The "10th Juror": Open-Set Standpoint Screening for Bureaucratic Bias Detection

Yuchen Miao[1], Zijun Wang[1], Chang Han[1], Yurui Shi[3], Mingtai Zhang[4], Siyang Xu[2,*]

[1]Sydney Smart Technology College, Northeastern University, China
[2]School of Computer and Communication Engineering, Northeastern University at Qinhuangdao, China
[3]Taiyuan University of Technology
[4]School of Resources, Environment and Materials, Guangxi University

Abstract

Presupposing the boundaries of bias is itself a form of bias. We study closed-loop bias governance for Dutch government documents, where a system must detect biased language, ground decisions in legal and contextual evidence, rewrite problematic sentences when intervention is warranted, and verify that the rewrite mitigates harm without distorting meaning. Existing methods face three challenges: (i) discriminative classifiers capture surface regularities but lack normative grounding; (ii) zero-shot LLMs often adopt generic viewpoints and over-flag ambiguous administrative language; and (iii) fixed taxonomies inherit the Closed-World Assumption, missing emerging local targets. We propose MARS-Gov, a standpoint-aware multi-agent framework that combines legal retrieval, open-set target screening, specialized jurors, conservative routing, and rewrite verification. When screening finds an uncovered group, MARS-Gov instantiates a dynamic "10th juror" to deliberate outside the fixed panel. On DGDB, MARS-Gov sets a new SOTA with 0.880 F1, outperforming the strongest zero-shot LLM detector by 20.2 points (29.8% relative) and the best supervised Dutch encoder by 6.8 points, while reducing unnecessary interventions to 2.5%. Leave-One-Category-Out (LOCO) evaluation recovers held-out categories with 85.1% Correct@1 and 93.8% Correct@3.

1 Introduction

Government documents are a primary interface between the state and its citizens. Their language therefore does more than transmit information: it marks who is treated as credible, risky, dependent, or entitled to care. Linguistic neutrality matters for institutional trust (Rizk and Lindgren, 2025; Eubanks, 2018), yet bias can persist even in restrained administrative prose (de Swart et al., 2025; Benjamin, 2019). It often appears not as overt toxicity (Waseem and Hovy, 2016), but as exclusionary naming, negative framing, presupposition, or uneven standards of responsibility (Otmakhova et al., 2024). A sentence may be formally polite and still attach suspicion, incompetence, or burden to a protected or socio-economically vulnerable group.

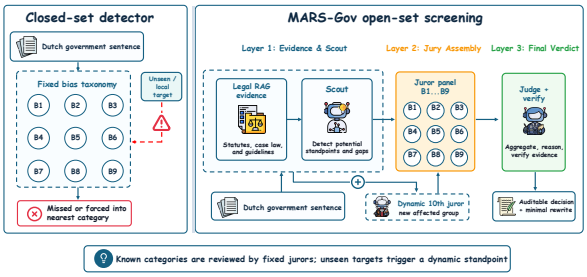

Figure 1: **Closed-set detection versus MARS-Gov open-set screening.** Fixed taxonomies may miss local targets or force them into a known class. MARS-Gov retrieves legal evidence, detects standpoint gaps, adds a dynamic 10th juror when needed, and sends grounded reports to the Judge for an auditable verdict and minimal rewrite.

This setting makes bias detection different from ordinary harmful-language classification. The relevant question is not only whether a phrase sounds negative, but whether it is normatively grounded, legally salient, and administratively faithful. A reliable system must identify the affected standpoint, distinguish a necessary eligibility condition from a stigmatizing proxy, and avoid rewriting official text in ways that change factual obligations. These requirements turn detection into a governance problem: the system must explain, intervene only when warranted, and verify that any intervention preserves meaning.

Existing approaches face three connected challenges. First, supervised encoders can be strong on closed benchmarks (Devlin et al., 2019; Chalkidis et al., 2020), but they mainly learn label regularities

*Corresponding author: xusiyang@neuq.edu.cn

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 12576–12594
October 24-29, 2026 ©2026 Association for Computational Linguistics

and provide limited evidence for why a decision is normatively justified (Blodgett et al., 2020). Second, zero-shot LLMs are more flexible (Gallegos et al., 2024), but they often adopt a generic fairness viewpoint, over-flag group mentions, and give plausible verdicts without showing which legal or administrative norm supports the decision (Resnik, 2025). Third, both paradigms usually assume that the ontology of bias is fixed before evaluation begins.

A deeper problem is therefore the fixed label space. DGDB contains 3,747 Dutch government-document sentences with binary labels organized around nine bureaucratic-bias categories. As Figure 1 shows, a detector trained on these categories is evaluated as if all relevant categories were known in advance (Scheirer et al., 2013). Administrative language is less tidy. Digital illiteracy stigma, regional disadvantage, care dependency, or new socio-economic stereotypes may appear before an annotation scheme names them. Treating such cases as negative examples silently preserves the Closed-World Assumption; representing them only through the nearest fixed standpoint may preserve a binary alert while losing the affected group's specific rationale. A governance detector therefore needs both known-category review and open-set target discovery.

We propose MARS-Gov (Multi-Agent Reasoning for Standpoint-aware Governance), a closed-loop framework for government text and a new SOTA system on DGDB. For each sentence, MARS-Gov first retrieves legal and administrative context through Normative Evidence Contextualization (NEC). It then performs Open-Set Standpoint Screening (OSS): a Scout identifies potentially affected groups, fixed jurors inspect the nine known categories, and a dynamic "10th juror" is created when the Scout finds a target outside the panel. An Alert-based Routing Gate (ARG) conservatively clears low-risk sentences and escalates uncertain cases to a grounded Judge. If the Judge flags bias, a Restoration Agent proposes a minimal rewrite, and the system verifies the rewrite with the same detection criteria. The result is not a single classifier with a longer prompt, but an auditable decision loop for detection, mitigation, and review.

Our contributions are:

- We formulate bureaucratic bias detection as closed-loop governance: detection, evidence retrieval, minimal rewriting, and verification.
- We introduce MARS-Gov, a legal-retrieval and juror-based framework that sets a new SOTA on DGDB.
- We show open-set and cross-lingual robustness through a dynamic "10th juror" and governance metrics for mitigation quality and over-governance.

2 Related Work

Bias in government documents is shaped by institutional practice as well as model behavior (Eubanks, 2018; Noble, 2018; Benjamin, 2019; Rizk and Lindgren, 2025). Legal NLP resources such as LEGAL-BERT, LexGLUE, and LegalBench-RAG support this setting (Chalkidis et al., 2020, 2022; Pipitone and Houir Alami, 2024), while DGDB shows that bureaucratic bias often appears through framing and presupposition rather than slurs (de Swart et al., 2025). This view is consistent with discourse, framing, stereotyping, and intersectionality research, which treats bias as situated and socially produced (Fairclough, 1992; van Dijk, 1993; Goffman, 1974; Entman, 1993; Otmakhova et al., 2024; Allport, 1954; Fiske, 1998; Crenshaw, 1989). For administrative review, this matters because the same phrase can be harmless in a procedural context and discriminatory when it assigns risk, competence, or trustworthiness to a protected or stigmatized group.

Supervised and benchmark-based bias detection. Computational bias detection has relied on association tests, debiasing methods, and BERT-style encoders (Caliskan et al., 2017; Bolukbasi et al., 2016; Devlin et al., 2019). Benchmarks such as StereoSet, BBQ, WinoBias, Social Bias Probing, and F2Bench probe stereotypes or fairness behavior with controlled examples (Nadeem et al., 2021; Parrish et al., 2022; Zhao et al., 2018; Marchiori Manerba et al., 2024; Lan et al., 2025). These resources are useful, and fine-tuned encoders remain strong closed-set baselines, but their fixed labels or demographic contrasts make them brittle for newly emerging administrative targets. They also give little support for the later governance question: what should be changed, and how can the system know that the change did not distort the original administrative meaning?

LLM-based detection, audits, and retrieval. LLMs can apply natural-language criteria without task-specific fine-tuning (Gallegos et al., 2024; Resnik, 2025). Audits nevertheless show hidden social associations, toxic continuations, long-form

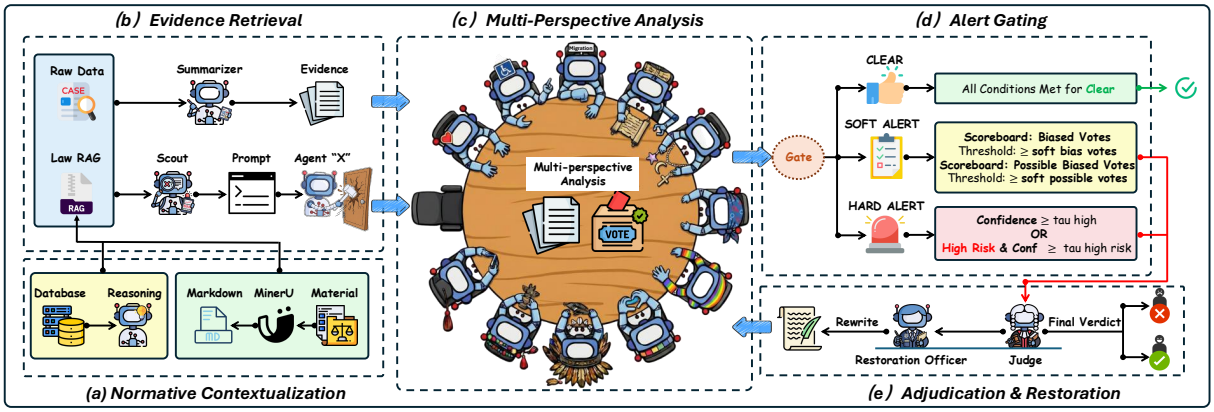

Figure 2: **Overview of MARS-Gov.** The framework retrieves normative evidence, gathers juror risk reports, routes only risky cases to a Judge, and verifies any minimal rewrite. Separating evidence, routing, adjudication, and restoration keeps the decision trace inspectable.

bias, and false positives around group mentions (Smith et al., 2022; Blodgett et al., 2020; Gehman et al., 2020; Jeung et al., 2025; Waseem and Hovy, 2016; Dixon et al., 2018). Retrieval can add legal evidence, but retrieval alone does not decide which standpoint should evaluate a sentence, how conflicting standpoints should be adjudicated, or whether a rewrite has repaired the problem. These unresolved steps are precisely where an apparently helpful LLM review can become hard to audit.

Open-set and agentic governance. Open-set recognition and out-of-distribution detection study cases outside known categories (Scheirer et al., 2013; Hendrycks and Gimpel, 2017); recent debiasing work applies related ideas to LMs (Rani et al., 2025). Multi-agent LLM systems coordinate specialized roles for software, simulation, safety review, and factuality (Wu et al., 2023; Li et al., 2023; Qian et al., 2024; Hong et al., 2024; Park et al., 2023; He et al., 2025; Ning et al., 2025). Most such systems define agent roles in advance. We combine legal grounding, open-set target discovery, standpoint-specialized agents, and closed-loop rewriting so that a newly affected group can receive a dedicated review standpoint instead of being represented only through the nearest fixed one.

