A Greek Government Decisions Dataset for Public-Sector Analysis and Insight

Giorgos Antoniou Giorgos Filandrianos
Aggelos Vlachos Giorgos Stamou
National Technical University of Athens, Greece
georgiosantoniou@mail.ntua.gr,geofila@ails.ece.ntua.gr
aavlachos@cslab.ece.ntua.gr,

Lampros Kollimenos
University of Thessaly, Greece
lkollimenos2001@gmail.com

Konstantinos Skianis
BLUAI, Greece
skianis.konstantinos@gmail.com

Michalis Vazirgiannis
Ecole Polytechnique, France
mvazirg@lix.polytechnique.fr

Abstract

We introduce an open, machine-readable corpus of Greek government decisions sourced from the national transparency platform *Diavgeia*[1]. The resource comprises 1 million decisions, featuring and high-quality raw text extracted from PDFs. It is released with raw extracted text in Markdown format, alongside a fully reproducible extraction pipeline.

Beyond the core dataset, we conduct qualitative analyses to explore boilerplate patterns and design a retrieval-augmented generation (RAG) task by formulating a set of representative questions, creating high-quality answers, and evaluating a baseline RAG system on its ability to retrieve and reason over public decisions. This evaluation demonstrates the potential of large-scale public-sector corpora to support advanced information access and transparency through structured retrieval and reasoning over governmental documents, and highlights how such a RAG pipeline could simulate a chat-based assistant capable of interactively answering questions about public decisions.

Due to its scale, quality, and domain coverage, the corpus can also serve as high-value pre-training or fine-tuning material for new Language Models (LMs) and Large Language Models (LLMs) respectively, including specialized models for legal and governmental domains, and as a foundation for novel approaches in domain adaptation, knowledge-grounded generation, and explainable AI. Finally, we discuss limitations, outline future directions, and make both the data and the code accessible [2].

CCS Concepts

• **Computing methodologies** → **Language resources**; • **Information systems** → **Data extraction and integration**.

Keywords

open government data, government transparency, public-sector AI

1 Introduction

Transparency, accountability, and data-driven supervision of public institutions are made possible by open government data, which has emerged as a key component of contemporary democratic governance. Governments can lessen information asymmetries, empower civil society, and encourage evidence-based policy evaluation by making administrative records accessible and machine-readable [4, 30]. Large-scale open-data programs throughout Europe and beyond have shown that open disclosure of official choices may boost anti-corruption efforts, increase public service quality, and build confidence [9]. One notable example of such institutional commitments is Greece's national transparency platform *Diavgeia*, which requires the disclosure of all administrative decisions made by public agencies, municipalities, and ministries.

Documents and data provided by *Diavgeia* differ from Greek law datasets because they capture the everyday operational decisions of the public administration, not the formal legal framework enacted by Parliament. Some well-known and widely used publicly available Greek law datasets and legislation repositories include: i) "Published laws from the National Printing Office[3]" includes metadata, issue numbers, etc.; ii) The official legislation search portal of Hellenic Parliament[4] for finding laws, bills, and enacted legislation; iii) the National Printing Office of Greece (Official Gazette / FEK)[5] search interface — for browsing the full official Government Gazette issues; iv)A consolidated-law dataset on Hugging Face: "Permanent Greek Legislation Code - Raptarchis[6]", a dataset consisting of approximately 47k legal resources of Greek legislation from 1834 to 2015 and is suitable for research/machine learning (ML) tasks. Laws and presidential decrees are structured, hierarchical, stable legal texts produced through a formal legislative process, while *Diavgeia* contains ministerial acts, procurement decisions, budget allocations, hiring approvals, board resolutions, and thousands of other administrative actions issued continuously by agencies across the state. As a result, *Diavgeia* is massive, dynamic, heterogeneous, and closer to administrative reality, whereas laws represent the normative legal order. This makes *Diavgeia* uniquely valuable for transparency, auditing, and empirical governance analysis, offering a granular, real-time view of how the public sector operates—far beyond what legal corpora alone can provide.

However, the existence of open data alone does not guarantee accessibility or meaningful public use. Government documents are

[1] https://diavgeia.gov.gr/
[2] https://github.com/y3nk0/diavgeia
[3] https://data.gov.gr
[4] https://www.hellenicparliament.gr/en/nomothetiko-ergo/anazitisi-nomothetikou-ergou
[5] https://search.et.gr/en/search-legislation
[6] https://huggingface.co/datasets/AI-team-UoA/greek_legal_code

often dispersed across formats such as PDFs and JSONs, inconsistently structured, and difficult for non-experts to interpret. Recent advances in natural language processing (NLP), information extraction, and retrieval systems provide new opportunities to transform raw administrative text into searchable, understandable, and actionable knowledge [5, 6]. Systems that leverage domain-adapted language models, question answering, or retrieval-augmented generation can substantially improve how citizens navigate public decisions, how journalists investigate patterns of spending or policy shifts, and how researchers analyze institutional behavior [15]. When deployed responsibly, such technologies offer a path toward more inclusive digital governance and stronger democratic participation.

In this work, we introduce the first large-scale, curated, and machine-readable corpus of Greek government decisions sourced from the *Diavgeia* platform. Our main contributions are:

- **A large-scale open dataset:** 1 million government decisions with normalized metadata, canonical identifiers, and high-quality Markdown text extracted from PDFs/JSONs, released in interoperable formats.
- **A fully reproducible extraction pipeline:** End-to-end tooling for harvesting, cleaning, normalizing, and structuring *Diavgeia* documents at scale.
- **Boilerplate analysis:** Empirical exploration of administrative patterns, decision types, institutional behaviors, and textual characteristics across agencies.
- **A retrieval-augmented generation (RAG) task:** A benchmark comprising representative questions, curated answers, and baseline RAG performance over the dataset.

This combination of open data infrastructure, analytical insights, and practical tools establishes a foundation for future work in legal NLP, digital governance, and democratic technology.

2 Related Work

Open government data has long been recognized as a foundation for transparency, accountability, and civic engagement. Early work established that making governmental documents publicly accessible enables independent auditing, reduces information asymmetries, and fosters trust in democratic institutions [4, 30]. Large-scale initiatives such as the European Data Portal have demonstrated the importance of standardisation and interoperability in realising the full societal value of public data [9]. In the Greek context, the *Diavgeia* transparency portal represents one of the most ambitious national-level efforts, mandating the publication of all administrative decisions. Yet despite its legal significance and substantial scale, technical exploitation of its content has historically been limited.

