Harvard Kennedy School Misinformation Review[1]
August 2026, Volume 7, Issue 5

DOI: https://doi.org/10.37016/mr-2020-210
Website: https://misinforeview.hks.harvard.edu/

Research Note

Monitoring misinformation, building discernment: A crowdsourcing study during the 2025 Australian federal election

We conducted a design evaluation exploring whether a structured crowdsourcing system using trained volunteers could monitor and categorize election-related misinformation during the 2025 Australian federal election campaign. Over six weeks, 90 active participants surfaced 592 potentially misinforming posts and provided over 3,000 assessments across multiple platforms and recurring narratives. Participants classified content with meaningful accuracy against an independent assessor, with inter-rater agreement varying by category. A follow-up three months later found that participants showed higher misinformation discernment than a separately recruited control group. Structured crowdsourcing may, therefore, serve as a scalable monitoring mechanism and, more tentatively, a potential pathway to building resilience to misinformation.

Authors: Ariel Kruger (1), Piers Howe, (2) Morgan Saletta (1)
Affiliations: (1) Hunt Laboratory for Intelligence and Security Studies, University of Melbourne, Australia, (2) Melbourne School of Psychological Sciences, University of Melbourne, Australia
How to cite: Kruger, A., Howe, P., & Saletta, M. (2026). Monitoring misinformation, building discernment: A crowdsourcing study during the 2025 Australian federal election. *Harvard Kennedy School (HKS) Misinformation Review*, *7*(5).
Received: March 23rd, 2026. Accepted: August 10th, 2026. Published: August 27th, 2026.

Research questions

- Can a trained volunteer crowd deployed during a live election identify and categorize election-related misinformation across major social media platforms?
- What forms of misinformation and themes did participants identify during the 2025 Australian federal election campaign, and how did these forms change over time?
- How consistently did participants classify the same content, and how well did their judgments match those of an independent assessor?
- Was participation associated with differences in participants' ability to distinguish true from false content several months later?

Research note summary

- This study deployed a structured crowdsourcing system in which trained volunteers monitored

[1] *A publication of the Shorenstein Center on Media, Politics and Public Policy at Harvard University, John F. Kennedy School of Government.*

their own social media feeds during the 2025 Australian federal election, submitting and peer-assessing potentially misleading posts using a shared classification framework.

- Volunteers surfaced hundreds of instances of election-related misinformation across major platforms. False content, unproven or conspiratorial claims, and out-of-context material dominated submissions. Narratives around electoral integrity grew in prominence towards polling day, while social justice, immigration, and foreign threat narratives persisted throughout.
- Participants showed meaningful convergence in applying the classification framework and achieved broad agreement with an independent assessor on a sample of submitted posts.
- Three months after the campaign, participants distinguished true from false simulated posts more accurately than a separately recruited control group. This difference persisted after adjusting for demographic differences between the groups.
- The method offers a non-partisan model that may support information literacy: by having participants log, categorize, and rate each other's contributions, it fosters structured engagement with misinformation without prescribing truth judgements.

Implications

Implication 1: Crowdsourcing as a scalable citizen-science model for monitoring misinformation

The 2025 Australian federal election, held in May under compulsory voting, returned the center-left Labor government amid widespread concern about election-related misinformation circulating on social media, a concern now common to elections worldwide. One increasingly discussed response is crowdsourcing: distributed users flagging and assessing questionable content, as in X's Community Notes. Such approaches have emerged as a potential complement to professional fact-checking, which is rigorous but difficult to scale during fast-moving events.

Monitoring misinformation during time-sensitive events requires identifying and classifying large volumes of content at speed. Automated systems currently struggle with contextual, culturally dependent, and multimodal content, particularly where evaluative judgments go beyond simple true/false distinctions (Feng et al., 2025; Kertysova, 2018; Santos, 2023). Professional fact-checking offers careful validation, but it is inherently labor-intensive and difficult to deploy at the speed and scale that live events demand (Zhao & Naaman, 2023).

The primary finding of our design evaluation study is that structured crowdsourcing can provide a practical mechanism for expanding monitoring capacity by distributing elements of this work across a large number of participants. In the system developed for this study, 90 active volunteers identified and submitted 592 potentially misleading posts encountered on their social media feeds and provided over 3,000 assessments using a shared classification framework. Because submissions and assessments were recorded through a structured platform, these distributed judgments were aggregated into a coherent dataset capturing misinformation types, confidence ratings and justifications, enabling analysis of narrative themes and targets, and capturing content directly from participants' own feeds without reliance on restrictive platform API's or scraping (Davidson et al., 2023).

This approach resembles citizen-science models used in fields such as environmental monitoring and astronomy, where distributed volunteers contribute observations that are aggregated into large-scale datasets (Bonney et al., 2009). The underlying model is similarly scalable: similar systems could be expanded to larger participant networks and applied to other domains, such as public health, environmental policy, or international crises, as demonstrated by citizen science applications to real-time weather hazard monitoring (Tan et al., 2022).

The model is also extensible across communities, not only domains. Misinformation monitoring is disproportionately oriented toward the Global North and English-language content, leaving the information environments of some communities under-observed (Mahl et al., 2026). The defining feature of our approach (that participants surface the misinformation they personally encounter) speaks directly to this gap. Where participants are recruited from a specific community and materials translated accordingly, the system could be deployed to surface and categorize the misinformation circulating within and targeting that community: content that centralized, English-language monitoring is poorly positioned to detect. In Australia, this gap is concrete: electoral misinformation targeting migrant and non-English-speaking communities—much of it on platforms such as WeChat and RedNote that sit outside mainstream monitoring and prohibit automated tools—is a recognized and under-addressed concern (Yang et al., 2025). Because these environments resist automated detection, community-embedded human monitoring is one of the few viable means of surfacing what circulates within them. A deployment drawing participants from an affected community could therefore build a clearer picture of the misinformation it actually encounters; information that bears directly on equal and informed democratic participation, yet rarely reaches centralized monitoring. We did not test a community-specific deployment; this is an application the design enables rather than a demonstrated outcome.

Implication 2: Crowdsourcing participation as a potential discernment intervention

Building resilience to misinformation is a challenge, and experts agree that interventions to improve the public's ability to identify misinformation should be prioritized (Kruger et al., 2024). Yet, such interventions often attenuate over time without reinforcement. Against this backdrop, we observed that participation in our study was associated with differences in participants' ability to distinguish true from false simulated social media content three months after the monitoring period concluded (compared with a control group). However, because participants were not randomly assigned, these differences may reflect pre-existing characteristics, such as political interest, motivation or baseline media-literacy, rather than the effects of participation itself.

If participation did contribute to this difference, there is a plausible mechanism. Sustained, repeated engagement with fact-checking has been shown to durably improve discernment of misinformation beyond the specific claims encountered, effectively functioning as a form of media literacy by familiarizing people with how misinformation works and how it can be evaluated (Bowles et al., 2025). Our system extends this logic from repeated exposure to repeated practice: rather than passively consuming corrections, participants actively classified real-world content, articulated their reasoning, and rated their confidence using shared analytical criteria. This active, effortful evaluation is consistent with research on psychological inoculation, which emphasizes the role of practice and active resistance in building durable resistance to misleading information (Roozenbeek & van der Linden, 2019).

Findings

Finding 1: A distributed crowd provided sustained monitoring of misinformation during the Australian 2025 federal election campaign.

Over the six-week monitoring period, 90 active participants monitored their social media environment, submitting 592 potentially misinforming posts and over 3,000 assessments. As is typical of crowdsourced projects (Sauermann & Franzoni, 2015; Swain et al., 2015), contributions were unevenly distributed: approximately 10% of participants accounted for 60% of submissions and 70% of assessments. Participants reported 1,228 hours of engagement, with an estimated 614 hours spent actively monitoring.

Thematic analysis showed that a small number of narratives accounted for a disproportionate share of submissions (Figure 1), while temporal analysis revealed that these narratives shifted over the course of the campaign (Figure 2). In particular, electoral integrity claims were present throughout but increased sharply in the final weeks, becoming the most frequently submitted theme around polling day. Each submission was also linked to a target where one was identifiable. Notably, the incumbent center-left Labor government was the most frequent target (53% of submissions with an identifiable target; see Appendix C).