3 Methodology

3.1 Task Definition: Closed-loop Bias Governance

We formulate *closed-loop bias governance* for a sentence $x \in \mathcal{X}$ from a government document or benchmark. In DGDB, x is a Dutch administrative sentence with an expert bias label. Unlike one-shot classification, the system may intervene and must verify the intervention. A pipeline $D(\cdot)$ returns an initial label $\hat{y} = D(x) \in \{0, 1\}$, an optional minimal restoration $x' = R(x)$, and a detection-only verification $\hat{y}' = D_{\text{verify}}(x')$ for non-empty restorations. The output also stores retrieved evidence and juror reports for audit. The task optimizes detection reliability, mitigation, and controlled side effects (§3.7). This framing treats a biased sentence as a governance failure only if the system can identify the issue, propose a faithful repair, and confirm that the repair is no longer judged biased.

Two constraints follow from this formulation. First, the system must explain why a sentence is problematic in terms that can be checked against external norms, rather than only against the model's internal label boundary. Second, the rewrite must remain narrow: it should remove the biased implication without changing the administrative fact pattern. We therefore treat evidence retrieval, standpoint review, adjudication, restoration, and verification as separate artifacts in the trace.

3.2 Overview: Pipeline-defined Detector

Figure 2 shows MARS-Gov as a pipeline-defined detector rather than a single classifier. This design keeps evidence, standpoint assignment, routing, and intervention separate enough to audit. It first builds a neutral evidence summary $E(x) = \text{NEC}(x; \mathcal{K})$ from knowledge base $\mathcal{K}$, then uses a Scout and juror panel to produce reports $\mathcal{J}(x) = \text{OSS}(x, E(x), \mathcal{G}(x))$. ARG clears low-risk cases and sends the rest to a Judge; if restoration is produced, D_{verify} applies the same detection criteria with rewriting disabled.

The trace records the retrieved evidence iden-

Algorithm 1: Inference in MARS-Gov.

Input: Sentence x; knowledge base $\mathcal{K}$; fixed jurors $\mathcal{B}_{\text{fix}} = \{B_1, \ldots, B_9\}$.
Output: Detection $\hat{y}$; optional restoration x'; verification $\hat{y}'$; trace $\mathcal{T}$.

$E \leftarrow \text{NEC}(x; \mathcal{K})$
$\mathcal{G} \leftarrow \text{Scout}(x, E)$
$\mathcal{U} \leftarrow \text{Uncovered}(\mathcal{G}, \mathcal{B}_{\text{fix}})$
$\mathcal{B}_{\text{dyn}} \leftarrow \{\text{Instantiate}(g, x, E) : g \in \mathcal{U}\}$
$\mathcal{J} \leftarrow \text{OSS}(x, E, \mathcal{B}_{\text{fix}} \cup \mathcal{B}_{\text{dyn}})$ `// one report per fixed or dynamic juror`
$a \leftarrow \text{ARG}(\mathcal{G}, \mathcal{J})$ `// rule-based routing gate`
$\hat{y} \leftarrow 0;\ x' \leftarrow \emptyset;\ \hat{y}' \leftarrow \emptyset$
if $a \neq$ CLEAR **then**
  $\hat{y} \leftarrow \text{Judge}(x, E, \mathcal{J})$
  if $\hat{y} = 1$ **then**
    $x' \leftarrow R(x, E, \mathcal{J})$ `// minimal restoration`
    if $x' \neq \emptyset$ **then**
      $\hat{y}' \leftarrow D_{\text{verify}}(x')$ `// same criteria; restoration disabled`
    end
  end
end
$\mathcal{T} \leftarrow \{E, \mathcal{G}, \mathcal{J}, a, \hat{y}, x', \hat{y}'\}$
return $(\hat{y}, x', \hat{y}', \mathcal{T})$

tifiers, the Scout's affected-group hypotheses, all juror reports, the routing action, the Judge rationale, the proposed restoration, and the verification result. These records are not presented as legal authority, but they make the model's path from sentence to intervention inspectable.

3.3 Normative Evidence Contextualization via RAG (NEC)

Governance decisions depend on policy definitions, protected attributes, and domain guidance, not only sentence semantics. NEC gives all downstream agents the same reference frame by applying three-stage RAG over $\mathcal{K}$ (Dutch equal-treatment law, antidiscrimination policy, accessibility guidance, online-harm and institutional-racism reports, inclusive-language guidance, and DGDB category notes): keyword retrieval, semantic retrieval, and chunk-level context expansion. Figure 3 summarizes these three retrieval stages. For input x, the retrieved evidence is

$$C(x) = \text{Read}\big(\text{Ret}_{\text{kw}}(x; \mathcal{K}),\ \text{Ret}_{\text{sem}}(x; \mathcal{K});\ \mathcal{K}\big). \tag{1}$$

NEC then produces a neutral summary

$$E(x) = \text{Summ}(x, C(x)). \tag{2}$$

The summarizer returns definitions, criteria, examples, and short caveats, but no verdict. This keeps NEC from becoming an implicit classifier and leaves later judgments traceable to $C(x)$ rather than to unobserved model priors. In administrative text, this separation is important: the same word can be neutral in a rule description and harmful when attached to a group as a proxy for risk or competence.

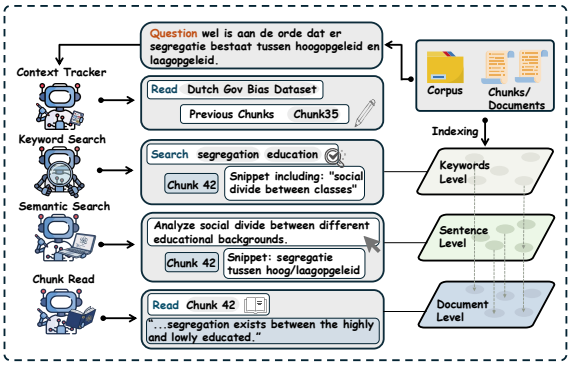

Figure 3: Three-layer retrieval in NEC.

3.4 Open-Set Standpoint Screening (OSS): Scout + Juror Panel

Bias governance is open-set because affected groups may not fit a fixed taxonomy. Given the input and evidence, the Scout identifies potentially implicated groups or sensitive attributes:

$$\mathcal{G}(x) = \text{Scout}(x, E(x)). \tag{3}$$

Let $\mathcal{B}_{\text{fix}} = \{B_1, \ldots, B_9\}$ denote the fixed panel, which covers Disability, Migration, Colonialism, Religion, Gender, LGBTQ+, Culture, Education/Class, and Welfare. The Scout maps every target in $\mathcal{G}(x)$ to this panel and collects the uncovered targets in

$$\mathcal{U}(x) = \{g \in \mathcal{G}(x) : \text{Cover}(g, \mathcal{B}_{\text{fix}}) = 0\}. \tag{4}$$

For every $g \in \mathcal{U}(x)$, OSS instantiates one temporary, evidence-conditioned juror:

$$\begin{aligned} \mathcal{B}_{\text{dyn}}(x, E) &= \{B_g : g \in \mathcal{U}(x)\}, \\ B_g &= \text{Instantiate}(g, x, E). \end{aligned} \tag{5}$$

Thus, a sentence with several uncovered standpoints receives a separate dynamic report for each standpoint rather than collapsing them into one generic analysis. DGDB predominantly assigns one primary category per sentence, so the common case creates a single additional juror, informally called the "10th juror"; this name is not an architectural limit. Dynamic jurors are not permanent

ontology classes, but temporary review standpoints grounded in the current sentence and evidence.

Each fixed or dynamic juror returns an evidence-grounded report:

$$\begin{aligned} \mathcal{B}(x, E) &= \mathcal{B}_{\text{fix}} \cup \mathcal{B}_{\text{dyn}}(x, E), \\ \mathcal{J}(x, E) &= \{J_B : B \in \mathcal{B}(x, E)\}, \\ J_B &= B(x, E). \end{aligned} \tag{6}$$

Here $J_B = (b_B, p_B, r_B, \rho_B)$ contains a binary verdict, confidence, severity, and a short rationale grounded in $E(x)$, optionally with supporting spans. These reports are passed to routing and restoration, giving the Judge several grounded views rather than one undifferentiated LLM answer.

This design also limits a common failure mode in open-set detection. The system may notice an unfamiliar affected group, but the final decision still passes through the same Judge, evidence, and verification steps as known categories. Open-set discovery therefore expands who can be reviewed, while the downstream standard for intervention remains unchanged.

3.5 Alert-based Routing Gate (ARG): Risk-sensitive Escalation

To reduce cost without lowering recall, MARS-Gov uses an *alert-based routing gate* (ARG)[1]. ARG maps $(\mathcal{G}, \mathcal{J})$ to CLEAR or ESCALATE, clearing only when all risk signals are weak and well-formed. It escalates under three conditions: *hard* alerts, when Scout confidence exceeds τ_{high} or a high-risk watchlist threshold; *soft* alerts, when multiple jurors give weaker but consistent signals; and *fail-safe* alerts, when any juror output is missing, malformed, or unparsable.

$$\text{ARG}(\mathcal{G}, \mathcal{J}) = \begin{cases} \textsc{Clear}, & \text{no alert fires}, \\ \textsc{Escalate}, & \text{otherwise}. \end{cases} \tag{7}$$

Escalated cases go to the Judge and, when needed, restoration. ARG is rule-based, so routing changes do not introduce extra backbone calls.

3.6 Decision and Restoration: Judge and Verify with Minimal Rewriting

For escalated cases, the Judge resolves juror conflicts using x, $E(x)$, and $\mathcal{J}(x, E(x))$. It grounds the final decision in explicit evidence and outputs a binary label

$$\hat{y} \;=\; \text{Judge}(x,\, E(x),\, \mathcal{J}(x, E(x))) \in \{0, 1\}. \tag{8}$$

The Judge sees raw reports and evidence, but not the ARG verdict, which keeps adjudication separate from threshold tuning. This matters in practice because conservative routing should affect cost and recall, not the legal rationale of the final verdict.

If $\hat{y} = 1$, a Restoration Agent generates a minimally edited sentence

$$x' \;=\; R(x, E(x), \mathcal{J}(x, E(x))). \tag{9}$$

The restoration removes biased framing while preserving factual content, meaning, and administrative register. It minimizes surface edits, returns exactly one revised sentence, and emits an empty string if no safe rewrite is possible. Empty restorations are treated as failed mitigation. The restored sentence is then verified:

$$\hat{y}' \;=\; D_{\text{verify}}(x'). \tag{10}$$

A governance action succeeds when $\hat{y} = 1, x' \neq \varnothing$, and $\hat{y}' = 0$, so the same detector clears the non-empty restoration. This prevents a rewrite from being counted as successful merely because it was produced.