Early efforts to exploit *Diavgeia* leaned primarily on semantic web tooling rather than full NLP. The civic-tech platform *PublicSpending.gr* semantically processed daily spending feeds to generate citizen-facing visualisations, although it focused mainly on descriptive analytics rather than deeper modelling [25]. More recently, the *d2kg* project proposed an OWL ontology that transforms *Diavgeia* decisions into a linked-data graph, enabling SPARQL-based retrieval across acts and legal references, yet leaving textual patterns, administrative intent, and authorial practice largely unexplored [20]. Comparable open-government corpora abroad provide methodological blueprints that *Diavgeia* research can adopt. The *TheyBuyForYou* knowledge graph integrates European procurement notices to study anomalies in public spending [22], while Germany's NIF4OGGD converts federal open-data dumps into the NLP Interchange Format, improving cross-dataset search and text reuse detection [21]. Rich legislative speech resources such as GerParCor Reloaded [1] and the ParlaMint Widened FOIA corpus [27] further illustrate how consistent schema design and preprocessing unlock large-scale diachronic political analyses.

Parallel to these open-data initiatives, the NLP community has made significant advances in processing legal and governmental documents. Recent work in legal NLP demonstrates the effectiveness of machine learning for multi-label classification, named-entity recognition, argument extraction, and domain-specific information retrieval [5]. The growing availability of legal benchmarks, such as LexGLUE [6], and the emergence of retrieval-augmented generation (RAG) systems tailored to complex knowledge-intensive tasks [15] have created opportunities to build citizen-oriented tools that simplify legal language and support public understanding of institutional decision-making.

Although *Diavgeia* represents a landmark for transparent governance, broader challenges identified in e-government studies remain relevant. Recent work highlights persistent issues around user experience, interface complexity, and data accessibility in general across public digital services [3, 24]. Many individuals struggle to navigate large document collections or locate relevant decisions, while users with disabilities or limited digital literacy face additional barriers. These usability constraints limit the platform's effectiveness in promoting transparency and enabling meaningful civic participation—underscoring the need for machine-readable resources, robust NLP pipelines, and citizen-facing retrieval or summarization tools.

3 Dataset

In this Section, we present the dataset, extraction techniques, final extracted data, and computed statistics.

3.1 *Diavgeia* API and Initial Data

Currently *Diavgeia* offers more than 71 million decisions. Using the *Diavgeia* OpenData API[7], a user is allowed to download a number of decisions, respecting specific request limits. The API provides a structured dataset comprising essential details of governmental decisions, including administrative metadata, decision dates, and legal references. This structured format ensures the data's reliability and accessibility, facilitating comprehensive analysis and insights into governmental operations. Each decision comes with structured info in an JSON format, along with the PDF. The main structure fields of a decision in the API is shown in Table 7 of the Appendix. We downloaded 1 million decisions from the year 2021, with the total size of PDFs and metadata going over 1TB of storage space. It is obvious that the size of this dataset, make it very difficult to be handled, and thus transforming it to text is the best approach to manage it efficiently. The size of the final extracted text files in

[7] https://diavgeia.gov.gr/api/help

compressed format is 4GBs, making it considerably more manageable than the original PDFs. We note that the data include personal names as well as names of entities and organizations. Since these records are publicly available, we do not perform any removal of personally identifiable information.

3.2 Text Extraction Comparative Analysis

Since the extraction process is very important in datasets like these, we also present a text extraction comparative analysis of different approaches. In order to further understand the text extraction process, we compare the performance of four methods: i) Tesseract OCR (CPU)[8], ii) Paddle OCR (GPU)[9], iii) Direct PDF Extraction (PyMuPDF4LLM[10]), and iv) an LLM-generated "Ground Truth" (GPT-5-nano), on a dataset of 200 Greek administrative PDF files. The primary goal was to find the most accurate method for text extraction, given the known issues with broken encoding in the source PDFs, which makes direct extraction unreliable. Due to the size of dataset, the approach we used PyMuPDF4LLM, mainly for two reasons: i) it was considerably faster compared to the others, and ii) the output is given in a Markdown format, which is considered optimal for training language models. An extensive analysis of time performance is presented in Table 8 of the Appendix.

3.3 Corpus-level Statistics

Table 1: Corpus Statistics for 1M *Diavgeia* documents.

Metric	Value
Number of Documents	1,061,438
Total Tokens	1,495,839,382
Unique Documents Read	1,061,438
Average Tokens per Document	5,455.62
Median Tokens per Document	2,806
Standard Deviation (Tokens per Document)	36,414.89
Maximum Tokens in a Document	19,169,263
Total Characters	2,005,743,579
Average Characters per Document	5,936.01
Total Sentences	38,740,934
Average Sentences per Document	36.50

The corpus statistics were computed using a high-performance processing pipeline designed for large-scale text analytics. All documents were pre-extracted as plain text, and tokenization was performed with the Rust-optimized tiktoken[11] library, which provides fast, LLM-compatible subword tokenization. To handle over one million files efficiently, we employed Python multiprocessing with batch dispatching across all CPU cores, enabling parallel reading and tokenization of documents. Additional metrics—such as character counts and sentence estimates—were computed using lightweight regular-expression processing to avoid the overhead of full NLP parsing. The combination of Rust-based tokenization and multi-core processing ensured that the entire *Diavgeia* corpus could be analyzed efficiently at scale.

The *Diavgeia* corpus comprises over 1 million documents totaling approximately 1.5 billion tokens, making it one of the largest publicly available Greek-language administrative datasets. The average document length is substantial (∼5,456 tokens), with a median of 2,806 and a very high maximum of 19 million tokens, reflecting the presence of long legal annexes, procurement contracts, technical descriptions, and multi-page governmental reports. The high standard deviation (36,415 tokens) confirms the extreme heterogeneity of document sizes, ranging from minimal administrative notices to highly detailed financial or regulatory texts. The corpus also contains over 2 billion characters and nearly 39 million sentences, indicating rich linguistic variety and significant complexity in sentence structure. All extracted statistics are shown in Table 1.

From an NLP perspective, these characteristics make *Diavgeia* particularly valuable for training and evaluating models on long-context processing, financial reasoning, multi-document aggregation, legal-style summarization, and entity extraction. The scale of 1.5B tokens is more than sufficient for continued pretraining or domain adaptation of modern small-to-medium LLMs, significantly improving performance on Greek public-administration tasks. At the same time, the heterogeneity and vocabulary richness pose challenges for tokenization, retrieval, and hallucination control, making this corpus a strong benchmark for robustness, RAG evaluation, and long-form sequence modeling. In short, the *Diavgeia* corpus is large and diverse enough to meaningfully support both applied NLP systems and research into domain specialization for LLMs.

Table 2: Top 5 organizations by number of *Diavgeia* decisions

Organization	# Docs
Aristotle University of Thessaloniki	24,997
National & Kapodistrian University of Athens	22,235
National Technical Chamber	17,412
Unified Social Security Institution	16,385
Ministry of Culture and Sports	15,383

Table 2 shows the top 5 organizations by number of *Diavgeia* decisions. These organizations are all large, central actors in the Greek public sector: two major universities (1st and 2nd), the national technical chamber (3rd), the unified social security institution (Εθνικός Φορέας Κοινωνικής Ασφάλισης in Greek) coming 4th, and the Ministry of Culture and Sports (Υπουργείο Πολιτισμού και Αθλητισμού) as 5th. They manage substantial budgets, infrastructure, and personnel, and are involved in continuous operational activity (procurement, project funding, appointments, committees, travel approvals, cultural events, maintenance, etc.). Because *Diavgeia* mandates publication of a wide range of such administrative acts, high-volume institutions naturally generate a very large number of decisions, which explains why they dominate the corpus in terms of document count.

