Evaluating Document Pre-processing Techniques for LLM-based Information Extraction

Matheus Leonel de Andrade[1], **Amir Youssef**[1], **Samuel Alves**[1]
Henrique Jamal[1], **Lucas Araújo Pereira**[1]
Sávio Salvarino Teles de Oliveira[1], **Deborah Silva Alves Fernandes**[1]

[1]Instituto de Informática – Universidade Federal de Goiás (UFG),
Alameda Palmeiras, Chácaras Califórnia, Goiânia, Brazil

{matheus.leonel,henrique.jamal}@discente.ufg.br
{amir.youssef.ia,samuelsch07,savioteles}@gmail.com
apereiral@ufg.br, deborah@inf.ufg.br

Abstract. *This paper evaluates binarization, skew correction, OCR-guided region selection, and their combinations for schema-constrained document extraction with Gemini 3 Flash and Gemma 4 31B. Experiments cover 564 real Brazilian documents, 17 types, 110 fields, and manually validated ground truth, using field-level F1, normalized Levenshtein similarity, and separate artifact and LLM-side latency. Artifact-generation time is measured separately because variants can be produced offline and is therefore excluded from LLM-side latency. Skew correction achieved the highest observed aggregate mean for Gemini 3 Flash, whereas the baseline did so for Gemma 4 31B; binarization and heuristic region selection often had lower F1. Without inferential tests, these descriptive results support no universal pre-processing policy.*

1. Introduction

Organizations in regulated domains process large volumes of compliance, training, safety, identification, and operational documents. These collections contain recurring fields, such as names, taxpayer identifiers, dates, companies, signatures, and document-specific attributes, but the information is embedded in heterogeneous layouts, scanned PDFs, and images with variable quality. Manual transcription is costly, and small extraction errors can invalidate structured records.

LLMs offer a practical alternative for schema-driven extraction. Recent work benchmarks realistic structured extraction and compares OCR-based and OCR-free document pipelines more broadly [Li et al. 2025, Tenckhoff et al. 2026, Shen et al. 2026] ; applied studies cover invoices and Brazilian digitized documents [Loukil et al. 2024, Almeida and Caminha 2024]. This study instead isolates specific OCR-oriented transformations within a fixed multimodal, schema-constrained architecture.

Document understanding pipelines commonly apply image transformations before extraction. Binarization may improve foreground/background contrast, skew correction may normalize page geometry, and OCR-guided region selection may reduce irrelevant visual content. These operations can also remove faint characters, alter visual cues, or crop required fields. Their effect under strict matching depends on document domains, institutional practices, acquisition conditions, and layouts.

This paper evaluates binarization, skew correction, and OCR-guided region selection, individually and in two-stage combinations, for LLM-based information extraction. The study uses a fixed extraction architecture, schemas, manually validated ground truth, prompting protocol, and scoring procedure.

It addresses the following research questions under this controlled protocol: **RQ1.** Do OCR-oriented pre-processing techniques improve field-level extraction quality relative to the original-input baseline? **RQ2.** How do these effects vary across the two evaluated models and the 17 document types? **RQ3.** Do two-stage combinations improve upon individual transformations, or do they compound representation losses?

The contributions are: (i) a controlled comparison over 564 real documents and 110 schema fields; (ii) joint analysis of exact-match quality, approximate similarity, artifact-generation latency, and LLM-side latency; and (iii) evidence that pre-processing effects vary across models and document types. The goal is not to generalize across multimodal LLM families, but to report controlled evidence for a proprietary model and an open-weights model accessed through the same managed API under the evaluated protocol.

2. Related Work

LLM-based information extraction has been studied in both benchmark and application settings. READoc evaluates realistic document structured extraction over heterogeneous PDFs, while LLMStructBench focuses on schema-constrained structured outputs [Li et al. 2025, Tenckhoff et al. 2026]. Applied work investigates LLM pipelines for invoices [Loukil et al. 2024]. In Brazilian contexts, studies have evaluated information extraction from digitized documents, initial court petitions, and legal documents with RAG [Almeida and Caminha 2024, Esashika et al. 2025, Aquino et al. 2024]. Recent BRACIS work extends this context through fine-tuning and LoRA for Portuguese Open Information Extraction and legal NER [Melo et al. 2024, Nunes et al. 2024]; keyword-extraction results also reinforce the need to evaluate extraction pipelines empirically [Zagatti et al. 2024].

