LLM-Agent AI for Cancer Fact-Checking on Online Platforms

Mohammed Al-Garadi[1], Shui Mauser[1]
[1]Department of Biomedical Informatics, Vanderbilt University Medical Center Nashville TN

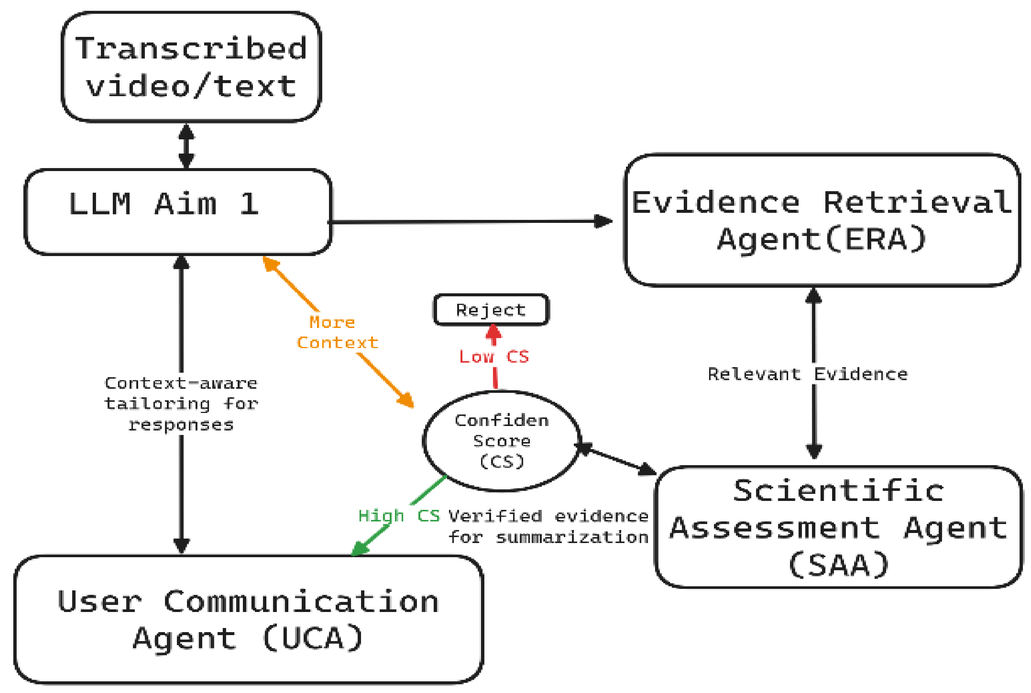
INTRODUCTION

- 30–5Cancer remains a leading cause of death in the U.S., with ~2M new cases and >600K deaths estimated in 2023.
- 0% of cases are preventable by reducing modifiable risk factors (tobacco, obesity, inactivity, diet, alcohol, UV exposure).
- Social media both spreads valuable cancer information and fuels cancer-related misinformation (CRM).
- Studies show ~77% of CRM posts contain harmful content, delaying treatment or promoting unsafe practices.
- While human-led corrections exist, scaling requires automated CRM detection, verification, and response systems—with built-in human oversight.

METHODS

We analyzed 1,000 YouTube videos (from 71,690 total) to develop an automated multi-agent system for detecting and verifying cancer-related misinformation. Audio was transcribed, text normalized, and a zero-shot Mistral-70 LLM applied for detection and claim generation. Verification used three agents— evidence retrieval agent(ERA), scientific assessment(SAA) via confidence scoring, and user communication for corrective reply. See Figure below for framework overview.

Figure 1 Multi-Agent Framework for Verifying and Responding to CRM



RESULTS

- The multi-agent system achieved strong performance across all components (Table below).
- The LLM detection model accurately identified CRM (Precision = 0.88, Recall = 0.70, F1 = 0.77).
- ERA sourced relevant biomedical literature (F1 = 0.68).
- The SAA verified health claims with 0.86 accuracy.
- The UCA corrective messages with 0.90 accuracy.

Component	Key Outcome
Detection (LLM)	**Precision = 0.88, Recall = 0.70, F1 = 0.77**
ERA	**Precision = 0.79, Recall = 0.60, F1 = 0.68**
SAA	**Fact act verification = 0.86**
UCA	**Corrective accuracy = 90%**

Discussion

Integrating LLMs with a multi-agent framework enables scalable, automated detection and correction of CRM on social media.
The system reduces manual workload, ensures evidence-based verification, and supports timely, accurate public health communication.
Each agent contributes a specialized function—retrieval, assessment, and communication—creating an end-to-end pipeline for misinformation management. Future work will refine model accuracy, expand to other platforms, and incorporate user-centered design for broader impact.

Limitations	Future Work
Single-platform dataset (YouTube only)	Expand to TikTok, Instagram, and other platforms
Limited manual annotation sample	Increase dataset size and annotation diversity
Zero-shot model not fine-tuned for domain	Fine-tune LLM on cancer-related corpora
Automated response not user-tested	Evaluate user trust and message comprehension

CONCLUSION

The multi-agent LLM framework was designed to detect, verify, and address cancer-related misinformation through automated, evidence-informed analysis. While current work focuses on YouTube data, future directions include expanding to other platforms, refining model performance, and incorporating user-centered evaluation.

VANDERBILT UNIVERSITY MEDICAL CENTER
Department of Biomedical Informatics

Mohammed.a.al-garadi@vumc.org