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

Dongwon Shin
Korea Advanced Institute of Science and Technology
Daejeon, Republic of Korea
godeastone@kaist.ac.kr

Sooel Son
Korea Advanced Institute of Science and Technology
Daejeon, Republic of Korea
sl.son@kaist.ac.kr

Abstract

Online question-and-answer (Q&A) platforms facilitate knowledge exchange through posted questions and answers. Recent advances in large language models (LLMs) have shown their strong capability in generating high-quality answers, leading to a recent surge in LLM-generated answers (LGAs) on Q&A platforms. In this paper, we conduct an in-depth analysis of how LGAs affect *Naver Knowledge iN*, the most popular Q&A platform in South Korea. To this end, we implement nine state-of-the-art LLM-generated text (LGT) detection methods and evaluate their performance on answers collected from *Naver Knowledge iN*. We then build an ensemble detector by stacking the three best-performing LGT detection methods, achieving an AUC of 0.9987 with a false positive rate below 1%. Using this LGA detector, we identify 75,558 LGAs among 1.46M answers. We find that LGAs tend to be longer, use more punctuation marks, and exhibit higher lexical diversity. However, LGAs do not show clear differences in user reactions, such as upvotes, downvotes, or selection rates by questioners. We also find that LGAs are primarily intended for knowledge sharing rather than personal experiences sharing. Finally, we observe a shift in the Q&A platform: questions increasingly move from simple fact-seeking to those involving complex contexts and seeking personal opinions or past experiences.

CCS Concepts

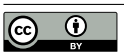
• **Computing methodologies** → **Natural language processing**; **Ensemble methods**; • **General and reference** → **Empirical studies**.

Keywords

LLM-generated text, LLM, Q&A platform, Ensemble

ACM Reference Format:
Dongwon Shin and Sooel Son. 2026. LLMs Killed Q&A Stars? Analyzing the Impact of LLM-Generated Answers on an Online Q&A Platform. In *Proceedings of the ACM Web Conference 2026 (WWW '26), April 13–17, 2026, Dubai, United Arab Emirates.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3774904.3792739

Resource Availability:
The source code of this paper has been made publicly available at https://doi.org/10.5281/zenodo.17291624.



WWW '26, Dubai, United Arab Emirates

ACM ISBN 979-8-4007-2307-0/2026/04
https://doi.org/10.1145/3774904.3792739

1 Introduction

In online Q&A platforms, users post questions, and other users with relevant expertise provide answers. These answers provide valuable knowledge across various domains, such as programming [1, 6], education [2, 5], and general knowledge [3, 4]. These Q&A platforms have been widely adopted, as they allow anyone to readily obtain answers to their questions. For instance, *Naver Knowledge iN* [3], the most popular Q&A platform in South Korea, has recorded over 970M questions and 617M answers since 2002.

However, the recent advancements of LLMs have affected the ecosystem of online Q&A platforms. As LLMs show outstanding performance in accurately answering users' questions [27, 31], users increasingly turn to LLM services instead of posting their questions on Q&A platforms [20, 46].

Previous studies have analyzed how LLM services affected Q&A platforms [14, 20, 30]. Specifically, Rio-Chanona *et al.* [20] examined the number of questions posted on Stack Overflow after the release of ChatGPT, demonstrating that user activities have decreased by 25% compared to counterfactual platforms. Burtch *et al.* [14] reported similar trends, with a 12% decline in Stack Overflow's daily web traffic after the release of ChatGPT. Furthermore, the number of questions related to simple and self-contained programming languages (e.g., Python, JavaScript, ReactJS) showed a substantial decline, whereas questions on more complex context-dependent frameworks (e.g., Spring, Spring-boot) exhibited marginal differences. Another line of research demonstrates that a variety of online services are suffering from LLM-generated content. For instance, LLMs can be used to generate fake news [24, 50], fake reviews on e-commerce sites [35, 49], and spam and phishing on social networks [8, 45]. These misuses of LGTs have led online Q&A platforms to prohibit them [7], instead of seeking ways to coexist.

However, none of the previous studies have examined the prevalence and characteristics of LLM-generated answers (LGAs) on Q&A platforms, or the associated user reactions and behaviors. Considering the overlapping nature between the functions of LLMs and Q&A platforms (i.e., posing questions and providing answers), it is important to understand how LLMs influence the Q&A platform ecosystems, paving the way for their potential symbiosis.

