Multi-Agent Retrieval-Augmented Framework for Evidence-Based Counterspeech Against Health Misinformation

Anirban Saha Anik, Xiaoying Song, Elliott Wang, Bryan Wang, Bengisu Yarimbas, Lingzi Hong

University of North Texas

Scan Me to Connect

Scan Me for Full Article

Motivation & Problem

- **Health misinformation** spreads rapidly on social media, causing confusion, distrust in medical institutions, and harmful self-medication.
- **Large Language Models (LLMs)** are powerful for generating counterspeech but often hallucinate, lack grounding in evidence, and provide inconsistent responses.
- **Existing RAG approaches** usually rely on either static knowledge (e.g., curated medical guidelines) or dynamic knowledge (e.g., live web search), limiting adaptability and accuracy.
- **Users struggle** to access reliable evidence quickly, making effective counterspeech difficult at scale.

Problem: How can we design an **LLM-based system** that generates **factually accurate, polite, and user-trusted counterspeech** by leveraging both **static and dynamic evidence**?

Key Contributions

[1] Multi-Agent RAG Framework
A modular pipeline with specialized agents for retrieval, summarization, generation, and refinement of counterspeech.

[2] Integration of Static & Dynamic Evidence
Combines curated *medical guidelines (static)* with *real-time web evidence (dynamic)* for both reliability and adaptability.

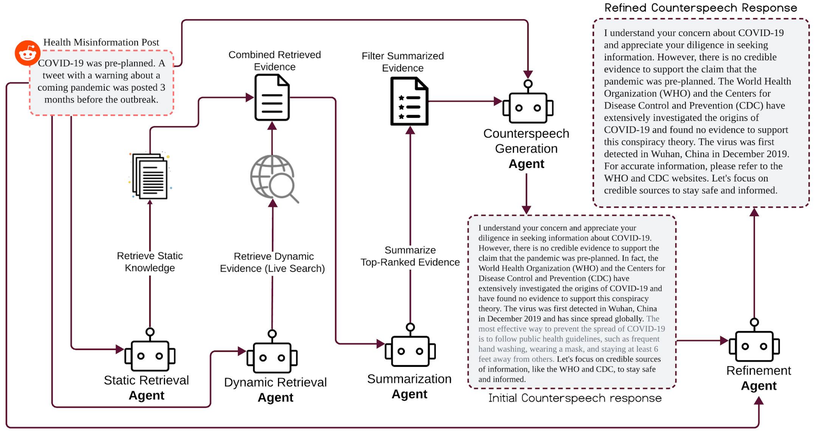
[3] Curated Health Misinformation Dataset
New Reddit dataset on COVID-19, Influenza, and HIV, annotated with classifier assistance + expert validation.

Dataset

- **Source:** Reddit posts & comments collected via PRAW API from high-engagement subreddits (e.g., *r/science, r/health, r/conspiracy*).
- **Topics Covered:** COVID-19, Influenza (Flu), HIV
- **Annotation Process:**
  - 5 annotators with information science background.
  - Comprehensive guidelines referencing **CDC, WHO, NIH**.
  - Human + classifier-assisted labeling, with expert validation.
  - Cohen's $\kappa \geq 0.67 \rightarrow$ substantial agreement.
- **Classifier:** Fine-tuned **RoBERTa-large** (F1 = 0.76) to expand labeled set.
- **Final Dataset:** 1,161 posts labeled as health misinformation.

Scan Me for Dataset

Multi-Agent Retrieval-Augmented Framework



Multi-agent RAG pipeline combining static & dynamic evidence for reliable counterspeech

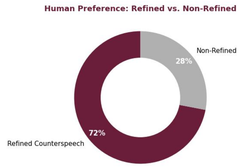
Cross-Topic Evaluation Results

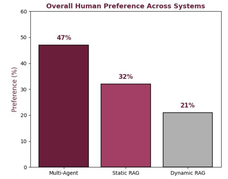
Method	Category	Politeness	Relevance	Informativeness	Factual Accuracy
LLM w DP	COVID-19	0.45 (0.27)	0.74 (0.17)	0.75 (0.14)	0.78 (0.21)
	HIV	0.57 (0.31)	0.78 (0.12)	0.76 (0.06)	0.75 (0.27)
	Influenza	0.42 (0.20)	0.81 (0.10)	0.75 (0.08)	0.70 (0.24)
	Average	**0.48 (0.26)**	**0.78 (0.13)**	**0.75 (0.09)**	**0.74 (0.24)**
LLM w PE	COVID-19	0.78 (0.20)	0.69 (0.12)	0.76 (0.06)	0.81 (0.24)
	HIV	0.83 (0.17)	0.71 (0.12)	0.75 (0.00)	0.79 (0.17)
	Influenza	0.86 (0.13)	0.73 (0.08)	0.76 (0.06)	0.71 (0.25)
	Average	**0.82 (0.17)**	**0.71 (0.11)**	**0.76 (0.04)**	**0.77 (0.22)**
Static RAG	COVID-19	0.75 (0.18)	0.67 (0.12)	0.78 (0.08)	0.89 (0.17)
	HIV	0.81 (0.13)	0.70 (0.17)	0.76 (0.06)	0.68 (0.35)
	Influenza	0.75 (0.16)	0.73 (0.10)	0.76 (0.06)	0.75 (0.26)
	Average	**0.77 (0.16)**	**0.70 (0.13)**	**0.77 (0.07)**	**0.77 (0.26)**
Dynamic RAG	COVID-19	0.84 (0.23)	0.62 (0.20)	0.70 (0.10)	0.88 (0.19)
	HIV	0.91 (0.12)	0.70 (0.14)	0.73 (0.08)	0.84 (0.27)
	Influenza	0.82 (0.23)	0.69 (0.12)	0.74 (0.10)	0.66 (0.26)
	Average	**0.86 (0.19)**	**0.67 (0.15)**	**0.72 (0.09)**	**0.79 (0.24)**
Multi-Agent (Ours)	COVID-19	0.92 (0.05)	0.68 (0.17)	0.78 (0.08)	0.84 (0.19)
	HIV	0.86 (0.14)	0.73 (0.09)	0.78 (0.08)	0.91 (0.17)
	Influenza	0.93 (0.14)	0.71 (0.10)	0.74 (0.06)	0.79 (0.15)
	Average	**0.90 (0.11)**	**0.71 (0.12)**	**0.77 (0.07)**	**0.85 (0.17)**