4 Qualitative Analysis: Boilerplate Study

Boilerplate removal plays a critical role in transforming raw administrative documents into information that is more accessible

[8] https://github.com/tesseract-ocr/tesseract
[9] https://github.com/PaddlePaddle/PaddleOCR
[10] https://pymupdf.readthedocs.io/en/latest/pymupdf4llm/
[11] https://github.com/openai/tiktoken

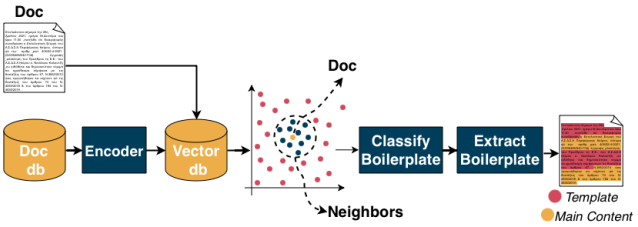


Figure 1: Our Boilerplate detection pipeline.

and meaningful for end users. Government decisions published on *Diavgeia* are often dense, technical, and formatted primarily for bureaucratic workflows rather than public readability, resulting in long documents where essential information is buried beneath recurring legal templates, procedural phrasing, and structural overhead. While boilerplate extraction has been explored in English [14, 28], the space remains significantly underexplored in for Greek administrative text, where document patterns, legal phrasing, and structural conventions differ markedly. An effective boilerplate removal system could automatically identify and separate core decision content from recurring legal scaffolding, enabling users to focus on the substantive elements of each document without navigating repetitive language. In this work, we introduce a Greek-language boilerplate extraction approach using LLM-based modelling, aiming to reduce textual friction and improve the interpretability of public-sector documents at scale.

4.1 Boilerplate Algorithm

Our method is illustrated in Figure 1. Each document is first stored in a structured document repository, from which its full textual stream is processed through an encoder to obtain fixed-length dense embeddings[12] . These representations are indexed into a vector database, enabling efficient k-NN retrieval across the corpus. At inference time, a query document is projected into the same embedding space and its nearest neighbours are obtained via cosine similarity. Semantically proximal documents are expected to share recurring administrative phrasing and structural patterns, allowing neighbourhood density to function as an implicit indicator of boilerplate presence.

Crucially, we intentionally avoid lexical similarity metrics such as keyword overlap, BM25 scoring [8], or edit-distance measures (e.g., Levenshtein), as these techniques overweight surface-form correspondence (see more in Section. Specifically, boilerplate templates can be short in proportion to the varying main body, which means lexical comparators tend to be dominated by document-specific content rather than structural repetition. As a result, vocabulary - based signals amplify noise instead of revealing template-consistency. Semantic embeddings, however, are better suited to the task: even when phrasing varies, documents derived from the same template frequently reference to similar public organisations, statutes, funding mechanisms or procedural entities. Thus, shared meaning rather than shared wording becomes the discriminative feature, enabling template-based clustering to emerge more naturally in embedding space 4.3.). The final processing stage consists of two LLM-based modules. The first model performs boilerplate classification, estimating the likelihood that the source document (or a segment thereof) originates from a reusable administrative template. The second model conducts boilerplate extraction, returning a segmented form of the input text in which each span is labelled either as template boilerplate or main content. Unlike highlight-based visual presentation, this stage yields a structural decomposition of the document, explicitly disentangling invariant legal scaffolding from case-specific informational material. Importantly, segmentation is carried out dynamically on a per-document basis, acknowledging that template magnitude, phrasing and placement vary significantly across decisions.

[12]Embeddings were generated using all-MiniLM-L6-v2.

4.2 Evaluation

To assess the effectiveness of our boilerplate extraction framework, we design an intrinsic evaluation procedure grounded in the following hypothesis. If two documents share the same boilerplate template and this template is correctly identified, then swapping the extracted main content between them should yield two documents structurally identical to the originals.

To obtain a reliable ground-truth basis for testing this hypothesis, we first conducted a manual annotation phase. Human annotators examined a subset of *Diavgeia* decisions and grouped together documents that were judged to share the same underlying administrative template. From these human-validated template clusters, we constructed an evaluation sample consisting of 100 document pairs.

For each document, our algorithm then extracts the boilerplate skeleton and the corresponding main content. We subsequently perform cross-swapping of the content segments within each pair. Under correct boilerplate segmentation, the reconstructed document should closely approximate its counterpart, since both are expected to differ only in their variable, non-boilerplate sections.

To quantify reconstruction quality, we compute the normalized word-level Levenshtein distance between the reconstructed document and the reference target, treating edit similarity as a measure of extraction accuracy. Lower distance values indicate higher structural alignment, reflecting successful separation of template and content. In addition, we compute the Boilerplate Extraction Rate (BER), defined as the ratio of correctly recovered boilerplate tokens to the total tokens in the document. This allowing us to measure how completely each system identifies and isolates the underlying template. The reconstruction error (RE) and the BER using GPT-5-mini[13] and GPT-5[14] based extraction models are reported in Table 3.

Table 3: Reconstruction error after boilerplate extraction and cross-content substitution.

Model	RE ↓	BER (%) ↑
GPT-5-mini	$0.0203_{\pm 0.0445}$	$58.19_{\pm 29.57}$
GPT-5	$0.0097_{\pm 0.0370}$	$75.60_{\pm 24.71}$

[13]gpt-5-mini-2025-08-07
[14]gpt-5-2025-08-07

The results in Table 3 show that both models are able to recover boilerplate templates with very high fidelity, as evidenced by the extremely low RE values. For GPT-5, an RE of 0.0097 at the word level implies that, for a document of 100 words, the reconstructed version differs on average in fewer than a single word-level edit (insertions, deletions or substitutions), i.e., under 1% of the text is affected after template extraction and content swapping. Even GPT-5-mini, with an RE of 0.0203, remains within a small deviation band of roughly 2% of the words. In conjunction with the high BER values observed, these results indicate not only minimal distortion during reconstruction, but also substantial boilerplate capture, meaning that the models recover a large share of template text and do so with very little structural loss.

Looking at boilerplate extraction performance overall, GPT-5 clearly outperforms GPT-5-mini, recovering on average 75.60% of boilerplate content compared to 58.19%, while simultaneously achieving lower RE. This indicates that GPT-5 not only preserves the underlying document structure almost perfectly, but also captures a substantially larger fraction of templated spans, resulting in a more complete and reliable separation between boilerplate and variable content.