Our contributions. In this work, we conduct an empirical study on *Naver Knowledge iN* to better understand the characteristics of LGAs and user responses involving those LGAs. To identify LGAs, we first implement and evaluate nine state-of-the-art LGT detection methods. We then ensemble the three best-performing methods to minimize the false positive rate, thus implementing the LGA detector. Our ensemble-based detector achieves an AUC of 0.9987 while keeping the FPR below 0.01.

For our large-scale analysis, we collect 1.46M answers from the *Naver Knowledge iN* platform and apply our ensemble detector to identify LGAs. The detection results demonstrate that 75,558 answers are LGAs, showing rapid growth from 2023. To assess their impact on the Q&A platform, we performed linguistic, user profile, and sentiment analyses.

We find that LGAs share several common characteristics: they are longer, contain more punctuation, and exhibit higher lexical diversity. Moreover, users posting LGAs tend to disclose their identities and are more likely to hold expertise badges, showing their intention to promote themselves through identifiable contributions. We also analyze user reactions—upvotes, downvotes, and answer selection rates by questioners. We find no clear differences between LGAs and human-written answers (HWAs), but comments on LGAs show more positive sentiment than those on HWAs. These findings suggest that LGAs have potential in eliciting favorable user responses, despite the prevailing negative perceptions of LLM usage on online Q&A platforms.

Moreover, we classify the intentions behind LGAs to identify posting motivations, and we analyze the categories of most-viewed questions by year to examine how question trends have shifted since the release of LLM services. The analysis results show that LGAs are focused more on *knowledge sharing* compared with HWAs, and that the most-viewed question categories have shifted from *basic information* requests to *context-specific troubleshooting* and *opinion/experience*. That is, questioners increasingly seek personal hands-on experiences on complex and context-dependent issues, an area where LLMs provide limited assistance.

In summary, we identify LGAs by ensembling recent LGT detection methods and conduct an in-depth analysis of their impact on *Naver Knowledge iN*. Our analysis results provide valuable insights into LGAs on the online Q&A platform, which have remained largely understudied. We find that LGAs are increasingly prevalent and exhibit distinct characteristics and purposes, focusing on knowledge delivery and information sharing. Also, the rise of LLM services has shifted user questions toward seeking more experience-based and personalized perspectives on complex issues. We hope this study offers data-driven insights into the current status of LGA usage on Q&A platforms and informs future directions for coexistence between human- and LLM-generated content.

2 Background and Related Work

2.1 Online Q&A Platforms

In the past, when people had questions, they generally found answers from books, experts, or by searching the Web. However, some questions are difficult to resolve in this way because they require complex contextual understanding (e.g., *After installing a driver on Windows, my computer shows a blue screen. How can I tell if it is a hardware or software problem?*) or rely on personalized experience and know-how (e.g., *Which laptop works best for programming if I have to carry it every day?*).

Online Q&A platforms rapidly emerged to fill these gaps by sharing the collective knowledge with users. People are willing to answer questions not only to share their knowledge but also to gain a reputation and promote themselves on the platforms. For instance, on *Naver Knowledge iN*, the most popular Q&A platform in South Korea, users earn *eXpert* badges only after a required number of their answers have been selected and additional qualifications (e.g., certifications) are met. They can further monetize their expertise by selling lectures or consulting services. Users with *eXpert* badges tend to provide high-quality answers, since highly upvoted answers appear at the top of question pages and help promote their paid services.

Meanwhile, beyond human-driven online Q&A platform, recent studies have addressed diverse Q&A tasks by leveraging LLMs [29, 32, 54]. These approaches improve Q&A reliability through structured summarization in retrieval-augmented generation (RAG) [29], knowledge graph-based reasoning [54], and the integration of visual information [32].

2.2 Detecting LLM-Generated Text

With the advancement of LLMs, LGTs have become prevalent across various services, raising concerns about integrity and safety. For instance, attackers leverage LGTs to spread misinformation [15, 28], conduct online fraud schemes [10, 39], and commit plagiarism [13, 34]. Therefore, distinguishing LGTs from human-written texts (HWTs) becomes important for trustworthy web platforms.

