



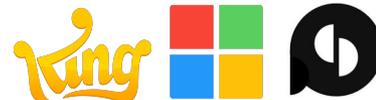
# A Practical Synthesis of Detecting AI-Generated Textual, Visual, and Audio Content



Lele Cao

Thanks to

MULTIMODAL AI  
@ ICDM 2025



# About Lele

**Senior Principal AI/ML Researcher  
Research Lead**  
@ King AI Labs, Microsoft

**Co-founder**  
@ CSpaper



## Goal of this work

- Building a unified view of how to detect AI-generated content across text, image, and audio domains.
- Provide practical industrial perspectives.

## What defines me

- 16-year industrial experience in telecom, e-commerce, private equity, gaming, geo-informatics.
- 10-year academic experience in machine learning, robotics, software engineering, HCI, and clinical science.
- Father of 2 boys; handyman, gardener and architect at home.

# Why Detect GenAI Content?

- Explosion of LLMs, diffusion, and TTS (text-to-speech) models  
→ challenges of misinformation, academic integrity, and authenticity.
- Real-world risks: deepfakes, plagiarism, synthetic news.
- Detection = Trust Infrastructure for digital ecosystems.  
Misinformation → Trust Erosion → Detection Need.



*Deloitte used AI in \$290,000 report for Australian government without declaration*



*A fabricated Pentagon explosion image got viral on social media.*



*An old man lost his pension after fraudsters called by replicating his son's voice.*

# Taxonomy of Detection Methods

- Statistical / Signal-Based  
*"Detecting the invisible noise – using statistical or signal fingerprints left by generative models."*
- Model-Likelihood Based (e.g., perplexity, curvature, reconstruction error)  
*"Let the model judge the model – measure how plausible the content is under real-world probability."*
- Supervised Classifiers  
*"Teach a model to spot fakes – train detectors on curated human-vs-AI datasets."*
- Provenance (Watermark / Fingerprint)  
*"Sign what you synthesize – tracing authenticity through embedded or metadata watermarks."*
- Retrieval / Consistency-Based  
*"Catch the copycat – retrieve, compare, and expose paraphrased or inconsistent AI outputs."*

# Detecting AI-Generated Text

- LLM Prompting (Zero-/Few-Shot Detection)  
*Simple but prompt-sensitive and model-dependent.*
- Linguistic & Statistical Signals (e.g., perplexity, n-gram patterns, stylistic features ...)  
*Strength: interpretability; Weakness: sensitivity to paraphrasing.*
- Training-Based Methods  
*Pros: high accuracy in-domain; Cons: overfit to specific models or datasets.*
- Watermarking Techniques  
*Embeds statistical patterns during generation to prove authenticity.*
- Retrieval-Based Approaches (Paraphrase Defense)  
*Builds databases of known LLM outputs to catch paraphrased variants.*

# Detecting AI-Generated Text

- Often, “*Detection isn’t one method – it’s a multi-layer defense, from simple prompting to robust watermarking.*”
- And you can apply **human-assisted approaches**: letting “I” have a say

ChatGPT generated text: Colors (top k):  10  100  1000

Microsoft Corporation is a global technology leader founded by Bill Gates and Paul Allen in 1975, renowned for its software products like Windows, Office, and its Azure cloud computing platform. The company also develops hardware such as the Surface line of tablets and laptops, and owns LinkedIn and GitHub.

Human authored text:

Motherbrain is EQT's proprietary investment platform driven by diversified big data and cutting-edge algorithms. EQT uses Motherbrain across the EQT platform to source deals and to help investment teams make better informed investment decisions. One of many analytical scenarios that Motherbrain helps tackle is defining the similarity between companies, which can be useful in tasks such as competitor mapping.

A visualization from [GLTR](#), which highlights the likelihood of each word being generated by an LLM, aiding human reviewers in detecting patterns that may indicate AI-generated text.

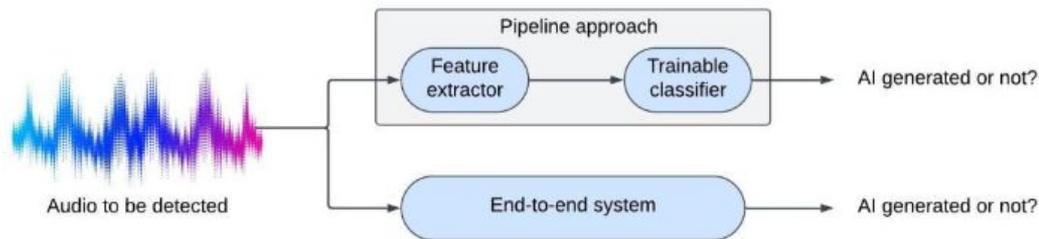
More robust and reliable after humans are trained for such tasks.

