Preprint Manuscript

A Comparative Evaluation of Demographic Bias in PII Detection Systems

Sneha Padmaraj, Mitchell Scott

Abstract

Personally Identifiable Information (PII) detection systems are increasingly relied upon to ease to the burden of large-scale manual document review to identify sensitive information in critical legal and regulatory workflows, such as identifying impacted individuals following a cyber breach. However, bias in PII detection systems creates disproportionate privacy risks if certain demographic groups are less likely to be detected and therefore protected. In this technical report, we set out to investigate whether PII detection systems and popular LLMs exhibit a bias against non-Western names, hypothesizing that Western names are detected more reliably than non-Western names. In our evaluations, the LLMs tested missed far fewer names than the traditional PII detection tools tested, and 4 of the 12 LLMs showed a statistically significant demographic bias - a higher rate of missed names missed for one group than another. We also test an in-house PII detection LLM, which was intentionally fine-tuned on a diverse mix of synthetic and real public international documents. Of the systems that did not show statistically significant demographic bias, this LLM achieved the highest measured recall (99.3%), while maintaining an economical footprint of only 8B parameters that supports large-scale deployment. These findings emphasize the importance of designing PII detection systems for overall accuracy in tandem with fair and equitable performance across demographic groups.

1. Background

Personally Identifiable Information (PII) detection systems detect data that can be utilized to distinguish a specific individual, such as their name, residential address, driver's license number or biometric data records [1]. PII detection systems are becoming a critical component in many enterprise workflows where organisations need to robustly identify sensitive personal information within large document sets [1]. For example, PII detection plays a critical role in preventing the disclosure of sensitive information during legal and regulatory document-review workflows, supporting privacy protection and compliance requirements when processing large collections of potentially sensitive documents [1][2].

Across these scenarios, organizations must carefully balance their legal or regulatory obligation to produce information with the need to prevent inadvertent exposure of personal or sensitive data [1]. Traditionally, identifying and redacting PII is a time-consuming process, often undertaken by highly qualified individuals at great cost due to the harms involved with accidental disclosure. As document volumes grow, more organizations are turning towards automated PII detection tools to process large datasets more efficiently and to strengthen privacy protection [1].

However, the use of automated systems introduces the inherent risk of algorithmic bias, including demographic bias. As used in this report, "demographic bias" refers to a statistically measured difference in system recall between demographic groups on the tested dataset, following similar methodology to Xiao et al. (2024) [3]. This definition is intended as a statistical description to help understand experimental differences in name detection recall. It does not constitute a normative judgment about any product, developer, or person, or a conclusion that generalizes beyond the test conditions described in the methodology. Something as small as a name that goes undetected is a privacy failure: the person's identity is not redacted and could result in their personal information being unlawfully disclosed [4]. If a model systematically misses names from a particular demographic group, that group receives a lower standard of privacy protection due to demographic bias.

To understand and assess the demographic bias in these tools, we performed an evaluation of industry-standard PII detection tools as well as state-of-the-art LLM models. Our study includes LLMs because they are increasingly used for information extraction and document review, and form the backbone of newly-released commercial PII detection systems [5]. Organizations with existing access to general-purpose LLMs may use them for document review workflows without any specialist tooling [6]. Given this growing use, it is important to assess whether traditional PII systems and LLMs exhibit significant demographic biases when performing PII detection tasks.

This led us to our research question: *Do LLMs and PII (Personally Identifiable Information) Detection Tools exhibit systematic bias in detecting names based on the demographic origin of those names - specifically, do they miss Non-Western names at a higher rate than Western names?*

2. Methodology

2.1. Hypothesis

Our investigation seeks to determine whether PII Extraction systems display demographic bias when detecting people in business documents.

Null Hypothesis (H0). The rate at which the model misses Western names is equal to the rate at which it misses Non-Western names: *Miss Rate Non-Western = Miss Rate Western*.

Alternate Hypothesis (H1). The rate at which the model misses Non-Western names is not equal to the rate at which it misses Western Names: *Miss Rate Non-Western ≠ Miss Rate Western.*

2.2. Test Conditions and Scope

Table 1. Systems and subgroups tested during this study.

SYSTEM SUBGROUP	SYSTEMS TESTED
Non-LLM PII Detection Tools	Microsoft Presidio, Google Cloud Data Loss Prevention (DLP)
Open-Source LLMs	Qwen3-8B, Qwen3.5-4B, Qwen3.5-35B-A3B, Qwen3.5-27B, Qwen3.6-27B, Gemma-4-31B-IT, NVIDIA-Nemotron-3-Nano-30B-A3B-BF16
Closed-Source LLMs	Gemini 2.5 Flash, GPT-4o mini, GPT-5.4 mini, Claude Sonnet 4.6, Consilio PII Model

14 systems were evaluated during this study. All non-LLM PII detection tools systems were evaluated as third-party products using their default, out-of-the-box configurations during the testing window of May and June 2026. LLMs were run with a temperature of 0 and reasoning/thinking features disabled, while fine-tuned models used the default prompt they were fine-tuned with. Each system was tested at the most current version publicly available to us at the time of testing, with version identifiers recorded where the provider discloses them. Some systems (for example, hosted commercial tools that do not disclose their underlying model) may change over time, so these results are specific to the version and date tested.

We were unable to compare some common PII detection systems used in eDiscovery. Some of these systems are designed to only identify whether a document contains references to a unique individual, rather than to extract every form of that individual's name. As a result, their outputs are not directly comparable to the exhaustive name-extraction task evaluated in this study. Similarly, some systems are designed to identify first names and surnames as separate entities but do not resolve which names belong to the same individual. Because the benchmark requires person-level name extraction and linkage, these systems were excluded from comparison as they are not directly comparable to the evaluated task.

The reported results reflect performance on the evaluated dataset under default configurations. Performance may vary across datasets, prompts, configurations, product versions, and levels of user expertise. Non-LLM PII detection tools in particular are typically deployed with custom recognizers, entity lists, and workflow-specific tuning in enterprise workflows; their default-configuration results may therefore understate production performance to a greater degree than LLM-based systems tested with standard prompts.

2.3. Dataset Makeup and Origin

The evaluation is based on a held-out proprietary PII benchmark dataset, originally developed to train and evaluate a PII extraction model for business documents commonly encountered in eDiscovery workflows. The corpus combines public business-document datasets with synthetic business documents designed to provide broad coverage of international locations, industries, and PII types [7].

Documents were annotated for 38 categories of PII using a large language model and validated through an LLM-as-a-judge quality assurance workflow followed by human review sampling. The corpus was split into training and test sets, with the training portion used to fine-tune the Consilio PII Model. The held-out test set of 5,646 documents forms the basis of the benchmark used in this study.

Of these 5,646 documents, 5,020 contain at least one annotated ground-truth person name, while 626 contain no annotated names. After cleaning and ground-truth construction, the dataset contains 19,769 ground-truth name entries across the 5,020 name-containing documents. The annotations consist of structured person entities, including first-name variants, last-name variants, and known aliases.

At the name-entry level, the dataset is approximately balanced across the binary demographic categories used in this study, with roughly 45.7% Western and 54.3% Non-Western names after excluding the most structurally imbalanced document sources.

2.4. Ground Truth Construction

The base dataset contained structured person annotations for each document. Each annotated person entity could contain one or more first-name variants, last-name variants, and alias forms.

To construct a name-level ground truth, we generated candidate full names from each detected person by taking all possible first-name × last-name combinations. For example, an entity annotated with first names `["John", "Johnny"]` and last name `["Smith"]` would produce the candidates `John Smith` and `Johnny Smith`.

Each candidate was then verified against the source document text. Rather than requiring the full name to appear as a contiguous phrase, we independently verified that both the first-name token and the last-name token appeared in the document using word-boundary regular expressions. Candidates where either component could not be found in the source text were discarded. Candidate names were first screened against the source document using token-level checks. This step required both the first-name and last-name components to appear somewhere in the document, but it was not treated as sufficient proof that the full name appeared as a valid person reference. Because independent token matching can create cross-contaminated candidates - for example, for the name “Sarah Smith”, combining Sarah from one person with Smith from another - all retained candidates were then passed through additional validation. We searched for contiguous and extended name forms where possible, deduplicated candidate names at the (document, name) level, and used the GPT-5.5 judge review to remove candidates where the first and last name components did not genuinely refer to the same person in the source text.

For candidates that passed this verification, we also searched for extended forms containing middle names, particles, or additional name tokens between the first and last name tokens, such as `Mohan Lal Sharma` or `Mary Beth Whalen`. Where a reversed family-name-first ordering appeared as a phrase in the document, that form was also included. All resulting entries were deduplicated at the `(document, name)` level.

2.5. Ground Truth Validation

Because the ground truth was generated from structured annotations rather than manually written name-level labels, we applied a separate validation workflow to ensure benchmark quality.

