Agent Orchestration - LLM for Legal Metadata Extraction: A Comparative Analysis of Efficiency and Precision

Luiz Anísio Batitucci[1,2], Luciane Inácia Lopes[1,2]
Rhodie Ferreira[1], Emerson Cabrera Paraiso[2,3]

[1]Superior Tribunal de Justiça (STJ), Brasília, Brazil
[2]PPGIa, Pontifícia Universidade Católica do Paraná (PUCPR), Curitiba, Brazil
[3]Pontifícia Universidade Católica do Paraná (PUCPR), Curitiba, Brazil
Correspondence: luizanisio@gmail.com

Abstract

This work introduces and evaluates JAMEX (Judicial Multi-Agent Metadata Extraction), a multi-agent pipeline for extracting structured metadata from Brazilian court decisions (*Espelho do Acórdão*), and compares it against a strong single-prompt baseline under an Information Retrieval-only (IR-only) setting. We first ran a pilot on 300 decisions and then reran the experiment on a stratified dataset of $n = 1{,}225$; completion rates varied across executions, yielding between 779–1,216 successfully completed instances, with non-completion concentrated in agentic configurations. Across re-executions, the accuracy impact of agents was *strategy-dependent*: GPT-5 improves over the baseline in multiple agentic strategies but not across all orchestration variants, while smaller models (Gemma3-12B/Gemma3-27B) show no robust gains. Orchestration refinements motivated by agent design literature (memory, planning and directed review) improved traceability, but performance remained sensitive to task decomposition and context splitting. Overall, JAMEX increases token usage and operational complexity, so deployment must balance accuracy, completion reliability, and cost for Portuguese legal metadata extraction.

1 Introduction

Agentic architectures built on top of large language models (LLMs) are moving from proof-of-concept demonstrations to production deployments, but controlled, domain-grounded evaluations against strong prompting baselines remain limited, particularly in non-English legal workflows where outputs must be strictly structured, auditable, and operationally reliable (Wang et al., 2024; Shen et al., 2023; Dong et al., 2024). This paper presents a systematic study in the Brazilian Portuguese legal domain, comparing a strong single-prompt baseline (few-shot, schema-constrained) against **JAMEX** (Judicial Multi-Agent Metadata Extraction), a multi-agent orchestration pipeline in which specialized agents plan, extract, cross-validate, and consolidate structured fields of a judgment metadata sheet (*Espelho do Acórdão*). Our baseline prompt design follows established prompting practices that use exemplars and explicit reasoning structure to improve instruction-following and extraction reliability in LLMs (Wei et al., 2022; Yu et al., 2023).

In public-sector workflows, high-volume court decisions must be converted into reliable and auditable metadata (e.g., legal thesis, cited precedents, legislative references) under governance and transparency requirements (Conselho Nacional de Justiça, 2024; Suriani and Pacheco, 2025; Moreira, 2023; Bezerra and Lopes, 2025). In parallel, automated extraction of legal attributes and terminology has been studied in the broader legal NLP literature, including settings where annotated data are scarce and the objective is to identify structured elements from unstructured proceedings (Adhikary et al., 2023; Breton et al., 2025). While Retrieval-Augmented Generation (RAG) is frequently discussed in NLP research (Lewis et al., 2020), many institutional pipelines operate under deterministic document retrieval, where the target decision is fetched by identifier from a governed corpus and injected verbatim into the prompt context. This design is consistent with practical constraints of long and often redundant Brazilian rulings, widely reported in Portuguese legal NLP benchmarks (de Vargas Feijó and Pereira Moreira, 2018; Lins et al., 2024). We adopt this Information Retrieval-only (IR-only) regime to isolate the contribution of orchestration (planning, schema validation, dependency management, directed review) from multi-source retrieval effects.

JAMEX decomposes the extraction of the *Espelho do Acórdão* into specialized agents (e.g., Legal Thesis, Cited Jurisprudence, Legislative References, Notes/Auxiliary Terms/Theme) and a final validator that enforces schema and institutional

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 727–737
April 13-16, 2026

rules. Related work has explored structured information extraction in Portuguese legal documents under weak supervision (e.g., role-filler entity extraction), but our setting targets institution-specific judgment metadata sheets with strict JSON schema enforcement and orchestration-dependent reliability trade-offs (Navarezi et al., 2022). This design is motivated by agentic-system literature emphasizing explicit planning, specialization, and consolidation (Shen et al., 2023; Wang et al., 2024), and by observability practices for analyzing agent behavior in production-like settings (Dong et al., 2024). However, decomposing a tightly coupled legal extraction task may also fragment critical context and amplify cascade errors across dependent fields, particularly for smaller models with limited instruction-following and self-correction capacity. Therefore, the central question is not whether agents *can* help, but under which conditions (model capacity, orchestration strategy, and task coupling) they provide measurable quality improvements over a strong monolithic baseline at acceptable operational overhead.

We evaluate field-level quality with judge-style metrics (precision, recall, F_1) using LLM-as-a-Judge rubrics aligned with institutional standards (Liu et al., 2023; Zhu et al., 2025), complemented by operational telemetry (tokens, latency, revision cycles, completion). Experiments are conducted on a stratified dataset of Brazilian Portuguese Criminal Law appellate decisions from an official public repository (Superior Tribunal de Justiça, 2022). The study quantifies the quality–cost trade-off of multi-agent orchestration under realistic constraints and reports how results vary across model families and orchestration variants.