We deliberately avoid free-form paraphrasing. In pilot runs, broad rewrites often improved tone while changing who did what, which is unacceptable for administrative records. The restoration agent is therefore instructed to preserve named entities, dates, obligations, eligibility criteria, and material facts unless the biased wording itself must be replaced.

We measure meaning preservation by the cosine similarity between sentence embeddings of the original and restored text:

$$\text{SF}(x, x') \;=\; \cos\big(\phi(x),\, \phi(x')\big)\,, \tag{11}$$

where $\phi(\cdot)$ is a fixed text encoder.[2] If no restoration is produced, $\text{SF}(x, x') = 0$.

3.7 Pipeline-aware Governance Metrics

F1 only assesses the initial decision on x. It does not show whether restoration resolves the problem or whether the system intervenes on neutral

[1] The concrete thresholds are selected on the validation set and reported in Appendix F.

[2] We employ `clips/e5-base-trm-nl`; weights are available at https://huggingface.co/clips/e5-base-trm-nl.

text. We therefore add metrics for the full detect–restore–verify loop. Given $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, let $\hat{y}_i = D(x_i)$, x'_i be the restored sentence, $\hat{y}'_i = D_{\text{verify}}(x'_i)$ when $x'_i \neq \varnothing$, and $m_i = \mathbb{I}[x'_i \neq \varnothing]$. We also compute $\text{SF}_i = \text{SF}(x_i, x'_i)$, with $\text{SF}_i = 0$ when no rewrite is produced. Bias Mitigation Rate (BMR) measures the share of correctly detected biased cases that are restored and cleared:

$$\text{BMR} = \frac{\sum_i m_i \, \mathbb{I}[y_i = 1 \wedge \hat{y}_i = 1 \wedge \hat{y}'_i = 0]}{\sum_i \mathbb{I}[y_i = 1 \wedge \hat{y}_i = 1]}. \tag{12}$$

Semantic Fidelity (SF) averages meaning preservation over successfully mitigated cases:

$$\overline{\text{SF}} = \frac{\sum_i m_i \, \text{SF}_i \cdot \mathbb{I}[y_i = 1 \wedge \hat{y}_i = 1 \wedge \hat{y}'_i = 0]}{\sum_i m_i \, \mathbb{I}[y_i = 1 \wedge \hat{y}_i = 1 \wedge \hat{y}'_i = 0]}. \tag{13}$$

Bias Governance Score (BGS) combines coverage, mitigation, and fidelity:

$$\text{BGS} = \text{Recall} \times \text{BMR} \times \overline{\text{SF}}. \tag{14}$$

Including Recall prevents a pipeline from scoring well by governing only easy positives. Over-Governance Rate (OGR) measures unnecessary intervention on non-biased inputs, equivalent to the false-positive rate on the initial detector:

$$\text{OGR} = \frac{\sum_i \mathbb{I}[y_i = 0 \wedge \hat{y}_i = 1]}{\sum_i \mathbb{I}[y_i = 0]}. \tag{15}$$

Together, BMR, $\overline{\text{SF}}$, BGS, and OGR evaluate mitigation quality and side effects beyond one-shot classification. We report these components alongside F1 because a model can look strong under ordinary classification while over-rewriting neutral text, or while producing rewrites that do not survive verification.

4 Experiments

We organize the experiments around five questions:
RQ1: Does MARS-Gov outperform fine-tuned, zero-shot, and agentic baselines?
RQ2: Can the Scout recover held-out bias categories outside the fixed juror panel?
RQ3: Which modules account for the gains in detection and governance quality?
RQ4: How much inference cost does ARG save without changing the detector?
RQ5: Does the review procedure remain useful across domains and languages?

4.1 Experimental Setup

Datasets.[3] We use DGDB (de Swart et al., 2025), a Dutch administrative benchmark with nine bias categories. To test vocabulary generalization, we evaluate on a held-out subset, DGDB-Unseen, containing new bias terms. We also use DALC (Caselli et al., 2021), a Dutch social-media offensive-language set; SBIC (Sap et al., 2020), an English social-bias inference corpus; and KoBBQ (Jin et al., 2024), a Korean BBQ-style bias benchmark.
Baselines. We compare fine-tuned PLMs (*BERTje*, *RobBERT*), zero-shot LLMs (*GPT-4o*, *Grok-3*, *Gemini 2.5 Pro*, *DeepSeek-R1*), and agentic controls: *SA-RAG* and *Standard Debate*. Two matched Gemini 2.5 Pro controls test whether stronger prompting can explain the gains. A fixed 12-shot CoT prompt uses nine biased demonstrations spanning the DGDB categories and three hard non-biased demonstrations, all selected from non-test data with fixed IDs and order. A *Single-Call Integrated Prompt* receives the same retrieved NEC evidence, task definition, and output requirements as MARS-Gov, but performs affected-group identification, classification, rationale generation, minimal rewriting, and self-checking in one LLM call. For governance metrics, all generative controls use the same frozen D_{verify} and fixed SF encoder; the Single-Call self-check does not determine BMR or BGS.

4.2 Overall Performance (RQ1)

Table 1 reports in-domain DGDB and DGDB-Unseen performance across four backbones[4]. DGDB-Unseen replaces selected DGDB bias terms with alternative unseen expressions while preserving labels and administrative context, testing whether detectors generalize beyond the original lexical cues.

❶ **Architecture Trumps Generic Debate.** Zero-shot models have high recall but plateau at F1 $\approx$ 0.63–0.68 because precision stays below 0.55. Standard Debate improves precision only modestly, and SA-RAG does not consistently surpass the best fine-tuned encoder. MARS-Gov reaches SOTA F1 (> 0.85) across all backbones, with Gemini 2.5 Pro obtaining the best mean F1 (0.880).
Matched prompting controls. The 12-shot CoT prompt raises Gemini 2.5 Pro F1 from 0.664 to

[3] Dataset statistics are summarized in Appendix A.
[4] Additional experiments with more backbone models are reported in Appendix B.

	Prec.	Rec.	F1	$F1_{OOD}$	ΔDrop	OGR↓	BMR	SF	BGS
Fine-tuned									
BERTje (de Vries et al., 2019)	0.842 ±.017	0.784 ±.021	0.812 ±.014	0.554 ±.032	-31.8% ±1.62	3.1% ±0.51	-	-	-
RobBERT (Delobelle et al., 2020)	0.839 ±.019	0.784 ±.018	0.811 ±.013	0.499 ±.034	-38.5% ±1.94	3.6% ±0.62	-	-	-
I. Prompted LLMs (Zero-/Few-shot)									
GPT-4o	0.514 ±.025	0.882 ±.021	0.649 ±.017	0.621 ±.024	-4.3% ±1.07	16.4% ±0.83	0.872 ±.016	0.821 ±.021	0.631 ±.021
Grok-3	0.484 ±.031	**0.920** ±.017	0.634 ±.019	0.598 ±.029	-5.7% ±1.18	19.1% ±0.94	0.864 ±.019	0.808 ±.024	0.642 ±.022
Gemini 2.5 Pro	0.524 ±.024	0.907 ±.018	0.664 ±.016	0.636 ±.022	-4.2% ±0.95	15.2% ±0.72	0.881 ±.014	0.827 ±.018	0.661 ±.020
Gemini 2.5 Pro (12-shot CoT)	0.642 ±.021	0.883 ±.019	0.743 ±.017	0.716 ±.021	-3.6%	9.6% ±0.61	0.914 ±.015	0.876 ±.018	0.707 ±.018
DeepSeek-R1	0.549 ±.023	0.887 ±.019	0.678 ±.015	0.652 ±.021	-3.8% ±0.83	13.4% ±0.65	0.893 ±.013	0.851 ±.016	0.674 ±.017
II. Standard Debate (No RAG)									
Standard Debate (GPT-4o)	0.587 ±.026	0.862 ±.020	0.698 ±.018	0.671 ±.026	-3.9% ±1.13	12.1% ±0.77	0.902 ±.017	0.868 ±.019	0.674 ±.020
Standard Debate (Grok-3)	0.549 ±.029	0.906 ±.019	0.684 ±.021	0.651 ±.031	-4.8% ±1.43	14.5% ±0.87	0.878 ±.019	0.851 ±.022	0.677 ±.024
Standard Debate (Gemini 2.5 Pro)	0.612 ±.023	0.886 ±.019	0.724 ±.018	0.696 ±.024	-3.9% ±0.96	10.7% ±0.69	0.908 ±.016	0.884 ±.018	0.711 ±.021
Standard Debate (DeepSeek-R1)	0.634 ±.022	0.882 ±.018	0.738 ±.017	0.711 ±.023	-3.7% ±0.86	10.2% ±0.63	0.921 ±.014	0.892 ±.017	0.724 ±.018
III. Integrated and Agentic Controls									
SA-RAG (GPT-4o)	0.667 ±.020	0.852 ±.019	0.748 ±.015	0.723 ±.021	-3.3% ±0.71	8.9% ±0.59	0.913 ±.015	0.876 ±.017	0.682 ±.017
SA-RAG (Grok-3)	0.619 ±.024	0.896 ±.017	0.732 ±.017	0.700 ±.024	-4.4% ±1.04	11.1% ±0.71	0.889 ±.018	0.862 ±.021	0.687 ±.019
SA-RAG (Gemini 2.5 Pro)	0.687 ±.019	0.876 ±.018	0.770 ±.014	0.748 ±.019	-2.9% ±0.62	7.9% ±0.51	0.918 ±.013	0.893 ±.014	0.718 ±.015
Single-Call Integrated Prompt	0.762 ±.019	0.869 ±.018	0.812 ±.015	–	–	6.7% ±0.55	0.927 ±.015	0.912 ±.017	0.735 ±.019
SA-RAG (DeepSeek-R1)	0.710 ±.018	0.876 ±.017	0.784 ±.013	0.762 ±.018	-2.8% ±0.53	7.2% ±0.47	0.929 ±.012	0.902 ±.013	0.734 ±.013
IV. Ours: MARS-Gov Framework									
MARS-Gov (GPT-4o)	0.844 ±.017	0.863 ±.015	0.853 ±.013	0.823 ±.019	-3.5% ±0.62	3.5% ±0.32	0.934 ±.011	0.939 ±.013	0.757 ±.014
MARS-Gov (Grok-3)	0.836 ±.019	**0.910** ±.013	0.871 ±.014	0.840 ±.022	-3.6% ±0.77	3.9% ±0.41	0.922 ±.013	0.928 ±.015	0.779 ±.018
MARS-Gov (Gemini 2.5 Pro)	0.865 ±.015	0.895 ±.013	**0.880** ±.012	**0.853** ±.017	-3.1% ±0.51	**2.5%** ±0.28	**0.951** ±.009	**0.949** ±.011	**0.808** ±.013
MARS-Gov (DeepSeek-R1)	**0.872** ±.016	0.884 ±.014	0.878 ±.013	0.857 ±.017	**-2.4%** ±0.47	2.6% ±0.29	**0.948** ±.011	0.962 ±.010	0.806 ±.015

Table 1: **Main DGDB results.** Darker green marks stronger performance; darker red marks larger ΔDrop or OGR. $F1_{OOD}$ is DGDB-Unseen F1; ΔDrop is the relative drop from DGDB F1. Single-Call was evaluated in-domain only. MARS-Gov gives the best mean F1/BGS with low governance intrusion.

