Aintelia Science Notes

ASNJ • BAUDER PRESS

Research Article

From Unstructured Documents to Intelligent Insights: An End-to-End AI Pipeline for Global Tax Compliance

Muhammed Şara[1,*], Ahmet Ak[2], Ayhan Boyacıoğlu[3], Yunus Taştutan[4]

1, 2, 3, 4 Sovos/F.I.T. R&D Center, İstanbul, Türkiye

Correspondence to:
Muhammed Şara –
muhammed.sara@sovos.com
+90 539 714 6546

How to Cite:
Actual Citation as taken from website

RECEIVED: 28/04/2026
ACCEPTED: 08/05/2026
PUBLISHED: 30/06/2026

ABSTRACT

Over 60 countries now mandate electronic invoicing, and this number has since exceeded 100 countries. Turkey is among the earliest adopters, with the Turkish Revenue Administration (GİB) generating more than 2 billion documents annually. For enterprises operating across multiple jurisdictions, meeting these obligations creates serious operational and legal challenges that become increasingly difficult to manage as transaction volumes increase. Traditional rule-based approaches were not built for this scale; they struggle with heterogeneous input formats and cannot easily absorb changes in jurisdiction-specific tax rules. This study presents an end-to-end artificial intelligence pipeline that addresses these challenges by transforming unstructured invoice documents (PDFs, images, and CSV files) into actionable compliance insights within a single, unified architecture. The pipeline integrates three tightly coupled components. The Smart Input Module uses a competitive ensemble of large multimodal models: Google Gemini Flash, Amazon Nova Lite, and OpenAI GPT-4.1 to extract tax-relevant fields from diverse document formats, achieving 92% overall extraction accuracy and 98% accuracy on legally critical fields, such as tax identification numbers and monetary amounts. The Cloud Processing Layer introduces a domain-specific XSLT deduplication technique that exploits a structural regularity unique to Turkish e-invoices: embedded rendering stylesheets that are nearly identical across documents from the same taxpayer. By replacing duplicate stylesheets with compact SHA-256 hash references while reinserting exact original bytes during reconstruction to preserve digital signature integrity, the system achieves 90% compression on document content, contributing to a 63% reduction in total storage footprint and extending the infrastructure runway from 18 months to over five years. An event-driven architecture built on RabbitMQ and Kubernetes autoscaling raises the peak throughput from 32 to 200 transactions per second, a 6.25× improvement. Redis caching further reduced the taxpayer lookup latency from 10 ms to 0.9 µs. The Sovi Intelligence Engine completes the pipeline by enabling natural language querying of processed tax data in Turkish, English, German, French, and Spanish. Based on an agentic architecture, Sovi decomposes each query into sub-steps: intent interpretation, schema mapping, and SQL generation, guided by a retrieval-augmented generation workflow that maps domain terminology to database columns. In the evaluation, the system achieved a 94% valid SQL generation rate and 82% semantic accuracy, with a mean response latency of 3.2 s and a user satisfaction score of 4.1 out of 5. Running in production today, the pipeline demonstrates that fully automated, AI-driven tax compliance is not a future aspiration but a current reality in Turkey.

KEYWORDS

Document AI, Tax Compliance, Large Language Models, Natural Language Processing, Electronic Invoicing

1. Introduction

Over 60 countries now mandate electronic invoicing, and this number has since exceeded 100 countries [1]. Turkey is among the earliest adopters; the Turkish Revenue Administration (GİB) requires all businesses above certain revenue thresholds to issue invoices electronically through its platform, generating more than 2 billion documents annually. For enterprises operating across multiple jurisdictions, meeting these obligations creates serious operational and legal challenges that become increasingly difficult to manage as transaction volumes increase.

Traditional approaches, such as manual data entry combined with rule-based validation, were not built for this scale. They struggle with the range of input formats encountered in practice, from scanned paper invoices to structured XML, and cannot easily absorb changes in jurisdiction-specific tax rules. Recent developments in document AI and large language models (LLMs) suggest a different path, where the extraction, validation, and analysis of tax documents can be handled with far less human intervention [2,3,4,5].

