Evaluating LLM-Based Resume Information Extraction: A Comparative Study of Zero-Shot and One-Shot Learning Approaches in Portuguese-Specific and Multi-Language LLMs

Arthur Rodrigues Soares de Quadros[1,4][a], Wesley Nogueira Galvão[2,4][b], Victória Emanuela Alves Oliveira[3,4][c], Alessandro Vieira[4][d] and Wladmir Cardoso Brandão[1,4][e]

[1]*Department of Computer Science, Pontifical Catholic University of Minas Gerais (PUC Minas), Brazil*

[2]*Department of Computer Science, Federal University of São Carlos (UFSCar), São Carlos, SP, Brazil*

[3]*Department of Computer Science, Federal University of Technology, Paraná (UTFPR), Campo Mourão, PR, Brazil*

[4]*Data Science Laboratory (SOLAB), Sólides S.A., Belo Horizonte, MG, Brazil*

Keywords: Large Language Models (LLMs), Information Extraction, Resume Screening, Zero-Shot Learning, One-Shot Learning, Prompt Engineering, LLM-as-a-Judge.

Abstract: This paper presents a comprehensive evaluation of Large Language Models (LLMs) in the task of information extraction from unstructured resumes in Portuguese. We examine six models, including both multilingual and Portuguese-specific variants, using 0-shot and 1-shot prompting strategies. To assess accuracy, we employ two complementary metrics: cosine similarity between model predictions and ground truth, and a composite LLM-as-a-Judge metric that weights factual information, semantic information, and order of components. Additionally, we analyze token cost and execution time to assess the practicality of each solution in production environments. Our results indicate that Gemini 2.5 Pro consistently achieves the highest accuracy, particularly under 1-shot prompting. GPT 4.1 Mini and GPT 4o Mini provide strong cost-performance trade-offs. Portuguese-specific models like Sabiá 3 achieves high average accuracy specially on 0-shot considering the cosine similarity metric. We also demonstrate how the inclusion of sections frequently missing in real resumes can significantly distort model evaluation. Our findings help determine model selection strategies for real-world applications involving semi-structured document parsing in a context of resume information extraction.

1 INTRODUCTION

Resume screening is a time-consuming task for Human Resources (HR) professionals (Aggarwal et al., 2021). To enable HR to focus on more strategic activities, there is a growing need for automation in this area (Balasundaram and Venkatagiri, 2020). Recent advancements in Natural Language Processing (NLP) models and Large Language Models (LLMs) have opened up new possibilities for leveraging highly capable generative AI models. These models offer a more robust approach compared to rule-based regular expressions, which can become overly complex when handling unstructured documents like resumes (Li et al., 2008).

Document information extraction (IE) typically relies on two primary methods: regular expressions and NLP approaches. Regular expressions employ a set of rules to search for specific string patterns within a sentence. This approach is well-suited for well-structured sentences or documents, as a series of regular expressions can effectively extract key information from predefined patterns (Li et al., 2008). In contrast, NLP approaches are more intricate. They involve generating numerical vectors from natural language sentences, enabling computers to interpret them. Each sentence is transformed into a sequence of numbers, which are then subjected to statistical calculations to analyze their syntax and semantics (Chowdhary and Chowdhary, 2020).

Several studies employ LLMs for IE in multiple contexts. In the context of Open Information Extraction for Portuguese, (Cabral et al., 2024) and (Melo et al., 2024) both propose comprehensible frameworks capable of extracting structured content from

[a] https://orcid.org/0009-0004-9593-7601
[b] https://orcid.org/0009-0001-8545-3126
[c] https://orcid.org/0009-0000-2777-4581
[d] https://orcid.org/0000-0002-9921-3588
[e] https://orcid.org/0000-0002-1523-1616

Soares de Quadros, A. R., Galvão, W. N., Oliveira, V. E. A., Vieira, A. and Brandão, W. C.
Evaluating LLM-Based Resume Information Extraction: A Comparative Study of Zero-Shot and One-Shot Learning Approaches in Portuguese-Specific and Multi-Language LLMs.
DOI: 10.5220/0013837900003985

In *Proceedings of the 21st International Conference on Web Information Systems and Technologies (WEBIST 2025)*, pages 313-324
ISBN: 978-989-758-772-6; ISSN: 2184-3252

any unstructured text, in a structure of tuples providing relationships between objects. (Cosme et al., 2024) reviews recent studies on IE, providing a systematic analysis on studies similar to ours in a multitude of research areas.

Although several studies explore the use of LLMs for information extraction when compared to traditional techniques, we lack detailed studies comparing LLMs on Portuguese-specific settings while using Brazilian LLMs to compare with multi-language LLMs in IE. Hence, this study explores information extraction on Portuguese resumes with multiple LLMs: ChatGPT, Google Gemini, and the Brazilian LLM Sabiá. The dataset used for information extraction was a set of 25 Portuguese resumes potentially containing information displayed in Table 1 in any order, with or without missing values. These resumes were collected as part of job application processes, and might contain null values for specific instances or several instances of a single section. We detail the dataset in Section 4.

This study aims to determine how effective are different LLMs and prompts on the structured information extraction of Portuguese resumes. To determine this, we explored multiple LLMs using zero-shot and one-shot approaches of information extraction on Portuguese resumes, evaluating its quality with a simple cosine similarity approach and a LLM-as-a-Judge approach for measuring accuracy. Both the example used in the prompt and the validation of results (ground truth) were conducted manually. In this study, we conduct a novel study comparing performance of resume information extraction tasks by Portuguese-specifc and multi-language LLMs. Our contributions with this study are as follows:

- **LLMs Perfomance Assessment:** We conduct a direct performance assessment of information extraction tasks by LLMs in a Portuguese setting, comparing multi-language and Portuguese-specific LLMs in zero-shot and one-shot settings.
- **IE Cost Measurement:** We measure effectiveness of each LLM model not only by accuracy, but also by computing time and monetary cost.

The remaining of this study is organized as follows: Section 2 contains a general background of LLM and Generative AI; Section 3 makes a direct analysis of similar studies to ours; 4 displays the methodology for this study; and Sections 5, 6 and 7 critically discuss our achieved results.

