T.R.I.M.: A Citizen-Led AI Platform for Municipal Budget Transparency

Presidential AI Challenge, Track II: Technical Implementation

Kundan Baliga[1], Aditya Byju[1], Aiden Xie[2], Matthew Kim[3]
[1]Neuqua Valley High School, Naperville, IL, United States
[2]Naperville North High School, Naperville, IL, United States
[3]Los Altos High School, Los Altos, CA, United States
Website: https://trytrim.org — Video: https://youtu.be/ozpOLFYALu4

***Abstract*—The opacity of municipal finance represents a critical barrier to civic engagement and government efficiency. While public spending data is technically "open" following federal legislation like the DATA Act, these standards rarely trickle down to local governance, leaving municipal data functionally inaccessible due to complexity, length, and inconsistent formatting. This paper presents T.R.I.M. (Taxpayer Resources Improvement Model), a comprehensive AI-powered auditing platform designed to bridge the gap between bureaucratic documentation and citizen oversight. T.R.I.M. introduces a Dual-Model Architecture, employing Gemini 3 Pro for high-reasoning semantic extraction and Gemini 3 Flash as an optional adversarial "LLM-as-a-Judge" validator. This pipeline ingests municipal PDF datasets, enforces strict citation requirements, and calculates a deterministic "Fiscal Efficiency Score." By automating the detection of fiscal anomalies and low-transparency spending, T.R.I.M. empowers citizens to perform forensic-level audits in seconds rather than weeks. This work aligns directly with national priorities regarding government efficiency and AI adoption, demonstrating how American artificial intelligence can be deployed to restore public trust and streamline the stewardship of taxpayer resources.**

***Index Terms*—Artificial Intelligence, Municipal Finance, Civic Auditing, Gemini 3 Pro, LLM-as-a-Judge, Government Efficiency, Natural Language Processing, Public Trust.**

I. Introduction: The Transparency Paradox

A. A Crisis of Trust in Mountain View

In late 2024 and early 2025, the Mountain View Whisman School District (MVWSD) in California became the epicenter of a profound civic crisis. Parents and taxpayers, already grappling with the high cost of living, were stunned by reports detailing the district's spending priorities. Hidden within the dense financial documentation were significant allocations for "executive leadership coaching" and "mindfulness" contracts for high-level administrators, expenses that seemed disconnected from the classroom needs of students.

The backlash was swift. In the November 2024 election, voters delivered a mandate for change, electing three new trustees who campaigned on a platform of fiscal oversight. However, the unrest did not end at the ballot box. By the spring of 2025, a recall effort was launched against a remaining board member, fueled by a community petition citing a "breach of trust" regarding financial transparency [1].

The central tragedy of this episode was not just the spending itself, but the obscurity that protected it. The relevant line items were buried deep in hundreds of pages of budget packets, obscured by fund accounting codes and jargon. It took months of dedicated forensic work by motivated citizens to uncover what should have been obvious in seconds.

B. The National Context

The events in Mountain View are not an anomaly; they are a symptom of a systemic failure playing out in communities across the United States. The gap between public resources and public understanding has never been wider.

With state and local government expenditures exceeding $3.5 trillion annually [2], the sheer volume of financial data makes manual oversight nearly

impossible. Too often, municipal and district budgets are published as PDFs exceeding 500 pages, filled with technical terms like “OPEB liabilities,” “encumbrances,” and “inter-fund transfers” that are inaccessible to everyday residents. Even motivated citizens with time and attention find themselves overwhelmed by the density of the data. As a result, inefficiency can go unnoticed for years—not because people don’t care, but because they lack the tools to see what is really going on.

When citizens do not understand how their tax dollars are spent, local governments lose their legitimacy. According to the Pew Research Center, public trust in government has hovered near historic lows [3]. This “Transparency Paradox”—where data is technically public but cognitively inaccessible—is the problem we set out to solve.