Research questions. **RQ1** Does a multi-agent pipeline yield higher field-level extraction quality than a single strong prompt under IR-only context?
RQ2 Does agentic orchestration improve operational robustness without prohibitive efficiency penalties?
RQ3 Which extraction fields benefit most from agent specialization?

Hypothesis. We hypothesize that the multi-agent pipeline yields statistically superior field-level F_1 scores compared to the single-prompt baseline. The formal statistical formulation is presented in Section 2.

This paper introduces a Portuguese legal benchmark for structured metadata extraction from Brazilian court decisions under an IR-only setting, enabling direct comparison between a strong schema-constrained single-prompt baseline and a multi-agent approach with strict JSON outputs. We present JAMEX (Judicial Multi-Agent Metadata Extraction), an orchestration pipeline with explicit planning, dependency-aware execution, schema validation, directed review, and auditable evidence linking through intermediate JSON artifacts and validator feedback. We also provide a controlled evaluation protocol that reports field-level judge-style quality metrics (precision, recall, and F_1) alongside operational telemetry (token usage, latency, revision cycles, and completion reliability). Finally, we report strategy-dependent results showing that agentic decomposition can improve traceability and, in some configurations, accuracy, but can also degrade performance when task coupling and context fragmentation outweigh the benefits, offering actionable guidance for deployment under cost and reliability constraints.

This work contributes a controlled comparison between a strong single-prompt baseline and multi-agent orchestration across multiple model families on real-world Brazilian Portuguese legal decisions, jointly analyzing quality, cost, and completion robustness. We also release a reproducible dataset-generation pipeline built on the STJ Open Data Portal (Superior Tribunal de Justiça, 2022), enabling future NLP research on Portuguese legal text without reliance on proprietary corpora.

2 Method

This section describes the methodological design used to evaluate JAMEX for pre-filling the *Espelho do Acórdão*, focusing on dataset construction, orchestration specification, evaluation protocol, and statistical verification. All code used in this work, including the dataset preparation pipeline extracted from the court's official repository, is available in a public GitHub repository[1].

Dataset and reference metadata. The experimental dataset comprises collegial (panel) judicial decisions (*acórdãos*) from an official public repository (Superior Tribunal de Justiça, 2022). Pipeline input consists exclusively of unstructured full-text documents, simulating a realistic scenario where no prior metadata are available.

For reference metadata, we use judgment metadata sheets (*espelhos de acórdão*) extracted by a

[1]Experiment repository: GitHub repository

proprietary model using prompts developed by domain experts from the indexing division. These sheets represent the expert-designed prompt reference metadata for each decision, enabling objective validation of model outputs against institutional standards.

The target output follows a strict JSON schema with seven canonical keys[2], including semantically complex fields such as `teseJuridica` (legal thesis) and `jurisprudenciaCitada` (cited precedents), as well as indexing and auxiliary fields: `referenciasLegislativas` (normative references), `tema` (themes), `termosAuxiliares` (auxiliary search terms), `notas` (annotated highlights), and `informacoesComplementares` (complementary information to the summary). This structure requires not only extraction but also normalization and cross-field consistency aligned with the institutional manual (Superior Tribunal de Justiça, 2021).

Sampling strategy. The experiments use a stratified set of $n = 1{,}225$ appellate decisions from the Criminal Law domain, spanning the period from January 1, 2023, to December 31, 2024. To reduce redundancy and amplify textual diversity, a semantic filter was applied using embeddings trained on decisions from the same court, retaining documents whose pairwise cosine distance exceeded $\theta = 0.15$ (i.e., cosine similarity ≤ 0.85). In addition to improving diversity, this filtering also reduced the overall volume to keep the experiment computationally feasible, as the court publishes approximately 5,000 decisions per week. This selection reduces the likelihood of near-duplicate samples, increasing corpus variance and mitigating overfitting to common linguistic patterns.

JAMEX architecture and orchestration pipeline. The experiment evaluates a multi-agent architecture designed for structured extraction of legal metadata, inspired by LLM orchestration frameworks (Wang et al., 2024) and distributing responsibilities across planning, execution, and validation phases. The system operates in a *stateless* manner between iterations, with inter-agent communication exclusively through JSON objects, ensuring auditability and reducing hallucination risks from shared context contamination.

As illustrated in Figure 1, the pipeline ingests the full-text decision and activates the **Fields Agent** (*Extraction Planning*), which selects the relevant schema fields and builds an execution graph $G = (V, E)$ over specialist agents and their dependencies. The orchestration mixes sequential and parallel steps: jurisprudence extraction is conditioned on thesis detection (skipped when thesis output is empty), while independent agents (Notes, TAP, Topics, and Legal References) run in parallel to reduce latency, each receiving only task-specific context.