2 LARGE LANGUAGE MODELS AND GENERATIVE AI

Recent advancements in LLMs have empowered NLP projects to extract information from documents using generative approaches of exceptional quality (Xu et al., 2023). LLMs are NLP systems trained on vast datasets, leveraging various statistical methods to maximize data likelihood. They generate data that is highly probable, conditioned on a given data sequence X and additional information provided through a prompt (Xu et al., 2023).

For information extraction from unstructured text, LLMs offer a significant paradigm shift compared to traditional rule-based or machine learning approaches. Their ability to understand context, semantic nuances, and generate structured output directly from free-form text makes them particularly suitable for complex tasks like resume parsing, where the information is often semi-structured and highly variable. This capability is central to our work, as we aim to leverage these generative properties to accurately identify and extract key data points from Portuguese resumes.

We can categorize LLM-based solutions into three general groups based on how they utilize examples in the prompt:

- **Zero-Shot:** In this scenario, the model is tasked with addressing a problem without any prior exposure to solution examples. It relies solely on its general knowledge base to generate a response. For resume extraction, a zero-shot approach would involve instructing the LLM to identify specific fields (e.g., name, contact, education) without providing example resumes or their corresponding extracted data.
- **One-Shot:** In this scenario, the model is presented with a single solution example and is expected to apply the learned concept to similar tasks. This could involve showing the LLM one resume and its extracted information, then asking it to process a new resume.
- **Few-Shot:** In this scenario, the model is given a few solution examples and needs to base its answers on them. This approach is often more robust for complex information extraction tasks, but can increase the overall cost because of the amount of input tokens.

The decision to employ zero-shot, one-shot, or few-shot learning depends on both the capabilities of the model and the complexity of the task itself. More sophisticated models may excel in zero-shot or one-shot scenarios, while complex tasks may need few-

Table 1: Explanation of each metric extracted in our study. To the left, we have the metric name, and to the right we have what the LLM should search for in the resumes for each key, with some being more straightforward than others.

Collected Metric	**Metric Composition**
Full Name	Full available name of the applicant.
Age / Date of Birth	Years of age, date of birth or both if available.
About	Text provided giving a brief biography and/or professional background of the applicant.
Contact	List of available e-mails and cellphone numbers.
Social Media	List of all social media links and personal website if available.
Marital Status	One of the possible "Single", "Married", "Divorced", etc.
Addresses	List of addresses comprised of street, neighborhood, city, state, country, and house/apartment number, if available.
Education	List of degrees related to formal education such as bachelors, masters and PhD's with information like degree, institution, period and associated link if available.
Work Experience	List of previous formal work experiences such as internships, part-time and full-time jobs with information like title, description (brief and detailed), company, period, and associated link if available.
Other Relevant Experience	List of relevant experiences that are not considered formal work or education and are not directly related to certificates, with title, description, institution/company, period, and associated link if available.
Other Courses or Certificates	List of certificates that are not related to formal education such as online platform certificates, with information like title, description, institution/company, period, and associated link if available.
Language Fluency	List of pairs language-proficiency containing each language and proficiency level cited in the resume.
Hard and Soft Skills	List of adjectives explicitly written in the resume, that can be considered a hard or soft skill.

shot learning to provide sufficient context (Chen et al., 2023).

2.1 Prompt Engineering

Employing more specific prompts related to task definitions significantly enhances the ability of LLMs to generate refined and contextually appropriate responses. By providing additional context within the prompt, the model gains a deeper understanding of the desired output, leading to improved content quality (Chen et al., 2023). In the context of information extraction from resumes, prompt engineering is crucial for defining the specific fields to be extracted, their desired format (e.g., JSON, YAML), and any constraints or instructions for handling missing or ambiguous data.

A variety of prompt engineering techniques can significantly enhance the capabilities of LLMs across numerous tasks. Techniques like Chain of Thought (CoT), where the LLM is prompted to show its reasoning steps before providing the final answer. Self-Consistency, Tree-of-Thoughts, and Graph-of-Thoughts are more advanced methods that can be employed to structure prompts effectively for even greater robustness (Sahoo et al., 2024).

2.2 LLM-as-a-Judge

The term LLM-as-a-Judge refers to the use of LLMs as evaluators for complex tasks (Gu et al., 2024). While human evaluations have a lower risk of failure, they are time-consuming, require considerable effort from specialists, and are costly to scale due to the limited availability of qualified evaluators.

This method offers a viable alternative to both human evaluations and traditional automated methods, providing distinct advantages in scalability, efficiency, and adaptability. LLM judges emulate the evaluation methods used by human judges but stand out for their sensitivity to the instructions specified in prompt models. During the evaluation process, the LLM judge generates textual decisions based on the presented case and converts them into quantitative metrics (Wei et al., 2024). Specifically, for resume extraction, the LLM judge receives the ground truth extracted information, and the LLM's extracted output. It then evaluates the correctness of the extracted fields, providing a quantitative score (in our case, 1 for correct, and 0 for incorrect) reflecting the quality of the extraction.

3 RELATED WORKS

Natural language is widely used nowadays, and extracting semantic information from it is crucial for deriving valuable insights (Grishman, 2015). IE plays a pivotal role in this process. While there is ongoing debate regarding the precise definition of NER (Marrero et al., 2013), it remains an essential component of IE's semantic focus. Various tools and methods, such as regular expressions and NLP frameworks, are employed to effectively extract information (Grishman, 2015).

Many studies propose information extraction frameworks on different document types (e.g., PDFs, websites), mostly using NER. (Carnaz et al., 2021) use NER and IE for criminal related documents. They use neural networks for automatically extracting rela-

tionships in criminal cases using a 5W1H IE method and then represent them in a graph structure. (Vieira et al., 2021) apply NER on the 1758 Portuguese Parish Memories manuscript. They use neural networks and manually annotate part of the dataset for evaluation. They provide an annotated dataset of the full manuscript enriched by their neural network. (Azinhaes et al., 2021) apply NER and IE for making a study on the army likeness on the Internet. This application is useful for understanding the reasoning for the current army reputation.

Notably, NLP and LLM approaches have recently emerged as powerful techniques for efficient IE (Xu et al., 2023). Several works propose the use of LLMs for IE. (Nguyen et al., 2024) explore the use of few-shot LLMs for skill extraction from unstructured texts. (Villena et al., 2024) propose employing zero-shot and few-shot LLMs to construct interactive prompts for NER, facilitating general information extraction from texts. (Herandi et al., 2024) combine supervised machine learning with LLMs to create an efficient NER system. Additionally, regular expressions can be a valuable tool for IE. Works like (G et al., 2023) and (Sougandh et al., 2023) integrate regular expressions with NLP to extract information from resumes.