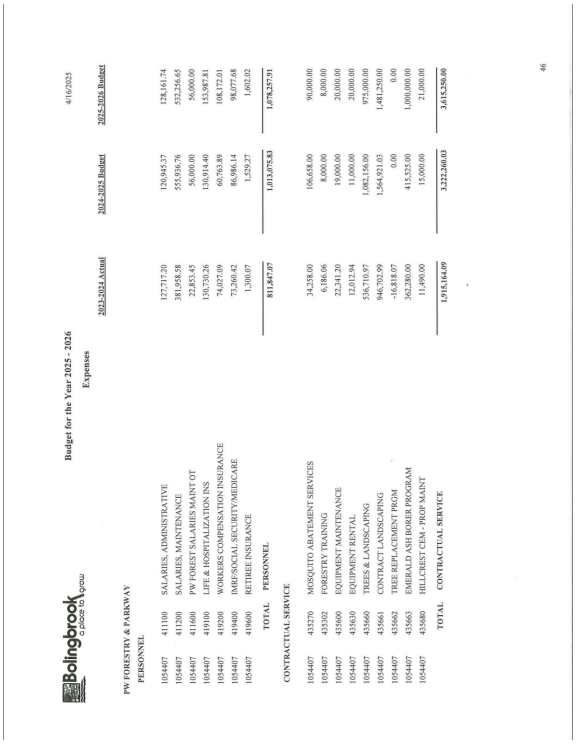
C. The T.R.I.M. Solution

T.R.I.M. (Taxpayer Resources Improvement Model) is an AI-powered auditing platform that converts municipal budget PDFs into *verifiable*, structured audit outputs for non-expert citizens. Rather than relying on city-provided dashboards or clean datasets, T.R.I.M. operates directly on the most common real-world artifact of local government finance: long, heterogeneous, and often poorly formatted budget and financial-report PDFs. As illustrated in Figure 1, the system reduces the practical barrier to oversight by turning hundreds of pages of accounting tables and narrative appendices into an interactive, evidence-linked audit.

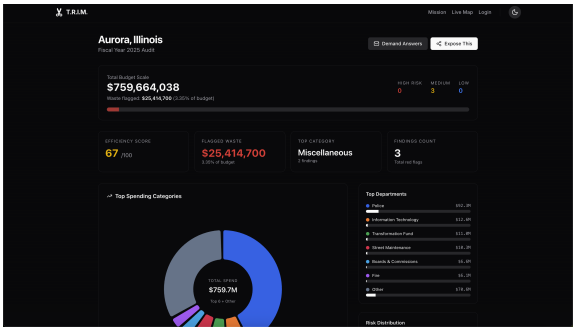
T.R.I.M. is designed around three core requirements:

(1) Evidence-backed extraction (not summarization). The platform does not simply “summarize” a budget. It extracts discrete spending items, program allocations, and fiscal claims into a strict schema (amounts, categories, departments, and context) *with mandatory page-level citations*. Each surfaced insight is therefore traceable to a specific location in the source document, allowing users to confirm claims immediately.

(2) Reliability under adversarial conditions. Municipal finance documents are uniquely hostile to automation: tables span pages, headers are inconsistent, numbers are reported in thousands, and category labels are ambiguous (e.g., “contractual



(a) Traditional budget spreadsheet



(b) T.R.I.M. interface

Fig. 1: Comparison of traditional budget reporting versus the T.R.I.M. user interface, demonstrating the reduction in cognitive load for citizens.

services” vs “third-party supplies”). To mitigate hallucination and misreads, T.R.I.M. uses a dual-model validation loop (described in Section IV-B) that checks whether extracted values and interpretations are consistent with the cited text before any results are persisted or scored.

(3) Deterministic scoring and explainability. Large language models are used only for *extraction and classification*. The final assessment is computed by a deterministic Fiscal Efficiency Score (Section V) that depends solely on validated, cited inputs. This design prevents “model mood” from

influencing results and ensures that identical source documents always produce identical scores.

Operationally, the user experience is intentionally simple: a citizen selects a municipality, T.R.I.M. locates official budget PDFs, runs the extraction/verification pipeline, and returns an audit dashboard with department-level risk indicators and flagged line items. Crucially, every flagged item is accompanied by a deep link that opens the original PDF to the supporting page, ensuring that the platform functions as a decision-support tool rather than an opaque authority. In effect, T.R.I.M. compresses the workflow of manual civic auditing—from searching, reading, and cross-referencing hundreds of pages—into a fast, reproducible, and source-grounded process.

II. Relevance to the Administration: America's AI Action Plan

T.R.I.M. is not simply a technical utility. More importantly, it serves as a manifestation of the core values prioritized by the President and the Administration regarding the modernization of government through technology.

A. Accelerating AI Adoption in Government

The Administration has consistently emphasized the need to utilize technology to streamline operations and enhance productivity. As outlined in *America's AI Action Plan*, the federal government is committed to accelerating the adoption of AI to solve complex national challenges [4]. T.R.I.M. serves as a digital force multiplier for this vision. By providing a scalable tool for budget transparency, it demonstrates how "productivity-enhancing AI uses" can be deployed to improve oversight capacity for both citizens and public officials without increasing bureaucratic overhead.

B. Enhancing Efficiency and Procurement Oversight

A central pillar of the Administration's reform agenda is the optimization of taxpayer resources (e.g., the Department of Government Efficiency initiative). T.R.I.M. focuses specifically on identifying anomalies and low-transparency spending—items that may be legal but lack sufficient detail or performance metrics. By automating the extraction of these items from unstructured documentation, the platform provides the raw data necessary for better procurement oversight and more efficient resource allocation across municipal departments.