4.3 Boilerplate Prevalence Analysis

To understand the structural characteristics of the corpus, we begin by estimating the degree to which boilerplate language appears within *Diavgeia* documents. Since running full-scale extraction across the entire dataset is computationally expensive due to the LLM components of our pipeline, we instead adopt an alternative exploratory strategy that also offers methodological insight into the behaviour and assumptions of the algorithm outlined above. Figure 2 illustrates the distribution of pairwise cosine distances between the document embeddings, providing an overview of the similarity relationships across the corpus. Distances were computed directly in embedding space, revealing how closely textual units are positioned relative to one another. The curve exhibits a sharp concentration of low-distance pairs, suggesting that individual documents tend to have many highly similar neighbours and comparatively fewer distant counterparts. Such behaviour is indicative of recurrent linguistic segments and duplicated structural patterns, where groups of documents share large overlapping spans of text. This clustering effect aligns with the presence of boilerplate material: if multiple documents are consistently situated near one another in semantic space, it is likely due to shared administrative templates, repeated procedural phrasing, or legally mandated formulations.

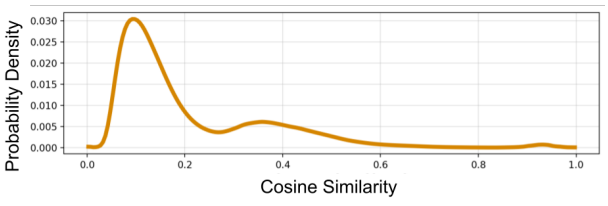


Figure 2: Pairwise cosine-distance distribution across document embeddings.

To further investigate this phenomenon, we applied k-means clustering to group documents based on their embedding similarity. For each cluster, we extracted the centroid document and its top-N nearest neighbors, and provided them to our algorithm (excluding the final boilerplate extraction step)[15] to determine whether the central document appeared to be generated from a boilerplate. The resulting predictions were manually inspected in order to verify true boilerplate presence and to discard false positives. We repeated the process for $k \in \{5, 10, 20, 30, 40, 50, 70, 100\}$, allowing us to assess how boilerplate detection behaves under varying levels of granularity. The proportion of clusters whose centroids were confirmed to contain boilerplate text is presented in Figure 3.

From Figure 3 we observe that the vast majority of k-means centroids systematically contain boilerplate. As k increases, we consistently uncover more boilerplate templates across clusters, with an ever larger share of centroids being classified as boilerplate-based. This indicates that increasing clustering granularity reveals additional templated structures rather than reducing their prevalence.

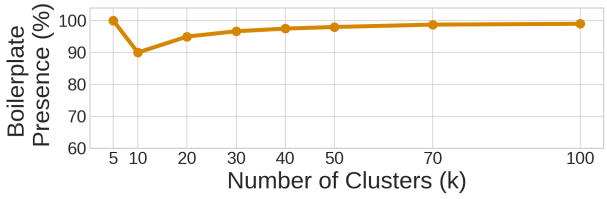


Figure 3: Boilerplate detection rate across different k-means cluster sizes.

5 Quantitative Evaluation: Citizen-Oriented Question Answering over *Diavgeia* via Retrieval Augmented Generation

To demonstrate the practical utility of the corpus for transparency and civic access, we formulate a retrieval-augmented generation (RAG) task in the form of a citizen-facing question answering system over *Diavgeia*. The goal is to answer natural-language questions about Greek government decisions, while grounding responses in the underlying acts and explicitly citing their unique identifiers (ADA codes).

5.1 Task Definition

Given a user query q in Greek (e.g., "Πόσα χρήματα διατέθηκαν για συμβάσεις τον Ιανουάριο 2024;"), the system must: (i) retrieve a small set of relevant decisions from the *Diavgeia* corpus, (ii) construct an evidence context from their text and metadata, and (iii) generate an answer a that is both correct and explicitly supported by the retrieved decisions, including ADA citations. The task assumes no prior legal expertise on the user side and is designed to support information needs such as amounts spent, responsible agencies, dates of publication, and descriptions of specific administrative acts.

[15] We used the gpt-5-mini-2025-08-07 model.

5.2 System Implementation

The deployed system follows a standard RAG architecture [15] adapted to the *Diavgeia* setting. All decisions collected via the official OpenData API are converted into structured text files that concatenate normalized metadata (e.g., ADA, protocol number, issue date, subject, organization, signers, financial fields) with the full PDF-extracted content. These documents are indexed in an Elasticsearch 7.11 instance using the Greek language analyzer, with BM25 as the ranking function.

At query time, we construct a retrieval query by concatenating the current user question with up to five previous turns from the ongoing conversation, enabling basic follow-up and context-aware queries. Elasticsearch performs a full-text BM25 search over the "content" field of the index and returns the top-k documents (we use $k = 8$ in our prototype). For each hit, we keep both the ADA identifier and the corresponding text segment, and we assemble a compact evidence context by concatenating the metadata header and a subset of the document text.

The generation component uses OpenAI GPT-4o-mini as the primary model, with GPT-4o employed for experiments requiring structured outputs. The prompt consists of: (i) a system message instructing the model to act as an assistant for *Diavgeia* documents and to always reference ADA codes, (ii) a short conversation history, (iii) the retrieved evidence, and (iv) the current user question. The model then produces either a streaming answer (for the interactive chat interface) or a parsed response conforming to a Pydantic schema with both a concise answer and a more detailed explanation including citations. All user–assistant exchanges, along with the detailed responses, are stored in Redis as compressed JSON, enabling stateful dialogue across turns.

5.3 Evaluation: Automated Question-Answer Pair Creation

To assess the accuracy and reliability of the RAG system, a comprehensive evaluation framework was implemented using a combination of automated question generation and semantic comparison techniques. The evaluation dataset was generated using GPT-4.1-mini to create domain-specific question-answer pairs from actual government documents. This approach ensures: i) Questions are grounded in real document content, ii) Ground truth answers are accurate and verifiable, iii) Questions cover diverse aspects (decisions, actions, participants, financial data, organizational information). We select 500 documents from the indexed *Diavgeia* corpus. For each document, the LLM generates: i) One specific, detailed question based on document content, ii) Ground truth answer with relevant details and iii) ADA reference number for traceability. Accuracy of 66.6% indicates the RAG system successfully retrieves and generates correct information for approximately two-thirds of queries. We also observe strong performance on financial queries (67% amount match), while moderate semantic similarity suggests responses capture key information even when phrasing differs. Summary statistics are shown in Table 4.

Table 4: Summary Statistics for *Diavgeia* QA Dataset (500 Question–Answer Pairs)

Metric	Value
Total Comparisons	500
Semantically Equivalent (≥ 70%)	333 (66.6%)
Not Equivalent (< 70%)	167 (33.4%)
Average Semantic Score	63.0%
Average TF-IDF Similarity	54.2%
Average Amount Match	67.0%

5.4 Manual Evaluation: Multi-Document Aggregation Queries

To complement the automated evaluation and to specifically assess the system's performance on complex multi-document queries, we conducted a manual evaluation focusing on organizational-level questions that require aggregation across multiple decisions. This setting reflects realistic user needs where information about a public entity is scattered across numerous *Diavgeia* documents.

