THE WEB CONFERENCE ACM

THE ACM WEB CONFERENCE 2026

29 June - 3 July 2026
Dubai, United Arab Emirates

www2026.thewebconf.org

Virtual EVENT

4488 - LLMs Killed Q&A Stars? Analyzing the Impact of LLM-Generated Answers on an Online Q&A Platform

Dongwon Shin

MOTIVATION

- **LLMs are reshaping Q&A platforms**
  - Users can now get direct answers to their questions from LLMs.
- **LLM-generated answers (LGAs) are rising**
  - They may be helpful, but can also be misleading or low-quality.
- **We empirically study this shift at scale** on *Naver Knowledge iN*, South Korea's largest online Q&A platform.

RESEARCH QUESTIONS

- **RQ1)** How **prevalent** are LGAs on the Q&A platform?
- **RQ2)** How do LGAs **differ from human-written answers**?
- **RQ3)** How do LGAs affect **users** and the **Q&A ecosystem**?

METHOD

- **Build a 100K labeled dataset**
  - Crawl 50K human-written answers (HWAs) before the release of ChatGPT and collect 50K LGAs from 10 LLMs.
- **Evaluate 9 LLM-generated text detectors**
  - The top-3 detectors are ReMoDetect [1], Text Fluoroscopy [2], and DeTeCtive [3].
- **Ensemble the best three methods**
  - Combine the top-3 detectors using a stacking ensemble.

Method	AUC ↑	FPR ↓	FNR ↓
ReMoDetect	0.9986	1.46%	**1.16%**
Text Fluoroscopy	0.9795	7.78%	3.86%
DeTeCtive	0.9231	3.62%	13.98%
Stacking Ensemble	**0.9987**	**0.98%**	1.48%

- **Deploy at scale**
  - Applied the ensemble detector to 1.46M answers from *Naver Knowledge iN*.

LIMITATIONS

- Detected LGAs are a conservative lower bound, as some LGAs may **evade detection**.
- Our findings are **correlational** and may be influenced by unobserved external factors.

REFERENCES

[1] ReMoDetect: Reward Models Recognize Aligned LLM's Generations, NIPS, 2024
[2] Text Fluoroscopy: Detecting LLM-Generated Text through Intrinsic Features, EMNLP, 2024
[3] DeTeCtive: Detecting AI-generated Text via Multi-Level Contrastive Learning, NIPS, 2024

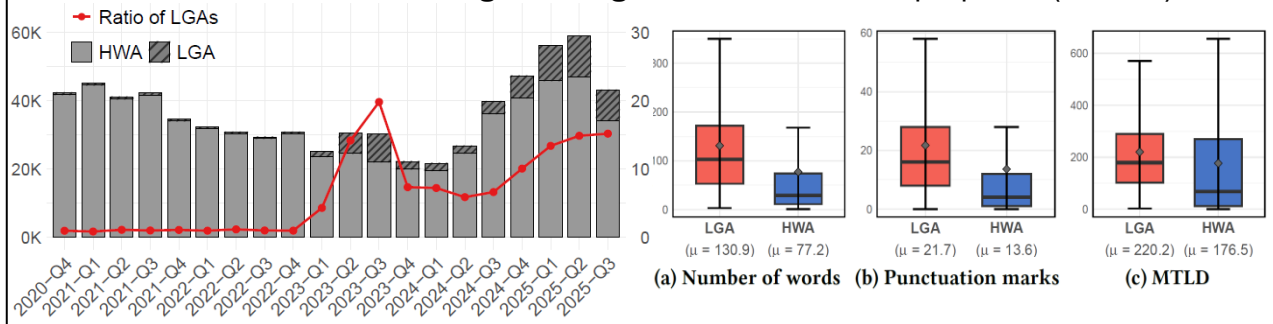
KEY FINDINGS

Finding 1) LGAs are common on Q&A platforms

- **LGAs account for 5.18%** of 1.46M answers, increased sharply after the release of ChatGPT, and peaked at 26.91% in 2023-Q3.

Finding 2) LGAs are longer, richer, and more knowledge-focused

- LGAs are **longer, richer,** and **more lexically diverse** than HWAs.
- LGAs focus more on **knowledge sharing** rather than on other purposes (71.59%).



Finding 3) Users react similarly to LGAs and HWAs, but Q&A usage is shifting

- We observe no significant differences in user feedback (upvotes and downvotes) between LLM-generated and human-written answers.
- Questions are **shifting** from basic information (C1) toward **context-specific (C5)** and **experience-based (C4) topics**.