Human Research Enhanced By AI

How to Execute and How to Protect Against Misinformation

Sofia & Toby Goldblatt
Renew

December 2025

Abstract: AI writing tools promise efficiency but deliver critical vulnerabilities in quality and security. Recent research reveals AI-generated content achieves only 40-49% professional standards while being compromised by as few as 250 malicious documents. This white paper examines these dual challenges and presents a framework for AI governance, infrastructure security, and organizational capability building to protect against misinformation.

The Dual Challenge: Quality and Security

AI writing tools promise efficiency but deliver two critical vulnerabilities: systematic quality deficits and susceptibility to data poisoning. Recent research from the University of Edinburgh Business School and the Alan Turing Institute reveals AI-generated content achieves only 40-49% professional standards while being compromised by as few as 250 malicious documents. For organizations producing high-stakes research, investment proposals, or policy documents, understanding these vulnerabilities and implementing robust governance frameworks is essential.

Figure 1: The AI Performance Gap

Evaluation Criteria	AI Performance	Expert Standard
Quality of Argument	40-49% ■	70%+ ✓
Evidence Currency	Outdated (1978-2020) ■	Current (2023-2025) ✓
Communication Style	Bullet points ■	Flowing prose ✓

Source: University of Edinburgh Business School, 2025

Three Critical Quality Gaps

1. Argument Quality: AI defaults to descriptive bullet points rather than analytical arguments. Research shows AI lists relevant factors without explaining interconnections or building evaluative conclusions. Strategic documents cite market forces without explaining competitive dynamics; investment proposals list opportunities without justifying prioritization.

2. Evidence Strength: Edinburgh research revealed AI citing sources from 1978, 1995, and early 2020s for 2025 challenges. Outdated evidence undermines credibility, particularly for market dynamics, regulatory environments, or social trends. AI name-drops examples (Patagonia, Shell) without detailed explanation.

3. Communication Clarity: Excessive subheadings and bullet points prevent coherent argument development. Research identified citation inconsistencies mixing referencing styles and lacking comparative language connecting ideas.

The Data Poisoning Threat

Beyond quality deficits, AI systems face security vulnerabilities. Research from the Alan Turing Institute, UK AI Security Institute, and Anthropic reveals that just 250 malicious documents can create backdoor vulnerabilities in LLMs—regardless of model size or training data volume. This finding challenges assumptions that larger models require proportionally more poisoned data.

Key finding: A 13B parameter model trained on 20× more data than a 600M model can be compromised by the same 250 poisoned documents. Attackers need only a fixed, small number—not a percentage—of training data.

These backdoors enable malicious behaviors: extracting sensitive data, degrading system performance, producing biased information, or bypassing security protocols. Since LLMs train on publicly available internet text, attackers can publish targeted content on webpages or blogs. Creating 250 malicious documents is trivial compared to millions, making this vulnerability accessible to potential attackers.

Figure 2: Data Poisoning Vulnerability Scale

Model Size	Training Data	Documents to Poison
600M parameters	~4M books	~250
13B parameters	~90M books	~250

Source: Alan Turing Institute / Anthropic / UK AI Security Institute, 2025

Execution Framework: The Optimal Workflow

Stage	AI Role (Speed)	Human Role (Quality)
Draft	Generate structure, identify topics	Define scope, provide prompts
Evidence	Locate potential sources	Verify currency, replace outdated data
Analysis	Present factors	Build arguments, reasoning chains
Refine	Basic formatting	Restructure, ensure consistency
Verify	—	Quality check, integrity review

Core Principle: AI provides speed; humans ensure credibility and protection against misinformation.

Five Protection Strategies Against Misinformation

1. Source Verification Protocol: Never accept AI sources without verification. Check publication dates, author credentials, and whether research has been superseded. Prioritize 2023-2025 sources for current topics. Access originals rather than AI summaries.

2. Argument Reconstruction: Convert descriptive bullet points into analytical prose. Build reasoning chains explaining why claims hold. Ask: How do factors interconnect? Which proves most significant?

3. Evidence Integration: Replace name-dropping with detailed explanation. Research specific practices, outcomes, mechanisms. Integrate statistics with clear interpretation.

4. Data Poisoning Awareness: Recognize AI training vulnerabilities. Question suspicious patterns, inconsistent outputs, or unexpected biases. Cross-reference multiple AI tools to identify potential

poisoning effects.

5. Expert Review for High-Stakes: Investment proposals, academic publications, policy briefs, and legal documents require expert human review. The 40-49% quality standard guarantees mediocrity in competitive contexts.