(Perot et al., 2024) proposes a new methodology leveraging LLMs for information extraction from Visually Rich Documents (VRD), such as invoices, tax forms, pay stubs, receipts, and more. The approach enables the extraction of singular, repeated, and hierarchical entities, both with and without training data, ensuring accuracy, anchoring, and localization of entities within the document. With high efficiency, generalization capability, and support for hierarchical entities, the methodology proves promising for practical applications across various document processing scenarios. Additionally, LLMs are also being applied to the extraction of complex information from scientific texts. (Dagdelen et al., 2024), for instance, proposes an approach that combines joint named entity recognition with relation extraction, using fine-tuning techniques on LLMs. This strategy holds significant potential for building structured databases derived from scientific literature.

Regarding resume IE for the Portuguese language, (Werner and Laber, 2024) explores neural networks for ensuring a correct resume structure. They do not focus on resume information parsing itself, but provide methods for defining the correct information order of the resume from any initial file structure. Major sections, similar to ours, specially “Personal Information”, “Education”, and “Work Experiences”. They want to ensure a given resume in provided in the correct information order to standardize the input data for other IE tasks, such as ours. Similar to our study, (Barducci et al., 2022) proposes an end-to-end framework for NER and IE for Italian resumes. Their experiments are similar to ours with regards to structured content extraction from resumes for faster resume processing. They do not directly use LLMs for IE, as they create their own neural network for NER and IE.

We have studies using LLMs for information extraction in Portuguese. But most of them apply LLMs in the context of Open Information Extraction. (Melo et al., 2024) investigate types of LLM finetuning, FFT (Full Fine Tuning) and LoRA (Low Rank Fine Tuning) for OpenIE in models of different scales, evaluating its trade-offs. (Cabral et al., 2024) explore few-shot approaches to finetune LLMs for OpenIE in Portuguese-specific tasks, outperforming commercial LLMs in the process. (Cosme et al., 2024) reviews several studies of LLM finetuning for multiple IE tasks.

In English, (Li et al., 2021) uses a BERT-based approach on a dataset of 700 english resumes annotated using the BIO method, achieving 91.41% precision on average extracting information on the features of name, designation, location, skills, college name, degree, companies worked at, and years of experience. (Gan and Mori, 2023) uses few-shot prompts with 25, 50, and 100 examples with different templates, using the T5 model with the methods of Manual Template and Manual Knowledge Verbalizer, achieving an f1-score of 78% in the extraction with 100-Shot.

4 METHODOLOGY

In this section, we explain how the experiments in this study were made. Our general methodology works as displayed in Figure 1.

Our methodology essentially pass through all resumes executing both zero-shot and one-shot methods, and after, we measure extraction accuracies using both cosine and LLM-as-a-Judge metrics. Algorithm 1 shows the step-by-step process we took throughout the extraction and evaluation process.

Essentially, we calculate the cosine similarity for each section using 768-dimensional vectors (768 is the default vector size) for all extracted parts of the section (as a single resume might have multiple work experiences or educational milestones, each are individually encoded by serafim-335 (Gomes et al., 2024)). We also determine a flag of “correct” and “incorrect” with an independent LLM judge. The embeddings for the cosine similarities are determined

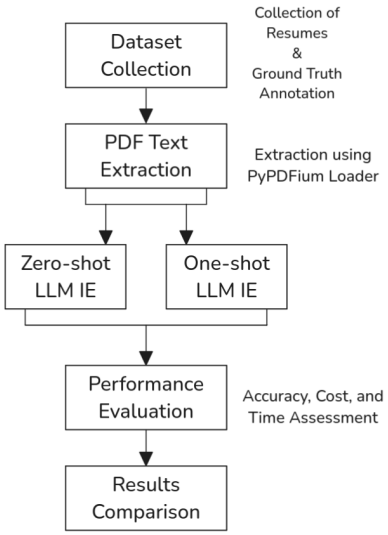

Figure 1: General methodology of the study. The text is extracted from a dataset of 25 resumes; then selected LLMs are applied both to zero-shot and one-shot prompts. After, we evaluate the extraction performance of each LLM using cosine similarity and LLM-as-a-Judge accuracy metrics.

Algorithm 1: Methodology Algorithm.

Input: Resumes R, Ground Truth GT, LLMs $LLMs$
Output: Extraction Accuracies Dictionary D
$L \leftarrow \{\}$;
$D \leftarrow \{\}$;
for *each* $r \in R$ **do**
  $text \leftarrow$ PyPDFium2(r);
  for *each* l *in* $LLMs$ **do**
    $E_{rl0} \leftarrow$ zero_shot($text$, l);
    $E_{rl1} \leftarrow$ one_shot($text$, l);
    $L \leftarrow \{L\} \cup \{E_{rl0}, E_{rl1}\}$;
for *each* E *in* L **do**
  $D[E_{cosine}] \leftarrow$ Cosine(E, GT_e);
  $D[E_{ai}] \leftarrow$ AI_as_a_Judge(E, GT_e);
return D

by the best performing state-of-the-art embedding for Brazilian Portuguese proposed in (Gomes et al., 2024) (serafim-335), while using the Qwen3:1.7b (Yang et al., 2025) model as the judge.

4.1 Dataset

The dataset used comprises 25 resumes in various formats. Each resume may contain different information about experiences, all of which are searched for, with missing information being denoted as null. This dataset was collected from recent job applications across diverse fields. Dataset statistics can be visualized in Table 2.

Table 2: Section and word counts across 25 resumes.

Section / Word Count	Total	Mean
Word Counts		
Words	80,849	3,234 ± 1,583
Resume Categories		
Name	25	1.00
Age / Date of Birth	16	0.64
About	17	0.68
Contact Information	25	1.00
Social Media	10	0.40
Marital Status	13	0.52
Addresses	18	0.72
Education	25	1.00
Work Experience	24	0.96
Other Relevant Experience	11	0.44
Other Courses / Certificates	15	0.60
Language Fluency	17	0.68
Skills	23	0.92

We observe frequent missing sections in the dataset, reflecting varied resume templates for LLM extraction. Among the 25 PDFs, 22 use unique layouts, ranging from one- or two-column formats, bullet points, or full paragraphs, with either explicit section labels (aligned with Table 1) or no clear divisions. This diversity enables evaluation across multiple input formats. The sample size of 25 was chosen to keep computational and manual annotation costs manageable while still enabling meaningful evaluation across different resume structures.