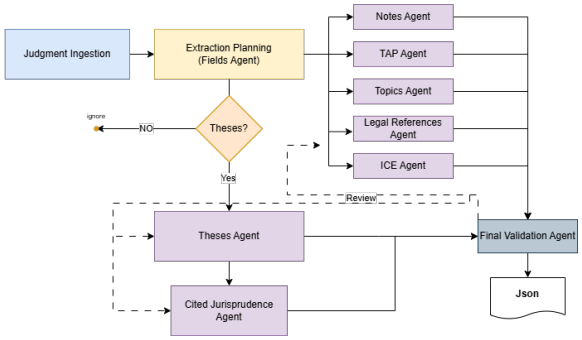

Figure 1: JAMEX orchestration flow showing specialist agents and their dependencies

All flows converge to the **finalValidationAgent**, which enforces schema conformance, data integrity, and rules derived from the institutional manual (Superior Tribunal de Justiça, 2021). JAMEX uses a **directed review** loop: when errors are found, the validator returns structured annotations to the responsible agent for refinement, allowing up to $r_{\max} = 2$ in some configurations and up to $r_{\max} = 5$ in others (including the re-execution) before final JSON consolidation. The entire run is logged as an auditable trail of intermediate JSON objects, validator feedback, and revision outcomes.

Observability and evaluation metrics. To support reproducibility and behavioral analysis, we adopt an observability strategy aligned with Dong et al. (2024), logging prompts and input context, intermediate artifacts and agent decisions, token usage and latency, and validation outcomes with revision counts. Evaluation combines *LLM-as-a-Judge* using GPT-5 (*reasoning medium*) with manual-derived rubrics to score field-level P, R, and F_1 and generate structured justifications (Liu et al., 2023; Zhu et al., 2025); ROUGE and BERTScore as secondary similarity diagnostics (Lin, 2004; Zhang et al., 2020); and operational telemetry (tokens τ, latency L, revision cycles r, and completion flags). An instance is *successfully completed* when

[2]An official *Espelho do Acórdão*: sample document

a schema-valid final JSON is produced within the revision limit. This design isolates orchestration effects relative to a strong single-prompt baseline and quantifies trade-offs between quality, completion reliability, and computational cost.

Statistical hypotheses were formulated to verify whether the agentic approach yields superior performance compared to the baseline in terms of extraction quality, measured by F_1 scores from LLM-as-a-Judge evaluations, assuming $\alpha = 0.05$.

Let F_1^{Agent} and F_1^{Base} denote paired distributions of field-aggregated F_1 scores for agentic and baseline approaches over the same documents. We test:

- **Null Hypothesis** (H_0)**:** $\mu_{F_1}^{\text{Base}} \geq \mu_{F_1}^{\text{Agent}}$
- **Alternative Hypothesis** (H_1)**:** $\mu_{F_1}^{\text{Base}} < \mu_{F_1}^{\text{Agent}}$

We apply the Wilcoxon signed-rank test for paired samples to compare Baseline vs Agent for the global set and separately per model family (GPT-5, Gemma3-12B, Gemma3-27B). We additionally report an effect size (Cohen's d) as a descriptive measure of the magnitude of paired differences. The Wilcoxon signed-rank test is appropriate for paired samples without assuming normality. We verified normality of paired differences using the Shapiro–Wilk test on ΔF_1, as shown in Table 1.

Let W denote the Wilcoxon test statistic and p the corresponding p-value. The decision rule for null hypothesis rejection follows:

- If $p < \alpha = 0.05$: reject H_0, indicating statistical evidence that the agentic approach exhibits superior performance.
- If $p \geq 0.05$: fail to reject H_0, indicating insufficient statistical evidence to conclude that the agentic approach is superior.

The study includes an initial pilot ($n = 300$) used to validate the experimental design and evaluation protocol, followed by a full-scale re-execution on the stratified dataset ($n = 1{,}225$). Completion rates varied by model and orchestration variant (between 779 and 1,216 successfully completed instances), with non-completion concentrated in agentic configurations. This gap is attributable to the multi-agent pipeline's cascading-failure design, where an invalid JSON response, an API timeout, or exhaustion of the per-agent revision budget in any of the specialized agents causes the entire instance to be discarded. We report results per orchestration variant where applicable, as orchestration effects were found to be strategy-dependent under re-execution.

3 Results and Discussion

The study compared six model configurations, spanning three model families (GPT-5, Gemma3-12B, Gemma3-27B) and two architectural approaches (Baseline and Agent-based). Statistical significance was assessed using the Wilcoxon signed-rank test for paired samples (applied to variant A only, where both architectures completed all n=300 documents per model family – see Section 2).

Table 2 consolidates overall performance across three orchestration variants. Variant A corresponds to the pilot configuration (support $n = 300$ per model), in which the agentic pipeline achieved higher mean F_1 across all evaluated families. Variants B and C correspond to full-scale executions on the stratified dataset ($n = 1{,}225$), with per-model support varying depending on completion rates (see Support column in Table 2). In these re-executions, the ordering is not stable: in Variant B, the agentic configuration attains higher mean F_1 for GPT-5, but underperforms the baseline for Gemma3-12B and Gemma3-27B; in Variant C, the baseline outperforms the agentic workflow across all families. These results indicate that orchestration effects are strategy-dependent rather than uniformly positive.

Beyond architectural differences, the variants also differ in review-loop design. Orchestrations A and B limit the reviewer to at most $r_{\max} = 2$ directed revision cycles, whereas orchestration C allows up to $r_{\max} = 5$ cycles and was designed to inject richer contextual information about prior revision requests into the reviewer prompt. This modification improves traceability and supports longer iterative correction, but the aggregated outcomes in Table 2 show that increasing review depth and revision context does not guarantee higher mean accuracy, especially when task decomposition fragments global context or amplifies cross-field dependencies.