0.743 and lowers OGR from 15.2% to 9.6%. Integrating all steps into one call further raises F1 to 0.812 and BGS to 0.735. The full staged procedure remains 6.8 F1 points and 7.3 BGS points higher than Single-Call, while reducing OGR from 6.7% to 2.5%. These comparisons isolate gains beyond demonstrations, prompt length, and retrieved context alone.

❷ DGDB-Unseen Exposes Lexical Fragility. Fine-tuned encoders look attractive on the original split: BERTje and RobBERT are compact and reach F1 ≈ 0.81, higher than most Standard Debate variants. Yet their $F1_{OOD}$ falls to 0.554 and 0.499 (ΔDrop = -31.8% and -38.5%). This is a serious weakness for bias detection, because social naming practices change and new stigmatizing terms appear after training. MARS-Gov is much more stable, keeping DGDB-Unseen F1 at 0.823–0.857 with ΔDrop between -3.6% and -2.4%.

❸ Improving the Precision-Recall Balance. Grok-3 gives high zero-shot recall (0.920) but high OGR (19.1%), showing that recall alone can produce excessive intervention. MARS-Gov (Gemini 2.5 Pro) keeps recall high (0.895), raises precision to 0.865, and lowers OGR to 2.5%. DeepSeek-R1 gives the highest MARS-Gov precision (0.872) and the lowest DGDB-Unseen drop (-2.4%).

Overall, the main gain comes from routing legal evidence and standpoint-specific reports into adjudication. Every backbone benefits more from MARS-Gov than from Standard Debate or SA-RAG, and the best configuration improves BGS to 0.808 while keeping unnecessary intervention low.

4.3 Controlled Open-set Evaluation with Held-out Categories (RQ2)

Open-set evaluation. DGDB-Unseen tests lexical robustness: whether a detector remains stable when familiar bias terms are replaced by unseen expressions. It does not show whether MARS-Gov can recover a bias category that is absent from the fixed juror panel. To test this stronger open-set claim, we design a leave-one-juror-out experiment over all 9 categories with Gemini 2.5 Pro; the setup is as follows.

Design. Each DGDB category is associated with a fixed juror and watchlist. We remove one juror–watchlist pair at a time and evaluate the corresponding test subset, excluding both the specialist and its lexical cues. We encode the Scout's predicted group and rationale together with fixed category prototypes using `clips/e5-base-trm-nl`, and rank the categories by cosine similarity. Correct@k indicates whether the held-out category appears among the top k predictions. The prototypes are fixed before evaluation and are not tuned by category; Appendix C provides the full texts.

Results. Table 2 shows that, without the Scout, the

Config.	Metric	B1 (Disability)	B2 (Migration)	B3 (Colonialism)	B4 (Religion)	B5 (Gender)	B6 (LGBTQ+)	B7 (Culture)	B8 (Education/Class)	B9 (Welfare)
w/o Scout	Rec.	0.794 ±.027	0.814 ±.024	0.778 ±.031	0.776 ±.028	0.821 ±.023	0.788 ±.026	0.807 ±.024	0.764 ±.029	0.812 ±.022
	F1	0.804 ±.021	0.826 ±.018	0.794 ±.025	0.789 ±.022	0.831 ±.017	0.806 ±.019	0.813 ±.018	0.782 ±.023	0.819 ±.017
w/ Scout (Ours)	Rec.	0.882 ±.018	0.901 ±.015	0.874 ±.021	0.871 ±.019	0.894 ±.014	0.881 ±.018	0.889 ±.016	0.866 ±.021	0.896 ±.015
	F1	**0.873** ±.015	**0.891** ±.012	**0.867** ±.017	**0.864** ±.015	**0.886** ±.011	**0.872** ±.014	**0.881** ±.013	**0.854** ±.017	**0.884** ±.012
	Correct@1 (%)	83.4 ±2.4	87.6 ±1.9	82.7 ±2.7	86.5 ±2.1	88.9 ±1.7	84.3 ±2.2	81.6 ±2.6	83.2 ±2.5	87.4 ±1.8
	Correct@3 (%)	92.5 ±1.5	96.3 ±0.9	91.8 ±1.7	94.7 ±1.2	95.6 ±1.1	93.4 ±1.4	92.7 ±1.6	92.1 ±1.6	95.2 ±1.1

Table 2: Leave-One-Category-Out (LOCO) evaluation on Gemini 2.5 Pro. Best F1 scores are bold. *Significance:* all w/ Scout gains over w/o Scout remain significant after Bonferroni correction over 9 categories (two-sided paired t-test over 3 runs; max corrected $p = 2.4\times10^{-2}$).

Configuration	Detection Quality Prec.	Rec.	F1	Governance BGS	Efficiency Tokens/Doc
Full MARS-Gov	**0.865** ±.015	**0.895** ±.013	**0.880** ±.012	**0.808** ±.013	**2,450** ±58
w/o Verify Loop	0.865 ±.015	0.895 ±.013	0.880 ±.012	0.646 ±.019 (↓ 20.0%)**	2,100 ±49
w/o Reasoning	0.842 ±.019	0.870 ±.017	0.856 ±.016	0.765 ±.021 (↓ 5.3%)*	2,380 ±62
w/o NEC	0.802 ±.021	0.847 ±.019	0.824 ±.018	0.710 ±.022 (↓ 12.1%)**	2,250 ±54
w/o Scout (10th)	0.851 ±.017	0.840 ±.019	0.845 ±.015	0.760 ±.018 (↓ 5.9%)*	2,150 ±44
w/o ARG	0.865 ±.015	0.895 ±.013	0.880 ±.012	0.808 ±.013	6,100 ±131 (↑ 149%)

Table 3: Ablation on DGDB with Gemini 2.5 Pro, averaged over 3 runs. *Significance:* marked BGS drops vs. Full MARS-Gov remain significant after Bonferroni correction (two-sided paired t-test; $^{**}p < 0.01$, $^{*}p < 0.05$; max corrected $p = 2.7\times10^{-2}$). w/o ARG is cost-only.

remaining jurors achieve an average F1 of 0.807: they often detect broad negative framing but fail to identify the held-out category. Scout-guided dynamic review raises average F1 to 0.875 and recovers the held-out category with 85.1% Correct@1 and 93.8% Correct@3. The gains are pronounced for categories with indirect cues, including Colonialism and Education/Class. This distinction matters for rewriting: a neighboring juror may flag the sentence as biased, but recovering the held-out standpoint provides the category-specific basis for revision.

4.4 Ablation & Efficiency Analysis (RQ3–RQ4)

Detection modules. Table 3 shows that Scout-guided dynamic juror review prevents static-juror consensus errors (F1 ↓ 0.035 without it and recall ↓ 5.5%). Removing NEC reduces F1 to 0.824, showing that legal grounding helps separate neutral bureaucratic phrasing from normative violation, especially in cases that are lexically neutral but procedurally stigmatizing.

Governance & efficiency. Removing verification drops BGS from 0.808 to 0.646 and inflates BMR to 98%, because rewrites are no longer checked against the same detector.[5] ARG filters about 65% of documents as CLEAR, reducing token use by roughly 60%. Figure 4 shows a stable elbow at $\tau = 2$, where over-governance falls with little recall loss across backbones.

[5]The ARG hyperparameter grid search is in Appendix F.

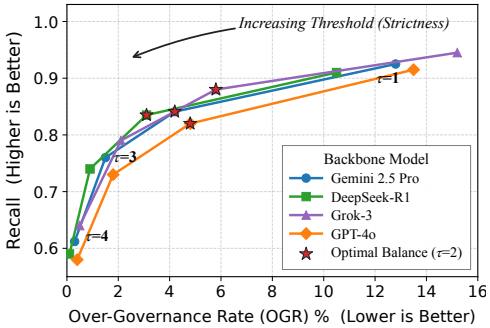

Figure 4: Recall vs. Over-Governance Rate (OGR) trade-off as threshold τ increases.

Configuration	Calls	Lat. (s)	Clear%	$/1k
Baselines (Gemini 2.5 Pro)				
Zero-shot LLM	1.0	1.8 (0.3)	—	0.85
Standard Debate	4.0	7.2 (0.9)	—	4.62
SA-RAG	1.0	3.4 (0.5)	—	2.46
MARS-Gov (with ARG)				
GPT-4o	12.6 (0.6)	20.5 (2.3)	64.2 (1.7)	10.04
Grok-3	12.8 (0.7)	18.7 (2.1)	63.5 (1.9)	13.39
Gemini 2.5 Pro	**12.4 (0.5)**	**19.2 (2.2)**	**64.8 (1.6)**	**7.35**
DeepSeek-R1	12.5 (0.5)	31.6 (3.2)	65.1 (1.8)	4.25
w/o ARG (Gemini)	28.6 (0.8)	41.5 (3.7)	—	18.30

Table 4: Per-sentence inference cost on DGDB (1,000 sentences, 3 runs; mean with σ). *Calls* counts all backbone calls; *Clear%* is the early-stop share.

Table 4 reports cost. The first-pass detector uses NEC, the Scout, and nine fixed jurors; dynamic jurors are called only for uncovered targets, one per target, and the rule-based gate adds no LLM call. With Gemini 2.5 Pro, ARG clears 64.8% of inputs early and reduces calls from 28.6 to 12.4, latency from 41.5s to 19.2s, and cost from $18.30 to $7.35 per 1,000 sentences. Since removing the gate leaves detection unchanged but roughly doubles cost, ARG acts as an efficiency component rather than a hidden classifier.

Configuration	DALC (Dutch, Social Media)			SBIC (English, Social Media)			KoBBQ (Korean, General)		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
Zero-shot LLMs	0.513 ±.028	0.873 ±.021	0.646 ±.018	0.486 ±.033	0.844 ±.023	0.617 ±.019	0.451 ±.036	0.803 ±.027	0.578 ±.024
Standard Debate	0.607 ±.025	0.851 ±.020	0.709 ±.019	0.583 ±.027	0.819 ±.021	0.681 ±.021	0.553 ±.029	0.811 ±.024	0.658 ±.023
SA-RAG	0.653 ±.021	0.862 ±.018	0.743 ±.016	0.626 ±.023	0.836 ±.019	0.716 ±.018	0.593 ±.027	0.821 ±.021	0.689 ±.019
Ours: MARS-Gov	**0.823** ±.015	**0.886** ±.013	**0.853** ±.012	**0.803** ±.018	**0.847** ±.016	**0.824** ±.015	**0.776** ±.021	**0.829** ±.018	**0.802** ±.017

Table 5: Generalization on DALC, SBIC, and KoBBQ using Gemini 2.5 Pro. Results average 3 runs; best F1 scores are bold. *Significance:* MARS-Gov beats SA-RAG after Bonferroni correction over 3 datasets (two-sided paired t-test; max corrected $p = 1.1 \times 10^{-2}$).

