARCA', 'ACIV', 'ARAR']` for an Air Conditioner, which might denote “Cassette Air Conditioner”, “Inverter Air Conditioner”, “Residential Air Conditioner”).

An example of an annotated product is:

- **Product Title:** `Ar Condicionado Cassete LG Round 36000 BTU/h Quente e Frio Monofásico AT-W36GYLP1 220 Volts`
- **Annotated Entity:** `'Ar Condicionado'`
- **Annotated Category:** `'AR'`
- **Annotated Subcategories:** `['ARCA', 'ACIV', 'ARAR']`

Therefore, beyond merely having a ground truth of attribute values for validation, a clear definition of the list of attributes to be extracted for each product type is essential. Table 1 illustrates several entities, their corresponding lists of attributes, and the annotation wave in which these entities were processed. Entities marked in the first annotation, and these entities formed the basis for initial experiments.

3.3. Sampling Methodology

To ensure the statistical validity and reduce selection bias, products were sampled randomly while considering existing classifications. This approach aimed to balance the sample set, accounting for the disparate volumes of certain product types in AI-PAVE-Br’s search index (e.g., 2 million “Tênis” vs. 17 thousand “Refrigerador”). The required sample size for each of the 20 product types was determined using Cochran’s formula for a large population:

$$n = \frac{Z^2 p(1-p)}{e^2} \tag{1}$$

where:

- n is the target sample size.
- Z is the Z-score corresponding to the desired confidence level.
- p is the estimated proportion of an attribute in the population.
- e is the desired margin of error.

Following the standard conservative approach, we used a 95% confidence level ($Z = 1.96$), a maximum variance assumption ($p = 0.5$), and a margin of error of 5% ($e = 0.05$). This calculation yields a sample size of $n \approx 385$, which we applied to each product type.

Table 1. Product Entities and Their Associated Attribute Lists (Wave 1 Annotation)

Entity	Attribute List
Ar Condicionado	Marca, Capacidade de Refrigeração, Tipo, Tecnologia, Ciclo, Potência, Voltagem
Fogão	Marca, Cor, Quantidade de Bocas, Instalação, Alimentação, Tipo de Acendimento, Tipo de Gás, Tipo de Forno, Voltagem
Guarda-Roupa	Marca, Cor, Tamanho, Tipo, Tipo de Porta, Quantidade de Portas, Quantidade de Gavetas, Material da Estrutura
Lavadora de Roupas	Marca, Cor, Capacidade de Lavagem, Voltagem
Notebook	Marca, Processador, Memória RAM, Capacidade do HD, Capacidade do SSD, Tamanho da Tela, Sistema Operacional, Capacidade da Placa de Vídeo, Resolução da Tela
Pneu	Marca, Largura do Pneu, Altura do Pneu, Aro, Quantidade de Pneus, Tipo de Veículo, Tipo de Terreno
Refrigerador	Marca, Cor, Material, Quantidade de Portas, Capacidade Líquida Total, Tipo de Degelo, Voltagem
Sofá	Marca, Cor, Tipo, Quantidade de Lugares, Revestimento, Tipo de Encosto, Tipo de Assento
TV	Marca, Polegadas, Resolução, Tecnologia, Conectividade, Sistema Operacional, Tipo de Tela, Assistente Virtual

It is important to note that the manual annotation of product data involves significant time and financial costs. For this reason, it was necessary to statistically limit the sample size to what is sufficient to ensure reliable results, while maintaining project feasibility. The complete Golden Set dataset, containing a comprehensive collection of annotated products, is available at `https://github.com/ai-luizalabs/AI-PAVE-Br`.

4. Product Attribute Value Extraction (PAVE) with LLMs

As already mentioned, PAVE is the task of identifying and extracting specific attributes and their corresponding values from unstructured product descriptions or titles. Unlike general NER, which focuses on predefined entity types (e.g., persons, locations), PAVE often deals with a wider, more dynamic range of attributes that are highly dependent on the product category.

4.1. Shifting from Traditional NER to LLM-based PAVE

Traditional NER methods, while effective for extracting basic entities, struggle with the nuances of product data, such as:

- **Contextual Dependence:** A `'10-inch'` value needs to be associated with a `'screen size'` attribute.
- **Novel Attributes:** New product features constantly emerge, requiring models to adapt.
- **Relationship Extraction:** Identifying not just values, but also their corresponding attributes (e.g., `'Bluetooth 5.0'` where `'Bluetooth'` is the attribute and `'5.0'` is the value).