Figure 3: Risk Assessment Matrix

Document Type	Financial/Reputational Risk	Expert Review
Internal notes	Low	Optional
Client communications	Medium	Recommended
Investment proposals	High	Essential
Academic publications	High	Essential
Legal/Policy documents	Critical	Mandatory

Decision Rule: If quality impacts funding, reputation, or legal standing, expert review is non-negotiable.

The Renew Solution: A Three-Pillar Framework

Renew provides a comprehensive framework to protect organizations from AI vulnerabilities while maximizing the technology's benefits. Our approach addresses both quality deficits and security threats through systematic governance, infrastructure protection, and capability building.

Pillar 1: Implement AI Governance Policies and Controls

Effective AI governance begins with clear policies defining acceptable use, quality standards, and accountability frameworks. Renew works with organizations to establish:

- **Usage Guidelines:** Define which tasks require AI assistance versus expert human oversight. Establish thresholds for when AI drafts require mandatory review based on document risk levels.

- **Quality Control Checkpoints:** Implement verification stages for argument quality, evidence currency, and source reliability. Mandate expert review for high-stakes documents affecting funding, reputation, or legal standing.

- **Compliance Frameworks:** Ensure AI usage aligns with industry regulations, data protection requirements, and professional standards. Document AI contributions for audit trails and intellectual property clarity.

- **Accountability Structures:** Assign responsibility for AI outputs, establish escalation procedures for quality concerns, and create feedback loops for continuous governance improvement.

Pillar 2: Secure AI Infrastructure

Given the data poisoning vulnerabilities revealed by Turing Institute research, infrastructure security becomes critical. Renew helps organizations implement:

- **Source Verification Systems:** Deploy automated tools to verify training data provenance, flag suspicious content patterns, and cross-reference outputs against known reliable sources.
- **Multi-Model Validation:** Use multiple AI systems from different providers to cross-check outputs. Inconsistencies between models may indicate data poisoning or quality issues requiring investigation.
- **Output Monitoring:** Implement continuous monitoring for unexpected behaviors, bias patterns, or quality degradation that could signal compromised models or emerging vulnerabilities.
- **Secure Deployment Practices:** Establish network segmentation, access controls, and data handling protocols that minimize exposure to malicious inputs while maintaining operational efficiency.

Pillar 3: Train and Increase Organizational Awareness

Technology alone cannot protect against AI misinformation—organizations need skilled personnel who understand vulnerabilities and recognize warning signs. Renew delivers:

- **Misinformation Recognition Training:** Teach staff to identify the three quality gaps (argument weakness, outdated evidence, poor communication) and data poisoning indicators like unexpected output patterns or suspicious source recommendations.
- **Critical Evaluation Skills:** Build organizational capability to assess AI-generated arguments, verify source currency and authority, and reconstruct descriptive content into analytical prose meeting professional standards.
- **Domain-Specific Expertise Development:** Provide specialized training for teams in business, academia, policy, and legal contexts—each facing unique AI risks requiring tailored mitigation strategies.
- **Continuous Capability Building:** Establish ongoing learning programs tracking AI evolution, emerging threats, and evolving best practices. Create communities of practice sharing lessons learned and successful protection strategies.

This integrated framework ensures organizations harness AI's efficiency benefits while maintaining the quality, security, and credibility essential for professional success. Governance provides structure, infrastructure delivers security, and training builds human capability—together creating resilient defenses against AI vulnerabilities.

Conclusion

Research evidence reveals AI's dual vulnerability: systematic quality deficits (40-49% professional standards) and security weaknesses (250 documents create backdoors). For high-stakes documents where credibility, funding, or legal standing depend on quality, accepting AI output without governance frameworks, security measures, and organizational capability creates unacceptable risk.

The optimal approach implements comprehensive AI governance, secures infrastructure against data poisoning, and builds organizational skill in identifying misinformation. Human research enhanced by AI—protected by systematic controls and expert oversight—represents the future of trustworthy knowledge work.

References

Edinburgh Business School (2025). *Evaluating an AI Generated Essay on the Contemporary Challenge of Global Inequality in 2025*. University of Edinburgh.

Souly, A., Rando, J., Chapman, E., Davies, X., Hasircioglu, B., Shereen, E., Mougan, C., Mavroudis, V., Jones, E., Hicks, C., Carlini, N., Gal, Y., & Kirk, R. (2025). Poisoning Attacks on LLMs Require a Near-constant Number of Poison Samples. Alan Turing Institute, UK AI Security Institute, and Anthropic. *arXiv:2510.07192*.

Alan Turing Institute (2025). LLMs may be more vulnerable to data poisoning than we thought. https://www.turing.ac.uk/blog/llms-may-be-more-vulnerable-data-poisoning-we-thought