MARTA: Machine Learning Auditing for Robust and Transparent (Public) Administration

Mariana Pinto[1] **Matilde Silva**[1] **Matilde Ferreira**[1] **Beatriz Félix**[1] **Iakovina Kindylidi**[2,3] **Marília Barandas**[1] **Hugo Gamboa**[1,4] **André Carreiro**[1]

[1]Fraunhorfer Portugal AICOS, Porto, Portugal [2]NOVA School of Law, 1099-032, Lisbon, Portugal [3]Vieira de Almeida Associados, 1200-151, Lisbon, Portugal [4]Laboratory of Instrumentation, Biomedical Engineering, Radiation Physics (LIBPhys-UNL), NOVA School of Science and Technology, NOVA University of Lisbon, 2829-516 Caparica, Portugal

Abstract

Artificial intelligence (AI) systems in public administration raise complex challenges for fairness, transparency, and legal accountability. This paper presents Machine Learning Auditing for Robust and Transparent Administration (MARTA), a practical methodology applied to a Smart e-Desk use case being developed for the Portuguese Tax Authority. MARTA proposes a multidisciplinary auditing framework combining technical, legal, and human-centred perspectives across five dimensions: bias, robustness, privacy, transparency, and human oversight. Using methods such as counterfactual bias inference, adversarial robustness evaluation, and privacy-risk analysis, the framework aligns with the HUDERIA methodology of the Council of Europe and extends it through deeper technical evaluation and regulatory mapping to the EU AI Act and GDPR. Results showcased a solid framework with minimal gender imbalance and robust to the most common data variations. While already integrating privacy-preserving and explainability methods, tests suggest partial sufficiency; thus, further refinement is advised. The study contributes a practical model for lawful and trustworthy AI auditing in the public sector. It demonstrates how rights-based principles can be translated into measurable audit procedures and actionable governance measures.

1 Introduction

Artificial Intelligence (AI) systems are increasingly embedded in public administration, supporting decision-making, information management, and citizen interaction processes. While promising efficiency and consistency, these deployments raises complex challenges concerning fairness, accountability, and respect for fundamental rights.

In the European Union, the Artificial Intelligence Act (AI Act) establishes new obligations for stakeholders across the entire AI value chain, including within the public sector. These obligations apply to model and AI system providers, deployers, distributors, and importers, depending on the system's assessed risk level. The Regulation requires demonstrable compliance with key principles such as non-discrimination, robustness, data governance, and human oversight. However, practical methods to assess these properties remain underdeveloped, particularly in operational contexts where AI systems are designed and maintained by third parties but deployed by public authorities.

Growing interest in algorithmic accountability and compliance has led to fast and rich research in fairness, explainability, and robustness testing. In spite of this, a gap persists between technical auditing methods and legal compliance frameworks. Current auditing frameworks provide organizational guidance but lack empirical tools for operational audits in public-sector AI systems (Raji et al. 2020; NIST 2023). Moreover, most auditing efforts occur post-deployment and overlook the socio-technical context of public administration.

Here, we approach these challenges through the case of an AI-assisted helpdesk being developed for the Portuguese Tax Authority - Autoridade Tributária e Aduaneira (AT). The system, referred to as the Smart e-Desk ("e-Balcão Inteligente" in Portuguese), relies on natural-language processing (NLP) models to classify incoming taxpayer queries and route them to the appropriate department, and to retrieve and rank relevant similar answers and frequently asked questions (FAQs) to support human operators. These models directly influence how taxpayer requests are processed and therefore constitute a critical interface between automation and administrative discretion.

1.1 Contributions

In this work, we propose an integrated auditing approach for AI systems in public services, illustrated through the Smart e-Desk use case. Our main contributions are:

- A multidimensional evaluation framework addressing bias and fairness, robustness, privacy, transparency, and human agency, combining quantitative model tests with qualitative analysis.
- Implementation of targeted auditing experiments, including counterfactual gender substitution for bias detection, perturbation-based robustness testing, anonymisation-pattern analysis for privacy risk, keyphrase-substitution experiments for explainability and pilot design guidelines for human oversight and agency.
- A cross-disciplinary reflection linking empirical findings to legal requirements under the AI Act, emphasizing transparency, reliability, and human oversight.

2 Background and Related Work

2.1 AI auditing and governance

Early work from Sandvig et al. (2014) on algorithm auditing and governance frames audits as empirical tools for detecting discrimination in socio-technical systems, adding to previous work on ethical technology assessment by Palm and Hansson (2006). Later, Brown, Davidovic, and Hasan (2021) emphasized that audits must also account for the surrounding decision-making context. In parallel, Raji et al. (2020) proposed an end-to-end internal auditing framework combining process documentation with technical tests to close the accountability gap between AI development and deployment. Risk-based governance frameworks such as the NIST AI Risk Management Framework (RMF) 1.0 (NIST 2023) and ISO/IEC 42001:2023 (ISO 2023) further operationalize auditing principles, while the OECD AI Principles (OECD 2019) and the UNESCO Recommendations on AI Ethics (UNESCO 2021) provide high-level dimensions of Trustworthy AI.

More recently, the Human Rights, Democracy and Rule of Law Impact Assessment for AI Systems (HUDERIA) methodology was adopted by the Council of Europe (2024), as a comprehensive framework for identifying and mitigating AI-related risks to fundamental rights and democratic governance. It is particularly relevant for public-sector applications, complementing existing technical and risk-management frameworks with a rights-based, legally grounded assessment perspective.

2.2 Regulatory landscape

The AI Act, the world's first comprehensive horizontal legal framework for AI regulating, applies across sectors, covering the entire AI value chain, from model and AI systems providers, deployers, distributors, and importers to actors outside the EU when their AI systems are placed on, or their outputs are used in the European market (European Union 2024). The AI Act establishes obligations for "high-risk" AI systems, including risk management, data governance, transparency, and human oversight, as well as transparency duties for AI systems that directly interact with individuals, irrespective of their risk classification. These obligations complement existing requirements in data protection, privacy, such as those stemming from the General Data Protection Regulation (GDPR) (European Union 2016), cybersecurity, product safety, as well as sector-specific regulations.

In the public sector, there are additional legal obligations that must be considered when assessing deployments of AI Systems. In this regard, the principles of lawfulness, transparency and duty to explanation, proportionality, and accountability, are at the core of Administrative Law.

The Ethics Guidelines for Trustworthy AI of the EU High-Level Expert Group (High-Level Expert Group on Artificial Intelligence 2019) summarizes seven requirements: human agency, technical robustness, privacy and data governance, transparency, fairness, societal well-being, and accountability, which remain widely used as auditing guidelines.

Building upon these frameworks, MARTA operationalizes their principles through an empirical audit of a citizen-service NLP system, explicitly aligning technical tests with a rights-based assessment model.

