Controlled AI for Accountable Data-to-Text Generation in Public Transparency Dashboards

Claudio Biancalana[1], Fabrizio Sbicca[1]

[1]*Italian National Anti-Corruption Authority (ANAC), Rome, Italy*
The opinions expressed in this paper are the authors' own and do not reflect the views of ANAC

Abstract

Public administrations increasingly rely on dashboards, interoperable data sources, and large-scale reporting systems, but the availability of more data does not automatically produce more intelligible public information. Transparency dashboards, such as the Italian *Piattaforma Unica della Trasparenza* (PUT), therefore require explanatory mechanisms that non-expert users can understand and institutions can defend. This paper contributes a **design paradigm for controlled AI** in high-accountability public-sector communication: a methodological framework for reproducible data-to-text generation that prioritises accountability, numerical faithfulness, and inspectability over linguistic fluency alone.

The contribution is not a new generative model nor merely a system implementation, but a reusable framework grounded in four principles: **controlled NLG**, **reproducible text generation**, **explainable ML**, and **governance-compliant AI** in public administration. The paradigm is operationalised through a rule-bound pipeline in which template-based natural language generation produces the surface text, while explainable machine learning is used only for structural decisions.

The pipeline comprises: (i) multi-source aggregation into an administration profile; (ii) template-based NLG with fixed thresholds and stable variant selection; (iii) explainable ML for block selection (committee of classifiers with boosting) and reliability scoring; and (iv) numerical validation. The approach does not rely on generative language models; every sentence derives from explicit rules, templates, or constrained choices.

Large-scale evaluation on 11,033 administrations shows Micro F1 = 0.901 for block selection, with ablation studies confirming the benefit of the full committee over single-model and rule-only baselines. A pilot human evaluation ($N = 50$ texts, three ANAC experts) reports high scores for factual accuracy, readability, and institutional appropriateness. The paradigm is transferable to other public-sector domains (e.g. tax compliance, healthcare quality, environmental monitoring) with minimal adaptation. Overall, the design supports governance-compatible explanatory automation for public-sector AI systems requiring accountability, transparency, and human oversight.

Keywords

Controlled natural language generation, Deterministic text generation, Explainable ML, AI governance, Public administration, Accountability, Transparency-by-design, Piattaforma Unica della Trasparenza

1. Introduction

Public administrations increasingly rely on dashboards, interoperable data sources, and large-scale digital reporting systems, but more data do not automatically produce more intelligible public information. In many administrative settings, the problem is no longer disclosure alone, but the ability to transform complex structured data into explanations that non-expert users can understand and institutions can defend [1]. In Italy, Legislative Decree 33/2013 mandates publication through the *Piattaforma Unica della Trasparenza* (PUT), which centralises indicators across municipalities, regions, ministries, and other bodies [2]. Yet dashboards remain numerically dense, and manual drafting of explanatory texts does not scale [3].

In this perspective, AI for the public sector should not be evaluated only by fluency or automation gains, but by its capacity to support reliable and inspectable communication. Generative systems can produce

Ital-IA 2026: Italian Conference on Artificial Intelligence, 2026, Italy
c.biancalana@anticorruzione.it (C. Biancalana); f.sbicca@anticorruzione.it (F. Sbicca)
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

fluent narratives [4], but stochastic generation makes reproducibility and numerical faithfulness difficult to guarantee in high-accountability settings [5]. This paper therefore presents a controlled AI framework for rule-bound data-to-text generation in public transparency dashboards, using PUT as a motivating case and focusing on governance-compatible explanatory automation [6, 7, 8].

The contribution is a methodological framework for accountable explanation systems, not a new generative model nor merely a system implementation. The framework combines controlled NLG, reproducible generation, explicit provenance of claims, and explainable ML for narrative structure: text is produced by templates and rules, while ML selects blocks and estimates reliability. Contributions are: (i) a framework for controlled text generation in public-sector transparency; (ii) an architecture with numerical validation and reliability scoring for human oversight; (iii) evaluation on 11,033 administrations (Micro F1 = 0.901), supporting deployment in **high-accountability** settings. The dashboard is available at https://dashboard-put.anticorruzione.it/. The paper then reviews related work, presents the methodology and architecture, reports evaluation results, and discusses limitations.