Pilot deltas are reported below for completeness, but re-execution confirms that effects may reverse across variants. In the full-scale orchestration variant C (Table 2), GPT-5 shows a negative delta ($\Delta F_1 = -0.0194$, -2.18%) while reducing dispersion (Std Dev 0.0516 $\rightarrow$ 0.0401), suggesting a stability-accuracy trade-off. Gemma3-27B and Gemma3-12B show larger negative deltas

Table 1: Normality test by model family — Shapiro-Wilk test on paired differences (A orchestration variant)

Family	n pairs	Shapiro W	Shapiro p	Normal Δ?
Global	900	0.9656	< 0.0001	No
GPT-5	300	0.9183	< 0.0001	No
Gemma3-12B	300	0.9888	0.0204	No
Gemma3-27B	300	0.9226	< 0.0001	No

Table 2: Overall model performance (mean F_1 score from LLM-as-a-Judge evaluation) across orchestration variants (A, B, C)

id	Architecture / Model	Type	Mean F_1	Std Dev	Median F_1	Support (n)
A orchestration variant - pilot						
1	base_gpt5	Base	0.8325	0.0721	0.8489	300
2	agents_gpt5	Agent	**0.9271**	0.0394	**0.9343**	300
3	base_gemma3(27B)	Base	0.7084	0.1043	0.7150	300
4	agents_gemma3(27B)	Agent	0.7244	0.1041	0.7266	300
5	base_gemma3(12B)	Base	0.6247	0.0966	0.6184	300
6	agents_gemma3(12B)	Agent	0.6494	0.1008	0.6486	300
B orchestration variant						
1	base_gpt5	Base	0.8918	0.0516	0.9028	1216
2	agents_gpt5	Agent	**0.9323**	0.0405	**0.9399**	1216
3	base_gemma3(27B)	Base	0.7260	0.0989	0.7266	1216
4	agents_gemma3(27B)	Agent	0.6689	0.1605	0.6921	1213
5	base_gemma3(12B)	Base	0.6432	0.0996	0.6375	1216
6	agents_gemma3(12B)	Agent	0.5864	0.1071	0.5866	1206
C orchestration variant						
1	base_gpt5	Base	**0.8918**	0.0516	**0.9028**	1216
2	agents_gpt5	Agent	0.8724	0.0401	0.8772	779
3	base_gemma3(27B)	Base	0.7260	0.0989	0.7266	1216
4	agents_gemma3(27B)	Agent	0.6822	0.0933	0.6844	1147
5	base_gemma3(12B)	Base	0.6432	0.0996	0.6375	1216
6	agents_gemma3(12B)	Agent	0.6183	0.1057	0.6242	1210

(-6.03% and -3.87%, respectively), consistent with higher sensitivity to context fragmentation and cross-field dependencies under multi-step decomposition.

Specifically, in the pilot configuration (variant A in Table 2), GPT-5 in the Agent architecture achieved an F_1 gain of **+11.36%** relative to GPT-5 Baseline, while Gemma3-27B exhibited a smaller gain of approximately **+2.3%** over its Baseline version (and Gemma3-12B a gain of **+3.95%**). In the pilot, paired comparisons indicate an advantage for using multiple specialized agents.

Let $\Delta F_1^{(m)} = F_1^{\text{Agent}}(m) - F_1^{\text{Base}}(m)$ represent the performance difference for model family m. We observe:

$$\Delta F_1^{(\text{GPT-5})} = 0.0946 \quad (11.36\%)$$

$$\Delta F_1^{(\text{Gemma3-27B})} = 0.0160 \quad (2.26\%)$$

$$\Delta F_1^{(\text{Gemma3-12B})} = 0.0247 \quad (3.95\%)$$

From a statistical perspective, these pilot improvements proved significant as presented in Table 3. The Wilcoxon signed-rank test for paired samples yielded $p < 0.05$ for all pilot comparisons, leading to rejection of the null hypothesis H_0 within this configuration. However, the hypothesis testing results in Table 3 refer to the pilot configuration only (variant A in Table 2). Under re-execution with orchestration variants B and C, mean differences can change magnitude or reverse direction; therefore, conclusions about superiority must be reported per orchestration strategy (and per model family) rather than assumed to generalize across designs.

In the pilot configuration (variant A), the Agent

Table 3: Hypothesis testing results (Wilcoxon signed-rank test)

Comparison	Mean F_1 (Base)	Mean F_1 (Agent)	Difference (Agent-Base)	*p-value*	Significant ($p < 0.05$)	Cohen's d	Effect Size
GLOBAL	0.7219	0.7670	+0.0451 (+6.25%)	< 0.0001	YES	0.4841	Medium
GPT-5	**0.8325**	**0.9271**	**+0.0946 (+11.36%)**	< 0.0001	YES	**1.1452**	Large
Gemma3-12B	0.6247	0.6494	+0.0247 (+3.95%)	< 0.0001	YES	0.2884	Medium
Gemma3-27B	0.7084	0.7244	+0.0160 (+2.26%)	**0.0200**	YES	0.1768	Small

approach also reduced F_1 variability for GPT-5 and Gemma3-27B, suggesting greater stability on this subset: for GPT-5, the standard deviation decreases from $\sigma = 0.0721$ (Base) to $\sigma = 0.0394$ (Agent). Because stability and mean accuracy can trade off under different orchestration variants (A-C), these dispersion results should be interpreted as configuration-specific rather than as a general property of agentic orchestration.

