White-Hat Testing for the Ballot Box: A Framework for Election AI Auditing

Chendi Wang[1] , **Jieying Chen**[2]
[1] Department of Political Science and Public Administration, Vrije Universiteit Amsterdam
[2] Department of Computer Science, Vrije Universiteit Amsterdam
chendi.wang@vu.nl, j.y.chen@vu.nl

Abstract

Recent research shows that conversational AI can shift voter preferences, with effects persisting for weeks. Yet frontier models exhibit a documented "persuasion–reliability tradeoff", producing hallucinated or systematically distorted election information. Despite these risks, election officials lack standardized tools to systematically evaluate AI systems before deployment. We propose *CivicAudit-Bench*, a stakeholder-guided auditing framework to stress-test large language models for civic hallucinations, false confidence, jurisdiction-dependent failure, and asymmetric refusals/accuracy. This framework introduces a modular, counterfactual, and severity-aware auditing methodology that integrates roll-call–based alignment modeling, entity-swap probing, and jurisdiction-conditional correctness criteria. Informed by engagement with the U.S. Election Assistance Commission, the toolkit consists of three modules: (1) *PoliBias-US*, a multi-indicator alignment screen combining Congressional roll-call ideology scaling with party-cue counterfactual sensitivity, persona robustness, and narrative-framing alignment; (2) *HalluBias-Election*, an evidence-linked benchmark that measures hallucinations, severity-weighted critical errors, and asymmetries via Entity-Swap Counterfactual Probing and a jurisdiction-safe completion criterion; and (3) *Disclosure-Test*, pre-registered experiments assessing whether transparency and calibrated-uncertainty disclosures reduce overreliance and attenuate persuasion without blocking legitimate civic information. CivicAudit-Bench outputs versioned audit scorecards and a coordinated white-hat disclosure workflow, advancing UN SDG 16 by strengthening democratic information integrity.

1 Problem Statement: Election AI Audit Gap

Pressing societal challenge: election information integrity in conversational AI. Democratic participation depends on timely, accurate, and accessible civic information, such as registration rules, eligibility requirements, voting modalities, polling-place lookup, and procedures for reporting and contestation. Digital intermediaries already shape political knowledge and misinformation dynamics, and the diffusion of generative AI systems adds a new layer of amplification and personalization to this information ecosystem [Vosoughi *et al.*, 2018; Jungherr, 2023]. Recent interdisciplinary syntheses argue that advanced AI systems can affect multiple democratic functions simultaneously, including epistemic integrity, citizen autonomy, political equality, and institutional legitimacy [Kreps and Kriner, 2023; Summerfield *et al.*, 2025]. In parallel, the technical community has documented that large language models can produce fluent outputs that are partially or wholly unsupported by evidence ('hallucinations'), and that these failures can be difficult for users to detect [Lin *et al.*, 2022; Huang *et al.*, 2025]. These properties are especially concerning in election administration contexts, where a small number of high-severity errors, e.g., incorrect deadlines, inaccurate eligibility guidance, or incorrect location instructions, can translate into real disenfranchisement and legal exposure. Throughout, *Election AI* denotes conversational AI that provides civic information (e.g., voter-facing chat assistants), not vote-tabulation systems or voting machines.

Why the risk is novel: persuasion at scale and the persuasion–reliability tension. Two developments make election AI auditing urgent. First, persuasion research increasingly shows that conversational AI can measurably shift political attitudes and voter preferences. Pre-registered experiments indicate that human–AI dialogues can shift candidate preference in multiple election contexts, with non-trivial persistence over time [Lin *et al.*, 2025]. Large-scale studies also show that post-training and rhetorical prompting can increase persuasion, but may reduce factual accuracy, highlighting a potential persuasion–reliability trade-off in conversational settings [Hackenburg *et al.*, 2025]. Related work suggests that large language models (LLMs) generated messages can persuade on policy issues at scale [Bai *et al.*, 2025], that AI-generated propaganda can be effective [Goldstein *et al.*, 2024], and that personalization may further amplify influence [Matz *et al.*, 2024; Hackenburg and Margetts, 2024]. Second, the broader information threat landscape includes synthetic media and deepfakes that can increase uncertainty and degrade trust, particularly when citizens cannot read-

ily distinguish authentic from synthetic content [Vaccari and Chadwick, 2020; Nightingale and Farid, 2022]. In this environment, conversational AI used for civic information can act as high-trust, high-reach intermediaries; both factual errors and systematic distortions can have disproportionate downstream effects.

Dual-use perspective. Importantly, conversational AI influence is not intrinsically malign: recent evidence suggests that AI dialogues can also reduce conspiracy beliefs in durable ways, implying that the same interactive affordances may be used for pro-social civic interventions when carefully designed [Costello *et al.*, 2024]. This dual-use reality strengthens (rather than weakens) the case for auditing: if election institutions experiment with helpful uses, they need tools that quantify when a system is simultaneously persuasive and error-prone. CivicAudit-Bench therefore treats persuasion, truthfulness, and operational safety as a coupled governance problem rather than separate research silos.

These risks motivate the need for technically grounded, pre-deployment auditing methods that can detect high-severity failures under realistic conversational conditions.

