

Beyond the Crowd: LLM-Augmented Community Notes for Governing Health Misinformation



Jiaying Wu^{*1}, Zihang Fu^{*1}, Haonan Wang¹, Fanxiao Li², Min-Yen Kan¹
^{*} Equal Contribution ¹ National University of Singapore ² Yunnan University

Post Flagged as Potentially Misleading:

 X User @X_User
 Homeopathy has been demonized for no one's benefit except Big Pharma. Thankfully, [@SecKennedy promises to speak to @MartyMakary to reintroduce homeopathy as an appropriate solution for a variety of health concerns](#). We look forward to this change!

Community Note Creation:

 There is little to no proof on the effectiveness of Homeopathy medicines. It is not recognized as a valid system, or outright banned in many countries.
<https://www.nhs.uk/conditions/homeopathy/#:~:text=Does%20homeopathy%20work%3F,treatment%20for%20any%20health%20condition>
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6399603/>
https://en.wikipedia.org/wiki/Regulation_and_prevalence_of_homeopathy

Helpfulness Rating:

 Helpful. It provides important context.
 Helpful. It cites high-quality sources.
 ...
 Status: Currently Rated Helpful

Motivation

Community Notes, X's crowd-sourced misinformation governance system, lets users **flag** misleading posts, **add** contextual notes, and **vote** on note helpfulness.

However, crowd wisdom takes time – and often never arrives.

From 30.7K notes on 25.5K health-related posts (2021–2025), we observe:

- ⌚ **17.6-hour median delay** before first Helpful/Not Helpful status is assigned
- 🚫 **87.9% of notes** never reach a final status ("Not Enough Ratings")

Pct.	Post Published → First Note	First Note → First Status
25%	3.4	3.6
50%	10.4	7.2
75%	23.0	18.4
90%	49.1	76.4

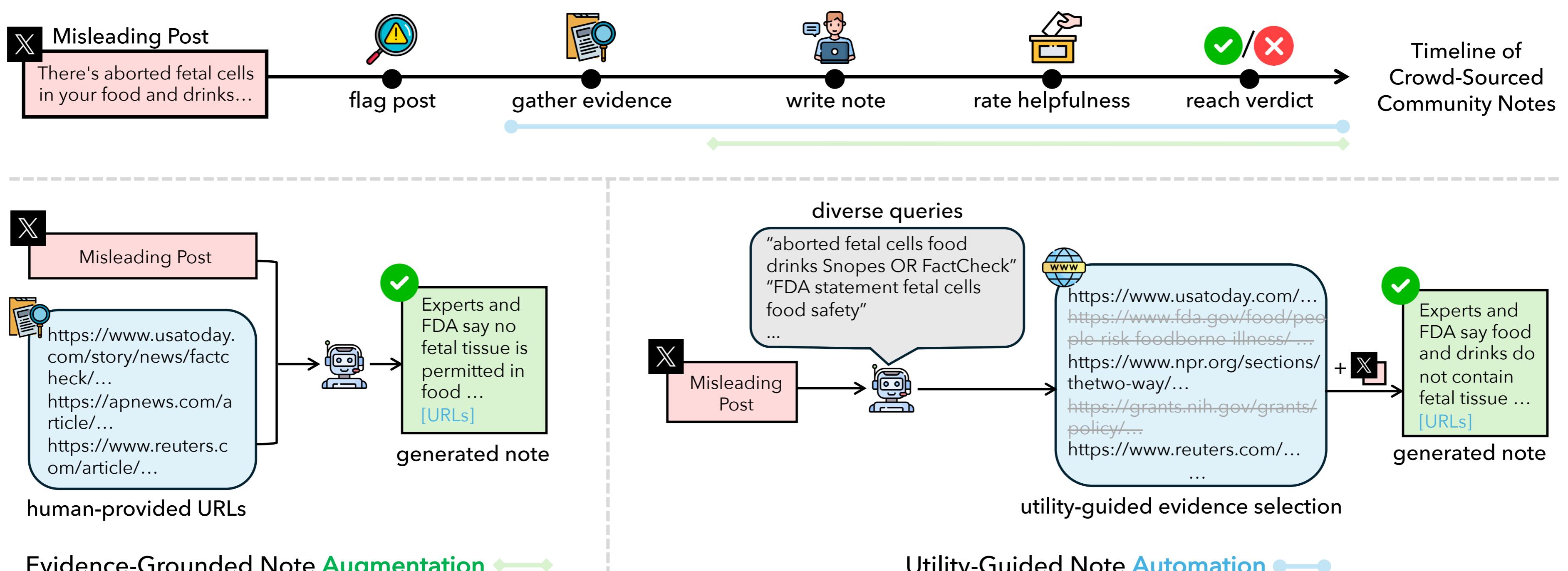
From Crowd-Sourced to CrowdNotes+: LLMs in Action

CrowdNotes+ bridges human curation and LLM automation, augmenting the crowd workflow with:

(1) **utility-guided evidence retrieval w/ diverse queries**, and (2) **evidence-grounded note synthesis**.

Hierarchical evaluation schema: **relevance** (context fit) → **correctness** (factual fidelity) → **helpfulness** (reader utility).

Result: faster, more reliable, and transparent community-based governance.



Effectiveness of Proposed Approach

Data. 1,268 human-curated health notes with crowd-confirmed helpfulness statuses (Helpful / Not Helpful)

Evaluation. **LLM-as-a-Judge** (GPT-4.1) for relevance + correctness; **domain-tuned 7B LLM** for helpfulness judgment

E1: Voting Loophole Surfaced

Our hierarchical evaluation uncovers systematic false positives in crowd votes: many human-written "Helpful" notes fail relevance or correctness (**-11.7% / -14.0% vs. crowd**), mainly from unsupported, misinterpreted, or over-generalized claims.

E2: CrowdNotes+ Surpass Humans

Across both augmentation & automation setups, CrowdNotes+ with **backbone LLMs > 14B** (e.g., o3, Grok-4, GPT-4.1, MedGemma-27B) generate notes that are more helpful than human-authored ones under our schema.

E3: LLM-Selected Evidence Wins

Overall, utility-guided, query-diversified retrieval with LLMs produces higher-quality evidence than human-provided sources: o3 wins **65.9%**, MedGemma-27B wins **57.6%**, based on LLM-judged pairwise results.