CS269 Final Project Report: Building a Framework for Explainability and Accountability for Albania's AI Minister, "Diella"

Chenyang Zhao
UCLA
zhaochenyang20@gmail.com

Genglin Liu
UCLA
genglinliu@gmail.com

Mingqi Zhao
UCLA
mizhao34@g.ucla.edu

Jaelyn Fan
UCLA
jfan981@ucla.edu

Abstract

Using Albania's AI Minister, "Diella," as a case study, this project addresses the risks of opaque "black box" models in high-stakes public governance. We argue that such systems must be inherently interpretable, not merely post-hoc explainable. We propose a framework mandating both technical interpretability and democratic governance principles. To demonstrate feasibility, we will build a prototype interpretable model for public procurement that provides transparent, human-readable justifications for its decisions, enabling true accountability.

1 Motivation

Albania's AI Minister, "Diella," who oversees public procurement, exemplifies the risks of using "black box" models for high-stakes governance. Such opaque systems lack accountability and can perpetuate hidden biases. Inspired by Cynthia Rudin's work (Rudin, 2019), we argue that post-hoc explanations are insufficient for decisions with major political and economic consequences. The goal of this project is to design a framework that requires the use of **inherently interpretable models**, ensuring that AI in the public service is transparent, fair, and subject to democratic scrutiny.

2 Related Work

In September 2025, the Albanian government appointed an artificial-intelligence system named *Diella* as Minister of Public Procurement, the world's first AI-driven ministerial position. Although introduced as an anti-corruption measure, the system's opaque decision-making process quickly sparked concerns about algorithmic accountability and democratic oversight (Reuters, 2025)(Arab News, 2025). Critics warned that delegating high-stakes governance powers to automated systems without built-in interpretability or institutional safeguards could blur lines of responsibility, erode public trust, and ultimately undermine democratic legitimacy. This case demonstrates the governance risks of deploying black-box AI systems without civic and institutional accountability mechanisms.

The theoretical basis for these concerns can be traced to (Rudin, 2019), who argues that black-box models are fundamentally incompatible with high-stakes domains such as healthcare, criminal justice, and public governance. Because their internal logic cannot be directly examined or verified, users must rely on post-hoc explanation methods to approximate a model's reasoning rather than truly understand its decision-making process. This "false transparency" creates the illusion of fairness without real accountability: external observers cannot detect biases or errors within the system, nor trace responsibility when failures occur. Hence, (Rudin, 2019) calls for the adoption of *inherently interpretable models*, whose decision logic is understandable, verifiable, and contestable by human stakeholders. Empirical research by (Slack et al., 2019) further supports this theoretical stance. Through adversarial experiments, they demonstrate that commonly used post-hoc explanation tools such as LIME and SHAP can be intentionally manipulated to produce misleading interpretations. In their experiments, biased models continued to discriminate against specific groups while appearing "fair" under LIME and SHAP explanations. This "packaged transparency" reveals how post-hoc methods can mask systemic bias, confirming (Rudin, 2019)'s claim that interpretability must be embedded in model design rather than added retroactively.

From a governance perspective, existing policy frameworks such as the OECD AI Principles (OECD, 2019) and the EU AI Act (European Commission, 2024) stress transparency and accountability but fail to specify technical standards for ensuring interpretability in practice.

3 Data

To build a model that accurately reflects Albania's specific procurement landscape, we shifted from theoretical proxies to the official procurement records for the 2025 fiscal year. This data was acquired directly from Albania's Public Procurement Agency (APP) portal (app.gov.al).[1]

3.1 Challenges and Data Preprocessing

This authentic dataset, however, presents several key challenges. First, all data is published in Albanian. To ensure analytic readability, we translated the column headers into English (e.g., Fondi_limit to limit_fund). We intentionally retained the original Albanian content for all records to maintain data integrity and avoid nuances lost in automated translation. Second, the dataset is a registry of awarded contracts, meaning it only contains information on *winning* bidders; data on unsuccessful participants is not included. This limits our analysis to the characteristics of successful bids rather than a direct classification of winners versus losers. Finally, the records lack explicit features for company history or company size, and external retrieval of these auxiliary data is difficult.

3.2 Feature Engineering

To overcome these data limitations, we performed feature engineering to derive behavioral and risk indicators directly from the available procurement records. We constructed features at both the **tender level** (capturing transactional risk within a single contract) and the **company level** (proxying for historical behavior across the 2025 fiscal year).