This literature is grounded in a broader tradition of language and domain adaptation for Brazilian Portuguese. BERTimbau and Sabiá show the relevance of Portuguese-focused pretraining [Souza et al. 2020, Pires et al. 2023], while GovBERT-BR targets governmental and administrative language [Silva et al. 2024]. Together, these works establish that linguistic and institutional context can affect Brazilian document processing. They primarily address text-based tasks or model adaptation, however, whereas this work keeps the extraction architecture fixed and isolates OCR-oriented transformations of the visual input.

The role of layout and OCR remains important in document understanding. Layout-aware models such as LayoutLM show that text and spatial information can be jointly useful [Xu et al. 2020], and LLM-based extraction with localization further highlights the value of structural cues [Perot et al. 2024]. At the same time, recent work questions whether OCR-centric pipelines are consistently beneficial in the multimodal LLM era [Shen et al. 2026]. Classical pre-processing remains relevant because OCR errors can propagate downstream [Smith 2007, Loukil et al. 2024], but its benefits for multimodal extraction cannot be assumed. Recent work also explores complementary direc-

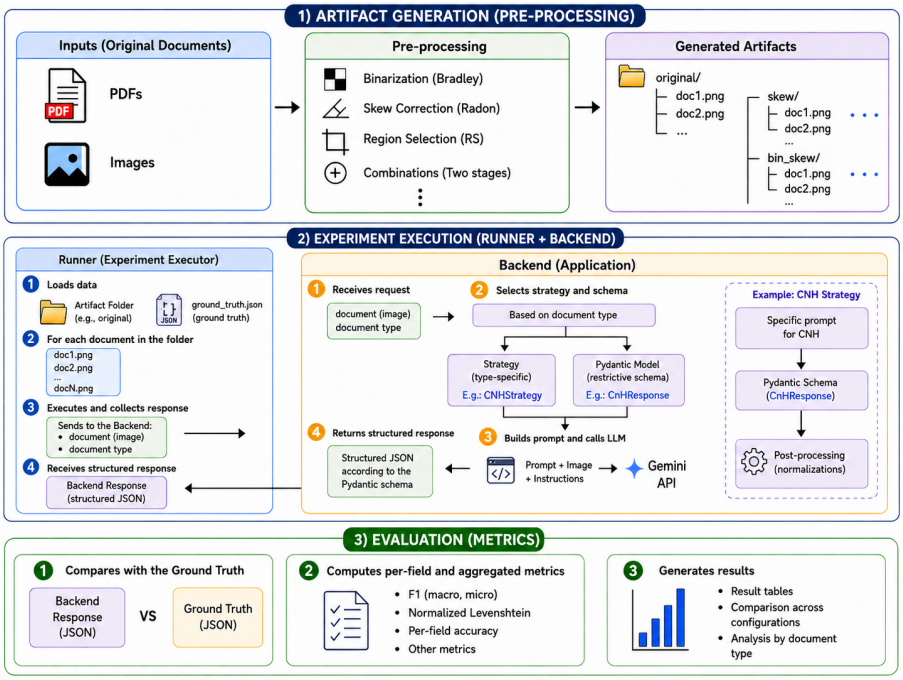


Figure 1. Experimental pipeline. Offline-generated document variants are processed by a fixed extraction system and evaluated against manually validated ground truth.

tions, including table-structure-aware multimodal extraction [Nunes et al. 2025]. These approaches address broader architectural choices, whereas this study changes only the document representation within a fixed extraction architecture rather than comparing all document understanding paradigms. In contrast to the broader OCR-versus-no-OCR question studied on large-scale real-world datasets in [Shen et al. 2026], the present work isolates three specific transformations and their combinations for fixed-schema extraction from Brazilian Portuguese compliance documents, with field-level quality and separate latency measurements.

3. Methodology

3.1. Pipeline and Evaluation

The protocol has five stages. First, the baseline and all transformed representations are generated from each document; combined artifacts apply the named transformations sequentially. Second, the runner sends a representation and its document type to the backend. Third, the backend selects the fixed prompt and Pydantic schema. Fourth, it submits the bytes to the selected multimodal model and returns a schema-constrained response. Finally, the runner compares that response with the ground truth. Representations of a given document therefore share the prompt, schema, and target; only the visual input changes within each model comparison. Figure 1 summarizes the pipeline.

