

Anthropic / Claude

Documented measures specific or directly relevant to Israel	Election policies and mechanisms documented globally or outside Israel	Information and additional measures requested for Israel
No public documentation of Israel-specific measures or policies. Anthropic's recent update on safeguarding elections covers the US midterms and Brazil.¹ Global safeguards relevant to Israeli users: Usage Policy is enforced through Safeguards Team monitoring. It prohibits personalized voter targeting, deceptive movements, automated outreach at scale that conceals its artificial origin, deceptive/false election content, voter suppression through deception or intimidation, and disruption of voting/certification.³ Advertising: Anthropic states Claude will remain ad-free, without adjacent sponsored links, advertiser-influenced responses or unsolicited third-party placements.⁵	Election safeguards and testing: Anthropic uses classifiers, human review, Threat Intelligence, Policy Vulnerability Testing and election evaluations. April 2026 tests covered 600 prompts plus multi-turn/autonomous influence operations; results are company-reported.^{1, 8} Country and region-specific preparation: In 2024 Anthropic reported 12+ election-test rounds across India, South Africa, Mexico, UK, France, EU and US, plus authoritative-source routing in the US, EU, UK and France. Testing identified outdated/inaccurate answers and uneven responses; Anthropic then tightened election policies, updated system prompts, adding knowledge-limit warnings and routing users to official sources.² Authoritative information and currentness in 2026: For 2026 Anthropic announced US TurboVote/Democracy Works and later Brazil banners, and tests web-search triggering for time-sensitive election queries; US results are company-reported.¹ Political-use rules: In 2025 Anthropic replaced its blanket campaign/lobbying ban with rules targeting deceptive/disruptive democratic activity and voter/campaign targeting.^{3, 4} Synthetic-content provenance: Anthropic announced future Claude text watermarking globally at launch; supported files will carry C2PA credentials. Its detection API is in private preview, with broader access planned.⁶ Threat disruption and transparency: Anthropic's September 2026 report describes nine influence operations, several election-timed, and says findings improve detection. Its Transparency Hub publishes election-integrity safeguards/monitoring.^{7, 8}	Local election readiness: Identify Israel election-monitoring teams, Hebrew/Arabic expertise, timing, escalation and Central Elections Committee coordination. If no local team exists, explain how safeguards/threat Intelligence functions are adapted to Israel. Authoritative election information: Implement Israel election banner or official-source routing using Central Elections Committee information. Test search triggering, accuracy, cutoff handling and corrections in Hebrew/Arabic and, and when practicable, also Russian/Amharic. Political misuse: Explain enforcement against personalized voter targeting, deceptive campaigns, false election information, synthetic political content intended to deceive, and voter suppression through deception or intimidation, including when conducted through multi-turn or agentic workflows. Clarify watermarking/C2PA coverage and detection access for media, fact-checkers and researchers. Reporting, escalation and local partnerships: Provide a priority reporting route for the Central Elections Committee and qualified civil society organizations, with response times and local-expert testing. Research and transparency: Publish Israel-specific and/or multilingual evaluation methods and limits, distinguishing single- from multi-turn/agentic tests. Enable independent replication and feasible privacy-preserving Israel misuse/enforcement data.

Sources

- 1 Anthropic, "An update on our election safeguards" (24 April 2026): <https://www.anthropic.com/news/election-safeguards-update>
- 2 Anthropic, "Elections and AI in 2024: observations and learnings" (12 December 2024): <https://www.anthropic.com/news/elections-ai-2024>
- 3 Anthropic, "Usage Policy" (effective 15 September 2025; accessed 29 September 2026): <https://www.anthropic.com/legal/aup>
- 4 Anthropic, "Usage policy update" (15 August 2025; changes effective 15 September 2025): <https://www.anthropic.com/news/usage-policy-update>
- 5 Anthropic, "Claude is a space to think" (4 February 2026): <https://www.anthropic.com/news/claude-is-a-space-to-think>
- 6 Anthropic, "How Claude's text watermark works" (14 August 2026; updated 1 September 2026): <https://www.anthropic.com/news/claude-text-watermark>
- 7 Anthropic, "Detecting and countering misuse of AI: September 2026" (10 September 2026): <https://www.anthropic.com/threat-intelligence-report-september-2026>
- 8 Anthropic, "Anthropic's Transparency Hub," Election Integrity section (last updated 23 July 2026; accessed 29 September 2026): <https://www.anthropic.com/transparency/voluntary-commitments>