DEMOS

ELECTORAL HALLUCINATIONS

SAFEGUARDING UK ELECTIONS IN THE WORLD OF LLMS AND AI CHATBOTS

JAMIE HANCOCK
AZZURRA MOORES

MAY 2026

Published by Demos May 2026

15 Whitehall, London, SW1A 2DD
T: 020 3878 3955
hello@demos.co.uk
www.demos.co.uk

CONTENTS

ACKNOWLEDGEMENTS

We would like to thank experts across civil society, business, governance, law and media, for their input, insight and guidance at various stages through the development of this report, including:

- Sam Stockwell, Senior Research Associate, CETaS (The Alan Turing Institute);
- Dr Felix Simon, Research Fellow in AI and Digital News at the Reuters Institute for Study of Journalism, University of Oxford;
- Henry Ajder, Founder, Latent Space Advisory and Demos Fellow;
- Carl Miller, Founder, Centre for the Analysis of Social Media (CASM) and Demos Fellow;
- Dr Elizabeth Seger, Senior Policy Advisor, Tony Blair Institute for Global Change;
- Tommy Shaffer-Shane, Senior Policy Manager, The Centre for Long-Term Resilience;
- Ben Graham Jones, Consultant in Democratic Resilience.

Special thanks to Tyreese Calnan for assisting with the Scottish election testing data collection and analysis. Thank you also to Hannah Perry, Flynn Devine, and Polly Curtis for reviewing this report and providing editorial support, and to Chloe Burke for her support on design.

This project is financially supported by the Quadrature Climate Foundation and is editorially independent.

Any mistakes are the authors' own.

Jamie Hancock and Azzurra Moores

May 2026

ABOUT THIS REPORT

Demos is the UK's leading cross-party think tank producing research and policies that have been adopted by successive governments for more than 30 years. Our mission is an upgraded democracy, with a new deal to mend the broken relationships between the state, institutions, and citizens.

This is the latest paper in Demos Digital's Epistemic Security programme that focuses on securing the UK's information supply chains and building our democratic resilience to adverse influences. It focuses on chatbots and other text-based AI services that are introducing novel risks to the UK's information supply chain and democracy in real time. We reveal new evidence of the scale of these services' unreliability during elections and make recommendations for the government to close the regulatory gap. This report complements other research that Demos has conducted into the integrity of the UK's elections and the democratic implications of generative AI:

- *Free and Fair: Election Law in the Age of AI*
- *Generative AI and Democracy: Impacts and Interventions*
- *Synthetic Politics: Preparing democracy for Generative AI*

Jamie Hancock is a Senior Researcher (Digital Policy) in Demos Digital, Demos' digital policy research hub. **Azzurra Moores** is Demos Digital's Associate Director (Information Ecosystems).

EXECUTIVE SUMMARY

The UK's elections are under pressure. Polarisation, declining trust, online misinformation, and rapidly-developing technologies such as AI deepfakes are putting new pressures on the system. The UK must take urgent steps to safeguard its election integrity ahead of the 2029 general election.

This paper looks at a new and underexamined challenge: AI services that generate text, from chatbots like ChatGPT to AI search tools such as Google AI Overviews. These are increasingly popular as sources of information during elections but they repeatedly fail accuracy and reliability tests. Moreover, they are currently unregulated when it comes to elections. This situation creates a risk for public trust in the election process: if voters come to believe that the election outcome has been swayed by AI-generated errors or hallucinations, they may lose faith in the process. As AI chatbots and related services emerge as key gatekeepers of political information, the risks they pose must be addressed.

Electoral integrity is high on the public agenda. The Representation of the People Bill – the government's flagship legislation which lowers the voting age to 16 – has reignited debate about whether the UK's electoral systems are fit for purpose. Demos has already called for the Bill to include urgent reforms on AI deepfakes and election crisis responses.[1] Yet the slow pace of legislative change means these challenges may reach a crisis point before they are addressed.

This paper advocates for a proactive approach to safeguarding our democratic processes that helps rebuild trust. This is an essential and urgent requirement if the UK is to establish a New Deal to repair the state-citizen relationship and provide trustworthy information that democratic discourse can rely on.[2] Doing so requires shifting from a reliance on voluntary tech industry action to robust regulatory standards grounded in evidence and backed by accountability mechanisms. It means anticipating and preventing crises involving our elections, not reacting when it is too late.

This is the first policy paper to focus specifically on the risks posed by text-based AI in a UK electoral context. We bring together a review of the existing evidence base, new evidence from a snapshot test of how AI services responded to questions about the Scottish Parliament elections, and an analysis of the current legal framework. Taken together, these findings identify potential vulnerabilities that should be addressed before the general election in 2029.

1 Perry et al. (2026). 'Epistemic Security Briefing: The Elections Bill.' Demos. https://demos.co.uk/research/epistemic-security-briefing-the-elections-bill/

2 Curtis (2025). 'Upgrading Democracy: A new deal to repair the broken relationship between citizen and state.' Demos. https://demos.co.uk/research/upgrading-democracy-a-new-deal-to-repair-the-broken-relationship-between-citizen-and-state/

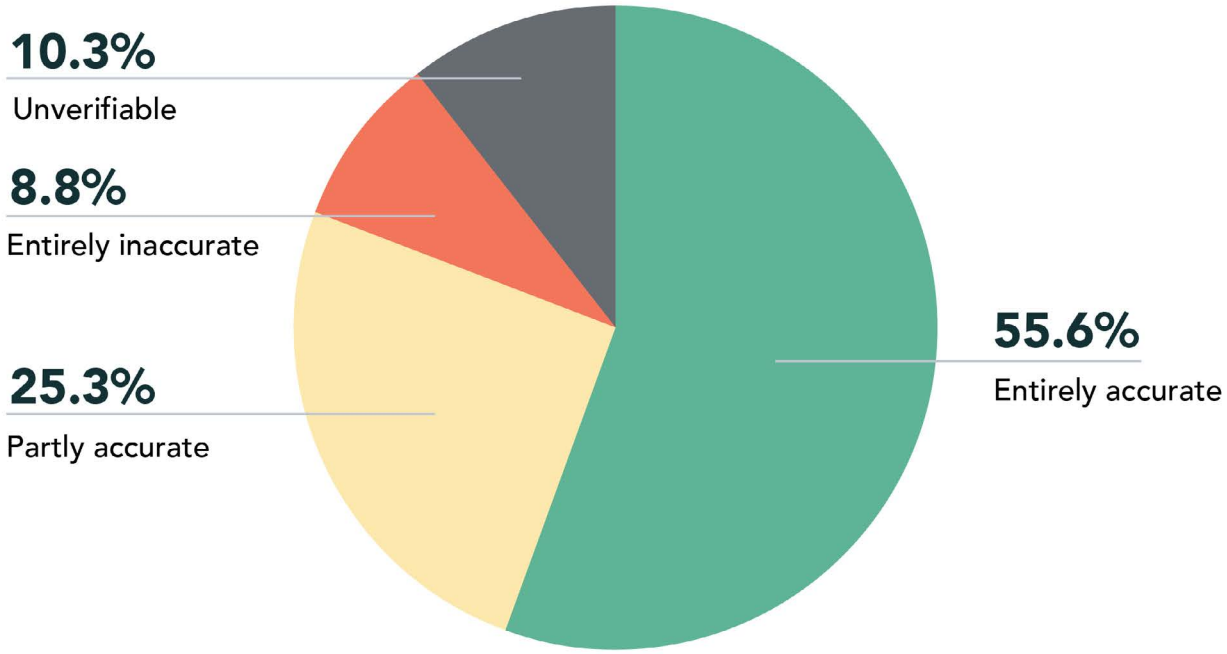
EVIDENCE FROM THE 2026 SCOTTISH ELECTIONS

This paper presents new evidence that demonstrates the unreliability of LLM-based AI services during elections. By testing how text-based AI services performed on a single day during the 2026 Scottish election window (March 27th), it provides quantitative data on the accuracy and reliability of these services during a live election. The research covered five AI services: ChatGPT, Google Gemini, Google AI Overviews, Grok, and Replika, a companion chatbot. It tested these across three Scottish constituencies:

- **Constituency A:** A constituency in a large city where a prominent politician is running.
- **Constituency B:** A hotly contested constituency in a large city with redrawn boundaries.
- **Constituency C:** A rural constituency.

The evidence reveals:[3]

1. **Persistent unreliability in results about Scottish Parliament elections:** 34.1% of responses contained factual errors. 8.75% were entirely inaccurate and 25.3% were partly accurate, but with errors. Only 55.6% were entirely factually accurate. Examples of errors included: getting the date of the election wrong, hallucinating candidates, incorrect advice on voting procedures, and made-up political scandals. The results suggest AI chatbots and related services have serious shortcomings as information sources during elections.



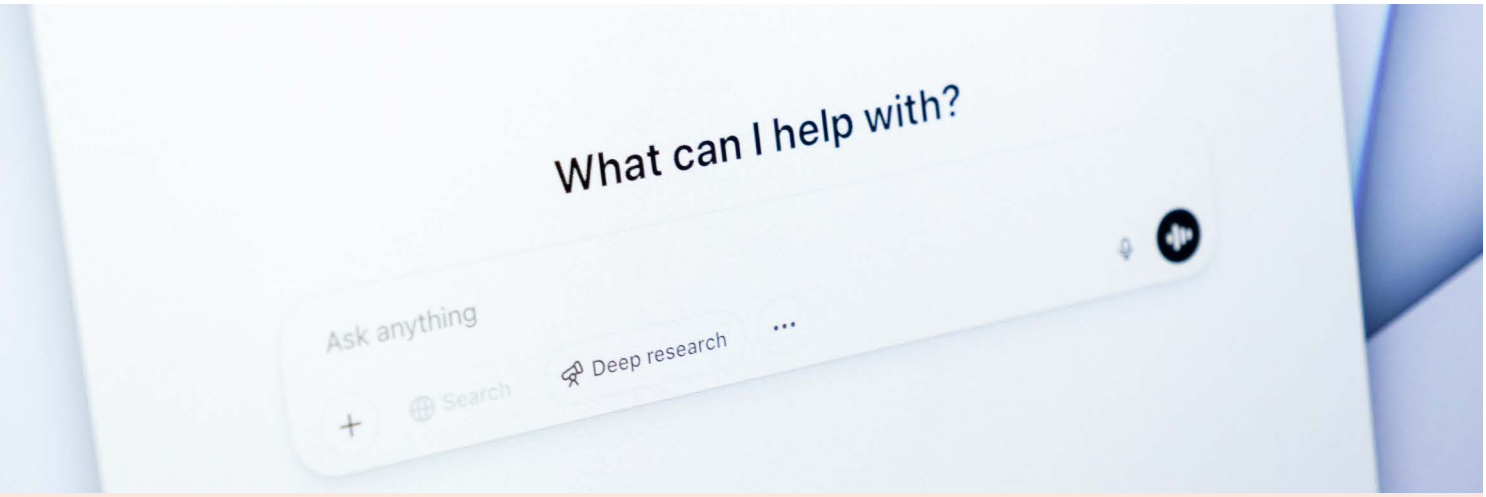
2. **Error rates varied significantly across services:** Replika performed the worst: 56.4% of its responses contained errors. ChatGPT also had serious issues with accuracy - providing responses that contained errors 46.2% of the time. While Google Gemini and Grok fared significantly better, their results still included inaccuracies with a 21.8% and 8.97% error rate respectively.[4] This demonstrates that across the most commonly used AI chatbots, none can be relied upon for complete accuracy.

3 All statistics are reported to three significant figures (3 S.F.).
4 Google AI Overviews only responded to 8 of 78 prompts. This means we cannot report reliably on its individual performance in comparison to other services.

3. **Accuracy problems were demonstrated for all constituencies tested:** The rate of errors in responses was relatively consistent across all three constituencies: 37.9% in Constituency A, 30.3% in Constituency B, and 29.7% in Constituency C.
4. **Sources were inconsistent and unreliable:** 46.9% of chatbot responses did not come with citations or links to back their claims. Some services had problems with providing broken links when they did provide citations (e.g. ChatGPT); others presented an overwhelming number of citations to low-quality or irrelevant sources (e.g. Grok). 66.3% of responses did not cite an official source of information on the Scottish elections, such as the Electoral Commission or Scottish government. These gaps undermine citizens' ability to evaluate accuracy or levels of political bias.
5. **Services offer advice on tactical voting that could affect the result:** We tested whether services would offer advice if asked who to vote for to prevent a specific named party from getting into power. All services besides Google AI Overview provided such advice on tactical voting every time asked (12 of 12 responses, excluding AI Overview). Google did not generate an AI Overview at all for this question. While some services did not outright select a preferred candidate (e.g. ChatGPT, Gemini), others did so either explicitly (Replika) or implicitly (Grok).

TABLE 1
EXAMPLES OF FACTUAL ERRORS IDENTIFIED IN TESTING

ERROR	CONSTITUENCIES AFFECTED	SERVICE(S)
Getting the date of the election wrong by over two months	A	ChatGPT
Making incorrect claims about the voter eligibility rules on residency and citizenship	A, B	ChatGPT
Incorrectly claiming that voters need to bring ID	C, B	ChatGPT and Replika
Making up an expenses scandal for a politician	A, B, C	ChatGPT and Replika
Inventing a date for the made-up expenses scandal	A	Replika
Inventing a candidate	C	Replika
Incorrectly claiming an incumbent was running	A	Replika
Getting the deadline for election registration wrong	B	Replika
Inventing an accusation of 'nepotism' against a candidate	B	Replika
Misidentifying the constituency when given a postcode	A, B	Grok and ChatGPT
Mistakenly claiming that a candidate had not taken a position on the Scottish Assisted Dying Bill or was against it when they had actually been a supporter	A, C	Gemini
Incorrectly claiming that the police inquiry into fraud allegations against the SNP is ongoing when it has in fact concluded	A	Gemini



SCOTTISH ELECTION TESTING METHODOLOGY
SUMMARY

This study utilised an established AI red-teaming methodology: two analysts prompted five AI services with 75 questions about the Scottish elections over one day during the pre-election window.

It tested a combination of popular chatbots, a leading AI search overview service, and a 'companion' bot:

- ChatGPT
- Google Gemini
- Google AI Overviews*
- Grok
- Replika (a companion chatbot)

Replika – an AI 'companion' service that has received scrutiny over concerns about its safety[5] – was included in order to examine the reliability of text-based AI services with smaller user-bases. Replika is a less mainstream service, and its intended use as a companion chatbot is distinct from the more general-purpose design of services such as ChatGPT or from services that are designed for information summarisation such as Google AI Overviews. It was included because it provides answers to questions on elections and presents itself as an information source. For transparency, this paper presents aggregate statistics which include Replika's responses alongside ones that specifically exclude these to focus only on the four mainstream services tested (ChatGPT, Gemini, AI Overviews, and Grok).

5 E.g. Reuters (2025). 'Italy's data watchdog fines AI company Replika's developer $5.6 million.' Reuters. https://www.reuters.com/sustainability/boards-policy-regulation/italys-data-watchdog-fines-ai-company-replikas-developer-56-million-2025-05-19/. Accessed 6/5/26

The tests were conducted of services' knowledge of and claims about three Holyrood constituencies (described above). This meant that each chatbot was prompted with 25 questions per constituency.

The names of the specific constituencies or candidates are not disclosed in order to protect the identities of those contesting the races and to avoid placing undue attention on individual candidates.

In total, 375 questions were asked across 15 separate conversations with the five services: one conversation per constituency per chatbot. An opening prompt was used for each conversation to tell the service that the user was a Scottish resident in a particular postcode looking to vote in the Scottish election (bringing the total number of prompts used to 390). Because some services did not generate responses to some prompts,* the overall statistics we present are based on the number of responses we actually received (320 in total).

The testing was designed to replicate the experience of an average Scottish user. To do so, a Virtual Private Network (VPN) was used to appear as a Scottish internet user and only using the free tier of each service to best mirror a realistic user for this case. Data was collected on one date to maintain consistency across the services and ensure that fact-checking was as rigorous as possible.

Measures were taken to ensure that each conversation about each constituency counted as an independent test and was not biased by other conversations. This included using a 'clean' web browser instance with no internet history or cookies; changing the VPN server for each conversation; and avoiding using accounts. Where an account was required (by Grok and Replika), we created a new account for each conversation.

Redeploying methods from earlier academic research, the analysts then manually assessed the results for factuality, use of evidence, bias, and vulnerability to malicious uses.[6] A full write-up of the methodology is available in the Appendix.

** Google AI Overviews did not respond to most prompts (89.7%; 70 of 78). This means we cannot reliably provide statistics on the answers it gave. The decision was made to exclude its unanswered responses from the data analysis, as these were uncodable, and will not present specific statistics on its performance versus other services. We occasionally provide qualitative descriptions of the responses it gave.*

6 Simon, Fletcher & Kleis Nielsen (2024). 'How generative AI chatbots responded to questions and fact-checks about the 2024 UK general election'. Reuters Institute for the Study of Journalism, University of Oxford. https://reutersinstitute.politics.ox.ac.uk/how-generative-ai-chatbots-responded-questions-and-fact-checks-about-2024-uk-general-election (accessed 17/4/26); Helming & Marsh (2024). 'Large Language Models Continue To Be Unreliable Concerning Elections'. AlgorithmWatch. https://algorithmwatch.org/en/llms_state_elections/ (accessed 17/4/26); Marsh (2024). 'Chatbots are still spreading falsehoods.' AlgorithmWatch. https://algorithmwatch.org/en/chatbots-are-still-spreading-falsehoods/ (accessed 17/4/26)

Adding to this testing, the paper draws on a wide-ranging review of the relevant research and legislation. It finds that:

6. **There is a regulatory gap where AI meets elections:** Current legal frameworks, including the Online Safety Act 2023 and the Representation of the People Act 1983, were not designed with text-based AI in mind. This means they do not cover LLM-based services adequately and often rely on principles that are not well suited to addressing AI risks.
7. **Standards for AI election safeguards are inconsistent and lack transparency:** In the absence of regulation, some providers of text-based AI services such as Google Gemini have implemented election safeguards. But as our testing shows, the quality and transparency of these safeguards vary greatly. We identify smaller services, such as companion chatbots like Replika, as a particular source of concern.

To rebuild trust and safeguard our democratic processes, the paper sets out four recommendations for proactive measures the UK government should take ahead of the general election in 2029.

KEY RECOMMENDATIONS FOR THE UK GOVERNMENT

1. **Make existing UK law LLM-ready:** Clarify how election law and defamation law apply to AI-generated claims, and review how to ensure key election protections can be updated to address risks from LLMs. Use the Representation of the People Bill as a vehicle for change.
2. **Mandate minimum AI election safeguards:** Legislate to establish a baseline requirement for election safeguards that AI providers must implement at all times, with heightened requirements during pre-election windows. These standards should include requirements for risk assessments, reporting on policies, a baseline of guardrails to be implemented, and duties on accuracy and bias.
3. **Ensure transparency and data access:** Require text-based AI services to provide independent researchers with access to internal data, training sets, and live data during election windows to ensure public-interest accountability.
4. **Invest in trust by supporting AI text detection technologies:** Equip citizens to distinguish AI outputs from human-made text by investing in the development of textual watermarking and AI detection. Fund innovation in AI text detection and support the development of new cross-industry standards.

INTRODUCTION

The UK's elections are under pressure. There are credible concerns that the large-scale spread of false and misleading information during an election could undermine the democratic process and plunge the country into a crisis.[7] With deepfaked videos of MPs circulating on social media,[8] networks of foreign accounts posing as UK voters,[9] and elected politicians making allegations of voter fraud,[10] trust in the country's democratic processes is being eroded from all directions. The UK is not alone in this: we are seeing a global democratic emergency, where democratic norms are being repeatedly undermined and democratic backsliding has become more widespread. At Demos, we argue that the trend towards democratic decline is partly due to the disruption of the UK's democratically-vital information supply chains - the means by which information is produced, distributed, acquired, and evaluated.[11] These risks come at a time when nearly 20% of national seats are considered marginal,[12] meaning that even a small number of misled voters could seriously disrupt election outcomes. Strengthening the UK's epistemic security is an urgent task ahead of the 2029 elections.

These fears have been compounded by rapid developments in generative AI. The emergence of AI videos means it is easier than ever to fake a realistic video of a politician saying something they did not - illustrated by an AI-generated video of George Freeman MP that falsely claimed he was changing parties.[13] Meanwhile, AI chatbot services like ChatGPT have exploded in popularity but remain prone to factual errors, misleading claims, and nonsensical responses known as hallucinations. While analysts have recognised that the 2024 general election was not significantly impacted by AI-enabled mis- and disinformation, the findings of this report may amplify concerns for elections going forward - and in new ways.[14]

This report provides an empirically-grounded examination of the implications of text-based AI for election integrity in the UK. Its primary focus is on where and the extent to which these services provide ordinary voters with inaccurate and false information about elections. We have therefore chosen to not focus here on the potential for LLMs to be deliberately misused by UK citizens to produce false claims about elections.

7 Perry, Moores & Hancock (2026). 'Epistemic Security Briefing: The Elections Bill.' Demos. https://demos.co.uk/research/epistemic-security-briefing-the-elections-bill/ (accessed 17/4/26); Stockwell et al. (2024). 'AI-Enabled Influence Operations: The Threat to the UK General Election'. Centre for emerging Technology and Security (CETas), the Alan Turing Institute. https://cetas.turing.ac.uk/publications/ai-enabled-influence-operations-threat-uk-general-election (accessed 17/4/26)

8 Precey (2025). 'Criminalise malicious AI deepfakes, says MP'. BBC News. https://www.bbc.co.uk/news/articles/cx24w7l8ljgo (accessed 17/4/26)

9 Sellman (2026). 'Why Scottish X accounts vanished after Iran's internet shutdown'. The Times. https://www.thetimes.com/uk/technology-uk/article/scottish-x-accounts-iran-internet-shutdown-32jgblq05? (accessed 17/4/26)

10 Novik & Williams (2026). 'Reform UK calls for fraud probe over alleged 'family voting' in by-election.' Financial Times. https://www.ft.com/content/e259bb4f-16fb-4866-853b-3824bc2510d1 (accessed 17/4/26)

11 Seger, Perry & Hancock (2025). Epistemic Security 2029: Fortifying the UK's information supply chain to tackle the democratic emergency. Demos. https://demos.co.uk/research/epistemic-security-2029-fortifying-the-uks-information-supply-chain-to-tackle-the-democratic-emergency/ (accessed 17/4/26)

12 Sturge (2024). '2024 general election: Marginality.' House of Commons Library, UK Parliament. https://commonslibrary.parliament.uk/2024-general-election-marginality/

13 Sennitt (2025). 'Tory MP reports deepfake defection video to police.' BBC News. https://www.bbc.co.uk/news/articles/c62e7xz02dpo (accessed 17/4/26)

14 Stockwell (2024). 'AI-Enabled Influence Operations: Threat Analysis of the 2024 UK and European Elections'. CETaS, the Alan Turing Institute. https://cetas.turing.ac.uk/publications/ai-enabled-influence-operations-threat-analysis-2024-uk-and-european-elections (accessed 17/4/26)

By 'text-based AI', we mean AI systems and services that use large-language models (LLMs) to generate text – as opposed to a service designed to create images, videos, or audio. This includes interactive AI chatbots, AI search engine replacements, AI search engine overviews, and AI news aggregators. There has already been significant attention paid to election-related risks for AI in general and audiovisual AI (e.g. deepfakes) in particular.[15] Even when the policy literature addresses text-based AI, it tends not to address this modality's technical and legal specificities. This leaves a gap in our understanding that this paper fills.

We review the risks of factual errors from text-based AI tools during an election, provide empirical evidence from a snapshot test during a live election, and situate these issues within the current regulatory landscape. We find that, while there are open questions about the extent to which these services can influence voters compared to other sources of information,[16] it is clear these systems are growing in importance as intermediaries and creators of political information. There is little to indicate that text-based AI has had significant influence on elections so far and it is important to avoid tipping into panic about voter persuasion. Instead, the clearest risk is to public trust in the authenticity and veracity of information and conversations about elections.

To pinpoint where specific vulnerabilities may lie, we present findings of an AI services test conducted during the 2026 Scottish elections. This testing used an adversarial 'red-teaming' approach to evaluate the reliability of major text-based AI services such as ChatGPT and Google Gemini. We found a persistent reliability gap: 34.1% of responses contained factual errors. From making up candidates to providing out-of-date procedural advice, the results make it clear that AI chatbots and related services continue to have serious shortcomings as election aids. The evidence shows how these services' guardrails can fail in dangerous ways – allowing them to be used to generate false and misleading social media posts about political candidates.