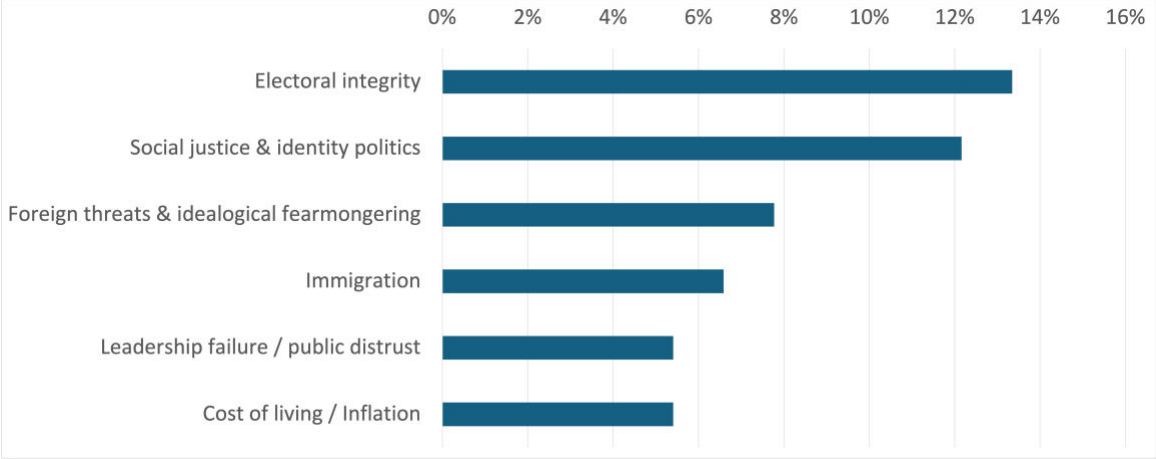

Figure 1. Top six misinformation themes identified during the 2025 federal election monitoring period. *Top themes accounted for 50.7% of all submissions.*

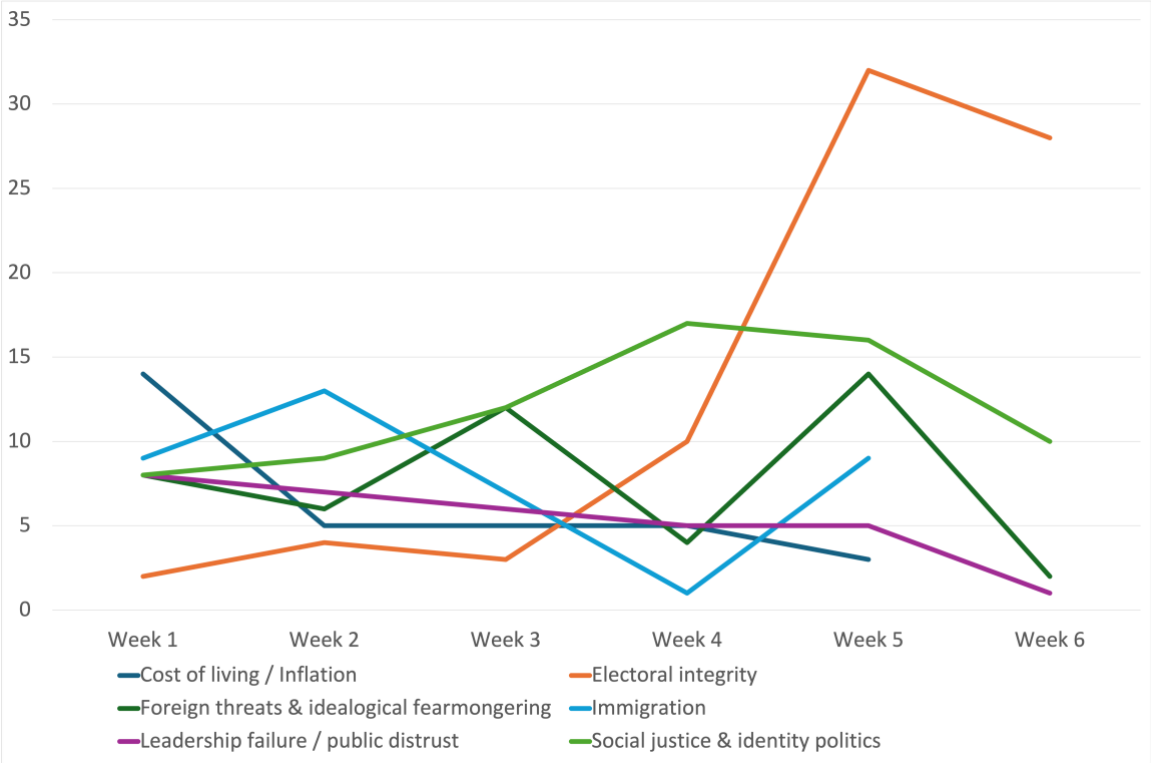

Figure 2. Weekly frequency of the six most common misinformation themes identified during the monitoring period.

Finding 2: The volunteer cohort provided broad cross-platform coverage that aligned with narratives independently documented by a major fact-checking organization.

The misinformation surfaced by participants was heavily concentrated on a small number of major social media platforms. The four platforms in Figure 3 accounted for over 90% of all submissions, with the remainder spread across YouTube, BlueSky, Xiaohongshu, Reddit, and Threads. However, because submissions were drawn from participants' own social media environments and the participant pool was ideologically skewed, the distribution likely reflects a partial rather than comprehensive view of the ecosystem. It also excludes misinformation in broadcast or print, except where it was later shared on social media.

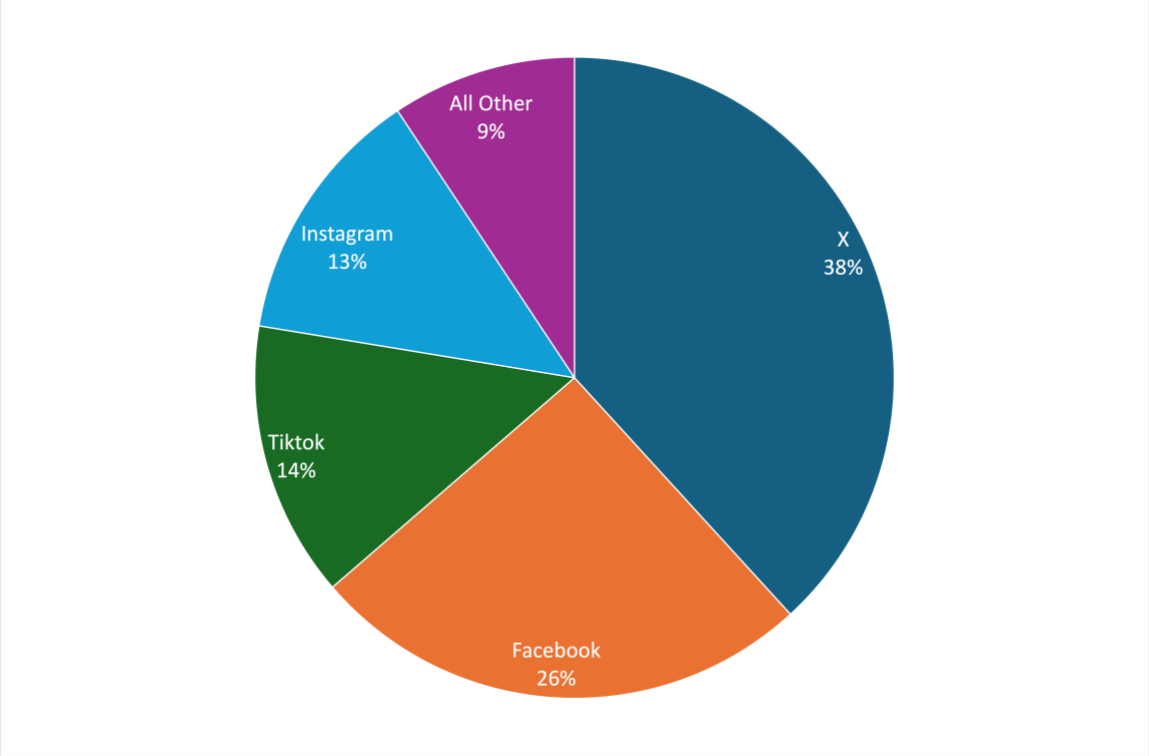

Figure 3. Concentration of misinformation detections across social media platforms.