This paper addresses this fragmentation by presenting a unified pipeline, deployed in production at Sovos, that spans the full compliance lifecycle—from raw document ingestion to natural language querying of processed tax records—integrating extraction, storage, and analytics within a single architecture.

2. Materials and Methods

The system consists of three modules: Smart Input, Cloud Processing, and Sovi Intelligence Engine, connected through a RabbitMQ event bus (see Figure 1). The evaluation uses a production dataset of 5,000 invoices drawn from approximately 1,000 unique recipient taxpayers and 50-100 distinct issuing companies, with the ground truth established through manual verification followed by automated validation tools.

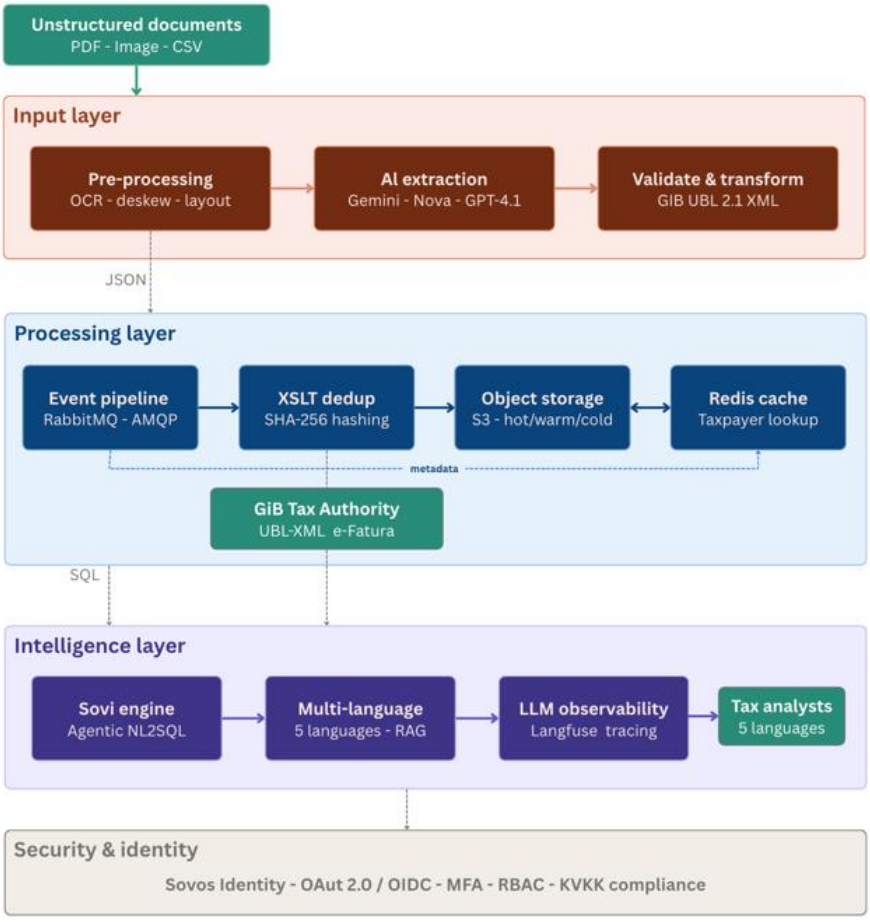

Figure 1: *End-to-end AI pipeline for tax compliance: Input Layer (orange), Processing Layer (blue), Intelligence Layer (purple), and Security Layer (gray).*

Smart Input Module

The Smart Input module transforms unstructured documents (PDF, JPEG/PNG, CSV) into GİB-compliant UBL 2.1 XML format. A format detection layer directs each document to the corresponding processing path based on the format and quality metrics. For image-based inputs, a three-stage pre-processing sequence applies deskewing, noise reduction, and contrast normalization before the layout analysis segments the document into header, line-item, and summary regions.