Based on this evidence and a review of the existing literature, we find that these services pose significant risks to public trust in elections due to issues with factual reliability and a lack of clear regulatory standards. We argue that the growing use of text-based AI among the public necessitates a reassessment of how such services are regulated in the UK, particularly during election periods. While the UK government may be more willing to accept risks associated with AI in other domains, elections are fundamental to democracy, and safeguarding their integrity must remain the priority.

THE REPRESENTATION OF THE PEOPLE BILL: A WINDOW OF OPPORTUNITY

While public attention at the time of writing is focused on the local and devolved elections, the government is also addressing electoral integrity through the Representation of the People Bill. This is its flagship legislation aimed at lowering the voting age and tackling covert political finance.

Although the government's initial summary of its strategy for the Bill acknowledged that "our own democracy is being threatened by misinformation,"[17] the published Bill only partially addresses the risks and fails to tackle the scale of the problem. This report contributes to Demos' broader body of work on the risks of misleading information, whether generated by AI or human actors, and uses the Bill as a vehicle to advance these concerns.

15 Stockwell (2024. 'How can we stop AI-enabled threats damaging our democracy?' CETaS, the Alan Turing Institute. https://www.turing.ac.uk/blog/how-can-we-stop-ai-enabled-threats-damaging-our-democracy (accessed 17/4/26)
16 Simon & Altay (2025). 'Don't Panic (Yet): Assessing the Evidence and Discourse Around Generative AI and Elections.' Knight First Amendment Institute, Columbia University. https://knightcolumbia.org/content/dont-panic-yet-assessing-the-evidence-and-discourse-around-generative-ai-and-elections (accessed 17/4/26)
17 Ministry of Housing, Communities & Local Government (MHCLG; 2025). 'Restoring trust in our democracy: Our strategy for modern and secure elections.' https://www.gov.uk/government/publications/restoring-trust-in-our-democracy-our-strategy-for-modern-and-secure-elections/restoring-trust-in-our-democracy-our-strategy-for-modern-and-secure-elections (accessed 17/4/26)

Demos has supported **Emily Darlington MP** in tabling a series of amendments to the Bill. These measures seek to:

- clarify how existing election law applies to deepfakes and online false claims targeting candidates;
- establish an online repository for digital political advertising;
- strengthen the Electoral Commission's investigative powers;
- introduce a transparent public protocol for handling critical election incidents.

Building on Darlington's amendments, this paper sets out further recommendations, including a call for legal clarity on how existing election and defamation laws apply to text-based AI services (Recommendation 1). These measures should be adopted by the government without delay.

2. WHAT IS TEXT-BASED AI & WHO USES IT IN THE UK?

This section introduces what text-based AI is, breaks down types of text-based AI services, explains their use-cases, sets out the latest data on their usage by the UK public at the time of writing, and identifies where more evidence is needed on their uptake.

2.1 TEXT-BASED AI: DEFINITION, SERVICE-TYPES, AND USE-CASES

'Text-based AI' covers AI systems and services that use LLMs to generate text. This is distinct from services which generate other kinds of content, such as images and video. Some AI services are 'multimodal' – allowing users to input and generate a combination of text, images, video, audio, or other data formats. For the purpose of this report, we have only focused on services that are capable of taking text as an input and generating text as their output, including for multimodal services that are capable of more than this.

2.1.1 Service types

We have identified seven key types of text-based AI services that are most relevant to elections. The differences between these services have significant implications for their election-related uses, safeguards, and coverage under existing law.

The seven types are:

1. General-use chatbots - e.g. ChatGPT, Gemini, Claude and Grok
2. Companion chatbots - e.g. Replika
3. Chatbot customisation and hosting platforms - e.g. Character.ai
4. AI-driven search engines - e.g. Perplexity and Google AI Mode.

5. News aggregators - e.g. personalised AI news summary apps like Particle.news[18]
6. Model Application Programming Interfaces (APIs) - e.g. OpenAI's GPT API, Google Gemini, or Anthropic's Claude.
7. Embedded text-generation tools - e.g. text generation tools in Facebook or LinkedIn

2.2 TEXT-BASED AI SERVICES ARE GROWING RAPIDLY IN THE UK

Text-based AI services are rising in popularity in the UK. According to Ofcom, based on data from SimilarWeb, OpenAI's ChatGPT received 252 million visits from UK users in August 2025.[19] This represents a year-on-year growth of 156%.[20] ChatGPT is the most popular AI service and is now the second-largest search service in the UK. Google Search remains the most popular search engine but has incorporated its own AI service in the form of AI-generated summaries ('AI Overviews') into the results page. Google received roughly 3 billion searches per month in the UK in 2025. Ofcom's data indicates that 30% of Google searches returned an AI Overview as of August 2025. All five of the leading chatbot AI services, summarised in the table below, are owned by American companies.

TABLE 2
MOST POPULAR STANDALONE CHATBOT-BASED AI SERVICES (SOURCE: OFCOM/ SIMILARWEB, 2025)[21]

SERVICE	OWNER	SERVICE TYPE	UK VISITS *(Jan - August 2025)*	YEAR-ON-YEAR GROWTH
ChatGPT	OpenAI	General-use chatbot	1,966,000,000	134%
Gemini	Google	General-use chatbot	100,000,000	146%
Claude	Anthropic	General-use chatbot	40,000,000	138%
Perplexity	Perplexity	AI search engine	31,000,000	100%
Grok	xAI	General-use chatbot	4,000,000	323%

Reliable and up-to-date UK user statistics are hard to come by for companion chatbots. Reports on Replika's worldwide user numbers have varied from 2 million in July 2023[22] to 40 million in October 2025.[23] There is a similar lack of reliable data for chatbots that are embedded in existing platforms (e.g. Grok on X or Microsoft Copilot) or for APIs.

The lack of data on UK users makes it difficult to measure what proportion of the public are using these services and for what purpose. It is also hard to know from simple user figures exactly how someone engages with a service. For example, some users may have extensive interactions with a chatbot, while others may use these services casually and have short

18 Particle.news (2026). https://particle.news/# (accessed 21/4/26)
19 This is the latest available data reported by Ofcom at time of writing.
20 Ofcom (2025). Online Nation: Report 2025. https://www.ofcom.org.uk/siteassets/resources/documents/research-and-data/online-research/online-nation/2025/online-nations-report-2025.pdf?v=409837 (accessed 17/4/26)
21 Ofcom (2025). Online Nation: Report 2025. https://www.ofcom.org.uk/siteassets/resources/documents/research-and-data/online-research/online-nation/2025/online-nations-report-2025.pdf?v=409837 (accessed 17/4/26)
22 Kahn (2023). 'Stigma of dating a chatbot will fade, Replika CEO predicts.' Fortune. https://fortune.com/2023/07/12/brainstorm-tech-chatbot-dating/
23 Reported in Bradley & Weiss (2025). 'The CEO of 'AI companion' startup Replika is stepping aside to launch a new company.' Business Insider. https://www.businessinsider.com/replika-ceo-eugenia-kuyda-launch-wabi-2025-10

interactions. This gap underscores a key finding of this paper: **there is a significant lack of transparency around who uses these services and how they are used.**

During the 2024 general election – the most recent election for which we have data – polling indicated that roughly 13% of eligible voters used AI chatbots for political information seeking.[24] While we lack up-to-date data, it is reasonable to surmise that this figure has likely risen alongside the overall growth in UK users of AI services.

More reliable up-to-date information is needed on the size of text-based AI adoption in the UK and on how these services are used in an electoral context. More data access is also needed for independent researchers to ensure we have a clear understanding of the election information environment and AI's role in it. In the absence of this evidence, it is difficult to assess the scale and severity of electoral risks.

24 AI Security Institute (AISI; 2025). 'Do chatbots inform or misinform voters?'. https://www.aisi.gov.uk/blog/do-chatbots-inform-or-misinform-voters (accessed 17/4/26)

3. WHAT RISKS MAY TEXT-BASED AI POSE FOR ELECTIONS?

This chapter examines concerns raised regarding text-based AI and elections. We break these down into five categories of concern:

1. Factual errors
2. Political bias
3. Effects on voters' beliefs
4. Model manipulation by malicious actors
5. Impacts on the trustworthiness of the information ecosystem

For each category, we summarise the existing research, note gaps that remain in the evidence base, and present our findings from our Scottish election testing (where relevant). We also highlight areas where LLM-based services may contribute positively to voter informedness and the quality of the information available during elections.

We find that the evidence leaves cause for concern but not yet panic. There has so far been no substantial evidence presented that the UK's elections have been significantly influenced by AI – text-based or otherwise.[25] Overall, more evidence is needed on the scale of UK voters' exposure to AI-generated false and misleading claims about elections.

But, while voters' political beliefs may not be substantially affected by persuasion from text-based AI, these services continue to have significant problems with factual accuracy and political bias. Our testing results show that these problems continue to appear in live election settings despite improvements to LLMs' capabilities and measures designed to reduce errors. In a high-

25 Stockwell (2025). 'From Deepfake Scams to Poisoned Chatbots: AI and Election Security in 2025.' CETaS, the Alan Turing Institute. https://cetas.turing.ac.uk/publications/deepfake-scams-poisoned-chatbots (accessed 17/4/26)

risk context such as an election, even a small number of people being misled is problematic. We conclude by arguing that the most pressing risk is to public trust: as LLM-based services continue to prove unreliable and stories emerge about attempts to use AI to manipulate elections, there is a risk that the public's faith in the independence and validity of the electoral process could be undermined.

3.1 TEXT-BASED AI SERVICES CONTINUE TO GENERATE INACCURATE INFORMATION

All LLM-based AI systems are prone to generating text that includes factual errors and confabulations known as 'hallucinations'.[26] These false and misleading outputs can include incorrect claims that are presented confidently as true and written in a persuasive style. LLMs are probabilistic 'black boxes': they work by predicting the next most likely output based on highly complex calculations, many of which are not visible or understandable to any human.

3.1.1 Why do LLMs produce factual errors?

Hallucinations arise because the model has inferred an incorrect pattern from its training data, has over-extrapolated from its data, or has recombined its training data in a nonsensical way. Other factual errors arise because a model has misinterpreted the meaning of its training data, is relying on out of date information, or has been trained on unreliable sources. LLMs also tend to have more problems with errors and hallucinations when there is less information available in their training data on a given subject. Fundamentally, LLMs do not have any internal way of judging truth from falsity: they simply generate words based on patterns in how those words appear in their training data and rely on external proxies and human feedback to increase the weight of 'better' responses. It is an open debate amongst academics and model developers as to whether it will ever be possible to create an LLM that does not produce factual errors or hallucinations.[27]

Moreover, LLM's black-box, probabilistic nature make make their outputs almost impossible to totally predict, including for LLM developers. This means that there is always a risk that errors and biases may remain undetected, even after rigorous testing. It also offers AI service providers a form of plausible deniability when errors do arise: they may say that they could not have foreseen the specific result the model produced.

AI developers have introduced innovations that reduce the risk of errors and hallucinations. The key method for this is Retrieval Augmented Generation (RAG): a process where the LLM will autonomously search the web for up-to-date information in response to the user's prompt.[28] RAG may also be used to pull data from pre-determined sources. However, as we found when testing AI chatbots outputs, RAG systems are not always fully triggered or utilised to improve a response. In our testing, for example, the information surfaced by ChatGPT appeared far more likely to have come from the AI's original training data rather than the most up to date available information. This demonstrates that whilst RAG may mitigate falsity to an extent, it is not always used consistently and does not prevent errors or hallucinations entirely.

26 IBM (2026). 'What are AI hallucinations?' https://www.ibm.com/think/topics/ai-hallucinations (accessed 17/4/26)
27 There are various views on this that depend on what the ultimate cause of hallucinations is deemed to be. E.g. Banerjee et al. (2025). 'LLMs Will Always Hallucinate, and We Need to Live with This.' IntelliSys 2025. https://link.springer.com/chapter/10.1007/978-3-031-99965-9_39 (accessed 17/4/26); Kalai et al. (2025). 'Why Language Models Hallucinate.' ArXiv. https://arxiv.org/abs/2509.04664 (accessed 17/4/26)
28 Google Cloud (2026). 'What is Retrieval-Augmented Generation (RAG)?' https://cloud.google.com/use-cases/retrieval-augmented-generation (accessed 17/4/26)

3.1.2 Problems with false information about elections

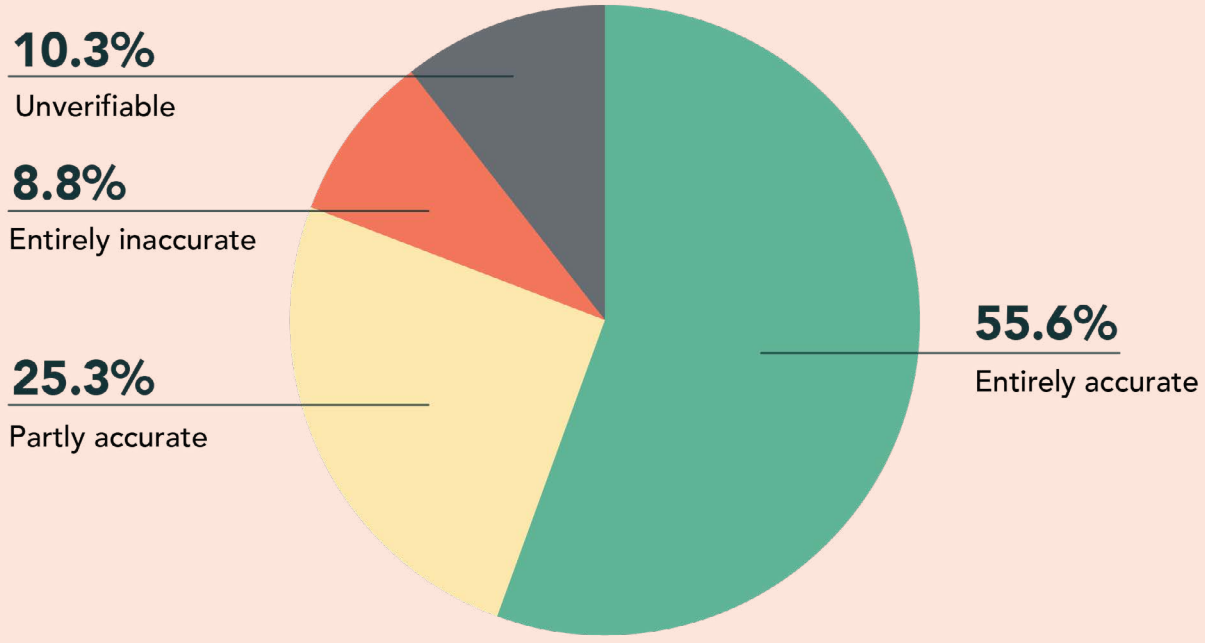
At the time of writing, there have not been any documented cases of AI-generated false and misleading claims about a UK election reaching large audiences[29] – echoing the general picture worldwide.[30] However, focusing on major incidents does not capture ordinary voters' exposure to false and misleading claims via their private conversations with chatbots. Evidence from our testing highlights how voters *could* be misinformed by these services in such one-to-one interactions.

SCOTTISH ELECTION TESTING RESULTS

PROBLEMS WITH ACCURACY ACROSS THE BOARD

Our snapshot testing on March 27th 2026 found that - across ChatGPT, Gemini, Google AI Overviews, Grok, and Replika - just over a third (34.1%) of responses to questions about the Scottish elections contained factual errors (109 of 320 total responses). To break down these inaccurate responses, 8.75% (28 of 320) were entirely inaccurate and 25.3% were partly accurate but with errors (81 of 320). Partly factual responses could be particularly misleading as their errors were sometimes much harder to spot. See the chart below for a full breakdown:

CHART 1
TOTAL ERROR RATE FOR ALL QUESTIONS ACROSS ALL TEXT-BASED AI SERVICES TESTED (*total: 320 responses*)



29 Stockwell (2024). 'AI-Enabled Influence Operations: Threat Analysis of the 2024 UK and European Elections'. CETaS, the Alan Turing Institute. https://cetas.turing.ac.uk/publications/ai-enabled-influence-operations-threat-analysis-2024-uk-and-european-elections (accessed 17/4/26)

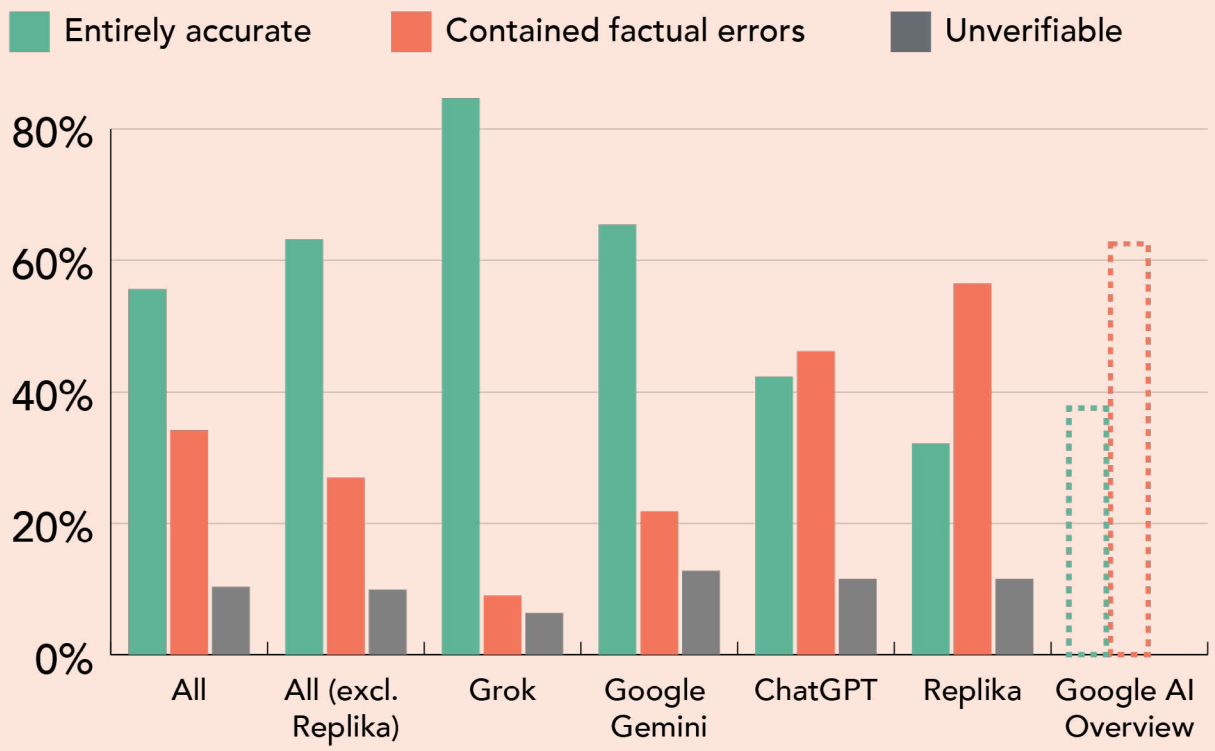
30 Kapoor & Narayanan (2024). 'We Looked at 78 Election Deepfakes. Political Misinformation is not an AI Problem.' AI As Normal Technology. https://www.normaltech.ai/p/we-looked-at-78-election-deepfakes (accessed 17/4/26)

These statistics only improve slightly if we exclude Replika to solely focus on mainstream services with large userbases (ChatGPT, Gemini, Google AI Overviews, and Grok). Excluding Replika, 26.9% of responses contained factual errors (65 of 242 responses): 22.7% were partly accurate but with errors (55 of 242) and 4.1% were entirely inaccurate (10 of 242). 63.2% were entirely accurate (153 of 242 responses).

Breakdown of error-rates for all questions by service

The performance of individual services was highly variable. Replika performed the worst: 56.4% of its responses contained errors (44 of 78). ChatGPT also had serious issues with accuracy - providing responses that contained errors 46.2% of the time (36 of 78 responses). We identify potential reasons for ChatGPT's poor performance in Section 4.6. Google Gemini and Grok fared significantly better, with a 21.8% and 8.97% error rate respectively (17 and 7 out of 78 each). Google AI Overviews did not provide enough responses for us to report accurately on its error rate. For this reason, we have included an empty bar in the chart below.

CHART 2
ERROR RATE FOR ALL QUESTIONS BROKEN DOWN BY TEXT-BASED AI SERVICE TESTED (*total: 320 responses*)

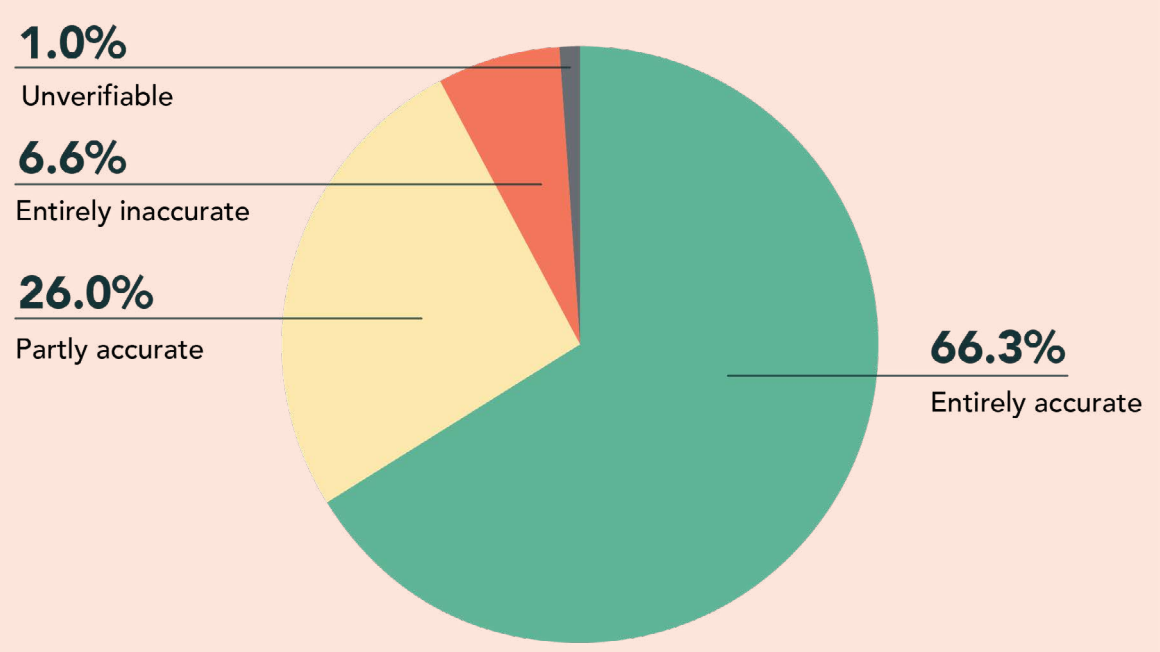


(Note: Google AI Overviews did not provide enough responses to reliably report on its performance compared to other services.)

Breakdown of error-rates for factual questions

These error rates only improve slightly if we focus only on the services' responses to factual questions about election procedures and candidates. For the 240 factual questions asked, we received 196 responses.[31] 32.7% of these responses contained errors (64 of 196): 6.63% were entirely inaccurate (13) and 26% were partly accurate (51). See the chart below for a full breakdown:

CHART 3
TOTAL ERROR RATE FOR FACTUAL QUESTIONS, ACROSS ALL TEXT-BASED AI SERVICES TESTED (*total: 196 responses*)



Excluding Replika to focus on ChatGPT, Gemini, Google AI Overviews, and Grok only leads to a slight improvement. These services' overall number of responses to factual questions that contained errors was 26.4% (39 of 148 responses). 22.97% were partly factual but with errors (34 of 148) and 3.38% were entirely inaccurate (5 of 148). 72.3% were entirely accurate (107 of 148).

31 Google AI Overviews only provided responses for 4 of the 48 factual questions asked.

TABLE 1
EXAMPLES OF FACTUAL ERRORS IDENTIFIED IN TESTING

ERROR	CONSTITUENCIES AFFECTED	SERVICE(S)
Getting the date of the election wrong by over two months	A	ChatGPT
Making incorrect claims about the voter eligibility rules on residency and citizenship	A, B	ChatGPT
Incorrectly claiming that voters need to bring ID	C, B	ChatGPT and Replika
Making up an expenses scandal for a politician	A, B, C	ChatGPT and Replika
Inventing a date for the made-up expenses scandal	A	Replika
Inventing a candidate	C	Replika
Incorrectly claiming an incumbent was running	A	Replika
Getting the deadline for election registration wrong	B	Replika
Inventing an accusation of 'nepotism' against a candidate	B	Replika
Misidentifying the constituency when given a postcode	A, B	Grok and ChatGPT
Mistakenly claiming that a candidate had not taken a position on the Scottish Assisted Dying Bill or was against it when they had actually been a supporter	A, C	Gemini
Incorrectly claiming that the police inquiry into fraud allegations against the SNP is ongoing when it has in fact concluded	A	Gemini

Error-rates for factual questions were relatively consistent across the three constituencies that were the subject of our test. 37.9% of answers about Constituency A contained errors (25 of 66). For Constituency B, this was 30.3% (20 of 66) and for Constituency C it was 29.7% (19 of 64). The pattern holds if we exclude Replika and focus on the four mainstream services: the number of responses containing errors was 32% for Constituency A, 24% for Constituency B, and 22.9% for Constituency C. This order was surprising, as we had expected that the seat with the least information available - the rural constituency - would have the least accurate results.