To assess whether our crowd-based method achieved coverage comparable to established fact-checking practice, we compared participant submissions with Australian Associated Press (AAP) fact checks published during the election period (see Appendix E for more detail). Analysis of 34 AAP fact-check articles showed a similar distribution of themes, with electoral integrity, government spending, and cost-of-living claims most prevalent. These correspond closely with the most common themes identified by participants (see Finding 1). Notably, every major election-related narrative documented by AAP was also present in the participant dataset; the crowd surfaced the same narratives that a professional fact-checking organization deemed significant enough to verify. This suggests the method can match an established fact-checking process for the narratives that process judged significant, even if it does not capture the full ecosystem.

Finding 3: Participants applied a structured classification framework that differentiated distinct forms of misinformation.

In addition to surfacing misinformation across platforms and narratives, participants classified each submission using a shared framework distinguishing among 17 categories, developed for usability by non—expert participants and refined against real posts during development (see Appendix A for more detail). As shown in Figure 4, the most common forms were false content, unproven or conspiratorial claims, and out-of-context material, which together accounted for nearly 60% of all submissions. Other forms (including rage-bait, cherry-picking, political defamation, altered media, and impersonation) appeared less frequently but were consistently identified across the monitoring period.

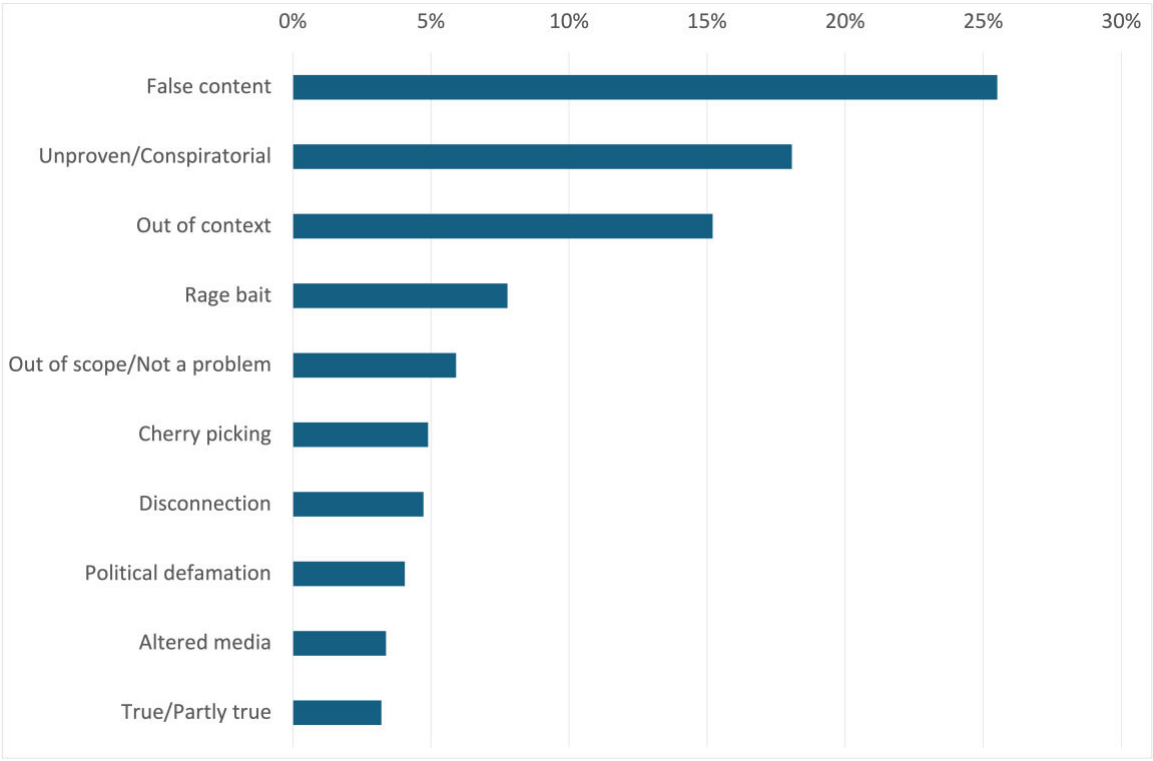

Figure 4. Distribution of primary misinformation forms assigned by participants. *False content, unproven claims, and out-of-context material accounted for the majority of submissions.*

Importantly, participants did not indiscriminately label all flagged content as misinformation. Approximately 5.9% of submissions were classified as out-of-scope or not problematic, and 3.2% were coded as true or partly true. This indicates that participants were exercising evaluative judgment, discerning what did *not* warrant a misinformation label.

Finding 4: Crowd classifications of verifiably false content approached fact-checking accuracy, while inter-rater agreement varied across categories.

To assess accuracy, we compared participant judgments against those of an independent assessor (a postgraduate researcher who documented the basis for each determination). Consistent with established approaches to evaluating crowd-based misinformation detection (e.g., Barbera et al., 2024), this check focused on the verifiable true/false dimension rather than the framework's more interpretive categories. Posts the crowd classified as false content (n = 24) or true (n = 6) were checked against the assessor's determination, since these rest on factual claims with a ground truth.

The assessor agreed with the crowd's classification in 81% of cases. Cohen's kappa was 0.42, indicating moderate agreement beyond chance (Landis & Koch, 1977). While the sample of posts is small, this level

of alignment is broadly consistent with recent findings that crowds can approach professional fact-checking accuracy under structured conditions (Barbera et al., 2024).

Among posts rated by multiple participants, mean agreement with the modal category was 64.7%, though because participants could see others' classifications before submitting their own, this may overstate independent agreement.

For posts yielding a unique majority classification (N = 387), agreement was highest for categories with clear structural features (e.g., political scams, unproven or conspiratorial claims, false content) and lower for interpretive ones such as rage-bait and cherry-picking (see Figure 5); these figures reflect agreement among raters, not accuracy against an external standard.

Confidence ratings of assessments (on a scale 0 – 10) followed a similar pattern. Confidence varied significantly across classification categories, $\chi^2(16) = 144.44$, $p < .001$. Clearer-cut categories such as political scams and false content tended to be rated more confidently than interpretive ones (see Appendix A).

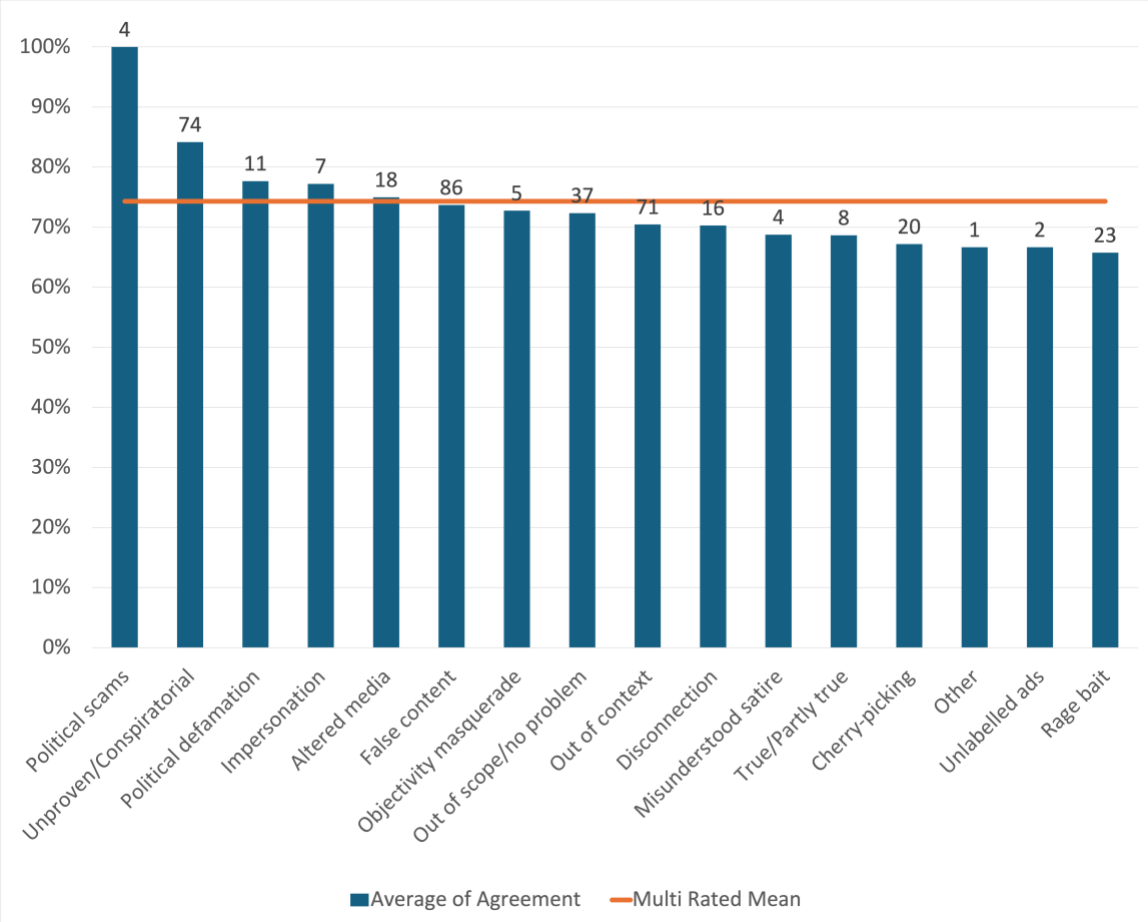

Figure 5. Majority agreement by category. *Agreement is measured as the proportion of raters selecting the modal category for each post. The x-axis displays misinformation categories, and the y-axis shows mean agreement (%). Values above bars indicate the number of posts, and the horizontal line shows the overall mean agreement.*

Finding 5: Participation in structured monitoring was associated with higher misinformation discernment.