C. Restoring Public Trust through Transparency

The "Transparency Paradox" identified in Section I undermines the legitimacy of local institutions. T.R.I.M. addresses this by creating a nonpartisan, evidence-based platform for civic engagement. It does not replace human judgment; rather, it informs it by flagging specific line items for review. This approach aligns with the national priority of developing AI that promotes public trust and ensures that the government remains responsive to the citizens it serves.

III. Background and Related Work

A. The Evolution of Open Government Data

The push for transparency is not new. The Digital Accountability and Transparency Act (DATA Act) of 2014 established federal standards for spending data [5]. However, these standards largely apply to *federal* agencies and awards. Municipal data remains standardized only loosely by the Governmental Accounting Standards Board (GASB). This creates a critical gap: while the federal government moves toward standardization, local school boards and city councils still publish budgets in idiosyncratic, non-machine-readable formats.

B. Limitations of Existing Tools

Existing platforms like *OpenGov* or *Tyler Technologies* provide dashboards for cities to display their own data. However, these tools suffer from two critical flaws:

1) **Voluntary Participation:** They rely on the city to pay for and input clean data. Cities with the most fiscal opacity are the least likely to purchase transparency software.
2) **Lack of Adversarial Review:** These tools are designed to showcase success, not find failure.

T.R.I.M. differs fundamentally because it works on raw, dirty data (PDFs) and assumes an adversarial stance, looking for anomalies rather than accepting the city's self-reported summaries.

TABLE I: Comparative Performance of LLMs on Long-Context and Agentic Tasks [7]

Benchmark Suite	**Description**	**Gemini 3 Pro**	**Claude Sonnet 4.5**	**GPT-5.1**
τ^2-bench	Agentic tool use	**85.4%**	84.7%	80.2%
FACTS Benchmark	Internal grounding	**70.5%**	50.4%	50.8%
Global PIQA	Commonsense reasoning	**93.4%**	90.1%	90.9%
MRCR v2 (8-needle) - Long context performance				
128k (average)		**77.0%**	47.1%	61.6%
1M (pointwise)		**26.3%**	n/a	n/a

IV. System Architecture

T.R.I.M. is a production-ready web application built on the Next.js 16 App Router framework [6]. The system architecture transforms unstructured PDF binary streams into structured, queryable databases of financial risk through a sophisticated multi-agent pipeline.

A. AI Model Selection and Justification

The selection of the underlying AI model was critical. Municipal budgets are long-context documents; a single budget file can easily exceed 200 pages. Traditional RAG (Retrieval Augmented Generation) approaches, which chunk documents into small pieces, fail in this domain because they lose the global context required to understand relationships between a summary table on page 4 and a line item on page 150.

We selected Gemini 3 Pro as our primary generation model. Its performance in long-context retrieval and agentic reasoning—as detailed in the official model card—makes it a viable candidate for ingesting municipal budgets [7]. Specifically, Gemini 3 Pro's ability to maintain high retrieval accuracy across large token windows allows the system to process multi-hundred-page documents without the loss of global context common in traditional RAG architectures.

B. The Dual-Model "Judge" Architecture

To combat the risk of hallucination—where an AI might invent spending items—we implemented an **LLM-as-a-Judge** pipeline [8]. When enabled, this architecture splits the workload between two distinct model instances:

1) Agent 1: The Auditor (Gemini 3 Pro): The first agent receives the PDF via the `@google/genai` file upload APIs. Its system prompt is configured to act as a forensic accountant. It extracts spending data into a strict JSON schema defined by Zod, utilizing `response_mime_type: application/json` to enforce type safety. It is optimized for high-reasoning capabilities to interpret complex accounting tables and infer meaning from column headers that may span multiple pages.

2) Agent 2: The Judge (Gemini 3 Flash): The second agent acts as an adversarial validator. We utilize Gemini 3 Flash for its speed and low cost. Rather than receiving a pre-extracted citation text map, the Judge is informed that structured evidence is unavailable and is instead tasked with verifying the Auditor's claims directly against the attached PDF source.

This verification loop is critical. As illustrated in Figure 2, the Judge acts as a firewall between the raw AI generation and the user database. The Judge performs the following logic:

1) **Verification:** It checks if `citation.text` provided by the Auditor actually exists on the `citation.pageIndex` of the source document.
2) **Fact-Checking:** It evaluates if the extracted `amount` and `category` align with the cited text. For example, if the Auditor flags $1M for "Parties" but the text says "Third Party Vendors," the Judge catches this nuance.
3) **Correction Strategy:** If the Judge detects a discrepancy, it does not merely flag the error; it **edits the audit payload** to remove the hallucinated item or correct the value before the data is ever committed to the database.

C. Application Walkthrough

To demonstrate the system's usability and transparent workflow, Figure 3 details the user journey from document ingestion to granular auditing.