At the **tender level**, we engineered fund_usage, defined as the ratio of winner_value to limit_fund (capped at 1.0), to measure how closely award values approach the stated budget. We also captured potential post-award adjustments via value_changed_at_contract_signing, computed as the difference between the final signed contract value and the winning bid. To represent **restricted competition**, we included boolean indicators for single-participation cases: is_single_bidder (only one submitted bid) and is_single_qualified_bid (only one qualified bid). Finally, we incorporated **procedural and compliance-related signals** through tender_duration_days (time between publication and closing) and is_over_budget (bids exceeding limit_fund).

At the **company level**, we aggregated statistics for each unique winning supplier (identified by winner_nipt) across all awarded tenders. These company-wide indicators include company_total_wins (total number of contracts won) and company_total_value (total monetary value of all contracts won), which together proxy for the supplier's **overall footprint in public procurement**. We additionally computed company_avg_fund_usage as the supplier's average fund_usage ratio, company_single_bid_win_rate as the fraction of wins occurring in single-bid tenders, and company_cancellation_rate as the proportion of awarded tenders that were later canceled.

4 Methodology

We employed a comparative methodology to evaluate the reliability of AI in procurement. We compared two distinct risk assessment models: a transparent, OECD-aligned rule-based framework (our proposed solution) and a black-box LLM baseline (representing the current trend).

4.1 Model 1: The Rule-Based Framework

To ensure transparency and auditability, we constructed a risk-scoring framework[2] grounded in the integrity standards outlined by the OECD Public Governance Policy Papers (OECD, 2023). The OECD framework emphasizes that procurement risks are multifaceted, requiring the identification of risks throughout the procurement cycle. Specifically, we decomposed integrity risk into four clear, measurable dimensions that map directly to observable data signals.

1. Integrity Risk (Signals of Restricted Competition) This dimension targets signals of restricted or suspicious competition, aligning with the OECD's focus on detecting collusion and bid-rigging. We primarily monitor cases with only one bidder, only one qualified bidder, or suppliers who exhibit a disproportionately high historical win rate in single-bid tenders. Patterns of winning without competition serve as primary indicators of potential collusion or market allocation, which are core concerns in OECD integrity guidelines.

[1] Original data source: https://www.app.gov.al/export-public-calls/

[2] The code for our framework is available at: https://github.com/jfan981/Albania-Procurement-Framework

2. Process Risk (Procedural Irregularities) This dimension captures anomalies in the tendering procedure itself, reflecting the OECD's emphasis on compliance and procedural fairness. The key indicators include the number of qualified bids and unusually low bidder participation relative to the specific contract type. When participation drops below expected levels for a given category (e.g., "Open Procedure"), it suggests the process may not have been fully open or accessible to the market, potentially violating equal treatment principles.

3. Financial Risk (Fiscal Red Flags) This dimension focuses on fiscal discipline and pricing anomalies, addressing the OECD's concerns regarding value for money and fiscal responsibility. We analyze `fund_usage` ratios approaching 1.0, contract cancellation rates, and recurring single-bid tendencies. Consistent bidding at the exact budget limit or frequent post-award cancellations can signal weak financial oversight or price manipulation risks that require closer scrutiny.

4. Delivery Risk (Supplier Reliability) This dimension measures a supplier's past performance and reliability, corresponding to the OECD's focus on operational and contract management risks. We aggregate data on a supplier's contract cancellation rate, past delivery failures, or patterns of problematic performance across multiple tenders. Repeated performance failures indicate that a supplier may be unfit for future government contracts, posing a direct threat to the successful delivery of public services.

4.1.1 Risk Level Assignment

To generate a final decision, we employ a two-step transformation. First, we normalize each of the four dimensions and combine them into a single score using logistic scaling. This compresses the output into a continuous, interpretable **Risk Score** $\in [0, 1]$. Second, to ensure stability in categorization, we use quantile-based cutoffs rather than arbitrary thresholds. This approach keeps risk categories stable even if the underlying score distribution shifts. Specifically, we classify the bottom 60% of scores as **Low Risk**, the next 25% (60%–85%) as **Medium Risk**, and the top 15% (85%–100%) as **High Risk**. The distribution of these scores is shown in Figure 1, where a meaningful tail extends into high-risk ranges, identifying cases requiring scrutiny.