To examine whether participation in structured monitoring was associated with improved evaluative judgment, we conducted a follow-up assessment approximately three months after the monitoring period concluded (see Appendix F). Participants drawn from the active monitoring cohort (treatment group; *n* =

56) were compared with a newly recruited control group (*n* = 132) using a 20-item test comprising simulated social media posts with established ground truth. Participants made binary true/false judgments for each item.

The treatment group achieved a mean accuracy of 73.8% (*SD* = 0.137), compared with 67.5% (*SD* = 0.128) in the control group. This difference corresponds to a moderate effect size (Cohen's *d* = 0.48, 95% CI [0.16, 0.79]). The groups differed on some demographic measures, but the difference in discernment persisted after adjusting for those on which they differed (age, employment, and education) in a trial-level model (see Appendix F). Because participants were not randomly assigned, however, the difference may still reflect unmeasured pre-existing characteristics. The pattern is best read as associational, consistent with the possibility that active engagement with misinformation is linked to improved discernment, though the design does not permit causal inference.

Methods

Design overview

This study was a field-based design evaluation of a structured crowdsourcing monitoring system during the 2025 Australian federal election. Rather than testing a tightly controlled intervention, the primary aim was to assess whether a trained volunteer cohort could surface and categorize election-related misinformation at scale, apply a shared framework with meaningful consistency, and generate structured outputs suitable for analysis. We also conducted an exploratory follow-up examining whether participation was associated with longer-term differences in misinformation discernment.

Participants and recruitment

Participants were recruited into two cohorts: a general public cohort (*n* = 401), recruited via targeted online advertising, and an expert cohort (*n* = 18), recruited through purposive and snowball sampling for professional or research experience with misinformation (median = ~6 years in the area), drawn from fields including psychology, public health, law, and science communication. Together, these yielded 419 individuals who completed onboarding and consented to participate. Eligibility required participants to be aged 18 or older, active social media users, with a strong interest in Australian politics.

All 419 participants were issued platform accounts. We define *active* participants as those who made at least one contribution to the database – a submission or an assessment; 90 were active in this sense, with the remainder onboarding but not contributing. The active group comprised both cohorts, 13 of the 18 recruited experts and 77 general public-contributors. The onboarding sample spanned a broad adult age range, skewed toward middle and older age groups, with comparatively high educational attainment and representation across Australian states (see Appendix B).

Volunteer-based detection raises the concern that raters may disproportionately flag content conflicting with their own views (Park et al., 2026). Although we ran targeted recruitment aimed at both left- and right-of-center participants, the active cohort skewed left-of-center, broadly reflecting Australia's electorate at the 2025 election (Australian Electoral Commission, 2025a, 2025b). Three features mitigated this: training on cognitive bias with instruction to set aside political views; required written justification for each assessment, which has been shown to aid misinformation identification regardless of ideology (Pennycook & Rand, 2019); and a peer-assessment structure that diluted individual skew across multiple judgments. These mechanisms cannot fully eliminate systematic bias, and the dataset may still reflect some ideological filtering in which posts were surfaced and how they were interpreted (see Appendix B).

Procedure

The monitoring phase ran for six weeks during the 2025 Australian federal election campaign (April 2 – May 16, 2025). Training consisted of slides, a participant "hub" where project updates and tips were posted, and a webinar on cognitive biases relevant to misinformation detection. Following onboarding and training, participants were granted access to a custom-built web platform and were instructed to monitor their own feeds as part of regular use and to submit posts they believed constituted election-related misinformation.

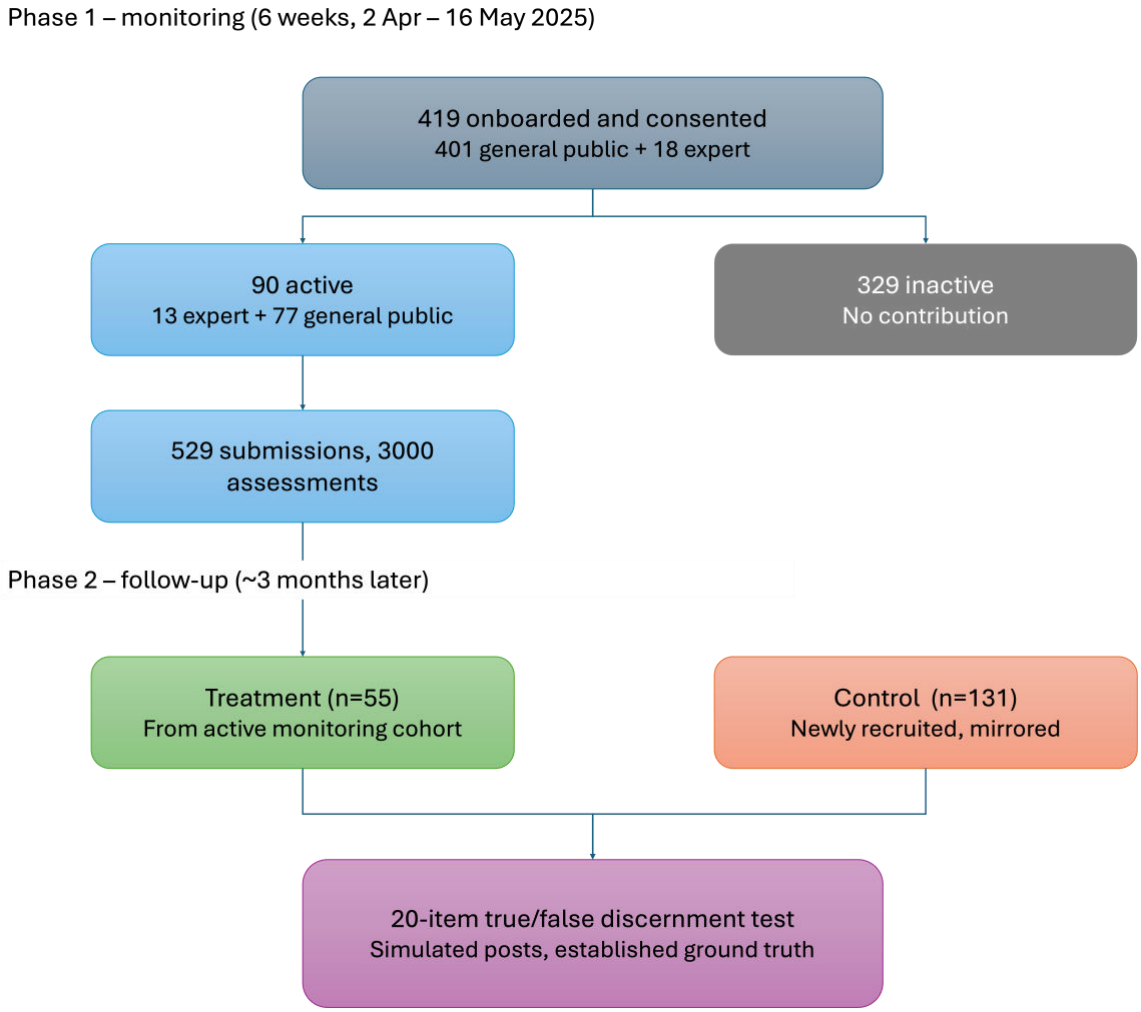

Figure 6. Two-phase study design and participant flow.