After constructing this initial name-level ground truth, we ran the generated entries through an LLM-as-a-judge validation pipeline to ensure that the final dataset did not omit relevant person entities, retain entities that were present in the structured ground truth but not actually present in the generated document text, or include cross-contaminated first-name/last-name combinations. Beyond verifying existing entries, the judge also performed a full scan of each document to find any person names present in the text but absent from the ground truth, with confirmed additions incorporated directly into the benchmark. We used GPT-5.5 as our judge (and consequently GPT-5.5 was excluded from later analysis). This review was especially important for the synthetic portion of the dataset, where structured annotations and generated document text can

occasionally diverge. We also applied sanity checks and filters to remove duplicate entries, invalid name forms, and annotation variants not evidenced in the document.

In addition to automated checks, a random sample of documents and their generated ground-truth entries was reviewed manually to verify that the pipeline was behaving as expected and to calibrate the LLM judge. This human sampling step was crucial for the synthetic portion of the dataset, where ground-truth entities were generated separately from the final document text. Cases where the LLM judge made mistakes were used to improve the validation process and correct the ground truths.

2.6. Name Classification

Each ground-truth name was classified into a demographic origin category using a two-stage process.

First, we extracted a canonical name for each annotated person entity. For each entity, we used the first available first-name variant and the first available last-name variant as the canonical form. Alias variants belonging to the same person later inherited the demographic label assigned to this canonical name.

Second, canonical names were sent in batches to GPT-5.5 with a classification prompt asking it to assign each full name to one of eight fine-grained demographic subgroups: Western-European, Eastern-European, East-Asian, South-Asian, Southeast-Asian, Middle-Eastern, African and Hispanic-Latino. The model was instructed to classify based on the full name combination, rather than the first or last name in isolation, and to choose the most statistically likely origin where names were ambiguous.

The eight subgroups were then mapped into the binary categories used for the main bias analysis.

Table 2. Binary classifications of demographic subgroups

BINARY LABEL	DEMOGRAPHIC SUBGROUPS
Western	Western-European, Eastern-European
Non-Western	East-Asian, South-Asian, Southeast-Asian, Middle-Eastern, African, Hispanic-Latino

Note: Eastern-European names were grouped with the Western category because they share similar linguistic and cultural naming conventions. Preliminary classification experiments showed inconsistent separation between Eastern- and Western-European labels.

2.7. Matching Algorithm

For each ground-truth name, the model output was checked for a matching predicted person entity.

For two-token names, such as “Priya Sundaram”, a match required the first token to appear in the predicted `first_name` field and the last token to appear in the predicted `last_name` field.

For names with three or more tokens, such as “Mohan Lal Sharma”, the first token was still required to match the predicted `first_name` field and the last token the predicted `last_name` field. However, all middle tokens were allowed to appear anywhere across the predicted entity’s name fields. This rule reflects the inherent ambiguity in how names are structured and recorded across cultures and data sources. For example, “Mohan Lal Sharma” may plausibly be represented as `first_name: "Mohan Lal", last_name: "Sharma"` or as `first_name: "Mohan", last_name: "Lal Sharma"`. Both are reasonable extractions of the same person, so the design of the matching algorithm treats middle tokens flexibly rather than requiring a single culturally-specific name segmentation.

Tokenization strips diacritics using NFKD normalization, removes punctuation, and lowercases ASCII text, while preserving non-Latin scripts such as Korean or Arabic so that international names are matched verbatim.

Each alias form is treated as a legitimate name entry. If a person appears in the ground truth under multiple documented name variants, missing any one of those variants counts as a miss for that specific entry.

2.8. Demographic Bias Metric Design

To ensure the bias metric captures genuine differential name recognition rather than unrelated failure modes, we apply a layered set of exclusions before computing any bias scores. These are applied in order: structural document exclusions first, then model-level exclusions, then name-level exclusions.

2.8.1. Structural Document Exclusions

These exclusions are applied identically across all 14 models. They remove documents where failures are driven by document-level properties rather than the demographic origin of names.

IDL documents (406 docs excluded). Documents sourced from the Industry Documents Library (IDL), a collection of historical corporate records from US-origin institutions were excluded. Two properties make these documents unsuitable for bias measurement here. First, they contain a severely imbalanced name distribution: 88.1% Western and 11.9% Non-Western, compared to 45.7% Western and 54.3% Non-Western in the rest of the dataset. This reflects the US-centric origin of the corpus. Including these documents would artificially inflate Western recall simply because there are far more Western names to detect. =Second, cross-label analysis confirmed that model failure rates within these documents are near-uniform across Western and Non-Western names, indicating failures are driven by document features such as poor OCR quality and non-standard formatting rather than differential name detection.

PII-Dense documents (199 docs excluded). Documents containing more than 8 ground-truth name entries were excluded. Under high name density, many systems experience severe structural failures - Presidio fails to detect at least one name in 76% of these documents, and Google DLP in 71%. These failures appear to be driven by document structure (typically structured tables or forms) rather than the cultural origin of the names.

After this stage, 4,415 documents and 15,243 names remained in the benchmark set.

2.8.2. Model-Level Exclusions

Documents where a model produced an unusable output, such as outputs that fail JSON schema validation, are excluded across the evaluation. Across all models there were only 8 error instances across 8 unique documents: 6 for the Consilio PII Model, 1 for Nemotron 30B-A3B, and 1 for Claude Sonnet 4.6.

After this stage, 4,407 documents and 15,224 names remained in the benchmark set.

2.8.3. Name-Level Exclusion

Initial-form names. Any GT name containing a single-letter token, such as "J. Smith", "D. Mehra", or "Priya S.", is excluded from the primary bias metric at the name level. These forms are inherently ambiguous: a model may legitimately choose not to extract a bare initial because it cannot confidently attribute it to a specific person. Since the proportion of initial-form names is comparable across both groups (Western: 5.4%, Non-Western: 3.6%), this exclusion does not introduce any demographic imbalance.

After all exclusions, the primary per-model bias metric is computed over 14,551 full-name entries. Documents that contain only initial-form names - or no ground-truth names at all - do not contribute to the bias score even if they remain in the pool. They are excluded entirely from our analysis.

After this stage 4,407 documents and 14,551 names remained in the benchmark set.

2.9. Evaluation Ground Truth Split

Slightly different ground truth versions were required to perform a fair analysis between PII detection systems (Google DLP, Presidio and the fine-tuned Consilio PII Model) and the 11 general-purpose LLMs.

PII detection systems extract names in addition to other forms of PII, such as email addresses or usernames. By design, where a person's name is only present within an email or username it will appear in that specific field and not in name fields. However, general-purpose LLMs in our evaluation perform name detection as a standalone task, extracting the

names of each person found anywhere in the document. They will infer a person's name from an email address or username, for example resolving linda.farr@acme.com to the name "Linda Farr".

To account for this, the base ground truth was split into two evaluation sets using a document-text verification step:

General Purpose LLM Ground Truth. 14,551 name entries across 4,351 documents were used for all 11 general-purpose LLM models. A name entry is retained if its first and last tokens appear within 60 characters of each other anywhere in the raw document text, including inside email addresses and username fields.

PII Detection System Ground Truth. 14,478 name entries across 4,345 documents were used for Google DLP, Presidio and the Consilio PII Model. This applies the same proximity check, but only after stripping email addresses and standalone username patterns from the document text. A name entry is excluded if its tokens only appear together inside an email local-part (e.g. linda.farr@acme.com) or a firstname.lastname-style username field. In 65 documents names were removed due to this additional processing step, and in 6 documents no names remained and therefore the document was removed from evaluation entirely.

2.10. Statistical Tests

2.10.1. Motivation and Choice of Statistical Tests

The goal of this study is to determine whether person-name extraction performance differs systematically between Western and Non-Western names. We are interested not only in how often a model misses names overall, but also whether those misses occur disproportionately for one demographic group across realistic and varied test data.

To measure this, we begin with the False Negative Rate (FNR), defined as the proportion of ground-truth names that a model fails to detect [8][9]. Comparing error rates across demographic groups is a standard approach in fairness evaluation and has been widely used in studies of biometric systems and machine learning models - for example, the NIST Face Recognition Vendor Test (FRVT) evaluates demographic performance differences using false negative rates across demographic groups [10]. In our setting, the FNR provides the most direct measure of whether one group experiences more missed names than another.

However, raw error rates alone do not tell us whether an observed difference is meaningful or simply due to random variation. Because each name instance has one of two possible outcomes, detected or missed, we model miss probability using logistic regression, which is widely used in AI fairness research to assess whether demographic characteristics are associated with different outcome probabilities. Recent work examining demographic effects in large language models has similarly employed logistic regression to quantify systematic differences in model behaviour across groups [10]. The primary effect size metric is the Odds Ratio, derived from the logistic regression coefficient. Odds ratios and confidence intervals are commonly reported alongside logistic regression models in fairness and bias studies because they provide an interpretable measure of the magnitude and reliability of observed disparities [11].