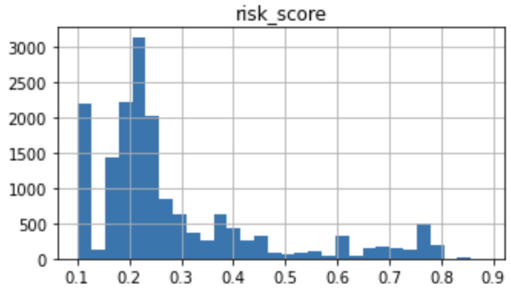

Figure 1: Risk Level Assignment Logic. The four normalized risk dimensions are logistically scaled into a continuous score in [0,1] and then categorized using quantile-based cutoffs: Low (0–60%), Medium (60–85%), High (85–100%).

As shown in Figure 1, most tenders cluster between 0.1–0.3, reflecting relatively low systemic risk in the procurement dataset. The long right tail motivates the use of quantiles instead of fixed thresholds.

4.2 Model 2: LLM-Based Inference (Baseline)

To investigate whether Large Language Models can perform integrity reasoning as reliably as our rule-based model, we evaluated GPT-4 on a subset of 300 tenders.

Constructing Natural-Language Summaries To ensure a fair comparison, we provided the LLM only with the raw signals used by our rule model. We generated a natural-language summary for each tender that included the contracting authority, procurement object, contract type, number of qualified bids, single-bidder status, supplier's historical win patterns, and the fund usage ratio. Crucially, no risk labels, hints, or rule-based outputs were provided to the LLM, ensuring its assessment was independent.

Evaluation Task We prompted the LLM to read the case summary and assign a risk level of **Low**, **Medium**, or **High**. Additionally, we required the model to provide a 2–4 sentence explanation justifying its decision based on the provided bidding, financial, and historical signals. This setup allows us to directly analyze where the LLM's free-form reasoning agrees or disagrees with the OECD-aligned framework.

5 Experiments and Results

We evaluated the models on a stratified subset of 300 samples. Our analysis focuses on the alignment between the LLM's assessment and the strict policy logic of the Rule-Based model.

5.1 Risk Distribution Analysis

Our rule-based model successfully captured meaningful variation across the four risk dimensions.

As shown in Figure 2, Integrity and Delivery risks are heavily skewed toward zero, reflecting that most suppliers have clean histories. However, Process risk shows a spike near 0.9 (indicating procedural irregularities), and Financial risk shows a wide spread, reflecting diverse fiscal behaviors.

5.2 Quantitative Analysis: LLM vs. Rule-Based

The alignment between the LLM and the Rule-Based model was only weak to moderate. This moderate correlation suggests the LLM uses fundamentally different criteria to assess risk than the OECD framework.

Metric	Value
Overall Accuracy	0.61
Macro-F1 Score	0.58
Spearman Correlation (ρ)	0.447

Table 1: Overall performance metrics comparing LLM predictions against the rule-based ground truth.

Class-Level Performance As shown in Figure 3, the confusion matrix reveals significant inconsistencies across risk categories. As detailed in Table 2, the LLM rarely identifies Low-risk cases correctly, with a recall of only 0.27. Instead, it frequently upgrades them to Medium or High. Conversely, for Medium-risk cases, the model acts as an "uncertainty bucket," capturing 88% of true Medium cases but with low precision (0.52), indicating it defaults to this label when unsure.

Risk Class	Precision	Recall	F1-Score
Low Risk	0.93	0.27	0.42
Medium Risk	0.52	0.88	0.65
High Risk	0.65	0.67	0.66

Table 2: Class-level performance metrics. Note the extremely low recall for Low-risk cases.

5.3 Systematic Biases

We identified clear directional biases in the LLM's errors. First, there is a **Strong Upward Bias on Low-Risk Cases**, where 73% of Low-risk cases (74/102) were incorrectly flagged as Medium or High. This high false-alarm rate dilutes the attention of auditors. Second, we observed a **Downward Bias on High-Risk Cases**, where 32% of High-risk cases were downgraded to Medium. This is critically dangerous, as it means the LLM underestimates complex, multi-factor red flags.

5.4 Qualitative Failure Mode Analysis

By analyzing the mismatched cases, we identified three recurring reasoning failures in the LLM:

1. Misinterpretation of Scale (The "Success is Suspicious" Fallacy): The LLM frequently treated "many contracts" or "high total value" as inherent evidence of corruption, even when the supplier had low cancellation rates and won in competitive environments. It effectively penalized successful companies, confusing market leadership with monopoly. This drove many Low-risk cases into the Medium/High categories.