Token and efficiency results in Table 4 refer to variant A (pilot). In this configuration, agentic orchestration substantially increased token usage relative to the baseline, with overheads ranging from +129% (GPT-5) to +245% (Gemma3-12B), consistently reducing efficiency (F_1 per 1,000 tokens) across model families. The same cost pattern (higher token usage and lower efficiency for Agent) was also observed in variants B and C, although Table 4 reports the detailed telemetry for the pilot setting. Given that accuracy effects vary by orchestration variant (A-C), deployment decisions should treat token overhead as a reliable cost factor and justify orchestration jointly by accuracy, completion reliability, and stability rather than by mean F_1 alone.

Table 4 shows that higher token consumption systematically reduced efficiency across all families in the pilot (e.g., $\Delta\eta = -0.0114$ for GPT-5 and $\Delta\eta = -0.0132$ for Gemma3-27B), reinforcing that orchestration introduces a non-trivial operational penalty in this configuration.

To contextualize costs in real-world deployment scenarios, we estimated monetary expenses based on current API pricing, as shown in Table 5.

Let C_{input} and C_{output} denote the costs per million tokens for input and output, respectively, in Brazilian Reais (BRL). The total cost per document can be approximated as:

$$\text{Cost}_{\text{doc}} = \frac{\tau_{\text{in}} \cdot C_{\text{input}}}{10^6} + \frac{\tau_{\text{out}} \cdot C_{\text{output}}}{10^6}$$

For GPT-5 Agent processing a typical document with $\tau_{\text{in}} \approx 75{,}000$ and $\tau_{\text{out}} \approx 11{,}000$ tokens, the estimated cost would be approximately (BRL):

$$\text{Cost}_{\text{doc}}^{\text{GPT-5}} \approx \frac{75{,}000 \times 1.25}{10^6} + \frac{11{,}000 \times 10.00}{10^6} \approx 0.20$$

While this per-document cost remains manageable, scaling to high-volume production scenarios (e.g., processing approximately 20,000 decisions per month) would incur substantial operational expenses, necessitating cost-optimization strategies such as prompt caching or hybrid architectures combining large and small models. These detailed cost estimates refer to orchestration variant A (pilot).

Textual similarity analysis: ROUGE and BERTScore. Beyond primary quality metrics (precision, recall, and F_1 via LLM-as-a-Judge), we analyzed outputs through the lens of textual similarity using ROUGE and BERTScore. Table 6 reports the detailed similarity results for orchestration variant A (pilot).

As expected, the Baseline architecture produced texts lexically closer to the reference metadata sheets than the Agent architecture. Let $S_{\text{ROUGE}}(\hat{y}, y)$ and $S_{\text{BERT}}(\hat{y}, y)$ denote ROUGE and BERTScore similarity functions, respectively, where $\hat{y}$ represents extracted text and y the reference metadata. Baseline outputs consistently achieved higher values:

$$S_{\text{ROUGE-L}}^{\text{Base}} = 0.2328 > S_{\text{ROUGE-L}}^{\text{Agent}} = 0.2008$$

$$S_{\text{BERT}}^{\text{Base}} = 0.6766 > S_{\text{BERT}}^{\text{Agent}} = 0.6540$$

However, these differences do not imply that Baseline is "better" in absolute terms, but rather that its outputs remain more literal and adherent to the reference text formulation than the reformulations introduced by agents. This apparent contradiction with F_1 superiority requires careful interpretation.

Table 4: Cost and efficiency (F_1 per 1,000 tokens)

Family	Mean Tokens (Base)	Mean Tokens (Agent)	Δ Tokens (%)	Δ F_1 (%)	η_{Base} (F_1/1k tok)	η_{Agent} (F_1/1k tok)	$\Delta\eta$
GPT-5	37,475	85,812	**+128.99%**	**+11.36%**	**0.0222**	0.0108	**-0.0114**
Gemma3-12B	36,508	126,117	+245.45%	+3.95%	0.0171	0.0051	-0.0120
Gemma3-27B	36,424	116,502	+219.85%	+2.26%	0.0194	0.0062	-0.0132

Table 5: Token pricing per model - BRL/1M tokens (quoted on November/2025; OpenAI and OpenRouter)

Model	Input Cost	Output Cost
GPT-5	1.25	10.00
Gemma3-27B	0.09	0.16
Gemma3-12B	0.04	0.13

Table 6: Performance synthesis by architecture (textual similarity metrics)

Metric	Mean Score (Base)	Mean Score (Agent)	Difference (pp)
BERTScore	**0.6766**	0.6540	-2.26
ROUGE-L	**0.2328**	0.2008	-3.20
ROUGE-2	**0.0801**	0.0482	-3.19

Examining cases individually reveals situations where the Agent approach reorganized or paraphrased metadata information, obtaining higher F_1 scores (by covering more relevant information or avoiding errors) while simultaneously distancing itself lexically from the reference text, thereby reducing ROUGE/BERTScore. This phenomenon confirms the complementary role of these metrics: they help identify that agents frequently generate more complete responses that are less "identical" to the baseline.