Predicted and expected dictionaries are compared over their union of keys. Values are lowercased, trimmed, stripped of diacritics, and normalized for internal spacing. Exact normalized matches determine field-level precision, recall, and F1; Levenshtein similarity is also used to capture near matches [Levenshtein 1966]. F1 captures schema-level exactness, while Levenshtein similarity distinguishes complete failures from outputs that are textually close but still unusable under strict normalization.

Each evaluated experiment was executed at least three times. Reported values are arithmetic means, and standard deviations are included in the relevant tables. No formal significance tests, confidence intervals, or effect sizes are reported. Artifact-generation latency denotes the time needed to produce a pre-processed representation. LLM-side latency denotes extraction request time after the representation is available and excludes offline artifact generation.

3.2. Models and Input Configuration

The evaluated models are Gemini 3 Flash and the open-weights Gemma 4 31B. Both were accessed through the same managed API service and credential; changing the model identifier selected the proprietary or open-weights model while retaining the same request path, document bytes, prompts, and Pydantic schemas. Hardware, quantization, and internal serving were provider-managed and outside the authors' control. The comparison therefore concerns two operational model configurations under one fixed extraction architecture, not controlled serving hardware or all multimodal model families.

The pipeline does not set temperature, top-p, top-k, a random seed, or an output-token limit, so each model used the service defaults in effect during the experiment. For `gemini-3-flash` and `gemma-4-31b`, these defaults include temperature 1.0, nucleus sampling with top-$p = 0.95$, and top-$k = 64$. No global resizing, re-encoding, or DPI normalization is applied; resolution, aspect ratio, and PDF rendering follow each transformation's implementation.

3.3. Dataset and Schemas

The dataset contains 564 real Brazilian compliance and operational documents in Portuguese, including flood safety certificates (Afogados), Arrais licenses, Occupational Health Certificates (ASO), Brazilian driver's licenses (CNH), operator badges (*crachá*), Personal Protective Equipment (EPI) records, excavator operator certificates (Escavadeira), employee registration forms (ERF), sailor licenses (Marinheiro), NR-5, NR-10, NR-10-SEP, NR-11, NR-12, NR-33, and NR-35 training certificates, and insurance documents. It includes mixed PDF and image inputs with substantial variation in layout, typography, and scan artifacts. Each document type has a fixed structured schema and manually validated ground truth. The schemas total 110 fields and include normalized dates, CPF/CNPJ identifiers, names, companies, signatures, document categories, and type-specific attributes. Table 1 summarizes the dataset, and Listing 1 shows the ground-truth structure used by the evaluator.

Because the documents contain sensitive personal data, the dataset cannot be publicly released.

3.4. Annotation Protocol

Ground-truth construction began with a study of every document type to identify information relevant to the extraction task. Target fields were standardized, including output formats for dates, CPF/CNPJ identifiers, and document-specific tie-breaking criteria. Prompts and schemas were then implemented from these rules.

Gemini 3 Flash was used for an initial broad extraction pass. Each resulting field was subsequently reviewed by a human annotator against the original document and cor-

Table 1. Number of documents per type in the dataset.

Type	Docs	Fields	Type	Docs	Fields
Afogados	47	7	NR 10	37	7
Arrais	16	6	NR 10-SEP	39	7
ASO	97	7	NR 11	27	7
CNH	28	6	NR 12	38	7
Crachá	18	6	NR 33	28	7
EPI	46	6	NR 35	38	7
Escavadeira	11	7	NR 5	6	7
ERF	44	7	Insurance	38	3
Marinheiro	6	6			
Total documents		564	Schema fields		110

Listing 1. Example ground-truth structure.

```
{
 "ground_truth": [{
  "doc": "document_001.pdf",
  "type": "document_type",
  "tags": ["pdf"],
  "responses": {
    "name": "employee name",
    "cpf": "000.000.000-00",
    "date": "2010-01-30",
    "signed": "False"
  }
 }]
}
```

rected when necessary. The model output therefore served only as an annotation initializer; the values used for evaluation were manually validated against the source documents.