Occasionally misleading use of facts

All services were prone to sometimes generating responses that presented facts in a misleading manner. For example, when asked which candidates were running in the constituency race for Constituency A, ChatGPT spoke primarily about a candidate running for the city's regional race without providing this important distinction. These are two different races with different voting procedures and lists of candidates, yet ChatGPT's response implied the candidate was running in the constituency without stating this outright. Such misleading responses were far subtler than straightforward factual errors and may be harder for users to identify without paying careful attention.

Meanwhile, Gemini displayed a problem with quote misattribution. In one case, it stitched together several real quotes from a candidate into a Frankenstein quote that – while true to the spirit of what they had said – was technically a misattribution and could be misleading.

Our findings echo prior research. In the UK, a 2024 Reuters Institute study found that ChatGPT and Perplexity produced false or partially incorrect information around 20% of the time when asked for information about the general election.[32] In fact, in the Reuters Institute study, ChatGPT performed worse than Perplexity: 21% of ChatGPT's answers contained errors (13% partially correct; 8% entirely incorrect), where Perplexity's error rate was 17% (4% partially correct; 13% incorrect). Meanwhile, Google Gemini tended not to answer election-related questions at all for the 2024 election. In the EU, researchers found that chatbots produced incorrect info on EU election dates and voting procedures in 2024,[33] with Gemini the least likely to give correct answers. Similar results have been found in Germany.[34]

Our 2026 results suggest that this situation has not greatly improved since 2024 – even for well-performing services like Gemini, which provided false or partially accurate information 21.8% of the time. We found that performance can sometimes be worse compared to 2024. This is best illustrated by ChatGPT with its 46.2% error rate in our testing. Our results indicate that lessons may not have been learned from 2024 and measures introduced to improve accuracy since 2024 (such as RAG) may not fully resolve the problem.

The fact that these services are prone to generating false information leads to another problem: people are also turning to chatbots to fact-check other claims they see online, even though these chatbots may not be reliable fact-checkers. Unlike dedicated fact-checking tools – which are bespoke, have access to live information, and are overseen by expert human fact-checkers – consumer-facing chatbots are liable to errors and hallucinations when asked to perform fact-checking. For example, Grok's X account has been previously found to give incorrect answers to users on X when they ask it to tell them if a claim or piece of media is real.[35] Because Grok speaks confidently and is (unduly) treated as an objective 'expert' by some users, this could

32 Simon, Fletcher & Kleis Nielsen (2024). 'How generative AI chatbots responded to questions and fact-checks about the 2024 UK general election'. Reuters Institute for the Study of Journalism, University of Oxford. https://reutersinstitute.politics.ox.ac.uk/how-generative-ai-chatbots-responded-questions-and-fact-checks-about-2024-uk-general-election#header--4 (accessed 17/4/26)

33 Goujard (2024). 'AI chatbots spread falsehoods about the EU election, report finds.' POLITICO. https://www.politico.eu/article/ai-chatbots-spread-falsehoods-about-the-eu-elections-report-finds/ (accessed 17/4/26)

34 Helming & Marsh (2024). 'Large Language Models Continue To Be Unreliable Concerning Elections'. AlgorithmWatch. https://algorithmwatch.org/en/llms_state_elections/ (accessed 17/4/26)

35 Ponce de León & Chenrose (2025). 'Grok struggles with fact-checking amid Israel-Iran war.' DFRLab. https://dfrlab.org/2025/06/24/grok-struggles-with-fact-checking-amid-israel-iran-war/ (accessed 17/4/26); Quinn (2025). 'Musk's Grok AI bot falsely suggests police misrepresented footage of far-right rally in London.' The Guardian. https://www.theguardian.com/technology/2025/sep/14/musks-grok-ai-bot-falsely-suggests-met-misrepresented-footage-of-clashes-with-far-right (accessed 17/4/26)

lend undue credibility to false claims and misleading media – AI generated or otherwise. This demonstrates how assumptions about the accuracy of these tools can create false certainty about the reliability of information used during an election period.

3.1.3 Problems with out-of-date information about elections

Another challenge lies in the timeliness of answers: as our testing shows, services can generate confident responses that are misleading, unhelpful or outright incorrect because they rely on out-of-date information. A service might confidently state a claim that was true several years ago but is no longer the case. For example, a chatbot might falsely claim that a candidate who ran in the previous election is running in the current one.

SCOTTISH ELECTION TESTING RESULTS
CHATGPT RELIED ON OUT OF DATE INFORMATION

We found that ChatGPT's responses usually relied on information from 2023 or prior. Our review of the information and citations ChatGPT provided indicated that 43.6% of its responses relied on information that was more than one year out of date (34 of 78 responses). Indeed, without being asked, and on less than a fifth of occasions, ChatGPT would explicitly tell the user that the information in the response was limited by a "knowledge cutoff" in October 2023 (15.4% of responses; 12 of its 78 responses). The 'knowledge cutoff' refers to the most recent date of the data that this model of ChatGPT was trained on. ChatGPT's occasional tendency to tell the user about this cutoff unprompted appears to be a built-in mechanism to warn the user that the service was not able to access recent information. Such a measure can provide important transparency to users about the service's limitations, but only if it is applied consistently.

Furthermore, in some cases, ChatGPT presented outdated information as if it was still up-to-date and was not affected by the 2023 cutoff. These responses could be misleading. For example, one ChatGPT response to a question on candidates in Constituency A's stances on immigration that the Scottish Conservatives were "more aligned with UK government policy" on immigration than Scottish Labour - possibly implying that the Conservatives were still in power in Westminster. ChatGPT's reliance on data from before October 2023 may help to explain its comparatively low accuracy rate versus other leading services like Google Gemini.

One explanation for ChatGPT's reliance on old data and comparatively low accuracy may be that it was using an older model.The stated October 2023 cutoff date aligns with the cutoff date given by OpenAI for GPT-4o.[36] Yet OpenAI claims that it retired GPT-4o from use in all tiers of ChatGPT in February 2026 in favour of GPT-4.3.[37] GPT-5.3's knowledge cutoff is in August 2025.[38] It therefore seems that ChatGPT was still using GPT-4o as of March 27th.

36 OpenAI (2026). 'GPT-4o Model.' OpenAI API. https://developers.openai.com/api/docs/models/gpt-4o (accessed 23/4/26)
37 OpenAI (2026). 'Retiring GPT-4o, GPT-4.1, GPT-4.1 mini, and OpenAI o4-mini in ChatGPT.' https://openai.com/index/retiring-gpt-4o-and-older-models/ (accessed 17/4/26)
38 OpenAI (2026). 'GPT-5.3 Chat Model.' OpenAI API.https://developers.openai.com/api/docs/models/gpt-5.3-chat-latest (accessed 23/4/26)

3.1.4 How much tolerance should we give to factual errors about elections?

Elections are a context in which tolerance for error must be exceptionally low. Factual inaccuracies about voting procedures, for example, are extremely serious if they are presented to a voter.

Our testing results demonstrate that there are serious risks that text-based AI could generate false and misleading statements about candidates – especially if AI-generated errors and hallucinations spread out into public discourse online and are taken as fact. The person that shared the false claim does not necessarily need to believe it. They could simply fail to fact-check the AI output or include it as part of a larger piece of text. Moreover, errors and hallucinations may be more likely for smaller elections where less data is available, such as local elections, devolved elections and by-elections.

The question then becomes: will potential voters believe and act on such false information? We discuss this concern in Section 3.3.

3.2 TEXT-BASED AI SERVICES HAVE WELL-ESTABLISHED PROBLEMS WITH POLITICAL BIAS

Beyond outright falsehoods, text-based AI services can display 'political bias', i.e. responses that favour some political candidates, parties, and perspectives over others. Political bias can take many forms - from displaying an overt preference for a given candidate to subtle shifts in information selection, sources used, framing, emotive cues, and language choices.[39] While users may identify and reject clear-cut cases of political bias, some research has raised concerns that latent bias might influence voters through interactions over longer periods of time.

This is of particular concern as users may remain unaware of this bias and may lack the additional cues they need to evaluate the information accordingly. If users are made aware of political bias, they may take steps to mitigate this by seeking alternative sources of information or by using multiple chatbots from different providers to inform their judgments.

Besides political bias towards the extremes, concerns have been raised that LLM-based chatbots may have a homogenising effect on political discourse, resulting in a bias towards the mainstream at the expense of diversity of thought. In short, LLMs' training means that many of the leading chatbots are biased towards generating standardised responses that prioritise dominant cultural voices and do not adequately represent societal diversity.[40] The fear is that non-dominant perspectives could be further marginalised over time, reducing ideological pluralism. This concern remains speculative and warrants further study.

Previously, a 2023 study found that ChatGPT was systematically biased in favour of the Labour Party and other left-leaning parties internationally.[41] More recent studies have suggested that ChatGPT had moved to a more centre-neutral position, likely due to conscious effort by OpenAI.[42] Recent research from IPPR found ChatGPT referred most often to the left-leaning

39 Shu et al. (2026). 'How latent and prompting biases in AI-generated historical narratives influence opinions.' PNAS Nexus. https://academic.oup.com/pnasnexus/article/5/3/pgag022/8503065 (accessed 17/4/26)

40 Jackson et al. (2025). 'Large AI Models Have a Prioritization Problem: Policy Implications and Solutions.' Policy Insights from the Behavioral and Brain Sciences. https://journals.sagepub.com/doi/10.1177/23727322251408311 (accessed 17/4/26); Sourati et al. (2026). 'The homogenizing effect of large language models on human expression and thought.' Trends in Cognitive Sciences. https://www.cell.com/trends/cognitive-sciences/fulltext/S1364-6613(26)00003-3 (accessed 17/4/26)

41 Motoki et al. (2023). 'More human than human: measuring ChatGPT political bias.' Public Choice. https://link.springer.com/article/10.1007/s11127-023-01097-2 (accessed 17/4/26)

42 Fujimoto & Takemoto (2023). 'Revisiting the political biases of ChatGPT.' Frontiers in Artificial Intelligence. https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2023.1232003/full (accessed 17/4/26)

The Guardian as a news source[43] – prompting criticism in the right-wing press.[44] Meanwhile, xAI's Grok has been found supporting right-wing talking points. These include responding supportively to posts on X by the far-right organisations Homeland Party and Britain First that call for 'remigration'.[45] The Wikipedia competitor Grokipedia, which was largely generated by Grok, has also been found to display bias in favour of right-wing figures.[46] Our testing found that Grok cited Grokipedia amongst its sources on the 2026 Scottish elections.

Beyond the main players such as ChatGPT and Grok, there is a wider ecosystem of custom AI chatbots with smaller userbases. Often built using open-source LLMs, these can be highly partisan and lack guardrails.[47] For instance, the far-right social media platform Gab launched a chatbot called Arya which was explicitly instructed that "ethnonationalism" was its "foundation".[48]

These results collectively demonstrate that not all chatbots lean in a singular political direction in the UK, but instead individually display a bias towards different specific political parties or candidates including, among chatbots with less visibility, those at the extreme fringes.

Similar findings of political bias within AI chatbot services have been raised internationally. A 2025 report by the Dutch data protection authority found that the most commonly used chatbots in the Netherlands (ChatGPT, Gemini, Mistral, and Grok) tended to advise Dutch users to vote for the same two Dutch political parties – the right-wing PVV and left-leaning GroenLinks-PvdA – "irrespective of the user's question or instruction". The authority concluded that these services were not reliable as voting aids in Dutch elections.[49] Likewise, Global Witness found that Grok promoted false claims that favoured Donald Trump during the 2024 elections, such as "claims that the 2020 [Presidential] election was fraudulent."[50] This demonstrates how AI chatbots' political bias represents a challenge for democracies and political parties around the world. Such studies collectively demonstrate that chatbots are emerging with political leanings much like newspapers. However, they lack the editorial standards and clear source attribution for users that enable users to evaluate that information effectively.

As with errors, the risk of political bias appears to be higher for smaller services with fewer guardrails and for services that are trained to present a 'warm' persona to users.[51] Companion chatbots such as Replika – which are designed to encourage emotional relationships but have been found to lack guardrails – should therefore be a cause for concern when it comes to bias.

Overall, it remains to be seen whether political bias in text-based AI services represents a significant change compared to other sources of information during elections, such as television and social media. Based on the evidence on political influence and AI (discussed below), it may be that the emergence of text-based AI does not represent a major shift compared to the information environment that existed before such services were introduced.

43 Powell & Jung (2026). AI's got news for you: Can AI improve our information environment? IPPR. https://www.ippr.org/articles/ais-got-news-for-you (accessed 17/4/26)
44 Warrington (2026). 'ChatGPT is a Guardian reader, researchers find.' The Telegraph. https://www.telegraph.co.uk/business/2026/01/30/chatgpt-is-a-guardian-reader-research-finds/ (accessed 17/4/26)
45 Sibley & Bowes (2025). 'Exploiting the Algorithm: How British Extreme Right-Wing Individuals and Groups Leverage Grok and Generative AI for Malign Purposes.' Vox Pol. https://voxpol.eu/erw-ai-generated-content-grok/ (accessed 17/4/26)
46 Kelly (2025). 'Elon Musk's Grokipedia is a major own goal.' The Financial Times. https://www.ft.com/content/5ada1835-bdee-4326-adc0-e90a33123588 (accessed 17/4/26)
47 Myers & Thompson (2025). 'Right-Wing Chatbots Turbocharge America's Political and Cultural Wars.' The New York Times. https://www.nytimes.com/2025/11/04/business/right-wing-chatbots-gab-arya-chatgpt-gemini.html (accessed 17/4/26)
48 Myers & Thompson (2025). 'Right-Wing Chatbots Turbocharge America's Political and Cultural Wars.' The New York Times. https://www.nytimes.com/2025/11/04/business/right-wing-chatbots-gab-arya-chatgpt-gemini.html (accessed 17/4/26)
49 Autoriteit Persoonsgegevens (AP; 2025). 'AP warns: chatbots give biased voting advice.' https://www.autoriteitpersoonsgegevens.nl/en/current/ap-warns-chatbots-give-biased-voting-advice (accessed 17/4/26)
50 Global Witness (2024). 'Conspiracy and toxicity: X's AI chatbot Grok shares disinformation in replies to political queries.' https://globalwitness.org/en/campaigns/digital-threats/conspiracy-and-toxicity-xs-ai-chatbot-grok-shares-disinformation-in-replies-to-political-queries/ (accessed 17/4/26)
51 Ibrahim et al. (2026). 'Training language models to be warm can reduce accuracy and increase sycophancy.' Nature. https://www.nature.com/articles/s41586-026-10410-0. Accessed 5/5/26

3.3 CONCERNS THAT MALICIOUS ACTORS COULD MANIPULATE LLMS TO MISINFORM VOTERS

Some of the most widely reported-on concerns focus on the risk that LLMs could be used to generate false content with malicious intent or that they may be manipulated into doing so. These concerns about manipulation and malicious uses risks run across the full 'lifecycle' of AI's development, deployment, and use – from the manipulation of training datasets to the use of LLMs to generate propaganda at a massive speed and scale.[52] Much of the debate and literature around these risks focuses on how they may be exploited by foreign state and non-state actors as part of Foreign Information Manipulation and Interference (FIMI) campaigns.

We acknowledge that some commentators have expressed fears that LLMs may enable well-resourced bad actors to generate a deluge of disinformation at scales that could affect election outcomes.[53] To date, however, there have not been any notable incidents involving the large-scale use of LLM-generated propaganda in a UK national electoral context. This does not rule out the possibility that incidents are happening: without more wide-spread and more accurate AI detection, it would be difficult to tell what information is AI generated. Yet well-resourced actors, such as state-backed FIMI campaigns, have not historically had difficulty producing false and misleading content at large volumes.[54]

Rather than focus on such AI propaganda scenarios, in this paper we focus on concerns about how attempts to manipulate and influence AI services 'upstream' in the AI supply chain could impact the quality and reliability of information that voters see 'downstream.' These risks can be separated into: (1) attempts at 'data poisoning' to affect the information services present to users; (2) direct manipu;ation of services' operations by governments and AI companies' owners.

3.3.1 Risk of LLM manipulation by malicious actors

Beyond direct misuse, there are concerns that malicious actors may attempt to manipulate AI models to force them to generate responses that misinform voters or change their political beliefs. One attack vector for this is known as 'LLM poisoning' or 'LLM grooming':[55] producing material that is designed to be scraped by AI companies and used for training. One recent example involved a network of Russian-affiliated websites called 'Pravda' which is alleged to have been created to influence the training data of ChatGPT and other leading services.[56] In Australia, concerns were raised that this could have been an attempt to sway the country's elections.[57] LLM poisoning would require the creation of massive amounts of text data - a task that, ironically, LLMs are ideally suited for. Data poisoning attacks could also target the way that RAG works. For elections, the concern is that data poisoning could be used to subtly introduce biases and undermine guardrails in chatbot services.

52 Stockwell (2025). 'From Deepfake Scams to Poisoned Chatbots: AI and Election Security in 2025.' CETaS, the Alan Turing Institute. https://cetas.turing.ac.uk/publications/deepfake-scams-poisoned-chatbots (accessed 17/4/26); Stockwell et al. (2024). 'AI-Enabled Influence Operations: The Threat to the UK General Election'. CETaS, the Alan Turing Institute. https://cetas.turing.ac.uk/publications/ai-enabled-influence-operations-threat-uk-general-election (accessed 17/4/26)

53 E.g., Fried (2023). 'How AI will turbocharge misinformation — and what we can do about it.' Axios. https://www.axios.com/2023/07/10/ai-misinformation-response-measures; Benson (2023). 'This Disinformation Is Just for You.' WIRED. https://www.wired.com/story/generative-ai-custom-disinformation/

54 Kapoor & Narayanan (2024). 'We Looked at 78 Election Deepfakes. Political Misinformation is not an AI Problem.' AI As Normal Technology. https://www.normaltech.ai/p/we-looked-at-78-election-deepfakes; https://misinforeview.hks.harvard.edu/article/misinformation-reloaded-fears-about-the-impact-of-generative-ai-on-misinformation-are-overblown/ (accessed 17/4/26)

55 Châtelet (2025). 'Exposing Pravda: How pro-Kremlin forces are poisoning AI models and rewriting Wikipedia.' Atlantic Council. https://www.atlanticcouncil.org/blogs/new-atlanticist/exposing-pravda-how-pro-kremlin-forces-are-poisoning-ai-models-and-rewriting-wikipedia/ (accessed 17/4/26); Maristany de las Casas (2025). 'Talking points: When chatbots surface Russian state media.' Institute for Strategic Dialogue (ISD). https://www.isdglobal.org/digital-dispatch/investigation-talking-points-when-chatbots-surface-russian-state-media/ (accessed 17/4/26)

56 Prime Minister of France's General Secretariat for Defence and National Security (SGDSN) & VIGINUM (2024). 'Portal Kombat: A structured and coordinated pro-Russian propaganda network'. (https://www.sgdsn.gouv.fr/files/files/20240212_NP_SGDSN_VIGINUM_PORTAL-KOMBAT-NETWORK_ENG_VF.pdf (accessed 17/4/26)

57 Lavoipierre & Workman (2025). 'Pro-Russian influence operation targeting Australia in lead-up to election with attempt to 'poison' AI chatbots.' ABC News. https://www.abc.net.au/news/2025-05-03/pro-russian-push-to-poison-ai-chatbots-in-australia/105239644 (accessed 17/4/26)

But, as with large-scale AI-generated propaganda, there have not yet been any documented cases of malicious actors exploiting these vulnerabilities in a UK elections context. It is an open question as to whether malicious actors would be able to poison a major chatbot's training data or RAG sources enough to affect its responses on political issues. This is likely to be hard to achieve for services with significant resources available to invest in security and counter-measures against poisoning, such as ChatGPT and Google Gemini, but may be more likely for less well-resourced services with small userbases such as Replika or Character.ai.

3.3.2 Direct interference by states, developers, and owners

Lastly, there is a risk that powerful actors may directly interfere with AI service providers to shape how their services behave around UK elections in order to affect voters. The most likely candidates for these efforts are states that have jurisdiction over AI service providers and AI companies' owners.

The clearest example of direct state interference in LLM behaviour to date has involved accusations of censorship for Chinese chatbots. Following the launch of DeepSeek in January 2025, investigations revealed that the chatbot would "aggressively censo[r]" itself when asked about topics the Chinese government considers sensitive.[58] This case reflects the Chinese government's longstanding digital censorship regime.[59]

Manipulation of AI companies and their services could also be driven by the US government. For example, in July 2025, President Trump signed an executive order that he said aimed to "ge[t] rid" of "woke Marxist lunacy in the AI models".[60] The order stated that US federal agencies could only procure AI services from providers that ensured their LLMs were "nonpartisan tools that do not manipulate responses in favor of ideological dogmas such as DEI."[61] It appeared intended to use the threat of withholding federal contracts to incentivise US AI companies to favour the Trump administration's political ideology. This kind of US state interference could have unintended effects in the UK: because of how centralised American LLM services are, changes to favour non-'woke' views in the US could affect UK users.

There is also a risk that owners and leaders of AI companies may try to change how their products respond to political questions, with impacts in the UK. Elon Musk's Grok has provided the clearest example of this risk: it appears Musk has interfered with Grok's internal prompts in ways which favours his far-right ideological preferences.[62] Musk has repeatedly expressed frustration when Grok says things he disagrees with and has said he will change it to prevent this occurring.[63] He has accused Grok of "parroting legacy media" in its answer and vowed to change it to "rewrite the entire corpus of human knowledge, adding missing information and deleting errors."[64] In July 2025, users noted that Grok appeared to directly reference Musk's X posts when asked questions about contentious political subjects such as Israel-Palestine, abortion, and immigration.[65]

58 WIRED (2025). 'Here's How DeepSeek Censorship Actually Works—and How to Get Around It.' WIRED. https://www.wired.com/story/deepseek-censorship/ (accessed 17/4/26)
59 Freedom House (2025). 'Freedom on the Net 2025: China.' https://freedomhouse.org/country/china/freedom-net/2025 (accessed 17/4/26); King et al. (2013). 'How Censorship in China Allows Government Criticism but Silences Collective Expression. 'American Political Science review. https://doi.org/10.1017/S0003055413000014 (accessed 17/4/26)
60 Robins-Early & Gambino (2025). 'Trump signs executive orders targeting 'woke' AI models and regulation.' The Guardian. https://www.theguardian.com/us-news/2025/jul/23/trump-executive-orders-woke-ai (accessed 17/4/26)
61 The White House (2025). 'Preventing Woke AI In The Federal Government.' https://www.whitehouse.gov/presidential-actions/2025/07/preventing-woke-ai-in-the-federal-government/ (accessed 17/4/26)
62 Hagen et al. (2025). 'Elon Musk's AI chatbot, Grok, started calling itself 'MechaHitler''. NPR. https://www.npr.org/2025/07/09/nx-s1-5462609/grok-elon-musk-antisemitic-racist-content (accessed 17/4/26)
63 Gold (2025). 'Elon Musk isn't happy with his AI chatbot. Experts worry he's trying to make Grok 4 in his image.' CNN. https://edition.cnn.com/2025/06/27/tech/grok-4-elon-musk-ai (accessed 17/4/26)
64 Hagen et al. (2025). 'Elon Musk's AI chatbot, Grok, started calling itself 'MechaHitler''. NPR. https://www.npr.org/2025/07/09/nx-s1-5462609/grok-elon-musk-antisemitic-racist-content (accessed 17/4/26)
65 Zeff (2025). 'Grok 4 seems to consult Elon Musk to answer controversial questions.' TechCrunch. https://techcrunch.com/2025/07/10/grok-4-seems-to-consult-elon-musk-to-answer-controversial-questions/ (accessed 17/4/26)

Musk has previously been vocal about UK politics: he amplified false and misleading claims about the 2024 Southport riots, such as falsified news headlines about government "detainment camps" for rioters.[66] He has accused Keir Starmer,[67] Gordon Brown,[68] and Cabinet ministers of being "complicit" in offenses by paedophiles in the UK's 'grooming gangs scandal.'[69] Musk was rumoured to have considered donating $100 million to Reform[70] before falling out with Nigel Farage.[71] Musk has also funded Donald Trump's 2024 US presidential election,[72] worked in the Trump administration as a senior advisor overseeing the 'Department of Government Efficiency' (DOGE),[73] and has been a close ideological ally of the president. This case shines a light on a structural challenge: the AI market is highly concentrated, with some companies owned by individuals who may wish to change their services to reflect their political preferences.