To overcome these challenges, we utilized Google's Gemini 1.5 Flash model for our LLM-based PAVE approach with a meticuosly created prompt to further adapt the model for the Brazilian e-commerce context. This model, known for its efficiency and strong performance across various

NLP tasks, allows for sophisticated semantic understanding and generative capabilities. Its large context window enables it to process extensive product descriptions and titles effectively. The main rationale behind our choice for our LLM-based PAVE approach is that LLMs address the limitations of NER approaches by inherently understanding context, possessing vast world knowledge, and exhibiting strong few-shot or zero-shot learning capabilities. Their generative nature allows them to not just classify, but also to generate structured outputs directly.

Given the vast number of products in our catalog, it is infeasible to annotate every single item. Therefore, defining a statistically significant sample size for validation of the proposed solution is essential. As already stated, to determine the sample sizes, we employed **Cochran's formula** for finite populations. Table 2 presents the total number of products for each entity and the corresponding calculated sample size.

Table 2. Product Entities and Their Calculated Sample Sizes

Entity	Total Products (N)	Sample Size (n)
Ar Condicionado	14,914	374
Fogão	11,880	372
Guarda-Roupa	144,859	383
Lavadora de Roupas	4323	352
Notebook	11,574	371
Pneu	10,041	382
Refrigerador	9330	369
Sofá	105,785	382
TV	7457	365

4.2. Prompt Engineering for PAVE

Our approach to PAVE leverages the power of LLMs through targeted prompt engineering. Instead of fine-tuning large models (which is computationally expensive and requires significant data), we craft specific prompts that guide the LLM to perform the desired extraction task. The prompt defines the task, provides context, specifies the desired output format, and can include few-shot examples if necessary. For our PAVE system, we developed tailored prompts for each type of attribute extraction (Entity). A typical prompt structure might include:

1. **Instruction:** Clearly define the task. "Extract the attributes <<list of attributes>> of the <<product type>> from the following product title, description, technical data sheet, and additional information.
2. **Context:** Provide the product title, description, technical data sheet, and additional information.
3. **Output Format Specification:** Define the desired structured output (e.g., JSON format with keys). This is crucial for consistent parsing.
4. **Examples (Few-Shot):** Include one or more input-output pairs to guide the model's understanding of the task and desired output style.

This specific prompting strategy allows the LLM to act as a highly intelligent parser and extractor, adapting its vast pre-trained knowledge to our specific domain requirements.

5. Experimental Setup and Results

Our evaluation focused on comparing the performance of our LLM-based PAVE system against traditional methods, using the Golden Set as the ground truth.

5.1. Dataset and Metrics

The Golden Set, comprising manually annotated products across 20 diverse categories, served as our primary evaluation dataset. For each product, we extracted attribute *key–value* pairs according to schema.

The performance metrics used were standard in information extraction:

Precision, The proportion of correctly extracted values among all extracted values; **Recall**, The proportion of correctly extracted values among all true values in the Golden Set; and **F1-score**, The harmonic mean of precision and recall, providing a balanced measure of performance. For multiple items could be present, we typically calculated set-based F1-score (exact match of the list or token-level F1 across all values). Moreover, regular expressions were employed to standardize some of the results. However, normalizing the extractor's output remains a key challenge in this type of AVE, presenting a significant obstacle for future work.

5.2. Comparison with the Traditional Method

Our baseline for comparison includes three parallel annotations per attribute in the *Golden Set*: values extracted by the **traditional system**, predictions from the **AI-PAVE-Br**, and **human-annotated values**, which serve as the gold standard. Human annotations were performed following a formal guideline, specifying the source of each value and whether external references were consulted. This setup enables a consistent and realistic comparison across methods.

5.3. Discussion of Results

The comprehensive evaluation of entity prediction performance is presented in Table 3 (F1-score) and Table 4 (Coverage). Our analysis compares a traditional baseline system (**Traditional Baseline**) with our advanced LLM-based PAVE approach (**AI-PAVE-Br**), which leverages product offer titles for enhanced context. This AI-PAVE-Br model represents the best performing LLM setup from our experiments.

