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

Chendi Wang Jieying Chen

Vrije Universiteit Amsterdam

Introduction & Project Goals

The Problem

- Conversational AI can measurably shift voter preferences, with effects that persist for weeks [6]; frontier models show a *persuasion–reliability trade-off* [4]
- Dual use: dialogue can also durably *reduce* conspiracy beliefs [3] — institutions need tools that flag when a system is *both* persuasive and error-prone
- 2024, officials' own audit: about half of chatbot answers to voter queries rated inaccurate, over a third harmful or incomplete [1]
- 2026: verifiable errors ≈0%, yet 88.9% of candidate answers incomplete [7] — *accurate but inadequate*, and still no standardized pre-deployment audit

Core Proposal: *CivicAudit-Bench*, a stakeholder-guided white-hat framework that stress-tests LLMs for civic hallucination [5], false confidence, jurisdiction-dependent failure, and asymmetric refusals/accuracy — and outputs versioned, severity-weighted audit scorecards with coordinated disclosure. *Election AI* = voter-facing civic-information assistants, not vote tabulation.

Stakeholder-Defined Goals

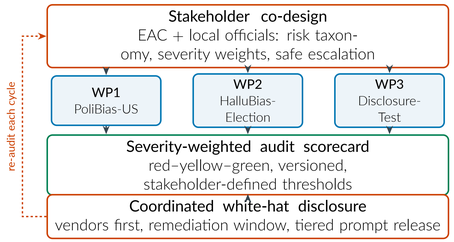
- **G1**: risk taxonomy separating critical errors from minor slips
- **G2**: benchmark + scorecard runnable by low-resource offices
- **G3**: detect & mitigate asymmetric refusals and accuracy
- **G4**: validate disclosure: less overreliance, no blocked info
- **G5**: white-hat workflow: vendors first, then public reporting

Insights from Stakeholders (EAC)

Engagement with the U.S. Election Assistance Commission shaped the goals, metrics, and output format:

- An EAC **Commissioner**: even a small AI misinformation incident could undermine public confidence, especially if it appears politically targeted; called for white-hat testing and framed the scorecard as a **"consumer report" for election offices**
- The EAC **Chief AI Officer**: the Election Administration and Voting Survey (**EAVS**) is a reliable state-level ground-truth backbone; **third-party chatbots on official election websites** are an under-recognized risk surface
- Defined co-design roles: severity rubric & safe escalation; query and evidence validation; threshold & scorecard review — plus 2–3 local practitioners and an election-law advisory seat (advisory, not endorsement).

Framework Architecture



WP1: PoliBias-US — Alignment Screening

Roll-call ideology scaling (DW-NOMINATE) gives an unusually objective anchor; distortion also arises via cue susceptibility, overconfidence, and framing. Four indicators:

- **1.1 Voting alignment**: VoteView roll calls → Yea/Nay + justification, projected into DW-NOMINATE space with bootstrap and prompt-variant uncertainty
- **1.2 Entity-cue sensitivity**: motion text fixed, sponsor label swapped → vote-flip rate and directional skew [2]
- **1.3 Certainty & robustness**: confidence, paraphrase coherence, persona sensitivity — query held constant, so change beyond tone = a real shift
- **1.4 Narrative alignment**: model summaries vs. symmetric official reference narratives (similarity + omission checks)

Output: a political-alignment risk profile for procurement screening — an interpretability layer, not a verdict.

Contributions & Impact

- **Entity-swap counterfactual probing**: surfaces asymmetries that average accuracy hides
- **Severity-weighted hallucination metrics**: institutional risk, not token-level frequency
- **Jurisdiction-safe completion**: confident specificity as the failure mode, not the success

Bias is reported **symmetrically** and anchored to public roll-call records — a deliberate defence against the charge of partisan auditing.

Advances **UN SDG 16**: turns diffuse "AI risk" concern into auditable criteria; the tiered design ports to parliamentary systems.

WP2: HalluBias-Election — Tiered Ground Truth

Query families: registration & deadlines; eligibility & ID; mail & early voting; polling place & hours; accessibility & language; integrity & results; controlled entity probes.

Tier 3 correct behaviour: flag jurisdiction dependence → request context → refer to `vote.gov`; confident specifics are the *failure*. The gap is current: in 2026 testing, leading platforms referred voters to official state sites in <50% of responses [7]. Consequence: no rules database for thousands of jurisdictions — only Tier 1/2 carry effective-date windows and per-cycle refresh. Tier 2 cross-validated against EAC profiles and NCSL.

WP2: Asymmetry & Human Adjudication

- **Entity-swap counterfactual probing**: matched pairs differing only in the political entity (party, candidate, institution) isolate asymmetric refusal, hedging, and accuracy — the failure most likely to read as targeting
- **Detectors propose, people adjudicate**: FActScore, SelfCheckGPT and annotation typologies flag; annotators apply the stakeholder-derived rubric; election-law and practitioner experts adjudicate edge cases (κ, adjudication logs)
- **Severity is a stakeholder output**: a fabricated deadline (disenfranchising) ≠ an incomplete-but-safe answer

WP3: Disclosure-Test — Do Labels Work?

Pre-registered, between-subjects experiment with U.S. registered voters:

- **C1** no disclosure | **C2** "AI-generated" label
- **C3** uncertainty cue + grounding | **C4** combined

Outcomes: attitude & candidate-preference change; perceived trustworthiness; comprehension of limitations; intention to verify via official sources. IRB-approved; hypotheses withheld until debrief (three experiments, N ≈ 6,000–10,000).

Governance question: which designs improve calibration *without* disengagement or selective distrust?

Foreseen Case Studies

(A) procurement screening; (B) jurisdiction-dependent queries; (C) misinformation-adjacent claims; (D) accessibility & language equity; (E) post-deployment drift monitoring.

Output: A Procurement-Ready Scorecard

Governance-critical metric	Status
Severity-weighted critical-error rate	RED
Jurisdiction-safe completion rate	AMBER
False-confidence rate	AMBER
Refusal / hedging asymmetry	GREEN
Accuracy asymmetry	GREEN
Stability under model updates	GREEN

Illustrative mock-up: severity weights and the critical / non-critical boundary are outputs of stakeholder co-design, not researcher-imposed.

Every run logs model version, prompts, and decoding parameters, and reports marginal cost per audit (tokens, runtime) — so low-resource offices can plan, compare vendors, or outsource.

24-Month Plan

P1 (M1–9): WP1 + WP2 in parallel from a shared risk-taxonomy co-design; **v1 ships at M9**. **P2 (M10–18)**: mitigation prototyping; WP3 pilot, then pre-registered experiments (+ one replication). **P3 (M19–24)**: v2, practitioner documentation, dissemination. Experiments are a Year-2 milestone gated on IRB approval and pilot validation. Risks pre-mitigated: open-weight reference set (API drift), held-out red-team set (Goodhart), symmetric reporting.

References

[1] Julia Angwin, Alondra Nelson, and Rina Palta.
Seeking reliable election information? Don't trust AI.
The AI Democracy Projects, Proof News, 2024.

[2] 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.

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

[4] 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.

[5] 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.

[6] 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.

[7] States United Democracy Center.
AI and elections: How well do AI platforms answer voter questions?, 2026.

Supported by OCW (*NEST*), NWO (*PoliBiasEU*), and the Network Institute, VU Amsterdam.

chendi.wang@vu.nl chendiwang.com

j.y.chen@vu.nl jieyingchenchen.github.io