A key challenge is that multiple names can appear within the same document, and names within a document share contextual factors such as OCR quality, formatting, language complexity, and document structure that make their miss outcomes correlated rather than independent. Ignoring this within-document clustering inflates the effective sample size and can produce anti-conservative p-values [12][13]. To account for this, we adopt a three-part approximation combining plain logistic regression, cluster-robust standard errors, and GEE-based clustering diagnostics, detailed in Section 2.10.2.

To assess statistical significance, we use two complementary tests: a Likelihood Ratio Test (LRT), which evaluates whether including name origin improves model fit, and a cluster-robust Wald test, which evaluates whether the estimated demographic effect remains significant after correcting for within-document dependence [12][14]. Reporting both tests provides a more robust assessment than relying on a single measure and follows common practice in statistical modelling and machine learning evaluation [12][14]. When both tests agree, we classify the result as significant; when they disagree, we classify it as non-significant. This conservative rule is particularly warranted because the LRT may be slightly anti-conservative under non-trivial within-document correlation [15], while the cluster-robust Wald test directly accounts for that dependence [12]. We also report the Intraclass Correlation Coefficient (ICC), which measures the extent to which misses are concentrated within the same document [16]. High ICC values indicate document-level failure patterns; low ICC values indicate that misses occur more independently at the individual-name level and justify treating the plain logistic likelihood approximation as reasonable [13].

2.10.2. Practical Approximation to GLMM

The dataset is structured at the name–document level, where each row contains: (1) a document identifier, (2) the extracted or ground-truth name, (3) a binary indicator for whether the name is Western or Non-Western, and (4) a binary outcome variable indicating whether the model successfully detected the name (`0`) or missed it (`1`).

Table 3. Example data structure for benchmark evaluation

DOCUMENT_ID	NAME	NONWESTERN	MISSED
doc_001	John Smith	1	1
doc_001	Priya Raman	1	1
doc_002	Wei Chen	1	0
doc_002	Michael Johnson	0	0

A natural first approach would be a two-proportion (z)-test comparing miss rates between Western and Non-Western names. Such a test treats each row as an independent Bernoulli trial and compares the overall proportion of misses across the two groups. However, this framework fails when data points are nested or clustered, violating the core assumption of statistical independence [13][17].

This assumption does not hold true in our data because multiple names may occur within the same document and these names are correlated. Names within a document share common contextual factors and are often drawn from the same social, geographic, or linguistic setting, creating dependence among observations within the document. They also share other language factors owing to OCR quality, formatting, language complexity, and document structure.

To account for within-document correlation, the ideal statistical framework is a mixed-effects logistic regression model (generalized linear mixed model; GLMM) with a document-level random intercept [18]:

$$\text{logit}\left(P(\text{missed}_{ij}=1)\right)=\beta_0+\beta_1\cdot\text{NonWestern}_{ij}+u_i$$

where:

- (i) indexes documents,
- (j) indexes names within documents,
- β_0 is the global intercept,
- $\beta_1$represents the effect of Non-Western name origin on miss probability,
- $\text{logit}\left(P(\text{missed}_{ij}=1)\right)=\beta_0$ is a binary indicator variable representing the origin classification of name j in document i. It takes the value 1 for Non-Western names and 0 for Western names. The coefficient β_1 therefore measures the change in miss probability associated with a name being Non-Western relative to a Western name,
- (u_j) is a document-level random intercept capturing shared document difficulty.

This allows us to determine whether non-Western names still more likely to be missed, while controlling for document-specific difficulty.

Fitting a binary GLMM is difficult because the model must estimate both the probability of the outcome and the variation between groups (the random effects) at the same time [19][20]. For binary outcomes, especially when events are rare or the data is sparse, this requires complex calculations that can be hard for software to solve reliably [21]. While some languages have well-established tools for this, Python's standard statistical libraries currently have more limited support for these models [22].

Therefore, we adopted a three-part alternative implemented using the python library `statsmodels` (v0.14) [12][13][23], combining plain logistic regression, cluster-robust standard errors, and GEE-based clustering diagnostics. This approach captures the primary effects of document-level dependence while remaining computationally stable [12][13].

2.10.2.1. Logistic Regression Model

We first fit a standard logistic regression using: `statsmodels.formula.api.logit`. The model predicts the binary miss outcome from a single indicator variable for Non-Western name origin:

$$\beta_0 + \beta_1 \cdot \text{NonWestern}_{ij}$$

where we assign 1 if the name is Non-Western and 0 if the name is Western.

The fitted coefficient, $\hat{\beta}_1$, represents the estimated increase in log-odds of a name being missed for Non-Western names relative to Western names.

2.10.2.2. Likelihood Ratio Test (Primary Significance Test)

To evaluate whether name origin significantly improves model fit, we performed a likelihood ratio test (LRT) between a full logistic model including `NonWestern`, and a null intercept-only model. The test statistic is:

$$\lambda = 2(L_{\text{full}} - L_{\text{null}})$$

where we fit 2 models: the Full model and the Null model.

Our research question is: "Does name origin (Western vs Non-Western) help explain whether a name gets missed?". The LRT answers this by comparing two competing explanations of the data.

The Null Model assumes that the name origins have no effect, i.e. our null hypothesis: Western names and Non-Western names have the same miss probability. The model only learns one overall miss rate. The equation is:

$$\text{logit}\left(P(\text{missed}_{ij} = 1)\right) = \beta_0$$

There is no Non-Western variable or β_1. In this model, there's no distinction between groups and it ignores the name type entirely. This means that each group gets the same prediction.

The Full Model assumes that name origins have an effect, i.e. our alternate hypothesis: Western names and Non-Western names do not have the same miss probability. One group will have a higher miss rate than the other. The equation is:

$$\text{logit}\left(P(\text{missed}_{ij} = 1)\right) = \beta_0 + \beta_1 \cdot \text{NonWestern}_{ij}$$

Here, we include the predictor `NonWesternij` so that the model can now make different predictions: Western names get one probability while Non-Western names get another probability.

The likelihood ratio statistic measures how much model fit improves when the $NonWestern$ predictor is added. Higher log-likelihood values indicate better fit, so a larger value of λ corresponds to a greater improvement of the full model over the null model. Under the null hypothesis (β_1=0), the statistic follows a $\chi^2_{(1)}$ distribution, from which the corresponding p-value is obtained [14].

2.10.2.3. Cluster-Robust Wald Test (Secondary Significance Test)

Because names within the same document are correlated, ordinary logistic regression standard errors underestimate uncertainty [10]. To correct for this, we refit the logistic model using document-clustered sandwich covariance estimators [10] via `cov_type="cluster"` and `cov_kwds={"groups":doc_id}`.

Here, `cov_type="cluster"` forces cluster-robust standard errors rather than ordinary standard errors in model training. This adjustment accounts for the possibility that observations within the same document (doc_id) may be correlated with one another, violating the assumption that all observations are independent. The argument `cov_kwds={"groups": doc_id}` specifies the clustering variable, so the model treats observations sharing the same doc_id as belonging to the same cluster when calculating standard errors. The standard errors are derived from a cluster-

robust variance-covariance matrix, which estimates variability using the residuals within each cluster and then takes the square root of the estimated variances. This typically leads to more reliable inference when data are grouped or repeated within documents [12].

The coefficient estimate $\hat{\beta}_1$ remains unchanged but inflates its standard error to account for within-document dependence. The Wald statistic is:

$$z = \frac{\hat{\beta}_1}{SE_{\text{cluster}}(\hat{\beta}_1)}$$

which asymptotically follows N[0,1] under the null hypothesis.

We use this test as a secondary check. When both the LRT and Wald tests give similar results, it suggests the effect is not caused by document-level clustering [24].

2.10.2.4. Intraclass Correlation (ICC) via GEE

To quantify the degree of within-document correlation, we fit a Generalized Estimating Equation (GEE) model using `statsmodels.genmod.generalized_estimating_equations.GEE` with a Binomial family (which is used because the outcome is binary), document-level grouping (which accounts for observations being clustered within the same document), and an Exchangeable correlation structure (which assumes that all observations within a document are equally correlated) [11].

The estimated exchangeable correlation parameter ($\hat{\alpha}$) approximates the intraclass correlation coefficient (ICC), which measures the average correlation between name outcomes in the same document [13][16]. Higher ICC values indicate stronger within-document dependence and justify the use of cluster-robust inference [12][13]. Low ICC values suggest that the plain logistic regression likelihood approximation remain reasonable.