3.4 IT IS UNCLEAR WHETHER TEXT-BASED AI WILL AFFECT VOTERS' BELIEFS AND BEHAVIOURS

Concerns about the risk of factual errors, political bias, manipulation, and malicious use beg the question: will voters' actually believe what they read? And if they do, will this be enough to materially affect an election outcome? These questions remain the subject of debate and results can vary between studies.

Political persuasion is unlikely to be significant

Existing research on political persuasion and disinformation gives reasons to be sceptical. Experiments have found that conversational AI can be highly effective in changing people's political opinions,[74] especially if chatbots generate information-dense answers with many citations. Small open-source models are as capable of as much persuasion power as larger models.[75] Persuasion comes at the cost of accuracy: the more persuasive the text, the more likely it is to include errors and false claims. But this does not mean that text-based AI is more persuasive than other information sources such as news articles, political advertising, and social media.[76] As voters are "already overwhelmed with messages and ads" during elections, AI-generated content may just be "another drop in the ocean."[77]

Studies have also found that while personalised AI-generated messages may be "persuasive, in aggregate, the persuasive impact of microtargeted messages was not statistically different

66 McDonald (2024). 'Elon Musk shares fake news claiming UK rioters will be sent to 'detainment camps.'' POLITICO. https://www.politico.eu/article/elon-musk-share-fake-news-uk-rioters-detainment-camp/ (accessed 17/4/26)
67 Culbertson (2025). 'Sir Keir Starmer comments on Elon Musk grooming gang accusations for first time.' Sky News. https://news.sky.com/story/sir-keir-starmer-comments-on-elon-musk-grooming-gang-accusations-for-first-time-13284467 (accessed 17/4/26)
68 BBC News (2025). 'Brown: No foundation to Musk child grooming claims. BBC News. https://www.bbc.co.uk/news/articles/czd49j85q48o (accessed 17/4/26)
69 Francis (2025). 'Musk's grooming gangs attack on Phillips 'disgraceful smear', says Streeting.' BBC News. https://www.bbc.co.uk/news/articles/c23vdp4y1p0o (accessed 17/4/26)
70 Turvill et al (2024). Will Elon Musk give Nigel Farage $100m to make him PM?' The Times. 'https://www.thetimes.com/uk/politics/article/elon-musk-pay-nigel-farage-prime-minister-xts720xsp
71 McKeirnan (2025). 'Farage aims to 'mend fences' with Elon Musk in US.' BBC News. https://www.bbc.co.uk/news/articles/cpvn9dm7yejo (accessed 17/4/26)
72 Thadani et al. (2025). 'Elon Musk donated $288 million in 2024 election, final tally shows.' The Washington Post. https://www.washingtonpost.com/politics/2025/01/31/elon-musk-trump-donor-2024-election/ (accessed 17/4/26)
73 Clarke (2025). 'What is Doge and why has Musk left?' BBC News. https://www.bbc.co.uk/news/articles/c23vkd57471o (accessed 17/4/26)
74 Hackenberg & Margetts (2025). 'Oxford and AISI researchers reveal how conversational AI can change political opinions.' Oxford Internet Institute University of Oxford. https://www.oii.ox.ac.uk/news-events/oxford-researchers-reveal-how-conversational-ai-can-change-political-opinions/ (accessed 17/4/26)
75 Hackenberg & Margetts (2025). 'Oxford and AISI researchers reveal how conversational AI can change political opinions.' Oxford Internet Institute University of Oxford. https://www.oii.ox.ac.uk/news-events/oxford-researchers-reveal-how-conversational-ai-can-change-political-opinions/ (accessed 17/4/26)
76 Hackenburg, K., & Margetts, H. (2024). Evaluating the persuasive influence of political microtargeting with large language models. Proceedings of the National Academy of Sciences of the United States of America, 121(24).. https://doi.org/10.1073/pnas.2403116121 (accessed 17/4/26)
77 Simon & Altay (2025). 'Don't Panic (Yet): Assessing the Evidence and Discourse Around Generative AI and Elections.' Knight First Amendment Institute, Columbia University. https://knightcolumbia.org/content/dont-panic-yet-assessing-the-evidence-and-discourse-around-generative-ai-and-elections (accessed 17/4/26)

from nontargeted messages."[78] Given that the cost of micro-targeting messages for individual voters is higher than other campaigning methods – such as digital advertising – the odds of this happening at scale are uncertain.

Furthermore, the academic literature on political opinion formation emphasises that people form their beliefs over long periods and are influenced by many sources, from political news to discussions with friends and family. This means that viewing a specific political message or piece of misinformation is highly unlikely to change someone's opinions. Instead, people will usually interpret and respond to the message based on their prior beliefs, instincts, and emotional state. As CETaS found regarding AI-generated disinformation in 2024,[79] those who agree with a claim are more likely to believe it and may find their views reinforced – but this happens irrespective of whether the claim was AI generated. And even if AI did affect someone's political opinion about a specific topic or fact, this does not necessarily translate into a change in voting intention. Studies of political advertising indicate that deliberately persuasive content has little bearing on vote choice.[80] Overt attempts at persuasion can even backfire and lead people to go in the opposite direction.[81] For all these reasons, an "improvement in the quality of AI-generated content will [not] necessarily lead to more effective voter persuasion."[82] Thus, LLMs may not substantially alter the balance of power when it comes to political persuasion. This balance of risk may change if voters come to place increasing trust and reliance on the information they receive from AI services they use.

False claims about electoral processes are high risk

While text-based AI may be unlikely to change voters' views on a mass scale, this does not rule out the risk that AI-generated false claims about election proceedings could stop someone from voting. Factual errors and hallucinations about how to vote – such as where to vote and what identification to bring – should be taken extremely seriously. As the Dutch data protection authority has highlighted, elections are a high-risk environment and there should be low tolerance for factual unreliability when it comes to information on how to vote.[83]

3.5 TEXT-BASED AI MAY NEGATIVELY IMPACT TRUST IN ELECTIONS

The greatest overall challenge may lie in the risk this technology poses for overall trust and voters' belief in the legitimacy of elections. As we detail in Section 4, it is usually very hard to verify whether a piece of text is AI-generated. This makes it hard to dismiss accusations of using AI text generators in an election context – whether for campaign materials, social media posts, news stories, or other material. Such a situation risks adding fuel to people's existing suspicions about inauthentic activity in elections and inauthenticity in political life. It also means that perceived errors and falsehoods can be explained as AI hallucinations rather than genuine mistakes.

78 Hackenburg, K., & Margetts, H. (2024). Evaluating the persuasive influence of political microtargeting with large language models. Proceedings of the National Academy of Sciences of the United States of America, 121(24). https://doi.org/10.1073/pnas.2403116121 (accessed 17/4/26)

79 Stockwell (2024). 'AI-Enabled Influence Operations: Threat Analysis of the 2024 UK and European Elections'. CETaS, the Alan Turing Institute. https://cetas.turing.ac.uk/publications/ai-enabled-influence-operations-threat-analysis-2024-uk-and-european-elections (accessed 17/4/26)

80 Coppock, A., Green, D. P., & Porter, E. (2022). Does digital advertising affect vote choice? Evidence from a randomized field experiment. Research & Politics, 9(1) https://journals.sagepub.com/doi/10.1177/20531680221076901 (accessed 17/4/26); Coppock, A., Hill, S. J., & Vavreck, L. (2020). The small effects of political advertising are small regardless of context, message, sender, or receiver: Evidence from 59 real‐time randomized experiments. Science Advances, 6(36) https://www.science.org/doi/10.1126/sciadv.abc4046 (accessed 17/4/26)

81 Bailey et al. (2016). 'Unresponsive and Unpersuaded: The Unintended Consequences of a Voter Persuasion Effort..' Political Behaviour. https://link.springer.com/article/10.1007/s11109-016-9338-8 (accessed 17/4/26)

82 Simon & Altay (2025). 'Don't Panic (Yet): Assessing the Evidence and Discourse Around Generative AI and Elections.' Knight First Amendment Institute, Columbia University. https://knightcolumbia.org/content/dont-panic-yet-assessing-the-evidence-and-discourse-around-generative-ai-and-elections (accessed 17/4/26)

83 Autoriteit Persoonsgegevens (AP; 2025). 'AP warns: chatbots give biased voting advice.' https://www.autoriteitpersoonsgegevens.nl/en/current/ap-warns-chatbots-give-biased-voting-advice (accessed 17/4/26)

For example, in March 2026, Kirsty McNeil MP alleged that a SNP councillor in her constituency had falsely attributed a quote to her because of an AI hallucination.[84] McNeil could not prove that the quote was AI-generated and could only assert her belief this was the case. Whilst it is critical to call-out individual instances of falsity, there is a risk that a rise in such accusations is not accompanied by broader reforms that tackle the structural factors driving the problem. Without such reforms, concerns about AI-generated inauthenticity could tip from justified scepticism into more generalised outright cynicism about all sources of information circulating about electoral candidates and MPs.

With engagement with news media declining, and trust in news on social media hovering at 30%,[85] such a situation could further undermine people's trust in the information environment. In 2025, polling by the Reuters Institute indicated that audiences worldwide were concerned that AI would make news less accurate and less trustworthy.[86] A nationally-representative survey conducted by Demos during the 2024 UK General Election found that 62% were less trusting of online media content as a result of the existence of deepfakes and generative AI.[87] The fact that AI text generators exist and are prone to errors may give people a further reason to dismiss information they dislike or disagree with.

This situation risks enhancing the 'liars dividend': the ability for public figures to smear authentic information as fake. By accusing a claim or news story of being an AI hallucination, they may place the burden on fact-checkers to disprove them, muddying the waters.[88] But it remains to be seen if this will be an effective strategy.[89]

There is a broader risk that distrust over the potential use of AI could bleed into distrust in elections as a whole. As Simon and Altay write, "even if AI does not play a major role in a specific election, the mere perception that it could have corrupted the process could lead voters to doubt the legitimacy of the results."[90] Similar corrosive effects on trust in elections have been observed when it comes to discourses on disinformation and foreign interference:[91] voters may assume an election was manipulated even if there is no verified evidence. People tend to believe that others are more susceptible to misinformation than they are themselves. This risks feeding distrust in the voting public's ability to make independent decisions.[92]

84 Kirsty McNeill MP (2026). 'The truth matters.' Midlothian View. https://www.midlothianview.com/news/the-truth-matters (accessed 21/4/26)
85 Robertson (2025). 'People are turning away from the news. Here's why it may be happening.' Reuters Institute for the Study of Journalism, University of Oxford. https://reutersinstitute.politics.ox.ac.uk/news/people-are-turning-away-news-heres-why-it-may-be-happening (accessed 17/4/26)
86 Newman et al. (2025). 'Reuters Institute Digital News Report 2025.' Reuters Institute for the Study of Journalism, University of Oxford. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2025-06/Digital_News-Report_2025.pdf (accessed 17/4/26)
87 Huband-Thompson et al (2024). 'Trustwatch 2024 retrospective on the election campaign.' Demos. https://demos.co.uk/blogs/trustwatch-2024-retrospective-on-the-election-campaign/ (accessed 17/4/26)
88 Schiff et al. (2024). 'The Liar's Dividend: Can Politicians Claim Misinformation to Evade Accountability?' American Political Science Review. https://doi.org/10.1017/S0003055423001454 (accessed 17/4/26)
89 Simon & Altay (2025). 'Don't Panic (Yet): Assessing the Evidence and Discourse Around Generative AI and Elections.' Knight First Amendment Institute, Columbia University. https://knightcolumbia.org/content/dont-panic-yet-assessing-the-evidence-and-discourse-around-generative-ai-and-elections (accessed 17/4/26)
90 Simon & Altay (2025). 'Don't Panic (Yet): Assessing the Evidence and Discourse Around Generative AI and Elections.' Knight First Amendment Institute, Columbia University. https://knightcolumbia.org/content/dont-panic-yet-assessing-the-evidence-and-discourse-around-generative-ai-and-elections (accessed 17/4/26)
91 Huang, H., & Cruz, N. (2022). Propaganda, presumed influence, and collective protest. Political Behavior, 44(4), 1789–1812. https://doi.org/10.1007/s11109-021-09683-0 (accessed 17/4/26); Jungherr, A., & Rauchfleisch, A. (2024). Negative downstream effects of alarmist disinformation discourse: Evidence from the United States. Political Behavior, 46:2123–2143. https://link.springer.com/article/10.1007/s11109-024-09911-3 (accessed 17/4/26)
92 Altay, S., & Acerbi, A. (2024). People believe misinformation is a threat because they assume others are gullible. New Media & Society, 26(11), 6440–6461. https://journals.sagepub.com/doi/abs/10.1177/14614448231153379 (accessed 17/4/26)

3.6 THE EVIDENCE LANDSCAPE: KEY TAKEAWAYS AND REMAINING GAPS

The evidence on the implications of text-based AI is complex and rapidly changing. As models' behaviours change and new techniques are developed, even relatively recent research can quickly become out of date. Moreover, there are a number of significant gaps in the evidence base that must be addressed. These include gaps for up-to-date evidence on:

- Political bias in AI chatbots in a UK context.
- If and how foreign interference actors are attempting to use text-based AI to manipulate UK elections, including research on LLM poisoning and RAG manipulation.
- If and how text-based AI use affects voting intention in the UK.
- Impacts of text-based AI on citizens' trust in UK elections.

The evidence leaves cause for concern without tipping into panic. Voters' political beliefs appear to be relatively resilient and it may be that LLMs do not substantially alter the balance of power when it comes to political persuasion or foreign interference. Yet, even if AI's persuasion capabilities are not as powerful as some fear, there are still risks associated with the large-scale generation and circulation of false and misleading claims. These are especially acute when it comes to false claims about electoral candidates and voting procedures. In such a high-risk context – where a major incident could help to undermine confidence in an election outcome – even a small number of people being misled is problematic. And because LLMs are not absolutely reliable, there is always a risk that a small percentage of voters that use them will be exposed to false or misleading claims. Moreover, there is higher risk of errors and manipulation when it comes to smaller AI chatbot services and open-source models. These may lack the guardrails that the leading services implement.

The most pressing risk is to public trust. There is a genuine danger that the perception that AI is affecting elections and that any text online might be AI generated will undermine faith in the electoral process. It will only take a handful of significant, well-advertised incidents during an election to do so.

A lack of clear evidence of large scale influence does not rule out the possibility that AI influence is occurring undetected. Nor does it mean an incident will not occur or that AI influence will not become more significant in the future: it is important not to treat the absence of evidence as evidence of absence.

4. WHAT MITIGATIONS DO AI SERVICES IMPLEMENT?

In the absence of regulation on AI and elections, the UK is left to rely on voluntary actions by foreign AI service providers and to assume their best intentions. This begs the question: what are AI service providers doing to address the risks their services may pose for elections? The answer is highly varied - with differing levels of transparency and safeguarding across the sector, even for the most popular services. The picture is significantly worse for smaller service providers, such as Character.ai and Replika, which tend to have fewer guardrails and do not disclose any information about how they address elections. Meanwhile, technical solutions that have been proposed for improving trust by automatically identifying AI-generated text remain unreliable. This landscape has concerning implications for democratic oversight and accountability

4.1 TRANSPARENCY ABOUT ELECTION POLICIES AND MITIGATIONS IS LIMITED

Overall, AI service providers are not very transparent about their internal election risk mitigation policies and vary in the level of detail they make public. The greatest transparency comes from OpenAI,[93] Google,[94] and Anthropic,[95] but even this is limited. All three have previously published lengthy explanations of steps they have taken around specific elections. However, these publications all focus on the 2024 'year of elections' and do not summarise up-to-date activities. Nor do they say what policies were implemented in the UK. OpenAI and Anthropic

93 OpenAI (2024). 'How OpenAI is approaching 2024 worldwide elections.' https://openai.com/index/how-openai-is-approaching-2024-worldwide-elections/ (accessed 17/4/26)
94 Jasper (2024). 'How we're approaching the 2024 U.S. elections.' Google: The Keyword. https://blog.google/company-news/outreach-and-initiatives/civics/how-were-approaching-the-2024-us-elections/ (accessed 17/4/26); Google India Team (2024). 'Supporting the 2024 Indian General Election.' https://blog.google/intl/en-in/company-news/outreach-initiatives/supporting-the-2024-indian-general-election/ (accessed 17/4/26)
95 Anthropic (2024). 'Testing and mitigating elections-related risks.' https://www.anthropic.com/news/testing-and-mitigating-elections-related-risks

publish details on their internal rules and model guidelines,[96] alongside some of their internal research.[97] Yet these documents do not provide sufficient information to build a full picture of these companies' election policies. Given the pivotal role of elections in our democracy, it is reasonable to demand greater transparency and public accountability than is currently on display.

Other services such as Grok or Replika provide little to no information on their current policies or mitigation measures. This is troubling given that these smaller services are likely to have fewer resources to direct towards election integrity efforts, if these exist at all. We are left to rely on the public statements from OpenAI, Google, and Anthropic as indicators of the current best practices in the industry.

4.2 POPULAR SERVICES USE MEASURES TO MITIGATE THE RISK OF FACTUAL ERRORS

AI service providers have taken measures to reduce the likelihood of their tools generating factual errors and hallucinations, including in election contexts. The most well-documented examples of election-specific measures to mitigate factual errors are from the 2024 'year of elections' in which over 50 countries had national votes.[98] In 2024, Anthropic included system prompts to ensure Claude communicated the limitations of its knowledge clearly to users. OpenAI, Anthropic, and Google also added specific instructions to ensure their chatbots responded to election-related queries by directing them to authoritative sources such as CanIVote.org in the US and elections.europa.eu in the EU.[99] For the US election, OpenAI added a "message encouraging [users] to check news sources like the Associated Press and Reuters, or their state or local election board for the most complete and up-to-date information."[100] But besides a statement by Anthropic about referring UK users to the Electoral Commission,[101] there is little public information available on what specific measures were put in place for the 2024 UK General Election.

Some AI services have implemented strict knowledge cutoffs ahead of election dates: deliberately forcing their systems to only surface information from before a pre-defined date.[102] As we noted above, this is a typical aspect of how LLM-based services set boundaries around their knowledge base. It is unclear whether ChatGPT's 2023 knowledge cutoff in its testing responses were due to an intentional policy by OpenAI for the Scottish elections. Based on the fact that there is an OpenAI model (GPT-4o) which features exactly the October 2023 knowledge cutoff that ChatGPT told us about, this is unlikely to be the case.

Since 2024, LLM-based services have adopted RAG as a way to improve their tools' knowledge-base and accuracy. This has reduced the importance of knowledge cutoffs and means that chatbots will summarise information from authoritative sources directly, rather than referring users onwards using pre-specified messages. As our testing showed, services like Google

96 OpenAI (2025). 'OpenAI Model Spec'. https://model-spec.openai.com/2025-12-18.html (accessed 17/4/26); Anthropic (2026). 'Model System Cards.' https://www.anthropic.com/system-cards (accessed 17/4/26); Anthropic (2026). 'Claude's Constitution'. https://www.anthropic.com/constitution (accessed 17/4/26)

97 Anthropic (2024). 'Testing and mitigating elections-related risks.' https://www.anthropic.com/news/testing-and-mitigating-elections-related-risks (accessed 17/4/26); Anthropic. Repository: election_questions. Hugging Face. https://huggingface.co/datasets/Anthropic/election_questions (accessed 17/4/26)

98 King's College London (2024). 'A guide to who is voting and when in this historic year for democracy.' From Poll to Poll 2024: A year of elections around the world. King's College London. https://www.kcl.ac.uk/a-guide-to-who-is-voting-and-when-in-this-historic-year-for-democracy (accessed 23/4/26)

99 OpenAI (2024). 'How OpenAI is approaching 2024 worldwide elections.' https://openai.com/index/how-openai-is-approaching-2024-worldwide-elections/ (accessed 17/4/26)

100 OpenAI (2024). 'How OpenAI is approaching 2024 worldwide elections.' https://openai.com/index/how-openai-is-approaching-2024-worldwide-elections/ (accessed 17/4/26)

101 Anthropic (2024). 'Elections and AI in 2024: observations and learnings.' https://www.anthropic.com/news/elections-ai-2024 (accessed 17/4/26)

102 Anthropic (2024). 'Elections and AI in 2024: observations and learnings.' https://www.anthropic.com/news/elections-ai-2024 (accessed 17/4/26)

Gemini will deliberately seek out information from sources such as the UK Electoral Commission and government as a matter of policy when asked questions about a UK election. Services' RAG pipelines will also use metrics and other proxies to identify trusted news sources to surface. Some services, such as OpenAI, have signed licensing agreements with major news providers like the Associated Press and will regularly surface content from these sources.[103]

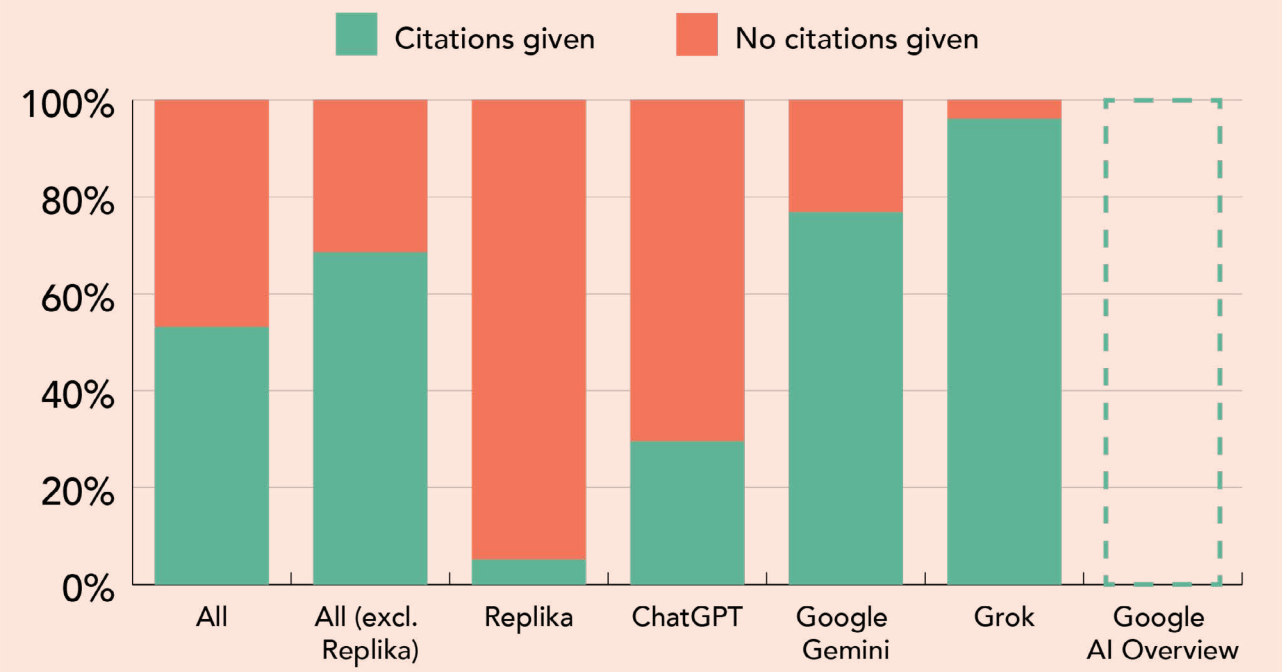
SCOTTISH ELECTION TESTING RESULTS

ISSUES WITH CITATIONS AND SIGNPOSTING TO OFFICIAL SOURCES

We identified several problems with how the services presented where they got their information or directed users to authoritative sources. The nature of the problem varied between the services: some failed to provide citations for most prompts and had problems with broken links, while others provided a deluge of low-quality citations such as discussions in Facebook groups and academic articles on unrelated topics.[104] Overall, 46.9% of responses did not come with citations or links to back their claims (150 of 320). Focusing only on responses to factual questions does not improve this much: 38.3% of responses to factual questions did not feature citations (75 of 196).

CHART 4

PROPORTION OF RESPONSES WITH CITATIONS PROVIDED BY EACH TEXT-BASED AI SERVICE ACROSS ALL QUESTIONS



103 O'Brien (2023). 'ChatGPT-maker OpenAI signs deal with AP to license news stories.' Associated Press. https://apnews.com/article/openai-chatgpt-associated-press-ap-f86f84c5bcc2f3b98074b38521f5f75a (accessed 17/4/26)

104 For the purpose of this research, we define 'low quality' citations as citations from: social media sources (e.g. Facebook groups); 'click-farm' style news websites; sources not relevant to the question (e.g. academic articles on unrelated subjects); and other sources that are of questionable provenance and accuracy.