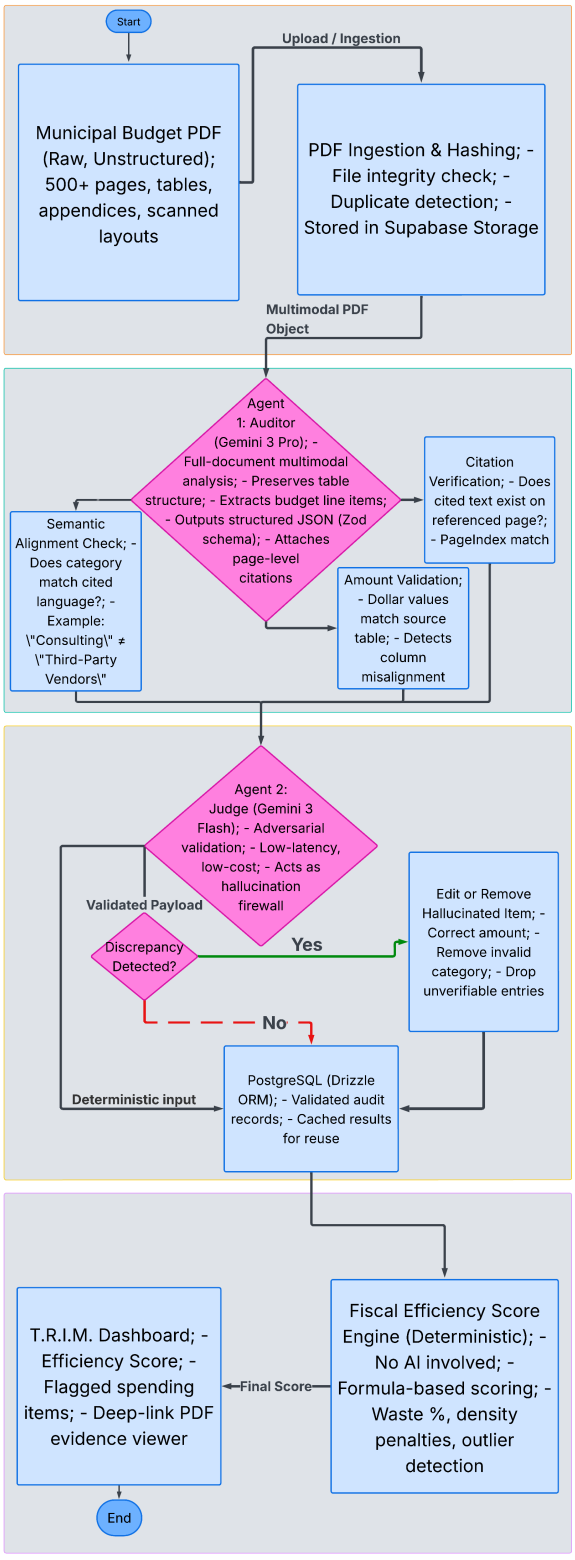


Fig. 2: The T.R.I.M. Validation Pipeline utilizing a high-reasoning Gemini 3 Pro auditor and a fast, adversarial Gemini 3 Flash judge.

D. Autonomous PDF Discovery and Ingestion

A primary challenge in municipal auditing is locating the correct source document amidst complex web architectures. T.R.I.M. implements a sophisticated two-stage autonomous discovery pipeline orchestrated by Server Actions in `src/app/actions/audit.ts`, designed to retrieve documents from official public sources using rate limits and caching to ensure respectful and efficient access.

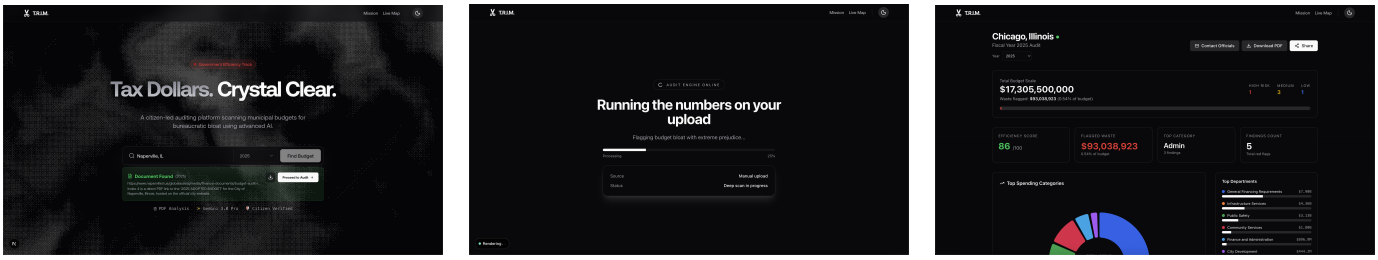
1) Stage 1: Search Queries and Model Selection: The system first generates targeted search queries (e.g., "ACFR [City, State] [Year]" and "[City, State] budget report [Year]"). It fetches the top 10 results via the Google Custom Search API. These results are passed to a Gemini-powered selection agent (`pickBudgetReport`), which evaluates the metadata and URLs to identify the most likely budget PDF. If a direct link (containing `.pdf`) is identified, it is returned immediately.