Methodology. The scenarios simulate real-world use cases in which users request organization-wide insights. Such queries require retrieving multiple related decisions, synthesizing their content, and performing higher-level aggregation. This simultaneously tests retrieval breadth, cross-document consistency, amount aggregation, signer extraction, and topic summarization.

Organization Selection. Five diverse government entities were selected to ensure variation in document types, amounts, and administrative roles:

- Κρατική Σχολή Ορχηστικής Τέχνης (5 decisions),
- Δ.Ε.Υ.Α Θήρας (5 decisions),
- Ερευνητικό Κέντρο Καινοτομίας [...] «Αθηνά» (6 decisions),
- Περιφερειακό Γενικό Νοσοκομείο Αθηνών (Λαϊκό) (3 decisions),
- Εγνατία Οδός (1 decision).

Question Template. Each organization was evaluated using four standardized queries:

(1) "How many and which decisions relate to organization X?"
(2) "What is the total amount spent for these decisions?"
(3) "Who are the signers for decisions related to the organization?"
(4) "What are the subjects/topics of these decisions?"

Ground Truth Extraction. Ground truth was manually derived from the indexed documents. This included:

- counting and listing all ADA codes for each organization,
- computing the total sum of financial amounts,
- compiling a list of unique signers and their positions,
- summarizing the topics of the corresponding decisions.

Response Collection. All queries were submitted through the web application interface. Responses were recorded in real time, and conversation context was intentionally preserved across related questions to reflect typical user interaction flows.

Human Evaluation. Three dimensions were assessed manually: completeness, accuracy, and reasoning quality. Each response was

compared directly against the ground truth, accompanied by a qualitative analysis to identify strengths (e.g., correct multi-document synthesis) and weaknesses (e.g., missed decisions, arithmetic errors, partial signer extraction). The overall statistics are shown in Table 5 and a full example is shown in Table 6.

Table 5: Stats on the Manual Multi-Document Evaluation

Metric	Result	Notes
Total Questions	20	4 questions × 5 organizations
Fully Correct	14	70.0%
Partially Correct	3	15.0%
Incorrect	3	15.0%
Overall Accuracy	85.0%	(Fully + 0.5×Partially)

5.5 Summary

The RAG system demonstrates strong performance across both evaluation paradigms. In the automated single-document evaluation, it achieves 66.6% accuracy on 500 diverse queries, with particularly strong results in financial extraction and reliable document identification through ADA codes. The manual multi-document evaluation further highlights the system's capabilities, reaching 85.0% accuracy on 20 organizational queries, with perfect signer identification (100%) and perfect topic summarization (100%), though arithmetic aggregation remains a weaker point (60%). Overall, the combined assessment shows that the system excels at retrieval and synthesis across heterogeneous administrative documents, while numerical reasoning and cross-document arithmetic remain areas for improvement.

5.6 Scope and Limitations

The current implementation indexes a subset of *Diavgeia* decisions and uses purely lexical BM25 retrieval, which may miss semantically relevant acts when queries are vague or paraphrased. The generator is constrained by a 1,500-token output limit and can occasionally over-generalise when evidence is sparse or noisy. Nonetheless, the system already supports a broad range of real-world queries (financial, organizational, temporal, and document-oriented) and serves as a reference implementation of a *Diavgeia* RAG task that future work can extend with hybrid retrieval, domain-adapted Greek language models, and richer evaluation protocols.

6 Challenges and Future Directions

Although our work lays the groundwork for computational transparency, there are still a number of obstacles to overcome. Robust handling of noisy OCR, varied document formats, and changing administrative schemas is necessary to guarantee the dependability of AI-driven studies. Additionally, models must be built to reduce hallucinations, maintain legal accuracy, and offer traceable reasoning—particularly when applied in high-stakes public sector situations. New techniques for verified reasoning, governance-aware language models, and controlled multi-agent interaction will be necessary to move towards novel AI systems. Future studies should investigate human-in-the-loop verification, counterfactual simulation, and multi-perspective oversight systems, as well as how these techniques might be included into formal auditing processes and cross-national transparency databases. When taken as a whole, these paths show the way towards reliable, scalable AI technologies that significantly improve transparency and accountability in contemporary governance. Next, we present two such systems that show exceptional potential for advancing transparency.

6.1 AI-Driven Public Decision Auditor

A promising and fundamentally new research direction enabled by large-scale open-government corpora such as *Diavgeia* is the development of an **AI Public Decision Auditor**: an automated system that performs continuous, data-driven oversight of government actions in real time. Rather than supporting only retrospective analysis or simple information access, such a system would operationalise proactive transparency by detecting irregularities, inconsistencies, and atypical behaviour at the moment decisions are published.

Recent work on knowledge-graph-based procurement platforms and anomaly detection has shown that open data can be used to identify unusual tenders, suspicious supplier patterns, and potential integrity risks in public spending [22, 23]. Machine learning has been applied to open-government data to detect improprieties and estimate corruption risk in public resource allocation [26]. In parallel, policy and governance communities have started to articulate frameworks for algorithmic transparency and accountability in the public sector [2, 17], and to explore how AI reshapes public auditing and oversight institutions [10].

An AI Public Decision Auditor would build on these foundations while going beyond domain-specific anomaly detection toward holistic, cross-cutting monitoring of the full decision stream. It would leverage statistical modelling, temporal pattern analysis, and large language models to identify signals of administrative risk or unusual institutional behaviour. Core analytical capabilities could include: **i) Anomaly detection in spending and procurement**: automatically flagging unusual financial patterns, inconsistent justifications, or deviations from agency- and category-specific historical norms, extending existing procurement-oriented anomaly detectors to the broader space of administrative acts. **ii) Inconsistency tracing across time**: detecting when newly published decisions contradict past rulings, commitments, or procedural habits of the same public body, or depart from legally relevant precedents. **iii) Procedural integrity checks**: identifying potentially irregular approvals, including missing or anomalous signatures, atypical approval chains, rushed publication timelines, or mismatches between an act's declared type and its legal or substantive content. **iv) Temporal and political clustering**: highlighting clusters of decisions that correlate suspiciously with election cycles, budget deadlines, procurement periods, or other political and administrative events, thereby supporting risk-based oversight of politically sensitive windows.