4.1.1 PDF Interpretation

The text content of the resumes was extracted using a document loader that processes PDF files[1]. Image-based content was ignored, and each page was extracted individually before being concatenated into a single text document. This resulting text was then incorporated into the prompts for IE.

4.2 LLMs Used

This study compared Google's Gemini 2.5 Pro (Comanici et al., 2025), and Gemini 2.5 Flash models with OpenAI's ChatGPT 4.1 Mini and ChatGPT 4o Mini (OpenAI, 2024), as multi-language LLMs. Both Gemini and ChatGPT are considered state-of-the-art language models and have consistently demonstrated strong performance across various tasks in numerous studies.

[1]For this we used PyPDFium2 (https://python.langchain.com/docs/integrations/document_loaders/pypdfium2/).

We also applied one Portuguese-specific LLM for information extraction: Sabiá 3.1 and Sabiá 3 (Pires et al., 2023). Sabiá is a Brazilian LLM trained on an extensive dataset in Brazilian Portuguese. This LLM showed great potential in comparison to ChatGPT, Claude, and Llama, with reduced costs while maintaining quality (Abonizio et al., 2024). Although we also have other Portuguese-specific LLMs, such as Tucano (Corrêa et al., 2024), we did not apply them because of their inherent constraints regarding the limits of the input and output tokens.

4.3 Prompt Engineering

We employed zero-shot and one-shot prompting techniques for each LLM model. The base prompt remained consistent, utilizing HTML notation to structure the following sections: Task (Information Extraction), consisting of Required Information to Extract (JSON keys), and Observation Notes (task details), Output Format (JSON), and Content (resume text). For one-shot prompts, additional sections for Input Example and Output Example were included, providing a concrete demonstration of the desired extraction task. A simplified version of the base prompt is presented below.

```
<Task>
Extract information from the text
of a resume provided after the tag
"Content". Necessary information:
* Name
* Age/Date of Birth
* About
* Contact Information:
  * Phone Numbers
  * E-mail addresses
* Social Media:
  * Name
  * Link
* Marital Status
* Addresses
* Education
* Work Experience
* Other Relevant Experience
* Other Courses or Certificates
* Language Fluency
* Skills:
  * Hard Skills
  * Soft Skills
Notes: {Notes or Details}
</Task>
<Output Format> JSON </Output Format>
<Example Input>
    {Example Input (if any)}
</Example Input>
<Example Output>
    {Example Output (if any)}
</Example Output>
<Content> {CV to be Extracted} </Content>
```

4.4 Evaluation Metrics

We employed two independent evaluation metrics to assess extraction accuracy: cosine similarity and LLM-as-a-Judge. Cosine similarity offers a nuanced evaluation by comparing the extracted text to the ground truth and calculating the average similarity across all sections. In contrast, the LLM-as-a-Judge metric adopts a "one-hot" approach, classifying each extraction as either correct (100% accuracy) or incorrect (0% accuracy) in three independent criteria: factual information, semantic information, and order. The final accuracy provided by LLM-as-a-Judge is the weighted average between all criteria, with 0.5 for factual information, 0.3 for semantic information, and 0.2 for order. As expected, the LLM-as-a-Judge metric tends to yield lower accuracy scores due to its stricter evaluation criteria.

The LLM-as-a-Judge is used with reasoning for each instance of all sections during the evaluation process. Below, we have a minimal example of answer for a single subsection of a resume.

```
<think>
First looking at the factual accuracy:
  The ground truth says "Pierre Lopes" and
the AI response exactly matches that.
  Since it's just comparing names - which
are objective facts - I should give 1 for
factual accuracy.
  Now for semantic accuracy: They're
identical so meaning is preserved perfectly.
No change in significance, so another
1 here as well.
  Finally checking order accuracy:
  The names are presented sequentially
without any particular order requirements
- just two words together. Since the
answer doesn't require specific ordering
of components, I can consider
this criterion met with a score of 1.
</think>
Factual: 1
Semantic: 1
Order: 1
```

From this response, we extract the numbers for each criterion, and get the weighted average with weights 0.5, 0.3, and 0.2 for factual, semantic, and order, respectively.

4.4.1 Cosine Similarity

Below we have the definition of both the cosine similarity (referred sometimes here as cosine accuracy)

and the average extraction accuracy in equations 1 and 2.

$$\cos(\theta) = \frac{A \cdot B}{||A||\,||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} B_i^2}} \quad (1)$$

$$\bar{y} = \frac{1}{M} \sum_{i=1}^{M} y_i \quad (2)$$

Equation 1 is used to calculate the cosine similarity between two vectors, A and B, representing the LLM-extracted text and the ground truth, respectively. The average similarity, $\bar{y}$, across all extracted sections, as calculated by Equation 2, serves as the overall extraction accuracy metric.

Extraction accuracy was aggregated by LLM and extraction metric for both zero-shot and one-shot approaches. This dual-level evaluation allowed for assessment of both the overall extraction quality of the metrics and the performance of the LLMs themselves.

To ensure optimal language representation, the vectorization of the CV section extractions was performed using the serafim-335 embedding (Gomes et al., 2024), state-of-the-art for Portuguese embeddings, specifically designed for the Portuguese language of the resumes. Serafim-335 vectorizes each major section extraction in a 768-dimensional vector, with the vector of the ground truth and extraction being compared for calculating the cosine similarity metric.

4.4.2 LLM-as-a-Judge

For each resume and CV section, we presented the ground truth, the full prompt with resume text, and the LLM-extracted section side-by-side. An adjusted predefined prompt was then used to query Qwen3:1.7b (Yang et al., 2025) to determine if the extracted section matched the ground truth in three independent criteria: factual information (names, dates, institutions need to be equal to ground-truth, and not missing), semantic information (be meaning needs to be equal, for example, “Bachelors of Science” and “BSc” are the same), and order (the sequence needs to be equal, for example, “April 2024, BSc” and “BSc, April 2024” are different, so it would result in 0.0). We chose Qwen3:1.7b because it is a capable yet light enough not to take too much time to run in a virtual machine. The virtual machine used for this evaluation contains 8 CPUs, 32 GB of RAM and a NVIDIA T4 GPU.