2. Related Work

Data-to-text has been studied in domains such as weather, finance, and sports [6, 4, 9]. Neural and LLM approaches produce fluent text but often lack reproducibility and reliable grounding of numerical claims [5, 10, 11], which is problematic where accountability is required. We align with **controlled NLG** and controllable generation [6, 12], but extend template-based generation with **explainable ML** for narrative structure, keeping the surface text rule-based and verifiable. We use tree-based ensembles (RF, GB) for interpretable block selection [13, 14, 15]. Governance and accountability in public-sector AI [3, 16, 8] motivate this design.

3. Methodological Framework

This work contributes a methodological framework for accountable data-to-text generation: a design paradigm, not a new generative model nor a system implementation alone. Let D be a structured indicator profile and T an explanatory text; the framework defines a reproducible mapping $G\colon D \to T$. Critically: **ML determines only the narrative structure** (e.g. which blocks to include); **text is produced by templates and rules**, so claims can be checked against their input data. Four principles guide the design: (1) **Stable generation**: same $D \Rightarrow$ same T. (2) **Controlled vocabulary and templates**: every sentence from rules, fixed templates, or constrained variant choice; no free-form generation. (3) **Provenance of claims**: every assertion links to a rule and a datum in D. (4) **Explainable ML for structure**: ML components are interpretable and inspectable [15, 8]. Section 4 describes the architecture implementing this framework.

4. Architecture

Figure 1 shows the pipeline. Flow: **Data sources** → **Profile** → **Template engine** → **ML block selector** → **Reliability** → **Validation** → **Dashboard**. (1) **Data aggregation.** Multi-source data are collected and normalised per administration from seven institutional sources: BDNCP,[1] SPL,[2] OIV,[3] BDAP,[4] SIOPE,[5]

[1] BDNCP: Banca Dati Nazionale dei Contratti Pubblici (ANAC).
[2] SPL: Servizi Pubblici Locali (ANAC).
[3] OIV: Organismi Indipendenti di Valutazione (ANAC).
[4] BDAP: Banca Dati Amministrazioni Pubbliche (MEF–RGS).
[5] SIOPE: Sistema Informativo sulle Operazioni degli Enti Pubblici.

Conto Annuale RGS,[6] and ISTAT Popolazione Residente.[7] (2) **Profile** D: 42 features—entity classification, 15 normalised indicator scores (0–100), 8 trend deltas, 12 comparative percentiles; intensity labels via ANAC-defined fixed thresholds. (3) **Template-based generation**: ∼120 templates in 11 blocks (8–12 variants per block [6]); variant selection by hash(IPA) modulo variants, stable across runs. Table 1 shows Trend-block examples. (4) **Block selection**: a committee of interpretable classifiers (RF, GB, ET) with boosting [13, 14] decides which blocks to include; weighted voting. (5) **Reliability scoring**: low scores flag cases for human review. (6) **Numerical validation**: consistency check before release. (7) **Publication**: no text reaches the dashboard without validation; governance can route cases back (dashed arc).