By transforming continuous streams of public-sector decisions into actionable, explainable signals, an AI Public Decision Auditor would establish a machine-driven accountability layer that is accessible to citizens, journalists, researchers, and watchdog organisations—even without technical expertise. This vision moves

Table 6: Results by Organization – ΚΡΑΤΙΚΗ ΣΧΟΛΗ ΟΡΧΗΣΤΙΚΗΣ ΤΕΧΝΗΣ **(5 decisions)**

Question	Ground Truth	Chatbot Response	Assessment
Count & List	5 decisions: ΡΦ9ΥΧ469ΗΥΖ-6ΩΛ; ΕΖΖΠ469ΗΥΖ-ΧΩ5; 6Ε6Ι469ΗΥΖ-1ΕΘ; ΨΒ41469ΗΥΖ-7ΝΣ; 9ΥΛ9469ΗΥΖ-Ι19	5 decisions: ΕΖΖΠ469ΗΥΖ-ΧΩ5, 9ΥΛ9469ΗΥΖ-Ι19, 6Ε6Ι469ΗΥΖ-1ΕΘ, ΨΒ41469ΗΥΖ-7ΝΣ, ΡΦ9Υ469ΗΥΖ-6ΩΛ	Perfect match – All 5 ADAs identified
Total Amount	73.225,56 €	52.736,56 + 10.416,00 + 60,00 + 8.680,00 + 1.333,00 = 73.225,56 €	Excellent – Correct sum with breakdown
Signers	Κόκκινος Χαράλαμπος, Διευθυντής	Χαράλαμπος Κόκκινος, Διευθυντής	Correct – Name order variation acceptable
Topics	(Various decision topics)	Rental, library services, software maintenance, legal services, past dues	Accurate summary
Performance: 4/4 correct – Exemplary multi-document handling			

beyond search, retrieval, or question answering, positioning open-government data as the substrate for *real-time, anticipatory oversight*. As such, it represents a radical and societally impactful next step for transparency and AI-for-governance research.

6.2 LLM-Based Multi-Agent Collective Oversight Simulator

Beyond single-agent auditors, a radical next step is a LLM-based multi-agent Collective Oversight Simulator for public decisions. In this paradigm, newly published administrative acts (e.g., *Diavgeia* decisions) are injected into a simulated ecosystem of specialized LLM agents. Each agent embodies a distinct institutional role or normative objective, and the agents *debate*, *negotiate*, and *stress-test* the decision before it is surfaced to end-users. Rather than passively flagging anomalies, the system performs proactive, scenario-based oversight, exposing weaknesses, and inconsistencies that may only emerge through interaction between multiple perspectives.

This vision builds on recent work showing that LLM-based multi-agent systems can autonomously cooperate, role-play, and coordinate on complex tasks [11, 29]. Role-playing frameworks such as CAMEL [16] demonstrate how distinct agent profiles and inception prompts can elicit complementary behaviours from the same underlying model, while general multi-agent conversation frameworks like AutoGen enable programmable interaction patterns between specialised agents and tools [29]. Multi-agent debate architectures further suggest that structured argumentation between LLMs can improve reasoning quality and truth-seeking [12], and recent surveys highlight the particular relevance of LLM-based multi-agent systems to domains requiring diverse expertise, such as legal and public-sector decision-making [7, 13].

Within this landscape, the Collective Oversight Simulator instantiates a set of persistent, LLM-based agents, for example: **Legal Coherence Agent**: checks compatibility with existing legislation, prior decisions, and cross-references, proposing counterfactuals where minor changes to the act would trigger conflicts. **i) Budget Compliance Agent**: simulates fiscal impact over time, tests alternative budgetary scenarios, and highlights combinations of decisions that risk overruns or hidden commitments. **ii) Procedural Integrity Agent**: systematically perturbs timelines, signatory chains, and act types to identify minimal changes that render the act procedurally irregular, thereby mapping its "fragility" with respect to administrative rules. **iii) Citizen Fairness Agent**: reasons about distributional effects across population groups, using fairness-inspired heuristics and counterfactual cases to surface potential inequities or disparate impacts.

Agents engage in a structured dialogue procedure in which they challenge one another's presumptions, suggest counterfactual changes to the choice, and either converge or fail to converge on a set of risk statements and justifications. The result is a multi-perspective report that identifies areas of agreement and disagreement among institutional responsibilities rather than a single score. In terms of methodology, this creates a new field of work at the nexus of computational public administration and LLM-based multi-agent systems: creating role specifications, debate protocols, and evaluation frameworks that make simulated oversight useful for practitioners and auditable by citizens, journalists, and regulators.

7 Conclusion

This work presented the first large-scale, processed corpus of Greek government decisions and demonstrated its value for computational transparency research. Our contributions include: (i) **a one-million–document open dataset** with normalized metadata, canonical identifiers, and high-quality Markdown text extracted from PDFs and XML; (ii) **a fully reproducible extraction pipeline** for harvesting, cleaning, and structuring *Diavgeia* documents at Web scale; (iii) **qualitative analyses** via boilerplate detection; and (iv) **a retrieval-augmented generation (RAG) benchmark** with representative questions, curated answers, and baseline performance, simulating **a citizen-facing tool** enabling natural-language querying of government decisions, demonstrating the dataset's practical value for transparency and civic empowerment.

Beyond these contributions, our findings highlight the transformative potential of AI as an accountability mechanism within digital governance ecosystems. As language models grow more capable of reasoning over long documents, cross-referencing administrative histories, and detecting irregularities, they can function

as computational aides that continuously surface inconsistencies, atypical patterns, or procedural anomalies across vast public decision streams. The scale and structure of the dataset released here provide a foundation for developing such systems in a rigorous and data-driven manner, especially in the case of low- to mid-resource languages like Greek, where limited NLP data are available [18, 19].

Looking forward, future models may incorporate role-specific reasoning, temporal analytics, counterfactual simulation, and debate-style multi-agent interaction to evaluate decisions from legal, financial, procedural, and fairness perspectives. These advances could give rise to proactive oversight tools—LLM-based auditors and multi-agent governance simulators—that support citizens, journalists, regulators, and civil-society organizations in holding institutions accountable.

Acknowledgments

This work was supported by the project ΥΠ2ΤΑ-0554078, implemented in the framework of the call "Visiting Professors/Visiting Researchers" (Επισκέπτες Καθηγητές/Επισκέπτες Ερευνητές, OPS TA code 5185218) of the Hellenic Ministry of Education, Religious Affairs and Sports, and funded by the European Union – NextGenerationEU under the Recovery and Resilience Facility of the National Recovery and Resilience Plan "Greece 2.0".

References

[1] Giuseppe Abrami, Mevlüt Bagci, and Alexander Mehler. 2024. German parliamentary corpus (GerParCor) reloaded. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 7707–7716.

[2] AI Now Institute. 2021. *Government by Algorithm: Public-Sector Algorithmic Accountability*. Technical Report. AI Now Institute. https://ainowinstitute.org

[3] Themis Beris and Manolis Koubarakis. 2018. Modeling and preserving Greek government decisions using semantic web technologies and permissionless blockchains. In *European Semantic Web Conference*. Springer, 81–96.

[4] John C Bertot, Paul T Jaeger, and Justin M Grimes. 2010. Using ICTs to create a culture of transparency: E-government and social media as openness and anti-corruption tools for societies. *Government information quarterly* 27, 3 (2010), 264–271.

[5] Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2019. Large-scale multi-label text classification on EU legislation. In *Proceedings of the 57th annual meeting of the association for computational linguistics*. 6314–6322.

[6] Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. LexGLUE: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 4310–4330.