2. Contextual Blindness (Zero-Value Bias): In cases where data fields were missing (value = 0.0), the LLM interpreted this as direct evidence of fraud ("zero-dollar contract"), whereas the rule-based model correctly treated it as a data artifact.

3. Inconsistent Signal Weighting (Middle-Label Bias): The LLM often downplayed critical red flags (like single-bidder status) if they were offset by irrelevant positive traits (e.g., "no history of cancellations"). This tendency to "average out" signals collapses clear extremes into the Medium category, avoiding decisive judgments.

6 Embedded Ethics Discussion

6.1 Module: "The Illusion of Intelligence in Governance"

As part of an AI curriculum, we propose a module titled "The Illusion of Intelligence in Governance." This module addresses the ethical risks of deploying generative AI in bureaucratic decision-making roles where accountability is paramount.

Problem Definition for Students: Students will be introduced to the concept of "Bureaucratic Hallucination"—the phenomenon where an LLM mimics the *style* and *tone* of an expert auditor without adhering to the *constraints* of law or policy. Using the "Diella" case study, we explain that while an LLM can write a convincing paragraph justifying a decision, its underlying logic is probabilistic, not rule-bound. This creates a "accountability gap": if

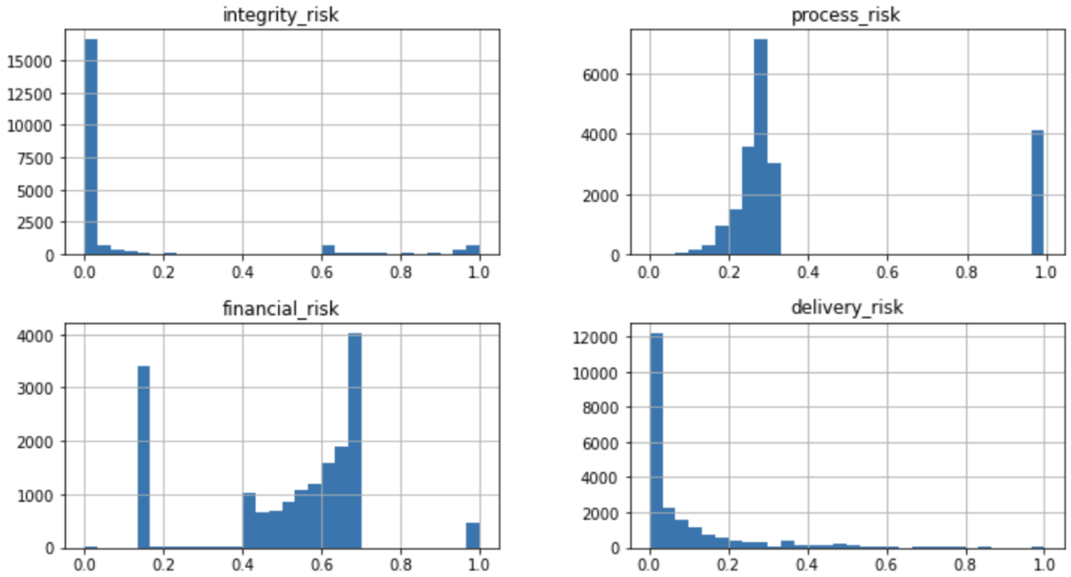

Figure 2: Distribution of Risk Dimensions. Integrity and Delivery risks are right-skewed (most cases clean), while Process and Financial risks show wider variance.

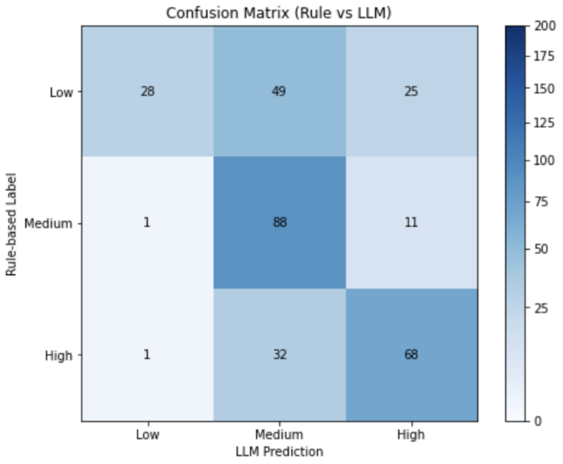

Figure 3: Confusion matrix comparing LLM predictions against rule-based risk labels on 300 samples. The LLM tends to over-assign the Medium risk label and shows low recall for Low-risk cases.

a citizen is denied a contract by an AI that "hallucinated" a rule violation, who is responsible?