2.3 Auditing Dimensions

Bias and Fairness The AI Act places strong emphasis on data quality as a prerequisite for responsible AI, particularly for high-risk systems. Providers must ensure that training, validation, and testing datasets are relevant, representative, free of known errors, and as complete as possible, thereby reducing the risk of discriminatory or misleading outputs (Article 10). Deployers, on the other side, should also ensure the quality of the input data during use (Article 26(4))

Surveys synthesize statistical and causal fairness definitions from global approaches such at group-level analysis to individual approaches such as counterfactual fairness (Mehrabi et al. 2021). Group fairness criteria encompass metrics such as demographic parity (Feldman et al. 2015), equality of opportunity (Hardt, Price, and Srebro 2016), and predictive parity (Chouldechova 2017), chosen based on contextual relevance and application intent to minimize potential harms. Counterfactual fairness (Kusner et al. 2017) frames bias testing via "what-if" substitutions of protected attributes, like sex or race.

Robustness High-risk AI systems must be designed and developed to achieve appropriate levels of robustness, resilience, and accuracy under foreseeable operating conditions (Article 15). Providers must implement and document technical measures that ensure system stability, reliability, and resistance to errors or adversarial manipulation. Deployers are responsible for operating the system in accordance with its intended purpose, including performing checks, monitoring accuracy, and reporting performance issues (Article 26(1) and (5)).

Text classification and retrieval models are notoriously brittle under small perturbations. TextAttack (Morris et al. 2020) employs adversarial transformations (typos, word swaps, sinonyms) and standardized evaluation recipes, that we adopt to quantify stability and performance degradation.

Transparency and Explainability Although the AI Act does not require explainability, it introduces transparency duties for providers and deployers. Providers of high-risk systems must make key information about system capabilities, limitations, and proper use available to deployers through technical documentation and user instructions (Article 13). In systems that interact directly with individuals, they must implement by design measures to ensure that the people are aware they are interacting with an AI system (Article 50(1)). Depending on the use case, deployers must ensure that individuals affected by an AI-supported decision, including operators of such systems, are informed about the use, logic, and possible impacts to their rights and freedoms (Articles 26(11) and 86).

Documentation artifacts like Datasheets for Datasets (Gebru et al. 2018) and Model Cards (Mitchell et al. 2019) promote traceability and contextual transparency of model behavior, core elements of modern audits. For local explanations, Jacovi and Goldberg (Jacovi and Goldberg 2020) emphasize the need for faithfulness and warn against plausible-but-misleading attributions, motivating our keyphrase-perturbation tests rather than reliance on keyphrases alone.

Privacy and Anonymization

In addition to the obligations related to data quality of the AI Act, the General Data Protection Regulation (GDPR), continues to apply where personal data is processed in the AI system lifecycle. Under the GDPR, controllers and processors must ensure lawfulness, fairness, transparency, data minimization, accuracy, storage limitation, and integrity and confidentiality. Both the AI Act and the GDPR emphasize the importance of privacy-enhancing design choices (such as use of synthetic data) to promote anonymization or effective pseudonymization, especially during training and testing, to reduce privacy risks while preserving utility.

Automated de-identification of unstructured text is well-studied in clinical NLP, from early rule-based systems to neural sequence models with strong F1 performance (Neamatullah et al. 2008; Dernoncourt et al. 2017). In the EU legal context, the GDPR's definition of pseudonymization (Art. 4(5)) clarifies that such data remain personal and must be protected accordingly, shaping how anonymization effectiveness and residual risk should be evaluated (European Union 2016).

Human Oversight and Impact

In high-stakes administrative settings, "human-in-the-loop" design must reflect organizational and procedural constraints. Studies of public-sector machine learning highlight gaps between research prototypes and operational practice, particularly regarding accountability and usability of transparency tools (Redding and O'Neil 2023). The simulation of real-world conditions in AI testing experiments involving end-users enables the detection of potential behavioral biases in operational practice, as well as resulting challenges (OECD 2025).

3 Smart e-Desk System Description

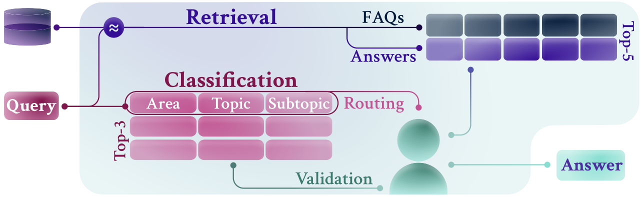

Figure 1: Smart e-Desk framework composed of a classification and a retrieval module and overseen by a specialized human operator.

The **Smart e-Desk** service leverages AI models to assist human operators in responding to taxpayer queries, with the goal of improving both the efficiency and quality of interactions. The collected data (September 2022 - September 2025) comprises approximately two million interactions between taxpayers and operators. Each query follows a three-tier hierarchical classification system, specifying the domain area, topic, and subtopic. The taxonomy includes 20 areas that branch out to 653 total labels.

Query classification was initially provided by taxpayers, which occasionally resulted in incorrect or inconsistent labeling. Misclassified issues were subsequently transferred across teams until resolution, sometimes by teams outside the appropriate domain. Another frequent source of noise arises when taxpayers pose multi-topic or cross-domain questions, which inherently leads to incomplete or ambiguous labeling.

To mitigate these issues, the dataset underwent extensive preprocessing to remove duplicates, near-duplicates, team-transfer cases, and statistical outliers. After filtering, the gold-standard dataset comprised 616,000 unique samples. Data were stratified by flattened labels and split into training and testing subsets following an 80/20 ratio.

The Smart e-Desk framework encompasses two primary AI-based components ran in parallel:

1. **Topic Routing through Query Classification** : screening the questions to the right team is an essential step to ensure the issues are properly processed. Incorrect assignment leads to infinite loops of forward issues between teams, which delays the response, affecting the taxpayer, and inputs additional burden to the different members of the teams that need to review the query.
2. **Retrieval-based Suggestion** : provides operators with similar historical cases and their respective resolutions to assist in decision-making. Incorrect or unhelpful suggestions may hinder the operator's resolution, by adding review times overheads and cluttering their screen with irrelevant information.

Each step of the workflow is supervised by a human operator, who is presented with multiple top-k suggestions: 3 for module (1) and 5 for (2). Operators may choose to follow, refine, or disregard these suggestions. This human-in-the-loop design enhances trustworthiness, accountability, and error mitigation. In accordance with the criteria of the EU AI Act, the system can be characterized as minimal-risk, given its objective and limited autonomy. Nonetheless, this classification does not preclude the adoption of best practices for bias and fairness, privacy, transparency, and human oversight.

The framework already incorporates a **keyphrase identification task** to provide contextual explanations for model predictions, enhancing operator understanding and trust. The system integrates an **anonymization module** that automatically redacts personally identifiable information prior to model processing or storage. The data are stored securely in the provider's cloud environment, accessible only through authenticated user credentials and a protected VPN connection. All information is encrypted both at rest and in transit within the cloud infrastructure. For model development and analysis, however, data are temporarily downloaded and processed locally. Although local storage remains protected, this step introduces a residual risk of data exposure, as the data are no longer shielded by the cloud's encryption and access controls.

In the present work, we focus primarily on auditing the query classification task, evaluating it in terms of robustness and bias. We further investigate their explainability and anonymization mechanisms within adversarial scenarios.