We evaluated three model providers on 5,000 production invoices: Google Gemini Flash, Amazon Nova Lite, and OpenAI GPT-4.1. Each extracted JSON output is then validated through check-digit verification of tax IDs (VKN/TCKN), cross-reference against the GİB taxpayer registry, and semantic consistency checks against Turkish VAT rules.

Cloud Processing Layer

The processing layer was built around an event-driven architecture using RabbitMQ (AMQP). Because the stages are decoupled, each stage can be scaled independently, which is a critical property during month-end demand spikes when transaction volumes rise sharply and unpredictably. Kubernetes Horizontal Pod Autoscaling adjusts the worker count from 4 to 24 replicas in response to the queue depth and CPU utilization metrics.

A significant optimization opportunity arose from the XSLT stylesheets embedded in Turkish e-invoices. These stylesheets allow recipients to view invoices in a browser without specialized software, but they are nearly identical across documents from the same taxpayer, a structural

regularity that is invisible to general-purpose compression algorithms. Our content-addressable storage strategy computes the SHA-256 hash of each stylesheet and replaces the duplicates with a compact reference pointer, storing the unique stylesheets separately for reconstruction. Reconstruction reinserts the exact original bytes, preserving the byte-level digital signature integrity, which was validated using the official GİB e-Fatura Görüntüleyici tool. Document metadata is persisted to PostgreSQL via the Outbox Pattern and replicated asynchronously to S3-compatible object storage with hot/warm/cold tiering (0–30, 30–365, 365–3,650 days).

Sovi Intelligence Engine

The Sovi Intelligence Engine provides natural language analytics for processed tax data. Rather than a single-step translation from query to SQL, Sovi decomposes each request into sub-steps: first interpreting the user's intent, then mapping it to the relevant schema elements, and finally generating and executing the SQL, following an established query decomposition approach [6, 7]. A Retrieval-Augmented Generation (RAG) workflow guides this categorization step, supplying a domain vocabulary that maps Turkish tax terminology – KDV (VAT), mükellef (taxpayer), stopaj (withholding) – to the corresponding database columns [8]. The system accepts queries in Turkish, English, German, French, and Spanish, and all LLM interactions are monitored via Langfuse for trace logging, cost-tracking, and quality metrics [9].

3. Result

Smart Input Evaluation

As Table 1 shows, all three providers hit 98% accuracy on critical fields – the ones that determine legal validity – although their overall scores and costs diverge considerably. Gemini Flash leads on overall accuracy at 92%, while Nova Lite offers the lowest cost at $0.02 per document, compared to $0.055 for GPT-4.1. Line-item extraction was the most challenging task, reaching only 89.3% because of merged cells and inconsistent column ordering across templates. Interestingly, errors showed no preference for any particular industry; they clustered around documents with unusual layouts regardless of the sector.

Table 1: *AI Model Comparison for Invoice Extraction*

Metric	Gemini Flash	Nova Lite	GPT-4.1
Overall Accuracy	92%	90%	89%
Critical Field Accuracy	98%	98%	98%
Avg. Processing Time	4.5 s	5.2 s	4.2 s
Cost per 1,000 docs (API)	$40	$20	$55

Processing Layer Evaluation

Migrating from batch processing to the event-driven architecture increased peak throughput from 32 to 200 TPS, a 6.25× improvement. Scaling is near-linear up to 16 workers, with diminishing returns beyond 20 workers attributable to HTTP server and CPU resource boundaries rather than database lock contention, indicating that the bottleneck is computational rather than state-dependent, and is addressable through standard horizontal scaling.