[7] Xi Chen, Mao Mao, Shuo Li, and Haotian Shangguan. 2025. Debate-feedback: A multi-agent framework for efficient legal judgment prediction. *arXiv preprint arXiv:2504.05358* (2025).

[8] Xiaoyin Chen and Sam Wiseman. 2023. Bm25 query augmentation learned end-to-end. *arXiv preprint arXiv:2305.14087* (2023).

[9] European Data Portal. 2020. Open Data Maturity Report. European Commission.

[10] Dolores Genaro-Moya, Antonio Manuel López-Hernández, and Mariia Godz. 2025. Artificial Intelligence and Public Sector Auditing: Challenges and Opportunities for Supreme Audit Institutions. *World* 6, 2 (2025). doi:10.3390/world6020078

[11] Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680* (2024).

[12] Zhe Hu, Hou Pong Chan, Jing Li, and Yu Yin. 2025. Debate-to-write: A persona-driven multi-agent framework for diverse argument generation. In *Proceedings of the 31st International Conference on Computational Linguistics*. 4689–4703.

[13] Cong Jiang and Xiaolei Yang. 2024. Agents on the Bench: Large Language Model Based Multi Agent Framework for Trustworthy Digital Justice. *arXiv preprint arXiv:2412.18697* (2024).

[14] Jurek Leonhardt, Avishek Anand, and Megha Khosla. 2020. Boilerplate removal using a neural sequence labeling model. In *Companion Proceedings of the Web Conference 2020*. 226–229.

[15] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems* 33 (2020), 9459–9474.

[16] Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems* 36 (2023), 51991–52008.

[17] OECD. 2024. *Algorithmic Transparency in the Public Sector: Emerging Practices and Policy Guidelines*. Technical Report. Organisation for Economic Co-operation and Development. https://oecd.org

[18] Stamatis Outsios, Christos Karatsalos, Konstantinos Skianis, and Michalis Vazirgiannis. 2020. Evaluation of Greek Word Embeddings. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 2543–2551. https://aclanthology.org/2020.lrec-1.310/

[19] Stamatis Outsios, Konstantinos Skianis, Polykarpos Meladianos, Christos Xypolopoulos, and Michalis Vazirgiannis. 2018. Word Embeddings from Large-Scale Greek Web Content. arXiv:1810.06694 [cs.CL] https://arxiv.org/abs/1810.06694

[20] Konstantinos Serderidis, Ioannis Konstantinidis, Georgios Meditskos, Vassilios Peristeras, and Nick Bassiliades. 2024. d2kg: An integrated ontology for knowledge graph-based representation of government decisions and acts: The Greek Programme Diavgeia case. *Semantic Web* 15, 5 (2024), 1677–1699.

[21] Mohamed Ahmed Sherif, Sandro Athaíde Coelho, Ricardo Usbeck, Sebastian Hellmann, Jens Lehmann, Martin Brümmer, and Andreas Both. 2014. NIF4OGGD-NLP Interchange Format for Open German Governmental Data.. In *LREC*. 3524–3528.

[22] Ahmet Soylu, Oscar Corcho, Brian Elvesæter, Carlos Badenes-Olmedo, Tom Blount, Francisco Yedro Martínez, Matej Kovacic, Matej Posinkovic, Ian Makgill, Chris Taggart, et al. 2022. TheyBuyForYou platform and knowledge graph: Expanding horizons in public procurement with open linked data. *Semantic Web* 13, 2 (2022), 265–291.

[23] Ahmet Soylu, Oscar Corcho, Brian Elvesæter, Carlos Badenes-Olmedo, Francisco Yedro Martínez, Matej Kovacic, Matej Posinkovic, Ian Makgill, Chris Taggart, Elena Simperl, et al. 2020. Enhancing public procurement in the European Union through constructing and exploiting an integrated knowledge graph. In *International Semantic Web Conference*. Springer, 430–446.

[24] Nikolaos Stylianou, Danai Vlachava, Ioannis Konstantinidis, Nick Bassiliades, and Vassilios Peristeras. 2022. Doc2KG: Transforming document repositories to knowledge graphs. *International Journal on Semantic Web and Information Systems (IJSWIS)* 18, 1 (2022), 1–20.

[25] Michalis N Vafopoulos, Marios Meimaris, Agis Papantoniou, Ioannis Anagnostopoulos, Giorgos Alexiou, Ioannis Avraam, Ioannis Xidias, Giorgos Vafeiadis, and Vasilis Loumos. 2012. Public spending: Interconnecting and visualizing greek public expenditure following linked open data directives. *Available at SSRN 2064517* (2012).

[26] Ramon Vaqueiro, Ana Vargas, Tatiana Escovedo, and Marcos Kalinowski. 2023. Machine Learning Applied to Open Government Data for the Detection of Improprieties in the Application of Public Resources. In *Proceedings of the XIX Brazilian Symposium on Information Systems*. 213–220.

[27] Gerda Viira, Maarten Marx, and Maik Larooij. 2024. ParlaMint Widened: a European Dataset of Freedom of Information Act Documents (Position Paper). In *Proceedings of the IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora (ParlaCLARIN)@ LREC-COLING 2024*. 171–172.

[28] Thijs Vogels, Octavian-Eugen Ganea, and Carsten Eickhoff. 2018. Web2text: Deep structured boilerplate removal. In *European Conference on Information Retrieval*. Springer, 167–179.

[29] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. 2024. Autogen: Enabling next-gen LLM applications via multi-agent conversations. In *First Conference on Language Modeling*.

[30] Anneke Zuiderwijk, Marijn Janssen, and Chris Davis. 2014. Innovation with open data: Essential elements of open data ecosystems. *Information polity* 19, 1-2 (2014), 17–33.

A Appendix

Next, we present some additional description of fields, analysis, and results.

Table 7 summarizes the core metadata fields available in *Diavgeia* decision records. For each field, we include the original Greek description as provided by the platform, along with an English translation to support international accessibility and reproducibility. These fields capture identifiers, organizational information, timestamps,

status indicators, and additional type-specific metadata, forming the basis for structured analysis of government decisions.

Table 8 presents a comparative overview of the execution time of different PDF text-extraction and OCR methods. It highlights the substantial performance differences between native PDF extraction tools, lightweight OCR engines, and more computationally intensive deep-learning-based OCR pipelines. By measuring the total processing time under identical conditions, the table helps identify which methods are most suitable for large-scale processing of Diavgeia documents.