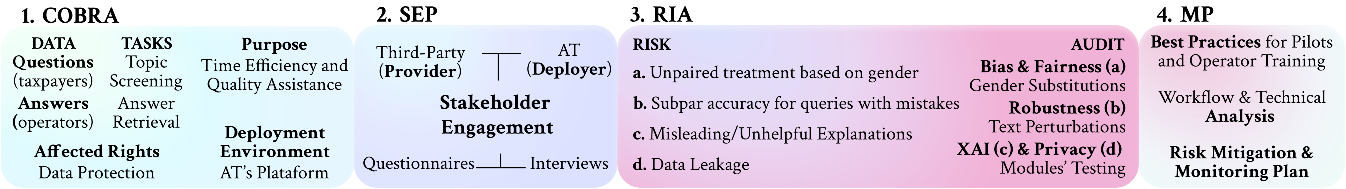

Figure 2: MARTA's integration of the HUDERIA framework based on the Smart e-Desk Use Case.

3.1 Socio-technical context

In the AI Act, human oversight is treated as both a design requirement and an organizational obligation. Providers must ensure by design that AI systems are technically capable of being effectively overseen by natural persons, including through appropriate interfaces, alerts, and control functions, that can support the human oversight obligations of the deployers (Article 14). As such, deployers should ensure operationally that adequate procedures are established and trained personnel is identified to monitor the system's functions, logs, and outputs, to intervene when necessary, and report any relevant incident (Article 26(2), (5) and (6)).

The Smart e-Desk system is operationalized by a diverse team with different technological proficiency, literacy and trust. This diversity urges for correct, thoughtful and complete processes to ensure the operators are not over or under reliant on the technology, and that they can navigate the new add-ons with ease, improving their workflow. The pilot phase involving the system's operatives and training sessions is just planned but concerns regarding excessive user expectations and mixed reactions to the system have been noted. User-involvement throughout AI development can mitigate and help address these challenges (OECD 2025).

4 MARTA Framework

4.1 Alignment with the HUDERIA Framework

The EU AI Act and the Council of Europe (2024)'s Framework Convention on Artificial Intelligence have emphasized the need to assess AI systems not only for technical performance but also for their implications for human rights, democracy, and the rule of law.

The HUDERIA framework, adopted by the Council of Europe, provides a structured methodology to evaluate such impacts across the full AI lifecycle (Council of Europe 2024). It was designed to guide both public authorities and private developers in identifying and mitigating risks before deployment, bridging the gap between technical audits and legal or ethical assessments. HUDERIA comprises four core phases, presented in Table 1 together with their mapping to our framework's auditing activities (see Figure 2).

MARTA integrates HUDERIA's objectives combining technical auditing, legal compliance analysis, and human-centred evaluation. While HUDERIA emphasizes procedural completeness and participatory governance, MARTA extends the framework through quantitative model testing and alignment with AI Act compliance documentation. Conversely, certain HUDERIA elements, such as deep stakeholder participation or continuous post-deployment monitoring, remain limited by our semi-external position and time-bound access to the system.

The adaptation of HUDERIA to MARTA's auditing workflow demonstrates how the former can be implemented in practice to produce evidence-based assessments that are both technically rigorous and legally grounded, contributing to the broader European agenda for trustworthy and accountable AI in administration.

4.2 Bias and Fairness

For the bias module, the model was tested to identify biased outputs based on the gender of the taxpayer, when explicit in the query.

Counterfactual Generation. In Portuguese, grammatical gender extends to inanimate objects. This linguistic property poses a challenge when generating gender-based counterfactuals (CF). Consequently, we found that a purely rule-based approach was insufficient to accurately flip the subject gender while maintaining lexical and grammatical agreement. To address this, we employed Qwen3-8B (Team 2025), instructing the model to generate two counterfactual variants for each input query by converting the identified animate subjects into their female and male forms. The prompting strategy directed the model to modify only animate entities, and preserve as much of the original query as possible. Queries lacking gendered entities were not changed. In cases involving mixed-gender references, the queries were modified to include only one consistent gender representation. For example:

Original: I'm a single ***mom***, is my ***son***...
Male CF: I'm a single ***dad***, is my ***son***...
Female CF: I'm a single ***mom***, is my ***daughter***...

Restricting CF generation to queries with a single identifiable subject gender enabled a more controlled basis for comparison between samples.

Following counterfactual generation, we applied a masked language model, BERTimbau Base (Souza, Nogueira, and Lotufo 2020), to restore grammatical agreement. The model was queried sequentially for each token, apart from changed tokens in the generation, using the original sentence as context for every prediction. This non-autoregressive setup prevented cascading edits and ensured consistency across predictions. Only high-confidence replacements (e.g., determiners) were accepted, effectively correcting local gender agreement without altering the semantics of the generated counterfactuals.

Representativeness of gendered samples. Distribution of male and female samples in the dataset. Cosine similarity

HUDERIA Phase	Purpose and Main Activities	MARTA Correspondence and Adaptations
Context-Based Risk Analysis (COBRA)	Define the system context, intended purpose, affected rights, and deployment environment.	Scoping of the Smart e-Desk models (classification and retrieval) within AT; identification of potential risks for fairness, privacy, and transparency; mapping to AI Act and GDPR obligations.
Stakeholder Engagement Process (SEP)	Identify and involve relevant actors (providers, deployers, users, and oversight bodies).	Structured engagement with AT and the third-party development team through questionnaires and interviews; no direct contact with end-user operators.
Risk and Impact Assessment (RIA)	Evaluate likelihood and severity of harms, combining qualitative and quantitative evidence.	Empirical bias detection (gender substitution), robustness tests (textual perturbations and uncertainty analysis), privacy leakage analysis, and explainability experiments. Complemented by legal expert review of AI Act and data-protection compliance.
Mitigation Plan (MP)	Define corrective measures, monitoring procedures, and accountability mechanisms.	Development of recommendations for pilots and operator training; proposals for oversight workflows, documentation practices, and continuous monitoring by AT and vendors.

Table 1: Mapping of HUDERIA phases (Council of Europe 2024) to MARTA's auditing activities.

between embeddings was used to classify each sample as originally male, female, or undefined, providing a quantitative measure of the dataset's representativeness.

Concept drift in the embedding space. The semantic stability of the classification model was studied by measuring the drift between the embeddings of original and counterfactual queries.

Group Fairness metrics. The practical utility was evaluated by applying them to our group fairness pipeline, measuring the deviation from ideal group fairness conditions. Specifically, we considered Equality of Opportunity (EO), introduced by Hardt, Price, and Srebro (2016) and Predictive Parity (PP), formalized by Chouldechova (2017). EO reflects parity of error rates (FPR and FNR) across gender groups, while PP captures parity in precision. Smaller absolute values of these metrics indicate lower bias.

4.3 Robustness

For the robustness module, we aimed to evaluate two real-world scenarios. To achieve this we conducted experiments that simulate noisy and unseen or uncommon data through **Input perturbations** (e.g. rephrasing, typos, and word swaps) to assess model stability and **Confidence calibration and out-of-distribution (OOD) detection**, measuring whether the model remains confident or uncertain when appropriate, reflecting that real-world deployments often encounter inputs outside the training distribution.

Dataset Attack We apply TextAttack transformations to simulate realistic input perturbations, including character-level typos (substitution, deletion, and insertion) and word-level paraphrases (synonym substitution and neighboring word swaps). These transformations are designed to reflect plausible real-world input variations (Martins and Silva 2004). We restrict each word to at most one modification, exclude stopwords from modifications, and apply character-level perturbations only to words with a minimum length of four characters, following the typographical error patterns reported in Flor et al. (2015), which offered useful guidance despite addressing a different research objective. A maximum of 10% of words per input were allowed to swap.