Extraction uses document-specific Pydantic schemas and prompts. Prompts define the target fields, require explicit evidence in the document, normalize dates to ISO 8601, preserve CPF/CNPJ formatting, and prohibit inference of missing values. For ambiguous documents, prompts include tie-breaking rules, such as prioritizing validity dates in licenses and insurance documents or admission dates in employee registration forms.

For reproducibility, the anonymous repository provides the document-specific prompt templates, Pydantic schema models, and keyword lists used by region selection: `https://anonymous.4open.science/r/codes-paper-eniac/`. The repository also includes the complete per-document-type tables so reviewers can inspect the values by document type in more detail.

3.5. OCR Preprocessing Techniques

Three pre-processing techniques are evaluated as independent configurations and as pregenerated combinations.

3.5.1. Binarization

Binarization is implemented in Python using `Pillow` and `NumPy`. Each input first undergoes format normalization, including conversion from RGBA to RGB with white-background composition when needed, followed by contrast enhancement with factor 1.5 and grayscale conversion. Three methods were implemented: Bradley, Otsu, and Sauvola. All were evaluated with Gemini 3 Flash; Bradley, which had the highest observed aggregate mean, was then fixed for Gemma 4 31B and the combined experiments.

Otsu dynamically selects a threshold that maximizes between-class variance [Otsu 1979]. Sauvola uses a 25-pixel local window, $k = 0.2$, and $R = 128$. Bradley uses a 41-pixel local window and contrast fraction $t = 0.15$, classifying a pixel as foreground when it is below the local mean multiplied by $(1 - t)$.

3.5.2. Skew Correction

Skew correction is implemented with `OpenCV`, `NumPy`, and `pdf2image` for PDF rasterization. Projection Profile, Radon, and Hough were evaluated with Gemini 3 Flash; Radon had the highest observed aggregate mean and was then fixed for Gemma 4 31B and the combined experiments.

All searches use an angular range of $\pm 15.0°$ and a refinement step of $0.1°$. Projection and Radon use a coarse-to-fine search with a $0.5°$ coarse step and a $\pm 2.0°$ refinement window. The coarse maximum width is 1000 pixels for Projection and 800 pixels for Radon. Projection selects angles by projection variance; Radon uses circular projection and interpolation order 1.

Hough-based correction converts each page to grayscale, extracts edges with Canny [Canny 1986], and detects candidate line segments using the probabilistic Hough transform [Kiryati et al. 1991]. Such line evidence is a common basis for document skew estimation [Amin and Fischer 2000]. Hough uses a vote threshold of 100, minimum line length of 100 pixels, and maximum line gap of 10 pixels. Detected angles are filtered to near-horizontal ranges, and the final estimate is the median of valid angles. Rotation is applied only when the absolute angle exceeds a minimum tolerance, using affine warping with replicated borders.

3.5.3. OCR-guided Region Selection

OCR-guided region selection (RS) uses Tesseract through `pytesseract` with Portuguese OCR configuration [Smith 2007]. For each document type, curated keywords identify relevant regions through exact and fuzzy matching. The fuzzy threshold is 0.72; candidate regions receive a 250-pixel vertical margin, are merged when overlap reaches 40%, and are cropped. Crops are concatenated for PDF inputs. Thus, RS is a specific heuristic based on OCR, keywords, fuzzy matching, margin expansion, merging, cropping, and concatenation.

3.5.4. Combined Configurations

The evaluated combinations are Skew+RS, Bin+RS, and Skew+Bin. The order in each name is the execution order: the first transformation produces the input for the second. Combined results therefore capture interactions between stages rather than the independent sum of their effects.

4. Results

The experiments compare Baseline, binarization (Bin.), skew correction (Skew), region selection (RS), and Skew+RS, Bin+RS, and Skew+Bin. For each document type, F1 and normalized Levenshtein similarity (Lev.) were averaged over three benchmark runs. Tables report the macro-average across the 17 type means; Baseline denotes the original representation, and aggregate σ values describe dispersion across these type means.

Table 2. Aggregate method comparison for Gemini 3 Flash. F1 and Lev. are percentages, ATG latency is in seconds, and σ is the sample SD across type means.