2.10.2.5. Effect Size and Confidence Intervals using Odds Ratio (Primary Bias Metric)

Effect size was summarized using the odds ratio (OR), derived from the logistic regression coefficient β_1. The Odds Ratio is defined as:

$$OR = e^{\hat{\beta}_1}$$

where $\hat{\beta}_1$ represents the estimated effect of the NonWestern_{ij} predictor in the logistic regression model. The coefficient is estimated in log-odds space, so exponentiating it converts it into an odds ratio that is easier to interpret.

The odds ratio compares the odds of a miss for Non-Western names relative to Western names:

$$OR = \frac{\text{Odds of miss for Non-Western names}}{\text{Odds of miss for Western names}}$$

An odds ratio greater than 1 indicates increased odds of a miss for Non-Western names. Unlike Bias %, which measures the absolute difference in false negative rates between groups, the odds ratio measures the relative disparity and allows comparison between models with different False Negative Rates [25].

We computed 95% confidence intervals using the cluster-robust standard error:

$$CI_{95\%} = \exp\left(\hat{\beta}_1 \pm 1.96 \times SE_{\text{cluster}}(\hat{\beta}_1)\right)$$

Confidence intervals excluding 1.0 provide an independent significance check consistent with the Wald test [10].

2.10.3. Final Metrics

We report 7 primary metrics.

Table 4. Metrics reported in this study.

Metric	Purpose	Interpretation
False Negative Rate (FNR)	The proportion of names that the model failed to detect within a demographic group.	Higher values indicate worse detection performance for that group.
Odds Ratio (OR)	The relative difference in miss odds between Non-Western and Western names.	OR > 1 indicates higher miss odds for Non-Western names; OR = 1 indicates no disparity; OR < 1 indicates higher miss odds for Western names.
Absolute Bias	The raw difference in False Negative Rates between Non-Western and Western names (FNR_NW − FNR_W), expressed in percentage points.	A positive value indicates Non-Western names are missed more often; a negative value indicates Western names are missed more often. Provides a direct measure of the gap, but because it depends on the model's overall miss rate, it should not be used to compare fairness across models with very different recall levels.
95% Confidence Interval (CI)	Quantifies uncertainty around the estimated odds ratio.	A CI that excludes 1.0 provides independent evidence of a meaningful disparity.
Likelihood Ratio Test (LRT)	Tests whether name origin significantly improves model fit.	A significant result ($p < 0.05$) indicates that name origin is associated with miss probability.
Cluster-Robust Wald Test	Tests whether the estimated effect remains significant after correcting for within-document correlation.	Serves as a robustness check for the LRT result.
Intraclass Correlation	Measures how strongly misses are clustered within documents.	High ICC indicates document-level failure patterns; low ICC indicates misses occur more independently across names.

3. Results

All results below reflect the dataset, configurations, and test conditions described in Section 2.

3.1 Summary of results

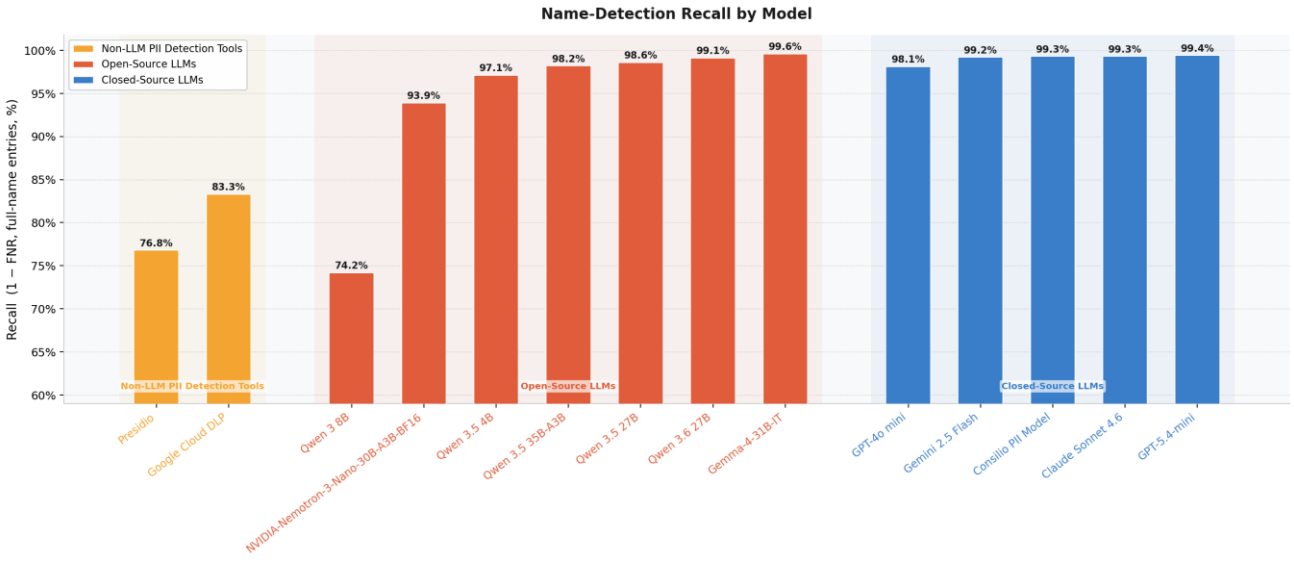


Figure 1. Name detection recall by system.

Table 5: Evaluation results, in descending order of recall.

	RECALL	*FNR_W*	*FNR_NW*	*LRT P*	*WALD P*	*SIG*	*OR (95% CI)*	*ICC*
Gemma-4-31B-IT	**99.6%**	**0.27%**	**0.52%**	0.0181	0.0432	*	1.91 (1.02-3.57) ‡	0.166
GPT-5.4-mini	99.4%	0.45%	0.77%	0.0144	0.0286	*	1.71 (1.06-2.75) ‡	0.085
PII Model	99.3%	0.67%	0.65%	0.8797	0.8904	ns	0.97 (0.62-1.51)	**0.068**
Claude Sonnet 4.6	99.3%	0.72%	0.62%	0.4342	0.5141	ns	0.85 (0.53-1.38)	0.151
Gemini 2.5 Flash	99.2%	0.63%	0.88%	0.0842	0.1370	ns	1.40 (0.90-2.17)	0.260
Qwen 3.6 27B	99.1%	0.65%	1.05%	0.0090	0.0336	**	1.62 (1.04-2.53) ‡	0.166
Qwen 3.5 27B	98.6%	1.31%	1.43%	0.5598	0.6065	ns	1.09 (0.79-1.49)	0.212
Qwen 3.5 35B-A3B	98.2%	1.72%	1.82%	0.6601	0.6814	ns	1.06 (0.81-1.38)	0.410
GPT-4o mini	98.1%	1.75%	1.94%	0.3906	0.4304	ns	1.11 (0.85-1.45)	0.078
Qwen 3.5 4B	97.1%	3.03%	2.81%	0.4332	0.5232	ns	0.93 (0.73-1.17)	0.311
NVIDIA-Nemotron-3-Nano-30B-A3B-BF16	93.9%	6.58%	5.73%	0.0331	0.0406	*	0.86 (0.75-0.99) ‡	0.123
Google Cloud DLP	83.3%	16.53%	16.81%	0.6572	0.6982	ns	1.02 (0.92-1.13)	0.326
Presidio	76.8%	23.01%	23.44%	0.5379	0.5820	ns	1.02 (0.94-1.12)	0.263
Qwen 3 8B	74.2%	25.96%	25.72%	0.7395	0.7926	ns	**0.99** (0.90-1.09)	0.794

*Interpretation: * $p < 0.05$; ** $p < 0.01$; ns = statistically non-significant; ‡ 95% confidence interval excludes 1.0. Note on default configurations. All systems were evaluated in their default, out-of-the-box configurations. Non-LLM PII detection tools such as Presidio and Google DLP are typically deployed in enterprise workflows with custom recognizers, entity lists, and workflow-specific tuning. Their default-configuration recall figures reported here may therefore understate production performance to a greater degree than LLM-based systems tested with standard prompts.*

Across the 14 evaluated models, 9 exhibited higher false negative rates (FNRs) for Non-Western names than for Western names, while 5 exhibited the reverse pattern. After accounting for within-document clustering, statistically significant evidence of demographic bias against Non-Western names was observed in three models: Gemma-4-31B-IT, GPT-5.4-mini, and Qwen 3.6 27B. These models produced odds ratios ranging from 1.62 to 1.91, indicating that Non-Western names had between 62% and 91% higher odds of being missed. One model, Nemotron-3-Nano-30B-A3B-BF16, showed statistically significant evidence of higher miss rates for Western names. Overall, 4/14 models evaluated showed statistically significant demographic bias.