1.1 Insights from Stakeholders

The proposed project is explicitly driven by election-administration needs rather than abstract benchmark performance. At the AI & Governance Conference at Washington University (October 2025), U.S. Election Assistance Commission (EAC) Commissioner emphasized that even a small AI misinformation incident could undermine public confidence, especially if it appears politically targeted, and called for 'white-hat' vulnerability testing with feedback to vendors and localities. The Commissioner framed the scorecard we propose as a "consumer report" for election offices. The EAC's Chief AI Officer identified the Election Administration and Voting Survey (EAVS) as a reliable state-level ground-truth backbone, flagged third-party chatbots on official election websites as a concrete deployment risk, and opened a recruitment pathway to local practitioners through the EAC's community of officials. Stakeholders hold defined co-design roles: severity-rubric and safe-escalation design, query and evidence validation, and threshold/scorecard review. We will additionally recruit 2–3 local practitioners and engage domain experts in election law, responsible disclosure, and HCI/trust. This collaboration directly shapes the project's goals, metrics, and dissemination format: the audit is designed to support deployment and governance decisions, not simply academic measurement.

Stakeholder-defined project goals. **G1**. Establish a stakeholder-validated risk taxonomy for election/civic conversational AI, distinguishing high-severity 'critical errors' from low-severity factual slips and identifying unacceptable failure modes (e.g., confident fabrication of jurisdiction-dependent rules). **G2**. Create an election-specific benchmark suite and scorecard that can be run by low-resource election offices or auditors, aligned with deployment decisions and coordinated disclosure norms. **G3**. Measure and mitigate asymmetric behavior (refusals, hedging, and accuracy) across matched entity-swapped queries to detect systematic distortions that may be perceived as political targeting. **G4**. Evaluate disclosure and transparency interventions as governance-relevant mitigations, assessing whether labels reduce overreliance and attenuate persuasion without encouraging disengagement or selective skepticism. **G5**. Produce an actionable implementation package: documentation, reporting template, and a responsible 'white-hat' workflow for sharing findings with vendors and election officials prior to public reporting.

Stakeholder workplan: goals, metrics, and outputs. Stakeholder collaboration is operationalized through structured engagements that determine benchmark scope, severity definitions, and deployment-oriented success criteria:

Risk taxonomy elicitation (Months 1–3). Using structured interviews and iterative card-sorting with EAC officials, rank institutionally concerning AI behaviors (e.g., deadline fabrication; asymmetric refusals) and translate them into query families, severity weights, and safe-escalation definitions.

Benchmark validation (Months 4–9). Practitioner review of query categories, edge cases, and evidence trails; validate correctness under jurisdiction-dependent conditions (answer vs. clarification vs. referral).

Thresholds and scorecard design (Months 7–10). Co-define red–yellow–green thresholds and critical-error budgets; iterate the practitioner-facing reporting template.

Disclosure and mitigation review (Months 10–18). Review disclosure formats and escalation guidance; ensure mitigations reduce overreliance without selective denial of legitimate civic information.

Responsible disclosure workflow (Months 10–24). Establish coordinated reporting so high-risk vulnerabilities are shared with vendors and stakeholders before public release, with remediation windows and tiered prompt release.

2 Strategies

CivicAudit-Bench is organized around four strategies intended to translate research insights into deployable election-governance infrastructure. (S1) Use stakeholder co-design to define the risk model, query taxonomy, and severity thresholds, avoiding the common mismatch between academic metrics and institutional needs. (S2) Combine holistic evaluation with domain-specific stress tests: general-purpose LLM benchmarks (e.g., MMLU [Hendrycks *et al.*, 2020], BIG-Bench [Srivastava *et al.*, 2023], HELM [Liang *et al.*, 2022]) demonstrate the field's capacity to standardize evaluation, but election administration requires targeted tests for critical procedural accuracy and safe escalation. (S3) Prioritize failure modes that matter for public trust, such as hallucinations, false confidence, and asymmetry, using factuality measures and hallucination detection methods developed in NLP [Lin *et al.*, 2022; Min *et al.*, 2023; Manakul *et al.*, 2023; Huang *et al.*, 2025]. (S4) Treat mitigation as part of auditing: election officials require not only a 'risk score' but also tested design patterns (grounding, escalation, and disclosure) that improve safety without creating asymmetric denial of legitimate civic information. This strategy is guided by prior work on trust and transparency in AI systems, which emphasizes that disclosure and explanation mechanisms must be evaluated empirically rather than assumed to work [Jacovi *et al.*, 2021; Buçinca *et al.*, 2021; Wittenberg *et al.*, 2024].

White-hat orientation and responsible disclosure. The project adopts a 'white-hat' auditing posture aligned with election-administration norms and responsible security research. Auditing results will be communicated first to stakeholders and model vendors through a coordinated disclosure protocol, with sufficient time for remediation before public reporting. This approach addresses a core concern: election offices benefit from diagnostics that lead to fixes, not adversarial demonstrations that increase mistrust or create perverse incentives. To balance transparency with misuse risk, CivicAudit-Bench will release core benchmark infrastructure and scoring code, while high-risk adversarial prompts and vulnerability details will follow tiered release practices (e.g., delayed release after remediation). Because coordinated disclosure depends on vendor cooperation, we set a fixed remediation window after which tiered public release proceeds; crucially, the jurisdiction-safe-completion guidance officials rely on does not require any single vendor's fix.

Benchmark lifecycle, versioning, and monitoring. Election AI evaluation is not a one-off certification task. Two dynamics require lifecycle-aware auditing: (i) election rules and administrative procedures evolve over time and vary across jurisdictions; and (ii) model behavior can change as vendors deploy updates, fine-tuning, or policy-layer modifications. CivicAudit-Bench therefore adopts a versioned benchmark design. Each query pack is tagged with (a) jurisdictional scope (federal vs. state/local), (b) effective date window, and (c) evidence sources used for validation. This enables periodic re-auditing and transparent comparison across time. The monitoring view is aligned with holistic evaluation principles in the LLM community, where repeated benchmarking has become standard practice [Hendrycks *et al.*, 2020; Srivastava *et al.*, 2023; Liang *et al.*, 2022]), but translated into operationally meaningful election-office thresholds (critical-error budgets, safe-escalation compliance).