Observing Table 3, a substantial improvement in F1-score is evident when transitioning from the Traditional Baseline to the LLM-based PAVE approach. The mean F1-score rose from **59.79** for the Traditional Baseline to **74.68** for AI-PAVE-Br. This overall trend highlights the superior semantic understanding and generalization capabilities of LLMs in extracting product entities. For a majority of entities (e.g., "Guarda-Roupa", "Notebook", "Pneu", "Sofá", "Armário", "Bicicleta", "Cadeira", "Celular", "Conjunto de Mesa e Cadeira", "Frigideira", "Desktop"), AI-PAVE-Br demonstrates a significant improvement in F1-score compared to the Traditional Baseline. While this enhancement is observed across most categories, it is also important to note instances where AI-PAVE-Br might exhibit lower F1-scores than the Traditional Baseline (e.g., "Lavadora de Roupas", "Refrigerador", "Tênis", "Camiseta"), indicating specific challenges or nuances within those categories that could benefit from further specialized prompt engineering or fine-tuning. These performance gaps are largely attributed to the quality, heterogeneity, and structural variability of the input text data, which directly affect the performance of prompting-based LLM approaches in PAVE. A central challenge lies in the lack of normalization and consistency in how attribute values are expressed. The same product model might appear as "WD11M4453JW/AZ", "Samsung WD11M Washer", "Samsung 11kg Inverter", or simply "WD11" making precise mapping difficult. Similar inconsistencies affect attributes such as dimensions or weight — e.g.,

Table 3. F1-Score for Entity Prediction Across Product Types

Entity	Traditional Baseline (F1)	LLM-based PAVE (AI-PAVE-Br) (F1)
Ar Condicionado	82.24	84.68
Fogão	76.63	75.69
Guarda-Roupa	49.74	71.05
Lavadora de Roupas	92.12	82.75
Notebook	73.44	93.19
Pneu	50.69	84.86
Refrigerador	80.28	74.61
Sofá	41.10	61.88
TV	75.06	76.16
Armário	57.41	77.31
Bicicleta	35.19	70.98
Cadeira	27.65	70.88
Caixa de Som	42.45	56.14
Celular	65.02	85.67
Conjunto de Mesa e Cadeira	42.49	70.33
Freezer	72.23	78.25
Frigideira	48.45	75.47
Microondas	85.53	85.45
Tênis	65.66	49.68
Desktop	37.26	88.75
Camiseta	54.73	53.57
Mean	**59.79**	**74.68**

"60x85x60 cm", "Height: 85cm; Width: 60cm" or "60 (W) x 85 (H) x 60 (D)" — all semantically equivalent but lexically diverse. In addition, multivalued or semantically ambiguous attributes introduce further complexity. Voltage may appear as "110V or 220V", "bivolt", "110/220V". In some cases, the product is truly bivolt, while in others it has a fixed voltage but the seller lists multiple values. Resolving such ambiguity often requires going beyond catalog data, consulting additional sources such as product manuals or official manufacturer specifications to verify whether the product is not actually bivolt but rather offered in different versions, and ensuring that the information is represented in a clear and normalized form. A similar problem exists for attributes like SIM card capacity, expressed in highly variable forms such as "dual chip", "2 chips", "dual SIM", "chip duplo". Likewise, color or size attributes are frequently presented as unordered lists ("black, blue, and gray"; "S, M, L, XL") with no explicit indication of which value corresponds to the item being sold. The lack of semantic anchoring in such cases makes it difficult to extract a single, reliable value and normalize it to a canonical form.

These challenges illustrate the inherent limitations of relying solely on textual cues for precise attribute extraction and normalization. However, despite such complexities, LLM-based systems demonstrate a remarkable ability to generate predictions across a wide variety of inputs. Table 4 provides insights into the models' ability to provide a prediction for each product. **Coverage** refers to the percentage of items in the dataset for which a model successfully generates a non-empty prediction, regardless of whether that prediction is correct. A higher coverage indicates

Table 4. Coverage for Entity Prediction Across Product Types

Entity	Traditional Baseline	LLM-based PAVE (AI-PAVE-Br)	Golden Set (Overall Coverage)
Ar Condicionado	71.91	82.12	94.80
Fogão	61.94	63.06	88.28
Guarda-Roupa	31.82	78.46	85.39
Lavadora de Roupas	89.77	76.10	98.47
Notebook	47.77	76.18	77.47
Pneu	31.41	94.66	84.02
Refrigerador	72.83	64.03	88.08
Sofá	35.97	77.88	83.42
TV	57.90	75.70	77.82
Armário	37.56	74.89	68.44
Bicicleta	16.70	53.66	61.05
Cadeira	20.25	69.88	82.49
Caixa de Som	41.59	73.62	67.80
Celular	45.97	76.37	86.21
Conjunto de Mesa e Cadeira	32.52	75.53	91.42
Freezer	57.03	71.04	92.43
Frigideira	37.27	77.04	93.03
Microondas	75.30	75.86	88.51
Tênis	52.04	43.16	94.40
Desktop	23.62	73.49	72.38
Camiseta	39.24	57.66	87.40
Mean	**46.71**	**71.96**	**83.96**