Formally, let $D : \mathcal{X} \to \{0, 1\}$ be a detector that maps an input text $x \in \mathcal{X}$ to a binary label, where 1 indicates an LGT and 0 indicates an HWT. There are two representative approaches to detecting LGTs: (1) zero-shot methods and (2) training-based methods.

Zero-shot methods. Zero-shot methods [11, 21, 23, 26, 40, 51, 53] take only the input text x and determine whether x is an LGT or an HWT without training. A simple zero-shot approach is to leverage a model's token-level logits to compute specific metrics, such as log-likelihood, log-rank, and rank [21, 26]. If the average of these metrics is higher than a chosen threshold, the input text was generated by an LLM because its wording closely follows the training distribution. For example, given a language model f_θ, a tokenized text $x = (x_1, \ldots, x_T)$, and a threshold τ, we define the average log-likelihood $\bar{\mathrm{L}}$ and the detection result $D_\tau(x)$ as follows:

$$\bar{\mathrm{L}}(x;\theta) = \frac{1}{T}\sum_{t=1}^{T} \log f_\theta(x_t \mid x_{<t}), \quad D_\tau(x) = \mathbf{1}\left\{\bar{\mathrm{L}}(x;\theta) \geq \tau\right\}.$$

Mitchell *et al.* [40] proposed DetectGPT, a perturbation-based detector. The key idea is that, for LGTs, the model's score at the original text $f(x)$ is consistently higher than at randomly perturbed texts $f(\tilde{x})$. Randomly perturbed texts are generated by masking a small fraction of tokens and refilling them with a mask-filling model such as T5 [43]. To formalize this, they define the *perturbation discrepancy* $\mathbf{d}(x, p_\theta, q)$ as follows:

$$\mathbf{d}(x, f_\theta, q) \;\triangleq\; \log f_\theta(x) \;-\; \mathbb{E}_{\tilde{x}\sim q(\cdot|x)}[\log f_\theta(\tilde{x})]$$

where $q(\cdot \mid x)$ denotes the perturbation function that generates a distribution over $\tilde{x}$. The quantity of $\mathbf{d}(x, p_\theta, q)$ is typically positive for LGTs, and near zero for HWTs. Bao *et al.* [11] proposed Fast-DetectGPT, an extension of DetectGPT. They devise a more efficient sampling procedure that avoids generating many perturbed variants. Fast-DetectGPT improves detection accuracy by 75% and increases detection speed by 340 times.

Training-based methods. Training-based methods [12, 16, 33, 36, 38, 48, 52] train classifiers to learn the intrinsic patterns distinguishing LGTs from HWTs. GPT-Sentinel [16] trains two classifiers based on RoBERTa and T5 architectures. For RoBERTa-Sentinel, a two-layer MLP is attached to the final [CLS] hidden state of a frozen RoBERTa encoder. For T5-Sentinel, since T5 is a sequence-to-sequence model, the task is framed as generating a label token. During training, the model is fine-tuned to produce the target token *positive* or *negative*. During inference, the decoder's vocabulary is restricted to these two tokens, and the label with the higher probability is selected.

Text Fluoroscopy [52] identifies the encoder layer of gte-Qwen whose projection into the vocabulary space yields the most divergent token distribution. Using the hidden states from that layer as input features, they train a three-layer MLP classifier.

DeTeCtive [22] adopts contrastive learning for LGT detection. It extracts embeddings from the training data using a SimCSE-RoBERTa encoder and optimizes a contrastive objective that pulls same-class embeddings together and pushes different-class embeddings apart. After training, the embeddings are stored in a feature database; at inference, test texts are embedded, and a K-NN search over the database is used to make the final prediction.

Lee *et al.* proposed ReMoDetect [33], which uses a *reward model*—an LLM trained to capture human preferences—to distinguish LGTs from HWTs. Specifically, the *reward model* assigns higher preference scores to LGTs than to HWTs. Based on this, they construct a mixed LGT-HWT dataset and further train the *reward model* with the following ordering constraint:

$$r_\phi(\text{LGT}) \;>\; r_\phi(\text{MIX}) \;>\; r_\phi(\text{HWT})$$

where $r_\phi(\cdot)$ denotes the *reward model*'s preference score, parameterized by ϕ. ReMoDetect attains superior AUROC with fewer model parameters and lower detection latency.