For instance, Table 7 contrasts ROUGE-L F_1 and LLM-as-a-Judge F_1 for selected Gemma3-12B fields. In *informacoesComplementares*, ROUGE-L F_1 drops from 0.41 to 0.34 while Judge F_1 rises from 0.35 to 0.40, indicating that agents reformulated complementary information more completely despite reduced lexical overlap with the reference. In contrast, *notas* shows a sharp decline in both metrics under the agentic approach, confirming that metric divergence is field-dependent rather than systematic. These patterns illustrate that agents may produce semantically richer extractions that score lower on lexical similarity, or they may genuinely degrade content in fields sensitive to context fragmentation.

In contrast, the Baseline approach tended to replicate document excerpts or prompt fragments more directly, maintaining elevated ROUGE/BERTScore values but occasionally missing opportunities to complete omitted information. This trade-off between lexical fidelity and semantic completeness represents a fundamental characteristic distinguishing the two architectural approaches.

The divergence between similarity metrics (favoring Baseline) and judge-based quality metrics (favoring Agent) underscores that textual similarity alone provides insufficient signal for evaluating structured extraction quality in legal domains. Effective metadata extraction requires not merely reproducing source text but rather identifying, organizing, and standardizing relevant information according to institutional schemas — a task better captured by rubric-based evaluation aligned with domain expertise.

Table 8 disaggregates LLM-as-a-Judge precision, recall, and F_1 by field and model family for orchestration variant C. At the global level, GPT-5 Baseline attains the highest F_1 (0.89), outperforming its agentic counterpart (0.87) by a narrow margin, while both Gemma3 families show a wider gap in favor of the Baseline. Field-level analysis reveals that the impact of orchestration is highly heterogeneous: for GPT-5, the Agent configuration improves *tema* (F_1: 0.98 $\rightarrow$ 0.99) and *termosAuxiliares* (F_1: 0.41 $\rightarrow$ 0.65), but degrades *notas* (F_1: 0.97 $\rightarrow$ 0.87) and *referenciasLegislativas* (F_1: 0.95 $\rightarrow$ 0.67). For smaller models, the agentic approach tends to improve recall at the expense of precision in most fields, with the largest negative impact observed in *informacoesComplementares* and *notas* for Gemma3-27B.

Synthesis: quality-cost trade-offs across model families. Cost–benefit trade-offs vary substantially by model. Robust models such as GPT-5 tend to benefit more from agentic decomposition because they better preserve task state across steps, follow iterative reviewer feedback, and maintain cross-field constraints under revision loops, whereas smaller models are more sensitive to context fragmentation and to the overhead introduced by multi-step control (Belcak et al., 2025).

We summarize this trade-off with $\gamma = (\Delta F_1 / F_1^{\text{Base}}) / (\Delta\tau / \tau_{\text{Base}})$, the quality–cost ratio. In the pilot setting, GPT-5 achieved $\gamma \approx 0.088$

Table 7: ROUGE-L F_1 vs. Judge F_1: Gemma3-12B (Base vs. Agent)

Field	ROUGE-L (Base)	ROUGE-L (Agent)	Judge F_1 (Base)	Judge F_1 (Agent)
informacoesComplementares	0.409	0.338	0.353	**0.402**
teseJuridica	0.451	0.458	0.764	**0.788**
notas	**0.683**	0.273	**0.668**	0.261

Table 8: Field-level precision, recall, and F_1 by model family and architecture – LLM-as-a-Judge, orchestration variant C. Support varies by model; see Table 2. Bold indicates the higher Base vs. Agent value within each field and model family.

Field	P (Base)	P (Agent)	R (Base)	R (Agent)	F_1 (Base)	F_1 (Agent)
GPT-5						
Global	**0.8677**	0.8235	0.9201	**0.9297**	**0.8918**	0.8724
informacoesComplementares	**0.9552**	0.9464	**0.9997**	0.9981	**0.9656**	0.9593
jurisprudenciaCitada	**0.9509**	0.9506	0.9704	**0.9868**	0.9580	**0.9671**
notas	**0.9735**	0.8448	0.9814	**0.9854**	**0.9653**	0.8706
referenciasLegislativas	**0.9507**	0.5735	**0.9612**	0.9276	**0.9467**	0.6688
tema	0.9822	**0.9916**	0.9937	**0.9992**	0.9799	**0.9927**
termosAuxiliares	0.3328	**0.5544**	0.5897	**0.8017**	0.4076	**0.6474**
teseJuridica	**0.9563**	0.9439	0.9887	**0.9962**	**0.9711**	0.9686
Gemma3-12B						
Global	**0.6061**	0.5533	0.6952	**0.7118**	**0.6432**	0.6183
informacoesComplementares	0.3414	**0.3833**	**0.9633**	0.9531	0.3525	**0.4020**
jurisprudenciaCitada	**0.7359**	0.6149	0.6768	**0.6915**	**0.6850**	0.6262
notas	**0.7237**	0.2047	0.8418	**0.8422**	**0.6678**	0.2605
referenciasLegislativas	**0.6470**	0.5866	0.7266	**0.7377**	**0.6402**	0.6080
tema	0.1875	**0.6929**	0.8782	**0.8955**	0.1863	**0.6575**
termosAuxiliares	0.2930	**0.3424**	0.4937	**0.5628**	0.3547	**0.4179**
teseJuridica	0.7204	**0.7224**	0.8422	**0.8905**	0.7638	**0.7875**
Gemma3-27B						
Global	**0.7471**	0.6105	0.7146	**0.7833**	**0.7260**	0.6822
informacoesComplementares	**0.8415**	0.2305	**0.9891**	0.9561	**0.8428**	0.2544
jurisprudenciaCitada	**0.8186**	0.7699	0.6938	**0.8041**	0.7264	**0.7668**
notas	**0.8069**	0.3369	0.8269	**0.9037**	**0.7076**	0.3919
referenciasLegislativas	**0.7144**	0.6027	0.7763	**0.8402**	**0.7016**	0.6666
tema	**0.9898**	0.5332	0.8887	**0.9065**	**0.8852**	0.5008
termosAuxiliares	0.3166	**0.3699**	0.5171	**0.6241**	0.3759	**0.4553**
teseJuridica	**0.8370**	0.7851	0.8249	**0.9281**	0.8175	**0.8435**