Table 2: *Storage Optimisation Results*

Metric	Before	After
Total data volume	150 TB	55 TB
Monthly growth rate	3 TB/month	650 GB/month
Full backup duration	4 days	1.5 days
Query latency (taxpayer lookup)	10 ms	0.9 µs

XSLT deduplication alone achieves a 90% size reduction in document content, outperforming general-purpose XML compression, which yielded only a 4× ratio on the same corpus. Combined with tiered storage and metadata optimization, the total volume decreased from 150 to 55 TB. Monthly growth dropped by 78%, extending the infrastructure runway from 18 months to over five years. Redis caching reduced taxpayer lookup latency by more than four orders of magnitude, from 10 ms to 0.9 µs.

Figure 2 illustrates the scaling behavior of the optimized system relative to the static baseline.

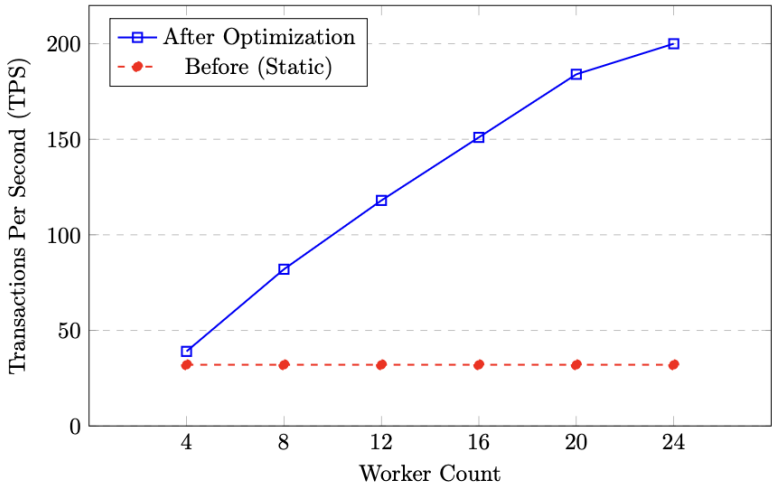

Figure 2: *System throughput scaling with the worker count. The optimised system shows near-linear scaling up to 16 workers.*

Sovi Intelligence Evaluation

Table 3 reports the Sovi performance metrics based on the model generation time as the primary latency measure, with user satisfaction derived from the Net Promoter Score responses normalized to a 5-point scale.

Table 3: *Natural Language Query Evaluation*

Metric	Value
Valid SQL syntax generation rate	94%
SQL semantic accuracy	82%
Avg. response latency (model generation)	3.2 s
User satisfaction (1–5, NPS-derived)	4.1

The 82% semantic accuracy is comparable with the performance of leading NL2SQL systems on domain-specific benchmarks [10]. The 12-point gap between syntactic validity (94%) and semantic accuracy (82%) reflects ambiguous GROUP BY selections in aggregation queries and off-by-one boundary errors in temporal comparisons. English queries (80% of the volume) achieved the highest success rate at 88%, followed by Turkish (15%) at 81%, while German/French/Other (5%) reached 61% likely reflecting the smaller test set for those languages rather than an inherent capability gap, though further evaluation is needed to confirm this.

4. Discussion and Conclusion

The most notable finding is that model selection has little bearing on legally critical fields: all three providers achieved 98% accuracy on tax IDs and monetary amounts. Differences emerge on less structured content such as line items, where cost-accuracy trade-offs become operationally significant. The second finding concerns storage: domain-specific XSLT deduplication achieved 10× compression on the same data that general-purpose algorithms compressed by only 4×, demonstrating that domain-specific regulatory knowledge can unlock optimisations that general-

purpose compression tools cannot achieve. Finally, the move to an event-driven architecture was not a marginal improvement; it transformed a system that topped out at 32 TPS regardless of added resources into one that scaled linearly with worker count up to 200 TPS.

An earlier project, the Indirect Tax Analytics (ITA) Framework, spent 18 months attempting to map data from over 30 global tax systems into a single canonical schema before the effort was discontinued. The maintenance burden grew faster than the value delivered: each new country or document type added not only new mappings but also new conflicts with existing ones. The Sovi engine takes the opposite approach, providing natural language abstraction over heterogeneous data stores at query time, rather than forcing schema homogeneity upfront. The result was a system that reached production in eight months and has remained stable since.