Domain Shift Analysis To assess robustness under uncertain inputs, we evaluate predictive uncertainty alongside threshold-independent and threshold-dependent metrics.

$$H[p(y|x)] = -\sum_{y\in\mathcal{Y}} p(y|x)\log_2 p(y|x) \tag{1}$$

We opted for the entropy of the predictive posterior distribution, modeled by Shannon entropy (1), a standard measure for single probability distributions.

For the threshold-independent, we adopt the Area Under the Confidence–Oracle (AUCO) error (Scalia et al. 2020), which quantifies the area between the theoretically optimal ordering (oracle) and the ordering induced by each uncertainty measure. Lower AUCO values indicate that the uncertainty estimation is closer to the oracle curve, and therefore a better predictor of uncertainty. For the special case of OOD datasets, we used the Area Under the Receiver Operating Characteristic (AUROC) metric, which represents the probability that a negative (OOD) example is assigned a lower confidence score than a positive (in-distribution) example.

For threshold-dependent, we followed studies that assess uncertainty estimation performance using a binary confusion matrix framework (Asgharnezhad et al. 2022; Tabarisaadi et al. 2022). In this context, predictions are categorized as correct or incorrect, and based on a confidence threshold, as certain or uncertain. Full definitions and formulas for AUCO and the binary confusion matrix framework are provided in Appendix B.

4.4 Transparency and Explainability

To examine whether the identified keyphrases can serve as faithful explanations for model predictions, we evaluate explanation faithfulness using Comprehensiveness and Sufficiency. We perturb the classifier's input in two complementary ways (DeYoung et al. 2020):

- Sufficiency: we replace all tokens not belonging to the keyphrases with the `<MASK>` token, assessing whether the remaining keyphrases alone are sufficient for the model to preserve its original prediction (i.e., maintain a similar confidence level).
- Comprehensiveness: we replace all tokens within the keyphrases with `<MASK>`, evaluating whether their removal substantially decreases the model's confidence, indicating that the keyphrases contain the information most relevant to the prediction.

4.5 Privacy and Anonymization

The system implements an anonymization procedure based on LLM instructions and supported by curated whitelists. In this audit, we first test the robustness of the anonymization layer by perturbing queries and verifying whether previously anonymized segments remain censored.

To detect potential leakage of sensitive information, we derive the *comprehensiveness* score to the anonymized tokens. Each sensitive token is replaced with `<MASK>`, and the model's confidence in its original prediction is measured. A substantial drop in confidence indicates that the model was relying on that token to make predictions, suggesting that the anonymized token still carries residual predictive information.

4.6 Human Oversight and Impact

A report providing recommendations tailored for the planning and conduction of the system's pilot was developed based on researchers' experiences on similar studies and established best practices. It was proposed that the pilot phase should be divided into two stages. As this will be the first interaction users have with the system, conducting moderated usability tests will help identify potential challenges that need to be addressed before proceeding to a longitudinal pilot that requires independent use. By integrating the system into users' workflows, the longitudinal pilots are expected to reveal challenges that may not appear in controlled settings.

For the study, we proposed a set of goals for the evaluation of the (i) system's usability, (ii) users' expectations alignment with the system's capabilities, (iii) users' trust in the system's outputs and (iv) system's impact and support on users' workflow. Beyond proposing the adoption of common usability metrics and instruments (such as the System Usability Scale (Brooke 2013)), additional research methods were suggested targeting the proposed goals, namely:

- Pre- and post-test questionnaires targeting performance expectations and after-use perceptions of changes in request resolution time, efficiency and utility of suggestions, and expected system use frequency, respectively;
- Evaluation of human operators' submitted responses;
- Interviews with team coordinators at the end of the longitudinal pilot to understand the system's broader organizational impact.

Moreover, the report provided general procedural recommendations, taking into account power relations inherent to labor contexts that may impact the pilot's results.

5 Results

5.1 Bias and Fairness

Using the similarity of the generated counterfactuals to the original samples, the dataset is composed of approximately 2.27% male samples, 0.52% female samples, and 15.78% mixed-gender samples. This skewed distribution is expected, as the male form in Portuguese functions as the neutral form when the gender is unknown.

Concept drift between the male and female counterfactuals was negligible (0.007) indicating that the model's internal representations are largely invariant to gender, which suggests the classifier is not substantially swayed by the subject's gender in the queries.

The utility tests also didn't reveal any meaningful bias; selecting the maximum noted imparity per class, the EO is $0.51p.p.$ and PP is $0.09p.p.$. When evaluating the results by area, the disparity increases, but it remains within an acceptable range, $EO = 5.50p.p.$ and $PP = 8.27p.p.$, both in favor of male subjects.

5.2 Robustness

Dataset Attack Table 2 summarizes the dataset attack results obtained across five different seeds (0, 1, 2, 3, and 5).

	Three-tier Classification (%)	Area (%)
Successful attacks	10.19 ± 0.04	1.64 ± 0.02
Original acc.	72.45 ± 0.02	97.34 ± 0.01
Acc. under attack	70.79 ± 0.06	96.79 ± 0.03
Avg. pert. words	9.00 ± 0.00	

Table 2: Dataset-level attack results showing the number of successful attacks, the original model accuracy, and the accuracy under attack. The average input length is 59 words.

On average, 9% of the input tokens were perturbed, resulting in an accuracy drop of $1.6p.p.$. This value is higher than desired, considering that the attacks represent plausible real-world variations.

We observed that the rate of successful attacks generally increased with the number of transformations applied to a single input, ranging from 8.2% to 13.8% (see Appendix B). Among transformation types, synonym swaps produced the highest success rate (12.0%), likely because WordNet synonym replacements may not preserve context, whereas word swaps produced the lowest rate (10.0%), possibly reflecting the frequent availability of paraphrases in Portuguese.

When the evaluation is restricted to the "Area" level, the model's accuracy is substantially higher, and the impact of perturbations is reduced to only $0.5p.p.$, indicating strong robustness in this scenario.

Domain Shift Analysis Visualizing uncertainty distributions for correctly and incorrectly classified samples provides insights into the robustness of the uncertainty estimation. Ideally, an increase in uncertainty should correspond to a decrease in classification performance, while

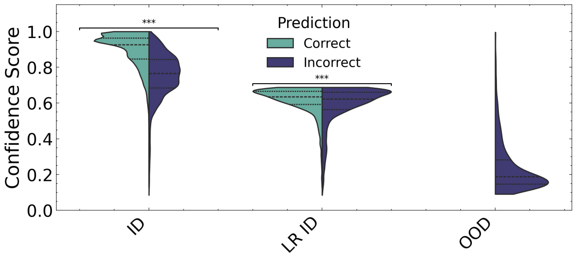

Figure 3: Confidence value distributions for correctly and incorrectly classified samples across different data sets: ID - Test set; LR-ID - Samples whose confidence scores fall at or below the 10th percentile; OOD - ASSIN2 test set.

for OOD samples, uncertainty should be maximized, reflecting the model's lack of confidence in unknown inputs. Figure 3 illustrates these confidence distributions, comparing in-distribution (ID) test set to Less Representative In-Distribution (LR-ID) and OOD samples. Here, the OOD subset corresponds to premise texts from the ASSIN2 test set (Real, Fonseca, and Oliveira 2020). The LR-ID subset is composed of samples whose confidence scores fall at or below the 10th percentile of all ID scores.