(each 1% token increase yields 0.088% F_1 improvement), while smaller models were markedly less favorable: $\gamma_{\text{Gemma3-27B}} \approx 0.010$ and $\gamma_{\text{Gemma3-12B}} \approx 0.016$. This indicates that model capacity moderates whether orchestration overhead is compensated by quality gains, so architecture selection must be conditioned on production constraints (cost, latency, and completion reliability).

These findings are not universal across orchestrations. In the full-scale re-execution, non-completion reduced the effective support for agentic configurations (as low as 779 for GPT-5 agents in variant C), and the direction of gains varied by orchestration variant: GPT-5 improved in some agentic strategies but not in all, and there are representative cases where the baseline outperforms the agentic workflow even for GPT-5 (e.g., $\Delta F_1 = -0.0194$ in variant C) despite reduced variance (Table 2).

Overall, JAMEX should be interpreted as a conditional approach: it can increase traceability and auditability, but it does not guarantee higher mean accuracy, particularly when decomposition fragments critical context or amplifies cross-field dependencies for smaller models. Across configurations, orchestration consistently increases token usage and reduces F_1/token efficiency, so deployment decisions should jointly balance accuracy, completion reliability, stability, and cost.

4 Conclusion

This work investigated JAMEX (Judicial Multi-Agent Metadata Extraction), a multi-agent architecture for pre-filling the *Espelho do Acórdão*, and compared it against a strong single-prompt baseline under an IR-only setting. The experimental protocol was validated in a pilot study (variant A) and re-executed on a stratified dataset of $n = 1{,}225$ Criminal Law appellate decisions; due to non-completion

in a subset of documents — predominantly in agentic configurations — effective support varied by model and orchestration variant (see Section 2).

Across experiments, agentic orchestration increased operational overhead and its impact on extraction quality was not uniformly positive across orchestration variants. While the pilot indicated higher mean F_1 under the agentic pipeline, re-execution showed that this effect can weaken or reverse depending on orchestration design and model family: GPT-5 benefited in some agentic strategies but not in all, and smaller models (Gemma3-12B/Gemma3-27B) showed no robust quality gains under agentic decomposition. Consequently, the evidence does not support a general rejection of H_0 across orchestrations, and H_1 is not confirmed as a general claim; rather, the results indicate that quality, robustness, and cost trade-offs depend on model capacity, task coupling, and the specific decomposition and review strategy.

Refinements motivated by agent design literature - memory, planning, and directed review - including richer validator memory (prior revision requests and explicit constraints from the institutional manual), improved traceability and reduced certain formatting and omission errors, but were insufficient to yield consistent accuracy gains across model families and orchestration variants. Practically, multi-agent decomposition appears most promising when subtasks are weakly coupled and the model reliably follows multi-step revision feedback; otherwise, a unified prompt that preserves global context can be preferable.

Future work should prioritize designs that explicitly model the interaction between task structure, model capacity, and orchestration strategy. A direct continuation is to formalize adaptive policies that decide when decomposition is warranted, which fields should be routed to specialist agents, and which model capacity is required to preserve global coherence while controlling token overhead. This naturally extends to hybrid and multi-model routing architectures, where the orchestrator selects the most cost-effective model per subtask under operational constraints, and to alternative dependency-management schemes that reduce cascade effects in sequential steps, potentially replacing brittle pipelines with more cooperative or jointly validated flows. Because smaller models did not show robust quality improvements under agentic decomposition — exhibiting lower F_1 and higher sensitivity to context fragmentation across orchestration variants — a complementary direction is task-specific adaptation of smaller models, including supervised fine-tuning on domain-specific data and schema-constrained objectives, aiming to improve extraction quality and instruction-following under iterative review. This direction is consistent with recent legal NLP work showing that LLM-based term/attribute extraction can be improved even when annotated resources are limited, provided the training objective is aligned with the extraction schema and domain constraints (Breton et al., 2025; Adhikary et al., 2023). As an extension, adopting multi-judge or "jury" evaluation protocols can mitigate single-evaluator variance and improve robustness (Verga et al., 2024).

Limitations

The study is limited by scope and by the evaluation regime. Restriction to Criminal Law appellate decisions improves homogeneity but limits external validity to other domains and monocratic decisions. Reference metadata sheets were generated by a proprietary model guided by expert prompts rather than direct human indexing, which may introduce systematic reference biases; in particular, prompt-generated references could favor baseline architectures whose outputs more closely mirror the same prompting style. Incorporating human-annotated ground truth would strengthen future comparisons. It is worth noting that a sample of extractions was performed and evaluated on the same legal corpus by experts during the development of the reference extraction prompt, providing partial — though not systematic — human validation of the reference outputs. Sequential dependencies can amplify errors across stages, and the observed sensitivity to context fragmentation highlights a risk of performance degradation when subtasks are strongly coupled. Finally, ROUGE and BERTScore have known limitations, and LLM-as-a-Judge remains partially subjective; reporting reliability analyses and triangulating with human judgments is necessary for stronger claims. Operationally, multi-agent orchestration increases token consumption and latency, and completion reliability varies across model families, requiring explicit cost-benefit and robustness analyses when scaling.