2) Stage 2: Agentic "Trail" Following: If Stage 1 fails to yield a direct PDF, the system initiates an agentic exploration using `followBudgetTrail` to navigate municipal websites. T.R.I.M. respects `robots.txt` and site-specific terms of service where applicable; if automated retrieval fails due to technical obfuscation or security measures, the system informs the user and provides an interface for manual PDF upload.

E. Ingestion, Persistence, and Data Privacy

Once a PDF is located or uploaded, the backend flow manages the pipeline with a focus on data integrity and privacy:

1) **Processing:** The PDF is downloaded or received and hashed to prevent duplicate processing. We use real-time polling to provide users with live feedback on the status of their audit.
2) **Privacy Posture:** T.R.I.M. primarily ingests official, public-domain municipal documents. For user-uploaded files, the system stores only the PDF (via Supabase Storage) and the extracted audit results (via PostgreSQL). Documents are processed via the Google Gemini API; we rely on provider-level data privacy standards and do not use uploaded data for model retraining.



(a) **Ingestion:** The user searches. (b) **Analysis:** Real-time feedback. (c) **Audit:** The final dashboard.

Fig. 3: T.R.I.M. App Walkthrough exposing each stage of the process, from document ingestion to evidence-backed conclusions.

3) **Persistence:** Extracted JSON data and hashes are cached. If a user requests a city that has already been audited, the system retrieves the validated data instantly, reducing redundant API calls and processing time.

F. Responsible Use and Escalation

T.R.I.M. is designed as a transparency tool, not an accusatory authority. To ensure responsible use:

- **Hypothesis Generation:** All flagged items are presented as hypotheses for review. Users are explicitly instructed to verify the provided citations and context before drawing conclusions.
- **Mitigation of Harm:** The system avoids targeting individual employees or residents. It focuses on aggregate departmental spending and program-level allocations.
- **Escalation Pathways:** The platform encourages users to use the generated insights as starting points for formal inquiries, such as FOIA requests, attendance at public budget meetings, or communication with municipal auditor offices.

V. Methodology: The Efficiency Score

Although we use AI to *extract* data, we do not use AI to *score* the city. Large Language Models can be subjective; to ensure fairness and repeatability, T.R.I.M. computes a deterministic "Fiscal Efficiency Score" (FES) implemented in TypeScript.

Let E denote total expenses and W the set of flagged items with amounts w_i and risk weights $\rho_i \in \{1.0, 0.6, 0.3\}$ (High/Medium/Low). Define total flagged spend T, raw spend percentage P, and risk-weighted percentage R:

$$T \triangleq \sum_{i \in W} w_i, \quad P \triangleq 100 \cdot \frac{T}{E}, \quad R \triangleq 100 \cdot \frac{\sum_{i \in W} \rho_i w_i}{E}.$$

We blend the two into a weighted spend percentage P_w:

$$P_w \triangleq 0.7P + 0.3R.$$

Note: The weights of 0.7 (raw spend) and 0.3 (risk-weighted) were determined empirically during beta testing. We prioritized the absolute volume of potential anomalies over the severity classification, as misclassification of risk (e.g., Medium vs. High) is more common than miscalculation of amounts.

We compute four penalties: spend (SP), volume (VP), concentration (CP), and budget concentration (BCP). Let H, M, L be the counts of High/Medium/Low items, $w_{(1)}, w_{(2)}$ the two largest flagged items, and let $D = \max_j d_j$ where d_j is the percent share of department j.

$$\begin{aligned} \text{SP} &\triangleq \min(70, 8P_w), \\ \text{VP} &\triangleq \min(12, 1.5H + 0.75M + 0.25L), \\ C &\triangleq \tfrac{w_{(1)}+w_{(2)}}{T}, \\ \text{CP} &\triangleq \min(8, \max(0, 20(C - 0.65))), \\ \text{BCP} &\triangleq \min(8, \max(0, 0.3(D - 45))). \end{aligned}$$

Finally, the Fiscal Efficiency Score is:

$$S \triangleq \text{clamp}_{[0,100]}\Big(\text{round}(100 - \text{SP} - \text{VP} - \text{CP} - \text{BCP})\Big).$$

This formulation guarantees objectivity: identical inputs always yield identical scores.

VI. Process: Iteration and Resilience

The development of T.R.I.M. was characterized by rapid iteration and the overcoming of significant engineering hurdles.