a more robust and comprehensive extractor, reducing instances where no relevant information is found. The mean coverage significantly increased from **46.71%** for the Traditional Baseline to **71.96%** for AI-PAVE-Br. This demonstrates that the LLM-based approach is considerably more robust in consistently generating entity predictions across the diverse product catalog, effectively filling gaps where traditional methods fall short. While the coverage for AI-PAVE-Br is substantially higher than the baseline, it still lags behind the ideal "Golden Set Overall Coverage" of **83.96%**, which represents the maximum possible coverage based on the annotated data, suggesting opportunities for further improvement in model robustness.

The overall success of our PAVE approach with LLMs can be attributed to:

- **Semantic Understanding:** LLMs' ability to grasp the nuanced meaning of product descriptions, even with linguistic variations or colloquialisms.
- **Contextual Reasoning:** Leveraging the broader context of the product title to infer attributes that might not be explicitly stated.
- **Prompt Engineering:** The careful design of prompts allowed us to steer the LLM's generative capabilities towards our specific attribute extraction goals and desired output formats.
- **High-Quality Golden Set:** The availability of a precise and comprehensive Golden Set was indispensable for robust model training (if applicable for fine-tuning) and, more importantly, for unbiased and accurate evaluation of model performance.

When evaluating deployment options for LLM-based solutions, both cost and latency were carefully considered. For open-weight models, deployment via Virtual Machines (VMs)—e.g., using custom containers on self-managed infrastructure—offered more predictable latency and potentially lower costs for long-term, high-volume use cases, despite requiring additional setup and maintenance effort.

In contrast, closed-weight models such as Google's Gemini are only accessible through managed services. In internal tests with Gemini 2.5 Flash-Lite, we issued 1,000 concurrent requests, with an average of 350–400 tokens per request. The model showed an average response time of 1.3 seconds, with occasional spikes reaching up to 30 seconds under specific conditions. These fluctuations were likely influenced by system load and rate limiting, which are typical of managed environments.

6. Conclusion and Future Work

This paper pioneers a specialized approach to Product Attribute Value Extraction (PAVE) for the Brazilian market, a domain largely overlooked by mainstream NLP research. We introduce **AI-PAVE-Br**, an LLM-based system specifically engineered to navigate the unique linguistic and structural complexities of product descriptions in Portuguese. A cornerstone of our contribution is the **Golden Set**, a meticulously annotated and publicly shared dataset spanning 20 product types. This set stands as the first high-quality benchmark for PAVE in Portuguese, providing an invaluable and objective asset for validating our system and enabling future research.

Our experimental results are conclusive: **AI-PAVE-Br**, guided by targeted prompt engineering, dramatically outperforms traditional NER methods. The system excels not only at extracting core product entities but also demonstrates remarkable capability in identifying complex "Category" and "Subcategory" attributes directly from Portuguese product titles. This success underscores the transformative potential of LLMs to move beyond Anglocentric models and provide scalable, highly accurate solutions for a major global market. By delivering both a novel system and a foundational benchmark, our work paves the way for a new wave of research and development in e-commerce AI for the Portuguese-speaking world. Future work will focus on several key areas:

- **Adaptive Prompting:** Investigating dynamic prompt generation strategies or few-shot learning techniques to improve performance on long-tail product categories or newly emerging attributes without requiring extensive re-annotation.
- **Fine-tuned SLMs:** Exploring the performance of **AI-PAVE-Br** in comparison to fine-tuned SLMs, such as mBERT [Devlin et al. 2019b] and BERTimbau [Souza et al. 2020], in the context of Brazilian e-commerce.
- **Addressing Performance Dips:** Further analysis and targeted optimization for categories where LLM performance did not surpass or even fell below traditional baselines.
- **Output Normalization Challenges:** Developing robust methods for normalizing the extractor's output, as inconsistent formatting of extracted values remains a significant challenge for downstream applications.

Ultimately, the combination of high-quality golden data and advanced LLM techniques paves the way for a more efficient, accurate, and scalable product data infrastructure, critical for enhanced user experiences and operational efficiency in e-commerce.

References

Abilio, R., Coelho, G. P., and da Silva, A. E. A. (2024). Evaluating named entity recognition: A comparative analysis of mono- and multilingual transformer models on a novel brazilian corporate earnings call transcripts dataset. *Applied Soft Computing*, 166:112158.