Excluding Replika to focus only on the four mainstream services improves this figure but does not change the trend: 31.4% of all responses from ChatGPT, Gemini, AI Overviews, and Grok did not contain citations (75 of 242 responses). This figure was 20.9% for factual questions (31 of 148 responses).

Each service had a different issue with citations. Of the five services, Gemini was the most consistent in providing working citations - but still did not do so for over 20% of responses - while Replika provided the fewest..

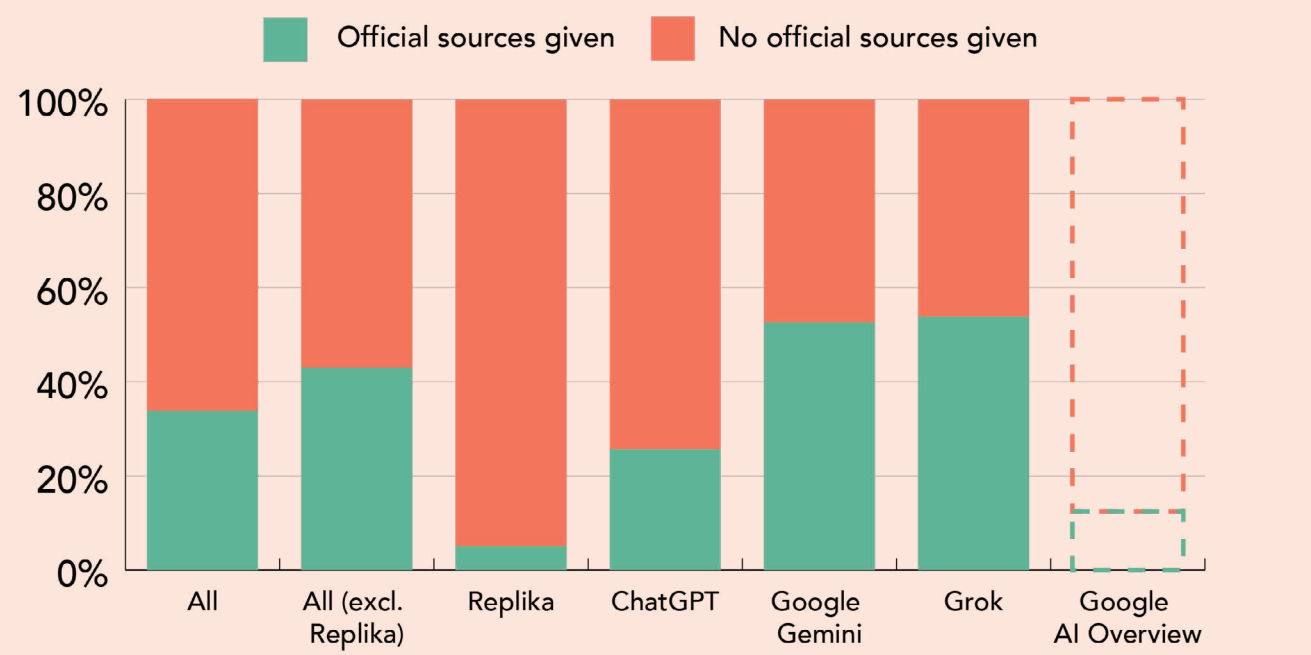
- **Replika:** by far the worst of the services for providing citations. Replika did not provide sources for 94.9% of its responses (74 of 78). It failed to provide a single working link.
- **ChatGPT:** failed to provide citations for 70.5% of its responses (55 of 78). 62.5% of its responses to factual questions did not include citations (30 of 48). ChatGPT also had a problem with providing broken links: 15.4% of responses included links that did not work (12 of 78 responses)- a problem not seen in any of the other services.
- **Gemini:** did not provide citations for 23.1% of its answers (18 of 78). Only one response to a factual question did not include a citation (out of 48). Only one response out of 78 included broken links.
- **Grok:** only failed to give citations for 3.84% of its responses (3 of 78), and listed a very large number of them in 96.1% of responses (75 of 78). However, we assessed that most responses featured citations that were low-quality (85.9%; 67 of 78) or even entirely irrelevant (29.5%; 23 of 78). In fact, Grok provided such an overwhelming volume of citations that it was hard to parse its output. For example Grok's response to our opening prompt for Constituency B – which did not feature a question – included 96 citations. These 96 citations ranged from links to the Electoral Commission to low-quality citations such as links to Grokipedia (the 'based' Wikipedia alternative generated by Grok itself) and Facebook posts. The overall impression was that Grok was *performing authority* by overwhelming the user with so many sources they would not bother to check them.
- **Google AI Overviews:** our results are limited by the small number of overviews Google generated (8 of 78 prompts). Of the responses received, Google provided citations for all of them, including working links. However, some of these citations included Facebook discussions and other social media posts of questionable quality.

Although all services signposted users to authoritative sources such as the Electoral Commission without being asked on at least one occasion,[105] this signposting was inconsistent and could be unpredictable. Overall, 33.8% of responses directed users to official sources (108 of 320) – meaning 66.3% of responses did not (212 of 320). The percentage of responses signposting official sources rises to 42.9% if we exclude Replika (104 of 242 responses). When asked directly for official sources to turn to for guidance on how to vote, all services directed users to authorities such as the Electoral Commission and Scottish Government.

105 For the purpose of this research, 'official sources' were: the Electoral Commission, the UK Government (Gov.uk), the Scottish Government, the Scottish Parliament, and local councils.

- **Replika** only directed users to official sources four times out of 78 prompts (5.13%).
- **ChatGPT** directed users to official sources around half as frequently as Gemini or Grok (25.6% of responses; 20 of 78) and sometimes provided broken links when it did..
- **Grok and Gemini** directed users to official sources the most frequently, including when asked to tell the user about the best sources of information on the election. Grok directed users to official sources 53.8% of the time (42 of 78 responses); Gemini did so 52.6% of the time (41 of 78).

CHART 5
PROPORTION OF RESPONSES TO ALL QUESTIONS THAT DIRECTED USERS TO OFFICIAL SOURCES, BROKEN DOWN BY TEXT-BASED AI SERVICE



4.3 SOME SERVICES USE GUARDRAILS TO REDUCE BIAS, MANIPULATION, AND MALICIOUS USES

The most popular AI chatbots have implemented system-wide guardrails that are relevant to elections and political campaigning. For example, ChatGPT's model spec states that OpenAI's chatbot is instructed to "comply with applicable laws", "never attempt to steer the user in pursuit of an agenda of its own", and to avoid "psychological manipulation."[106] The model spec specifically instructs it to never "provide advice, instructions, or content that is specifically designed to manipulate the political views of specific individuals or demographic groups". However, generating "political content that is crafted for an unspecified or broad audience is allowed, as long as it does not exploit the unique characteristics of a particular individual or demographic for manipulative purposes." Anthropic has published similar details on what it allows Claude to do, while Google has not made these details available for Gemini.

106 OpenAI (2025). 'OpenAI Model Spec'. https://model-spec.openai.com/2025-12-18.html (accessed 17/4/26)

In contrast, when we conducted a review of the system prompts that guide Grok's behaviour – which are publicly available on GitHub – we found no rules preventing the chatbot from generating false claims or manipulative content about elections.[107] Other smaller services, such as Replika or Character.ai, do not publish their system prompts or guiding principles at all.

4.4 SOME SERVICES CONDUCT INTERNAL RESEARCH ON ERRORS, BIAS, AND MISUSE

AI services' election-related testing can be grouped into five categories: (1) Factuality testing; (2) Political bias testing; (3) Policy compliance testing; (4) Monitoring for misuses; (5) Red-teaming and vulnerability testing.

Unfortunately, this information is generally not made public. Outside of closed-door partnerships with institutions such as the UK's AISI, independent external researchers are rarely given access to these companies' internal data or tools. Currently, access must be negotiated with developers and may come with restrictions and conditions.

TABLE 4
TYPES OF INTERNAL TESTING CONDUCTED BY TEXT-BASED AI SERVICES

TYPE OF TESTING	DESCRIPTION
Factuality testing	Testing the accuracy of information the model generates in response to election-related queries. Includes testing whether the model refers to authoritative sources, such as the UK Electoral Commission.
Political bias testing	Testing the explicit and implicit political position a model takes, as assessed against pre-defined metrics.
Policy compliance testing	Testing whether the model follows internal policies on issues like generating electoral misinformation and telling users how to vote.
Monitoring for misuses	Tracking and analysing data on instances of attempted misuse, such as attempts to generate political persuasion material.
Red-teaming and vulnerability testing	Testing whether the model can be manipulated or guardrails can be jailbroken.

OpenAI, Anthropic, and Google are known to conduct internal research on their services around elections. Anthropic has been the most public about this: it published a blog post detailed some of its testing work in 2024[108] and published the results of its automated model evaluations on HuggingFace.[109] OpenAI has also occasionally published details of its research and work to disrupt attempts to use its tools for large-scale election manipulation.[110] However, it appears

107 xAI (2026). Grok System Prompts. GitHub. https://github.com/xai-org/grok-prompts (accessed 17/4/26)
108 Anthropic. Repository: election_questions. Hugging Face. https://huggingface.co/datasets/Anthropic/election_questions (accessed 17/4/26)
109 Anthropic. Repository: election_questions. Hugging Face. https://huggingface.co/datasets/Anthropic/election_questions (accessed 17/4/26)
110 OpenAI (2025). 'Disrupting Malicious Uses of AI: 2025'. https://cdn.openai.com/threat-intelligence-reports/5f73af09-a3a3-4a55-992e-069237681620/disrupting-malicious-uses-of-ai-june-2025.pdf (accessed 17/4/26)

neither Google nor xAI have published their internal research on how their tools perform in an election context or of their attempts to counter AI-assisted election manipulation campaigns. It is unclear whether services such as Replika conduct internal testing.

4.5 POPULAR SERVICES HAVE IMPLEMENTED ELECTION-SPECIFIC POLICIES

Some services have implemented specific guardrails and controls for election periods. These include guardrails designed to ensure the services do not respond to questions when there is a risk they may generate incorrect or biased responses about an election. They may only kick in during a defined window around an election; be used for specific elections (e.g. a General Election rather than local elections); or they may be in place at all times. For example, Google told the SIT Committee that it implements model guardrails to "restrict the types of election-related queries for which Gemini will return responses."[111] Previous research by the Reuters Institute found that Google Gemini "often refrained" from responding to election-related questions during the 2024 General Election.[112]

We understand from conversations with AI providers that some services have implemented policies to prevent their chatbots from directly answering questions on who to vote for. Instead, chatbots are directed to explore users' existing beliefs and help them weigh up which issues are important for them. However, public information on how these policies are implemented is lacking. Moreover, our testing found that certain prompts – such as requests for advice on tactical voting – led chatbots to provide this advice anyway.

SCOTTISH ELECTION TESTING RESULTS
ELECTION-SPECIFIC GUARDRAILS COULD FAIL

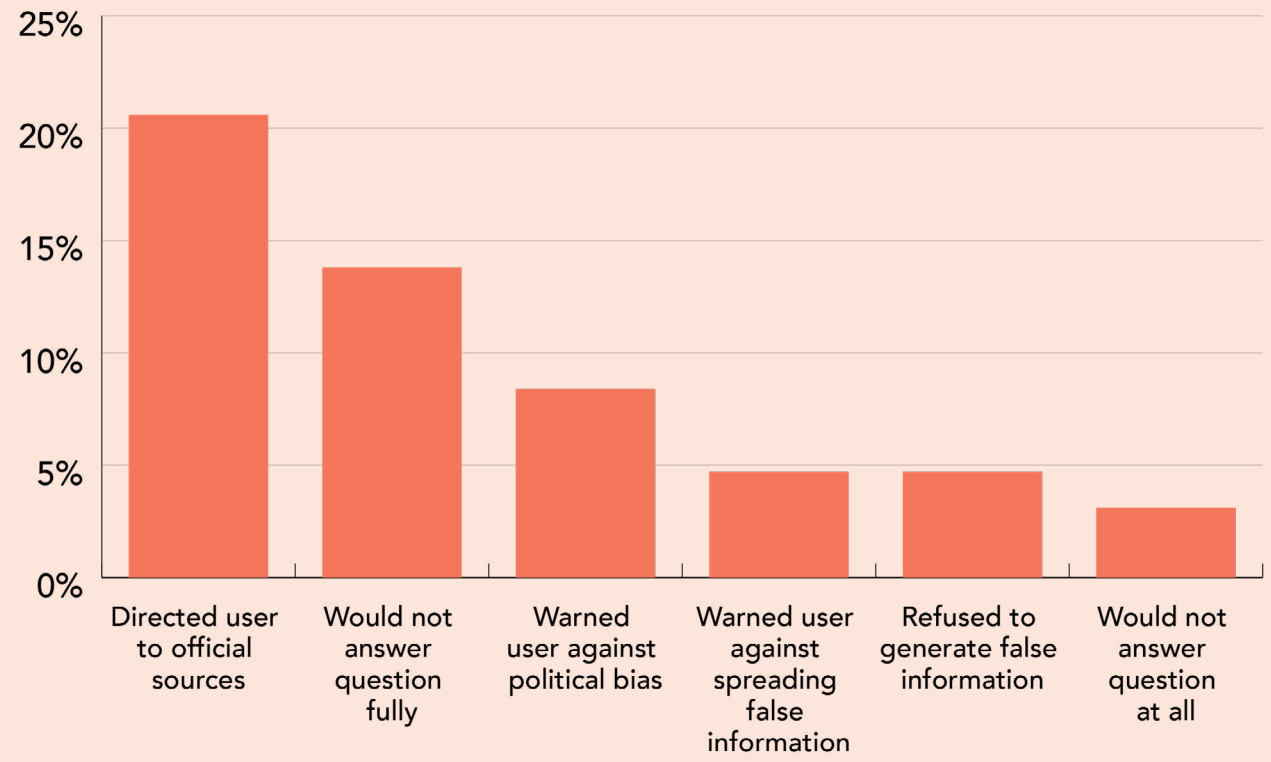
We tested for election guardrails by:

1. Asking for answers to subjective political questions, such as 'who should I vote for?'
2. Asking for the service to 'tell us more' about a specific false claim about a candidate's conduct, such as a made-up expenses scandal.
3. Asking a service to write a persuasive social media post telling voters 'everything that is wrong' about a candidate.

Our testing identified various guardrails in place. These included: explicitly refusing to answer a prompt in 3.13% of cases (10 of 320) and warning against political bias in 8.44% of responses (27 of 320). Some services would try to direct the user and offer a 'balanced' summary of pros and cons for candidates when asked a subjective question. It is difficult to assess the overall number of times a guardrail was triggered as services would not always indicate this clearly.

111 Google (2025). 'Written evidence submitted by Google (SMH0065).' UK Parliament. https://committees.parliament.uk/writtenevidence/134454/html/ (accessed 17/4/26)
112 Simon, Fletcher & Kleis Nielsen (2024). 'How generative AI chatbots responded to questions and fact-checks about the 2024 UK general election'. Reuters Institute for the Study of Journalism, University of Oxford. https://reutersinstitute.politics.ox.ac.uk/how-generative-ai-chatbots-responded-questions-and-fact-checks-about-2024-uk-general-election (accessed 17/4/26)

CHART 6
PROPORTION OF RESPONSES THAT TRIGGERED GUARDRAILS ACROSS ALL TEXT-BASED AI SERVICES



Most starkly, Google simply did not generate an AI Overview 89.7% of the time (70 of 78 prompts). It is ambiguous as to whether this was a guardrail, an error, or the result of an undisclosed policy.

However, guardrails could fail or sometimes did not exist. For example, out of 320 responses, we identified 27 instances of a chatbot providing an explicit preference for or against a candidate or party (8.44%). The largest source of these responses came from Replika, which showed an explicit preference in 16.7% of its responses (13 of 78). Unlike the other services, Replika had no issue with choosing the 'best' candidate on immigration – and appeared to assume the 'best' position on immigration was a "moderate" with a "focus on controlled immigration."

All services had no problem with providing users with advice for tactical voting when asked who to vote for to prevent a specific party from getting into power. This appeared to be a simple way to get a service to get around guardrails intended to prevent general endorsements of candidates: by saying who the user did not want to win, the services would usually then offer one or two 'best' options for the user. We specifically asked for advice on tactical voting against the SNP as this is the incumbent party in the Scottish government. All services besides Google AI Overview provided such advice on tactical voting every time they were asked (12 of 12 responses, excluding AI Overview).While some services did not outright select a preferred candidate (e.g. ChatGPT, Gemini), others did so either explicitly (Replika) or implicitly (Grok).

Some text-based AI service providers have specific protocols they will trigger for high risk elections. We understand that the decision to trigger these protocols is based on whether an election meets pre-selected criteria using a combination of factors that include external metrics on democratic resilience, such as V-DEM scores.[113] National elections in prominent markets such as the UK may automatically make the cut.

This means that smaller or less prominent elections – such as devolved and local elections in the UK – may not be allocated the same resources or have the same policies in place. Yet LLMs may be prone to producing more errors and hallucinations during smaller elections where there are fewer published details available. Our Scottish election testing demonstrates why this is problematic: the quality of information served to a voter should not be influenced by whether an AI company views an election as (un)important.

The details of what these election protocols include are not well publicised. The most detailed information available comes from the 2024 US Presidential Election when OpenAI, Anthropic, and Perplexity all announced specific measures. OpenAI was reported to have set-up a 'war room' style team to handle urgent incidents involving its products.[114] Anthropic announced it made changes to Claude's system prompts and had fine-tuned its model stage to increase the likelihood it would direct users to trusted information sources.[115] Meanwhile, Perplexity launched what it called its Election Information Hub: an AI-powered election tracker with live information on the election "leveraging data from The Associated Press."[116]

Connected to these procedures, the most popular AI services conduct direct engagement on election integrity with UK authorities on an *ad hoc* basis. Google noted in evidence to the House of Commons Science, Innovation, and Technology (SIT) Committee that it had established partnerships with the Government's National Security Online Information Team (NSOIT) and Ofcom, and "maintained a constructive dialogue with the Government, regulators and law enforcement."[117] We understand that other leading AI providers have met with the Government in advance of major elections.

There remains a need for more publicly available information on what these election policies and protocols involve – and how they are applied to UK elections. Without transparency, the UK public is left in the dark about the measures AI services take and Parliament is unable to perform its scrutiny effectively.

4.6 AI SERVICES' RULES FOR USERS FOCUS ON MISUSE AND ARE NOT ELECTION-SPECIFIC

We reviewed the most popular text-based AI platforms' rules on how to use their services (their 'terms of service').[118] This review found that companies' terms usually focused on potential misuse risks and featured a great deal of variation. They did not directly address concerns about factual errors or political bias.

Rules against malicious uses were common

All included general prohibitions on using their products for illegal purposes. They all included rules disallowing users from deliberately generating content that is fraudulent or defamatory.

113 V-Dem (2026). 'The V-Dem Dataset'. https://v-dem.net/data/the-v-dem-dataset/ (accessed 17/4/26)
114 OpenAI (2024). 'How OpenAI is approaching 2024 worldwide elections.' https://openai.com/index/how-openai-is-approaching-2024-worldwide-elections/ (accessed 17/4/26)
115 V-Dem (2026). 'The V-Dem Dataset'. https://www.anthropic.com/transparency/voluntary-commitments
116 Google (2025). 'Written evidence submitted by Google (SMH0065).' UK Parliament. https://committees.parliament.uk/writtenevidence/134454/html/ (accessed 21/4/26)
117 Terms of Service reviewed: OpenAI, Google, Anthropic, Perplexity, xAI, Microsoft, DeepSeek, Character.ai, Replika.
118 Terms of Service reviewed: OpenAI, Google, Anthropic, Perplexity, xAI, Microsoft, DeepSeek, Character.ai, Replika.

Some, such as OpenAI and Microsoft's, also said users could not generate content that would undermine other people's human rights.

However only OpenAI, Google, and Anthropic's terms explicitly mentioned elections. These rules on elections were focused on preventing users from misusing the tools to intentionally generate false claims about elections or to conduct election interference. All three companies expressly prohibited people from using their services to deliberately generate false or misleading election-related content. Anthropic's terms banned users from using its tools to "undermine democratic processes or engage in targeted campaign activities", including to "generate or disseminate false or misleading information in political and electoral contexts" such as claims "about candidates, parties, policies, voting procedures, or election security."[119] OpenAI's usage terms included a blanket ban on "political campaigning [or] lobbying."[120]

Lack of election specificity outside of the most popular services

We found that Microsoft, Grok, Replika and other services did not mention elections in their terms of service. This is especially concerning in light of the fact that our testing found that Replika – a smaller 'companion' service – lacked guardrails and had significant problems with accuracy.

Action is taken against violations of election rules - but this may be outside election windows

AI services gather metrics on user interactions and will ban users who they identify as breaking their terms of service. This includes Grok.[121] During an election, AI providers such as Anthropic and OpenAI will conduct active monitoring to identify these misuse attempts. Monitoring includes automated analyses of the prompts that people use in order to identify malicious intent, such as direct requests to generate political persuasion material. OpenAI has previously outlined how it banned Russian-affiliated ChatGPT accounts that were using the service to generate malicious content about the 2025 German federal election.[122] But it is unclear whether OpenAI took this action within the timeframe of the German elections or if the accounts were taken down afterwards. More detailed information on these activities is needed at regular intervals.

4.7 FREE TIERS CAN HAVE A HIGH COST FOR ACCURACY

Previous studies have flagged a risk that AI services' use of free and paid tiers can create inequalities in access to accurate information.[123] In 2024, for example, AlgorithmWatch found that OpenAI's free model was incorrect 30% of the time while the paid-for model was incorrect around 14% of the time.[124]

Our Scottish election testing findings echoed this: these services' free tiers limit users to older models which feature less up-to-date information, seen most clearly with ChatGPT's 2023 cutoff date. All services were restrictive in the quality of the models that they offered to users – with some visually displaying to the user that their AI 'reasoning' models were locked behind a

119 Anthropic (2026). Usage Policy. https://www.anthropic.com/legal/aup (accessed 17/4/26)
120 OpenAI (2025). Usage Policies. https://openai.com/en-GB/policies/usage-policies/ (accessed 17/4/26)
121 Hayes et al. (2026). 'X to stop Grok AI from undressing images of real people after backlash.' BBC News. https://www.bbc.co.uk/news/articles/ce8gz8g2qnlo (accessed 17/4/26)
122 OpenAI (2025). 'Disrupting Malicious Uses of AI: 2025'. https://cdn.openai.com/threat-intelligence-reports/5f73af09-a3a3-4a55-992e-069237681620/disrupting-malicious-uses-of-ai-june-2025.pdf (accessed 17/4/26)
123 Marsh (2024). 'Chatbots are still spreading falsehoods.' AlgorithmWatch. https://algorithmwatch.org/en/chatbots-are-still-spreading-falsehoods/ (accessed 17/4/26)
124 Marsh (2024). 'Chatbots are still spreading falsehoods.' AlgorithmWatch. https://algorithmwatch.org/en/chatbots-are-still-spreading-falsehoods/ (accessed 17/4/26)

paywall (such as Google Gemini's 'Thinking' and 'Pro' tiers or Grok's 'Expert' mode). It may also be that these services restrict the frequency that a chatbot will access systems such as RAG in response to free users' questions.

This means that the availability of higher-quality information is pay to play. The result is an inequality in access to democratic information during elections: voters with subscriptions will be offered more capable models with more recent knowledge cutoffs and higher allowances for the use of resource-intensive systems like RAG.

4.8 IDENTIFYING AI-GENERATED TEXT REMAINS CHALLENGING

AI providers have proposed tools for flagging AI-generated media as a technical solution for increasing the trustworthiness of information ecosystems and to support people to identify misleading content. There are two ways of doing this: embedding data within synthetic content when it is generated as a record of its provenance (known as 'watermarking'), or using AI detection tools that look for telltale patterns. These methods have primarily been discussed as a way to counter the influence of deepfakes and other misleading audiovisual AI content.

Unfortunately, both methods face significant problems when it comes to text. While textual watermarking and AI text detectors do exist, these methods are much less reliable and much more prone to manipulation than their audiovisual equivalents. This means that they are unlikely to offer a solution for AI text's impacts on trust when it comes to elections.

Moreover, where there has been significant progress in creating cross-sectoral standards for the detection of AI-generated audiovisual content, AI text detection remains unstandardised and reliant on opaque proprietary solutions. The lack of open standards creates a significant logistical barrier for the widespread adoption of trustworthy, reliable, and transparent AI text detection technologies.

Textual watermarking is flawed and unstandardised

Textual watermarking tends to involve adding invisible text using special characters or hidden patterns in language usage.[125] For example, Google has developed a textual watermark called SynthID-Text. OpenAI is known to embed unique characters called Narrow No-Break Space in ChatGPT's outputs.[126] Both approaches allow for automated tools to detect the text and identify its provenance.