4.5 Cross-Domain and Cross-Lingual Generalization (RQ5)

We test transfer on DALC, SBIC, and KoBBQ using balanced 1,000-instance samples and Gemini 2.5 Pro. Table 5 shows that MARS-Gov remains strongest across all three datasets. The gain mainly comes from precision: the review loop rejects unsupported group-risk inferences left by zero-shot detection, adding 20–30 precision points while keeping recall comparable. Absolute F1 remains below in-domain DGDB because the external benchmarks shift genre, label assumptions, and language.

These results support the portability of the review procedure, not of Dutch law itself: deployment in another jurisdiction would still need a local legal and policy evidence base.

5 Conclusion

We presented MARS-Gov, a standpoint-aware framework for closed-loop bias governance in Dutch government documents. Rather than treating bias detection as one-shot classification over a fixed taxonomy, MARS-Gov combines legal retrieval, open-set target screening, juror-based adjudication, conservative routing, and rewrite verification. Experiments show stronger performance than fine-tuned, zero-shot, and generic debate baselines, with higher open-set recovery and BGS, better cross-domain and cross-lingual transfer, and lower over-governance. These results suggest that reliable governance NLP requires more than stronger classifiers: it requires legally grounded evidence, explicit standpoint reasoning, and verification after interventions. The DGDB-Unseen results sharpen this point: compact encoders remain competitive on familiar terms, but lose reliability when public language shifts. Future work will extend the framework to additional jurisdictions and reduce inference cost for broader evaluation and lower-cost deployment.

Limitations

Our framework remains dependent on domain-specific normative resources. Although the external experiments suggest that the core mechanism can transfer across languages and domains, MARS-Gov's strongest setting is Dutch administrative text because its evidence base is built around Dutch anti-discrimination law and administrative guidance. Applying the framework to other jurisdictions or platforms therefore requires localized legal and policy resources, not prompt translation alone; otherwise, the system may produce justifications that are linguistically plausible but normatively incorrect.

A second limitation is inference cost. The multi-agent pipeline is more expensive and slower than lightweight discriminative models, even with the routing gate. This cost is acceptable for high-stakes public-sector review, where false positives and false negatives can both have institutional consequences, but it limits use in high-throughput or real-time settings. Reducing the number of model calls without weakening legal grounding or verification remains an important direction for future work.

Finally, BGS is an aggregate governance metric rather than a complete diagnostic. Because $\mathrm{BGS} = \mathrm{Recall} \times \mathrm{BMR} \times \overline{\mathrm{SF}}$ (Eq. 14), different systems can obtain similar scores through different trade-offs between detection coverage, mitigation success, and meaning preservation. We therefore report the individual components alongside BGS in Table 1. BGS should be interpreted as a compact summary of governance performance, not as a substitute for component-level analysis.

Ethics Statement

This work addresses a socially beneficial but high-stakes NLP setting: identifying potentially biased language in Dutch government documents and proposing minimally edited alternatives. MARS-Gov is not intended to act as an autonomous decision-maker or to automatically rewrite official

records. In any real deployment, it should be used only as a human-in-the-loop decision-support system: it may flag potentially problematic passages, present evidence and rationales, and suggest candidate rewrites for review by qualified experts such as legal professionals, policy officers, or ethics reviewers. The original text must remain preserved in the auditable record, especially in historical or contested cases, so that affected individuals retain access to recourse and institutions remain accountable for past wording. We also do not claim that neutralized wording eliminates discriminatory practice. A rewrite should be treated as a signal for substantive institutional review, including review of policy rules, risk indicators, and administrative procedures.

The normative resources used by MARS-Gov, including watchlists, criteria, and protected-group standpoints, are not fixed truths. They should be transparently documented, versioned, periodically audited, and revised with input from affected communities, domain experts, and legal stakeholders. Since LLMs can reproduce social stereotypes, model confidence should not be treated as correctness, and high-risk outputs require manual verification, error analysis, and monitoring for disparate impact.

Finally, the system has dual-use risks. An unrestricted rewriting tool could be misused to sanitize discriminatory policy language without changing the underlying policy. For this reason, we release the code and prompts as research artifacts rather than as a turnkey tool for live administrative deployment, and we do not recommend use in fully automated government decision pipelines.

Acknowledgments

This work was supported by the National College Student Innovation and Entrepreneurship Training Program under Project No. 202619145026; the Open Project Program of Marine Ecological Restoration and Smart Ocean Engineering Research Center of Hebei Province under Grant No. HBMESO2507; the Shijiazhuang–Northeastern University Science and Technology Cooperation Special Project under Grant No. NEUS2025-01-003; and the Scientific Research Project of Hebei Education Department under Grant No. QN2024167.

References

Gordon W. Allport. 1954. *The Nature of Prejudice*. Addison-Wesley.

Ruha Benjamin. 2019. *Race After Technology: Abolitionist Tools for the New Jim Code*. Polity.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Tommaso Caselli, Arjan Schelhaas, Marieke Weultjes, Folkert Leistra, Hylke Van Der Veen, Gerben Timmerman, and Malvina Nissim. 2021. DALC: the Dutch abusive language corpus. In *Proceedings of the 5th Workshop on Online Abuse and Harm (WOAH 2021)*, pages 54–66. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. LexGLUE: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland. Association for Computational Linguistics.

Kimberlé Crenshaw. 1989. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum*.

Milena de Swart, Floris den Hengst, and Jieying Chen. 2025. Detecting linguistic bias in government documents using large language models. In *WWW 2025: Proceedings of the ACM on Web Conference 2025*, pages 5034–5044. Association for Computing Machinery.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. Bertje: A dutch bert model. *arXiv preprint arXiv:1912.09582*.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. RobBERT: a dutch RoBERTa-based language model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3255–3265, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long and Short Papers)*, Minneapolis, Minnesota. Association for Computational Linguistics.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73. ACM.

Robert M. Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of Communication*.

Virginia Eubanks. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press.

Norman Fairclough. 1992. *Discourse and Social Change*. Polity Press, Cambridge, UK.

Susan T. Fiske. 1998. Stereotyping, prejudice, and discrimination. In Daniel T. Gilbert, Susan T. Fiske, and Gardner Lindzey, editors, *The Handbook of Social Psychology*, 4 edition, pages 357–411. McGraw-Hill.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md. M. Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Rui Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics.

Erving Goffman. 1974. *Frame Analysis: An Essay on the Organization of Experience*. Harper & Row.

Yuxiang He, Jian Zhao, Yuchen Yuan, Tianle Zhang, Wei Cai, Haojie Cheng, Ziyan Shi, Ming Zhu, Haichuan Tang, Chi Zhang, and Xuelong Li. 2025. Aetheria: A multimodal interpretable content safety framework based on multi-agent debate and collaboration. *Preprint*, arXiv:2512.02530.

Dan Hendrycks and Kevin Gimpel. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*.

Sirui Hong, Mingchen Zhuge, Jiaqi Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2024. Metagpt: Meta programming for a multi-agent collaborative framework. In *International Conference on Learning Representations*.

Wonje Jeung, Dongjae Jeon, Ashkan Yousefpour, and Jonghyun Choi. 2025. Large language models still exhibit bias in long text. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26147–26169, Vienna, Austria. Association for Computational Linguistics.

Jiho Jin, Jiseon Kim, Nayeon Lee, Haneul Yoo, Alice Oh, and Hwaran Lee. 2024. KoBBQ: Korean bias benchmark for question answering. *Transactions of the Association for Computational Linguistics*, 12:507–524.

Tian Lan, Jiang Li, Yemin Wang, Xu Liu, Xiangdong Su, and Guanglai Gao. 2025. F²Bench: An open-ended fairness evaluation benchmark for LLMs with factuality considerations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2031–2046, Suzhou, China. Association for Computational Linguistics.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for "mind" exploration of large language model society. *arXiv preprint arXiv:2303.17760*.

Marta Marchiori Manerba, Karolina Stanczak, Riccardo Guidotti, and Isabelle Augenstein. 2024. Social bias probing: Fairness benchmarking for language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14653–14671, Miami, Florida, USA. Association for Computational Linguistics.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. Stereoset: Measuring stereotypical bias in pretrained language models. In *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)*. Association for Computational Linguistics.

Yucheng Ning, Xixun Lin, Fang Fang, and Yanan Cao. 2025. Mad-fact: A multi-agent debate framework for long-form factuality evaluation in llms. *Preprint*, arXiv:2510.22967.

Safiya Umoja Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press.

Yulia Otmakhova, Shima Khanehzar, and Lea Frermann. 2024. Media framing: A typology and survey of computational approaches across disciplines. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15407–15428, Bangkok, Thailand. Association for Computational Linguistics.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, Samuel R. Bowman, and Alex Warstadt. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*. Association for Computational Linguistics.

Nicholas Pipitone and Ghita Houir Alami. 2024. LegalBench-RAG: A benchmark for retrieval-augmented generation in the legal domain. *arXiv preprint arXiv:2408.10343*.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. Chatdev: Communicative agents for software development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Arti Rani, Shweta Singh, Nihar Ranjan Sahoo, and Gaurav Kumar Nayak. 2025. Open-DeBias: Toward mitigating open-set bias in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 25027–25051, Suzhou, China. Association for Computational Linguistics.

Philip Resnik. 2025. Large language models are biased because they are large language models. *Computational Linguistics*, 51(3):885–906.

Aya Rizk and Ida Lindgren. 2025. Automated decision-making in public administration: Changing the decision space between public officials and citizens. *Government Information Quarterly*, 42(3):102061.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Walter J. Scheirer, Anderson Rocha, Archie Sapkota, and Terrance E. Boult. 2013. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1757–1772.

Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. 2022. "I'm sorry to hear that": Finding new biases in language models with a holistic descriptor dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9211. Association for Computational Linguistics.

Teun A. van Dijk. 1993. Principles of critical discourse analysis. *Discourse & Society*.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*. Association for Computational Linguistics.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. 2023. Autogen: Enabling next-gen LLM applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics.