A. Phase 1: The Naive Approach

Initially, we attempted to convert PDFs to plain text using Python libraries like `PyPDF2` before processing. However, municipal budgets rely heavily on complex, multi-column tables. Converting these

to plain text destroyed the semantic relationship between column headers (e.g., "2024 Actual" vs "2025 Projected") and the data. The AI began hallucinating because the spatial layout was lost.

B. Phase 2: Multimodal Breakthrough

We migrated to Gemini's multimodal capabilities [9]. By passing the PDF as a file object, the model "sees" the layout, preserving the tabular structure. This improved extraction accuracy of tabular data significantly.

C. Phase 3: The Latency Problem

Analyzing a 500-page document takes time—often 60 to 90 seconds. In web development, this is an eternity. Users would abandon the site assuming it had crashed. *Solution:* We implemented a real-time polling architecture with optimistic UI updates.

1) **Optimistic Progress Tracking:** The frontend displays a detailed progress bar (utilizing `framer-motion`) that informs the user of specific stages ("Scanning PDF...", "Verifying Citations...", "Generating Score..."). This is implemented using a duration-based easing function synchronized with the average processing time, ensuring a smooth user experience while the backend executes.
2) **Local Caching:** We implemented `Dexie.js` to store user session data in `IndexedDB`. If a user refreshes the page or returns later, the browser instantly loads their previous audit without hitting the server.

D. Phase 4: Building Trust through Evidence

Early testers noted they didn't trust the AI's claims. They wanted proof. *Solution:* We engineered a "Deep Link" PDF viewer. By extracting the `pageIndex` during the AI pass, the frontend constructs a URL fragment (e.g., `#page=42`) and opens the source PDF in a modal iframe directly to the evidence. This feature transformed T.R.I.M. from a novelty into a verified auditing tool.

VII. Case Study: City of Chicago, Illinois

To demonstrate the real-world applicability of T.R.I.M. at scale, we applied the system to the City of Chicago's 2025 adopted municipal budget overview. Chicago represents an ideal stress test for the platform: it is one of the largest municipal governments in the United States, with a highly diversified spending structure spanning public safety, infrastructure, pensions, debt service, and citywide administrative functions.

To simulate the "Transparency Paradox," we ingested the City's official Budget Overview PDF, a 200+ page document combining high-level financial summaries, departmental narratives, and capital planning explanations into a single, heterogeneous artifact. While publicly available, the document requires extensive cross-referencing to understand how major funding categories relate to one another, particularly across operating and capital contexts.

A. Findings

T.R.I.M. processed the complete Chicago Budget Overview in approximately 60–90 seconds of model processing time following upload. The Gemini 3 Pro Auditor identified several areas of low transparency where large expenditure totals are aggregated under broad labels that obscure underlying structure without careful manual review.

One prominent example surfaced in the *General Financing Requirements* section, where multi-billion-dollar expense categories—including pension contributions, debt service, and other citywide obligations—are summarized into high-level budget buckets. While these figures are accurate and legally mandated, their aggregation makes it difficult for a non-expert reader to trace how costs originate or evolve across departments without navigating multiple sections of the document. T.R.I.M. flagged these entries as "low-transparency" items and linked them directly to their originating pages for user verification.

The system also highlighted ambiguity at the boundary between operating and capital spending. The Budget Overview discusses the Capital Improvement Program (CIP) separately from the annual operating budget and explicitly explains their relationship, including distinct funding sources and planning horizons. However, narrative descriptions of capital-related expenditures appear near operating budget summaries, creating a realistic risk of misinterpretation during manual review.

Rather than presenting these findings as abstract summaries, T.R.I.M. surfaced each flagged item with page-level citations and deep links, enabling users to immediately inspect the surrounding context within the original PDF.

B. Validation in Action

During an initial extraction pass, the Auditor agent tentatively categorized a capital-related expenditure description as discretionary operating spending due to its proximity to operating budget narrative text. This misclassification illustrates a common failure mode when interpreting large municipal documents: capital investments are often described alongside operating context without explicit labeling on every page.

The Gemini 3 Flash Judge detected the discrepancy by cross-referencing the cited page with the document's Capital Improvement Program discussion and the section explaining the relationship between capital planning and the annual operating budget. Upon verification, the Judge correctly reclassified the item as long-term infrastructure spending and removed it from the "high-risk" review set before the Fiscal Efficiency Score was computed.

This correction demonstrates the practical value of the dual-model "LLM-as-a-Judge" architecture when operating on complex, narrative-heavy government documents where context spans dozens of pages.

C. Outcome

Following validation, T.R.I.M. generated a comprehensive city-level audit dashboard for Chicago, including a deterministic Fiscal Efficiency Score, department-level transparency indicators, and evidence-linked review flags. Users could interactively explore why the score was assigned, inspect areas of budget aggregation or ambiguity, and trace every flagged item back to its precise location in the source PDF within seconds.