Family	Variant	F1 $\pm\sigma$	Lev. $\pm\sigma$	ATG Lat. $\pm\sigma$	Decision
Binarization	Bradley	**79.08 ± 8.60**	**91.59 ± 4.93**	9.95 ± 1.91	Selected
Binarization	Otsu	69.86 ± 22.61	83.38 ± 17.16	9.54 ± 1.83	Lower quality
Binarization	Sauvola	78.44 ± 8.97	91.53 ± 5.38	**7.96 ± 2.51**	Fastest
Skew correction	Projection	92.47 ± 4.90	96.36 ± 3.90	7.74 ± 1.27	Competitive
Skew correction	Radon	**93.08 ± 4.91**	**96.66 ± 3.92**	7.88 ± 1.05	Selected
Skew correction	Hough	92.28 ± 4.61	96.25 ± 3.81	**7.04 ± 1.29**	Fastest

4.1. Variant Selection

Table 2 compares alternative implementations of binarization and skew correction using Gemini 3 Flash before the broader cross-model comparison.

Bradley achieved the highest observed mean F1/Lev. (79.08/91.59), while Sauvola had the lowest artifact-generation latency (7.96 s); Bradley was therefore fixed for later configurations. Radon likewise achieved the highest observed skew mean F1/Lev. (93.08/96.66), while Hough had the lowest artifact latency (7.04 s), so Radon was fixed. These descriptive selections used Gemini only; the close skew means and type-level variation may indicate that the extraction pipeline is more sensitive to the presence of skew correction than to the specific skew-correction algorithm. The supplementary tables retain the document-specific variation behind this decision.

4.2. Individual-Technique Results

Table 3 shows a model-dependent effect through the deltas relative to the baseline. For Gemini 3 Flash, Skew has the highest aggregate F1 (93.08; ΔF1 = +2.17), whereas Bin. and RS fall to 79.08 and 82.13. For Gemma 4 31B, Baseline is strongest (87.82), and all individual techniques reduce aggregate F1, with Skew closest to baseline (84.40; ΔF1 = -3.42). The supplementary per-type tables indicate that these aggregate effects are not uniform: binarization sharply degrades CNH for both models; RS improves ERF for both models but strongly degrades the exploratory Marinheiro subset ($n = 6$) and NR 10; and NR 12 favors RS for Gemma while Skew favors Gemini. These contrasts support a conditional interpretation of pre-processing rather than a single globally preferred representation.

For Gemini 3 Flash, skew correction improves F1 over Baseline for Afogados, Escavadeira, ERF, Marinheiro, NR 10, NR 12, and NR 35, while CNH and EPI do not improve. This pattern suggests that geometric normalization is useful only when page alignment is a relevant obstacle. Binarization is less stable: Gemini CNH falls from 94.64 to 52.90 and Insurance from 99.12 to 66.67; Gemma CNH falls from 97.02 to 34.21. The supplementary tables also show degradation for EPI, ERF, and Insurance under binarization. A plausible explanation is that thresholding can collapse low-contrast color differences, thin characters, signatures, or field boundaries that remain useful in the original representation, particularly in scanned or signed forms and identification-like documents.

RS also demonstrates why aggregate means require per-type inspection. It im-

Table 3. Aggregate individual-technique comparison. F1 and Lev. are percentages, deltas are technique minus baseline, and σ is the sample SD across type means.

Model	Technique	F1 $\pm\sigma$	ΔF1	Lev. $\pm\sigma$	ΔLev.
Gemini 3 Flash	Baseline	90.92 $\pm$ 6.00	–	94.82 $\pm$ 5.23	–
Gemini 3 Flash	Bin.	79.08 $\pm$ 8.60	-11.84	91.59 $\pm$ 4.93	-3.24
Gemini 3 Flash	Skew	**93.08 $\pm$ 4.91**	**+2.17**	**96.66 $\pm$ 3.92**	**+1.84**
Gemini 3 Flash	RS	82.13 $\pm$ 15.91	-8.79	87.60 $\pm$ 13.61	-7.23
Gemma 4 31B	Baseline	**87.82 $\pm$ 5.82**	–	**94.01 $\pm$ 5.33**	–
Gemma 4 31B	Bin.	75.23 $\pm$ 12.63	-12.59	85.65 $\pm$ 10.69	-8.36
Gemma 4 31B	Skew	84.40 $\pm$ 8.33	-3.42	89.79 $\pm$ 8.94	-4.22
Gemma 4 31B	RS	71.26 $\pm$ 20.30	-16.55	77.39 $\pm$ 21.07	-16.63

Table 4. Aggregate combined-technique comparison. F1 and Lev. are percentages, deltas are technique minus baseline, and σ is the sample SD across type means.