# Visual Content Detection

- Observation-based cues (first-line defense)  
*physical, physiological, stylistic anomalies.*
- Model-based:  
*GAN/diffusion artifact analysis (Fourier/frequency cues).*
- Watermarking:  
*DWT/DCT/SVD, SynthID, Stable Signature.*
- Temporal forensics for video:  
*optical flow, viseme mismatch (lips don't lie), photoplethysmography.*



# Detecting AI-Generated Audio



- Pipeline classifiers (e.g., MFCC/LFCC + CNN)
- End-to-end deep models (e.g., SincNet, AASIST, Wav2Vec2, HuBERT)
- Watermarking frequency modulation (AudioSeal, SynthID-Audio)

## Challenges:

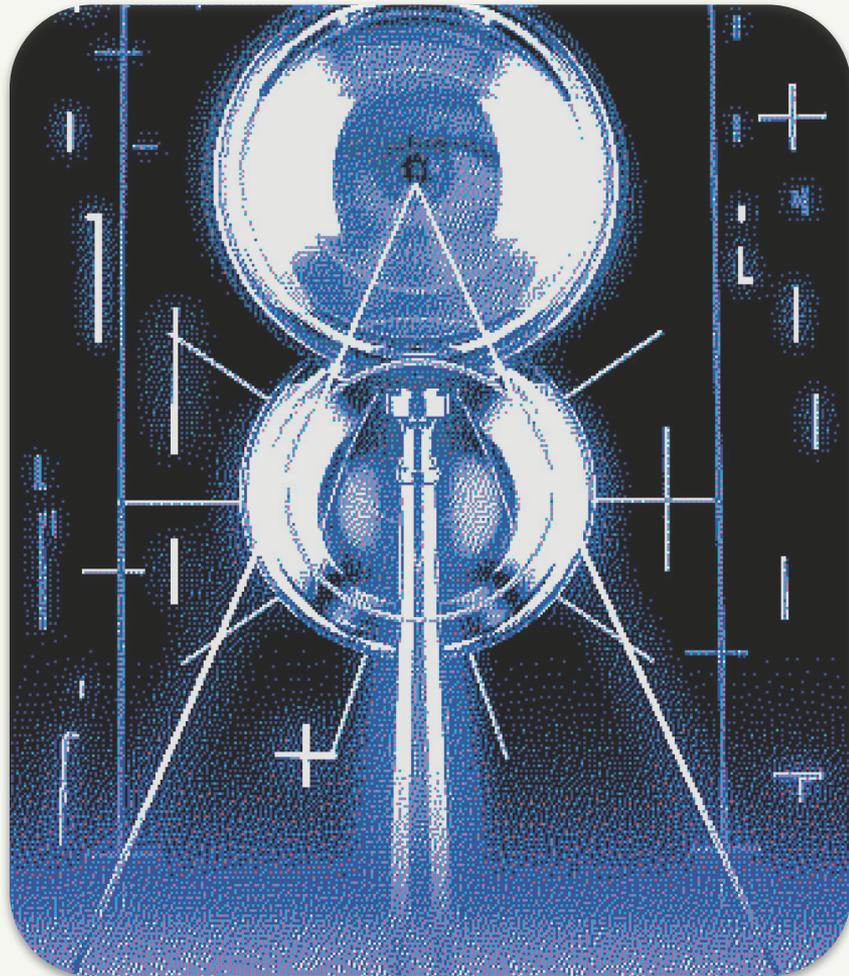
- Latency & real-time response for fraud calls  $\leq 200$  ms pipeline.
- Multilingual robustness: accent & noise drop accuracy (ASVspoof 2021 LA/PA tracks).
- Domain generalization  $\rightarrow$  train on diverse TTS pipelines.
- Adversarial perturbations: noise or codec shift can fool detectors.

# Cross-Modal Insights

- Arms race with generators
- Vulnerability to perturbations
- Watermarking as partial solution
- Multimodal ensemble detection
- Ethical/privacy trade-offs



*"No silver bullet — effective detection combines modalities, metadata, and governance."*



# Future

# Direction

Governance & provenance (C2PA)

Robustness to paraphrasing & compression.

Domain generalization for unseen generators.

Explainability → Human-AI collaboration.

## More details, examples and use cases:

- A Practical Guide to Detect GenAI Content
- A Practical Synthesis of Detecting AI-Generated Textual, Visual, and Audio Content
- Watermarking for AI Content Detection: A Review on Text, Visual, and Audio Modalities



# CSPaper

## A real-world vertical case

<https://cspaper.org>

AI agent system to help researchers to obtain fast and conference-specific review for their research papers.



A Short Paper:

[CSPaper Review: Fast, Rubric-Faithful Conference Feedback](#)

CSPaper.org Beta

My Reviews

---

**YOUR REVIEW**

**Desk Rejection Assessment:**

Paper Length  
Topic Compatibility  
Minimum Quality

**Expected Review Outcome:**

Summary  
Strengths and Weaknesses  
Quality  
Clarity  
Significance  
Originality  
Questions  
Ethical concerns  
Limitations  
Overall  
Confidence

**How a Critical Reviewer Might Rate**

Quality  
Clarity  
Significance  
Originality  
Overall  
Confidence

---

Sign out

---

**Attention Is All You Need**

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones Polosukhin

Reviewed on June 29, 2025 [View Paper](#)

---

Reviewed as if submitted to NeurIPS 2025 - ma

**Desk Rejection Assessment:**

*Disclaimer: Anonymity and formatting are not checked. This assessment focuses on correctness, novelty, clarity, and scientific quality per NeurIPS desk rejection criteria.*

**Paper Length**

Pass

**Topic Compatibility**

Pass

The paper presents a new sequence transduction model for applications in natural language processing (machine translation, parsing). These subjects fall squarely within the scope of machine learning, deep learning architectures, NLP, and evaluation.