Methodological contributions. CivicAudit-Bench contributes three technical advances: counterfactual entity-swap probing for detecting asymmetric failures, severity-weighted hallucination metrics grounded in institutional risk, and jurisdiction-safe completion as an operational criterion.

3 Methods

3.1 WP1: PoliBias-US: Multi-Indicator Alignment Screening

Rationale. Roll-call-based ideology scaling, i.e. the projection of voting behavior into a low-dimensional ideological space, as in DW-NOMINATE, provides an unusually objective anchor for political alignment measurement, but election-relevant distortions can also arise through party-cue susceptibility, overconfident stances, and narrative framing choices. WP1 therefore augments vote-alignment with three complementary stance signals to produce a 'political alignment risk profile' useful for procurement screening and for interpreting WP2's operational-risk results [Carroll *et al.*, 2009; Wang *et al.*, 2023; Bang *et al.*, 2024; Gallegos *et al.*, 2024].

Task 1.1 Voting alignment (roll-call projection). Data. We use publicly available U.S. Congressional roll-call votes from VoteView and focus on Congresses and issue areas most relevant to civic information and election governance [Lewis *et al.*, 2019]. For each roll call, the model receives a neutral summary of the motion plus short pro/con arguments drawn from public congressional summaries, then chooses Yea/Nay/Present and provides a brief justification. Mapping into ideology space. We project vote vectors into DW-NOMINATE space and quantify uncertainty via bootstrap and prompt-variant sensitivity checks. Outputs include an ideology location estimate with uncertainty, domain contribution diagnostics (e.g., voting rights, election security), and stability across time/prompt variants.

Task 1.2 Entity cue sensitivity (party-label counterfactuals). To distinguish content-based positioning from label-driven partisanship, we construct matched prompts where the motion text is held fixed but the attributed sponsor/coalition label is swapped. We report an Entity Cue Sensitivity Index: (i) vote-flip rate under label swaps, (ii) directional skew of flips, and (iii) differences in justification tone. This counterfactual design builds on causal perspectives on entity bias in language models [Wang *et al.*, 2023] and on our prior voting-record benchmarking work [Chen *et al.*, 2026].

Task 1.3 Ideological certainty, coherence, and robustness. Beyond direction, governance risk depends on how confidently and how stably a model expresses political positions. We therefore measure (i) certainty (e.g., elicited probability of Yea/Nay, or token-level confidence where available), (ii) coherence across paraphrases of the same motion, and (iii) persona sensitivity (e.g., 'as an election administrator' vs. 'as a campaign volunteer') to detect role-driven ideological shifts. To separate stylistic persona from a substantive reliability shift, we hold the factual query constant across personas and track accuracy and refusal: changes beyond tone signal a real shift, not stylistic adaptation. This complements surveys of LLM bias and fairness that emphasize robustness and measurement validity [Blodgett *et al.*, 2020; Navigli *et al.*, 2023; Gallegos *et al.*, 2024].

Task 1.4 Narrative alignment (framing beyond voting). Some election-civic risks arise not from explicit 'votes' but from how models frame contested issues (what arguments are foregrounded, omitted, or treated as settled). For a stakeholder-selected subset of contested civic topics (e.g., mail voting, voter ID, election security), we curate symmetric reference narratives from official sources (e.g., party platforms and authoritative election-administration materials) and compare model summaries/argument maps to these references using semantic similarity and omission checks. The goal is not to enforce a 'correct ideology' but to quantify systematic narrative proximity and imbalance as an additional screening indicator, aligned with the literature emphasizing that political bias manifests both in what is said and how it is said [Bang *et al.*, 2024] and with holistic evaluation principles for LLM behavior across scenarios [Liang *et al.*, 2022].

WP1 outputs and integration with WP2. WP1 produces a compact profile: ideology location (with uncertainty), cue sensitivity, certainty/coherence, and narrative alignment. The profile is intended as a screening and interpretability layer rather than a standalone verdict. In later phases, we test convergence: models that appear highly cue-sensitive or narra-

tively imbalanced in WP1 should also exhibit higher asymmetry and critical-error vulnerability in WP2 under matched election-civic prompts.

3.2 WP2: HalluBias-Election: Hallucination, Asymmetry, and Critical-Error Auditing

Rationale. Hallucination and factuality failures are widely documented in generative models, motivating both taxonomies and automated detection methods [Huang *et al.*, 2025]. However, election administration requires domain-specific operationalization, building on our hallucination-bias pipeline: the key concern is not just whether a response contains any unsupported text, but whether it contains high-severity, actionable errors and whether failures are systematically asymmetric across political entities or topics.

Benchmark design. The benchmark comprises stakeholder-validated query families covering (i) voter registration and deadlines; (ii) eligibility and ID rules; (iii) mail/absentee and early voting; (iv) polling-place lookup and hours (jurisdiction-conditional); (v) accessibility and language assistance; (vi) election integrity and results reporting; and (vii) controlled 'entity' prompts for counterfactual asymmetry tests.

Gold standards and evidence. We operationalize ground truth through a tiered architecture. *Tier 1* covers federal/universal rules with a single correct answer. *Tier 2* covers state-level procedures, grounded in the EAC's Election Administration and Voting Survey (EAVS) and Policy Survey covering 50 states spanning registration, mail/absentee voting, ID, early voting, and felon disenfranchisement. *Tier 3* covers jurisdiction- or user-dependent queries, where the correct answer is not a cached fact but *safe escalation*. Query construction is therefore concrete: EAVS items become natural-language voter phrasings per state, and each state's own response is the gold standard. We cross-validate against EAC state-level data, use `vote.gov` as the referral target, and triangulate against NCSL databases.