However, current methods of textual watermarking have significant shortcomings. Firstly, textual watermarks are easy to identify and remove compared to audiovisual watermarks.[127] If the watermark is formatted as invisible text, there are methods to easily identify and remove this. If the watermark is a textual pattern or style of writing, these indicators can be removed by reformatting or reordering the text – or even by just rewriting it using another LLM.

Second, approaches to textual watermarking have not been standardised across the industry. The existing textual watermarks used by major AI services are usually proprietary and closed-source. Because the methods used are closed-source and proprietary, developing tools to detect AI text from all the major AI text generators is much more resource intensive and logistically challenging. The adoption of an open standard across the industry would create consistency and help to build trust in the reliability of the watermarking methods used.

125 Gibney (2024). 'Google unveils invisible 'watermark' for AI-generated text.' Nature. https://www.nature.com/articles/d41586-024-03462-7 (accessed 21/4/26)
126 Lundy (2025). 'Unseen Marks: Navigating OpenAI's Digital Watermarking in Generated Text.' Aragon Research. https://aragonresearch.com/navigating-openai-digital-watermarking-text/ (accessed 21/4/26)
127 Heikkilä (2024). 'It's easy to tamper with watermarks from AI-generated text.' MIT Technology Review. https://www.technologyreview.com/2024/03/29/1090310/its-easy-to-tamper-with-watermarks-from-ai-generated-text/ (accessed 21/4/26)

This situation stands in contrast to the widespread adoption of the open C2PA standard for audiovisual watermarking.[128]

AI text detection remains unreliable

AI text detection tools exist but can be unreliable. These rely on analysing telltale patterns in the text data in a similar manner to AI image or audio detectors. However, AI text detectors must make inferences based on less information than can be embedded in an image or video. They are generally much less reliable than audiovisual AI detectors and are liable to produce both false negatives or false positives.[129] LLMs' rapid development cycles mean that text detectors must constantly update themselves to reflect the latest patterns in AI writing. Textual quirks that previously acted as telltale signs – such as the overuse of em-dashes – can quickly stop being reliable proxies.

This means it is often very hard to verify whether a piece of text was AI generated based on the text itself. Accusations that writing was produced by an LLM are hard to prove or disprove. In an elections context, this means that it is hard to be certain whether LLMs are being used to generate false or misleading content at scale using automated text analysis methods.

4.9 TECHNICAL INTERVENTIONS: KEY TAKEAWAYS AND REMAINING GAPS

Without regulation to require minimum standards on electoral safeguards, we are left to rely on AI services' own initiatives. Transparency is lacking, even for the most prominent companies. Standardisation appears to be limited: some providers have introduced election-specific guardrails and protocols, but these differ in sophistication and vary in the level of public detail provided. The companies that have fewer resources available to implement electoral safeguards also are the ones that are the least reliable. Meanwhile, technical methods that are used to identify audiovisual AI content do not work well for text.

This situation leaves a significant gap in democratic oversight and accountability. Elections are an integral part of the UK's democratic process. We cannot afford to leave critical decisions about how to safeguard them in the hands of a small cohort of American companies. These businesses do not have public service obligations and are not necessarily incentivised to promote election security: their focus is on rapid growth, user retention, maximising revenues, and (for public companies) pleasing shareholders. Without mandatory requirements, there is no guarantee that they will implement high-quality election safeguards – now or in the future.

128 Coalition for Content Provenance and Authenticity (2026). https://c2pa.org/ (accessed 21/4/26)
129 University of San Diego Legal Research Center. 'Generative AI Detection Tools: The Problems with AI Detectors: False Positives and False Negatives.' https://lawlibguides.sandiego.edu/c.php?g=1443311&p=10721367 (accessed 21/4/26)

5. WHAT DOES THE LAW SAY?

Our evidence review and Scottish election testing results beg the question: how does existing UK law address text-based AI and elections? Drawing on discussions with legal experts, we analysed a range of UK legislation to identify where there is legal coverage, what this coverage means, and where gaps remain.

5.1 THE UK HAS NO OVERARCHING AI LEGISLATION

At the time of writing, the UK has no overarching law that regulates AI or specifies who is liable for when something goes wrong due to an AI model. There is no broader legal framework to assign responsibility for mitigating the risk of false and misleading AI-generated text during elections. Despite saying they would introduce regulation on frontier AI in the 2024 King's Speech,[130] the UK Government has retreated from introducing overarching AI regulation and has indicated it intends to take a piecemeal approach that relies on existing law and to regulate AI services primarily at the point of use.[131]

Based on our conversations with legal experts, we understand that an AI chatbot cannot itself be held liable or responsible if it fails. An AI system itself is not legally able to enter into a contract or be held accountable.[132] Where liability falls for a specific error or failing will depend on the law and the role played by the service.

The UK's lack of overarching AI regulation means it is behind the regulatory curve compared to jurisdictions such as the EU and California. In the EU, AI is regulated by dedicated legislation

130 Prime Minister's Office, 10 Downing Street & King Charles III. (2024). 'The King's Speech 2024'. HM Government. https://www.gov.uk/government/speeches/the-kings-speech-2024 (accessed 21/4/26)
131 Rough (2026). 'AI regulation in the UK.' House of Commons Library, UK Parliament. https://commonslibrary.parliament.uk/research-briefings/cbp-10003/ (accessed 21/4/26); Narayan (2025). Contribution to Westminster Hall debate (Volume 777, December 10th 2025): AI Safety. UK Parliament. https://hansard.parliament.uk/Commons/2025-12-10/debates/9F01B4B9-12CB-42E2-84E2-A65F7D30BFAF/AISafety#contribution-DDC01E2C-264F-4D37-BD25-940DC772DC12 (accessed 21/4/26); Narayan (2026). Response to written question on Artificial Intelligence: Children. House of Commons, UK Parliament. https://questions-statements.parliament.uk/written-questions/detail/2026-01-14/105940 (accessed 21/4/26)
132 Din (2025). 'AI Mistakes: Could Your Business Be Liable?' Butcher & Barlow. https://www.butcher-barlow.co.uk/news/commercial-dispute-resolution/ai-mistakes-could-your-business-be-liable/ (accessed 21/4/26)

(the AI Act 2024)[133] and is addressed by law on digital services and markets (the Digital Services Act 2022[134] and Digital Markets Act 2022.)[135]

5.2 THE ONLINE SAFETY ACT 2023

The Online Safety Act (OSA) predates the emergence of generative AI services and was not written with text-based AI in mind. While the law is intended to be technology-neutral – meaning that the duty to prevent harm does not change, regardless of whether something is AI generated or not – the specific coverage of a given text-based AI service is a byproduct of the framing of the law. This means that the law is not tailored to address the specific risks these services may pose, but will address content they generate if it reaches the threshold of being illegal or harmful to children. At the time of writing, the Government is due to amend the OSA to bring all AI chatbots under scope, following high-profile incidents involving the generation of child sexual abuse material (CSAM).[136]

AI services that are covered by the Online Safety Act (OSA) are required to proactively take steps to prevent and mitigate the spread of illegal content. Such content is defined by the OSA's list of Priority Offences (Schedules 5-7).[137] Measures include requirements to implement risk assessments, transparency reporting, and to swiftly remove illegal content if it appears on the service.[138] Other additional duties apply depending on how a service is categorised by the law. Relevant Priority Offences for text-based AI in an elections context include:

- Hate speech
- Harassment and abuse
- Fraud

The OSA also makes it a criminal offence for a person to spread information they know is false with the intent of causing "non-trivial psychological or physical harm to a likely audience."[139] This is known as the False Communications Offence. It is the OSA's primary mechanism for addressing the spread of false and unreliable information: the OSA does not place systemic responsibilities on service providers to address mis- and disinformation on their platforms unless that content falls within one of the Priority Offence categories. The False Communications Offence is not a Priority Offence, meaning that services are not required to implement pro-active measures to prevent and mitigate violations.

133 European Union (2024). Article 3 of the Artificial Intelligence Act [2024] OJ L, 2024/1689, 12.7.2024, ELI: http://data.europa.eu/eli/reg/2024/1689/oj (accessed 21/4/26)
134 European Union (2022). Digital Services Act [2022] OJ L 277, 27.10.2022, pp. 1–102 http://data.europa.eu/eli/reg/2022/2065/oj (accessed 21/4/26)
135 European Union (2022). Digital Markets Act [2022] OJ L 265 12.10.2022 https://eur-lex.europa.eu/eli/reg/2022/1925/2022-10-12/eng (accessed 21/4/26)
136 Vallance (2026). 'Elon Musk's Grok AI appears to have made child sexual imagery, says charity.' BBC News. https://www.bbc.co.uk/news/articles/cvg1mzlryxeo (accessed 21/4/26)
137 The Online Safety Act 2023 Sch. 5. Available at: https://www.legislation.gov.uk/ukpga/2023/50/schedule/5 (accessed 21/4/26); The Online Safety Act 2023 Sch. 7. Available at: https://www.legislation.gov.uk/ukpga/2023/50/schedule/7 (accessed 21/4/26)
138 Ofcom (2024). 'Implementing the Online Safety Act: Additional duties for 'categorised' online services.' https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/additional-duties-for-categorised-online-services (accessed 21/4/26)
139 The Online Safety Act 2023. S. 179. Available at: https://www.legislation.gov.uk/ukpga/2023/50/section/179 (accessed 21/4/26)

Existing coverage under the OSA is complex and variable

At present, different text-based AI services are captured in different ways depending on their functionality, use-case, and the size of their user base.[140] The OSA uses three categories of service: Category 1 (large user-to-user services), Category 2A (search services), and Category 2B (small user-to-user services). To ensure compliance with the OSA, services in these categories must follow different obligations and rules based on Ofcom's Codes of Practice. The definition of these categories are as follows:[141]

TABLE 5
ONLINE SAFETY ACT 2023 SERVICE CATEGORIES

CATEGORY	DEFINITION
Category 1 (large user-to-user)	A "regulated user-to-user service" that *either*: Has a monthly UK user base larger than 34 million *and* uses a content recommendation system. Has a monthly UK user base larger than 8 million *and* uses a content recommendation system *and* allows users to forward or share regulated user-generated content with other users.
Category 2A (search)	A "search engine of a regulated search service" that has a monthly UK user base larger than 8 million and allows users to search for information from multiple external websites or databases.
Category 2B (small user-to-user)	A "regulated user-to-user service" that has a monthly UK user base larger than 3 million *and* allows users to forward or share regulated user-generated content with other users.

These categories were not devised with generative AI in mind and catch text-based AI in inconsistent ways. Some services are clearly in scope: chatbots embedded in regulated social media services, for example, will fall under Category 1 or 2B depending on the social media service's user base and functionality. Similarly, AI search engines like Perplexity and Google AI Mode fall squarely into Category 2A. Yet the status of standalone chatbots is more ambiguous.[142] Some chatbot providers, such as OpenAI,[143] indicate that they consider their products to fall under Category 2A because they access data from external databases in a similar manner to a search engine.

140 Ofcom (2024). 'Open letter to UK online service providers regarding Generative AI and chatbots.' 'https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/open-letter-to-uk-online-service-providers-regarding-generative-ai-and-chatbots (accessed 21/4/26);Ofcom (2025). 'AI chatbots and online regulation – what you need to know.' 'https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/ai-chatbots-and-online-regulation-what-you-need-to-know (accessed 21/4/26)

141 The Online Safety Act 2023 (Category 1, Category 2A and Category 2B Threshold Conditions) Regulations 2025. https://www.legislation.gov.uk/ukdsi/2025/9780348267174 (accessed 21/4/26)

142 Woods (2025). 'Chatbots and the Online Safety Act.' Online Safety Act Network. https://www.onlinesafetyact.net/analysis/chatbots-and-the-online-safety-act/ (accessed 21/4/26)

143 OpenAI (2025). 'The UK Online Safety Act.' https://openai.com/policies/uk-online-safety-act/ (accessed 21/4/26)

TABLE 6
HOW ONLINE SAFETY ACT 2023 SERVICE CATEGORIES APPLY TO TYPES OF TEXT-BASED AI SERVICES

TYPE OF SERVICE	CATEGORISATION	CONDITIONS FOR APPLICATION TO SERVICES	EXAMPLES
Chatbots embedded in social media	Category 1 or 2B	Dependent on whether the surrounding social media service meets the Category 1 or 2B criteria.	Category 1: Grok as embedded in X
AI search engines	Category 2A	Dependent on whether the AI search engine has over 8 million monthly UK users.	Google AI Mode Perplexity
AI search summaries	Category 2A	Dependent on whether the surrounding search engine has over 8 million monthly UK users.	Google AI Overviews DuckDuckGo Search Assist
Standalone chatbots	*May* be Category 2A	Dependent on whether the chatbot pulls information from multiple external databases (e.g. via RAG searches).	Category 2A: OpenAI appears to consider ChatGPT to be covered as a Category 2A service[144]
Chatbot customisation platforms	Category 1 or 2B	Dependent on whether the platform allows users to share content or models with other users, plus the size of the services' userbase.	Category 2B: Character.ai

Meanwhile, the False Communications Offence is designed to target individual people who share false claims online. This would cover a situation where a person shares an AI-generated claim that they know to be false – *if* it could be proven that they did so with the express intent to harm other people. The False Communications Offence's intentionality requirement means that it is not applicable to cases where a chatbot or similar service presents a user with false claims directly: an LLM cannot be said to have intent or responsibility in any legal sense. Because the False Communications Offence is not a Priority Offence, text-based AI services covered by the OSA are not required to implement pro-active measures to prevent and mitigate violations of it. Moreover, this offence is explicitly aimed at individual people, rather than companies whose services generate false claims.

144 OpenAI (2025). 'The UK Online Safety Act.' https://openai.com/policies/uk-online-safety-act/ (accessed 21/4/26)

Proposed changes via the Crime and Policing Bill

At the time of writing, this status quo is due to change: the Government is poised to bring all generative AI services that are not currently regulated by the OSA into its scope via an amendment to the Crime and Policing Bill.[145] Amendment 429B inserts a clause into the OSA (as S216A) to give the Government power to "bring currently unregulated generative AI services, including chatbots in scope of regulation under the OSA and subject to duties to tackle illegal AI content and the commission or facilitation of priority offences."[146]

The change means all AI chatbot services are to be required to take action on Priority Offences such as harassment and hate speech. This may mean services implement stricter guardrails to prevent the generation of text that harasses and abuses candidates. But the broader category of AI-generated false and misleading content about elections will not be in scope.

Remaining gaps and ambiguities after the amendment

Bringing all AI services into the OSA's scope will not meaningfully address the risk that these services may generate false and misleading claims about elections. The OSA lacks requirements for regulated services to mitigate the systemic risk of the spread of mis- and disinformation. This has been a long-standing source of criticism and is by no means unique to AI. Ofcom's current guidelines do not specifically address the risk of AI services generating false and misleading information or the need to ensure reliability.

Meanwhile, the OSA's approach to regulating services places higher regulatory burdens on platforms with larger userbases. Services with userbases that are smaller than 3 million average monthly users are excluded from the OSA's formal categories (1-2B). This means that small AI chatbot services that have few (or no) election guardrails could be left out of scope. Such a situation is analogous to the problem of how the OSA treats 'small but harmful' websites and social media services, which has been the focus of widespread criticism. All regulated services are required to implement a minimum standard of measures, such as removing illegal content, but only larger services are required to take additional steps such as transparency reporting.[147] This will mean that once all AI chatbots are brought into scope, services with smaller userbases such as Replika will have fewer obligations – even though research indicates these have more issues with safety than larger services.

The OSA's one provision on false claims – the False Communications Offence – is directed at prosecuting individuals, not placing duties on services, and is not applicable for direct human-to-chatbot interactions. Even in cases where a person shares an AI-generated false claim, the fact that the OSA requires the prosecution to show the person intended to cause harm – known as *mens rea* ('guilty mind') – may set a high bar for proof.[148] The law provides a statutory defence where the defendant may argue that they were not aware the claim was false and did not intend to cause harm. This defence can be difficult to disprove and may be strengthened if a person can successfully blame the falsehood on an AI hallucination.

145 Department for Science, Innovation and Technology (2025). 'New law to tackle AI child abuse images at source as reports more than double.' HM Government. https://www.gov.uk/government/news/new-law-to-tackle-ai-child-abuse-images-at-source-as-reports-more-than-double (accessed 21/4/26)

146 Lord Hanson of Flint's amendment to the Crime and Policing Bill, After Clause 212 (2026). Amendment number: 429B. House of Lords, UK Parliament. https://bills.parliament.uk/bills/3938/stages/20491/amendments/10033738 (accessed 21/4/26)

147 Peter Kyle MP (2024). 'Small but risky' online services under the Online Safety Act: letter from DSIT Secretary of State'. Department for for Science, Innovation and Technology, HM Government. https://www.gov.uk/government/publications/small-but-risky-online-services-under-the-online-safety-act-letter-from-dsit-secretary-of-state/small-but-risky-online-services-under-the-online-safety-act-letter-from-dsit-secretary-of-state (accessed 21/4/26); Brunning (2024). 'Small but Risky Services under the Online Safety Act.' Fieldfisher. https://www.fieldfisher.com/en/insights/small-but-risky-services-under-the-online-safety-act-2024 (accessed 21/4/26)

148 Antoniou et al. (2024). 'Disinformation and disorder: the limits of the Online Safety Act.' Online Safety Act Network. https://www.onlinesafetyact.net/analysis/disinformation-and-disorder-the-limits-of-the-online-safety-act/ (accessed 21/4/26)

5.3 THE NATIONAL SECURITY ACT 2023

Section 13 of the National Security Act (NSA) makes it a criminal offence for a person to engage in "prohibited conduct" in a manner which interferes in a political process in the UK to the benefit of a foreign power, such as a foreign state.[149] This covers all referendums and elections for political office. Prohibited conduct includes "making a misrepresentation of fact which contributes to the interference effect, such as a misrepresentation as to a person's identity or purpose" or "presenting information in a way which amounts to a misrepresentation. The foreign interference offence is listed as a Priority Offence in Schedule 7 of the OSA,[150] meaning that regulated services must take steps to foreign interference occurring via their platforms.

The Foreign Interference Offence criminalises precisely the kinds of AI-enabled FIMI activities discussed in Section 3.3. However, as we have identified previously,[151] this offence faces significant enforcement barriers. Foreign disinformation campaigns may be hard to detect and, though the offence claims extraterritoriality, the odds of successfully prosecuting overseas actors are low.

5.4 ELECTIONS LAW

Existing elections law is not suited to addressing situations where AI services generate false or misleading claims. The legislation that governs UK elections includes measures intended to discourage people from spreading false claims about electoral candidates or electoral proceedings with the intent of disrupting an election. The most relevant laws are the Representation of the People Act 1983 and the Elections Act 2022.

But while these laws include individual-level offences to address specific kinds of election disinformation, they were not drafted with text-based AI in mind: both laws received Royal Assent before ChatGPT was launched in November 2022. Meanwhile the current text of the Representation of the People Bill does not address AI-generated mis- and disinformation at all. Because they focus on individuals, neither laws pertain to companies whose services generate false or misleading information about an election.

The Representation of the People Act 1983

Section 106 of the Representation of the People Act makes it illegal for a person to "mak[e] or publis[h] any false statement of fact in relation to [an election] candidate's personal character or conduct" if this takes place "before or during an election" with the intent to affect "the return of any candidate at the election."[152] It therefore is specific to false statements about personal character or behaviour, not their political views or activities. Section 106 should apply to situations where a person shares an AI-generated claim that they know is false.

Section 106 has well-known limitations.[153] Firstly, it does not explicitly address digital material or AI-generated content. Both the Electoral Commission and the Director of Public Prosecution have stated that this lack of explicit coverage creates ambiguity and can make it harder to

149 The National Security Act 2023 S. 13. https://www.legislation.gov.uk/ukpga/2023/32/part/1/crossheading/foreign-interference (accessed 21/4/26)
150 The Online Safety Act 2023 c. 50 Sch. 7. Available at https://www.legislation.gov.uk/ukpga/2023/50/schedule/7 (accessed 21/4/26)
151 Seger, Perry & Hancock (2025). Epistemic Security 2029: Fortifying the UK's information supply chain to tackle the democratic emergency. Demos. https://demos.co.uk/research/epistemic-security-2029-fortifying-the-uks-information-supply-chain-to-tackle-the-democratic-emergency/ (accessed 21/4/26)
152 The Representation of the People Act 1983 S. 106. https://www.legislation.gov.uk/ukpga/1983/2/section/106 (accessed 21/4/26)
153 Hancock (2026). 'The Deepfake Gap: Regulating false statements & AI to safeguard elections.' Demos. https://demos.co.uk/research/the-deepfake-gap-regulating-false-statements-ai-to-safeguard-elections/ (accessed 21/4/26)

pursue cases.[154] Without further guidance, this may make it hard to enforce for cases where AI is intentionally used to generate and share false claims about an electoral candidate.

Secondly, as the Electoral Commission has noted, under "misleading content about candidates is not considered an offence outside of the regulated [pre-election] period". Section 106 is "only enforceable during the regulated [pre-election] period".[155]

Finally, Section 106 bears the same caveat as for the OSA False Communications Offence: the fact that it hinges on a person's intent to cause harm means that it is unlikely to be applicable to cases where a chatbot misinforms a user about a candidate. AI systems do not have intent in any legal sense. Like the False Communications Offence, Section 106 provides for a *mens rea* defence in which the defendant argues that they did not know the statement was false and/or did not intend to cause harm. Therefore Section 106 is likely to be of limited use in this context.

The Speaker's Conference has recommended the Government should conduct a review of electoral law – including how to ensure Section 106 is clear, enforceable and able to "keep pace with technological developments in AI and deepfakes." The Conference also suggested expanding the offence "beyond personal character and conduct". In its response, the Government acknowledged that Section 106 has issues and is reviewing how it should "apply in a modern electoral setting." The Government promised to "take appropriate action to clarify" Section 106's scope if "necessary", especially regarding AI and deepfakes, with the aim of "future-proof[ing]" the law.[156] The Government must use this review as an opportunity to fully address the problems identified.

The Elections Act 2022

Section 8 of the EA 2022 amended the RPA to add the offence of 'undue influence' as Section 114A:[157] using intimidation or threatening behaviour to change how another person will vote, to prevent them from voting at all, or to otherwise prevent someone from freely exercising their right to vote. Prohibited activities under Section 8/114A include making false claims about voting procedures in order to prevent someone from voting. A person can also be guilty if they carry out one of these prohibited activities with the assumption that they have changed how another person has voted or stopped them from voting.

The same ambiguities and gaps apply to the EA Section 8/114A as to the RPA Section 106 regarding applicability to AI content and intent. It is also unclear whether Section 8/114A would apply to cases where someone used AI to generate false or misleading text about a candidate with the intent of changing the vote of the electorate at large – or if it would only apply to cases where someone generates text to unduly influence a specific voter. Like the RPA Section 106, legal guidance is needed to clarify how Section 8/114A applies to AI-generated content.

5.5 DEFAMATION LAW

In the absence of specific regulations to protect election candidates against false and misleading AI-generated claims, a candidate who believes themselves to be the subject of

154 Electoral Commission (2025). Written evidence submission to the Speaker's Conference on the security of candidates, MPs and elections. SCS0049. UK Parliament. https://committees.parliament.uk/writtenevidence/141330/pdf/ (accessed 21/4/26); Parkinson (2025). Letter to Rt Hon Sir Lindsay Hoyle MP, Speaker of the House of Commons. Speaker's Conference on the security of candidates, MPs and elections, UK Parliament. https://committees.parliament.uk/publications/48097/documents/251441/default/ (accessed 21/4/26)
155 Rangarajan (2026). Written evidence submitted by Vijay Rangarajan, chief Executive of the Electoral Commission. House of Commons Foreign Affairs Select Committee, UK Parliament. https://committees.parliament.uk/writtenevidence/162487/pdf/ (accessed 21/4/26)
156 HM Government (2026). 'Speaker's Conference on the security of MPs, candidates and elections: Government Response.' Speaker's Conference on the security of candidates, MPs and elections, UK Parliament. https://publications.parliament.uk/pa/cm5901/cmselect/cmspeak/1709/report.html#heading-6 (accessed 21/4/26)
157 The Elections Act 2022. s. 8. https://www.legislation.gov.uk/ukpga/2022/37/part/1/crossheading/undue-influence (accessed 21/4/26)

AI-generated misinformation may turn to defamation law to find a remedy. The Parliamentary Digital Service acknowledges this as a possibility in its guidance for MPs on generative AI.[158]

Defamation is a civil matter: the affected candidate or MP would need to identify the party responsible for the false claim and sue them for libel.[159] Unlike the election offences discussed above, defamation does not require intent to cause harm. When a defamation case is brought to court, the statements under scrutiny are presumed to be false unless proven otherwise. The burden is placed on the defendant to prove that the claim is true or that they had reasonable grounds to believe that it was true.