For each submission, participants provided the original URL and selected a primary misinformation category (with an optional secondary category), rated their confidence, and gave written reasoning. Screenshots were uploaded to preserve content and submissions could be edited after posting.

Submitted posts were available for peer assessment: other participants could review the content, classification and reasoning before recording their own classification, confidence rating and justification. Participants were encouraged to apply independent judgment using the shared framework. Assessments were visible within the system, and discussion was possible on a separate platform (Slack).

Approximately three months after monitoring concluded, a follow-up compared a treatment group drawn from the active monitoring cohort (n = 56) with a newly recruited control group (n = 132), recruited using mirrored eligibility criteria, on the same 20-item true/false test (see Finding 5; Appendix F).

Measures and analysis

Primary category selections (Appendix A) were used in the reliability and distributional analyses. All submissions were reviewed by both authors using a thematic coding approach; narrative themes and institutional targets were developed from the data and refined iteratively. Independent initial coding surfaced differences in interpretation, which were discussed and resolved through consensus. The coding should be understood as an iterative qualitative analysis rather than a fully independent coding procedure.

Classification consistency was measured as the proportion of raters assigning the same primary category to a post. Accuracy was assessed against an independent assessor on 30 verifiable true/false posts (see Finding 4).

In the follow-up study, discernment was operationalized as accuracy on the 20-item test. We compared treatment and control accuracy using Welch's t-test (which does not assume equal group sizes or variances) and, using a linear mixed-effects model, tested whether the difference persisted after adjusting for age, employment, and education, the variables on which the groups differed, with random intercepts for participant and item (see Appendix F). Cohen's d is reported as a descriptive effect size.

Bibliography

Australian Electoral Commission. (2025a). *2025 Federal Election*. https://results.aec.gov.au/31496/Website/HouseDefault-31496.htm

Australian Electoral Commission. (2025b). *Turnout by state*. https://results.aec.gov.au/31496/Website/HouseTurnoutByState-31496.htm

Barbera, D. L., Maddalena, E., Soprano, M., Roitero, K., Demartini, G., Ceolin, D., Spina, D., & Mizzaro, S. (2024). Crowdsourced fact-checking: Does It actually work? *Information Processing & Management*, *61*(5), Article 103792. https://doi.org/10.1016/j.ipm.2024.103792

Bonney, R., Cooper, C. B., Dickinson, J., Kelling, S., Phillips, T., Rosenberg, K. V., & Shirk, J. (2009). Citizen science: A developing tool for expanding science knowledge and scientific literacy. *BioScience*, *59*(11), 977–984. https://doi.org/10.1525/bio.2009.59.11.9

Bounegru, L., Gray, J., Venturini, T., & Mauri, M. (2018). *A field guide to "fake news" and other information disorders*. Public Data Lab. https://doi.org/10.5281/zenodo.1136272

Bowles, J., Croke, K., Larreguy, H., Liu, S., & Marshall, J. (2025). Sustaining exposure to fact-checks: Misinformation discernment, media consumption, and its political implications. *American Political Science Review*, *119*(4), 1864–1887. https://doi.org/10.1017/S0003055424001394

Davidson, B.I., Wischerath, D., Racek, D. et al. (2023). Platform-controlled social media APIs threaten open science. *Nature Human Behaviour*, *7*, 2054–2057. https://doi.org/10.1038/s41562-023-01750-2

Feng, X., Luo, J., Yang, Y., El Baz, D., & Shi, L. (2025). Health misinformation detection: Approaches, challenges and opportunities. *INQUIRY: The Journal of Health Care Organization, Provision, and Financing*, *62*. https://doi.org/10.1177/00469580251384784

Kertysova, K. (2018). Artificial intelligence and disinformation: How AI changes the way disinformation is produced, disseminated, and can be countered. *Security and Human Rights*, *29*(1–4), 55–81. https://doi.org/10.1163/18750230-02901005

Kruger, A., Saletta, M., Ahmad, A., & Howe, P. (2024). Structured expert elicitation on disinformation, misinformation, and malign influence: Barriers, strategies, and opportunities. *Harvard Kennedy School (HKS) Misinformation Review, 5*(7). https://doi.org/10.37016/mr-2020-169

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Mahl, D., Zeng, J., Schäfer, M. S., Egert, F. A., & Oliveira, T. (2026). "We follow the disinformation": Conceptualizing and analyzing fact-checking cultures across countries. *The International Journal of Press/Politics*, *31*(2), 264–290. https://doi.org/10.1177/19401612241270004

Park, S., Lee, J. Y., McGuinness, K., Fisher, C. & Fulton, J. (2026). People rely on their existing political beliefs to identify election misinformation, *7*(1). *Harvard Kennedy School (HKS) Misinformation Review*. https://doi.org/10.37016/mr-2020-195

Pennycook, G., & Rand, D. G. (2019). Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences*, *116*(7), 2521–2526. https://doi.org/10.1073/pnas.1806781116

Pew Research Center. (2026, June 10). *Political typology quiz: Which of the 9 groups are you?* https://www.pewresearch.org/politics/quiz/political-typology/

Roozenbeek, J., & van der Linden, S. (2019). The fake news game: Actively inoculating against the risk of misinformation. *Journal of Risk Research*, *22*(5), 570–580. https://doi.org/10.1080/13669877.2018.1443491

Santos, F. C. C. (2023). Artificial Intelligence in automated detection of disinformation: A thematic analysis. *Journalism and Media*, *4*(2), 679–687. https://doi.org/10.3390/journalmedia4020043

Sauermann, H., & Franzoni, C. (2015). Crowd science user contribution patterns and their implications. *Proceedings of the National Academy of Sciences*, *112*(3), 679–684. https://doi.org/10.1073/pnas.1408907112

Swain R., Berger A., Bongard J., & Hines, P. (2015). Participation and contribution in crowdsourced surveys. *PLOS ONE*, *10*(4), Article e0120521. https://doi.org/10.1371/journal.pone.0120521

Tan, M. L., Hoffmann, D., Ebert, E., Cui, A., & Johnston, D. (2022). Exploring the potential role of citizen science in the warning value chain for high impact weather. *Frontiers in Communication*, *7*, Article 949949. https://doi.org/10.3389/fcomm.2022.949949

Wardle, C., & Derakhshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policy making* (Report No. DGI(2017)09). Council of Europe. https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c

Yang, F., Heemsbergen, L., & Fordyce, R. (2025). Contextualizing critical disinformation during the 2023 Voice referendum on WeChat: Manipulating knowledge gaps and whitewashing Indigenous rights. *Harvard Kennedy School (HKS) Misinformation Review, 6*(5). https://doi.org/10.37016/mr-2020-185

Zhao, A., & Naaman, M. (2023). Insights from a comparative study on the variety, velocity, veracity, and viability of crowdsourced and professional fact-checking services. *Journal of Online Trust and Safety*, *2*(1). https://doi.org/10.54501/jots.v2i1.118

Funding
Funding for the project was provided by the Susan McKinnon Foundation.

Competing interests
The authors declare no competing interests.

Ethics
This study was approved by the University of Melbourne Human Research Ethics Committee (Approval ID: 30659). All participants provided informed consent prior to participation in both the monitoring phase and the follow-up discernment study. Demographic categories for gender (male, female, non-binary/third gender, prefer not to say) and age were defined by the investigators based on standard survey conventions. Gender and age data were collected to characterize the sample and assess representativeness, not as analytic variables. Ethnicity data were not collected.

Data availability
All materials needed to replicate this study are available via the Harvard Dataverse: https://doi.org/10.7910/DVN/UZ42LX. Free-text justification fields and participant account identifiers have been removed to protect anonymity; restricted materials are available on reasonable request, subject to ethics approval.

Appendix A: Participant classification framework and category level confidence

Before assigning a formal classification, participants were instructed to apply an initial evaluative screen referred to as the "sniff test." This step was designed to prompt reflective judgement and orient participants toward core norms of ethical democratic discourse. The sniff test did not generate a recorded variable; rather, it functioned as a heuristic guide to determine whether a post appeared to fall afoul of basic standards of truthful and constructive public communication.