The LLM-as-a-Judge evaluation uses a detailed version of the following prompt.

```
You are evaluating the output of an AI model
by comparing it to a ground truth.
[BEGIN DATA]
************
[Section]: {section}
************
[Ground Truth Answer]: {correct_answer}
************
[AI Answer]: {ai_answer}
************
[END DATA]
Evaluate the AI answer using three
independent criteria, returning only "0"
(incorrect) or "1" (correct), with no
explanation, for each:
- Factual Accuracy: Objective details.
   Are Names, Dates, Institutions correct?
- Semantic Accuracy: Phrase Meaning.
   Is the overall meaning the same?
- Order Accuracy: Extracted Sequence.
   Is the order of extraction the same?
```

The LLM-as-a-Judge evaluation was used as the final accuracy of our experiments, as it contains a more nuanced approach for measuring accuracy of extraction than the cosine metric.

4.4.3 Statistical Significance

Due to non-normal accuracy distributions and unequal group sizes, we used the Kruskal-Wallis test to compare models. This was followed by Dunn’s post hoc test with Bonferroni correction to assess pairwise differences. We analyzed cosine scores per section across 25 resumes, totaling over 1,000 observations. All tests used a 5% significance threshold.

5 RESULTS AND DISCUSSION

5.1 Accuracy Metrics

Regarding accuracy, Figure 2 displays the average cosine similarity between extracted content and the manually extracted ground-truth, per section. Sections such as Name and Contact Information achieved values close to or equal to 1.0 across all models and configurations. In contrast, more open-ended sections like Other Relevant Experiences and About showed substantial variation across models. Gemini 2.5 Pro obtained the best overall results for 1-shot prompts, particularly in Education, Work Experience, and Skills, often exceeding 0.9 similarity. Sabiá 3.1 for 0-shot prompts showed notably lower performance in sections like Other Relevant Experiences, with values below 0.4.

Additionally, Figure 3presents results based on a composite metric that aggregates factual, semantic, and order-based accuracy into a weighted average of

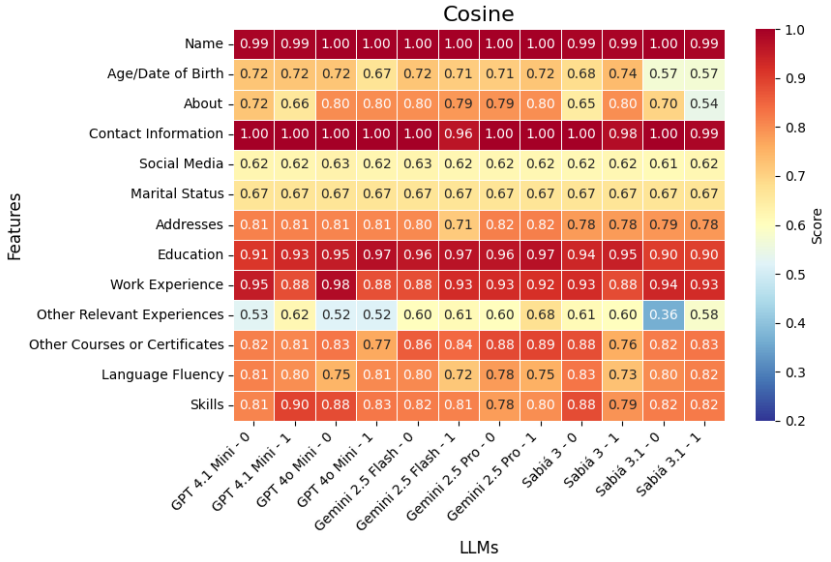

Figure 2: Accuracy per LLM, resume section for both 0-Shot and 1-Shot calculated with the Cosine Similarity metric.

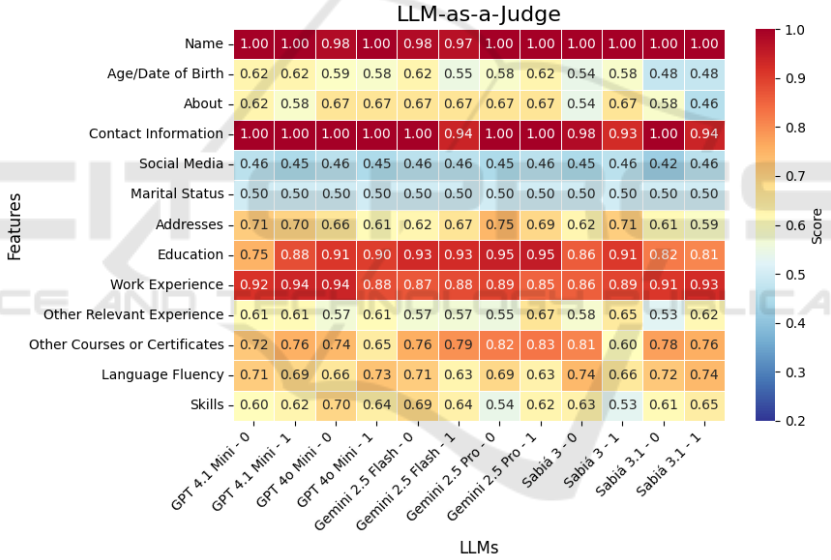

Figure 3: Accuracy per LLM, resume section for both 0-Shot and 1-Shot calculated with the LLM-as-a-Judge metric.

0.5, 0.3, and 0.2, respectively. The overall trends are similar to the cosine-based results but with sharper distinctions between models. Once again, GPT 4.1 Mini stands out as one of the top performers with the 1-shot prompt. Most models maintained high accuracy in objective sections but performed worse in descriptive or frequently absent sections.

It is important to note that, due to our methodology, missing sections in the resume were assigned an accuracy of 0.0, which significantly impacts the overall averages. Specially sections such as Social Media and Marital Status (around half missing). Other section often have between 5% and 50% missing. This leads to apparent poor performance in those frequently missing sections, but in the four particular sections that are always present: Name, Contact Information, Education, and Work Experience.