Model	Technique	F1 $\pm\sigma$	ΔF1	Lev. $\pm\sigma$	ΔLev.
Gemini 3 Flash	Skew+RS	63.71 $\pm$ 14.91	-27.21	73.01 $\pm$ 12.85	-21.81
Gemini 3 Flash	Bin+RS	78.20 $\pm$ 12.90	-12.72	86.36 $\pm$ 10.79	-8.46
Gemini 3 Flash	Skew+Bin	**81.28 $\pm$ 7.40**	**-9.64**	**95.69 $\pm$ 3.88**	**+0.87**
Gemma 4 31B	Skew+RS	61.23 $\pm$ 15.55	-26.59	69.44 $\pm$ 14.15	-24.57
Gemma 4 31B	Bin+RS	74.45 $\pm$ 14.64	-13.37	82.63 $\pm$ 13.43	-11.39
Gemma 4 31B	Skew+Bin	**77.19 $\pm$ 9.50**	**-10.63**	**93.91 $\pm$ 6.26**	**-0.11**

proves ERF for Gemini and improves ERF, Escavadeira, and NR 12 for Gemma, but strongly degrades Marinheiro and NR 10 for both, with NR 10-SEP also showing a large drop in the supplementary tables. Successful crops may reduce irrelevant content, whereas OCR errors, incomplete keyword coverage, or region merging can exclude required evidence. Marinheiro illustrates this failure mode: RS alone removes too much evidence, but applying skew correction before RS partially recovers F1 relative to RS alone, which is consistent with highly angled inputs affecting the OCR-guided crop.

Table 4 shows Skew+Bin as the best combination for both models (F1/Lev. 81.28/95.69 and 77.19/93.91), although it remains below the strongest individual setting in F1. Its Levenshtein values are close to, or slightly above, baseline, especially for Gemini (ΔLev. = +0.87), indicating that many outputs may remain textually similar while failing strict exact-field matching. Skew+RS is weakest, consistent with geometric changes amplifying OCR-based selection failures, while Bin+RS is intermediate, and no combination surpasses the best individual configuration in aggregate F1, indicating that sequential transformations can accumulate information loss.

Table 5 reports observed API-request time after each representation was available, excluding offline preprocessing. RS had the lowest mean for both models (6.94 s and 5.39 s), while Skew+RS had the highest (16.32 s and 9.83 s). Within-model rep-

Table 5. Observed API-request latency (s). Standard deviations are sample SDs across the 17 type-level means.

Metric	Gemini 3 Flash						
	Baseline	Bin.	Skew	RS	S&RS	B&RS	S&B
Mean	7.49	9.85	7.76	**6.94**	16.32	7.78	8.61
Standard Deviation	1.32	6.07	1.42	1.41	5.24	2.14	4.06
Metric	Gemma 4 31B						
	Baseline	Bin.	Skew	RS	S&RS	B&RS	S&B
Mean	6.87	6.84	9.09	**5.39**	9.83	7.12	7.63
Standard Deviation	0.83	0.45	1.55	1.67	2.47	2.08	0.97

resentation comparisons are more defensible because the API path and model are fixed. Absolute differences between models remain descriptive and service-specific, not a controlled hardware or serving benchmark. For on-demand use, applicable artifact-generation and LLM-side times can be added. Unusually high values may reflect input complexity, page/crop concatenation, visual load, or processing overhead, but these factors were not isolated.

Research-question synthesis. For RQ1, Table 3 shows an aggregate gain only for Gemini with skew correction, while Gemma's baseline is strongest. For RQ2, the results show substantial model and document-type variation. For RQ3, Table 4 shows that no two-stage combination exceeds the strongest individual configuration in aggregate F1. These answers are descriptive evidence under the evaluated protocol rather than universal claims.