**Minimum Quality**

Pass

All required major sections are present: Abstract, Introduction, Model Architecture (including Experiments/Tables of quantitative results), Discussion/Conclusion. The paper provides empirical and comparative evidence, describes a complete and original method, and shows improvements over prior work. No major methodological or clarity flaws are present.

---

**Expected Review Outcome:**

*Disclaimer: Only the main paper is subject to review, in accordance with the Track.*

**Summary**

# Usage



# Will Your Paper Be Accepted? Find Out In A Minute.

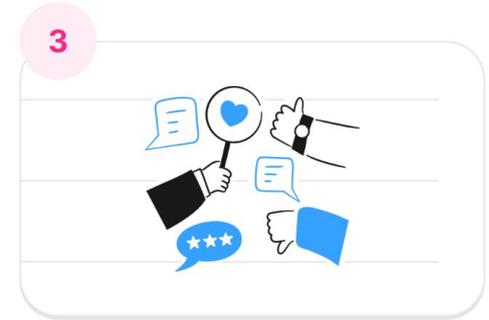
## How it works



Go to [CSpaper.org](https://cspaper.org)



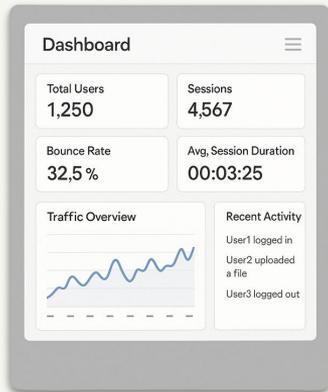
Upload your paper



Submit for review

# User

# Feedback



Survey

How satisfied are you with our product?

Very Unsatisfied  Neutral  Satisfied  
 Very Satisfied

How often do you use our product?

Daily  Weekly

What is your favorite feature?

Any additional comments?

Submit

Can I customize the reviewer's profile?

Can you support conference xyz?

**Could you analyze the extent to which this paper was likely authored or assisted by a large language model (LLM)?**

# A few-shot LLM-Prompting Pilot

provide a more insightful analysis of the component interactions in their rebuttal, particularly the failure case of the channel-mixing plus patching model.

## Confidence Level

4 High Confidence: The reviewer is confident in their assessment. They are familiar with the relevant literature, though they may not be a leading expert in the specific subfield.

## GenAI Content Analysis:

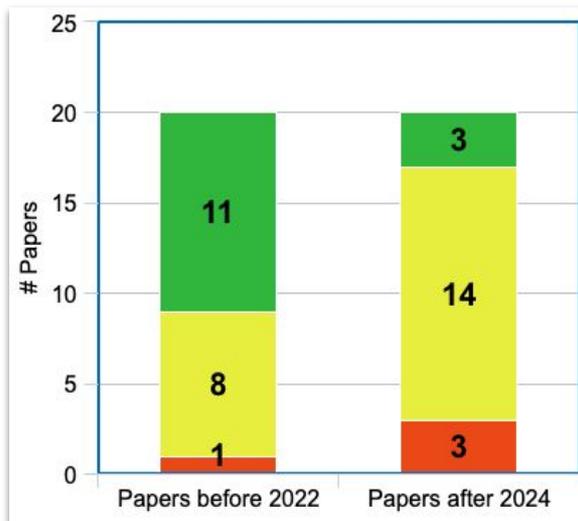
Partial/Moderate: Likely phrasing or structural traits at places.

The paper uses very standard, clear, and formulaic language typical of academic writing in machine learning, which can sometimes overlap with the style of high-quality LLMs. For instance, sentences are often structured as "We propose X, which is based on Y and Z" (Abstract) or "This technique is conceptually simple: ..." (Section 3.2). Phrases like "We would like to highlight," "In essence," and "As evident in" are common connectors. While this is not definitive proof, the overall text is extremely clean and follows a very logical, almost template-like flow from motivation to method to results, which could suggest assistance from an advanced writing tool. However, the core ideas and the experimental design are clearly human-driven, so any AI involvement would likely be at the level of polishing and phrasing.

Reviewed on November 08, 2025  Gemini-2.5-Pro

Your rating: 

None/Minimal  Partial/Moderate  Extensive/Intensive 





# Thanks for listening!



<https://cspaper.org>



**Get 10 Free CSPaper Reviews!**

Email [support@cspaper.org](mailto:support@cspaper.org) with "ICDM2025" as the email subject. (Valid till Nov 30th)