5.2 Cost Metrics

Figure 4 presents the total costs, in US dollars (USD), associated with input and output tokens during the extraction process performed by different LLMs using 0-shot and 1-shot prompting strategies.

We notice that the price for Gemini 2.5 Pro is naturally the highest, as this is technically the most power-

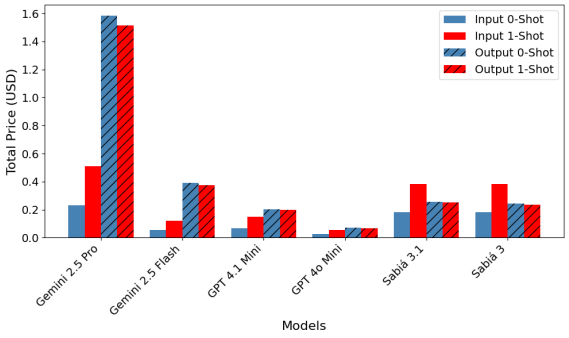

Figure 4: Total token prices in US Dollars for both 0-shot and 1-shot experiments.

ful model tested. The GPT 4.1 Mini and GPT 4o Mini are the cheapest models overall. This is expected as both are simplified models. And both Sabiá 3.1 and 3 contain essentially the same prices, as the input and output token prices for both models are the same.

Table 3 shows the average execution duration as well as a 1-sigma CI.

Table 3: Mean duration and 1-sigma confidence interval for each model and prompt configuration.

Model	Duration (s)	
	0-Shot	1-Shot
Gemini 2.5 Pro	43.57 ± 16.19	43.05 ± 14.29
Gemini 2.5 Flash	23.24 ± 6.42	26.11 ± 7.54
GPT 4.1 Mini	24.10 ± 9.20	28.49 ± 18.01
GPT 4o Mini	21.98 ± 7.20	21.79 ± 7.46
Sabiá 3.1	28.85 ± 9.99	28.37 ± 8.71
Sabiá 3	93.13 ± 53.37	86.44 ± 61.77

The Sabiá 3 model exhibited the longest times, surpassing 90 seconds, while the other models ranged between 20 and 45 seconds. The use of 1-shot prompting generally do not negatively affect the time needed for the experiments execution.

5.3 Aggregated Results

Table 4 displays aggregated accuracies for each model/prompt/metric groups including all features, while all missing values are set as 0.

Table 4: Average accuracy across all features using cosine similarity and LLM-as-a-Judge.

Model	Cosine		Judge	
	0-Shot	1-Shot	0-Shot	1-Shot
Gemini 2.5 Pro	0.811	0.818	0.722	0.731
Gemini 2.5 Flash	0.811	0.796	0.722	0.707
GPT 4.1 Mini	0.798	0.801	0.710	0.720
GPT 4o Mini	0.811	0.795	0.722	0.709
Sabiá 3.1	0.768	0.772	0.689	0.687
Sabiá 3	0.805	0.791	0.701	0.699

As shown in Table 4, when considering all features – including those that are frequently missing – average accuracy scores tend to be lower. This is expected, as our methodology assigns a score of 0.0 to any section that is missing in the resume. Under these conditions, the Gemini 2.5 Pro model achieves the highest overall accuracy for both metrics, with a cosine similarity of 0.818 and a Judge score of 0.731 under the 1-shot setting. GPT 4.1 Mini also performs competitively, particularly in the 1-shot setting with a cosine score of 0.801 and a Judge score of 0.720. The Sabiá models lag behind across both metrics and prompting strategies, with the lowest Judge scores observed in the Sabiá 3.1 configuration.

Table 5 displays aggregated accuracies for each model/prompt/metric groups excluding features containing mostly null values. While missing values are still set as 0, accuracies are higher because there are fewer null values present.

Table 5: Average accuracy excluding sparse features using cosine similarity and LLM-as-a-Judge.

Model	Cosine		Judge	
	0-Shot	1-Shot	0-Shot	1-Shot
Gemini 2.5 Pro	0.867	0.877	0.806	0.820
Gemini 2.5 Flash	0.866	0.856	0.813	0.793
GPT 4.1 Mini	0.853	0.865	0.789	0.813
GPT 4o Mini	0.864	0.845	0.812	0.801
Sabiá 3.1	0.831	0.857	0.795	0.804
Sabiá 3	0.882	0.835	0.808	0.771

Table 5 presents the same metrics excluding features with predominantly null values. As expected, removing these sparsely populated sections increases the average scores for all models. The differences are significant, with the cosine metric improving by approximately 5 to 6 percentage points, while the LLM-as-a-Judge metric is improved by 8 to 10 percentage points. Notably, Sabiá 3 shows a significant improvement in cosine similarity under the 0-shot setting, reaching 0.882 – the highest among all models in this filtered setup. Gemini 2.5 Pro still maintains the best performance overall in the 1-shot approach with LLM-as-a-Judge, reinforcing its strong extraction capabilities across present and consistently structured sections. Across both tables, 1-shot prompting generally leads to marginal gains in accuracy, although the improvements are not uniform across models or metrics.

Table 6 displays the best-performing models for each metric: Cosine Similarity, LLM-as-a-Judge Accuracy, Cost, and Execution Time.

Table 6 summarizes the best-performing models across the four key dimensions: accuracy (both cosine similarity and LLM-as-a-Judge), cost, and exe-

Table 6: Best-performing models by metric and prompt type ignoring mostly null features.

Metric	Best Model	
	0-Shot	1-Shot
Cosine	Sabiá 3	Gemini 2.5 Pro
LLM-as-a-Judge	Gemini 2.5 Flash	Gemini 2.5 Pro
Cost	GPT 4o Mini	GPT 4o Mini
Execution Time	GPT 4o Mini	GPT 4o Mini

cution time. All models achieve high accuracies overall when sparse features are not included in the calculations. In particular, Gemini 2.5 Pro consistently achieved high accuracy in both cosine and judge-based metrics, particularly with the 1-shot prompt strategy, and Sabiá 3 achieved the highest accuracy in the 0-shot setting with the cosine metric. On the efficiency side, as expected, GPT 4o Mini, being the smallest model, delivered the lowest total cost and fastest response times, regardless of prompt type. These results reinforce the trade-off between performance and resource consumption, with some models offering balanced outcomes while others specialize in either speed or accuracy.