Appendix Index

1. Dataset Overview (p. 12)
2. Extended Benchmarks across Model Families (p. 12)
   - Analysis of Extended Results (p. 13)
   - Cross-Lingual Reasoning (p. 13)
3. LOCO Prototypes (p. 13)
   - Fixed Prototype Texts for Deterministic Mapping (p. 13)
4. Representative End-to-End Decision Traces (p. 14)
5. Knowledge Base Construction and RAG Strategies (p. 15)
   - DGDB NEC Retrieval Setting (p. 15)
   - Knowledge Base Composition (p. 15)
   - Implementation Parameters (p. 15)
6. ARG Hyperparameter Sensitivity Grid (p. 15)
   - Performance Heatmaps (F1 vs. OGR) (p. 15)
7. English and Dutch Prompt Templates (p. 16)

A Dataset Overview

Tables 6 and 7 summarize the datasets used in the experiments.

B Extended Benchmarks across Model Families

This appendix tests whether MARS-Gov's gains depend on a single frontier backbone. We evaluate

Statistics	DGDB	DGDB-Unseen
Language	Dutch	Dutch
Domain	Gov. docs	Gov. docs
Unit	Sentence	Sentence
Corpus size	3,747	DGDB-derived
Evaluation size	Test split	Held-out subset
Construction	Original split	Bias-term substitution
Evaluation balance	Original split	Label-preserving
Targets / labels	9 bias cats.	Same 9 bias cats.

Table 6: Statistical overview of the DGDB-derived datasets.

Statistics	DALC	SBIC	KoBBQ
Language	Dutch	English	Korean
Domain	Twitter	Social	QA
Unit	Tweet	Post / tuple	MCQ
Corpus size	8,156	44,671 / 147,139	76,048
Evaluation size	1,000	1,000	1,000
Evaluation balance	500 / 500	500 / 500	500 / 500
Targets / labels	Abusive binary	Bias tuple	Bias contexts

Table 7: Statistical overview of the external evaluation datasets.

18 LLMs, grouped into *Frontier SOTA*, *Strong Generalist*, and *Mid/Efficiency* tiers. Table 8 reports the corresponding zero-shot and MARS-Gov results on DGDB.

B.1 Analysis of Extended Results

The extended results show broad gains across backbones, but the final ceiling still depends on reasoning capacity. Models at or above the GPT-4 Turbo / Llama-3-70B tier reliably exceed the fine-tuned BERTje baseline, including open-weight options such as *Qwen-2.5-72B* and *DeepSeek-V3*.

B.2 Cross-Lingual Reasoning

Native vs. English Reasoning. Figure 5 compares Dutch prompting with English prompting after DeepL translation. All scores are averaged over 3 independent runs.

Dutch prompting gives a small but consistent F1 advantage across all four backbones ($+0.8\%$ to $+2.3\%$), mainly because administrative terms such as *"afstand tot de arbeidsmarkt"* and *"leefvorm"* lose part of their meaning after translation.

C LOCO Prototypes

To support reproducibility of the deterministic open-set evaluation, we provide the full fixed prototype texts $e(c)$ used for cosine-similarity mapping. LOCO uses `clips/e5-base-trm-nl` for prototype ranking, whereas NEC uses the separate retrieval embedding model described in Appendix E.3.

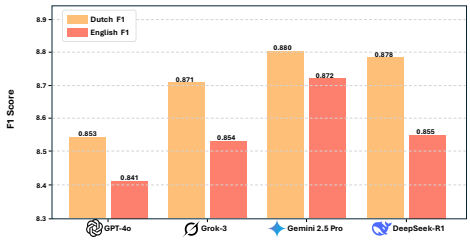

Figure 5: Comparison of F1 scores between Native Dutch and Translated English prompts. Native prompting consistently outperforms English across all backbones, avoiding the semantic loss associated with translating specific administrative terminology ($p < 0.05$, paired Student's t-test over 3 independent runs).

C.1 Fixed Prototype Texts for Deterministic Mapping

The following prototype vectors were fixed before the experiment based on dataset statistics and definitions. The Scout's generated `group` and `rationale` were mapped against these to calculate Correct@1 and Correct@3.

- **B1 (Disability):** *"beperkingen, handicap, rolstoelgebonden, dwerg, doventolk, medische toestand, fysieke of mentale beperking."*
- **B2 (Migration):** *"migratie, allochtoon, vluchteling, nieuwkomer, asielzoeker, integratie, westers, niet-westers."*
- **B3 (Colonialism):** *"kolonialisme, slavernijverleden, inheems, etnisch, blank, zwart, ras, exotisch, primitief."*
- **B4 (Religion):** *"geloof, religie, christen, moslim, islam, joods, hoofddoek, extremisme."*
- **B5 (Gender):** *"gender, geslacht, man, vrouw, meisje, jongen, seksisme, traditionele rolpatronen."*
- **B6 (LGBTQ+):** *"seksualiteit, geaardheid, homo, hetero, queer, transgender, non-binair, lhbtiq+."*
- **B7 (Culture):** *"cultuur, traditie, bi-cultureel, tussenpositie, culturele achtergrond, minderheden."*
- **B8 (Education/Class):** *"onderwijs, klasse, achterstandsleerling, mbo, laagopgeleid, probleemwijk, sociale status."*
- **B9 (Welfare):** *"algemeen, armoede, machtsongelijkheid, fobie, sociaal-economische kwetsbaarheid."*

Model Family	Model Name	Standard Zero-shot Prec	Standard Zero-shot Rec	Standard Zero-shot F1	Ours: MARS-Gov Prec	Ours: MARS-Gov Rec	Ours: MARS-Gov F1	Governance BGS
Frontier SOTA	**Gemini 2.5 Pro**	0.524 ±.024	0.907 ±.018	0.664 ±.016	**0.865** ±.015	0.895 ±.013	**0.880** ±.012	**0.808** ±.013
	DeepSeek-R1	0.549 ±.023	0.887 ±.019	0.678 ±.015	**0.872** ±.016	0.884 ±.014	0.878 ±.013	0.806 ±.015
	Claude 3.5 Sonnet	**0.562** ±.022	0.872 ±.020	**0.683** ±.016	0.860 ±.017	0.882 ±.015	0.871 ±.013	0.787 ±.016
	Grok-3	0.484 ±.031	**0.920** ±.017	0.634 ±.019	0.836 ±.019	**0.910** ±.013	0.871 ±.014	0.779 ±.018
	GPT-4o	0.514 ±.025	0.882 ±.021	0.649 ±.017	0.844 ±.017	0.863 ±.015	0.853 ±.013	0.757 ±.014
Strong Generalist	Claude 3 Opus	0.531 ±.026	0.872 ±.021	0.660 ±.018	0.852 ±.018	0.867 ±.016	0.859 ±.014	0.749 ±.018
	GPT-4 Turbo	0.494 ±.031	0.862 ±.020	0.628 ±.021	0.836 ±.019	0.852 ±.018	0.844 ±.015	0.717 ±.017
	DeepSeek-V3	0.482 ±.033	0.846 ±.023	0.614 ±.020	0.811 ±.022	0.832 ±.019	0.821 ±.017	0.682 ±.021
	Qwen-2.5-72B	0.472 ±.036	0.852 ±.024	0.608 ±.023	0.819 ±.021	0.839 ±.019	0.829 ±.016	0.696 ±.019
	Mistral Large 2	0.491 ±.034	0.828 ±.026	0.616 ±.022	0.806 ±.022	0.826 ±.019	0.816 ±.018	0.654 ±.022
	Llama-3-70B	0.463 ±.037	0.837 ±.024	0.596 ±.023	0.800 ±.023	0.820 ±.020	0.810 ±.017	0.668 ±.023
	— Performance Threshold: Fine-tuned BERTje (F1 = 0.812) —							
Mid/Efficiency	Gemini 1.5 Pro	0.478 ±.038	0.821 ±.027	0.604 ±.024	0.781 ±.024	0.798 ±.021	0.789 ±.018	0.612 ±.022
	Qwen-2.5-32B	0.422 ±.041	0.798 ±.028	0.552 ±.026	0.742 ±.026	0.762 ±.023	0.752 ±.019	0.575 ±.024
	Claude 3 Haiku	0.412 ±.043	0.778 ±.031	0.539 ±.029	0.718 ±.027	0.738 ±.023	0.728 ±.021	0.539 ±.027
	Gemini 1.5 Flash	0.401 ±.042	0.789 ±.030	0.532 ±.027	0.703 ±.027	0.723 ±.025	0.713 ±.021	0.521 ±.027
	Mixtral 8x7B	0.382 ±.045	0.762 ±.033	0.509 ±.031	0.671 ±.029	0.694 ±.026	0.682 ±.023	0.486 ±.029
	GPT-3.5 Turbo	0.352 ±.048	0.748 ±.036	0.479 ±.033	0.633 ±.032	0.654 ±.027	0.643 ±.024	0.448 ±.031
	Llama-3-8B	0.321 ±.051	0.712 ±.037	0.442 ±.035	0.591 ±.033	0.611 ±.029	0.601 ±.026	0.397 ±.033

Table 8: **Extended Leaderboard on DGDB (In-Domain).** Models are ranked by their MARS-Gov F1 score. The red dashed line marks the performance threshold of the fine-tuned BERTje baseline ($F1 = 0.812$). **Observation:** While MARS-Gov improves performance across model families (avg. Δ F1 $+0.20$, larger for frontier backbones than for smaller ones), a minimum reasoning capability threshold (approx. GPT-4 Turbo level) is required to surpass supervised baselines.

D Representative End-to-End Decision Traces

We provide three representative traces that illustrate the main decision paths: severe intersectional bias, false inference from language proficiency, and a neutral procedural case. These traces explain the decision mechanism; aggregate detection and ablation results are reported in Tables 1 and 3. The full prompt definitions used to produce these traces are available in `agent_bias/common/prompts.py`.

Case T1: Intersectional welfare-fraud profiling

Input. *"Om uitkeringsfraude tegen te gaan, moet de gemeente strengere huisbezoeken uitvoeren bij allochtone alleenstaande moeders in de Schilderswijk, aangezien deze groep historisch gezien een gebrek aan administratief inzicht heeft."*

Trace. The evidence layer retrieves anti-discrimination and profiling guidance. The Scout identifies migration, gender, and neighborhood/class proxies. Jurors for migration, gender, class, and welfare flag the sentence because it links protected and socio-economic attributes to fraud risk and administrative incapacity.
Decision. The Judge returns BIASED; the rewrite preserves the fraud-control goal but replaces group targeting with objective risk indicators.

Case T2: Language proficiency as intent inference

Input. *"De aanvrager toonde tijdens het gesprek een gebrekkige beheersing van de Nederlandse taal; dit duidt op een gebrek aan integratie en een verhoogde kans op opzettelijke misleiding."*

Trace. The Scout identifies non-native speakers as the affected group. Jurors flag the unsupported causal chain from language ability to integration failure and intentional deception.
Decision. The Judge returns BIASED; the rewrite reframes the observation as a communication-support need while preserving the administrative fact.

Case T3: Objective income threshold

Input. *"Om in aanmerking te komen voor de huurtoeslag, mag het gezamenlijke toetsingsinkomen niet hoger zijn dan €32.500 per jaar."*

Trace. The evidence layer identifies a statutory eligibility threshold. The Scout finds no protected or stigmatized group target beyond the general applicant population. Jurors treat the sentence as a neutral criterion.
Decision. The Judge returns NON-BIASED, and restoration is skipped.