Tables 9, 10, 11 show results by organization on manual evaluation regarding multi-document aggregation queries. Our evaluation highlights clear differences in performance across tasks. Signer identification achieves perfect accuracy, consistently extracting names and positions across organizations and handling name variations without errors. Topic summarization also performs exceptionally well, producing coherent and semantically accurate summaries, sometimes even offering more detail than the reference summaries. Document enumeration performs well overall, correctly identifying most relevant decisions, though it occasionally retrieves extra or incorrect documents—an issue that appears mainly when organization names are not unique. Financial aggregation is the weakest area, succeeding only when monetary amounts are explicitly listed. It struggles with implicit calculations or incomplete data, and a large deviation in the ΛΑΙΚΟ case (28% error) suggests a systematic limitation in numerically grounded reasoning across financial documents.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009

Table 7: Description of Decision Fields

Field	Description (GR)	English Translation
ada	Αριθμός Διαδικτυακής Ανάρτησης	Web Posting Number
protocolNumber	Αριθμός Πρωτοκόλλου	Protocol Number
subject	Θέμα πράξης	Subject of the decision
Issue Date	Ημερομηνία έκδοσης σε μορφή Unix Timestamp	Issue date (Unix timestamp, ms)
decisionTypeId	Κωδικός του τύπου πράξης	Decision type ID
organizationId	Κωδικός φορέα	Organization ID
unitIds	Λίστα κωδικών των μονάδων που εμπλέκονται	IDs of units involved
signerIds	Λίστα κωδικών των τελικών υπογραφόντων	IDs of official signatories
extraFieldValues	Ειδικά πεδία ανάλογα με τον τύπο πράξης	Type-specific metadata fields
submissionTimestamp	Ημερομηνία/ώρα τελευταίας τροποποίησης σε Unix Timestamp	Last modification timestamp (ms)
status	Κατάσταση πράξης (PUBLISHED, PENDING REVOCATION, REVOKED, SUBMITTED etc.)	Status of the decision
versionId	Αριθμός έκδοσης	Version number

Table 8: Speed Comparison of PDF Extraction and OCR Methods

Method	Total Time (s)	Avg. Time per Page (s)	Relative Speed
Extraction (PyMuPDF)	0.07	~0.00017	~2500× Faster
Paddle OCR (Zoom 2)	423.69	0.97	1×
Tesseract OCR (Zoom 2)	429.43	0.99	0.97×
Tesseract OCR (Zoom 3)	523.76	1.20	0.81×
Paddle OCR (Zoom 3)	896.10	2.06	0.47×
GPT-5-nano (Ground Truth)	~25,200 (7 hrs)	~126	0.007×

Table 9: Results by Organization – Δ.Ε.Υ.Α ΘΗΡΑΣ (5 decisions)

Question	Ground Truth	Chatbot Response	Assessment
Count & List	5 decisions: ΨΣ02ΟΕΨΠ-ΛΔΤ; 6Μ6ΨΟΕΨΠ-ΛΗ2; 98Π9ΟΕΨΠ-ΞΨΕ; 66ΤΙΟΕΨΠ-ΕΟΤ; 6Λ98ΟΕΨΠ-ΩΣΙ	6 decisions: ΨΣ02ΟΕΨΠ-ΛΔΤ, 6Μ6ΨΟΕΨΠ-ΛΗ2, 98Π9ΟΕΨΠ-ΞΨΕ, 66ΤΙΟΕΨΠ-ΕΟΤ, 6Λ98ΟΕΨΠ-ΩΣΙ + 1 extra	Minor error – Retrieved 5 correct + 1 incorrect
Total Amount	381,22 €	381,22 €	Perfect
Signers	ΖΩΡΖΟΣ ΕΥΑΓΓΕΛΟΣ, Διευθυντής	Ευάγγελος Ζώρζος, Προϊστάμενος Τμήματος και Διευθυντής	Correct with additional detail
Topics	ΚΡΑΤΗΣΕΙΣ ΤΡΑΠΕΖΩΝ	Κρατήσεις τραπεζών	Correct
Performance: 3.5/4 – One extra document retrieved but calculations correct			

Table 10: Results by Organization – ΕΡΕΥΝΗΤΙΚΟ ΚΕΝΤΡΟ ΚΑΙΝΟΤΟΜΙΑΣ **[...]** «ΑΘΗΝΑ» **(6 decisions)**

Question	Ground Truth	Chatbot Response	Assessment
Count & List	6 decisions: 6ΥΒΘ469ΗΞΩ-ΣΚΟ; 6ΨΨΞ469ΗΞΩ-ΦΧΗ; 6ΑΝ7469ΗΞΩ-7Ε3; 9Ο5Β469ΗΞΩ-ΦΜΙ; 9ΗΓΕ469ΗΞΩ-0Τ3; ΨΨΝΛ469ΗΞΩ-5ΩΞ	6 decisions: 6ΨΨΞ469ΗΞΩ-ΦΧΗ, 9Ο5Β469ΗΞΩ-ΦΜΙ, 6ΑΝ7469ΗΞΩ-7Ε3, ΨΨΝΛ469ΗΞΩ-5ΩΞ, 6ΥΒΘ469ΗΞΩ-ΣΚΟ, 6ΥΜΝ469ΗΡΙ-Ρ66	Partial – 5/6 correct, 1 wrong substitution
Total Amount	11.170,33 €	11.170,33 €	Perfect despite wrong ADA
Signers	Κατσαμάνης Νάσος, ΙΩΑΝΝΗΣ ΕΜΙΡΗΣ, ΓΑΡΟΦΑΛΑΚΗΣ ΜΙΝΩΣ, Στύλιος Χρυσόστομος	Νάσος Κατσαμάνης (Αναπληρωτής Διευθυντής), Μίνως Γαροφαλάκης (Διευθυντής ΙΠΣΥ), Ιωάννης Εμίρης (Γενικός Διευθυντής)	~ Incomplete – Missing 1 signer
Topics	(List of 6 topics)	Software subscriptions, elevator maintenance, H/M installations, cybersecurity subscription, courier services, call for proposals	Good summary
Performance: 2.5/4 – Retrieval issue but good synthesis			

Table 11: Results by Organization – ΠΕΡ.ΓΕΝ. ΝΟΣΟΚΟΜΕΙΟ ΑΘΗΝΩΝ **(**ΛΑΪΚΟ**) (3 decisions)**

Question	Ground Truth	Chatbot Response	Assessment
Count & List	3 decisions: 9ΓΨΒ4690Ω8-ΥΤΧ; 6ΝΞΕ4690Ω8-ΑΗΖ; 935Δ4690Ω8-ΑΤΓ	3 decisions: 6ΝΞΕ4690Ω8-ΑΗΖ, 9ΓΨΒ4690Ω8-ΥΤΧ, 935Δ4690Ω8-ΑΤΓ	Perfect
Total Amount	338.580,80 €	243.380,00 €	Incorrect – Major calculation error (72% accuracy)
Signers	ΡΟΪΔΗΣ ΘΕΟΦΑΝΗΣ, Διοικητής	ΘΕΟΦΑΝΗΣ ΡΟΪΔΗΣ, Διοικητής	Correct
Topics	ΑΠΟΦΑΣΗ ΑΝΑΛΗΨΗΣ ΠΙΣΤΩΣΗΣ	Payroll commitments, allowances, ΕΦΚΑ contributions, training, facilities maintenance	More detailed than GT
Performance: 3/4 – Document retrieval perfect but aggregation failed			