Acknowledgments

The authors thank João Paulo de Franco Alcantara (Jurisprudência, STJ) for providing the baseline ex-

traction prompt that served as a foundation for this study, and the Superior Tribunal de Justiça (STJ) for making judicial decisions and metadata publicly available through its Open Data Portal. The authors acknowledge the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) for its role in fostering graduate education and scientific research in Brazil. The authors also acknowledge the Pontifícia Universidade Católica do Paraná (PUCPR) for the academic environment and institutional support that enabled this research. Finally, the authors acknowledge Financiadora de Estudos e Projetos (Finep), grant FINEP ProInfra 2021 Ref: 259/2022.

References

Subinay Adhikary, Sagnik Das, Sagnik Saha, Procheta Sen, Dwaipayan Roy, and Kripabandhu Ghosh. 2023. Automated attribute extraction from legal proceedings. *Artificial Intelligence and Law*. Also available as arXiv:2310.12131.

Peter Belcak, Greg Heinrich, Shizhe Diao, Yonggan Fu, Xin Dong, Saurav Muralidharan, Yingyan Celine Lin, and Pavlo Molchanov. 2025. Small language models are the future of agentic ai. *arXiv preprint arXiv:2506.02153*.

Alessandro de Souza Bezerra and Luciane Cavalcante Lopes. 2025. Controle externo baseado em llm: AvaliaÇÃo empÍrica do prume ai. *ARACÊ*, 7(9):e8409.

Julien Breton, Mokhtar Mokhtar Billami, Max Chevalier, Ha Thanh Nguyen, Ken Satoh, Cassia Trojahn, and May Myo Zin. 2025. Leveraging llms for legal terms extraction with limited annotated data. *Artificial Intelligence and Law*.

Conselho Nacional de Justiça. 2024. *O uso da inteligência artificial generativa no Poder Judiciário brasileiro: relatório de pesquisa*. CNJ, Brasília.

Diego de Vargas Feijó and Viviane Pereira Moreira. 2018. Rulingbr: A summarization dataset for legal texts. In *Computational Processing of the Portuguese Language: 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings*, volume 11122 of *Lecture Notes in Computer Science*, pages 255–264, Cham. Springer.

Liming Dong, Qinghua Lu, and Liming Zhu. 2024. Agentops: Enabling observability of llm agents. *arXiv preprint arXiv:2411.05285*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81. Association for Computational Linguistics.

Alex Aguiar Lins, Cecilia Silvestre Carvalho, Francisco Das Chagas Jucá Bomfim, Daniel de Carvalho Bentes, and Vládia Pinheiro. 2024. CLSJUR.BR - a model for abstractive summarization of legal documents in Portuguese language based on contrastive learning. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 321–331, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522. Association for Computational Linguistics.

Érica Barbosa Sousa Moreira. 2023. A base de dados de jurisprudência do stj: histórico, estrutura e critérios de catalogação. Dissertação (mestrado profissional em direito, regulação e políticas públicas), Universidade de Brasília, Brasília. Acesso em: 27 set. 2025.

Lucas M. Navarezi, Kenzo Miranda Sakiyama, Lucas de Souza Rodrigues, Caio M. O. Robaldo, Gustavo Rocha Lobato, Paulo Arantes Vilela, Edson Takashi Matsubara, and Eraldo R. Fernandes. 2022. Entity extraction from portuguese legal documents using distant supervision. In *Computational Processing of the Portuguese Language - 15th International Conference, PROPOR 2022, Fortaleza, Brazil, March 21–23, 2022, Proceedings*, volume 13208 of *Lecture Notes in Computer Science*, pages 166–176. Springer.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. In *Advances in Neural Information Processing Systems*, volume 36, pages 37654–37688.

Superior Tribunal de Justiça. 2021. *Manual de Inclusão de Acórdãos na Base de Dados*. Secretaria de Jurisprudência, Brasília.

Superior Tribunal de Justiça. 2022. Portal de dados abertos do stj. Disponibiliza dados de interesse público do Tribunal em formato aberto e legível por máquina. Acesso em: 04 dez. 2025.

Fernanda Suriani and Eduardo Pacheco. 2025. Transforming justice: The rise of ai in brazilian courts. In *Proceedings of the 26th Annual International Conference on Digital Government Research (dg.o 2025)*. Digital Government Society.

Pat Verga, Sebastian Hofstätter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. 2024. Replacing judges with juries: Evaluating llm generations with a panel of diverse models. *arXiv preprint arXiv:2404.18796*.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, and 1 others. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Zihan Yu, Liang He, Zhen Wu, Xinyu Dai, and Jiajun Chen. 2023. Towards better chain-of-thought prompting strategies: A survey. *arXiv preprint arXiv:2310.04959*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations (ICLR)*.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2025. Judgelm: Fine-tuned large language models are scalable judges. In *International Conference on Learning Representations (ICLR)*. Spotlight Presentation.