E Knowledge Base Construction and RAG Strategies

This appendix summarizes the NEC knowledge base $\mathcal{K}$ and retrieval settings used by MARS-Gov.

E.1 DGDB NEC Retrieval Setting

For DGDB, NEC retrieval uses the source inventory released with the code. It combines legal and policy foundations, accessibility and administrative communication guidance, reports on online discrimination and institutional racism, inclusive-language resources, and DGDB category reasoning notes.

E.2 Knowledge Base Composition

The curated repository $\mathcal{K}$ is structured into three functional layers:

- **Legal and Policy Core (Grondslag):** *Grondwet* Chapter 1, AWGB core articles, OCW antidiscrimination agenda material, the Dutch national programme against discrimination and racism, and ECRI Netherlands extracts.
- **Administrative and Accessibility Guides (Overheidsrichtlijnen):** accessible voting plans, election communication guidance, and official inclusive and accessible writing guides.
- **Sociocultural and Inclusive-Language Resources (Inclusiviteit):** online-discrimination and harmful-behavior reports, institutional-racism summaries, emancipation monitoring, DEBIAS vocabulary entries, inclusive communication glossaries, cultural-sector language guidance, and DGDB reasoning entries.

Table 9 summarizes how this inventory supports the DGDB retrieval setting.

E.3 Implementation Parameters

NEC utilizes `text-embedding-3-small` for RAG dense retrieval. This is separate from the `clips/e5-base-trm-nl` encoder used for semantic fidelity and LOCO prototype ranking. The released implementation first retrieves keyword matches, then dense semantic matches, and finally expands selected hits to adjacent chunks from the same source. Long policy documents use per-source character windows, while DGDB reasoning entries are split by markdown headings. At inference, the default setting selects $k = 5$ seed fragments before local context expansion, ensuring the Summary Agent synthesizes a contextually grounded, neutral evidence base.

Dimension	DGDB NEC retrieval
Primary Objective	Detecting implicit bias in official policy texts.
Core Domain	Legal and political administrative text (*high-stakes*).
Inventory	Dutch constitutional and equal-treatment law; antidiscrimination policy; accessibility and election communication guidance; online-discrimination, institutional-racism, emancipation, and inclusive-language resources; DGDB reasoning entries.
Selection Rationale	Grounds the review in legal norms while covering administrative, accessibility, and inclusive-language contexts that commonly shape implicit bias.

Table 9: **NEC retrieval construction for DGDB.** MARS-Gov prioritizes normative sources and official writing guidance for administrative bias review.

The three retrieval stages are complementary: keyword matching preserves precise legal and administrative terminology, dense retrieval recovers paraphrases, and local expansion restores nearby definitions or caveats. Source identifiers and chunk positions are retained in the decision trace, allowing the retrieved evidence to be inspected independently of the verdict. The Summary Agent condenses this material but cannot classify the sentence, keeping retrieval errors separate from later reasoning errors.

F ARG Hyperparameter Sensitivity Grid

We tune ARG on the validation set with Gemini 2.5 Pro. Five parameters govern escalation: τ_{high}=0.6 immediately flags `HARD` cases, $\tau_{\text{high_risk}}$=0.7 applies a stricter watchlist threshold, τ_{mid}=0.7 triggers `SOFT` jury review, $N_{\text{soft_bias}}$=2 confirms bias in `SOFT` mode, and $N_{\text{soft_poss}}$=1 marks possible bias for manual review. The heatmap varies the two most influential parameters: τ_{high} and $N_{\text{soft_bias}}$.

F.1 Performance Heatmaps

Figure 6 illustrates the interaction between these ARG parameters.

Selected Point. Low thresholds bypass the jury too often, while high thresholds increase calls without improving F1. The selected setting ($\tau_{\text{high}} = 0.6, N_{\text{soft_bias}} = 2$) reaches F1 $= 0.880$ with OGR $= 2.5\%$.

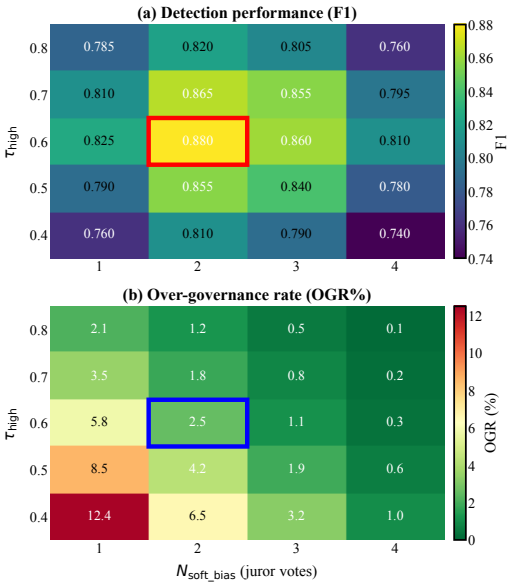

Figure 6: **ARG Hyperparameter Grid Search.** (a) F1 Score and (b) Over-Governance Rate (OGR) across varying τ_{high} and juror-vote thresholds. The selected configuration ($\tau_{\text{high}} = 0.6, N_{\text{soft_bias}} = 2$) is highlighted.

G English and Dutch Prompt Templates

This appendix gives the English and Dutch prompt templates for the main MARS-Gov agents. The templates are formatted for readability; placeholders such as `{sentence}`, `{evidence}`, and `{jury_outputs}` denote runtime fields. Prompt boundaries mirror the pipeline stages and restrict each agent to a specific role and output contract. Evidence summarization, standpoint discovery, adjudication, restoration, and verification therefore remain separately inspectable. Malformed outputs are exposed to the routing logic rather than interpreted heuristically.

G.1 English Prompt Templates

P1. NEC Summary Agent

System role. You are a context summarizer for governance text screening. Select and summarize only the context fragments that are directly relevant to the input sentence. Be objective and factual: no verdict, no diagnosis, no classification, and no interpretation of intent. Use only information that follows literally or unambiguously from the supplied context.

User task. Given `{query}` and retrieved `{context}`, produce a short neutral evidence summary before any bias decision is made.
Output contract. Return **3–6 bullets**. Each bullet must contain one paraphrased context fragment and one neutral relevance label: *definition*, *criterion*, *guideline*, *example*, *caveat*, *scope*, or *terminology*. If no relevant evidence exists, return exactly one bullet stating that no sufficiently relevant context was found. Do not mention source names, links, file names, or citations.

P2. Scout Agent for Open-Set Standpoint Screening

System role. Identify which social groups may be referenced or affected by the sentence. Focus on groups that are explicitly mentioned or strongly implied, especially groups outside the fixed nine-category panel. Do not judge whether the sentence is biased; only detect and describe affected groups.
Known panel. B1 disability/functional limitations; B2 migration/refugees; B3 colonial history/slavery narratives; B4 religion, especially Islam; B5 gender and gender roles; B6 LGBTQ+; B7 cultural or ethnic groups; B8 education/class; B9 welfare and social vulnerability.
Policy-language cues. Treat context-sensitive expressions as possible group markers when they function that way in administrative text. Examples include *parents* in education/youth-policy contexts, *background* or *minorities* for cultural/ethnic groups, religious identifiers such as *Islam* or *headscarf*, and migration terms such as *newcomer* or *non-Western*.

User task. For `{sentence}`, return exactly one JSON object:

```
{ "detected_groups": [
{ "group_label": "...",
"is_known_category": true/false,
"mapped_jury": "B1|...|B9|none",
"evidence": ["..."],
"x_prompt_hint": "..." } ] }
```

Constraints. Use short generic group labels; add multiple objects when multiple groups are implicated; write `x_prompt_hint` only for a newly surfaced group; return an empty list when no relevant group is found.

P3. Juror Agents

Common user prompt. Each juror receives `{sentence}` and the NEC evidence summary `{evidence}`. The juror returns exactly one JSON object with `agent_id`, `group_labels`, `label`, `confidence`, `bias_signals`, `analysis`, and `risk_score`.
Label contract. **Biased** means negative framing, stereotyping, exclusion, or dehumanization directed at the juror's target group. **PossiblyBiased** means a concrete textual cue exists but evidence is not decisive. **Non-biased** means neutral, factual, or positive toward the target group. **Uncertain** means insufficient or ambiguous evidence.
Grounding rules. Do not label a sentence biased solely because a watchlist term appears. The juror must cite a short signal from the sentence or evidence summary. If no signal can be cited, use **Uncertain**. Calibrate `risk_score`: high for prohibited language or direct threat/criminality links, medium for clear negative framing, and low for

weak or doubtful signals.

Dynamic 10th juror template. You represent {group_label}. Flag possible open-set bias for this group even when other jurors consider the risk low. Look for subtle exclusion, stereotyping, marginalization, presupposition, neutral-sounding policy frames, implicit causality, group generalizations, and loaded labels. Use only the sentence and evidence summary; if evidence is weak, choose **Uncertain** and state what is missing. Runtime hint: {rationale}.

P4. Fixed Juror Standpoints

The fixed jurors share P3's output contract but differ in standpoint:

- **B1 Disability:** watch for disability framed as burden, cost, risk, dependency, incapacity, or exclusion from participation.
- **B2 Migration:** watch for massing metaphors, threat framing, securitization, blame, and unsupported links to disorder, abuse, or incompatibility.
- **B3 Colonialism:** watch for civilizing frames, gratitude or debt frames, agency erasure, euphemisms, and softened accounts of coercion or exploitation.
- **B4 Religion:** watch for collective blame, essentialism, links between religion and threat, and problematization of religious practice without nuance.
- **B5 Gender:** watch for stereotyped roles, male-default norms, objectification, vulnerability framing, and positive stereotypes that preserve hierarchy.
- **B6 LGBTQ+:** watch for hetero/cisnormativity, pathologization, moral-panic framing, and neutral-sounding exclusion from rights or services.
- **B7 Culture/Ethnicity:** watch for essentialist culture claims, us/them framing, exoticization, hierarchy, and ethnicity linked to threat or deviance.
- **B8 Education/Class:** watch for condescension, status hierarchy, deficit assumptions, and access rules that treat higher-status participation as default.
- **B9 Welfare:** watch for burden framing, victim blaming, criminalization, dehumanization, and treating poverty, unemployment, or homelessness as uniform stereotypes.

P5. Judge Agent

System role. Decide the final binary label for linguistic bias in a Dutch policy-style sentence: biased (1) or non-biased (0). Apply the definition strictly: a sentence is biased if and only if **Condition A** holds, or **Condition B1 and Condition B2** both hold.
Condition A: prohibited terms. The sentence contains a universally inappropriate term, unless it is a direct quotation or is explicitly used to explain or criticize harmful language.
Condition B1: contextual bias. The sentence exhibits stereotyping, exclusion, power imbalance, or prejudice, either explicit or implicit.
Condition B2: directed at a group. The bias is directed at a specific social group that is explicitly mentioned or unambiguously implied.