Model reliability varied substantially across systems. Higher-recall models generally exhibited lower intraclass correlation coefficients (ICCs), indicating that errors occurred as isolated misses rather than as complete document-level failures. In contrast, lower-performing systems such as Qwen 3 8B, Google Cloud DLP, and Presidio exhibited higher ICC values, reflecting stronger within-document clustering of errors and a greater prevalence of whole-document failure modes.

The Consilio PII Model achieved the highest recall (99.3%) of any system that did not show statistically significant evidence of demographic bias (OR = 0.97, LRT p = 0.880, Wald p = 0.890).

Absolute differences in FNR between demographic groups were generally small, ranging from +0.10 to +0.43 percentage points among models with higher Non-Western miss rates. Among statistically significant models showing higher Non-Western miss rates, absolute gaps ranged from +0.25 to +0.40 percentage points. Relative measures such as odds ratios therefore provide important complementary information, particularly for highly accurate systems where small absolute differences can correspond to substantial relative disparities. Conversely, statistical evidence for demographic disparities was generally weaker among lower-performing systems, where elevated overall error rates made group-specific effects more difficult to distinguish from baseline model error.

3.2 Repeatability analysis

Given time and cost constraints, each of the tested models were evaluated on the benchmark dataset once. To verify the stability of our evaluation, we selected four models representing a range of observed significance levels and ran each three

times on the full dataset: Qwen3.6-27B (very significant, LRT $p < 0.01$), GPT-5.4-mini (significant, LRT $p < 0.05$), Gemini 2.5 Flash (marginally non-significant, LRT $p = 0.0842$), and Qwen3.5-4B (non-significant, LRT $p = 0.4332$). Each run in this experiment was a fully independent inference on the benchmark dataset, in addition to the original evaluation run.

Table 5. Repeatability analysis results.

MODEL	OR MEAN ± SD	SIGNIFICANCE CLASSIFICATION ACROSS RUNS
Qwen3.5-4B	0.93 ± 0.008	ns / ns / ns
Gemini 2.5 Flash	1.38 ± 0.019	ns/ ns /ns
GPT-5.4-mini	1.88 ± 0.242	** / * * * / **
Qwen3.6-27B	1.52 ± 0.106	* / ** / *

*Interpretation: * $p < 0.05$; ** $p < 0.01$; ns = statistically non-significant.*

Across all repeated runs, recall varied by at most 0.04 percentage points and odds ratios showed standard deviations of 0.02–0.44 across runs. Results were stable in both direction and magnitude.Upon inspection, wider variance for GPT-5.4-mini and Qwen3.6-27B reflects run-to-run variation in a small number of borderline documents rather than instability in the underlying bias signal. Given the minimal recall variance and consistent significance effects across all models on this dataset, we find it reasonable to rely on single run results for the remaining 10 models.

3.3 Qwen Size Scaling

We conducted a deeper analysis into our Qwen3.5 family of models to investigate if there was any correlation between the size of the model and the demographic bias it contains. Here, given similar recall, we measure the absolute bias %. Model effective size was measured as either the dense model weights size, or for mixture of experts models the geometric mean of total and active parameters.

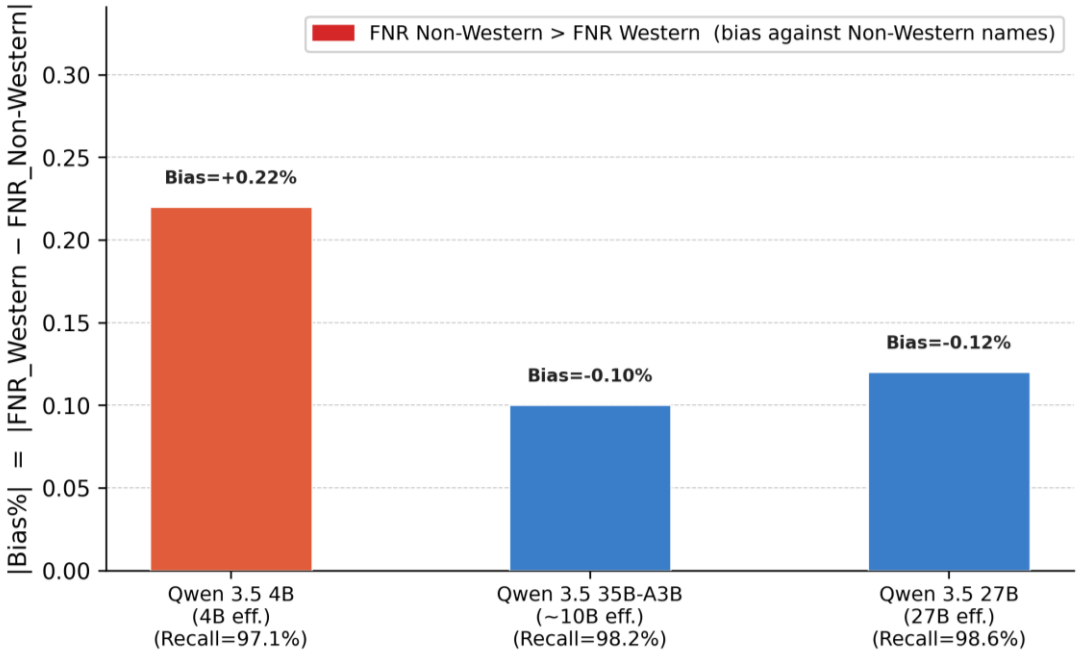


Figure 2. Effect of LLM size on measured demographic bias in Qwen 3.5 model family.

Table 6. Comparison of evaluation metrics for Qwen 3.5 model family.

MODEL	EFFECTIVE SIZE	FNR_W	FNR_NW	RECALL	ABSOLUTE BIAS	LRT P	SIG	OR (95% CI)
Qwen3.5-4B	4B	3.03	2.81%	97.1%	-0.22%	0.4332	ns	0.93 (0.73–1.17)

Qwen3.5-35B-A3B	~10B	1.72%	1.82%	98.2%	+0.10%	0.6601	ns	1.06(0.81–1.38)
Qwen3.5-27B	27B	1.31%	1.43%	98.6%	+0.12%	0.5598	ns	1.09(0.79–1.49)

Within the Qwen 3.5 family, increasing effective model size improved overall extraction recall but did not produce statistically significant demographic disparities in any direction. The 27B and 35B variants achieved higher recall than the 4B model (98.6%, 98.2%, and 97.1% respectively). The 4B model showed a marginal tendency toward higher miss rates for Western names (OR = 0.93, 95% CI: 0.73–1.17), while the larger variants showed small tendencies in the opposite direction - Qwen 3.5 27B (OR = 1.09, 95% CI: 0.79–1.49) and Qwen 3.5 35B-A3B (OR = 1.06, 95% CI: 0.81–1.38). None of these effects reached statistical significance after accounting for within-document clustering (all LRT $p > 0.43$).

These observations are consistent with recent work showing that scaling model size alone does not necessarily improve fairness outcomes and that the evolution of bias depends strongly on the characteristics of the training data [20].

3.4 Fine Tuned Model and Base Model Comparison

This experiment shows the improvements made by our fine-tuned Consilio PII model vs its base model, Qwen3-8B.