5. Discussion

The results indicate that OCR-oriented pre-processing should not be treated as a universally beneficial step for multimodal, schema-constrained information extraction. Its effect depends on the document type, transformation, and model, consistent with work questioning the universal value of OCR-centric processing [Shen et al. 2026].

The clearest contrast is model-dependent: skew correction raises Gemini's aggregate F1 from 90.92 to 93.08, whereas Gemma's original-input baseline remains strongest at 87.82. Thus, Gemini appears to benefit more from geometric normalization in this collection, while Gemma either handles original layouts more effectively or is more sensitive to the transformation. Binarization reduces mean F1 for both models despite Bradley being its strongest variant. Thresholding may remove low-contrast characters, alter boundaries, or amplify compression artifacts; under exact field matching, small changes in names, dates, and identifiers become errors.

RS is likewise asymmetric: it helps selected document/model pairs but strongly degrades others, such as Marinheiro and NR 10. Its OCR, keyword matching, margin expansion, merging, and cropping stages create multiple opportunities to exclude evidence. This result concerns the evaluated heuristic and does not rule out learned or layout-aware localization.

The combined configurations reinforce this document- and model-dependent in-

terpretation. Skew+Bin is the strongest combined setting for both models, with mean F1/Lev. of 81.28/95.69 for Gemini 3 Flash and 77.19/93.91 for Gemma 4 31B, but neither exceeds the best individual configuration. The gap between its high Levenshtein similarity and lower F1 indicates outputs that remain textually close yet fail exact matching. Skew+RS is weakest, suggesting that geometric correction can compound heuristic cropping failures.

Operationally, the results favor policies validated by model and document type rather than mandatory pre-processing. Original inputs remain the safest default when gains are inconsistent, while skew correction is a candidate for rotation-affected classes. The fixed architecture attributes these differences to input representation rather than model adaptation.

Latency adds a practical trade-off. RS is the fastest LLM-side configuration on average for both models (6.94 s and 5.39 s), but sacrifices quality. Gemini's skew correction instead combines the best individual-technique F1 with latency near baseline. End-to-end on-demand cost includes both artifact generation and LLM-side time, although artifacts can be prepared offline.

5.1. Limitations

The evidence is limited to two models and one Brazilian Portuguese collection; other prompts, schemas, domains, languages, and multimodal families may produce different effects. Bradley and Radon were selected with Gemini before the cross-model experiments, which may favor variants compatible with that model. A single-run Gemma diagnostic retained their ordering, but was not part of the repeated main evaluation. RS is a keyword-driven heuristic and therefore does not represent learned or layout-aware localization. Initializing annotations with Gemini may have retained residual formatting, normalization, or ambiguity-resolution bias despite human review; field-level correction rates were not recorded.

The reported standard deviations summarize dispersion across type-level means, but the study reports no significance tests, confidence intervals, or effect sizes. It also excludes OCR-first and supervised baselines, Portuguese-specific pretraining, fine-tuning, LoRA, and ensembles, which address broader architecture or adaptation questions. Privacy precludes releasing documents or before/after images that may expose personal information. Finally, LLM-side latency excludes offline artifact generation and should not be read as end-to-end on-demand latency.

6. Conclusion

This paper evaluated three OCR-oriented transformations and their combinations for schema-constrained multimodal extraction from heterogeneous Brazilian documents. Skew correction improves aggregate F1 for Gemini 3 Flash, but Gemma 4 31B performs best on original inputs; effects also vary substantially by document type. Skew+Bin is the strongest combination, yet no combination surpasses the best individual configuration.

By holding prompts, schemas, models, and evaluation fixed, the study isolates input representation as the source of the observed differences. The results also show that preserving the original visual evidence can be preferable to automatically applying transformations designed for conventional OCR.

The findings therefore support conditional rather than uniform pre-processing. Future work should broaden model coverage, add OCR-first and supervised baselines, compare OCR-free and hybrid pipelines, improve layout-aware localization, analyze errors by field class, and provide confidence intervals and anonymized qualitative failure analysis.