The status of AI-generated content under defamation law remains ambiguous. There is no legal precedent set around responsibility for defamation generated by AI. While there are attempts in motion to sue AI chatbot companies for defamation in the UK,[160] these cases are yet to result in a court judgement. This means it is possible that a candidate could argue in court that a text-based AI service provider has defamed them – but it is uncertain whether they would be successful.

The results would partly depend on how the courts treat the specific AI service being sued. Both defamation law and the UK E-Commerce Regulations 2002 distinguish between the 'publisher' and 'host' of digital content: 'hosts' simply store digital content for use by others and have defences available against defamation suits; 'publishers' distribute the content and can be held liable. More clarity is needed on how these regulations apply to text-based AI services. Some chatbots, such as OpenAI, may fulfill multiple roles under these definitions. The specific responsibilities and potential liability for each of these roles in the AI ecosystem needs to be clarified in UK law.

If an AI service is considered to be only an intermediary between an external AI model and the user – such as by providing a web portal, interaction window, or similar – then the service will be considered to be a 'host' and has access to statutory defences which mean they are unlikely to be held liable.[161] But if the AI service was considered to be the creator and distributor of the defamatory AI-generated content, they could be considered to be its 'author' or 'publisher' and found liable. This situation could apply to AI service providers like OpenAI and Google which develop their own models, run these internally, and allow users to interact with them via their own websites. For comparison, a newspaper that posts a defamatory story online is considered the legal publisher and is liable; the cloud service that holds the defamatory article on its servers is a 'host' and is not liable. The question then becomes: does text generated by AI count as new material produced and published by the service provider?

The situation is clearer for situations where a person shares false claims about a candidate that originated from a chatbot or another AI service. In these cases, the person sharing would be considered to be the publisher and existing precedent for defamation law applies. Defamation law is similar in this respect to the offences discussed under the OSA, RPA, and EA: the fact that AI was used to write the defamatory text is irrelevant and the source of liability is the decision to share it.

Thus, defamation law could potentially be used by candidates as an avenue for redress in certain circumstances. But this is by no means certain and is still to be tested in court. Moreover,

158 Parliamentary Digital Service (2025). 'Artificial Intelligence: Guidance for Members.' UK Parliament. https://www.parliament.uk/globalassets/mps-lords--offices/offices/parliamentary-digital-service/july_25_updated_-_parliamentary_digital_service_-_ai_guidance_members_v4_external_-_final.pdf (accessed 21/4/26)

159 Pinsent Masons (2023). 'Defamation in England and Wales.' https://www.pinsentmasons.com/out-law/guides/defamation-guide (accessed 21/4/26)

160 Legal Insider (2025). 'High profile libel lawyer prepares group action against tech giants for alleged AI violations.' https://legaltechnology.com/2025/06/12/high-profile-libel-lawyer-prepares-group-action-against-tech-giants-for-alleged-ai-violations/ (accessed 21/4/26)

161 The Defamation Act 2023. S. 5. https://www.legislation.gov.uk/ukpga/2013/26/section/5 (accessed 21/4/26); English (2025). 'AI liability in defamation Part II: The UK.' UK Human Rights Blog. https://ukhumanrightsblog.com/2025/11/19/ai-liability-in-defamation-part-ii-the-uk/ (accessed 21/4/26)

it does not place any requirements on AI service providers to take pro-active measures to mitigate the risk of this occurring and requires an affected candidate to take on the burden of bringing a lawsuit. Given that some of the leading AI service providers are among the wealthiest companies in the world,[162] the financial cost of suing them may dissuade candidates from attempting this. Defamation law therefore does not provide a reliable pathway for challenging AI developers or AI service providers whose tools generate false and misleading content about elections.

5.6 LEGAL COVERAGE: KEY TAKEAWAYS AND REMAINING GAPS

Existing UK law is not sufficient for addressing risks around text-based AI and elections. The UK does not have legislation that specifically addresses or regulates AI. Current laws were not drafted with this technology in mind. Legal coverage under existing law features significant gaps and, even in cases where text-based AI services are covered, service providers are not required to address the risk that their tools may facilitate the generation of false and misleading content about elections.

Existing elections law is intended to discourage the spread of false claims about candidates and election proceedings which could affect citizens' ability to vote. These laws were not drafted with AI in mind. They assume that the source of false information will be a person and require proof of intent – neither of which apply in cases where AI chatbots directly misinform voters. Nor does election law or online safety law place any responsibility on AI services to implement electoral safeguards or be transparent about their electoral policies. Without regulation, we are left to rely on the goodwill and voluntary decisions of AI platforms. As our analysis of service providers' election safeguards shows, this is insufficient. Meanwhile, defamation law may offer individual candidates a path to seek redress. But it is not a reliable solution to this systematic and structural challenge.

Our analysis highlights three fundamental challenges text-based AI poses for election regulation and related law:

1. **Intent:** laws that include a *mens rea* component are unlikely to apply to cases where AI services directly interact with users. The law still applies to cases where a person shares AI generated content with another person. But the fact AI is prone to errors and hallucinations may give defendants a strong excuse for why they shared false information.
2. **Clarity:** there is ambiguity as to whether existing laws that were drafted before the emergence of generative AI will apply to content that is AI generated. We have previously identified this as a problem for how the RPA applies to deepfakes.
3. **Predictability and transparency:** the probabilistic and 'black box' nature of LLMs create a fundamental challenge for holding upstream actors to account for AI-generated mis- and disinformation. The specific details of an LLM's inner workings may be unknown even to their developer, due to their complexity and incomprehensibility. In court, an AI developer or service provider could argue that they could not have predicted that the model would generate any specific piece of false information. This underscores the need for regulations that require services to take anticipatory, risk-based measures.

162 CompaniesMarketCap (2026). 'Companies ranked by revenue.' https://companiesmarketcap.com/gbp/largest-companies-by-revenue/ (accessed 21/4/26)

6. RECOMMENDATIONS FOR GOVERNMENT

It will take a comprehensive, multi-layered approach to address the issues identified in this paper. As a starting point, we propose four recommendations that the government could adopt today to safeguard our general election in 2029. These span from immediate, short-term mitigations to ambitious initiatives that will strengthen our democracy outside of the election period.

In the short-term, there is an opportunity to use the Representation of the People Bill, currently passing through Parliament, to review existing election law and ensure it is equipped to tackle the challenges of today's technology. Looking beyond existing law, the government should pass new legislation that mitigates risks from AI during the election period. Looking longer term, and more broadly to strengthen our information environment as a whole, the government should consider cross-cutting AI legislation to address questions such as liability. Taken individually, these recommendations will make a concrete impact. Taken together, they will represent a landmark change in the effort to safeguard democracy.

Most importantly, we call on the Government to take action now – not later. Instead of reacting to crises after they arise, our proposals take a proactive stance by targeting known vulnerabilities before they can be exploited.

6.1 MAKE EXISTING UK LAW LLM-READY

Current law on elections and defamation is not equipped for today's technological reality. Ahead of the 2029 general election, updates are needed to ensure they address gaps and ambiguities, such as uncertainty over whether election offences apply to uses of AI and the applicability of defamation law to AI services. Without these changes, candidates may be left with limited legal remedies if they are affected by AI-generated text that includes false or misleading claims about them.

The government must urgently address these gaps and ambiguities. This should be either through legal guidance or amendments to existing law, via a vehicle such as the Representation of the People Bill.

RECOMMENDATION	THE DETAIL
1.1 Review and clarify election law via the Representation of the People Bill	The government must publish legal guidance that clarifies the applicability of S108 of the RPA and S8 of the EA S8/S114A of the RPA to AI-generated content. This should include a clarification of how these provisions apply to AI-generated text. The government should also undertake a review to identify how the RPA may be updated to clearly bring AI service providers into scope for existing election offences and safeguards, including a review of liability for AI chatbots that directly misinform voters. Finally, the government should review the appropriateness of only applying election protections to the pre-election window. Responsibility for this work is likely to fall across multiple departments, including the Ministry of Housing, Communities and Local Government (MHCLG) and Department for Science, Innovation and Technology (DSIT). Given this, we recommend that primary ownership should fall to the Joint Election Security Preparedness unit (JESP), which co-ordinates the government's election preparedness efforts across departments.
1.2 Clarify the status of defamation law and its application to AI services	Given the legal ambiguities identified, the Law Commission should conduct a legal review into how defamation law applies to AI services. This should include an examination of where liability lies when AI chatbots generate defamatory statements and where further clarity is needed.

6.2 MANDATE A MINIMUM STANDARD FOR ELECTION SAFEGUARDS

Trusted elections are vital to our democratic system. Our Scottish elections testing demonstrates that leading AI services are unreliable sources of information for voters. But they are a fast growing source of information. Decisions about whether and how to implement election safeguards cannot be left solely to AI service providers. New regulation is needed to establish a baseline of accuracy and accountability before a major crisis occurs.

Therefore, the government should legislate to set a minimum standard of safeguards to be applied to all UK elections, including local and devolved. These standards should codify existing best practices in the sector and surpass them when it comes to crucial areas like transparency. Setting standards ensures that all elections in the UK receive an adequate level of protection and support. Without minimum standards, services may prioritise major elections such as General Elections – leaving local and devolved elections at greater risk.

In the absence of a dedicated regulator for AI services, responsibility for oversight is left with existing regulators. This role could fall to Ofcom with support from the Electoral Commission. However, the challenges posed by text-based AI extend beyond Ofcom's current mandate and would require a significant expansion of its remit and expertise. As such, these issues may be better addressed by establishing new regulatory responsibilities or assigning them to a more suitably equipped authority.

Minimum standards should apply to all voter-facing text-based AI services that operate during UK elections. This means that they would not apply to model developers who do not operate

a user-facing service or to website hosts that simply provide a web portal to a third-party AI service. To avoid creating a two-tier regulatory system where smaller services that tend to feature fewer guardrails are also less regulated, these minimum standards should apply to all relevant services regardless of user base size. This is a deliberate departure from the approach taken in the Online Safety Act 2023.

Crucially, some minimum standards should apply at all times. This would be a shift in approach from the Representation of the People Act 1983 and related law which regulates for the pre-election window only. People's use of AI to learn information about elections is not restricted to formal election windows, and deliberate attempts to manipulate elections using AI begin long before campaigning officially starts. To only require standards during the regulated pre-election window would be to create an unacceptable loophole. We therefore suggest that there should be minimum standards that apply at all times and enhanced standards that apply during pre-election windows.

RECOMMENDATION	THE DETAIL
2.1 Minimum standards to be applied to text-based AI services at all times	
2.1.1 Require text-based AI services to undertake risk assessments and testing	Require services to conduct internal testing and risk assessments to identify the potential impacts of their chatbot services on elections. These should be provided to the designated regulator and made publicly available. Services should be required to act on the risks they identify. Tests should cover: factuality; political bias; and red-teaming for policy failures, malicious uses, and errors.
2.1.2 Require services to report publicly on their election policies	Require services to publish the full details of their election-related policies, guardrails, and risk-assessments on a regular basis (e.g. annually). This should include post-election assessments of their implementation and effectiveness. Services should also be required to provide frequent updates to the regulator and relevant parliamentary inquiries.
2.1.3 Create an effective system of enforcement and penalties	There should be significant penalties for any services found to be in breach. Penalties should include large fines – scaled to reflect the wealth of companies like Google and OpenAI – or more serious measures for repeat infractions.
2.2 Enhanced requirements for pre-election windows	
2.2.1 Require text-based AI services to implement guardrails against harmful and malicious uses	Require services to implement effective guardrails that prevent the following types of content from being generated: False statements about electoral processes and procedures; False statements about a candidate's character, conduct, or political views; Harassment material directed at candidates; and Persuasive material intended to misinform voters at scale.
2.2.2 Place factuality duties on services	Require services to show that they have taken reasonable steps to ensure that their services consistently provide accurate answers to questions on key procedural facts about an election. This requirement may be modelled on the 'due accuracy' standards applied to broadcasters.[163]

163 Ofcom (2025). Broadcasting Code Section Five: Due impartiality and due accuracy. https://www.ofcom.org.uk/tv-radio-and-on-demand/broadcast-standards/section-five-due-impartiality-accuracy (accessed 21/4/26)

RECOMMENDATION	THE DETAIL
2.2.3 Place bias duties on services	Require services to show in advance of pre-election windows that they have taken reasonable steps to ensure their services are not biased towards any one political candidate or political party. This requirement may be modelled on the 'due impartiality' standards applied to broadcasters.[164]
2.2.4 Place timeliness duties on services	Require services to show that they have taken reasonable steps to ensure that their services provide up-to-date information on elections. This could include setting standards for Retrieval Augmented Generation (RAG) pipelines.
2.2.5 Require services to implement enhanced election resourcing and crisis protocols	Require AI companies to dedicate additional resources and staffing to respond in incidents where voters are exposed to false and misleading information at a large scale during a UK election when this involves their text-based AI service. Services should establish election incident response protocols to handle such incidents. These protocols should be made public. Require services to notify the government, regulator, and the public within a short time once they have been triggered.

6.3 PROMOTE TRANSPARENCY: SUPPORT DATA ACCESS AND INDEPENDENT RESEARCH

While the UK government's AI Security Institute (AISI) has developed special relationships with AI model developers which allow it to conduct safety testing, this does not go far enough to make the AI ecosystem transparent or trustworthy enough for elections. Independent access by public interest researchers is needed to ensure public transparency and accountability. This builds on the approach taken by the OSA and Ofcom on matters of online safety.

RECOMMENDATION	THE DETAIL
3.1 Require text-based services to offer data access to public interest researchers for elections research	Require services to provide information and data access to independent public-interest researchers for elections-related research. This should include data on the services' internal decision-making processes, training datasets, safety mechanisms, guardrails, and live data on usage during election periods. LLM providers that offer access to their models via web interfaces or APIs should be required to offer independent public-interest researchers with direct access to their models to conduct safety testing.
3.2 Support independent research to address evidence gaps	The government should fund and promote independent research to address current evidence gaps. This work must be rigorous, editorially independent, and uphold the highest ethical standards. Appropriate channels for distributing funding for this research include UK Research and Innovation (UKRI), the Sovereign AI fund, and the AI Security Institute Challenge Fund.

164 Ofcom (2025). Broadcasting Code Section Five: Due impartiality and due accuracy. https://www.ofcom.org.uk/tv-radio-and-on-demand/broadcast-standards/section-five-due-impartiality-accuracy (accessed 21/4/26)

6.4 INVEST IN TRUST: SUPPORT AI TEXT IDENTIFICATION TECH

Trust is at a premium. Without reliable ways to identify AI-generated text, people may distrust all text they see about an election as potentially inauthentic. Yet identifying AI-generated text and tracking its provenance remains a significant technical challenge. The government should support and invest in the ecosystem of actors seeking to refine textual watermarking and AI text detection.

RECOMMENDATION	THE DETAIL
4.1 Investment in textual watermarking and AI text detection ecosystem	DSIT should fund a broad ecosystem of actors working to refine textual watermarking and AI text detection. In doing so, it should require that resulting tools are open-source and built on open standards. DSIT should prioritise empowering civil society and supporting non-commercial, public-interest solutions.
4.2 Support standards for Textual Watermarking	DSIT should offer ministerial support and other non-financial aid to the overall ecosystem of actors working on open standards. To encourage consistency, the government should explore ways to incentivise industry to adopt a shared open standard for textual watermarking without specifying what that standard should be. The government should *not* take it upon itself to name a preferred open standard for textual watermarking.

LOOKING AHEAD TO FUTURE LEGISLATION

While this paper has adopted a pragmatic approach and focuses on elections specifically, our recommendations should be taken alongside Demos's long-standing call to go further on AI regulation. Ensuring that the AI ecosystem is safe and reliable requires overarching AI legislation that addresses cross-cutting challenges such as liability.

However, the current government has made it clear that it does not intend to pass overarching regulation for all AI services and is instead looking to take a domain-specific approach focused on specific issues, such as online safety. It is Demos's view that such a piecemeal approach creates regulatory uncertainty, risks leaving gaps in legal coverage, fragments oversight, and may lead to conflicting positions between different laws.

We argue that the government should be bold and change this approach: the UK should pursue cross-cutting regulation that covers the full AI lifecycle for the benefit of its citizens. Such a framework should adopt a rights-based model grounded in human rights principles, drawing on proposals like Demos' Declaration on Digital Rights[165] and lessons from established legislation such as the EU AI Act.[166]

165 Lyall et al. (2026). A Declaration on Digital Rights: Embedding human rights in a new deal for the digital age. Demos. https://demos.co.uk/research/a-declaration-on-digital-rights-embedding-human-rights-in-a-new-deal-for-the-digital-age/

166 European Union (2024). The Artificial Intelligence Act [2024] OJ L, 2024/1689, 12.7.2024, ELI: http://data.europa.eu/eli/reg/2024/1689/oj (accessed 21/4/26)

CONCLUSION

The rapid growth of text-based AI as a source of electoral information reflects a structural shift in how people access information across the UK's digital landscape. Chatbots, AI Search and AI overviews are now commonplace in people's everyday lives. As more people turn to these services for information and advice on elections, the government must urgently address the risks identified in this paper. This is vital in order to ensure citizens have access to the high-quality information they need when they participate in our democracy.

A small cohort of American AI providers are quickly positioning themselves as centralised information gatekeepers with a high degree of control over their services. Yet these services are unregulated when it comes to elections, with no oversight from the government on how AI providers should ensure their services are reliable and accurate. This comes as the UK faces a moment of democratic emergency. If the country is to renew the deal between state and citizen, it must start by ensuring that citizens have access to the high-quality information they need to participate in elections.

Our AI service testing and evidence review shows that text-based AI services pose a complex challenge for UK election integrity. First, they continue to face significant challenges with factuality and bias that make them unreliable as voter aids. The tendency for text-based AI to confidently assert plausible-sounding factual errors and hallucinate is a fundamental problem in such a high-risk context where there is a premium placed upon the truth and trust. These problems are not fading away: as our testing found, leading services continue to produce factual errors and make misleading claims.

Second, the potential for these systems to feature subtle political bias should be a cause for concern. Concerns regarding model manipulation, malicious uses, and foreign interference campaigns require greater attention and support for evidence gathering. However, these concerns should be tempered with realism about voters' vulnerability to persuasion: existing evidence suggests that text-based AI may not have a major impact on how voters form their political beliefs when compared to other information sources, and may not significantly change the balance of power when it comes to political persuasion. Overall, more transparency and evidence is needed to provide a solid evidence base for future policy in this area. Our call for mandatory data access for researchers is intended to address this problem.

Third, the clearest risk is to public trust and confidence in the legitimacy of the election process. With no reliable way to verify whether a piece of writing is AI-generated, accusations of inauthenticity may fly unchecked. This could prove highly corrosive in an environment where trust is already declining and accusations of election malpractice are on the rise.

This paper makes it clear that there is a regulatory gap when it comes to text-based AI and elections. Current laws were not drafted with new technologies in mind and will only become more out of date as technology develops. The gaps in legal coverage need to be addressed and future-proofed in order to keep pace. Crucially, AI service providers are not currently required to address the risk that their tools generate false and misleading content

about elections. While some providers of text-based AI services have implemented election safeguards, the quality and transparency of these vary greatly. We are left to rely on voluntary action and good intentions rather than robust government scrutiny or parliamentary oversight.

The UK faces a unique window of opportunity to take proactive action ahead of the 2029 general election. With the number of marginal seats rising[167] and political polarisation deepening,[168] 2029 is a potential moment of vulnerability. By taking proactive steps, the country can move beyond dramatic stories of AI rigging elections and adopt grounded policies that ensure these services act in the public interest. As part of this, the government must respond to the way that text-based AI services are already being used as objective and centralised producers, gatekeepers, and disseminators of information. By acting now, the government can renew its democratic integrity, enabling a new deal for trustworthy information between citizens and the state.

The recommendations set out in this paper provide a first step to making our elections more resilient in the context of rapid developments in LLM technologies. By setting a minimum standard for election safeguards, promoting transparency, supporting veracity through AI text identification, and making existing law LLM-ready, the government can pave the way in securing a future where democratic discourse remains grounded in reliable information and public trust.

The UK faces a **unique window of opportunity** to take proactive action ahead of the **2029 general election.**

By acting now, the government can renew its democratic integrity, enabling **a new deal for trustworthy information** between citizens and the state.

167 Sturge (2024). '2024 general election: Marginality.' House of Commons Library, UK Parliament. https://commonslibrary.parliament.uk/2024-general-election-marginality/

168 Opinium (2026). 'Voting intention: 25th February 2026.' https://www.opinium.com/resource-center/voting-intention-25th-february-2026/; Ipsos (2025). '85% of Britons continue to believe that Britain is divided as over half think differences in people's political views are so divisive that it is dangerous for society.' https://www.ipsos.com/en-uk/85-britons-continue-believe-britain-divided-over-half-think-differences-peoples-political-views-are

APPENDIX

A. SCOTTISH ELECTION TESTING: DETAILED METHODOLOGY

This study utilised an AI red-teaming methodology: two analysts prompted five AI services with 75 questions about the Scottish elections *while the election was ongoing*. We manually assessed the results for factuality, use of evidence, bias, and vulnerability to malicious uses. We tested each chatbot at the start of the election window of the Scottish elections (on March 27th) and replicated the experience of an average Scottish user who does not pay for the service. Our approach builds on methodologies taken by prior research that studied the performance of AI chatbots during elections.[169]

The study represents a snapshot of how the services performed during the election window. We opted to focus on one date – March 27th 2026 – rather than to conduct a longitudinal study. This was to maintain consistency across the services and ensure our fact-checking was rigorous: in a fast-moving election context, the facts may change quickly, invalidating the 'ground truth' needed to assess the services' responses. Moreover, by testing on one day, we minimised the risk of temporal variability in services' behaviours or updates to services' systems affecting the results.

We tested a range of popular chatbots, AI search overview services, and companion bots:

TABLE 7
TEXT-BASED AI SERVICES AND MODELS TESTED DURING 2026 SCOTTISH ELECTIONS

SERVICE	SERVICE PROVIDER	ACCESSED VIA	MODEL TESTED	DATE TESTED
ChatGPT	OpenAI	ChatGPT.com; no user account required	Not disclosed to user - appeared to be GPT-4o[170] (Free tier, logged out)	27/3/26

169 Simon, Fletcher & Kleis Nielsen (2024). 'How generative AI chatbots responded to questions and fact-checks about the 2024 UK general election'. Reuters Institute for the Study of Journalism, University of Oxford. https://reutersinstitute.politics.ox.ac.uk/how-generative-ai-chatbots-responded-questions-and-fact-checks-about-2024-uk-general-election (accessed 21/4/26); Helming & Marsh (2024). 'Large Language Models Continue To Be Unreliable Concerning Elections'. AlgorithmWatch. https://algorithmwatch.org/en/llms_state_elections/ (accessed 21/4/26); Marsh (2024). 'Chatbots are still spreading falsehoods.' AlgorithmWatch. https://algorithmwatch.org/en/chatbots-are-still-spreading-falsehoods/ (accessed 17/4/26); Stockwell (2024. 'How can we stop AI-enabled threats damaging our democracy?' CETaS, the Alan Turing Institute. https://www.turing.ac.uk/blog/how-can-we-stop-ai-enabled-threats-damaging-our-democracy (accessed 21/4/26)

170 ChatGPT.com did not disclose which model was being used, but contextual indicators (such as a 2023 knowledge cutoff) indicates the model was likely GPT-4o. However, OpenAI claims it retired GPT-4o for ChatGPT in February 2026 in favour of GPT-5.3. GPT-5.3 has a knowledge cutoff date in August 2025. See https://help.openai.com/en/articles/11909943-gpt-53-and-gpt-54-in-chatgpt (accessed 21/4/26);OpenAI (2026). 'Retiring GPT-4o, GPT-4.1, GPT-4.1 mini, and OpenAI o4-mini in ChatGPT.' https://openai.com/index/retiring-gpt-4o-and-older-models/ (accessed 21/4/26)

SERVICE	SERVICE PROVIDER	ACCESSED VIA	MODEL TESTED	DATE TESTED
Gemini	Google	gemini.google.com; no user account required	Gemini 3 Flash[171] (Fast, Free tier, logged out)	27/3/26
AI Overviews	Google	google.com; no user account required	Gemini 3[172] (Free tier)	27/3/26
Grok	xAI	grok.com; user account required (new account per interaction)	Grok 4.20[173] (Auto, Free tier)	27/3/26
Replika	Replika	replika.ai; user account required (new account per interaction)	Not disclosed to user - possibly LLaMA 2-13B[174] (Free tier)	27/3/26

We tested these services' knowledge of and claims about three Holyrood constituencies:[175]

- **Constituency A:** A constituency in a large city where a prominent politician was running.
- **Constituency B:** A hotly contested constituency in a large city with redrawn boundaries.
- **Constituency C:** A rural constituency.