Regarding the threshold-independent measures, the AUCO for the ID set was 0.058, indicating that the model effectively quantifies its own confidence, as it is close to the perfect ordering defined by the true error. For the OOD dataset, the AUROC reached 98.97%, demonstrating performance close to that of an ideal detector.

To evaluate the practical application of uncertainty estimation, we selected the confidence threshold to maximize the F1-Score, allowing the classifier to abstain from making predictions in cases of high uncertainty. Rejecting 16% of the inputs yields a $7.5p.p.$ improvement in accuracy (from 72.5% to 80.0%), showing that the framework can benefit from selective prediction. Additional results are provided in Appendix B.

5.3 Transparency and Explainability

Since our keyphrases are binary (either a token belongs to a keyphrase or does not), masking inputs with a low number of keyphrases can create extreme cases. To address this, we also evaluate a random masking baseline, where non-keyphrase tokens (excluding special tokens) are masked randomly, with the same number of tokens as the keyphrases (or the maximum possible if the number of keyphrases exceeds the number of non-keyphrase tokens).

When keyphrases are removed, the average drop in prediction probability is $56.5p.p.$ (Comprehensiveness), indicating that keyphrases are important for the model's predictions. For Sufficiency, the average drop is $13.6p.p.$, where lower is better. Although this drop is smaller than for Comprehensiveness, it still exceeds $10p.p.$, suggesting that the keyphrases capture relevant features but lack full explanatory sufficiency, and that local explanations may not fully reflect the model's reasoning.

The random non-keyphrase masking baseline further supports this conclusion: the score is significantly lower than Comprehensiveness, at $1.8p.p.$, corroborating the importance of the keyphrases.

5.4 Privacy

Anonymizer Attack When running perturbed sentences through the anonymizer, it was verified that only 17.5% of the sensitive information segments were still successfully censored (exact match). This value shows the anonymizer is susceptible to bypassing sensitive information when small perturbations occur, such as plausible typos made by the taxpayer. By averaging the model confidence for the sensitive tokens, we don't note a significant drop from the original samples when the model correctly identifies the tokens ($0.1p.p.$ drop from 85.6% original token-level average confidence). However, the model is more confident in its results for the remainder of the sentence with an average drop of $5.7p.p.$ from 94.0%, which suggests the model is contextually less certain.

Comprehensiveness Analysis The model does not seem to rely heavily on sensitive information since its masking does not lead to a large drop in confidence (Comprehensiveness of $2.2p.p.$). The baseline score is slightly higher, $4.9p.p.$, as randomly masking tokens not marked as sensitive can occasionally remove relevant tokens. This suggests that the risk of overfitting to sensitive attributes is reduced.

6 Discussion and Recommendations

6.1 Legal and Ethical Relevance

The AI Act, together with the broader legal framework, establishes both technical and organizational requirements for responsible AI governance across the entire AI lifecycle. To operationalize these requirements, the AI Act adopts a pragmatic approach, promoting accountability along the AI value chain. Accordingly, assessing conformity with legal frameworks and ethical standards requires analyzing both technical controls and organizational measures. In the Smart e-Desk System case, the technical audit present several metrics with clear legal and ethical relevance:

Bias metrics (CF, EO, PP, drift): informative for data governance compliance under the AI Act and GDPR, relevant to fairness duties under Administrative Law, and directly connected to the ethical principle of non-discrimination.

Robustness: closely aligned with the AI Act's robustness, accuracy, and cybersecurity obligations, and consistent with core principles of trustworthy AI.

Uncertainty and abstention: essential for operationalising **meaningful human oversight**, supporting both legal obligations (AI Act Art. 14) and ethical expectations regarding human-centred decision-making.

Comprehensiveness and sufficiency scores: useful for assessing the boundaries of explainability, thereby supporting compliance with the transparency duties of the AI Act and the justification requirements of Administrative Law.

Anonymisation accuracy: directly relevant for GDPR principles of data protection by design, data minimisation, and

Nature	Provider	Deployer
Bias & Fairness	**(S)** Counterfactual augmentation as mitigation measure.	**(G)** Monitor demographic proxies (GDPR-compliant) for fairness drift; **(S)** Periodic linguistic audits to identify representational and allocation harms.
Robustness	**(G)** Provide robustness documentation; **(S)** Perturbation-based data augmentation for training.	**(G)** Adversarial testing integrated into lifecycle risk management.
Human Oversight & Impact	**(S)** Confidence-based AI-aid reject option.	**(G)** Operator training and intervention traceability; **(G)** Periodic review of abstention thresholds; **(S)** Monitor the operators satisfaction.
Transparency & Explainability	**(G)** Include model-internal explanation methods; **(G)** Contextualization warnings in documentation.	**(G)** Keep updated online policies (TC, privacy).
Privacy & Anonymization	**(G)** Strict data minimization in the AI and data lifecycle; **(S)** Finetune pseudonymization with adversarial attacks.	**(G)** Regular leakage audits and monitoring.

Table 3: Summary of recommendations from the Smart e-Desk AI audit, indicating whether each measure is a general best practice (G) or specific (S) to the Smart e-Desk context, and the intended stakeholder (provider or deployer).

data security.

6.2 e-Desk System Analysis

Bias and Fairness Although the analyses reveal negligible gender drift and balanced performance, ethical and regulatory frameworks require continuous monitoring of bias, especially in public-sector environments.

Robustness The model demonstrates high robustness at the hierarchical "Area" level. Nonetheless, given its public-sector context, the inherent sensitivity of fiscal information, additional robustness, resilience, and cybersecurity safeguards are advisable.

Human oversight and Impact The results show a strong ability to identify low-confidence or OOD cases. To complement these results, meaningful human oversight can be further operationalized.

Transparency and Explainability The keyphrase-based explanations module evaluation indicates that explanations capture meaningful features but not the full decision rationale. In this regard, additional measures to promote transparency can be implemented, considering that in the AI Act all systems interacting directly with individuals are subject to transparency requirements (Art. 50), also reinforced by the Administrative Law requirements.

Privacy The analysis raises warning signs in relation to privacy-preserving by design methodologies. Considering the system's deployment and its inherent sensitivity, additional measures are advised in the AI and data lifecycle.

6.3 Mitigation Plan

Table 3 presents a consolidated set of recommendations, containing general best practices (G) and others specific to the Smart e-Desk context (S), along with associated monitoring and mitigation measures (see the extended list in Appendix D). These recommendations align with HUDERIA's RIA and MP phases, making audit findings actionable.

7 Conclusion

We introduce MARTA, a multidimensional auditing framework for evaluating AI systems in public administration that integrates technical, legal, and organizational analysis.

Following the HUDERIA methodology, this work maps each step to practical methodologies and strategies across the central audit pillars: fairness, robustness, transparency, privacy, and human oversight. By identifying which technical tests carry legal significance, MARTA contributes to the emergence of national AI governance mechanisms that bridge the gap between legal frameworks and operational system evaluation.

