ProofPols: Making trust inspectable

Team The Strongest

Track

Track 1 — Restoring Trust in the Age of Synthetic Media

1 Problem Context

Advances in generative AI have made it trivial to create realistic text, images, audio, and video. This has eroded a fundamental pillar of the digital world: trust. Synthetic media enables misinformation, impersonation, scams, and reputational harm at scale[1].

Current solutions are fragmented, opaque, and often overconfident[3, 2]. They typically conflate two different questions:

- Where does this content come from?
- Are the claims in this content factually reliable?

These must be treated as independent dimensions.

2 Product Concept

ProofPols is a trust infrastructure that independently assesses:

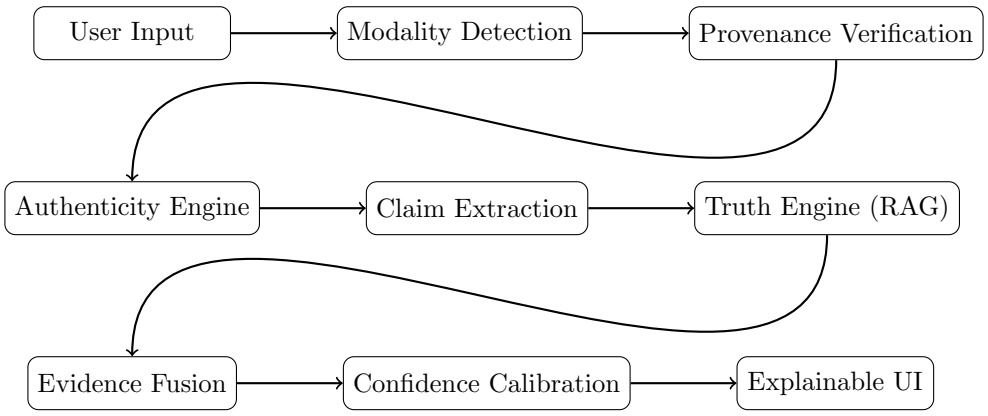
1. **Authenticity** — whether content is real, AI-generated, or edited.
2. **Credibility** — whether explicit claims associated with content are supported by evidence.

ProofPols does not attempt to decide what users should believe. Instead, it provides interpretable evidence, conservative confidence, and explicit uncertainty to support informed human judgment.

3 User Flow

1. User uploads content (text, image, audio, or video) and optional claim context.
2. The system detects modality and preprocesses the input.
3. The Authenticity Engine evaluates origin and integrity.
4. If a textual claim is present, the Truth Engine evaluates its credibility.
5. Results are displayed as two independent panels with explanations and limitations.

4 System Architecture



5 Authenticity Engine

The Authenticity Engine assesses whether content is:

- Likely Real
- Likely AI-Generated
- Likely Edited / Mixed
- Uncertain

It uses converging evidence from:

- Cryptographic provenance (C2PA, content credentials)[4]
- Physical and signal constraints (camera physics, audio acoustics)
- Temporal consistency (video)[7, 8]
- Generative-prior behavioral tests (diffusion reconstruction stability)[5, 6]
- Limited contextual signals

Evidence is fused using Bayesian accumulation with capped likelihood ratios and conflict dampening. Confidence is conservatively bounded per modality.

6 Truth Engine

The Truth Engine evaluates the credibility of explicit or implicit textual claims associated with content.

Pipeline

1. Claim extraction and decomposition[9]
2. Evidence retrieval using Retrieval-Augmented Generation (RAG)[10] from:
    - Wikipedia / Wikidata
    - News APIs
    - Government and scientific datasets
3. Cross-source agreement analysis
4. LLM-based comparison between claims and retrieved evidence
5. Conservative scoring with uncertainty

Truth is never assigned directly to images, audio, or video — only to claims expressed about them.

7 Outputs

Each analysis produces:

- Verdict
- Confidence score (bounded)
- Evidence summary
- Known limitations

8 User Experience

- Drag-and-drop upload interface
- Two-panel trust display (Authenticity and Credibility)
- Visual confidence indicators
- Expandable evidence explanations
- Explicit uncertainty messaging

Designed for non-technical users.

9 Ethics and Privacy

- Stateless analysis
- No biometric storage
- No author fingerprinting
- No opaque trust scores
- Explicit uncertainty and limitations[12, 13]
- Human review encouraged for high-stakes use

10 Impact and Metrics

Impact

- Reduced misinformation spread
- Lower scam and impersonation success
- Improved user awareness and caution

Metrics

- Calibration accuracy[11]
- False positive rate
- User comprehension surveys
- Reduction in blind sharing behavior

11 Scalability

- Modular pipeline per modality
- Cloud and edge deployable
- Extendable to new content types and platforms

12 Prototype / Demo

We provide a web-based demo that:

- Accepts all media types
- Shows dual trust panels
- Uses mock and real backend endpoints
- Demonstrates realistic user interaction

13 Technical Stack

Backend

- Django (core backend and orchestration)
- REST APIs for forensic services
- Modular services for each analysis pipeline

ML and Forensics

- PyTorch for model inference
- OpenCV for image/video processing
- librosa / torchaudio for audio
- Diffusion-based behavioral testing

Truth Engine

- Retrieval-Augmented Generation (RAG)
- Vector search (FAISS / similar)
- LLM for comparison only

Frontend

- HTML, CSS, JavaScript
- Professional, minimal UI

14 Conclusion

ProofPols restores trust not by making strong claims, but by providing clear evidence, conservative confidence, and explicit uncertainty. It enables users to navigate a world saturated with synthetic media without relying on blind faith or opaque systems.

References

[1] Chesney, R., & Citron, D. (2019). *Deepfakes and the New Disinformation War*. Foreign Affairs.

[2] Gragnaniello et al. (2021). Detecting GAN-generated images with CNNs. IEEE.

[3] OpenAI (2023). Why AI detection is unreliable.

[4] Coalition for Content Provenance and Authenticity (C2PA). `https://c2pa.org`

[5] Dhariwal, P., & Nichol, A. (2021). Diffusion Models Beat GANs on Image Synthesis. NeurIPS.

[6] Meng et al. (2021). SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. ICLR.

[7] Korshunov & Marcel (2018). Vulnerability of face recognition to deep fake attacks. BTAS.

[8] Chung & Zisserman (2016). Out of Time: Automated Lip Sync in the Wild. ACCV.

[9] Thorne et al. (2018). FEVER: a large-scale dataset for fact extraction and verification. NAACL.

[10] Lewis et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. NeurIPS.

[11] Guo et al. (2017). On Calibration of Modern Neural Networks. ICML.

[12] Selbst et al. (2019). Fairness and Abstraction in Sociotechnical Systems. FAT*.

[13] Raji et al. (2020). Closing the AI Accountability Gap. FAT*.