Hallucination detection stack. We will combine (a) fine-grained factuality scoring methods such as FActScore for long-form outputs [Min *et al.*, 2023], (b) black-box hallucination detection such as SelfCheckGPT [Manakul *et al.*, 2023], and (c) annotation-backed hallucination typologies [Ji *et al.*, 2024; Gu *et al.*, 2024]. We will additionally benchmark against TruthfulQA-style prompts for common misconceptions [Lin *et al.*, 2022] adapted to election contexts.

Entity-swap counterfactual probing. To detect asymmetric treatment, we construct matched query pairs that differ only in the political entity (e.g., 'Democratic' vs. 'Republican'; named candidates; or institutions). This counterfactual design isolates entity-specific differences in refusal rates, hedging, and accuracy, while rhetorical or 'soft' bias that does not hinge on a factual swap is captured instead by Task 1.4's narrative-alignment screen. The approach is informed by causal perspectives on entity bias [Wang *et al.*, 2023] and by bias benchmark methodology in NLP [Parrish *et al.*, 2022; Czarnowska *et al.*, 2021].

Jurisdiction-conditional correctness. For Tier 3 queries whose correct answer depends on jurisdiction or user-specific information, the benchmark treats 'correct' behavior as (i) flagging jurisdiction dependence, (ii) requesting necessary context, and (iii) directing the user to authoritative lookup resources rather than fabricating specifics. Consequently the benchmark needs no rules database for thousands of jurisdictions: for Tier 3 we test escalation behavior, while only the stable Tier 1/2 layers are versioned with effective-date windows and refreshed each cycle, bounding maintenance.

Operational risk metrics. The scorecard reports: hallucination rate by category; false-confidence rate; severity-weighted critical-error rate; refusal/hedging asymmetry index; accuracy asymmetry index; and 'jurisdiction-safe completion' rate for location-dependent prompts.

Outputs. WP2 delivers an election-office-facing red–yellow–green scorecard and a technical audit report suitable for vendors and auditors, enabling before-purchase model comparison and post-deployment monitoring.

Query construction and annotation. WP2 pairs each query family with (i) an authoritative evidence trail and (ii) an annotation rubric distinguishing incomplete-but-safe answers, incorrect procedural answers, and high-severity critical errors. Severity weights and the critical/non-critical boundary are not researcher-imposed: they are outputs of the stakeholder co-design, grounded in our EAC contact's distinction between fabricating a deadline or polling place (directly disenfranchising) and an incomplete-but-safe response. Labeling is two-stage: annotators apply the rubric and weights, then domain experts (election-law and practitioner) adjudicate conditional edge cases. This human-in-the-loop step keeps the framework from reducing to one LLM grading another: detectors propose, people adjudicate. We report inter-annotator agreement (κ) and adjudication logs per query family, and test metric stability across paraphrases.

Model set and cost transparency. CivicAudit-Bench is designed to evaluate both open-weight models and API-based frontier models because election offices may face trade-offs between cost, controllability, and safety policies (refusals/guardrails). For reproducibility, we will log model versions, system prompts/policy settings where disclosed, and sampling parameters. We also report the marginal cost per audit run (tokens, runtime) so that low-resource jurisdictions can plan evaluations and decide whether to outsource audits to third parties.

3.3 WP3: Disclosure-Test: Validating Transparency Interventions for Elections

Rationale. Disclosure and transparency are frequently proposed governance interventions for AI-generated political content, but empirical evidence shows that label design and user interpretation matter [Wittenberg *et al.*, 2024; Gamage *et al.*, 2025]. Research on trust in AI further suggests that disclosure can either improve calibration or create misplaced skepticism depending on context [Jacovi *et al.*, 2021; Eslami *et al.*, 2019].

Design. We will conduct a pre-registered, between-subjects experiment with U.S. registered voters, testing how different disclosure formats affect perceived credibility, reliance, and persuasion in election-relevant dialogues. Conditions include: (C1) no disclosure; (C2) source label ('AI-generated'); (C3) uncertainty cue + grounding prompt (e.g.,

WP	Deliverables (stakeholder-facing)
WP1	Ideology + cue-sensitivity + robustness profile; procurement screening summary.
WP2	Severity-weighted scorecard; evidence-linked query packs; disclosure report template.
WP3	Pre-registered disclosure effect estimates; recommended label/interface patterns.

Table 1: WPs mapped to outputs for election stakeholders.

'I may be wrong; please verify with official sources'); (C4) combined disclosure (source + uncertainty + verification). Participants give informed consent and are told they are interacting with AI-generated election information; following IRB partial-disclosure practice, specific hypotheses are withheld until a post-task debrief to preserve ecological validity.

Outcome measures. We measure (i) attitude change and candidate preference where applicable; (ii) perceived trustworthiness; (iii) comprehension of limitations; and (iv) behavioral intention to verify via official sources. The design is informed by work on cognitive forcing functions to reduce overreliance on AI [Buçinca *et al.*, 2021] and on explainable AI experience design [Liao *et al.*, 2020].

Why this matters. If persuasion can be scaled through dialogue [Lin *et al.*, 2025; Hackenburg *et al.*, 2025] and personalization [Matz *et al.*, 2024], governance-relevant disclosure should be tested under realistic conversational conditions.

Outputs. WP3 produces evidence-based recommendations for election offices and regulators regarding which disclosure designs increase calibration and reduce overreliance without inadvertently increasing polarization or selective distrust.

4 Foreseen Case Studies

CivicAudit-Bench will be demonstrated through five foreseen case studies that reflect realistic election-administration adoption pathways and stress-test conditions.