Based on the public sector use case, the Portuguese Tax Authority's Smart e-Desk, the framework employs fairness metrics, robustness evaluations, explanation faithfulness tests, and anonymization analyses that provide empirical evidence directly aligned with HUDERIA's normative commitments to non-discrimination, transparency, privacy, and accountable administration. As a result, we demonstrated how MARTA can faithfully uncover both strengths, such as negligible bias, low susceptibility to linguistic perturbations, and well-calibrated uncertainty; and limitations, including less resilient anonymization, partial explanation sufficiency, and oversight challenges.

Future work will extend the audit to the system's retrieval model, followed by the development of continuous monitoring procedures and the application of MARTA to additional administrative AI systems.

Ethical Statement.

This work audits an AI system used in public administration with the goal of strengthening transparency, fairness, and accountability. All data were provided by the Portuguese Tax Authority and processed under a formal data-sharing agreement. The findings and recommendations aim to reduce risks of bias, privacy breaches, and over-reliance on automated suggestions, thereby supporting more equitable and trustworthy public services. At the same time, we recognize that auditing itself may surface limitations that, if unaddressed, could affect citizens' rights or institutional decision-making. Our intent is to promote responsible development and deployment of AI systems in ways that safeguard fundamental rights and democratic values.

Acknowledgments.

The authors acknowledge the Portuguese Tax Authority, Autoridade Tributária e Aduaneira (AT), for their essential collaboration, including data provision and coordination with the development team. The authors also thank AXIANS, responsible for the backbone of Smart e-Desk System, for sharing their models for auditing. Both organizations demonstrate a clear commitment to system improvement, transparency, and the responsible use of AI in public administration, in line with data protection and emerging AI governance frameworks.
This work was supported by Fundação para a Ciência e Tecnologia (FCT) under the project "Machine Learning Auditing for Robust and Transparent Administration (MARTA)"[2024.07646.IACDC/2024], funded through the Inteligência Artificial, Ciência dos Dados e Cibersegurança de relevância na Administração Pública (Artificial Intelligence, Data Science and Cybersecurity of relevance to Public Administration) call.

A Bias and Fairness

Uncertainty Analysis

Table 4 compresses the general confidence for original and counterfactual samples. Results point to negligible variation between groups, meaning the gender is not solely disrupting the models confidence in its queries. The variation lightly increases when comparing confidence in the Area classification tier, where the mean an median point for incorrect samples is lower for male samples, however, the difference (0.009 and 0.008) is still minimal and can be attributed to noise.

B Robustness

Dataset Attack: Additional Results

Table 5 shows the impact of the number of different transformations (NDT) on the rate of prediction change. In general, increasing the number of transformations applied leads to a higher change rate. An exception occurs for 7 transformations, which only correspond to 16 samples and also have the highest average number of words per input, potentially affecting the results.

When focusing solely on the Area-level classification, the number of transformations applied has little effect on the change rate, suggesting that the model is relatively robust at this coarser level.

Table 6 shows that all perturbation types result in substantially higher prediction change rates in the three-tier classification (TTC) than in the Area-level, with TTC shifts ranging from 10.0% to 12.0%, while Area changes remain between 1.1% and 1.7%. Synonym Swap produces the largest TTC change, likely due to its semantic impact, whereas character-level operations (deletion, insertion, substitution, QWERTY errors) also introduce sensitivity, though their effects are relatively similar to one another. In contrast, Area predictions remain remarkably stable across all transformations. We also examined whether differences between transformations could be attributed to the number of different transformations applied per input, given that this factor previously showed an effect on the change rate, but the average number was similar across transformations, ruling it out as an explanatory factor.

	Mean	Std. Dev.	Median	
		All Labels		
Corr	0.894	0.096	0.926	Og
	0.894	0.097	0.925	Male
	0.893	0.097	0.925	Female
Inc	0.752	0.131	0.765	Og
	0.751	0.134	0.765	Male
	0.753	0.134	0.767	Female
		Area		
Corr	0.861	0.115	0.890	Og
	0.860	0.116	0.889	Male
	0.860	0.116	0.889	Female
Inc	0.622	0.196	0.658	Og
	0.611	0.203	**0.646**	Male
	0.620	0.202	**0.654**	Female

Table 4: Classification model score confidence, measured with entropy, for original and counterfactual samples(male and female).

Domain Shift Analysis

Framework The AUCO error quantifies the area between the theoretically optimal ordering (oracle) and the ordering induced by each uncertainty measure and is defined as:

$$AUCO = \int_0^1 (conf_r^u - conf_r^o)\, dr \tag{2}$$

where $conf^u$ and $conf^o$ denote the confidence curves for the given uncertainty estimation and the oracle, respectively, and r is the fraction of rejected samples.

The binary confusion matrix framework characterizes the relationship between correctness and certainty by assigning each prediction to one of four categories: (i) True Certainty (TC): correct and certain; (ii) True Uncertainty (TU): incorrect and uncertain; (iii) False Uncertainty (FU): correct and uncertain; and (iv) False Certainty (FC): incorrect and certain. From these, we compute Uncertainty Accuracy (UAcc), Uncertainty Sensitivity (USens), Uncertainty Specificity (USpec), and Uncertainty Precision (UPrec).

$$UAcc = \frac{TU + TC}{TU + TC + FU + FC} \tag{3}$$

NDT	1	2	3	4	5	6	7
Count	22,966	38,997	33,639	19,118	7,356	1,304	16
NW	25	42	63	92	133	176	200
A (%)	1.4	1.3	1.4	1.4	1.6	1.4	0.0
TTC (%)	8.2	9.7	10.6	12.1	12.6	13.8	5.3

Table 5: Rate of prediction change as a function of the number of different transformations (NDT) for three-tier classification (TTC) and Area-level classification (A). Count corresponds to the number of samples, and NW corresponds to the average number of words per input.

Transformation	TTC(%)	A (%)	NDT	Count
Synonym Swap	12.0	1.7	4	33,334
Char. Insertion	11.5	1.1	4	627
Adj. Char. Swap	11.2	1.4	3	47,226
Char. Subst.	11.0	1.4	3	60,539
Character Deletion	10.8	1.4	3	54,947
QWERTY Subst.	10.5	1.4	3	111,911
Adj. Word Swap	10.0	1.4	4	14,481

Table 6: Impact of perturbation types on the percentage of prediction changes for three-tier classification (TTC) and Area-level classification, along with the average number of different transformations per input (NDT) and the total number of perturbed samples (Count).

$$USen = \frac{TU}{TU + FC} \tag{4}$$

$$USpec = \frac{TC}{TC + FU} \tag{5}$$

$$UPrec = \frac{TU}{TU + FU} \tag{6}$$

Results We applied the non-parametric Mann-Whitney U test (McKnight and Najab 2010) for unpaired groups. Both the ID and LR-ID sets showed statistically significant differences at the 0.05 significance level. However, when assessing practical significance using Cliff's Delta effect size (Sawilowsky 2009), the ID set exhibited a large effect ($d = 0.656$), while the LR-ID set showed only a small effect ($d = 0.123$).

Regarding threshold-dependent evaluation, Figure 4 depicts the predictive uncertainty evaluation metrics as the uncertainty threshold varies. The annotated point on each curve indicates the selected confidence threshold based on the best F1-Score. Table 7 summarizes the results for each metric.