The sniff test asked participants to consider whether the content violated any of the following norms of ethical democratic discourse:

- **Truthfulness:** Does the content appear truthful, or does it make claims that seem clearly or verifiably false, exaggerated, or unsupported?
- **Quantity:** Does the content provide sufficient relevant information, or does it appear vague, incomplete, or selectively detailed in a way that obscures key facts?
- **Relevance:** Does the content address the issue it claims to address, or does it distract, deflect, or misdirect?
- **Clarity:** Is the message clear and straightforward, or is it confusing, ambiguous, or unnecessarily complex?
- **Authenticity:** Does the contribution appear genuine and contextually appropriate (e.g., originating from a legitimate participant in Australian political discourse), or does it raise concerns about impersonation or deceptive identity?
- **Helpfulness:** Does the content appear intended to contribute constructively to democratic discourse, or does it seem designed to disrupt, inflame, or derail discussion?

If a post was judged to fall afoul of one or more of these norms, participants were required to assign it a primary classification category from a predefined framework of 17 misinformation types (with the option to select a secondary category where relevant). The framework was designed to distinguish among different mechanisms by which content may mislead, rather than simply reducing posts to a binary true/false judgement. The framework was developed pragmatically rather than derived from a single theoretical model, with the priority being categories that non-expert participants could readily understand and apply. We began from categories common to existing public-facing typologies of misinformation (e.g., Bounegru et al., 2018; Wardle & Derakhshan, 2017), producing an initial list of overlapping types. Three researchers then independently coded a development set of 24 social media posts using this list, comparing classifications and resolving disagreements through discussion. Where posts were not adequately captured by the initial categories, additional categories were added inductively to reflect the forms of misinformation present in the material. This process produced the final framework of 17 categories.

***Table A1**. Misinformation classification categories and descriptions.*

Category	Description
False content	Factually incorrect information, either wholly or partially false.
Out of context	Accurate or partly accurate information that misleads due to missing or distorted context.
Unproven/conspiratorial	Claims made without evidence (or using only anecdotal evidence), explanation, or conspiracies, e.g., baseless accusations or attributing events to secretive or deceptive forces without credible evidence.
Disconnection	Misleading headlines, visuals, or captions that do not accurately represent the content.
Altered media	Deepfakes, images, videos, or audio that have been altered, edited, or AI-generated to mislead.
Rage bait	Content designed to provoke strong emotional reactions for engagement—such as anger, outrage, or division—often by using inflammatory language, misleading framing, or exaggeration.
Impersonation	Misleading use of names, images, or branding to mimic real people, organizations, or media.
Bogus authority	Content that falsely claims or distorts authority, credentials, or expertise to mislead audiences. E.g., pretending to be an expert, or claiming expertise in an area they don't have.
Objectivity masquerade	Egregious examples of unlabeled partisan opinion, editorial, or analysis masquerading as factual, fair, or objective news, misleading audiences about its neutrality. Most likely to be from fringe media outlets.
Political defamation	False or defamatory claims intended to harm the reputation of political candidates, parties, or election officials.
Misunderstood satire	Satirical or parody content mistaken for real news, often due to unclear labelling or users spreading posts from parody accounts as factual reports.
Cherry-picking	Content that selectively highlights specific facts, quotes, or data while ignoring relevant context or contradictory information.
Unlabeled ads	Political ads lacking proper authorization or disclosure, potentially misleading viewers.
Political scams	Fraudulent schemes that exploit political events, candidates, or causes to deceive individuals into giving money.
Out of scope/not a problem	Not related to the Australian election or politics generally OR passes the sniff test.
True/partly true	The content is partly or wholly accurate.
Other	Fails the sniff test, but doesn't fit any of the categories.

Category level confidence

Across assessments, raters recorded their confidence (0–10) in each classification. Differences in mean confidence across categories were tested using a linear mixed-effects model, with classification category as a fixed effect and a random intercept for rater to account for the uneven number of assessments contributed by each participant (N = 3,037 assessments). Category significantly predicted confidence, $\chi^2(16) = 144.44$, $p < .001$, and a non-parametric Kruskal–Wallis test returned the same conclusion, $\chi^2(16) = 116.89$, $p < .001$. The test indicates that confidence varied across categories but does not identify which categories differed from one another; pairwise comparisons were not conducted, given the number of categories and the small number of assessments in several of them.

Table A2. *Mean confidence rating across categories.*

Category	**Mean confidence**	***N***
Political scams	8.36	11
False content	7.6	664
Misunderstood satire	7.56	39
Altered media	7.5	147
Disconnection	7.42	139
Unproven/Conspiratorial	7.38	362
Political defamation	7.3	92
Out of context	7.23	477
Impersonation	6.94	77
True/partly true	6.86	158
Out of scope/not a problem	6.84	321
Objectivity masquerade	6.81	53
Rage bait	6.71	260
Cherry-picking	6.68	174
Unlabeled ads	6.58	12
Bogus authority	5.14	7
Other	4.36	45

Appendix B: Participant demographics

Expanded basic demographics

This appendix presents expanded demographic information for participants who completed onboarding and provided consent to participate in the study (N = 419). All demographic questions were optional; therefore, totals vary slightly across items, and percentages are calculated using the number of valid responses for each question.

Table B1. *Demographic characteristics of onboarding participants.*

Variable	Category	*n*	%
Age group	18–24	24	5.8%
	25–34	58	13.9%
	35–44	84	20.2%
	45–54	96	23.1%
	55–64	120	28.8%
	65–74	24	5.8%
	75+	10	2.4%
Gender	Female	199	47.8%
	Male	203	48.8%
	Non-binary / third gender	8	1.9%
	Prefer not to say	6	1.4%
Highest education	Bachelors degree	148	35.7%
	Doctorate	42	10.1%
	High school graduate	27	6.5%
	Less than high school	6	1.4%
	Masters degree	104	25.1%
	Professional degree	17	4.1%
	Some university	71	17.1%
Location (state/territory)	Northern Territory	4	1.0%
	Tasmania	10	2.4%
	Outside Australia	12	2.9%
	South Australia	19	4.6%
	Western Australia	24	5.8%
	ACT	26	6.3%
	Queensland	76	18.3%
	New South Wales	91	21.9%
	Victoria	153	36.9%
Employment status	Not working (other)	17	4.1%
	Not working (disabled)	24	5.8%
	Not working (looking for work)	41	9.9%
	Not working (retired)	47	11.3%
	Working (self-employed)	48	11.6%
	Working (paid employee)	238	57.3%

Political alignment scoring rubric

Participants' responses to a set of attitudinal and "feeling thermometer" questions were converted to numerical scores, which were then summed to produce an overall political alignment score. Negative scores indicate a more left-leaning orientation, positive scores indicate a more right-leaning orientation, and values near zero indicate centrist or mixed views. Several items were adapted from Pew Research Center's political typology surveys (Pew Research Center, 2021). Weights were intentionally set asymmetrically for items where one response option signals a strong ideological position while the opposing option reflects a broadly mainstream view that carries less diagnostic weight. For example, endorsing "too much openness risks losing national identity" was treated as a stronger rightward signal than endorsing "openness is essential to Australian identity" was a leftward one, since the latter reflects a widely held position across the Australian political spectrum. Symmetric weights in such cases would overstate the ideological significance of mainstream responses.

Table B2. *Political leaning scoring rubric.*

Question	Response Option	Score
If you had to choose, would you rather have:	Larger government providing more services	−3
	Smaller government providing fewer services	+3
	No answer / Other	0
Which of the following statements is closest to your viewpoint?	Openness is essential to Australian identity	−1
	Too much openness risks losing national identity	+7
	No answer / Other	0
In general, would you say that experts who study a subject for many years are:	Usually worse than non-experts	+7
	Neither better nor worse	+3
	Usually better	0
	No answer	0
How much more needs to be done, if anything, to ensure that all Australians have equal rights and opportunities, regardless of race, ethnicity or religion?	None at all	+8
	A little	+4
	A moderate amount	0
	A lot	−4
	A great deal	−8
How much of a problem, if any, would you say each of the following are in Australia today? - People being too easily offended by things others say	Major problem	+6
	Minor problem	0
	Not a problem	−6
How much of a problem, if any, would you say each of the following are in Australia today? - People saying things that are offensive to others	Major problem	−6
	Minor problem	0
	Not a problem	+6
Which statement do you agree with most?	Corporations make a fair and reasonable amount	+3
	Corporations make too much profit	0

Which statement do you agree with most?	Private schools receive a fair share of government money.	+4
	Private schools receive too much funding from the government and more money should go to public schools.	−4
Should Australia increase its defense spending to counter potential threats in the Asia-Pacific region?	Strongly agree with increasing	+4
	Somewhat agree	+1
	Neither agree nor disagree	0
	Somewhat disagree	−5
	Strongly disagree	−7

Attitudes toward political parties scoring rubric

Participants rated their feelings toward political parties/groups from 0 (extremely negative) to 10 (extremely positive). Each rating was re-scaled so that positive feelings toward right-leaning parties/groups increased the score, while positive feelings toward left-leaning parties/groups decreased it (and vice versa).