Study A. Procurement screening for voter-information chatbots. We will simulate a pre-deployment procurement decision by evaluating multiple candidate models (open-weight and API-based) on CivicAudit-Bench. Outputs include a scorecard and a narrative report translating metrics into procurement-relevant risk statements. This case study operationalizes stakeholder concerns about low-resource decision-making and creates a repeatable 'minimum audit' checklist.

Study B. Jurisdiction-dependent procedural queries. We will focus on high-severity prompts where rules vary by state/county (polling place, ID rules, deadlines). The key outcome is jurisdiction-safe completion: whether the model escalates to official lookup tools rather than hallucinating specifics. This case study responds directly to evidence that misinformation spreads quickly online and that procedural errors can have outsized impact [Vosoughi *et al.*, 2018].

Study C. Synthetic media and misinformation-adjacent claims. We will evaluate model responses to misinformation-adjacent election claims and synthetic-content scenarios, emphasizing non-amplification, evidence-based correction, and safe referral to authoritative sources. The case study is motivated by documented threats from deepfakes and synthetic political media [Vaccari and Chadwick, 2020; Pawelec, 2022] and by findings that synthetic content can appear trustworthy to users [Nightingale and Farid, 2022] and that LLMs can sometimes 'disinform' users better than humans [Spitale *et al.*, 2023].

Study D. Accessibility and language-support stress tests. To align with 'leave no one behind' priorities, we will evaluate how models handle accessibility accommodations (e.g., disability-related voting assistance, accessible ballot options) and basic language-assistance queries. Beyond average accuracy, we will report whether critical-error, false-confidence, or refusal rates are systematically higher on these prompts—an operational equity risk for marginalized voters.

Study E. Post-deployment drift monitoring. We demonstrate how the benchmark supports periodic re-auditing as models or policies change. Using versioned query packs (Section 2.3), we compare successive model snapshots and flag meaningful degradations or shifts in refusal/asymmetry patterns, enabling election offices to treat vendor updates as auditable events rather than opaque changes.

5 Expected Results and Impacts on Pressing Societal Challenges

Expected results are both technical and institutional. Technically, CivicAudit-Bench will deliver a benchmark suite, scoring code, and reporting templates that enable systematic, replicable evaluation of election-civic conversational AI. Institutionally, it will provide election offices with a concrete method to translate general concerns about 'AI risk' into auditable criteria, supporting safer adoption and procurement.

Deliverables include: (D1) PoliBias-US multi-indicator screening module; (D2) HalluBias-Election query suite, annotation rubric, and severity-weighted risk metrics; (D3) a red–yellow–green scorecard with stakeholder-defined thresholds; (D4) a coordinated disclosure workflow and documentation for vendors and officials; and (D5) experimental evidence on disclosure designs that improve calibration and reduce overreliance.

Broader impacts map directly to UN SDG 16 (peace, justice, and strong institutions) by strengthening the informational integrity of democratic processes, improving institutional capacity to evaluate emerging technologies, and reducing unequal burdens created by procedural misinformation. The project also contributes to a growing research agenda on AI and democracy that calls for empirically grounded assessments of democratic harms and benefits [Jungherr, 2023; Summerfield *et al.*, 2025; Kreps and Kriner, 2023].

6 Evaluation Criteria

Evaluation is organized around the following criteria:
(E1) Technical validity. WP1's voting-alignment estimates will be assessed for stability across prompt variants and compared to established roll-call estimation approaches [Poole and Rosenthal, 2011; Clinton *et al.*, 2004; Imai *et al.*, 2016]. We will additionally evaluate WP1's entity cue sensitivity and narrative-alignment signals for robustness across paraphrases and persona frames, and for convergence with entity-bias diagnostics and holistic evaluation principles [Wang *et al.*,

2023; Bang *et al.*, 2024; Liang *et al.*, 2022]. WP2's hallucination and factuality measures will be cross-validated across methods (FActScore, SelfCheckGPT, annotation-based labels) to avoid over-reliance on any single detector [Min *et al.*, 2023; Manakul *et al.*, 2023; Ji *et al.*, 2024].

(E2) Benchmark reliability. We will report inter-annotator agreement for severity labels and compute robustness to query paraphrases and time drift.

(E3) Governance relevance. Stakeholder reviewers will assess whether the scorecard is interpretable and whether thresholds align with procurement decision needs. Usability feedback will be collected from local election practitioners, focusing on the clarity of 'what to do next' recommendations.

(E4) Mitigation efficacy. WP3 will estimate disclosure effects on reliance and persuasion under pre-registered analysis plans, reporting effect sizes and heterogeneity by political knowledge and trust predispositions.

7 Challenges and Limitations

Several challenges are anticipated. First, election rules change over time and vary across jurisdictions; maintaining benchmark currency requires a lightweight update process and careful scoping to avoid encouraging models to guess local rules. Second, model behavior can drift as vendors update systems; CivicAudit-Bench therefore emphasizes reproducible testing and periodic re-auditing rather than one-time certification. Third, political "bias" is a contested concept; we treat ideology screening as an uncertainty-aware diagnostic and report results consistent with critical perspectives in NLP bias research [Blodgett *et al.*, 2020; Gallegos *et al.*, 2024]. Fourth, black-box evaluation cannot fully reveal model internals; we mitigate this by triangulating measurement approaches [Liang *et al.*, 2022] and emphasizing observable operational harms (critical errors and false confidence) over speculative mechanisms. Fifth, the benchmark is U.S.-focused by design (EAVS and U.S. roll-call data), though the tiered, modular architecture ports to parliamentary or multi-party systems—a natural extension of the team's parliamentary-records work [Chen *et al.*, 2026]. Sixth, scope is text-based; multimodal misinformation (AI-generated images and memes) is left to future work.