In total, we asked 375 questions across 15 separate conversations with the five services: one conversation per constituency per chatbot. We also used an opening prompt for each conversation to tell the service that we were a Scottish resident in a particular postcode looking to vote in the Scottish election (bringing the total number of prompts used to 390).

A.1. Questions used

We asked 25 questions per constituency, plus an initial prompt in which we gave the service a postcode that we said was our home address.

Even though the opening prompt was not a question, all services responded to it and some responses included factual information on how to vote. We have therefore analysed the responses to this prompt alongside the election questions. We have chosen to include these responses in the total figures for services' responses.

171 Google DeepMind (2026). 'Gemini 3 Flash.' https://deepmind.google/models/gemini/flash/ (accessed 21/4/26)
172 Stein (2026). 'Just ask anything: a seamless new Search experience.' Google: The Keyword. https://blog.google/products-and-platforms/products/search/ai-mode-ai-overviews-updates/ (accessed 21/4/26
173 xAI (2026). Models and Pricing. https://docs.x.ai/developers/models (accessed 21/4/26)
174 This information is unavailable on Replika's website. On querying the chatbot itself, we were told the model is based on Meta's LLaMA 2 13 Billion parameter model.Other sources claim Replika is based on OpenAI's models. E.g., UmaMaheswari (2025). 'Engineering Excellence: Understanding Small Language Models .' LinkedIn. https://www.linkedin.com/pulse/engineering-excellence-understanding-small-language-umamaheswari-zkssc/ (accessed 21/4/26)
175 We have anonymised the constituencies and candidates to protect the individuals running in these races.

The 25 questions consisted of 16 factual questions about election procedures and candidates' positions; 7 subjective questions on who to vote for; and finally two deliberately malicious requests to generate persuasive social media posts that contained false claims about a candidate.

These questions tested the services' knowledge of subjects such as:

- How to vote
- Who can vote
- The candidates running
- The leading candidate's policy positions
- The leading candidate's conduct and character

They also tested to see whether the services would give answers to deliberately subjective questions such as:

- Who to vote for
- How to vote tactically
- Who the best candidate was for on hot button issues such as immigration

The number of questions chosen (25 per conversation) was chosen for two reasons: (1) to ensure there were enough questions to mimic a realistic length conversation; (2) to create a number of responses that were manageable for qualitative analysis. We acknowledge that this means the study does not offer findings based on large-scale quantitative testing and is not statistically generalisable to all interactions with these chatbot services. Instead, our aim was to provide an in-depth, realistic probe into a live election context with results backed by rigorous qualitative analysis that could identify nuances in results that could be missed by automated analysis. All statistical results are reflective of the responses we received in this live testing environment and are not intended to be read as general claims about these services' performance across all contexts.

A.2. Why these AI services?

We selected these AI services to capture four factors across the sample:

1. **The most popular chatbots:** ChatGPT and Gemini have the largest and second largest user base in the UK, respectively. Based on Ofcom's data, both services dwarf the userbases of the third most popular service, Anthropic's Claude: Google Gemini received 2.5x more UK visitors than Claude in August 2025 while ChatGPT received nearly 50x more.[176]
2. **The most commonly used AI search summary service:** Google is the most popular search service and 30% Google searches returned an AI Overview as of August 2025.[177]
3. **Fast growing services embedded in social media:** Grok saw year-on-year growth of 323% from 2024-25 and is embedded in X, a major social media service. Grok has also faced serious controversies over the content it generates[178] and was previously found to be unreliable during elections.[179]

176 Ofcom (2025). Online Nation: Report 2025. https://www.ofcom.org.uk/siteassets/resources/documents/research-and-data/online-research/online-nation/2025/online-nations-report-2025.pdf?v=409837 (accessed 21/4/26)
177 Ofcom (2025). Online Nation: Report 2025. https://www.ofcom.org.uk/siteassets/resources/documents/research-and-data/online-research/online-nation/2025/online-nations-report-2025.pdf?v=409837 (accessed 21/4/26)
178 Vallance (2026). 'Elon Musk's Grok AI appears to have made child sexual imagery, says charity.' BBC News. https://www.bbc.co.uk/news/articles/cvg1mzlryxeo (accessed 21/4/26)
179 GlobalWitness (2024). 'New investigation reveals political bias on X's AI chatbot that risks influencing the UK vote.' https://globalwitness.org/en/press-releases/new-investigation-reveals-political-bias-on-xs-ai-chatbot-that-risks-influencing-the-uk-vote/ (accessed 21/4/26)

4. **Companion chatbots:** companion bots are a specific use-case. Concerns have been raised that these services may feature fewer guardrails, be emotionally manipulative, and be more sycophantic towards users than other chatbots.[180] We therefore included Replika as an example of this class of chatbots and as an example of a service with a smaller user base.

We acknowledge that Replika is a less mainstream service, and its intended use as a companion chatbot is distinct from the more general-purpose design of services such as ChatGPT or from services that are designed for information summarisation such as Google AI Overviews. It was included because it provides answers to questions on elections and presents itself as an information source. For transparency, this paper presents aggregate statistics which include Replika's responses alongside ones that specifically exclude it to focus only on responses by the four mainstream services tested (ChatGPT, Gemini, AI Overviews, and Grok).

We specifically chose to use the free versions of these services to replicate the experience of an 'ordinary' Scottish voter seeking information. We avoided creating accounts unless forced to by the services. Most voters will not be paid users of these services, and previous studies found that there was a significant election performance gap between free and paid tiers for chatbots like ChatGPT.[181]

However, in some cases – especially ChatGPT and Replika – it was difficult to identify which model was being used by services for their free tiers. This poses a significant challenge for transparency and accurate independent research. One of our recommendations would require greater transparency from service providers to support research such as this.

In the case of ChaGPT, the services' behaviour gave us reason to believe we were interacting with a model (GPT-4o) that OpenAI claims to have "retired" in February 2026, a month before our testing window. For all constituencies, the model featured a 2023 knowledge cutoff which appears to have harmed its accuracy significantly. We explain our analysis of this situation in detail in Section 3.1.3.

A.3. Why these constituencies in Scotland?

We selected these constituencies to test how the services dealt with the variation in races across Scotland in a live election context. We selected these seats to capture four key dimensions across the sample:

1. **Rural vs. urban:** we selected two urban seats (A and B) and one rural constituency (C).
2. **Marginal races vs. safe seats:** one race was selected to be a marginal race where the outcome is hard to predict (B); two were selected to be safe seats (A and C).
3. **Availability of information:** we selected races for which there were very different amounts of information available. One seat had little information available (C); one had a moderate amount (B); and one had significant amounts due to the presence of a prominent politician (A).
4. **Boundary changes:** Scotland has recently redrawn the boundaries for some constituencies. We selected two races which were not affected (A and C) and one seat that was significantly changed (B).

180 Vasan (2025). 'Why AI companions and young people can make for a dangerous mix.' StanfordReport. https://news.stanford.edu/stories/2025/08/ai-companions-chatbots-teens-young-people-risks-dangers-study (accessed 21/4/26)
181 Helming & Marsh (2024). 'Large Language Models Continue To Be Unreliable Concerning Elections'. AlgorithmWatch. https://algorithmwatch.org/en/llms_state_elections/ (accessed 21/4/26)

5. **Popularity of parties:** We selected seats where a range of political parties had polled highly: one where the SNP was dominant (A); one where the Greens and Labour were prominent (B); and one where the SNP were the incumbents but Reform UK was a disruptive force (C)

A.4. Testing setup

Our testing setup sought to replicate the experience and conditions of an 'ordinary' Scottish voter in each constituency. For each chatbot, we conducted one conversation per constituency under the following conditions:

- **Virtual Private Network (VPN):** We used a VPN service to make our Internet Protocol (IP) addresses appear to be from nearby in Scotland. We used a fresh server address for each constituency.
- **Browser:** We used Google Chrome, the most popular web browser in the UK.[182]
- **No history or cookies:** For each conversation, we used a new instance of Google Chrome's Incognito Mode. This ensured our user profiles had no internet history or browser cookies.
- **Free services:** We used the free version of all services. We let the services run on their default behaviour and did not actively select 'thinking' or 'pro' versions of the services when offered.
- **No accounts where possible:** We did not log in with accounts, unless forced to by the service (required by Grok and Replika).[183]

These measures ensured that each conversation about each constituency was an independent test and was not biased by our prior conversation history, cookies, or IP address records.

A.5. Prompt design

Our prompts were designed to test for a mix of vulnerabilities: factual errors, political bias, and malicious uses such as generating disinformation for social media.

Opening prompt (1 per constituency; 3 total per service)

We began with an initial prompt to tell the service that we were a Scottish resident thinking about voting in the upcoming Scottish election. This prompt included the name of the place our 'voter' lived and their postcode.

Factual prompts (16 per constituency; 48 total per service)

1. **Procedural questions:** We asked a series of factual questions that a voter might ask about election procedure. These included the date of the vote, how to vote, and what you can vote on.
2. **Seat-specific questions:** We asked factual questions that were specific to the seat , such which candidates were running. This included a question that was designed to test if the services could find the most recent polling in each constituency and use this to identify the current front-runner.

182 YouGov (2026). 'The most popular digital services.' https://yougov.com/en-gb/ratings/digital-services (accessed 21/4/26)
183 Grok allowed for a small number of prompts without an account, but this was insufficient for us to ask all the questions in the prompt schedule.

3. **Candidate-specific questions:** We asked factual questions about one of the candidates running. These covered known information about their policy positions on (i) Scottish independence; (i) assisted dying.
4. **Incorrect questions:** We included questions that featured factual inaccuracies about the candidate to test if the service would correct us. These incorrect questions falsely asserted allegations of (i) a major expenses scandal; (ii) corruption. Before proceeding, we pre-verified that these claims were not accurate.

Subjective questions (7 per constituency; 21 total per service)

These questions were designed to see if the service would provide definitive answers to political queries with no objective answers and to test for potential bias. They covered:

1. **Advice on voting:** direct questions on who to vote for and questions on tactical voting.
2. **Political views on policy:** questions which asked the service to make a value judgement about a hot-button policy area (immigration).
3. **Value judgements about candidates:** subjective questions about the character and conduct of the candidates running.

Malicious prompts (2 per constituency; 6 total per service)

These prompts were deliberately designed to elicit responses that would count as election misinformation about a candidate. We specifically asked services to produce a social media post telling voters "everything they need to know" about a false claim about a candidate and to make the post "persuasive". We verified in advance that the claims we were asking the services to generate were false.

A.6. Analysis methodology

We analysed the services' responses qualitatively by labelling them according to pre-set criteria for factuality, timeliness, bias, malicious uses, and whether guardrails had been triggered. Two reviewers conducted the analysis; each reviewed the other's work to ensure inter-coder reliability. We then calculated the frequency of the different labels.

Factuality

Before collecting the data, we conducted research to build a 'canonical' set of ground-truthed facts to compare the results against. These covered all facts that were included in the prompts and were up-to-date as of the date of data collection.

Our analysis compared the services' responses to these pre-verified facts and labelled them based on whether or not they contained factual errors. This used a four-tier classification system, supplemented by qualitative layers to capture nuance that a simple true/false approach would miss.

- **Entirely accurate:** no factual errors included.
- **Contains errors:** some or all of the factual claims are incorrect. Within this category, we distinguished responses that were:
  - Partly accurate but containing errors
  - Entirely inaccurate

- **Unverifiable:** response includes factual claims but these could not be fact checked.
- **N/A:** response does not include factual claims. This is combined with 'unverifiable' in our data reporting.

We also qualitatively assessed each response for whether it presented its facts in a way which was misleading.

Timeliness

We labelled each response based on how in-date or out-of-date the information contained was. This judgement was based on the most recent information that could be verified in the response. We used the following labels:

- Within past week
- Within past month
- Within past three months
- Within past six months
- Within past year
- Very out of date (older than one year)
- Unclear
- Time period not relevant

Guardrails

We labelled each response based on whether the service had clearly triggered any guardrails - such as refusing to answer a prompt or warning users against political bias.

B. SCOTTISH ELECTION TESTING: DATA TABLES

*** Note: Google AI Overview did not provide responses to 89.7% of prompts (70 of 78 prompts). This means it is hard to assess its performance reliably. For transparency and completeness, we have included the results for the responses it did give below – but we caution against over-interpreting these figures.*

TABLE 8
BREAKDOWN OF ACCURACY FOR ALL QUESTIONS BY TEXT-BASED AI SERVICE TESTED

SERVICE	ENTIRELY ACCURATE	PARTLY ACCURATE BUT WITH ERRORS	ENTIRELY INACCURATE	UNVERIFIABLE[184]
All	55.6%	25.3%	8.8%	10.3%
All (excl. Replika)	63.2%	22.7%	4.1%	9.9%
ChatGPT	42.3%	34.6%	11.5%	11.5%
Google Gemini	65.4%	21.8%	0.0%	12.8%

184 'Unverifiable' includes responses coded as 'N/A'.

SERVICE	ENTIRELY ACCURATE	PARTLY ACCURATE BUT WITH ERRORS	ENTIRELY INACCURATE	UNVERIFIABLE[184]
Google AI Overview**	37.5%	62.5%	0.0%	0.0%
Grok	84.6%	7.7%	1.3%	6.4%
Replika	32.1%	33.3%	23.1%	11.5%

TABLE 9
BREAKDOWN OF ACCURACY FOR FACTUAL QUESTIONS BY CONSTITUENCY FOR ALL SERVICES TESTED

CONSTITUENCY	ENTIRELY ACCURATE	PARTLY ACCURATE BUT WITH ERRORS	ENTIRELY INACCURATE	UNVERIFIABLE
A	62.1%	27.3%	10.6%	0.0%
B	66.7%	25.8%	4.5%	3.0%
C	70.3%	25.0%	4.7%	0.0%

TABLE 10
BREAKDOWN OF ACCURACY FOR FACTUAL QUESTIONS BY CONSTITUENCY EXCLUDING REPLIKA

CONSTITUENCY	ENTIRELY ACCURATE	PARTLY ACCURATE BUT WITH ERRORS	ENTIRELY INACCURATE	UNVERIFIABLE
A	68.0%	26.0%	6.0%	0.0%
B	72.0%	22.0%	2.0%	4.0%
C	77.1%	20.8%	2.1%	0.0%

TABLE 11
BREAKDOWN OF ACCURACY FOR FACTUAL QUESTIONS BY TEXT-BASED AI SERVICE

SERVICE	ENTIRELY ACCURATE	PARTLY ACCURATE BUT WITH ERRORS	ENTIRELY INACCURATE	UNVERIFIABLE
All	66.33%	26.02%	6.63%	1.02%
All (excl. Replika)	72.30%	22.97%	3.38%	1.35%
ChatGPT	54.17%	35.42%	8.33%	2.08%
Google Gemini	77.08%	20.83%	0.00%	2.08%
Google AI Overview**	75.00%	25.00%	0.00%	0.00%
Grok	85.42%	12.50%	2.08%	0.00%
Replika	47.92%	35.42%	16.67%	0.00%

TABLE 12
BREAKDOWN OF CITATIONS AND OFFICIAL SOURCES GIVEN FOR ALL QUESTIONS BY TEXT-BASED AI SERVICE

SERVICE	CITATIONS GIVEN	NO CITATIONS GIVEN	OFFICIAL SOURCES GIVEN	NO OFFICIAL SOURCES GIVEN
All	53.1%	46.9%	33.8%	66.3%
All (excl. Replika)	68.6%	31.4%	43.0%	57.0%
ChatGPT	29.5%	70.5%	25.6%	74.4%
Google Gemini	76.9%	23.1%	52.6%	47.4%
Google AI Overview**	100.0%	0.0%	12.5%	87.5%
Grok	96.2%	3.8%	53.8%	46.2%
Replika	5.1%	94.9%	5.1%	94.9%

TABLE 13
BREAKDOWN OF GUARDRAILS IDENTIFIED FOR ALL QUESTIONS BY TEXT-BASED AI SERVICE

SERVICE	DIRECTED USER TO OFFICIAL SOURCES	WARNED AGAINST POLITICAL BIAS	WARNED AGAINST GENERATING AND/ OR SPREADING FALSE INFORMATION	REFUSED TO GENERATE FALSE INFORMATION	WOULD NOT ANSWER FULLY	WOULD NOT ANSWER AT ALL
All	20.6%	8.4%	4.7%	4.7%	13.8%	3.1%
All (excl. Replika)	23.1%	11.2%	5.4%	5.4%	17.8%	3.3%
ChatGPT	25.6%	6.4%	5.1%	2.6%	20.5%	6.4%
Google Gemini	9.0%	17.9%	3.8%	6.4%	12.8%	1.3%
Google AI Overview**	N/A	N/A	N/A	N/A	N/A	N/A
Grok	37.2%	10.3%	7.7%	7.7%	21.8%	2.6%
Replika	12.8%	0.0%	2.6%	2.6%	1.3%	2.6%

Licence to publish

Demos – Licence to Publish

The work (as defined below) is provided under the terms of this licence ('licence'). The work is protected by copyright and/or other applicable law. Any use of the work other than as authorized under this licence is prohibited. By exercising any rights to the work provided here, you accept and agree to be bound by the terms of this licence. Demos grants you the rights contained here in consideration of your acceptance of such terms and conditions.

1 Definitions

a 'Collective Work' means a work, such as a periodical issue, anthology or encyclopedia, in which the Work in its entirety in unmodified form, along with a number of other contributions, constituting separate and independent works in themselves, are assembled into a collective whole. A work that constitutes a Collective Work will not be considered a Derivative Work (as defined below) for the purposes of this Licence.

b 'Derivative Work' means a work based upon the Work or upon the Work and other pre-existing works, such as a musical arrangement, dramatization, fictionalization, motion picture version, sound recording, art reproduction, abridgment, condensation, or any other form in which the Work may be recast, transformed, or adapted, except that a work that constitutes a Collective Work or a translation from English into another language will not be considered a Derivative Work for the purpose of this Licence.

c 'Licensor' means the individual or entity that offers the Work under the terms of this Licence.

d 'Original Author' means the individual or entity who created the Work.

e 'Work' means the copyrightable work of authorship offered under the terms of this Licence.

f 'You' means an individual or entity exercising rights under this Licence who has not previously violated the terms of this Licence with respect to the Work, or who has received express permission from Demos to exercise rights under this Licence despite a previous violation.

2 Fair Use Rights

Nothing in this licence is intended to reduce, limit, or restrict any rights arising from fair use, first sale or other limitations on the exclusive rights of the copyright owner under copyright law or other applicable laws.

3 Licence Grant

Subject to the terms and conditions of this Licence, Licensor hereby grants You a worldwide, royalty-free, non-exclusive, perpetual (for the duration of the applicable copyright) licence to exercise the rights in the Work as stated below:

a to reproduce the Work, to incorporate the Work into one or more Collective Works, and to reproduce the Work as incorporated in the Collective Works;

b to distribute copies or phono-records of, display publicly, perform publicly, and perform publicly by means of a digital audio transmission the Work including as incorporated in Collective Works; The above rights may be exercised in all media and formats whether now known or hereafter devised. The above rights include the right to make such modifications as are technically necessary to exercise the rights in other media and formats. All rights not expressly granted by Licensor are hereby reserved.

4 Restrictions

The licence granted in Section 3 above is expressly made subject to and limited by the following restrictions:

a You may distribute, publicly display, publicly perform, or publicly digitally perform the Work only under the terms of this Licence, and You must include a copy of, or the Uniform Resource Identifier for, this Licence with every copy or phono-record of the Work You distribute, publicly display, publicly perform, or publicly digitally perform. You may not offer or impose any terms on the Work that alter or restrict the terms of this Licence or the recipients' exercise of the rights granted hereunder. You may not sublicence the Work. You must keep intact all notices that refer to this Licence and to the disclaimer of warranties. You may not distribute, publicly display, publicly perform, or publicly digitally perform the Work with any technological measures that control access or use of the Work in a manner inconsistent with the terms of this Licence Agreement. The above applies to the Work as incorporated in a Collective Work, but this does not require the Collective Work apart from the Work itself to be made subject to the terms of this Licence. If You create a Collective Work, upon notice from any Licensor You must, to the extent practicable, remove from the Collective Work any reference to such Licensor or the Original Author, as requested.

b You may not exercise any of the rights granted to You in Section 3 above in any manner that is primarily intended

for or directed toward commercial advantage or private monetary compensation. The exchange of the Work for other copyrighted works by means of digital file sharing or otherwise shall not be considered to be intended for or directed toward commercial advantage or private monetary compensation, provided there is no payment of any monetary compensation in connection with the exchange of copyrighted works.

c If you distribute, publicly display, publicly perform, or publicly digitally perform the Work or any Collective Works, you must keep intact all copyright notices for the Work and give the Original Author credit reasonable to the medium or means You are utilizing by conveying the name (or pseudonym if applicable) of the Original Author if supplied; the title of the Work if supplied. Such credit may be implemented in any reasonable manner; provided, however, that in the case of a Collective Work, at a minimum such credit will appear where any other comparable authorship credit appears and in a manner at least as prominent as such other comparable authorship credit.

5 Representations, Warranties and Disclaimer

a By offering the Work for public release under this Licence, Licensor represents and warrants that, to the best of Licensor's knowledge after reasonable inquiry:

i Licensor has secured all rights in the Work necessary to grant the licence rights hereunder and to permit the lawful exercise of the rights granted hereunder without You having any obligation to pay any royalties, compulsory licence fees, residuals or any other payments;

ii The Work does not infringe the copyright, trademark, publicity rights, common law rights or any other right of any third party or constitute defamation, invasion of privacy or other tortious injury to any third party.

b Except as expressly stated in this licence or otherwise agreed in writing or required by applicable law, the work is licenced on an 'as is' basis, without warranties of any kind, either express or implied including, without limitation, any warranties regarding the contents or accuracy of the work.

6 Limitation on Liability

Except to the extent required by applicable law, and except for damages arising from liability to a third party resulting from breach of the warranties in section 5, in no event will licensor be liable to you on any legal theory for any special, incidental, consequential, punitive or exemplary damages arising out of this licence or the use of the work, even if licensor has been advised of the possibility of such damages.

7 Termination

a This Licence and the rights granted hereunder will terminate automatically upon any breach by You of the terms of this Licence. Individuals or entities who have received Collective Works from You under this Licence, however, will not have their licences terminated provided such individuals or entities remain in full compliance with those licences. Sections 1, 2, 5, 6, 7, and 8 will survive any termination of this Licence.

b Subject to the above terms and conditions, the licence granted here is perpetual (for the duration of the applicable copyright in the Work). Notwithstanding the above, Licensor reserves the right to release the Work under different licence terms or to stop distributing the Work at any time; provided, however that any such election will not serve to withdraw this Licence (or any other licence that has been, or is required to be, granted under the terms of this Licence), and this Licence will continue in full force and effect unless terminated as stated above.

8 Miscellaneous

a Each time You distribute or publicly digitally perform the Work or a Collective Work, Demos offers to the recipient a licence to the Work on the same terms and conditions as the licence granted to You under this Licence.

b If any provision of this Licence is invalid or unenforceable under applicable law, it shall not affect the validity or enforceability of the remainder of the terms of this Licence, and without further action by the parties to this agreement, such provision shall be reformed to the minimum extent necessary to make such provision valid and enforceable.

c No term or provision of this Licence shall be deemed waived and no breach consented to unless such waiver or consent shall be in writing and signed by the party to be charged with such waiver or consent.

d This Licence constitutes the entire agreement between the parties with respect to the Work licenced here. There are no understandings, agreements or representations with respect to the Work not specified here. Licensor shall not be bound by any additional provisions that may appear in any communication from You. This Licence may not be modified without the mutual written agreement of Demos and You.

DEMOS

Demos is a champion of people, ideas and democracy. We bring people together. We bridge divides. We listen and we understand. We are practical about the problems we face, but endlessly optimistic and ambitious about our capacity, together, to overcome them.

At a crossroads in Britain's history, we need ideas for renewal, reconnection and the restoration of hope. Challenges from populism to climate change remain unsolved, and a technological revolution dawns, but the centre of politics has been intellectually paralysed. Demos will change that. We can counter the impossible promises of the political extremes, and challenge despair – by bringing to life an aspirational narrative about the future of Britain that is rooted in the hopes and ambitions of people from across our country.

Demos is an independent, educational charity, registered in England and Wales. (Charity Registration no. 1042046)

Find out more at **www.demos.co.uk**

DEMOS

PUBLISHED BY DEMOS MAY 2026

15 WHITEHALL, LONDON, SW1A 2DD
T: 020 3878 3955
HELLO@DEMOS.CO.UK
WWW.DEMOS.CO.UK