Table B3. *Transformation formulae for converting party feeling-thermometer ratings into political alignment scores. Ratings were centered at 5 (neutral) and scaled by party-specific weights.*

Political group	Transformation formula
Liberal Party	(rating − 5) × 0.4
Labor Party	(rating − 5) × −0.2
Nationals	(rating − 5) × 0.5
Greens	(rating − 5) × −0.4
Teals	(rating − 5) × −0.2
Independents	(rating − 5) × −0.1
One Nation	(rating − 5) × 0.6

Political alignment distribution

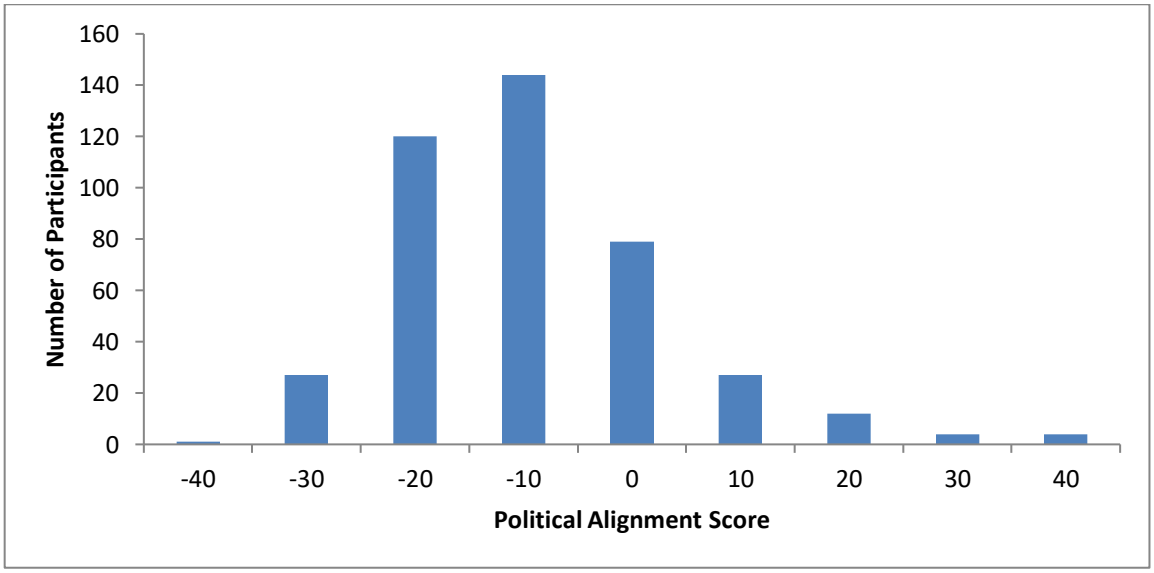

Figure B1. Distribution of political alignment scores for all consenting participants.

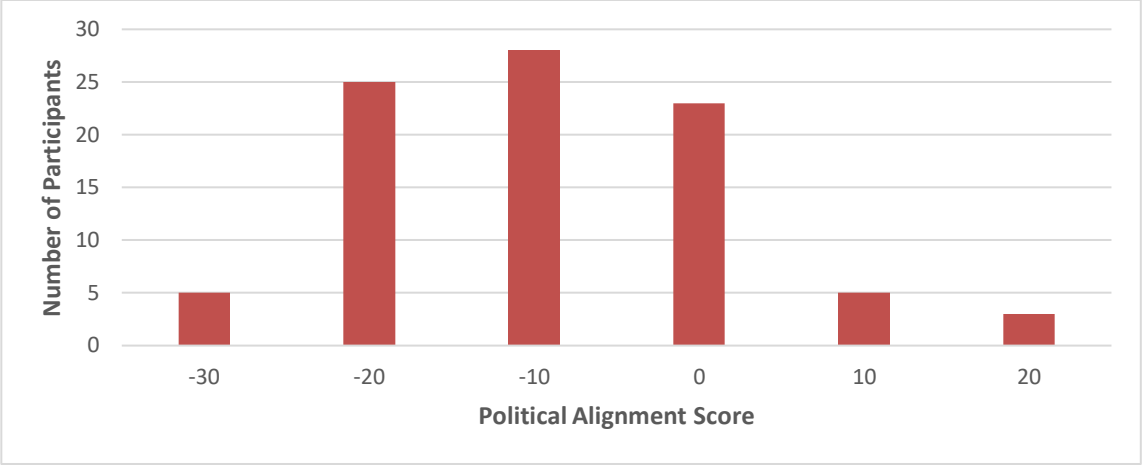

Figure B2. Distribution of available political alignment scores for active participants (n = 89).

Mitigating ideological bias in detection

The cohort's left-of-center skew (see distributions above) raises the concern, noted in the body, that raters may flag content conflicting with their own views (Park et al., 2026). Two points contextualize this. First, the skew broadly mirrors the Australian electorate at the time: the 2025 election was held under compulsory voting with roughly 90% turnout (Australian Electoral Commission, 2025b) and returned a landslide for the center-left Labor Party (Australian Electoral Commission, 2025a), so a left-leaning pool is less anomalous than in a voluntary-turnout setting. Second, our attempt to balance the cohort through separate targeted recruitment campaigns aimed at left- and right-of-center users did not overcome the skew, which suggests it reflects the composition of the politically engaged social media population we drew from rather than a recruitment artifact. The mitigations described in the body—bias training, required written justification (Pennycook & Rand, 2019), and dilution of individual skew through peer assessment—reduce but cannot eliminate this risk.

Appendix C: Target and narrative theme coding

Misinformation posts identified during the monitoring period were coded using a structured codebook designed to capture key narrative characteristics of the content. Each post was coded for two dimensions: target and narrative theme. The target refers to the organization, political party, institution, or individual that the content attempts to discredit or malign. The narrative theme captures the central narrative claim conveyed by the post. Not all posts contained a clear target. The target coding scheme and results are provided below; the full narrative theme codebook and complete frequency tables are available in the project repository: https://doi.org/10.7910/DVN/UZ42LX.

Target categories

***Table C1**. Targets of misinformation submissions, grouped by type.*

Targets	Type
Political parties and political actors	Labor party
	Liberal party
	The Greens
	Teal Independents
	Australia's Voice
	Trumpet of Patriots
	Lidia Thorpe
Media Organizations	ABC
	Channel 9
Institutions	Australian Electoral Commission (AEC)
	Electoral system
	South Australia Police
Political coalitions or combined targets	Labor / Greens
	Labor / Liberal
	Greens / Teal Independents

Target analysis results

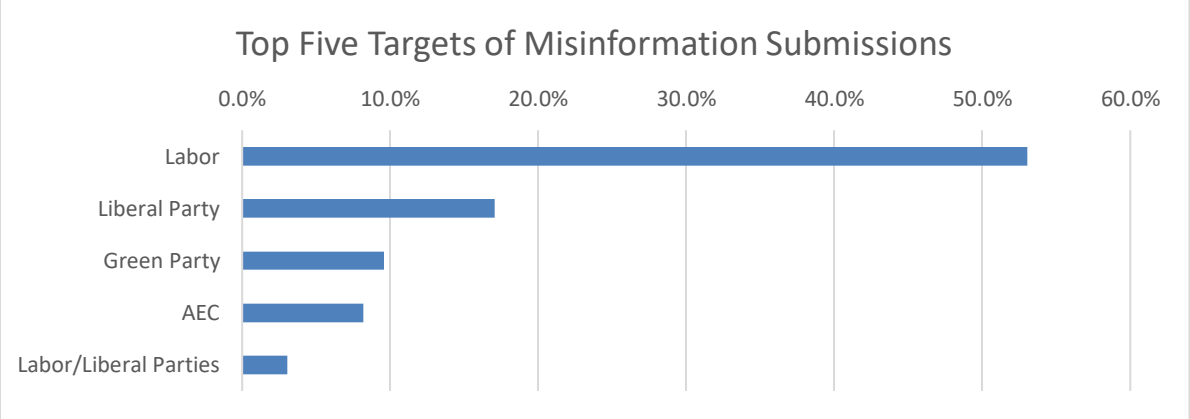

Figure C1. Top five targets of misinformation submissions, as a proportion of the 428 submissions with an identifiable target.
A further 164 submissions (27.7% of all submissions) had no clear target and are not shown.