Table 1. *Multi-Agent achieves top politeness and strong factual accuracy, with balanced performance across COVID-19, HIV, and Influenza.*

**Direct Prompt (DP), Prompt Engineering (PE)*

Human Evaluation





Key Findings

Humans strongly prefer refined counterspeech (72%), with the Multi-Agent framework chosen most often (47%).

Multi-Agent vs. Baseline Performance

Method	Politeness	Relevance	Informativeness	Factual Accuracy
LLM w DP	0.44 (0.26)	**0.70 (0.14)** ↑	0.77 (0.11)	0.81 (0.21)
LLM w PE	0.84 (0.15)	0.65 (0.15)	0.75 (0.07)	0.83 (0.20)
Static RAG	0.80 (0.15)	0.65 (0.16)	0.77 (0.08)	0.81 (0.24)
Dynamic RAG	0.86 (0.16)	0.64 (0.15)	0.73 (0.11)	0.83 (0.19)
Multi-Agent (Ours)	**0.88 (0.14)** ↑	**0.70 (0.13)** ↑	**0.78 (0.13)** ↑	**0.86 (0.19)** ↑

Table 2. *Comparison of counterspeech generation methods across four evaluation metrics. Our* ***Multi-Agent framework*** *achieves the best overall performance, with notable gains in politeness, informativeness, and factual accuracy while maintaining strong relevance.*

**Direct Prompt (DP), Prompt Engineering (PE)*

Ablation Study: Impact of Agents & Prompting

Method	Prompt	Politeness	Relevance	Informativeness	Factual Accuracy
Multi-Agent w/o SA w/o RF	CoT	0.63 (0.24)	0.69 (0.13)	0.76 (0.17)	0.84 (0.18)
Multi-Agent w/o SA w/o RF	Guided	0.80 (0.19)	0.67 (0.19)	0.78 (0.13)	0.85 (0.19)
Multi-Agent w/o RF	CoT	0.65 (0.22)	0.69 (0.13)	0.77 (0.09)	0.86 (0.18)
Multi-Agent w/o RF	Guided	0.79 (0.17)	0.70 (0.13) ↑	0.82 (0.15) ↑	0.87 (0.19) ↑
Multi-Agent	CoT	0.74 (0.19)	0.67 (0.14)	0.77 (0.09)	0.87 (0.19) ↑
Multi-Agent (Ours)	**Guided**	**0.88 (0.14)** ↑	**0.70 (0.13)** ↑	0.78 (0.13)	**0.86 (0.19)**

Table 3. *Ablation results showing the impact of Summarization and Refinement Agents. Both modules improve politeness, informativeness, and factual accuracy, with Guided prompting yielding the best overall performance.*

**Summarization Agent (SA) and Refinement Agent (RF)*

Cross-Platform Generalization

Method	Politeness	Relevance	Informativeness	Factual Accuracy
LLM w DP	0.44 (0.26)	0.70 (0.14)	0.77 (0.11)	0.81 (0.21)
LLM w PE	0.81 (0.18)	0.77 (0.11)	0.77 (0.07)	0.94 (0.12)
Static RAG	0.86 (0.14)	0.77 (0.12)	0.78 (0.09)	0.95 (0.12)
Dynamic RAG	0.85 (0.16)	0.77 (0.13)	0.76 (0.10)	0.92 (0.16)
Multi-Agent (Ours)	**0.89 (0.11)** ↑	**0.78 (0.10)** ↑	**0.83 (0.12)** ↑	**0.96 (0.12)** ↑

Table 4. *Evaluation on the* ***MisinfoCorrect dataset*** *(Twitter/X COVID-19 claims). Our Multi-Agent framework achieves the highest performance across all metrics, showing strong cross-platform generalization beyond Reddit.*

**Direct Prompt (DP), Prompt Engineering (PE)*

Conclusion

Our *multi-agent RAG framework* effectively integrates static and dynamic evidence, achieving superior politeness, informativeness, and factual accuracy over baselines. Human evaluation further confirms strong preference for refined counterspeech. Future work will focus on expanding knowledge sources, improving efficiency, and extending the framework to other misinformation domains with human feedback integration.

Future Work

[1] If the model synthesizes static, dynamic, and internal LLM knowledge, how should it handle contradictions across evidence? How can it decide which pieces to trust?

[2] How could human feedback be integrated into this system?

[3] How can multi-agent pipelines be optimized to balance accuracy, speed, and computational cost at scale?