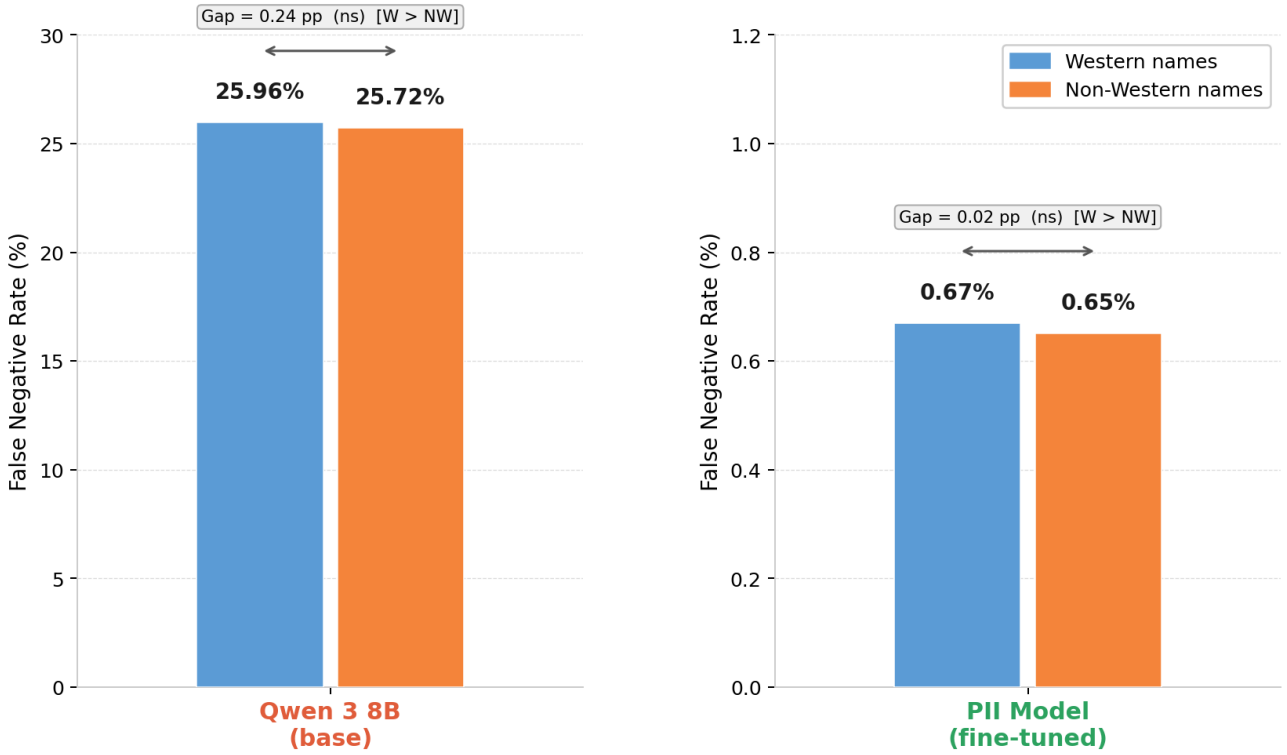


Figure 3. Comparison between Qwen3-8B and Consilio PII Model (fine-tuned version of Qwen3-8B)

Table 7. Comparison of evaluation metrics for Qwen 3.5 model family.

METRIC	QWEN 3 8B (BASE)	CONSILIO PII MODEL (FINE-TUNED)
Overall recall	74.2%	**99.3%**
FNR - Western	25.96%	**0.67%**
FNR - Non-Western	25.72%	**0.65%**
LRT p-value	0.7395(ns)	0.8795(ns)
Wald p-value (Cluster-Robust)	0.7926 (ns)	0.8904(ns)
Odds Ratio (NW vs W Miss)	**0.99** (0.90–1.09)	0.97 (0.62–1.51)
ICC (Within-Document Correlation)	0.794	**0.068**

Fine-tuning on a mix of synthetic and real public international documents improved recall from 74.2% to 99.3% and virtually eliminated empty-output documents (Qwen3-8B found no person entities in ~1900 documents). Notably, our fine-tuned Consilio PII model achieved these scores despite performing a substantially more complex task of detecting 38 distinct PII entity types simultaneously. It also maintained the non-significant demographic bias outcome.

Neither the base model nor the fine-tuned model shows significant evidence that name origin affects miss rates (LRT p = 0.740 and p = 0.880 respectively). The biggest improvement is in reliability - the ICC drops from 0.794 to 0.068, showing that the base model often failed entire documents and missed every name within them, whereas the fine-tuned model's remaining misses are much more isolated.

3.5 Origin and Source Type Comparison

We compare Non-LLM PII Detection Tools, Open-source LLMs, and Closed-Source LLMs to evaluate whether different architectures, training strategies, and data sources lead to distinct demographic bias patterns, or whether the observed disparities persist consistently across fundamentally different approaches to PII extraction.

We use the Symmetric Miss Rate Difference (SMRD) which measures the relative gap between Western and Non-Western miss rates, normalized by their combined miss rate, so that fairness can be compared across models without the logarithmic scaling of the Odds Ratio metric.

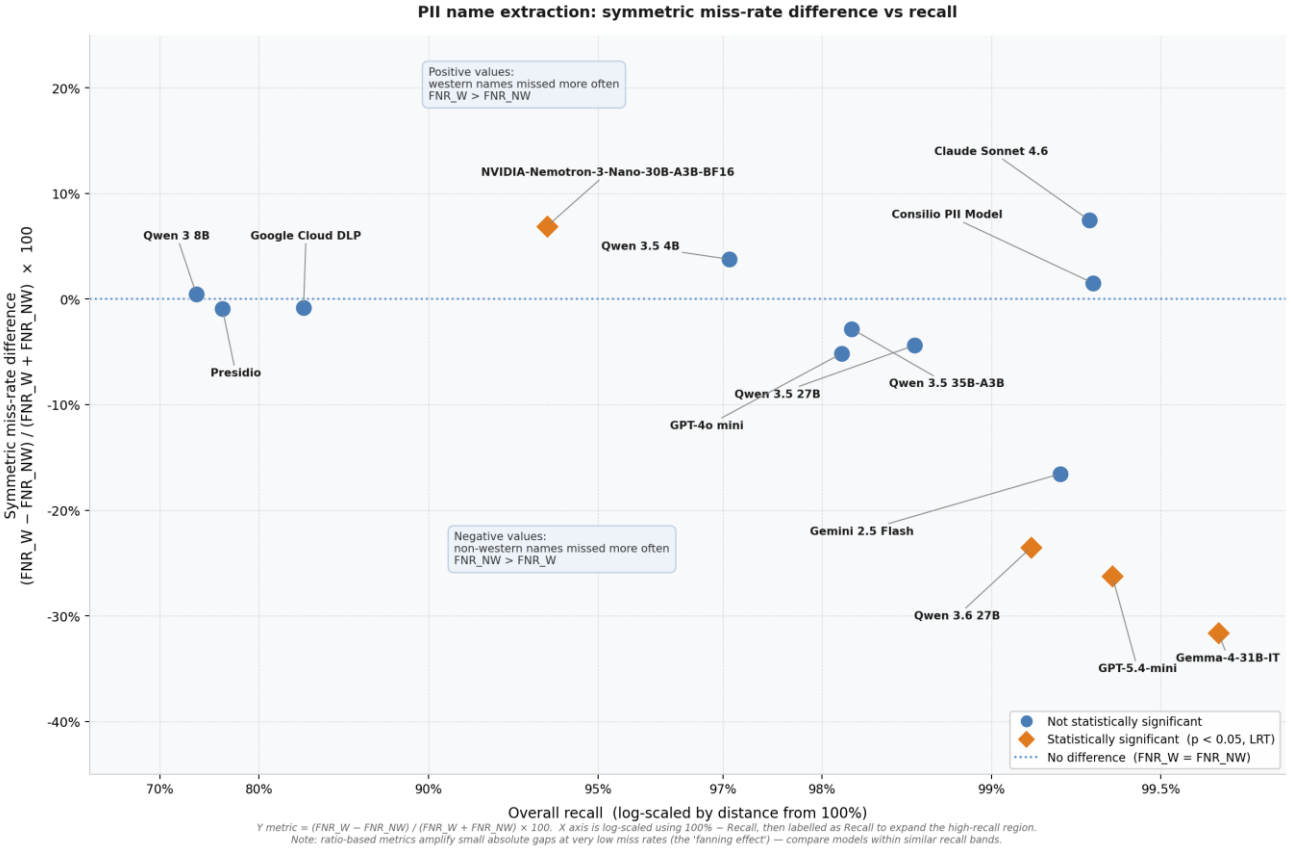


Figure 4. Symmetric miss-rate difference on name extraction task. SMRD = (FNR_W − FNR_NW) / (FNR_W + FNR_NW) × 100. Negative values indicate Non-Western names missed more often; Positive values indicate Western names missed more often.

Table 8. Subgroup categorisation of systems.

MODEL	SYSTEM TYPE	DEVELOPER ORIGIN
Gemini 2.5 Flash	Closed-Source LLM	Western
GPT-4o mini	Closed-Source LLM	Western
GPT-5.4-mini	Closed-Source LLM	Western
Claude Sonnet 4.6	Closed-Source LLM	Western
Consilio PII Model	Closed-Source LLM	Western
Qwen 3.6 27B	Open-Source LLM	Non-Western
Qwen 3.5 35B-A3B	Open-Source LLM	Non-Western

Qwen 3.5 27B	Open-Source LLM	Non-Western
Qwen 3 8B	Open-Source LLM	Non-Western
Qwen 3.5 4B	Open-Source LLM	Non-Western
Gemma-4-31B-IT	Open-Source LLM	Western
NVIDIA-Nemotron-3-Nano-30B-A3B-BF16	Open-Source LLM	Western
Presidio	Non-LLM PII Detection Tool	Western
Google Cloud DLP	Non-LLM PII Detection Tool	Western

Table 9. Subgroup results on benchmark.

GROUPING	SUBGROUP	N	AVERAGE OR	SIGNIFICANT MODELS
System type	Closed-Source LLMs	5	1.21	1/5
System type	Open-Source LLMs	7	1.21	3/7
System type	Non-LLM PII Detection Tools	2	1.02	0/2
Developer origin	Western	9	1.21	3/9
Developer origin	Non-Western	5	1.14	1/5

When grouped by model origin and source type, no clear relationship emerges between where a model was developed and how it performs across demographic groups. Both open-source and proprietary systems contain examples of models with statistically significant demographic disparities as well as models with little or no measurable bias. Of the four models that reach statistical significance, three - GPT-5.4-mini, Gemma 4 31B, and Qwen 3.6 27B - show higher miss rates for Non-Western names (OR = 1.71, 1.91, and 1.62 respectively). The fourth, Nemotron 30B-A3B, is significant in the opposite direction (OR = 0.86), meaning Western names are missed more often by that model. The magnitude and significance of bias vary substantially between models.