Appendix D: Platform workflow and data fields

Participants interacted with a custom web-based platform used to submit and assess potential misinformation posts. The platform allowed users to log content encountered on social media, categorize the issue using a classification framework, and assess submissions made by other participants. Each post could be evaluated by multiple participants.

Data fields for submissions and assessments

Table D1. *Data fields recorded for each submission and assessment.*

Field	Description
Post URL	Link to the original social media post
Screenshot	Image capture of the content submitted by the participant
Initial classification	Category assigned by the submitting participant (for assessments)
Misinformation category	Category participant believes the content falls under
Confidence rating	Self-reported confidence in the classification
Justification	Short written explanation of the classification
Content type	Whether the post contained text/image or video
Additional assessments	Classifications and confidence ratings submitted by other participants

Appendix E: Assessment of fact-checking coverage

To assess whether the misinformation narratives surfaced by participants corresponded with narratives identified by professional fact-checking organizations, we compiled a dataset of election-related fact checks published by Australian Associated Press (AAP) FactCheck.[2] Because election misinformation typically circulates well before polling day, we deliberately used a wider collection window (January 9–May 14, 2025). Each fact check was reviewed and coded using the same narrative theme codebook applied to participant submissions. Coding focused on the central narrative claim addressed by the fact check rather than the specific wording of the claim or the individual post investigated. This allowed comparison between narratives identified through crowdsourced monitoring and those documented through professional fact-checking.

In total, 34 AAP fact checks were published during the monitoring period. Table E1 summarizes their distribution across narrative themes. Electoral integrity claims were the most frequently fact-checked category (*n* = 11, 32.4%), with a notable concentration in the post-election period (May 3–14), driven by claims about ballot counts, voter turnout, and electoral fraud. Government spending (*n* = 5, 14.7%) and cost of living and inflation (*n* = 4, 11.8%) were the next most common categories. Together, these three themes accounted for approximately 59% of all professional fact checks in the dataset. The remaining ten themes each appeared in one or two fact checks.

Table E1. *Distribution of AAP FactCheck articles by narrative theme (January 9 – May 14, 2025; N = 34).*

Narrative theme	Fact checks (*n*)	% of total
Electoral integrity	11	32.4%
Government spending	5	14.7%
Cost of living / Inflation	4	11.8%
Energy policy	2	5.9%
Health & education funding	2	5.9%
National security	2	5.9%
Tax policy fearmongering	2	5.9%
Anti-government spending	1	2.9%
Housing crisis	1	2.9%
Immigration	1	2.9%
Impersonation	1	2.9%
Misconduct allegation	1	2.9%
Nuclear energy	1	2.9%
Total	34	100%

[2] See: https://www.aap.com.au/factcheck/

Appendix F: Follow-up discernment study

The discernment test

Participants in the treatment group were drawn from individuals who had taken part in the monitoring phase and consented to be recontacted for follow-up research; *n* = 56 completed the follow-up assessment. The control group (*n* = 132) was recruited using channels and eligibility criteria mirroring the original cohort, creating a comparison group broadly similar to the original participant pool while ensuring control participants had not taken part in the monitoring exercise.

Both groups completed the same online discernment test of 20 items (10 true, 10 false), administered through an online survey platform, with participants indicating whether they believed each item was "true" or "false." The items covered a range of topics and were not specifically focused on politics or elections, in order to assess general misinformation discernment rather than recall of claims encountered during the monitoring phase.

The items were purpose-built rather than drawn from real social media, because constructed items can be assigned an unambiguous ground truth—a determination that is often not possible for authentic posts, whose veracity may be contested or context-dependent. Each item was therefore designed to be verifiably true or false while resembling the format and style of typical news-sharing posts, including the visual conventions of particular outlets.

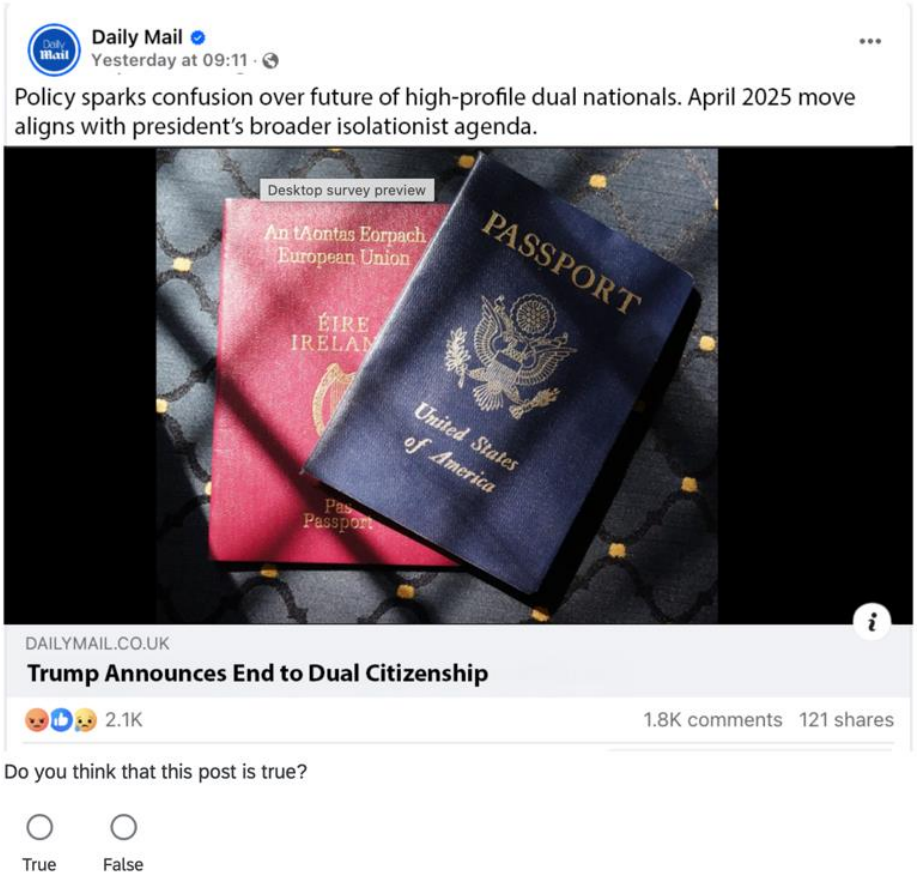

Figure F1. Example simulated Facebook-style news post used in the discernment test.

Group differences and adjusted comparison

The treatment group (n = 56) answered 73.8% of test items correctly (SD = 0.137), compared with 67.5% (SD = 0.128) in the control group (n = 132); a difference equivalent to a moderate effect size (Cohen's d = 0.48, 95% CI [0.16, 0.79]). Because the two groups were unequal in size, we compared mean accuracy using Welch's t-test, which does not assume equal variances or balanced samples. The difference was statistically significant, $t(97.8) = 2.91$, $p = .005$. This indicates that the higher discernment observed in the treatment group is unlikely to reflect sampling variation alone.

The treatment and control groups differed on several demographic characteristics (Table F1). Treatment participants were younger, more likely to be in paid employment, and more highly educated than the control group, which was older and included a high proportion of retirees. The groups were similar in their distribution across Australian states, though the treatment group included somewhat more women. Because age, employment, and education could each plausibly be associated with misinformation discernment, these three characteristics were included as covariates in the adjusted analysis reported below.

Table F1. *Sample characteristics. Percentages are calculated among participants who reported each characteristic. Four treatment participants did not provide demographic information and are excluded from the adjusted analysis below.*

Characteristic	Treatment	Control
Aged 55 or older	38%	73%
In paid employment	67%	43%
Bachelors degree or higher	88%	63%

To examine whether this difference could be explained by the demographic differences between the groups, we fitted a linear mixed-effects model. The model predicted each true/false response from the item's actual veracity, the participant's group (treatment or control), and the interaction between the two, with age, employment, and education included as covariates and random intercepts for each participant and each item. The interaction between veracity and group is the term of interest: it captures whether being in the treatment group was associated with a sharper distinction between true and false items. This interaction was positive and statistically significant (b = 0.12, SE = 0.03, t = 3.96, $p < .001$), indicating that the treatment group discriminated true from false content more effectively than the control group, after adjusting for the demographic variables on which the groups differed. Because participants were not randomly assigned, however, this adjustment accounts only for measured differences and does not establish that participation caused the improvement.