5.4 Statistical Significance

We applied the Kruskal-Wallis test to the cosine similarity scores across all models and prompting strategies. The result was highly significant ($H = 269.97$, $p < 0.001$), indicating performance differences between groups. To identify which models differ, we ran Dunn's post hoc test with Bonferroni correction. Figure 5 shows the pairwise comparisons. Several model combinations exhibit significant differences ($p < 0.05$), especially between the Sabiá models and Gemini 2.5 Pro/GPT 4.1 Mini.

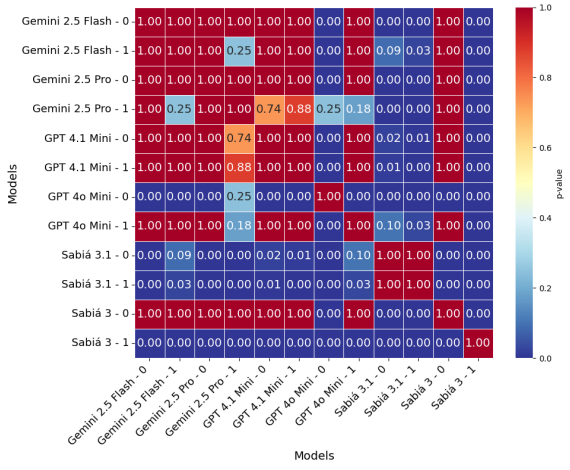

Figure 5: Post hoc Dunn's test (p-values, Bonferroni-corrected) comparing cosine similarity across model–prompt pairs.

5.5 Discussions

The results presented suggest several practical implications for production use. While 1-shot prompting generally yields slight improvements in accuracy, especially for stronger models like Gemini 2.5 Pro and GPT 4.1 Mini, the gains are modest and not always consistent across all metrics or models. Therefore, in resource-constrained scenarios or latency-sensitive environments, 0-shot prompting may still offer a favorable cost-performance trade-off, especially for models like GPT 4o Mini.

The comparison between Portuguese-specific models (Sabiá 3 and 3.1) and multilingual models highlights a clear gap in performance. While Sabiá 3 reached the highest cosine similarity in the 0-shot setting after filtering sparse features, its overall performance – especially under the LLM-as-a-Judge metric – remains behind that of multilingual models. This indicates that while language-specific models can excel in certain structured sections, they may still require improvements in general semantic understanding and reasoning consistency.

Regarding the inclusion of sparse features, our analysis shows that their presence can significantly lower average accuracy scores, due to the methodology assigning a score of 0.0 to missing sections. When these features (e.g., Social Media, Marital Status, About, Addresses) are excluded, accuracy metrics increase substantially. This highlights the importance of aligning evaluation metrics with realistic use cases: if certain sections are optional or rarely present in real data, including them in the evaluation may distort the perceived performance of LLMs.

In summary, the choice of model and prompting strategy should consider the trade-offs between accuracy, cost, and speed, as well as the nature of the expected input data. For production deployments that target structured, always-present fields, even mid-tier models may suffice with 0-shot prompts. However, for broader coverage and higher consistency, especially when handling semi-structured or descriptive fields, stronger models with 1-shot prompting may remain the best choice.

5.6 Ethical Considerations

The use of LLMs for resume information extraction raises important ethical concerns. Automated extraction pipelines may inadvertently perpetuate or amplify existing biases present in training data, particularly regarding gender, race, age, or disability. This is especially critical when models are used to support recruitment or selection decisions, where fairness

and transparency are paramount. Furthermore, the processing of personal documents like resumes must comply with data privacy regulations, such as LGPD or GDPR, ensuring informed consent, data minimization, and secure handling. Developers and practitioners should adopt fairness-aware modeling practices, audit outputs regularly, and ensure that model predictions do not become opaque filters in high-stakes human resource processes.

6 LIMITATIONS

Both our accuracy metrics does not account for weights in different sections, meaning, for example, "Name" and "Work Experience" accuracies both account for the same, even if both have completely different content both in structure and size. Also, our convention to when a specific section of a resume is empty in both (when there is no section content to compare), we treat it as 0.0 accuracy. This partially limit our assessment of the models' extraction, as we might undervalue or overvalue different sections. Our results might also be limited by the dataset used, as we did not explore open datasets for resume IE.

In order to reduce costs, our LLM-as-a-Judge approach does not take into account the response context (i.e., the resume content), meaning the judge can become limited in some cases. The evaluation by LLMs approach 3 different metrics, factual, semantic, and order information, but is still binary for each, in the sense of each metric being either 0 or 1. We did not explore more nuanced metrics for accuracy using LLM-as-a-Judge. Also to reduce costs, we did not explore the most advanced models of OpenAI, as prices for the preview of GPT 4.5 is 60 and 15 times higher than Gemini 2.5 Pro for input and output tokens, respectively.

7 CONCLUSION AND FUTURE WORKS

In this work, we evaluated the performance of six LLMs in extracting structured information from unstructured resumes written in Portuguese. We tested each model using both 0-shot and 1-shot prompts and applied two distinct accuracy metrics: cosine similarity and a weighted mean approach using LLM-as-a-Judge (with Qwen3:1.7b). Our experiments were conducted on 25 real-world resumes, and included a cost analysis of token consumption and execution time.

Our findings show that Gemini 2.5 Pro consistently outperformed other models in both accuracy metrics, particularly in the 1-shot setting. GPT 4.1 Mini also delivered competitive accuracy with significantly lower costs. The Sabiá models showed competitive results, with higher overall accuracy in some cases, but in some open-ended section, it showed lower overall accuracy in both metrics. A cost analysis highlighted GPT 4o Mini as the most economical option in both prompt settings, with faster execution times and reduced token usage. This result was expected as this model is the smallest tested. While Gemini 2.5 Pro and Flash are the heaviest models, and end up being more costly, but is still very fast, with the slowest model being Sabiá 3.

Future work may include expanding the dataset to cover more diverse resume formats and testing fine-tuned models specifically adapted to the task of resume IE. This analysis can provide valuable insights on how general LLMs compare to targeted models, designed for IE. We can also compare targeted models and LLMs with traditional extraction methods based on regular expressions and measure better the quality of recent techniques.

REFERENCES

Abonizio, H., Almeida, T. S., Laitz, T., Junior, R. M., Bonás, G. K., Nogueira, R., and Pires, R. (2024). Sabi\'a-3 technical report. *arXiv preprint arXiv:2410.12049*.