This case study illustrates T.R.I.M.'s core value proposition: transforming a sprawling, opaque municipal budget into a readable, explainable, and verifiable audit artifact—without relying on city-provided dashboards or curated datasets, and while preserving human judgment as the final authority.

VIII. Evaluation and Validation

To quantify the reliability of T.R.I.M., we developed a rigorous evaluation protocol contrasting AI extraction against a manually verified ground truth. This section details our methodology, the dataset selection, and the comparative performance of our Dual-Model architecture versus a standard single-model approach.

A. Evaluation Protocol: Defining the Human Baseline

To establish a "Human Baseline," we selected 50-page random samples from four diverse municipal budgets. The four authors of this paper independently annotated every line item considered "discretionary" or "high-risk" according to a shared rubric identical to that provided to the AI system.

The validity of the Ground Truth was enforced through strict consensus and accuracy criteria:

- **Ground Truth Consensus:** Items were included in the validated set only if at least 3 out of the 4 annotators agreed on both the classification (e.g., "Waste" vs. "Essential") and the associated dollar amount.
- **Hit/Miss Criteria:** A "Hit" was recorded only if the AI extracted:
  1) The correct dollar amount (within a 1% tolerance for rounding differences);
  2) The correct spending category/classification;
  3) The correct page-level citation.

  Any deviation in these three fields resulted in a "Miss."

B. Dataset Selection and Justification

We validated T.R.I.M. against four municipalities chosen not for convenience, but to represent distinct structural challenges in auditing:

1) **Chicago, IL (Legacy Complexity):** A massive urban budget characterized by entrenched legacy pension obligations, large-scale debt service, and highly aggregated financing categories that require extensive cross-referencing to interpret correctly.
2) **San Francisco, CA (High Variance):** Represents a high-cost-of-living area where high dollar amounts for services might be flagged

as anomalies by a naive model but are standard for the region.
3) **New York City, NY (Scale):** Used to stress-test the token context window with the largest municipal budget in the US.
4) **Aurora, IL (Suburban/Mixed):** Represents a mid-sized city with heterogeneous budget formatting, including scanned, non-native PDF tables that require OCR or multimodal interpretation.

C. Correction Efficacy: The Judge's Role

The introduction of the "Judge" agent (Gemini 3 Flash) provided the most significant boost in reliability. Table II demonstrates concrete examples where the Auditor (Gemini 3 Pro) made semantic errors that the Judge successfully caught and corrected before the user ever saw the data.

TABLE II: Examples of "Auditor" Errors Corrected by the Judge

Example A: Category Misread (San Francisco)
Auditor: { "item": "Community-Based Services (Port)", "amount": $4,600,000, "flag": "Waste" } *Context:* The phrase "community-based services" was misread as a general grant program. ↓ **Judge:** RECLASSIFIED to "Program Support / District Services (Medium Risk)" *Reason:* "The cited text ties the increase to programs at the Fisherman's Wharf Community Benefit District; amount is valid, but 'waste' framing is unsupported."
Example B: Directionality Error (Chicago)
Auditor: { "item": "Cultural Grants and Resources", "amount": $12,612,120, "claim": "cost savings" } *Context:* Narrative language around investments/initiatives was incorrectly summarized as savings. ↓ **Judge:** CORRECTED CLAIM to "spending allocation" (kept amount) *Reason:* "The cited section describes a funded program/budget line, not a reduction or savings."

D. Quantitative Results

Table III presents the performance metrics of the full T.R.I.M. pipeline against a solo Gemini 3 Pro pass. The metrics are defined as follows:

1) **Total Budget Extraction:** The recall rate of relevant line items present in the Ground Truth set.
2) **Risk Classification:** The accuracy of the AI in matching the human consensus on the *nature* of the expense (e.g., correctly identifying a luxury retreat vs. standard travel).
3) **Citation Accuracy:** The percentage of extracted items where the provided page index matched the physical location in the PDF.
4) **Hallucination Rate:** The percentage of extracted items that were factually non-existent or gross misinterpretations of the source text.

TABLE III: T.R.I.M. Accuracy vs Human Baseline

Metric	**Gemini 3 Pro + Flash Judge**	**Gemini 3 Pro (Solo)**
Total Budget Extraction	**90.9%**	63.4%
Risk Classification	**94.5%**	92.0%
Citation Accuracy	**73.1%**	52.0%
Hallucination Rate	**6.4%**	18.8%

The data reveals that while the Solo model is capable of finding items (63.4% extraction), it struggles significantly with grounding them (52.0% citation accuracy) and frequently invents or misinterprets data (18.8% hallucination). The Dual-Model architecture drastically reduces hallucinations to 6.4% and improves citation accuracy to 73.1%. The remaining error rate primarily stems from extremely poor quality scanned documents where OCR fails entirely, a limitation we address by designing the UI to force human verification via deep links.