User task. Use only {sentence}, {evidence}, and raw {jury_outputs}. Prioritize quoted language signals and consistent multi-agent signals over vague impressions; this is not simple majority voting. Use confidence and risk scores only as weighting hints.
Output contract. Return exactly one JSON object:

```
{ "label": 0/1, "confidence": 0-1,
"rationale": "max 3 sentences; cite
decisive signals/agents" }
```

When jurors conflict, prefer quoted prohibited terms, dehumanizing or threat/burden framing, and higher-confidence analyses only when they are grounded in concrete signals.

P6. Restoration Agent

System role. Rewrite Dutch policy-style sentences to remove linguistic bias while preserving original meaning, facts, administrative register, and policy intent. Assume the sentence has already been flagged as biased upstream; the task is mitigation, not reclassification.
Edit policy. Make the smallest possible edits that neutralize the bias: replace prohibited or problematic terms, remove stigmatizing metaphors, remove stereotyping or dehumanizing framing, and use neutral or person-first wording where appropriate. Keep the output to one formal Dutch sentence. Do not add new facts, actors, numbers, causal claims, or assumptions.

User task. Given {sentence}, {evidence}, {aggregate}, {orig_label}, and {sample_id}, return exactly one JSON object:

```
{ "rewritten_text": "...",
"new_label": 0, "original_label":
{orig_label}, "id": "{sample_id}",
"notes": "1-2 key edits" }
```

Fidelity rule. Preserve all concrete facts from the original sentence unless the factual wording itself is the biased framing. The rewrite must not become an explanation, bullet list, or multi-sentence paraphrase.

G.2 Dutch Prompt Templates

D1. NEC Summary Agent

Systeemrol. Je bent een Nederlandstalige context-samenvatter voor screening van bestuursteksten. Selecteer en vat alleen contextfragmenten samen die direct relevant zijn voor de gegeven zin. Werk strikt objectief en feitelijk: geen oordeel, geen diagnose, geen classificatie en geen interpretatie van intentie. Gebruik alleen informatie die letterlijk of ondubbelzinnig uit de aangeleverde context volgt.

Gebruikerstaak. Gegeven {query} en de opgehaalde {context}, maak een korte neutrale evidence-samenvatting voordat een bias-beslissing wordt genomen.

Outputcontract. Geef **3–6 bullets**. Elke bullet bevat een geparafraseerd contextfragment en een neutraal relevantielabel: *definitie*, *criterium*, *richtlijn*, *voorbeeld*, *aandachtspunt*, *nuance/afbakening* of *terminologie*. Als er geen relevante evidence is, geef exact een bullet die meldt dat geen voldoende relevante context is gevonden. Noem geen brondata, links, bestandsnamen of citaties.

D2. Scout Agent voor Open-Set Standpoint Screening

Systeemrol. Identificeer welke sociale groepen mogelijk worden genoemd of geraakt door de zin. Focus op groepen die expliciet genoemd worden of sterk geïmpliceerd zijn, vooral groepen buiten het vaste panel van negen categorieën. Geef geen oordeel over bias; detecteer en beschrijf alleen mogelijke doelgroepen.

Vast panel. B1 beperking/functionele beperking; B2 migratie/vluchtelingen; B3 koloniale geschiedenis/slavernijnarratieven; B4 religie, vooral islam; B5 gender en genderrollen; B6 LGBTQ+; B7 culturele of etnische groepen; B8 onderwijs/klasse; B9 welzijn en sociale kwetsbaarheid.

Cues in beleidstaal. Behandel contextgevoelige termen als mogelijke groepsmarkeringen wanneer zij zo functioneren in bestuurstekst. Voorbeelden zijn *ouders* in onderwijs- of jeugdbeleid, *achtergrond* of *minderheden* voor culturele/etnische groepen, religieuze markeerders zoals *islam* of *hoofddoek*, en migratietermen zoals *nieuwkomer* of *niet-westers*.

Gebruikerstaak. Voor {sentence}, retourneer exact een JSON-object:

```
{      "detected_groups":          [
{      "group_label":          "...",
"is_known_category":      true/false,
"mapped_jury":      "B1|...|B9|none",
"evidence":                  ["..."],
"x_prompt_hint": "..." } ] }
```

Beperkingen. Gebruik korte generieke groepslabels; voeg meerdere objecten toe wanneer meerdere groepen geraakt worden; vul x_prompt_hint alleen voor een nieuw opgekomen groep; retourneer een lege lijst wanneer geen relevante groep wordt gevonden.

D3. Juror Agents

Gemeenschappelijke gebruikerstaak. Elke juror ontvangt {sentence} en de NEC evidence-samenvatting {evidence}. De juror retourneert exact een JSON-object met agent_id, group_labels, label, confidence, bias_signals, analysis en risk_score.

Labelcontract. **Biased** betekent negatieve framing, stereotypering, uitsluiting of ontmenselijking gericht op de doelgroep van de juror. **PossiblyBiased** betekent dat er een concreet taalsignaal is, maar dat het bewijs nog niet doorslaggevend is. **Non-biased** betekent neutraal, feitelijk of positief ten opzichte van de doelgroep. **Uncertain** betekent onvoldoende of ambigue informatie.

Groundingregels. Label een zin niet als biased alleen omdat een watchlistterm voorkomt. De juror moet een kort signaal uit de zin of evidence-samenvatting kunnen citeren. Als dat niet kan, kies **Uncertain**. Kalibreer risk_score: high voor verboden taal of directe dreiging/criminaliteitskoppeling, medium voor duidelijke negatieve framing, en low voor zwakke of twijfelachtige signalen.

Dynamische 10e juror-template. Jij vertegenwoordigt {group_label}. Signaleer mogelijke open-set bias voor deze groep, ook wanneer andere jurors het risico laag inschatten. Let op subtiele uitsluiting, stereotypering, marginalisering, vooronderstellingen, neutraal klinkende beleidsframes, impliciete causaliteit, groepsgeneralisaties en geladen labels. Gebruik alleen de zin en evidence-samenvatting; kies **Uncertain** wanneer evidence zwak is en benoem wat ontbreekt. Runtime-hint: {rationale}.

D4. Vaste Juror-standpunten

De vaste jurors delen het outputcontract van D3, maar verschillen in standpoint:

- **B1 Beperking:** let op beperking als last, kostenpost, risico, afhankelijkheid, onbekwaamheid of uitsluiting van participatie.
- **B2 Migratie:** let op massametaforen, dreigingsframes, securitisering, schuldtoewijzing en ongefundeerde koppelingen aan wanorde, misbruik of onverenigbaarheid.
- **B3 Kolonialisme:** let op beschavingsframes, dankbaarheids- of schuldframes, agency-uitwissing, eufemismen en verzachte beschrijvingen van dwang of uitbuiting.
- **B4 Religie:** let op collectieve schuld, essentialisme, koppelingen tussen religie en dreiging, en problematisering van religieuze praktijk zonder nuance.
- **B5 Gender:** let op stereotiepe rollen, mannelijke standaardnormen, objectificatie, kwetsbaarheidsframing en positieve stereotypen die hiërarchie bewaren.
- **B6 LGBTQ+:** let op hetero/cisnormativiteit, pathologisering, morele-paniekframing en neutraal klinkende uitsluiting van rechten of diensten.
- **B7 Cultuur/Etniciteit:** let op essentialistische cultuurclaims, wij/zij-framing, exotisering, hiërarchie en etniciteit gekoppeld aan dreiging of deviant gedrag.
- **B8 Onderwijs/Klasse:** let op neerbuigendheid, statushiërarchie, deficit-aannames en toegangseisen die hogere-statusdeelname als norm behandelen.
- **B9 Welzijn:** let op lastframing, victim blaming, criminalisering, ontmenselijking en het behandelen van armoede, werkloosheid of dakloosheid als uniform stereotype.

D5. Judge Agent

Systeemrol. Beslis het finale binaire label voor taalkundige bias in een Nederlandse beleid-achtige zin: biased (1) of non-biased (0). Pas de definitie strikt toe: een zin is biased als en alleen als **Voorwaarde A** geldt, of als **Voorwaarde B1 en Voorwaarde B2** allebei gelden.
Voorwaarde A: verboden termen. De zin bevat een universeel ongepaste term, tenzij het gaat om een direct citaat of om expliciet uitleggen/kritiseren van schadelijk taalgebruik.
Voorwaarde B1: contextuele bias. De zin bevat stereotypering, uitsluiting, machtsongelijkheid of vooroordeel, expliciet of impliciet.
Voorwaarde B2: gericht op een groep. De bias is gericht op een specifieke sociale groep die expliciet genoemd of ondubbelzinnig geïmpliceerd is.

Gebruikerstaak. Gebruik alleen `{sentence}`, `{evidence}` en ruwe `{jury_outputs}`. Geef voorrang aan geciteerde taalsignalen en consistente multi-agent signalen boven vage indrukken; dit is geen gewone meerderheidsstemming. Gebruik confidence en risk scores alleen als weegsignalen.
Outputcontract. Retourneer exact een JSON-object:

```
{ "label": 0/1, "confidence": 0-1,
"rationale": "max 3 zinnen; citeer
doorslaggevende signalen/agents" }
```

Wanneer jurors conflicteren, geef voorrang aan geciteerde verboden termen, ontmenselijkende of dreiging/last-framing, en analyses met hogere confidence alleen wanneer zij gegrond zijn in concrete signalen.

D6. Restoration Agent

Systeemrol. Herschrijf Nederlandse beleid-achtige zinnen om taalkundige bias te verwijderen met behoud van oorspronkelijke betekenis, feiten, administratief register en beleidsintentie. Ga ervan uit dat de zin eerder in de pipeline al als biased is gemarkeerd; de taak is mitigatie, niet herclassificatie.
Editbeleid. Maak de kleinst mogelijke edits die de bias neutraliseren: vervang verboden of problematische termen, verwijder stigmatiserende metaforen, verwijder stereotypering of ontmenselijkende framing, en gebruik waar passend neutrale of person-first formuleringen. Houd de output bij een formele Nederlandse zin. Voeg geen nieuwe feiten, actoren, getallen, causale claims of aannames toe.

Gebruikerstaak. Gegeven `{sentence}`, `{evidence}`, `{aggregate}`, `{orig_label}` en `{sample_id}`, retourneer exact een JSON-object:

```
{ "rewritten_text": "...",
"new_label": 0, "original_label":
{orig_label}, "id": "{sample_id}",
"notes": "1-2 kernedits" }
```

Fidelityregel. Behoud alle concrete feiten uit de oorspronkelijke zin, tenzij de feitelijke formulering zelf de biased framing is. De herschrijving mag geen uitleg, bulletlijst of meerzinnige parafrase worden.