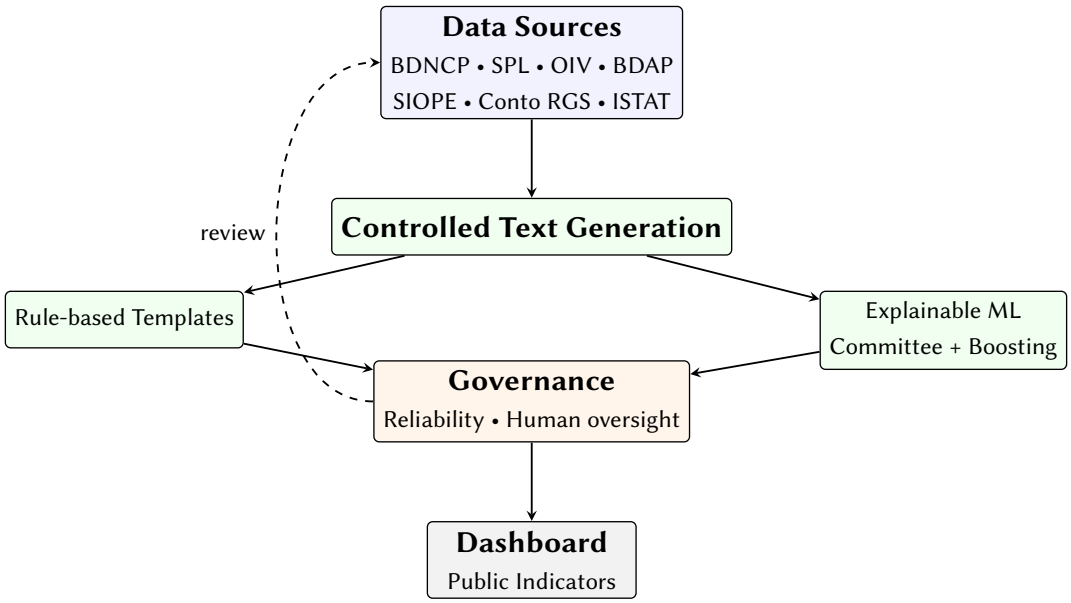


Figure 1: Controlled AI pipeline: data → templates + ML block selection → governance → dashboard.

Table 1
Example template variants for the Trend block; selection is determined by threshold on normalised trend index.

Variant	Condition	Example surface form
A	$trend_idx \in [0, 33)$	"The indicator shows a downward trend over the reference period."
B	$trend_idx \in [33, 66]$	"The indicator remains stable compared to the previous year."
C	$trend_idx \in (66, 100]$	"The indicator shows an upward trend over the reference period."

5. Experimental Evaluation

Why evaluate block selection. Once the narrative structure (which blocks to include) is fixed, templates and rules produce the surface form without further learning. Thus the main variable affecting explanation quality is block selection; evaluating the classifier effectively evaluates the explanation structure. High classifier accuracy means the right narrative components are included or omitted in line with expert annotation; inspectability follows from the constrained generation design.

[6] Conto Annuale RGS: personnel and public employment expenditure data (RGS–MEF).
[7] ISTAT Popolazione Residente: resident population data, Istituto Nazionale di Statistica.

Dataset. 11,033 administrations (2019–2024), 42 features per profile (Section 4), expert-annotated labels. Split: 70% train, 15% val, 15% test (stratified); see Table 2.

Model. Tree-based models (RF [13], GB [14], Extra Trees) provide interpretable decision structures [15], which support institutional review. The committee combines them for robustness; each member is trained for multi-label (One-vs-Rest) block selection; predictions are combined by weighted majority voting, with weights learned via AdaBoost on the validation set.

Metrics. Precision, Recall, F1 (macro and micro); Micro F1 is the main metric.

Results. The committee was evaluated using 5-fold stratified cross-validation in addition to the held-out test set. Test set metrics are reported in Table 2: Micro F1 = 0.901 (95% CI [0.893, 0.909], bootstrap $B = 1000$), Micro Precision = 0.908, Micro Recall = 0.894. Cross-validation yielded Micro F1 $= 0.899 \pm 0.008$ (mean $\pm$ std across folds), demonstrating stability. Reliability scoring correlates with ground truth at $\rho = 0.78$ ($p < 0.001$, $n = 11{,}033$); 82% of cases below reliability threshold $\tau = 0.6$ (chosen on the validation set) had errors vs. 9% above, directing human review effectively. Residual errors concentrate on context-dependent blocks (Territorial comparisons, Significant years); the reliability score and threshold $\tau = 0.6$ direct human review to the most uncertain cases. Table 2 reports F1 per narrative block: Introduction, Profile diagnosis, and Synthetic indices reach F1 > 0.92; Territorial comparisons and Significant years reach F1 ≈ 0.84–0.85.