IX. Limitations and Future Work

A. Accuracy Under Extreme Budget Complexity

While T.R.I.M. demonstrates strong performance across large municipal budgets, accuracy remains bounded by the inherent complexity and heterogeneity of real-world financial documents. Residual errors arise primarily in edge cases, including densely packed multi-year comparison tables where column headers span multiple pages, or ambiguous fund codes defined only once in appendices.

Importantly, T.R.I.M. is intentionally designed as a *decision-support* system, not an autonomous authority. All flagged items are presented with direct PDF citations and deep links, enabling users to independently verify claims before drawing conclusions.

B. Autonomous PDF Discovery and Ingestion

The current pipeline relies on a combination of the Gemini agent and targeted web retrieval to locate official budget PDFs. However, this process

is frequently hindered by modern web architectures and security measures. Many municipal websites employ client-side rendering and dynamic content (often referred to as complex web architectures), which can obscure direct links from simple crawlers. Furthermore, sophisticated anti-bot protections and browser fingerprinting defenses often block automated access to public records. Future work will focus on integrating headless browser environments (for example, a server dedicated to Playwright calls that our serverless website can communicate with) and secure connection protocols to navigate these obstacles, ensuring that T.R.I.M. can autonomously ingest data from even the most technically obfuscated municipal repositories.

C. Structured Data Integration

T.R.I.M. currently operates on unstructured PDFs, which represent the most common but least machine-friendly format for municipal budgets. As governments gradually adopt structured reporting standards such as XBRL, future iterations of T.R.I.M. will ingest these sources alongside PDFs. Hybrid ingestion has the potential to further reduce citation ambiguity and hallucination rates.

X. Ethical Reflection and Educational Impact

Working on T.R.I.M. fundamentally deepened our understanding of the "Black Box" nature of Large Language Models. Initially, we treated the AI as an oracle—assuming that if it said a city allocated funds to a "luxury retreat," it was true. Our early failure cases taught us that LLMs are probabilistic, not deterministic; they will prioritize completing a pattern over factual accuracy if not constrained.

This realization shifted our development focus from "Prompt Engineering" to "System Engineering." We learned that responsible AI use requires building guardrails (like the Judge agent) that assume the model is wrong until proven right. Ethically, we also wrestled with the weight of our output. Flagging an expenditure for review is serious. This motivated our decision to enforce mandatory deep-linking: AI should never be the final judge, only the investigator. By forcing the user to view the source PDF, we ensure that human judgment remains the ultimate arbiter of accountability.

XI. Conclusion

T.R.I.M. proves that the complexity of government is not an insurmountable obstacle. By combining the semantic power of Gemini 3 Pro, the adversarial validation of the LLM-as-a-Judge architecture, and the principles of open democracy, we have built a tool that restores power to the taxpayer.

In an era where every dollar counts, T.R.I.M. ensures that we can finally answer the question: "Where did the money go?" It is a solution built for the people, by students, ready to serve the priorities of a more efficient, accountable, and transparent America.

Acknowledgments

The authors used AI-based tools to assist with drafting, editing, and refining portions of the manuscript, as well as for programming support during implementation, including LaTeX refinement (via Gemini 3 Pro) and code assistance (via GPT-5.2-Codex). The overall system design, methodological choices, experimentation, evaluation framework, and interpretation of results were led, implemented, and validated by the authors.

References

[1] Z. Y., "Mountain View school board member Devon Conley faces recall attempt," *Mountain View Voice*, June 13, 2025. [Online] Available: https://www.mv-voice.com/election/2025/06/13/mountain-view-whisman-school-board-member-devon-conley-faces-recall-attempt/.

[2] U.S. Census Bureau, "2023 Annual Survey of State and Local Government Finances," Washington, D.C., released Jun. 2025.

[3] Pew Research Center, "Public Trust in Government: 1958–2025," Dec. 4, 2025. [Online]. Available: https://www.pewresearch.org/politics/2025/12/04/public-trust-in-government-1958-2025/.

[4] U.S. Government, "America's AI Action Plan: Accelerating AI Adoption in Government," 2025. [Online]. Available: https://www.ai.gov/action-plan.

[5] Digital Accountability and Transparency Act of 2014, Pub. L. No. 113–101.

[6] Next.js, "App Router Documentation," Vercel, 2024. [Online]. Available: https://nextjs.org.

[7] Google DeepMind, "[Gemini 3 Pro] External Model Card - Dec 4, 2025 - v7 DT update," Jan. 2026. [Online]. Available: https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf.

[8] L. Zheng *et al.*, "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena," *arXiv preprint* arXiv:2306.05685, 2023.

[9] G. Team, "Gemini 1.5: Unlocking multimodal understanding across millions of tokens," *arXiv preprint* arXiv:2403.05530, 2024.