8 Implementation Plan and Needs

8.1 Timeline (24 Months)

The plan is staged with explicit dependencies and intermediate deliverables rather than independent tracks.

Phase 1 (Months 1–9): foundations. WP1 (public VoteView roll-call data) and WP2 (EAVS and other authoritative sources) run in parallel, sharing the Months 1–3 risk-taxonomy co-design that fixes the query taxonomy, severity rubric, and IRB materials. We release CivicAudit-Bench v1 at Month 9 for early feedback.

Phase 2 (Months 10–18): mitigation and human-subjects evaluation. Mitigation prototyping runs alongside WP3 piloting (Months 10–12), then the main pre-registered experiments (Months 13–18), with one replication across a second sampling frame. The Phase 1 benchmark and scorecards are already in hand, reducing WP3 risk.

Phase 3 (Months 19–24): integration and dissemination. Produce CivicAudit-Bench v2, practitioner documentation and training, and disseminate via briefings, webinars, and an update-ready playbook.

The project reuses working infrastructure from the PoliBias project and the validated pipeline in [Chen *et al.*, 2026]; and the versioned design ships a usable v1 at an intermediate milestone, so the contribution does not hinge on every later component landing on schedule.

8.2 Resources

The main cost drivers are (i) participant recruitment incentives and platform fees for WP3, and (ii) human labor costs, including full-time research associate effort and research assistant time for benchmark construction, annotation, and maintenance. To budget realistically, the project will treat human experiments as a Year-2 milestone gated on IRB approval and pilot validation. For typical online survey studies (10-15 minutes) paid at ethical rates, per-complete costs are on the order of a few USD on academic recruiting platforms (participant pay plus platform fees). As a concrete planning range, three experiments totaling N = 6,000-10,000 responses implies a respondent-cost budget on the order of tens of thousands of USD (higher if representative state-level quotas are required).

8.3 Outputs and Dissemination

Code, core benchmark datasets, and documentation will be released openly, with a practitioner-oriented 'how to run an audit' guide and a standardized scorecard report suitable for procurement workflows. High-risk adversarial prompts will follow a tiered release policy with coordinated disclosure to affected stakeholders.

8.4 Governance, QA, and Reproducibility

Reproducibility and traceability. Each audit run will log: model name/version (as disclosed), decoding parameters, system prompts/policy settings where available, and the exact benchmark version. Each query pack will include an evidence trail (official source links and effective dates) so that stakeholders can distinguish model errors from rule updates.

Annotation QA. For stakeholder-prioritized "critical" query families, we will use double annotation with adjudication. We will report inter-annotator agreement, maintain an adjudication log of disputed cases, and version the rubric so that updates remain transparent across election cycles.

Release governance and misuse controls. The toolkit will be open-source, but high-risk adversarial prompts will follow a tiered release strategy aligned with coordinated disclosure. Public releases will prioritize diagnostic categories and safe behaviors (grounding, non-amplification, escalation) rather than providing reusable attack recipes.

Success criteria (operational). Beyond accuracy, we will treat the following as governance-critical metrics: (i) severity-weighted critical-error rate; (ii) jurisdiction-safe completion rate; (iii) asymmetric refusal/hedging/accuracy indices under entity-swap counterfactuals; and (iv) the stability of these metrics under model updates.

8.5 Risk Management

Risk: API Volatility. Commercial models (e.g., GPT) may change safety filters mid-audit, invalidating benchmarks. *Mitigation:* We maintain a parallel "Reference Set" of open-weight models (Llama 3, Mistral) hosted on local GPU infrastructure to distinguish model drift from benchmark drift.
Risk: Adversarial Adaptation. Vendors might optimize against public benchmarks (Goodhart's Law). *Mitigation:* We maintain a held-out "Red Team" dataset of adversarial prompts that are never published, shared only with verified election officials under NDA.
Risk: Political Polarization. The audit tool could be accused of partisan bias. *Mitigation:* The framework reports bias symmetrically (flagging both Left-Skew and Right-Skew) and relies on ground-truth voting records rather than researcher labels to establish ideological baselines.

9 Scalability and Economic Sustainability

Scalability is addressed at three levels. Methodologically, the benchmark architecture is modular: jurisdictions can add local query packs and gold standards while preserving metrics for cross-jurisdiction comparison. Institutionally, the scorecard and reporting template are designed for election offices with limited technical capacity and can be used by third-party auditors on behalf of smaller jurisdictions. Economically, the core toolkit will be open-source, minimizing licensing barriers and enabling community maintenance; periodic benchmark updates (e.g., per election cycle/Congress) are designed as low-cost data refreshes rather than model retraining.

10 Project Team Description

10.1 Core Team and Complementary Expertise

Dr. Chendi Wang (Co-PI) is a tenured Assistant Professor at Vrije Universiteit Amsterdam (Political Science and Public Administration), with expertise in political institutions, political behavior, computational social science, and experimental methods. He co-leads scientific direction from a governance and democratic-institutions perspective, including stakeholder co-design, risk taxonomy development, and the definition of institutionally meaningful evaluation criteria. He leads the human-subjects and disclosure evaluation component (WP3) and co-leads cross-package validity assessment and interpretation.
Dr. Jieying Chen (Co-PI) is a tenured Assistant Professor at Vrije Universiteit Amsterdam (Computer Science) specializing in knowledge representation and reasoning and bias detection in LLMs. She is the PI of *PoliBiasEU* (NWO) on multilingual political-bias benchmarking using parliamentary voting records. She co-leads scientific direction from an AI methodology perspective, leads the diagnostic framework underlying (benchmark architecture, robustness diagnostics, and reproducible auditing pipelines) and the foundations of WP1 and WP2, and co-leads integration.