Because SMRD is normalized by the combined miss rate, it amplifies small absolute gaps at very low overall miss rates (the fanning effect) - a model with 99.6% recall and a 0.25pp gap can produce an SMRD as large as a model with a much larger absolute disparity at lower recall. SMRD should therefore be read alongside overall recall and raw FNR values rather than in isolation.

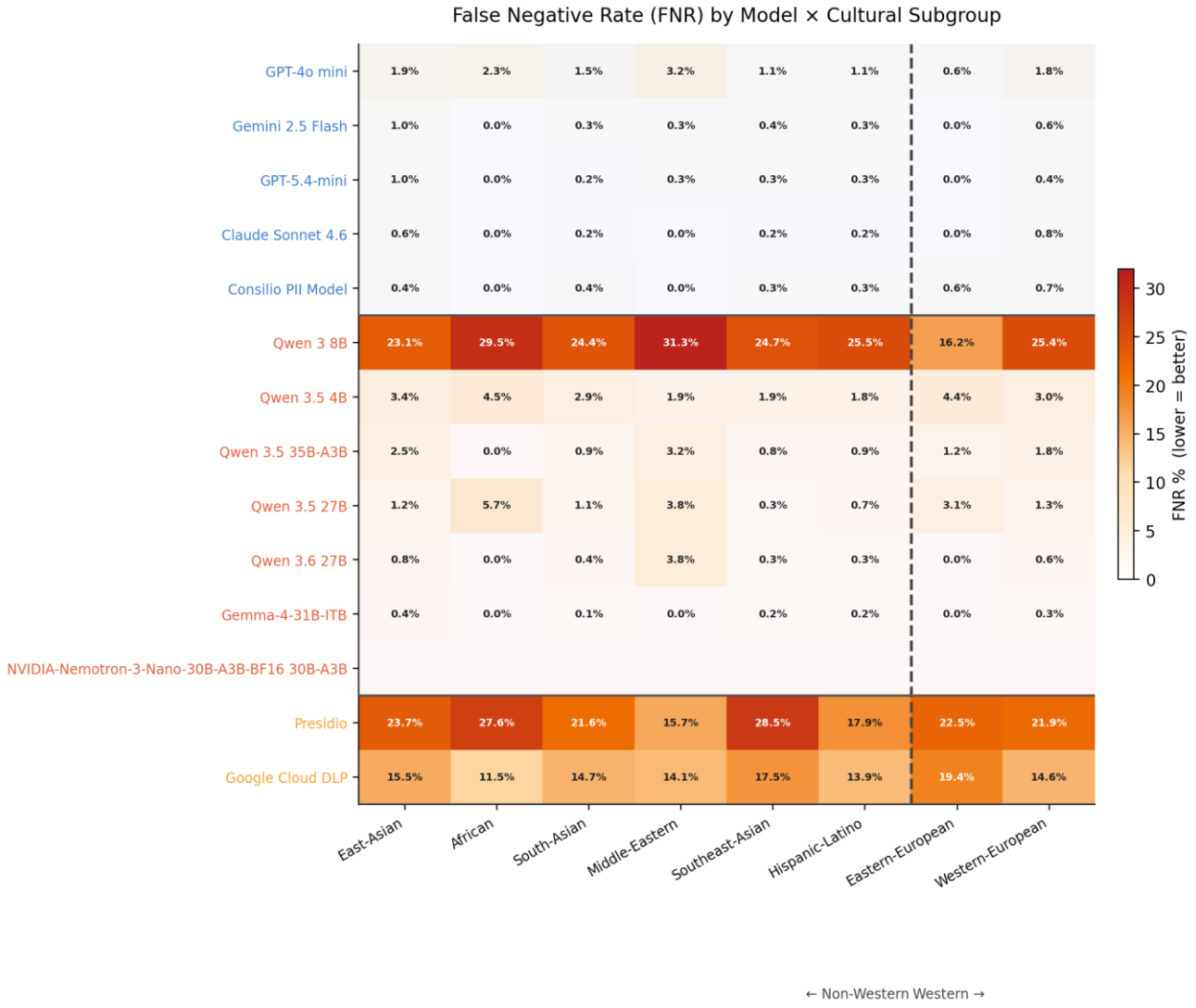
The results also suggest that model origin alone is not a reliable predictor of fairness outcomes. For example, Qwen 3.6 27B exhibits statistically significant Non-Western disparity despite strong overall recall (99.1%), while the other four Qwen models evaluated show no significant bias. On the other hand, the fine-tuned Consilio PII Model demonstrates that targeted training on a diverse, domain-specific dataset can achieve strong performance without introducing statistically significant evidence of bias. These findings indicate that training data composition, fine-tuning objectives, and model behaviour are likely to play a larger role in determining fairness outcomes than whether a model is open-source, closed-source, Western-developed, or non-Western-developed.

3.6 Subgroup Analysis

Overall recall level had a larger effect on subgroup performance than subgroup membership itself. High-recall models - including all four closed-source LLMs, the best-performing open-source LLMs (Gemma 4 31B, Qwen 3.6 27B), and the Consilio PII Model achieved uniformly low false negative rates across all eight subgroups, with nearly all values below 3%. Within this group, Middle-Eastern names were the most consistently elevated subgroup, reaching 3.2% for GPT-4o mini and 3.8% for Qwen 3.5 27B and Qwen 3.6 27B, though absolute differences between subgroups remained small throughout.

By contrast, the two Non-LLM PII Detection Tools - Presidio and Google DLP - exhibited substantially higher miss rates across every subgroup, indicating a broad recall limitation rather than a concentrated demographic effect. For Presidio, the highest miss rates were observed for Southeast-Asian (28.5%) and African (27.6%) names; for Google DLP, Eastern-European names were the most missed (19.4%), followed by Southeast-Asian (17.5%). Qwen 3 8B showed a similar pattern of uniformly elevated miss rates, peaking at 31.3% for Middle-Eastern names, consistent with its overall low recall driven largely by high volumes of empty model outputs.

Figure 5. Comparison of False Negative Rate (FNR) across demographic subgroup.



4. Discussion

The null hypothesis - that miss rates for Non-Western names are equal to miss rates for Western names - was rejected at $p < 0.05$ for 4 of the 14 evaluated systems. Three of these showed higher miss rates for Non-Western names: Qwen 3.6 27B (OR = 1.62, p = 0.009), GPT-5.4-mini (OR = 1.71, p = 0.014), and Gemma 4 31B (OR = 1.91, p = 0.018), with odds ratios ranging from 1.62 to 1.91. One model, Nemotron 30B-A3B, reached significance in the opposite direction (OR = 0.86, p = 0.033), meaning Western names were missed more often. For the remaining 10 systems, no statistically significant demographic bias was detected - though non-significance does not establish the absence of demographic bias.

In enterprise PII detection settings like litigation discovery, a missed name carries consequences that are asymmetric with the cost of a correct detection. A document whose custodian or subject is not recognised may be incorrectly withheld, excluded from a privilege review, or left unredacted in a regulatory production. Where a system showed up to 1.91 times higher odds of missing Non-Western names than Western names, as measured in this evaluation, that disparity could result in records containing people from certain cultural backgrounds receiving a lower standard of protection in real workflows. In practice, this risk scales with the diversity of the custodian population: large multinational matters, immigration proceedings, or employment litigation involving globally distributed workforces are contexts where this bias would have the greatest impact. These findings argue that demographic fairness should be treated as an explicit, measurable evaluation criterion when selecting or auditing PII extraction tools. A tool can achieve 98%+ recall and still demonstrate a significant miss rate disparity approaching 2× by cultural origin, as several models in this study did.

Two results were not anticipated at the outset. First, recall and fairness proved largely orthogonal. The intuition that more capable models would naturally be fairer was not supported. Second, within the Qwen 3.5 family, increasing effective parameter count from 4B to 27B improved name extraction recall measurably but had very minor effects on the Odds Ratio

(0.93 → 1.06 →1.09), which shifted from a marginal Western-direction effect to a marginal Non-Western-direction effect. None of these differences reached statistical significance, so this should be treated as an observed trend rather than a confirmed finding. If a training data composition is driving this directional shift, however, it may be reinforced rather than diluted with model scale. Further research investigating the relationship between bias metrics and model scaling is required.