Table 2
Experimental results (final committee model): aggregate metrics (left), dataset split (centre), F1 by block (right). 95% CI from bootstrap ($B = 1000$); 5-fold CV Micro F1 0.899 ± 0.008.

Metric	Value	95% CI
Micro Prec.	0.908	–
Micro Rec.	0.894	–
Micro F1	0.901	[0.893, 0.909]
Macro Prec.	0.887	–
Macro Rec.	0.872	–
Macro F1	0.879	–

Split	N	%
Train	7,723	70
Val.	1,655	15
Test	1,655	15

Block	F1
Introduction	0.93
Profile diagn.	0.92
Synthetic idx.	0.93
Territ. comp.	0.84
Signif. years	0.85

5.1. Baseline Comparisons and Ablation Study

We compared four configurations: (a) full committee (RF+GB+ET, AdaBoost); (b) single RF; (c) majority-vote committee (no AdaBoost); (d) rule-only heuristic (no ML). Table 3 reports results; the full committee outperforms all baselines; the rule-only baseline confirms that learned block selection is necessary. This work does not aim to outperform generative LLMs (e.g. GPT-4 or T5), but to demonstrate an **alternative design paradigm** for settings where governance or regulatory requirements demand reproducibility, verifiability, and human oversight [8]. Future work may include comparison with generative approaches on a PUT sample and a human-annotation baseline (e.g. Cohen's κ) for block selection.

Table 3
Ablation study: full committee vs. single RF, majority-vote committee (no AdaBoost), and rule-only heuristic. Test set metrics.

Configuration	Micro F1	Macro F1	Micro Prec.	Micro Rec.
Full committee (RF+GB+ET, AdaBoost weights)	0.901	0.879	0.908	0.894
Single Random Forest	0.872	0.851	0.881	0.863
Majority-vote committee (no AdaBoost)	0.887	0.864	0.895	0.879
Rule-only heuristic (no ML)	0.814	0.792	0.831	0.798

5.2. Text Quality Evaluation

Text quality is determined by block selection accuracy (above) and template correctness (verified by design). We conducted a pilot human evaluation: $N = 50$ texts, three ANAC experts, 5-point Likert on Factual Accuracy, Readability, and Institutional Appropriateness (mean scores 4.3, 4.1, 4.5; Krippendorff's $\alpha \approx 0.71$). We recommend larger-scale validation before wider deployment. BLEU, ROUGE, and BERTScore are not applicable—there is no reference corpus of human-written PUT explanatory texts; domain-specific controlled NLG is evaluated via task-oriented and human judgements.

6. Discussion

The results support that rule-based NLG with explainable ML for block selection and reliability scoring can produce explanatory text suitable for **high-accountability** public transparency. The approach makes each narrative reproducible and reviewable, which is essential under public-sector governance requirements [7, 8, 5].

An important implication is that reliability scoring should be seen not only as a model output, but as an operational governance mechanism. By routing uncertain cases to human review, the pipeline preserves institutional responsibility while still allowing scalable production of explanatory texts, consistently with trustworthy AI principles [17].

Transferability. The same design applies to analogous public-sector domains with minimal adaptation: (a) tax compliance dashboards; (b) healthcare quality reports; (c) environmental monitoring bulletins. Required changes are limited to a domain-specific schema, a retrained block-selection committee, and updated templates; the core principles (stable outputs, provenance of claims, explainability, human oversight) stay unchanged. The paradigm is suitable for low-resource institutional settings where governance-compatible generation is required.

7. Limitations

The controllability–expressive richness trade-off is accepted by design; verifiability is primary for high-accountability deployment. Evaluation is limited to one domain (PUT, Italy); residual errors on context-dependent blocks are partly mitigated by the committee and reliability module. Pilot human evaluation ($N = 50$) warrants larger-scale validation before wider deployment.