10.2 Work-Package Leadership

WP1 (PoliBias-US): co-led by Wang and Chen. Wang leads political-measurement validity (roll-call scaling, uncertainty reporting, interpretation), while Chen leads robustness/engineering (multi-signal indicators, counterfactual instrumentation, scalable implementation).
WP2 (HalluBias-Election): led by Chen with Wang as domain lead. Chen leads the benchmark infrastructure (query versioning, scoring stack, automation, release engineering). Wang leads domain scoping, severity rubric design, and practitioner-facing reporting formats.
WP3 (Disclosure-Test): led by Wang with Chen supporting instrumentation and logging. Wang leads study design, preregistration, analysis, and governance interpretation; Chen supports experimental interface implementation, reproducibility, and measurement pipelines.

10.3 Staffing and Execution Capacity

We budget for (i) a full-time research associate for dataset construction, annotation, and benchmark maintenance, and (ii) part-time student assistants for query expansion, evidence linking, and documentation. The design keeps core auditing runs feasible for low-resource stakeholders (minimal dependencies, transparent cost accounting).

10.4 Stakeholder Engagement and Collaboration

Stakeholder engagement is built into the project lifecycle. We will conduct structured consultations with the U.S. Election Assistance Commission (EAC) and recruit 2–3 local election practitioners (county/state) for usability feedback and threshold validation. For operational realism and responsible disclosure, we will seat a project advisory board including a dedicated election-law expert who validates jurisdiction-conditional correctness and edge cases, alongside HCI/trust and coordinated-disclosure experts. EAC will provide advisory input on institutional priorities and acceptable risk thresholds. Stakeholder participation is advisory only and does not constitute institutional endorsement.

10.5 Project Management

The Co-PIs hold weekly check-ins, monthly milestone reviews against the 24-month plan and quarterly stakeholder briefings, with a biannual advisory council meeting on progress, risks and alignment. Releases follow an internal QA checklist (data provenance, model/version logging, reproducibility scripts), and the benchmark uses semantic versioning for comparable results over time.

Ethical Statement

Election-related AI failures can affect rights, equality, and trust. WP3 human-subjects research runs under IRB review with informed consent and partial disclosure of hypotheses until debrief, preserving ecological validity. To avoid partisan targeting, asymmetry metrics are computed symmetrically and bias is reported regardless of direction. We use coordinated disclosure for high-risk vulnerabilities and tiered release for adversarial prompts; for misinformation-adjacent cases we score response quality (grounding, non-amplification, safe escalation) rather than reproducing false narratives, and treat disclosure and explanation interfaces as human-factors interventions requiring empirical validation.

Acknowledgements

This work was supported by the Dutch Ministry of Education, Culture and Science (OCW) in the scope of the project *Navigating the Storm: European Political Contestation in Geopolitical Transformation (NEST)* and by the Dutch Research Council (NWO) through the project *PoliBiasEU: A Scalable and Multilingual Benchmark for Political Bias Detection in Large Language Models*, and partially supported by the Network Institute, Vrije Universiteit Amsterdam, through the Academy Assistants Program.

References

[Bai *et al.*, 2025] Hui Bai, Jan G Voelkel, Shane Muldowney, Johannes C Eichstaedt, and Robb Willer. Llm-generated messages can persuade humans on policy issues. *Nature Communications*, 16(1):6037, 2025.

[Bang *et al.*, 2024] Yejin Bang, Delong Chen, Nayeon Lee, and Pascale Fung. Measuring political bias in large language models: What is said and how it is said. *arXiv preprint arXiv:2403.18932*, 2024.

[Blodgett *et al.*, 2020] Su Lin Blodgett, Solon Barocas, Hal Daumé Iii, and Hanna Wallach. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*, 2020.

[Buçinca *et al.*, 2021] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-computer Interaction*, 5(CSCW1):1–21, 2021.

[Carroll *et al.*, 2009] Royce Carroll, Jeffrey B Lewis, James Lo, Keith T Poole, and Howard Rosenthal. Measuring bias and uncertainty in dw-nominate ideal point estimates via the parametric bootstrap. *Political analysis*, 17(3):261–275, 2009.

[Chen *et al.*, 2026] Jieying Chen, Karen de Jong, Andreas Poole, Jan Burakowski, Elena Elderson Nosti, Joep Windt, and Chendi Wang. Uncovering political bias in large language models using parliamentary voting records. *arXiv preprint arXiv:2601.08785*, 2026.

[Clinton *et al.*, 2004] Joshua Clinton, Simon Jackman, and Douglas Rivers. The statistical analysis of roll call data. *American Political Science Review*, 98(2):355–370, 2004.

[Costello *et al.*, 2024] Thomas H Costello, Gordon Pennycook, and David G Rand. Durably reducing conspiracy beliefs through dialogues with ai. *Science*, 385(6714):eadq1814, 2024.

[Czarnowska *et al.*, 2021] Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. Quantifying social biases in nlp: A generalization and empirical comparison of extrinsic fairness metrics. *Transactions of the Association for Computational Linguistics*, 9:1249–1267, 2021.

[Eslami *et al.*, 2019] Motahhare Eslami, Kristen Vaccaro, Min Kyung Lee, Amit Elazari Bar On, Eric Gilbert, and Karrie Karahalios. User attitudes towards algorithmic opacity and transparency in online reviewing platforms. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2019.

[Gallegos *et al.*, 2024] Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179, 2024.

[Gamage *et al.*, 2025] Dilrukshi Gamage, Dilki Sewwandi, Min Zhang, and Arosha K Bandara. Labeling synthetic content: User perceptions of label designs for ai-generated content on social media. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–29, 2025.