7. Acknowledgments

This work was supported by the ANEEL Research and Development (R&D) Program under Project PD-02949-0124/2024 - SafeAI: Plataforma Inteligente de Comunicação, Monitoramento e Gestão de Segurança", funded by Foz do Chapecó Energia and developed in partnership with the UFG, CEIA, 4VANTS and Solumar. This work was also supported by the National Institute of Science and Technology (INCT) in Responsible Artificial Intelligence for Computational Linguistics and Information Treatment and Dissemination (TILD-IAR) grant number 408490/2024-1.

References

Almeida, F. and Caminha, C. (2024). Evaluation of entry-level open-source large language models for information extraction from digitized documents. In *Anais do XII Symposium on Knowledge Discovery, Mining and Learning*, pages 25–32, Porto Alegre, RS, Brasil. SBC.

Amin, A. and Fischer, S. (2000). A document skew detection method using the hough transform. *Pattern Analysis and Applications*, 3(3):243–253.

Aquino, I., dos Santos, M. M., Dorneles, C., and Carvalho, J. T. (2024). Extracting information from brazilian legal documents with retrieval augmented generation. In *Anais Estendidos do XXXIX Simpósio Brasileiro de Bancos de Dados*, pages 280–287, Porto Alegre, RS, Brasil. SBC.

Canny, J. (1986). A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679–698.

Esashika, R., Figueiredo, C., and Melo, T. (2025). Structuring information from initial petitions using llms: A study in brazilian courts of justice. In *Anais do XXII Encontro Nacional de Inteligência Artificial e Computacional*, pages 546–557, Porto Alegre, RS, Brasil. SBC.

Kiryati, N., Eldar, Y., and Bruckstein, A. M. (1991). A probabilistic hough transform. *Pattern Recognition*, 24(4):303–316.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Li, Z., Abulaiti, A., Lu, Y., Chen, X., Zheng, J., Lin, H., Han, X., Jiang, S., Dong, B., and Sun, L. (2025). READoc: A unified benchmark for realistic document structured extraction. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21889–21905, Vienna, Austria. Association for Computational Linguistics.

Loukil, F., Cadereau, S., Verjus, H., Galfre, M., Salamatian, K., Telisson, D., Kembellec, Q., and Le Van, O. (2024). Llm-centric pipeline for information extraction from invoices. In *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*, pages 569–575.

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

Nunes, G., Rolla, V., Pereira, D., Alves, V., Carreiro, A., and Baptista, M. (2025). Benchmarking table extraction: Multimodal LLMs vs traditional OCR. In *Proceedings of the 1st Joint Workshop on Large Language Models and Structure Modeling (XLLM 2025)*, pages 8–15, Vienna, Austria. Association for Computational Linguistics.

Nunes, R., Puttlitz, L., Boll, A., Spritzer, A., Freitas, C., Balreira, D., and Tavares, A. (2024). An ensemble of llms finetuned with lora for ner in portuguese legal documents. In *Anais da XXXIV Brazilian Conference on Intelligent Systems*, pages 127–140, Porto Alegre, RS, Brasil. SBC.

Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66.

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.

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.

Shen, J., Peiyue, Y., Ghosh, A., Mai, Y., and Dahlmeier, D. (2026). OCR or not? rethinking document information extraction in the MLLMs era with real-world large-scale datasets. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 385–396, Rabat, Morocco. Association for Computational Linguistics.

Silva, M., Oliveira, G., Costa, L., and Pappa, G. (2024). Govbert-br: A bert-based language model for brazilian portuguese governmental data. In *Anais da XXXIV Brazilian Conference on Intelligent Systems*, pages 19–32, Porto Alegre, RS, Brasil. SBC.

Smith, R. (2007). An overview of the tesseract ocr engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pages 629–633.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: Pretrained bert models for brazilian portuguese. In *Intelligent Systems: BRACIS 2020 Proceedings*, page 403–417. Springer-Verlag.

Tenckhoff, S., Koddenbrock, M., and Rodner, E. (2026). Llmstructbench: Benchmarking large language model structured data extraction.

Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., and Zhou, M. (2020). Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 1192–1200. ACM.

Zagatti, F., Lucrédio, D., and Caseli, H. (2024). Unsupervised statistical keyword extraction pipeline: Is llm all you need? In *Anais da XXXIV Brazilian Conference on Intelligent Systems*, pages 460–475, Porto Alegre, RS, Brasil. SBC.