This production deployment demonstrates that end-to-end AI-driven tax compliance is technically feasible and operationally viable at enterprise scale. Priority directions for future work include proactive anomaly detection to shift the system from reactive audit support toward pre-emptive risk management, and multi-jurisdictional expansion as the EU VAT in the Digital Age (ViDA) initiative extends comparable mandates to European markets.

STATEMENTS AND DECLARATIONS

Supplementary Material

No supplementary material accompanies this article.

Acknowledgements

The authors thank the engineering team at Sovos Turkey for their contributions to the system development.

Author Contributions

M. Şara: Conceptualization, Methodology, Software, Investigation, Writing – Original Draft, Writing – Review & Editing. A. Ak: Software, Validation, Data Curation. A. Boyacıoğlu: Software, Investigation, Visualization. Y. Taştutan: Methodology, Formal Analysis, Writing – Review & Editing.

Conflicts of Interest

The authors declare that there is no conflict of interest. The authors are affiliated with Sovos/F.I.T. R&D Center, whose production system is described in this study.

Data, Code & Protocol Availability

The invoice dataset used in this study contains proprietary and confidential taxpayer information and cannot be publicly shared. Aggregated results and methodology details are available from the corresponding author upon reasonable request.

Funding Received

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

ORCID IDs

Muhammed Şara https://orcid.org/0000-0002-8353-3544
Ahmet Ak https://orcid.org/0009-0001-2660-860X
Ayhan Boyacıoğlu https://orcid.org/0009-0001-2906-2526
Yunus Taştutan https://orcid.org/0009-0001-1262-1838

REFERENCES

1. Billentis, "E-invoicing/e-billing: Digitisation & automation," Market Rep., 2023. [Online]. Available: https://www.billentis.com/
2. Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, and M. Zhou, "LayoutLM: Pre-training of text and layout for document image understanding," in Proc. 26th ACM SIGKDD Conf. Knowl. Discov. Data Min., 2020, pp. 1192–1200, doi: 10.1145/3394486.3403172.
3. Y. Huang, T. Lv, L. Cui, Y. Lu, and F. Wei, "LayoutLMv3: Pre-training for document AI with unified text and image masking," in Proc. 30th ACM Int. Conf. Multimedia, 2022, pp. 4083–4091, doi: 10.1145/3503161.3548112.
4. D. Wang, N. Ravi, C. Arber, S. Bhatia, and A. Chakravarthi, "DocLLM: A layout-aware generative language model for multimodal document understanding," in Proc. ACL, 2024, pp. 2551–2568.
5. L. Li, Y. Zhang, and L. Chen, "A survey on large language models for recommendation," World Wide Web, vol. 27, no. 5, pp. 1–35, 2024, doi: 10.1007/s11280-024-01291-2.
6. A. Plaat, T. Muller, and J. van den Herik, "Agentic large language models: A survey," J. Artif. Intell. Res., vol. 82, pp. 345–412, 2025, doi: 10.1613/jair.1.16853.
7. M. Pourreza and D. Rafiei, "DIN-SQL: Decomposed in-context learning of text-to-SQL with self-correction," in Adv. Neural Inf. Process. Syst., vol. 36, 2023.
8. A. Plaat, A. Wong, and J. van den Herik, "Multi-step reasoning with large language models: A survey," ACM Comput. Surv., vol. 57, no. 3, pp. 1–38, 2024, doi: 10.1145/3717843.
9. Langfuse GmbH, "Langfuse: Open source LLM engineering platform," Software documentation, 2023. [Online]. Available: https://langfuse.com/
10. J. Li et al., "Can LLM already serve as a database interface? A big bench for large-scale database grounded text-to-SQL," in Adv. Neural Inf. Process. Syst., vol. 37, 2024.