Aggarwal, A., Jain, S., Jha, S., and Singh, V. P. (2021). Resume screening. *International Journal for Research in Applied Science and Engineering Technology*, 134:66–88.

Azinhaes, J., Batista, F., and Ferreira, J. (2021). ewom for public institutions: application to the case of the portuguese army. *Social Network Analysis and Mining*, 11(1):118.

Balasundaram, S. and Venkatagiri, S. (2020). A structured approach to implementing robotic process automation in hr. In *Journal of Physics: Conference Series*, volume 1427, page 012008. IOP Publishing.

Barducci, A., Iannaccone, S., La Gatta, V., Moscato, V., Sperlì, G., and Zavota, S. (2022). An end-to-end framework for information extraction from italian resumes. *Expert Systems with Applications*, 210:118487.

Cabral, B., Claro, D., and Souza, M. (2024). Exploring open information extraction for portuguese using large language models. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pages 127–136.

Carnaz, G., Nogueira, V. B., and Antunes, M. (2021). A graph database representation of portuguese criminal-related documents. In *Informatics*, volume 8, page 37. MDPI.

Chen, B., Zhang, Z., Langrené, N., and Zhu, S. (2023). Unleashing the potential of prompt engineering in large language models: a comprehensive review. *arXiv preprint arXiv:2310.14735*.

Chowdhary, K. and Chowdhary, K. (2020). Natural language processing. *Fundamentals of artificial intelligence*, pages 603–649.

Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al. (2025). Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Corrêa, N. K., Sen, A., Falk, S., and Fatimah, S. (2024). Tucano: Advancing Neural Text Generation for Portuguese.

Cosme, D., Galvão, A., and Abreu, F. B. E. (2024). A systematic literature review on llm-based information retrieval: The issue of contents classification. In *Proceedings of the 16th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (KDIR)*, pages 1–12.

Dagdelen, J., Dunn, A., Lee, S., Walker, N., Rosen, A. S., Ceder, G., Persson, K. A., and Jain, A. (2024). Structured information extraction from scientific text with large language models. *Nature Communications*, 15(1):1418. Publisher: Nature Publishing Group.

G, G. M., Abhi, S., and Agarwal, R. (2023). A hybrid resume parser and matcher using regex and ner. In *2023 International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT)*, pages 24–29.

Gan, C. and Mori, T. (2023). A few-shot approach to resume information extraction via prompts. In *International Conference on Applications of Natural Language to Information Systems*, pages 445–455. Springer.

Gomes, L., Branco, A., Silva, J., Rodrigues, J., and Santos, R. (2024). Open sentence embeddings for portuguese with the serafim pt* encoders family. In Santos, M. F., Machado, J., Novais, P., Cortez, P., and Moreira, P. M., editors, *Progress in Artificial Intelligence*, pages 267–279, Cham. Springer Nature Switzerland.

Grishman, R. (2015). Information extraction. *IEEE Intelligent Systems*, 30(5):8–15.

Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., Wang, Y., and Guo, J. (2024). A survey on llm-as-a-judge. *ArXiv*.

Herandi, A., Li, Y., Liu, Z., Hu, X., and Cai, X. (2024). Skill-llm: Repurposing general-purpose llms for skill extraction. *arXiv preprint arXiv:2410.12052*.

Li, X., Shu, H., Zhai, Y., and Lin, Z. (2021). A method for resume information extraction using bert-bilstm-crf. In *2021 IEEE 21st International Conference on Communication Technology (ICCT)*, pages 1437–1442.

Li, Y., Krishnamurthy, R., Raghavan, S., Vaithyanathan, S., and Jagadish, H. (2008). Regular expression learning for information extraction. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 21–30.

Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato, J., and Gómez-Berbís, J. M. (2013). Named entity recognition: fallacies, challenges and opportunities. *Computer Standards & Interfaces*, 35(5):482–489.

Melo, A., Cabral, B., and Claro, D. B. (2024). Scaling and adapting large language models for portuguese open information extraction: A comparative study of fine-tuning and lora. In *Brazilian Conference on Intelligent Systems*, pages 427–441. Springer.

Nguyen, K. C., Zhang, M., Montariol, S., and Bosselut, A. (2024). Rethinking skill extraction in the job market domain using large language models. *arXiv preprint arXiv:2402.03832*.

OpenAI (2024). Gpt-4o system card.

Perot, V., Kang, K., Luisier, F., Su, G., Sun, X., Boppana, R. S., Wang, Z., Wang, Z., Mu, J., Zhang, H., Lee, C.-Y., and Hua, N. (2024). Lmdx: Language model-based document information extraction and localization. *ArXiv*.

Pires, R., Abonizio, H., Almeida, T., and Nogueira, R. (2023). Sabiá: Portuguese large language models. In *Anais da XII Brazilian Conference on Intelligent Systems*, pages 226–240, Porto Alegre, RS, Brasil. SBC.

Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., and Chadha, A. (2024). A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*.

Sougandh, T. G., Reddy, N. S., Belwal, M., et al. (2023). Automated resume parsing: A natural language processing approach. In *2023 7th International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)*, pages 1–6. IEEE.

Vieira, R., Olival, F., Cameron, H., Santos, J., Sequeira, O., and Santos, I. (2021). Enriching the 1758 portuguese parish memories (alentejo) with named entities. *Journal of Open Humanities Data*, 7:20.

Villena, F., Miranda, L., and Aracena, C. (2024). llmner:(zero— few)-shot named entity recognition, exploiting the power of large language models. *arXiv preprint arXiv:2406.04528*.

Wei, H., He, S., Xia, T., Wong, A., Lin, J., and Han, M. (2024). Systematic evaluation of llm-as-a-judge in llm alignment tasks: Explainable metrics and diverse prompt templates. *ArXiv*, abs/2408.13006.

Werner, M. and Laber, E. (2024). Extracting section structure from resumes in brazilian portuguese. *Expert Systems with Applications*, 242:122495.

Xu, D., Chen, W., Peng, W., Zhang, C., Xu, T., Zhao, X., Wu, X., Zheng, Y., Wang, Y., and Chen, E. (2023). Large language models for generative information extraction: A survey. *arXiv preprint arXiv:2312.17617*.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al. (2025). Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.