8. Conclusion

This work proposes a design paradigm for controlled AI in public-sector communication: a methodological framework for accountable automated explanations [6], formalising $G\colon D \to T$ under four design principles and implementing it as a **governance-compliant** pipeline (template-based NLG + explainable ML for block selection). The paradigm targets **high-accountability** public sector AI [3, 8]. Evaluation on 11,033 administrations (Micro F1 = 0.901) supports feasibility. Beyond the PUT case, the results suggest that public institutions can improve the intelligibility of structured data without delegating factual wording to open-ended generation. This is particularly relevant where explanations must be understandable to citizens while remaining reviewable by the administration that publishes them. Future work: replication in other domains, comparison with generative approaches, and hybrid designs preserving controllability.

9. Declaration on Generative AI

The authors employed ChatGPT exclusively for orthographic and grammatical proofreading of the final manuscript draft. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] A. Zuiderwijk, M. Janssen, G. van de Kaa, K. Poulis, The quality of open data for public sector use, Journal of Systems and Information Technology 17 (2015) 268–292. doi:10.1108/JSIT-09-2014-0064.

[2] Repubblica Italiana, Decreto legislativo 33/2013: Riordino della disciplina riguardante il diritto di accesso civico e gli obblighi di pubblicità, trasparenza e diffusione di informazioni, 2013. URL: https://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:decreto.legislativo:2013;33, gazzetta Ufficiale della Repubblica Italiana, n. 80 del 5 aprile 2013.

[3] B. W. Wirtz, J. C. Weyerer, C. Geyer, Artificial intelligence and the public sector—applications and challenges, International Journal of Public Administration 42 (2019) 596–615. doi:10.1080/01900692.2018.1498103.

[4] A. Gatt, E. Krahmer, Survey of the state of the art in natural language generation: Core tasks, applications and evaluation, Journal of Artificial Intelligence Research 61 (2018) 65–170. doi:10.1613/jair.5477.

[5] J. Maynez, S. Narayan, B. Bohnet, R. McDonald, On faithfulness and factuality in abstractive summarization, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), Association for Computational Linguistics, Online, 2020, pp. 1906–1919. doi:10.18653/v1/2020.acl-main.173.

[6] E. Reiter, R. Dale, Building Natural Language Generation Systems, Cambridge University Press, Cambridge, UK, 2000. doi:10.1017/CBO9780511519857.

[7] High-Level Expert Group on Artificial Intelligence, Ethics guidelines for trustworthy AI, European Commission, 2019. URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[8] European Union, Regulation (eu) 2024/1689 of the european parliament and of the council of 24 june 2024 laying down harmonised rules on artificial intelligence, 2024. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689, eU AI Act. Official Journal of the European Union, L 2024/1689.

[9] X. Lin, et al., Data-to-text generation: A survey, IEEE Transactions on Knowledge and Data Engineering (2022).

[10] Z. Ji, et al., Survey of hallucination in natural language generation, ACM Computing Surveys (2023).

[11] N. Dziri, et al., Faithfulness in natural language generation, in: Transactions of the ACL, 2022.

[12] Y. Wang, et al., Controllable text generation: A survey, Artificial Intelligence Review (2024).

[13] L. Breiman, Random forests, Machine Learning 45 (2001) 5–32. doi:10.1023/A:1010933404324.

[14] J. H. Friedman, Greedy function approximation: A gradient boosting machine, The Annals of Statistics 29 (2001) 1189–1232. doi:10.1214/aos/1013203451.

[15] C. Molnar, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, Lulu.com, 2019. URL: https://interpretable.ml, 2nd ed. 2022.

[16] T. Sun, R. Medaglia, Artificial intelligence in government: A review, Government Information Quarterly (2023).

[17] OECD, OECD principles on artificial intelligence, OECD Legal Instruments, 2019. URL: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.