CS269 Final Project Mid-Quarter 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 extensive feature engineering to derive behavioral and risk indicators directly from the available data. We created new features at both the tender level (capturing transactional risk) and the company level (proxying for historical behavior).

Tender-Level Features These features capture risks within a single transaction. We engineered:

- `fund_usage`: The ratio of `winner_value` to `limit_fund`, capped at 1.0.
- `value_changed_at_contract_signing`: The difference between the final contract value and the winning bid.
- `is_single_bidder`: A boolean flag for tenders with only one submitted bid.
- `is_single_qualified_bid`: A boolean flag for tenders with only one qualified bid.
- `tender_duration_days`: The time between publication and closing.
- `is_over_budget`: A boolean flag for bids exceeding the `limit_fund`.

Company-Wide Features To build a proxy for company history and behavior, we aggregated statistics for each unique winning company (identified by `winner_nipt`) across the 2025 dataset. This includes:

- `company_total_wins`: Total number of contracts won.
- `company_total_value`: Total monetary value of all contracts won.
- `company_avg_fund_usage`: The company's average `fund_usage` ratio.
- `company_single_bid_win_rate`: The percentage of a company's wins that were single-bid tenders.
- `company_cancellation_rate`: The percentage of a company's awarded tenders that were later canceled.

Finally, the fully engineered dataset was split into training (80%), validation (10%), and test (10%) sets for model development.[2]

4 Methodology

Our methodology integrates a rule-based interpretable risk framework with a comparative evaluation against large language model (LLM) inference. This dual-track design allows us to assess both the benefits of transparency and the reliability limitations of black-box generative reasoning in public decision-making.

4.1 Interpretable Risk Construction

We construct a transparent and auditable **Risk Level Score** to quantify potential irregularities in each procurement record. The score aggregates three interpretable dimensions: (1) *competition risk*, (2) *budget risk*, and (3) *historical risk*. Each component corresponds to measurable attributes within the dataset such as `company_total_wins`, `fund_usage`, or `company_single_bid_win_rate`.

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

[2] The complete data processing script and resulting datasets are available at: https://github.com/jfan981/Albania-Procurement-Framework

Competition Risk. This dimension evaluates how competitive a bidding process is. It increases when `company_total_wins` is low or when `is_single_bidder` equals one. Intuitively, more winning bids imply lower risk, whereas single-bid or near-monopoly cases indicate limited market competition or potential favoritism.

Budget Risk. This dimension reflects fiscal discipline and transparency. Higher values of `fund_usage` (i.e., winning bid relative to `limit_fund`), or flags such as `is_over_budget` and `value_changed_at_contract_signing`, signal elevated risk. These patterns often imply weak budget enforcement or post-hoc contract renegotiation.

Historical Risk. This dimension captures systemic patterns in supplier behavior. Companies with high `company_single_bid_win_rate` or `company_cancellation_rate` are assigned greater risk, reflecting recurring irregularities or low market accountability.

Each dimension produces a normalized score between 0 and 1 and contributes linearly to a composite risk index, which is then logistic-scaled to remain within a consistent range. The weights assigned to competition, budget, and historical components are tunable by auditors, allowing context-dependent emphasis on market fairness, fiscal prudence, or institutional integrity. This structure ensures that every element of the score remains explicitly interpretable and directly traceable to the raw data fields.

4.2 LLM-Based Inference and Reliability Comparison

To assess the reliability gap between interpretable and opaque reasoning, we compare our rule-based model against an LLM-driven inference approach. For each procurement record, we generate a natural-language summary of all key attributes (e.g., `procedure_type`, `contract_type`, `num_bids_qualified`, `fund_usage`) and prompt a large language model (e.g., GPT-4) to classify whether the case appears “risky,” providing an accompanying textual rationale.

We then evaluate:

- **Agreement rate:** the proportion of cases where the LLM classification aligns with the categorical risk level of our interpretable model.
- **Consistency:** the stability of LLM outputs under paraphrased prompts and varying temperature settings.
- **Explanation quality:** the extent to which LLM rationales remain factual, verifiable, and non-hallucinatory.

This comparison exposes the contrast between a deterministic, data-grounded reasoning process and an opaque, generative one. While LLMs can produce superficially plausible justifications, their reasoning often lacks consistency and auditability, revealing limitations that are unacceptable in high-stakes governance applications.

4.3 Implementation Plan

The interpretable model will be implemented in `Python` using `pandas` and `numpy` for feature engineering and computation. Visualization modules will decompose each record’s `Risk Level Score` into its contributing components for auditor review. The LLM inference pipeline will rely on standardized prompts and temperature-controlled inference for reproducibility. Evaluation metrics include agreement rate, Spearman correlation between numeric risk levels, and qualitative audits of explanation faithfulness.

This methodology directly bridges *technical interpretability* with *institutional accountability*, showing that transparent, rule-based models can provide more reliable and auditable judgments than opaque LLMs in high-stakes public procurement.

5 Expected Timeline (Mid-Quarter)

Weeks	Task
Week 1–2	Literature review (Completed) Dataset identification and acquisition (Completed)
Week 3–4	Draft initial policy framework and model prototype (Completed)
Week 5–6	Data collection, preprocessing, and feature analysis for interpretability (In Process)
Week 7–8	Build interpretable model prototype and rule-based risk-score construction Evaluate model performance
Week 9–10	Integrate and revise full paper Final review, formatting, and submission
Week 11	Finalize & Submit Report

Table 1: Project timeline.

6 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, and developed the structural framework of the report.
- **Mingqi Zhao:** Conducted literature research and construct the *Related Work* section, identifying and analyzing relevant academic sources.
- **Jaelyn Fan:** Collected and processed the procurement dataset, performing data cleaning and feature preparation for model development.
- **Genglin Liu:** Designed the methodology, including model selection, experimental structure, and technical implementation plan.

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.

Reuters. 2025. Albania appoints ai bot minister to tackle corruption. https://www.reuters.com/technology/albania-appoints-ai-bot-minister-tackle-corruption-2025-09-11/. 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*.