UAcc	USen	USpec	UPrec
78.0	39.1	92.8	67.3

Table 7: Performance evaluation metrics (%) for classification.

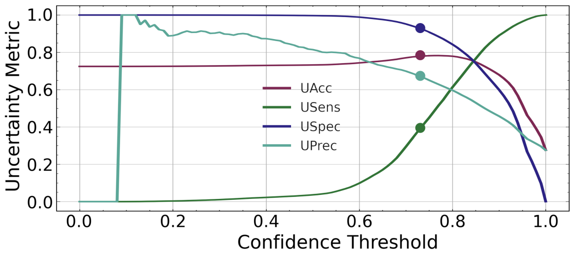

Figure 4: Uncertainty performance measures for varying threshold values. The chosen threshold for each method is denoted by a data point superimposed on each line plot.

Multi-intent Inputs

Framework We iteratively generate multi-intent inputs by randomly sampling k distinct classes from the remaining test-set labels and concatenating one randomly selected sample from each class. After constructing each multi-intent query, the selected samples are removed from the pool. This process continues until fewer than k unique classes remain.

To evaluate the model's performance on these multi-intent prompts, we use Mean Average Precision at K (MAP@K). Each prompt is constructed to contain exactly K ground truth intents, and the model returns a ranked list of its top-K predicted labels. For each input, we compute Average Precision at K (AP@K) by considering the predicted labels in order and checking whether each prediction is among the true intents. The MAP@K is then obtained by averaging the AP@K scores across all inputs.

Results As shown in Table 8, when combining two categories per input ($K = 2$), the model shows difficulty in retrieving both ground truth labels among the top two predictions ($MAP@2 = 39.2\%$), although it often retrieves at least one ($MAP@1 = 66.0\%$). As expected, increasing the number of intents to K=3 further decreases performance ($MAP@3 = 27.8\%$), highlighting the added difficulty of ranking multiple target labels within a single mixed prompt.

Performance improves substantially when evaluating at the area level, where we filtered previous samples to only consider samples with all different general topics. Here, the model achieves $MAP@2 = 54.0\%$ and $MAP@1 = 96.9\%$ for $K = 2$, and $MAP@3 = 40.0\%$ for $K = 3$. This indicates that, while the model struggles with fine-grained multi-intent predictions, it maintains strong robustness when predicting coarse-grained semantic areas.

Interestingly, even in the area-level evaluation, the top

K	Metric	Three-tier Classification (%)	Area (%)
2	MAP@2	39.2	54.0
	MAP@1	66.0	96.9
3	MAP@3	27.8	40.0
	MAP@1	61.8	96.2

Table 8: Multi-intent evaluation results using MAP@K for K=2 and K=3. The total number of generated inputs is 60,280 for K=2 and 39,768 for K=3. When considering predictions at the area level, only inputs containing K distinct general topics are considered, resulting in 51,638 inputs for K=2 and 24,465 inputs for K=3.

predictions occasionally fall within the same general topic, despite the input containing multiple areas. This occurs most frequently for the class appearing first in the input, but also affects other cases, possibly due to unequal segment lengths within the concatenated input.

C Transparency and Explainability: Additional Experiments

Framework To assess how well keyphrases represent their corresponding Area, we embedded all keyphrases using the paraphrase-multilingual-MiniLM-L12-v2 model from SentenceTransformers (Reimers and Gurevych 2019). The embeddings were L2-normalized, reduced to 50 dimensions using UMAP (Allaoui, Kherfi, and Cheriet 2020), and clustered with HDBSCAN (minimum cluster size of 10) (McInnes, Healy, and Astels 2017). The objective was not to optimize the number of clusters, but to examine whether semantically coherent clusters aligned with Area-level categories. To quantify this alignment, we computed cluster purity (the proportion of keyphrases belonging to the majority Area label within each cluster) and normalized entropy, both weighted by cluster size.

Because keyphrases can appear multiple times in the dataset, clustering was performed on the set of unique keyphrases, while purity and entropy were computed over all occurrences. This means the evaluation reflects how frequently each keyphrase co-occurs with an Area-level category.

Results The clustering achieved a weighted-average purity of 64.0% and a normalized entropy of 46.7%, indicating that many clusters contained a mix of Area-level categories. These results likely reflect the inherent ambiguity or general scope of some keyphrases, as well as semantic overlap between Areas. Overall, the findings suggest that keyphrases provide only partial alignment with Area-level categories: they capture meaningful semantic relationships but are not sufficient indicators of Area-level distinctions.

Some limitations may also stem from the embedding and dimensionality-reduction pipeline rather than from the keyphrases themselves. Nevertheless, manual inspection of several clusters revealed clear semantic coherence, indicating that the clustering did capture meaningful similarities in the data.

D Recommendations

The following measures compromise the full, detailed list of recommended measures, complementing the concise summary presented in the main text. They can be put into place to comply with the legal obligations and, in the case of Smart e-Desk AI system, as best practice, during the use of the system. In the bullet points (specific or general case; stakeholder). Each bullet indicates whether the recommendation is general (G) or specific (S) and the relevant stakeholder (provider (P) or deployer (D)) formatted as **(recommendation type; stakeholder)**.

Bias and Fairness

- **(S; P)** Accounting for low percentage of female queries, it is also recommended to populate the training set with curated counterfactuals as a preventive measure;
- **(G; D)** Periodic fairness checks across real-world inputs, not only synthetic counterfactuals;
- **(G; D)** Logging of demographic proxies (where appropriate and compliant with GDPR) to detect emerging drift;
- **(S; D)** Domain review to identify linguistic patterns (e.g., Portuguese gender-neutral male plural forms) that might mask subtle exclusion or representational harms.

Robustness

- **(S; P)** Incorporate perturbation-based data augmentation during training to improve robustness against small, realistic input variations;
- **(S; P)** Stress testing with multilingual, orthographic, and dialectal variations, expanding beyond WordNet-based synonym attacks;
- **(G; P)** Request of provider robustness documentation detailing expected degradation under perturbations;
- **(G; D)** Adversarial testing integrated into lifecycle risk management.

Human Oversight and Impact

- **(S; P)** Incorporate a confidence-based rejection option by design, supported by documented thresholds;
- **(G; D)** Periodic review of the abstained and fallback thresholds based on real-world inputs;
- **(S; D)** Ensure that employees using and interacting with the Smart e-Desk AI system are adequately trained to be able to understand the functions, limitations, and possible risks of the tool, as well as the possibility of automation bias, to ensure that they meaningfully review abstained cases, maintain traceability of interventions and fallback decisions;
- **(G; D)** Early user involvement during AI system's development to increase trust, adoption, and enable early identification of issues because limited end-user involvement during the system's development may lead to misaligned user expectations and concerns regarding the system's adoption.

Transparency and Explainability

- **(S; P)** Include model-internal interpretability techniques (e.g., gradient-based token attributions) to enhance transparency and provide explanations that better reflect the system's actual decision-making process;
- **(S; P)** Evaluate explanations for consistency and stability across perturbations and domains;
- **(G; P)** Include contextualization warnings in user documentation when explanations require surrounding context;
- **(G; D)** Use explanations as decision-support, not decision-determinative, consistent with human oversight obligations;
- **(G; D)** The existing online policies (terms and conditions and privacy policies) of the deployer should be evaluated and updated, in light of the use of the Smart e-Desk System.