Proposed Solution for Students: We introduce the concept of "Inherent Interpretability" as an ethical requirement. Students will learn that for high-stakes government functions, the model's architecture must allow for direct inspection of its decision path (e.g., decision trees, linear scorecards) rather than relying on post-hoc explanations of a neural network.

Actionable Coding Assignment: "Audit the Auditor" To translate this theory into practice, students will complete a two-part assignment:

1. **Part A: The Rule-Based Auditor (Python).** Students are given a dataset of 50 procurement records (some with clear red flags like single-bidders, others benign). They must write a Python function that implements a strict legal rubric (e.g., `if bids < 2: flag_risk()`) to grade these contracts deterministically. 2. **Part B: The LLM Auditor (Prompt Engineering).** Students then prompt an LLM (e.g., GPT-4) to grade the same 50 records using a natural language summary. They must ask the LLM to "act as an expert auditor" and provide a rationale. 3. **Part C: The Ethics Audit (Reflection).** Students compare the outputs. They must identify at least three specific cases where the LLM's fluent explanation contradicted their strict Python rule (e.g., the LLM forgave a single-bidder violation because the company was "experienced").

This assignment viscerally demonstrates the danger of replacing explicit rules with probabilistic text generation, reinforcing the ethical mandate for transparency in public-sector AI.

7 Conclusion

Our investigation into the "Diella" framework reveals that while LLMs can generate fluent justifications for procurement decisions, they are currently unsuitable for autonomous oversight. They exhibit systematic biases—confusing scale with corruption and softening critical red flags. In contrast, our Rule-Based model, though simpler, offers the consistency, auditability, and adherence to policy required for democratic governance.

Specifically, the LLM demonstrated a dangerous "middle-label bias," collapsing clear high-risk and low-risk cases into a safe "Medium" category, effectively nullifying the utility of a risk alert system. Furthermore, its inability to contextualize data artifacts (like zero values) and its penalization of successful market leaders ("scale bias") suggest it lacks the nuanced understanding of market dynamics required for this task.

Future work should focus on hybrid systems where rule-based logic is used to flag risks deterministically, while LLMs are restricted to summarizing those flags for human auditors, rather than making the judgment themselves. Ultimately, for high-stakes public sector decisions, the "black box" must remain open.

8 Contribution Statement

This project is a collaborative effort by all team members, developed through shared discussions on research direction and project framing. Individual contributions are as follows:

- **Chenyang Zhao:** Designed the overall project proposal, defined the research direction, developed the structural framework, proofread the final report.

- **Mingqi Zhao:** Conducted literature research and construct the *Related Work* section, identified and analyzed relevant academic sources, drafted and proofread the final report.

- **Jaelyn Fan:** Collected and processed the procurement dataset, performed data cleaning and feature preparation for model development, drafted the final report.

- **Genglin Liu:** Designed the methodology, including model selection, experimental structure, and results analysis, proofread the final report.

All members participated in joint discussions, writing, and editing to ensure consistency between the technical, policy, and analytical components of the project.

References

Arab News. 2025. Albania's ai minister diella sparks debate on algorithmic accountability. https://www.arabnews.com/node/2616288/world. Accessed: 2025-11-02.

European Commission. 2024. Regulatory framework for artificial intelligence (eu ai act). https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai. Accessed: 2025-11-02.

OECD. 2019. Oecd principles on artificial intelligence. https://www.oecd.org/en/topics/sub-issues/ai-principles.html. Accessed: 2025-11-02.

OECD. 2023. Managing risks in the public procurement of goods, services and infrastructure. OECD Public Governance Policy Papers 33, OECD Publishing, Paris.

Reuters. 2025. Albania appoints ai bot minister to tackle corruption. https://www.reuters.com/technology/albania-appoints-ai-bot-minister-tackle-corruption-2025 Accessed: 2025-11-02.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Preprint*, arXiv:1811.10154.

Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. 2019. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. *arXiv preprint arXiv:1911.02508*.