Adamson, A. S. and Welch, H. G. (2019). Machine learning and the cancer-diagnosis problem — no gold standard. *New England Journal of Medicine*, 381(24):2285–2287.

Brinkmann, A., Baumann, N., and Bizer, C. (2024a). *Using LLMs for the Extraction and Normalization of Product Attribute Values*, page 217–230. Springer Nature Switzerland.

Brinkmann, A., Baumann, N., and Bizer, C. (2024b). Using llms for the extraction and normalization of product attribute values. In Tekli, J., Gamper, J., Chbeir, R., and Manolopoulos, Y., editors, *Advances in Databases and Information Systems*, pages 217–230. Springer Nature Switzerland.

Brinkmann, A. and Bizer, C. (2025). Automated self-refinement and self-correction for llm-based product attribute value extraction. *arXiv preprint arXiv:2501.01237*.

Brinkmann, A., Shraga, R., and Bizer, C. (2025). Extractgpt: Exploring the potential of large language models for product attribute value extraction. In *International Conference on Information Integration and Web Intelligence*, pages 38–52. Springer.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners (openai). `https://arxiv.org/abs/2005.14165`.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019a). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1*, pages 4171–4186.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019b). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North*. Association for Computational Linguistics.

Fang, C., Li, X., Fan, Z., Xu, J., Nag, K., Korpeoglu, E., Kumar, S., and Achan, K. (2024). Llm-ensemble: Optimal large language model ensemble method for e-commerce product attribute value extraction. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 2910–2914, New York, NY, USA. Association for Computing Machinery.

Gong, J. and Eldardiry, H. (2024). Multi-label zero-shot product attribute-value extraction. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 2259–2270. ACM.

Joachims, T. (1999). Making large-scale svm learning practical. advances in kernel methods-support vector learning. b. schokopt et al.

Kim, H., Kim, J.-E., and Kim, H. (2024). Exploring nested named entity recognition with large language models: Methods, challenges, and insights. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8653–8670.

Lafferty, J., McCallum, A., Pereira, F., et al. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Icml*, volume 1, page 3. Williamstown, MA.

Luo, F., Xiao, H., and Chang, W. (2011). Product named entity recognition using conditional random fields. In *2011 Fourth international conference on business intelligence and financial engineering*, pages 86–89. IEEE.

Neuberger, J., Ackermann, L., van der Aa, H., and Jablonski, S. (2025). A universal prompting strategy for extracting process model information from natural language text using large language models. In Maass, W., Han, H., Yasar, H., and Multari, N., editors, *Conceptual Modeling*, pages 38–55, Cham. Springer Nature Switzerland.

Sabeh, K., Kacimi, M., Gamper, J., Litschko, R., and Plank, B. (2024). Exploring large language models for product attribute value identification. *arXiv preprint arXiv:2409.12695*.

Silva, D. F., Silva, A. M. e., Lopes, B. M., Johansson, K. M., Assi, F. M., de Jesus, J. T. C., Mazo, R. N., Lucrédio, D., Caseli, H. M., and Real, L. (2021). *Named Entity Recognition for Brazilian Portuguese Product Titles*, page 526–541. Springer International Publishing.

Souza, F., Nogueira, R., and Lotufo, R. (2020). *BERTimbau: Pretrained BERT Models for Brazilian Portuguese*, page 403–417. Springer International Publishing.

Srinivas, M., Krishna Reddy, S. V., NM, M., and Miyazawa, H. (2024). Evaluation of chatgpt, gemini and llama-2 for e-commerce product attribute extraction. In *Proceedings of the 2024 10th International Conference on e-Society, e-Learning and e-Technologies (ICSLT)*, pages 43–48.

Sugiyama, A., Harumoto, K., Kawashima, M., and Matsumoto, Y. (2010). Attribute value extraction from semi-structured web documents. *IEICE transactions on information and systems*, 93(10):2626–2633.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Wasilewski, A. (2024). Functional framework for multivariant e-commerce user interfaces. *Journal of Theoretical and Applied Electronic Commerce Research*, 19(1):412–430.

Yang, L., Wang, Q., Wang, J., Quan, X., Feng, F., Chen, Y., Khabsa, M., Wang, S., Xu, Z., and Liu, D. (2023). MixPAVE: Mix-prompt tuning for few-shot product attribute value extraction. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9978–9991, Toronto, Canada. Association for Computational Linguistics.