The most notable positive finding was the effect of targeted fine-tuning in improving name detection recall without demonstrating statistically significant bias. The Consilio in-house PII model, fine-tuned from the same Qwen3-8B base that produced the worst recall results in this study (74.2%), achieved 99.3% recall - a gain of more than 25 percentage points over its base model. This is particularly notable as it was run with the prompt template that it was fine-tuned on, requiring it to extract 38 types of PII simultaneously – a more difficult task than the standalone name extraction task other LLMs in this study performed. This indicates that fine-tuning an LLM on diverse mix of international data examples may help to reduce the risk of demographic bias in LLM-based systems, while also improving overall performance. Further comparative fine-tuning studies would help to clarify this effect.

5. Limitations and areas for future work

Interpretation of demographic bias metrics becomes more challenging for systems with poor overall recall. Models such as Qwen3-8B and several traditional PII detection tools exhibited substantially higher false negative rates than the strongest-performing systems, meaning that a large proportion of names were missed regardless of demographic subgroup. In such settings, demographic disparities may be more difficult to detect statistically because subgroup-specific effects are obscured by the overall volume of errors. Consequently, a non-significant bias result should not necessarily be interpreted as evidence of fairness; it may instead reflect limited statistical power to distinguish demographic effects from general system underperformance. Bias estimates are therefore most informative when considered alongside overall recall and error rates.

Some observed failures, particularly in the Qwen models, appeared to be formatting errors rather than true recognition failures. For example, Arabic mononyms were sometimes placed in the wrong JSON field, while Chinese full names were occasionally merged into a single field. Non-LLM Detection Systems also exhibited issues with correctly isolating names from surrounding text or punctuation. Under the current evaluation methodology, these cases are counted as misses even when the model appears to have identified the correct name. This means that some of the measured demographic disparity may reflect structured-output compliance issues rather than genuine differences in name recognition. Future work could evaluate both schema-strict matching and a more flexible "name detected anywhere in the output" approach to better separate formatting failures from recognition failures.

Due to the limited number of public eDiscovery datasets we constructed our own as part of this research, which was split into a training set and a held-out testing set. While efforts were made to ensure that the dataset represents business documents typically found in an eDiscovery use case, there may be idiosyncrasies due to the method of the dataset's construction. We encourage future work to validate these results across new datasets in the eDiscovery domain and other related PII detection scenarios.

The Consilio PII Model was fine-tuned on the training partition of the same proprietary dataset whose held-out test partition forms the benchmark used in this study. This decision was made as no public eDiscovery-type datasets with sufficient demographic diversity and scale were found at the time of writing. Although the training and test sets contain no overlapping documents, the Consilio PII Model may benefit from exposure to the same underlying data distribution, including document formats, synthetic-generation patterns, annotation conventions, and the mix of international name origins represented in the corpus. All other systems evaluated had no prior exposure to data from this pipeline. This experimental setup means that the Consilio PII Model's benchmark performance may not be directly comparable to that of the other evaluated systems on this specific dataset, and readers should exercise caution when drawing competitive inferences from the recall figures alone. To fully isolate the effect of fine-tuning on demographic fairness from any dataset-familiarity advantage, future work should evaluate the Consilio PII Model against independent, externally constructed benchmarks.

The ground truth was constructed through a rigorous multi-stage pipeline: automated candidate generation from structured entity annotations, document-level verification confirming that both name tokens appear in the source text, and a final LLM-as-a-judge review pass to remove spurious or cross-contaminated entries. To provide an additional quality check, a random sample of documents was manually reviewed by a human evaluator and was found to be broadly consistent with the automatically generated labels. While these steps were designed to maximise reliability, no automatically constructed ground truth is entirely free of error. In particular, if any stage of the pipeline produces errors that correlate with demographic group, those errors could partially confound the demographic bias measurements reported here. We consider

this a low-probability but non-negligible risk, and systematic quantification of ground-truth error rates remains an avenue for future work.

References

[1] Lison, P., Pilán, I., Sanchez, D., Batet, M., & Øvrelid, L. (2021). Anonymisation models for text data: State of the art, challenges and future directions. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 4188–4203). Association for Computational Linguistics.

[2] Allard, T., Béziaud, L., & Gambs, S. (2020). *Online publication of court records: Circumventing the privacy-transparency trade-off*. arXiv. https://arxiv.org/abs/2007.01688

[3] Yuxin Xiao, Shulammite Lim, Tom Joseph Pollard, and Marzyeh Ghassemi. 2023. In the Name of Fairness: Assessing the Bias in Clinical Record De-identification. *In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 123–137.

[4] Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics*, 50(3):1097–1179.

[5] Brown, C. (2026, March 9). *Introducing Relativity aiR for Data Breach Response*. Relativity Blog. https://www.relativity.com/blog/introducing-relativity-air-for-data-breach-response/

[6] Martin, L., Whitehouse, N., Yiu, S., Catterson, L., & Perera, R. (2024). *Better Call GPT, comparing large language models against lawyers*. arXiv. https://doi.org/10.48550/arXiv.2401.16212

[7] Industry Documents Library. (n.d.). *Datasets*. University of California, San Francisco. https://www.industrydocuments.ucsf.edu/resources/datasets/

[8] Jacquemet V. A statistical model of false negative and false positive detection of phase singularities. *Chaos*. 2017 Oct; 27(10):103124.

[9] Marina Sokolova, Guy Lapalme. A systematic analysis of performance measures for classification tasks, *Information Processing & Management*, Volume 45, Issue 4, ISSN 0306-4573.

[10] Grother, P., Ngan, M., & Hanaoka, K. (2019). Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects. National Institute of Standards and Technology (NIST). *NIST Interagency Report* 8280.

[11] Sekkat, C., Deroy, O., & Mercier, H. (2024). Sonos Voice Control Bias Assessment Dataset: A Methodology for Demographic Bias Assessment in Voice Assistants. arXiv.

[12] Cameron, A. C., & Miller, D. L. (2015). A Practitioner's Guide to Cluster-Robust Inference. *Journal of Human Resources*, 50(2), 317–372

[13] KUNG-YEE LIANG, SCOTT L. ZEGER, Longitudinal data analysis using generalized linear models, *Biometrika*, Volume 73, Issue 1, April 1986, 13–22.

[14] Wilks, S. S. 1938. The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *Annals of Mathematical Statistics* 9, 1 (1938), 60–62.

[15] Hedges, L. V. 2007. Correcting a Significance Test for Clustering. *Journal of Educational and Behavioral Statistics* 32, 2 (2007), 151–179.

[16] Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*. 1979 Mar;86(2):420-8.

[17] Hubbard AE, Ahern J, Fleischer NL, et al. To GEE or not to GEE: comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. Epidemiology. 2010; 21(4):467-474.

[18] Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390–412.

[19] Breslow, N. E., & Clayton, D. G. (1993). Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association*, 88(421), 9–25.

[20] Bolker BM, Brooks ME, Clark CJ, et al. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends Ecol Evol.* 2009; 24(3):127-135.

[21] Pinheiro, J. C., & Bates, D. M. (1995). Approximations to the Log-Likelihood Function in the Nonlinear Mixed-Effects Model. *Journal of Computational and Graphical Statistics*, 4(1), 12–35.

[22] Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1–48.

[23] Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with Python. In *Proceedings of the 9th Python in Science Conference* (pp. 57–61).

[24] Hauck, W. W., & Donner, A. (1977). Wald's Test as Applied to Hypotheses in Logit Analysis. *Journal of the American Statistical Association*, 72(360a), 851–853.

[25] Peters, U., & Chin-Yee, B. (2025). Generalization bias in large language model summarization of scientific research. *Royal Society Open Science*, 12(4), 241776.

Disclaimer. *This study reflects the results of independent testing conducted by Consilio LLC using the methodology described herein, as of 14/06/2026. This is a preprint manuscript. It has not been submitted for peer review as of the date of publication. The authors intend to submit this work for peer review, and this document will be updated to reflect any changes resulting from that process. All findings are based on the specific test conditions, model versions, prompt configurations, and evaluation criteria identified in this paper. Results may vary under different conditions, configurations, or model versions. This paper reports observed outcomes for the specific models tested and does not purport to characterize the overall quality, reliability, or fitness for purpose of any third-party product or service. All systems were evaluated using their default, out-of-the-box configurations at the versions and on the dates identified herein; results may not reflect a product's performance when configured, tuned, or deployed for a specific production workflow. A finding that a system did not exhibit a statistically significant demographic bias does not establish that the system is unbiased or fair. All product and company names are trademarks of their respective owners, are used here for identification and comparison only, and this paper does not imply any affiliation with, endorsement by, or sponsorship from any third party. The models referenced are the property of their respective owners. Consilio makes no representation that the results presented here are generalizable beyond the specific test parameters described, and no claim of overall product superiority is intended. Readers are encouraged to conduct their own independent testing before making procurement or implementation decisions.*