Privacy and Anonymization

- **(S; P)** Document the pseudonoymization and possible anonymization pipeline in the system's technical documentation and carry a data protection impact assessment;
- **(S; P)** Implement multi-layer anonymization to increase the success rate of the anonymizer (81.5%);
- **(G; P)** Strict minimization throughout the AI and data life-cycle should be ensured and duly documented;
- **(G; D)** Conduct regular leakage audits, especially under domain shifts or prompt variations.

Some recommendations might fit multiple categories depending on the evaluating team. We observed that technical and legal teams might categorize certain measures under different pillars, yet they complement each other in practice. For instance, implementing uncertainty or rejection thresholds to route low-confidence cases to human oversight supports effective system monitoring. While these thresholds could relate to fairness, robustness, transparency, or privacy, we categorize them under human oversight, reflecting their primary function. This example illustrates the complexity of auditing AI systems and how interdisciplinary teams can approach the same recommendation from different perspectives while achieving complementary outcomes.

References

Allaoui, M.; Kherfi, M. L.; and Cheriet, A. 2020. Considerably Improving Clustering Algorithms Using UMAP Dimensionality Reduction Technique: A Comparative Study. In El Moataz, A.; Mammass, D.; Mansouri, A.; and Nouboud, F., eds., *Image and Signal Processing*, 317–325. Cham: Springer International Publishing. ISBN 978-3-030-51935-3.

Asgharnezhad, H.; Shamsi, A.; Alizadehsani, R.; Khosravi, A.; Nahavandi, S.; Sani, Z. A.; Srinivasan, D.; and Islam, S. M. S. 2022. Objective evaluation of deep uncertainty predictions for covid-19 detection. *Scientific Reports*, 12(1): 1–11.

Brooke, J. 2013. SUS: a retrospective. *Journal of Usability Studies*, 8: 29–40.

Brown, S.; Davidovic, J.; and Hasan, A. 2021. The algorithm audit: Scoring the algorithms that score us. *Big Data & Society*, 8(1).

Chouldechova, A. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, 5(2): 153–163. PMID: 28632438.

Council of Europe. 2024. HUDERIA – Risk and Impact Assessment of AI Systems on Human Rights, Democracy and the Rule of Law. https://www.coe.int/en/web/artificial-intelligence/huderia-risk-and-impact-assessment-of-ai-systems.

Dernoncourt, F.; Lee, J. Y.; Uzuner, ; and Szolovits, P. 2017. Identifying Protected Health Information in Patient Notes with Recurrent Neural Networks. In *AMIA Annual Symposium Proceedings*, volume 2017, 1130.

DeYoung, J.; Jain, S.; Rajani, N. F.; Lehman, E.; Xiong, C.; Socher, R.; and Wallace, B. C. 2020. ERASER: A Benchmark to Evaluate Rationalized NLP Models. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4443–4458. Online: Association for Computational Linguistics.

European Union. 2016. Regulation (EU) 2016/679: General Data Protection Regulation (GDPR). Official Journal of the European Union L119.

European Union. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council on Artificial Intelligence (AI Act). Official Journal of the European Union L1689.

Feldman, M.; Friedler, S.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and removing disparate impact. arXiv:1412.3756.

Flor, M.; Futagi, Y.; Lopez, M.; and Mulholland, M. 2015. Patterns of misspellings in L2 and L1 English: a view from the ETS Spelling Corpus. volume 6.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Daumé III, H.; and Crawford, K. 2018. Datasheets for datasets. In *Proceedings of the 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning*.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of Opportunity in Supervised Learning. In Lee, D.; Sugiyama, M.; Luxburg, U.; Guyon, I.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

High-Level Expert Group on Artificial Intelligence. 2019. Ethics Guidelines for Trustworthy AI.

ISO. 2023. ISO/IEC 42001:2023 Artificial Intelligence Management System.

Jacovi, A.; and Goldberg, Y. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? *Proceedings of ACL 2020*, 4198–4205.

Kusner, M. J.; Loftus, J.; Russell, C.; and Silva, R. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, 4066–4076.

Martins, B.; and Silva, M. 2004. Spelling Correction for Search Engine Queries. volume 3230, 372–383. ISBN 978-3-540-23498-2.

McInnes, L.; Healy, J.; and Astels, S. 2017. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11): 205.

McKnight, P. E.; and Najab, J. 2010. Mann-Whitney U Test. *The Corsini encyclopedia of psychology*, 1–1.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6): 1–35.

Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I. D.; and Gebru, T. 2019. Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229.

Morris, J.; Lifland, E.; Yoo, J. Y.; Grigsby, J.; Jin, D.; and Qi, Y. 2020. TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP. In *Proceedings of EMNLP 2020: System Demonstrations*, 119–126.

Neamatullah, I.; Douglass, M.; Lehman, L.-W.; Reisner, A.; Villarroel, M.; Long, W. J.; Clifford, G. D.; Moody, G. B.; Mark, R. G.; et al. 2008. Automated de-identification of free-text medical records. *BMC Medical Informatics and Decision Making*, 8(1): 32.

NIST. 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0). Technical report, National Institute of Standards and Technology.

OECD. 2019. OECD Principles on Artificial Intelligence.

OECD. 2025. Governing with Artificial Intelligence: The State of Play and Way Forward in Core Government Functions. Technical report, OECD Publishing, Paris.

of Europe, C. 2024. Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law.

Palm, E.; and Hansson, S. O. 2006. The case for ethical technology assessment (eTA). *Technological Forecasting and Social Change, Volume 73, Issue 5*.

Raji, I. D.; Smart, A.; White, R. N.; Mitchell, M.; Gebru, T.; Hutchinson, B.; Smith-Loud, J.; Theron, D.; and Barnes, P. 2020. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 33–44.

Real, L.; Fonseca, E.; and Oliveira, H. G. 2020. The assin 2 shared task: a quick overview. In *International Conference on Computational Processing of the Portuguese Language*, 406–412. Springer.

Redding, E.; and O'Neil, C. 2023. Street-level algorithms: Public sector machine learning and the limits of accountability. *AI & Society*.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Sandvig, C.; Hamilton, K.; Karahalios, K.; and Langbort, C. 2014. Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms.

Sawilowsky, S. S. 2009. New effect size rules of thumb. *Journal of modern applied statistical methods*, 8(2): 26.

Scalia, G.; Grambow, C. A.; Pernici, B.; Li, Y.-P.; and Green, W. H. 2020. Evaluating scalable uncertainty estimation methods for deep learning-based molecular property prediction. *Journal of chemical information and modeling*, 60(6): 2697–2717.

Souza, F.; Nogueira, R.; and Lotufo, R. 2020. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

Tabarisaadi, P.; Khosravi, A.; Nahavandi, S.; Shafie-Khah, M.; and Catalão, J. P. 2022. An Optimized Uncertainty-Aware Training Framework for Neural Networks. *IEEE transactions on neural networks and learning systems*, 1–8.

Team, Q. 2025. Qwen3 Technical Report. arXiv:2505.09388.

UNESCO. 2021. Recommendation on the Ethics of Artificial Intelligence.