[Goldstein *et al.*, 2024] Josh A Goldstein, Jason Chao, Shelby Grossman, Alex Stamos, and Michael Tomz. How persuasive is ai-generated propaganda? *PNAS nexus*, 3(2):pgae034, 2024.

[Gu *et al.*, 2024] Yuzhe Gu, Ziwei Ji, Wenwei Zhang, Chengqi Lyu, Dahua Lin, and Kai Chen. Anah-v2: Scaling analytical hallucination annotation of large language models. *Advances in Neural Information Processing Systems*, 37:60012–60039, 2024.

[Hackenburg and Margetts, 2024] Kobi Hackenburg and Helen Margetts. Evaluating the persuasive influence of political microtargeting with large language models. *Proceedings of the National Academy of Sciences*, 121(24):e2403116121, 2024.

[Hackenburg *et al.*, 2025] Kobi Hackenburg, Ben M Tappin, Luke Hewitt, Ed Saunders, Sid Black, Hause Lin, Catherine Fist, Helen Margetts, David G Rand, and Christopher Summerfield. The levers of political persuasion with conversational artificial intelligence. *Science*, 390(6777):eaea3884, 2025.

[Hendrycks *et al.*, 2020] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

[Huang *et al.*, 2025] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.

[Imai *et al.*, 2016] Kosuke Imai, James Lo, and Jonathan Olmsted. Fast estimation of ideal points with massive data. *American Political Science Review*, 110(4):631–656, 2016.

[Jacovi *et al.*, 2021] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 624–635, 2021.

[Ji *et al.*, 2024] Ziwei Ji, Yuzhe Gu, Wenwei Zhang, Chengqi Lyu, Dahua Lin, and Kai Chen. Anah: Analytical annotation of hallucinations in large language models. *arXiv preprint arXiv:2405.20315*, 2024.

[Jungherr, 2023] Andreas Jungherr. Artificial intelligence and democracy: A conceptual framework. *Social media+ society*, 9(3):20563051231186353, 2023.

[Kreps and Kriner, 2023] Sarah Kreps and Doug Kriner. How ai threatens democracy. *Journal of Democracy*, 34(4):122–131, 2023.

[Lewis *et al.*, 2019] Jeffrey B Lewis, Keith Poole, Howard Rosenthal, Adam Boche, Aaron Rudkin, and Luke Sonnet. Voteview: Congressional roll-call votes database. *See https://voteview. com/(accessed 27 July 2018)*, 2019.

[Liang *et al.*, 2022] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

[Liao *et al.*, 2020] Q Vera Liao, Daniel Gruen, and Sarah Miller. Questioning the ai: informing design practices for explainable ai user experiences. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–15, 2020.

[Lin *et al.*, 2022] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 3214–3252, 2022.

[Lin *et al.*, 2025] Hause Lin, Gabriela Czarnek, Benjamin Lewis, Joshua P White, Adam J Berinsky, Thomas Costello, Gordon Pennycook, and David G Rand. Persuading voters using human–artificial intelligence dialogues. *Nature*, pages 1–8, 2025.

[Manakul *et al.*, 2023] Potsawee Manakul, Adian Liusie, and Mark Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 9004–9017, 2023.

[Matz *et al.*, 2024] Sandra C Matz, Jacob D Teeny, Sumer S Vaid, Heinrich Peters, Gabriella M Harari, and Moran Cerf. The potential of generative ai for personalized persuasion at scale. *Scientific Reports*, 14(1):4692, 2024.

[Min *et al.*, 2023] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, 2023.

[Navigli *et al.*, 2023] Roberto Navigli, Simone Conia, and Björn Ross. Biases in large language models: origins, inventory, and discussion. *ACM Journal of Data and Information Quality*, 15(2):1–21, 2023.

[Nightingale and Farid, 2022] Sophie J Nightingale and Hany Farid. Ai-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences*, 119(8):e2120481119, 2022.

[Parrish *et al.*, 2022] Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. Bbq: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, 2022.

[Pawelec, 2022] Maria Pawelec. Deepfakes and democracy (theory): How synthetic audio-visual media for disinformation and hate speech threaten core democratic functions. *Digital society*, 1(2):19, 2022.

[Poole and Rosenthal, 2011] Keith T Poole and Howard L Rosenthal. *Ideology and congress*, volume 1. Transaction Publishers, 2011.

[Spitale *et al.*, 2023] Giovanni Spitale, Nikola Biller-Andorno, and Federico Germani. Ai model gpt-3 (dis) informs us better than humans. *Science Advances*, 9(26):eadh1850, 2023.

[Srivastava *et al.*, 2023] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on machine learning research*, 2023.

[Summerfield *et al.*, 2025] Christopher Summerfield, Lisa P Argyle, Michiel Bakker, Teddy Collins, Esin Durmus, Tyna Eloundou, Iason Gabriel, Deep Ganguli, Kobi Hackenburg, Gillian K Hadfield, et al. The impact of advanced ai systems on democracy. *Nature Human Behaviour*, pages 1–11, 2025.

[Vaccari and Chadwick, 2020] Cristian Vaccari and Andrew Chadwick. Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social media+ society*, 6(1):2056305120903408, 2020.

[Vosoughi *et al.*, 2018] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *science*, 359(6380):1146–1151, 2018.

[Wang *et al.*, 2023] Fei Wang, Wenjie Mo, Yiwei Wang, Wenxuan Zhou, and Muhao Chen. A causal view of entity bias in (large) language models. *arXiv preprint arXiv:2305.14695*, 2023.

[Wittenberg *et al.*, 2024] Chloe Wittenberg, Ziv Epstein, Adam J Berinsky, and David G Rand. Labeling ai-generated content: promises, perils, and future directions. *An MIT Exploration of Generative AI*, 2024.