3 Motivation and Research Questions

3.1 Motivation

A vast volume of prior research reports declining user activities on online Q&A platforms as users increasingly turn to LLM services [20, 46]. However, most work focuses on migration from Q&A platforms to LLM services, rather than on how LGTs are used within the platforms themselves.

Typically, LGTs within the online platform can raise integrity concerns. For instance, the U.S. Federal Trade Commission (FTC) has moved to ban fake reviews and testimonials generated by AI [17], and Amazon restricts authors from publishing more than three books per day due to the surge of AI-generated content [18].

In a similar vein, Stack Overflow—the most popular Q&A platform for programming—banned posting content generated by generative AI models (i.e., LGTs) [7]. Their concerns include (1) fabrication of false or misleading information, (2) excessive noise that obscures useful content, and (3) additional risks (e.g., security, optimization). However, recent advances in LLMs have improved accuracy, concision, and safety in their responses [19, 41], partially addressing such concerns. This suggests that online Q&A platforms may require well-specified policies for accommodating LGAs, rather than banning them across the board.

Table 1: Nine LLMs used for generating LLM-generated answers (LGAs) and their release dates.

Family	Model	Release Date
GPT	gpt-3.5-turbo	2023-03
	gpt-4.1	2025-04
	gpt-5-nano	2025-08
Gemini	gemini-1.5-flash-002	2024-09
	gemini-2.0-flash	2025-02
	gemini-2.5-flash	2025-06
Llama	Llama-3-Instruct-Turbo	2024-04
	Llama-4-Scout-17B-16E-Instruct	2025-04
	Llama-4-Maverick-17B-128E-Instruct-FP8	2025-04

To this end, we first study the presence and characteristics of LGAs within a Q&A platform. We then further examine user experience with LGAs, their underlying purposes, and recent shifts in question trends, taking a first step toward how users produce and consume LGAs.

3.2 Research Questions

To conduct an in-depth analysis of how LGAs affect Q&A platforms, we present five research questions. These research questions focus on (1) analyzing the characteristics of LGAs on the Q&A platform (**RQ1, RQ2**), and (2) examining user behaviors (**RQ3, RQ4, RQ5**).

RQ1. How prevalent are LGAs on the online Q&A platform? Before analyzing the impact of LGAs on the online Q&A platform, we first identify how many answers are generated by LLMs. For this, we employ state-of-the-art LGT detection methods and examine how the proportion of LGAs changes over time.

RQ2. What are the characteristics of LGAs? If LGAs are prevalent, we wonder whether they have common characteristics in their linguistic patterns, lexical diversity, and author profiles.

RQ3. How do users react to LGAs? Kabir *et al.* [30] have shown that people prefer human answers to ChatGPT answers over 65% of the time. We investigate whether this tendency persists in our target Q&A platform by examining the number of upvotes and downvotes, the rate of answers selected by the questioners, and the sentiment analysis of comments on answers.

RQ4. What is the purpose of LGAs? There exist diverse objectives in posting answers on Q&A platforms, such as knowledge sharing, advertising, and sharing personal experiences. We classify the purpose of LGAs and assess user preferences for each purpose.

RQ5. Why do users still use Q&A platforms instead of LLM services? Although recent LLM services provide direct, high-quality answers, users still post thousands of questions and tens of thousands of answers on *Naver Knowledge iN* everyday. To investigate this, we classify the most-viewed question categories in recent years, track which categories dominate over time, and analyze which types of questions elicit user interactions.

4 Identifying LGAs and HWAs

Before analyzing the impact of LGAs on the online Q&A platform, we need to identify which answers are generated by LLMs and which are written by humans.

Table 2: Performance of LGT detection methods categorized into zero-shot and training-based approaches. We use bold to highlight the best-performing method and underlines to indicate the second- and third-best.

Method	Acc.	Prec.	Rec.	FPR (↓)	FNR (↓)	AUC
Log-likelihood	77.04%	73.70%	84.16%	30.08%	15.84%	0.8463
Log-rank	75.98%	73.05%	82.40%	30.44%	17.60%	0.8341
Rank	58.08%	72.85%	25.86%	9.65%	74.14%	0.5852
DetectGPT	80.58%	79.19%	82.96%	22.16%	14.74%	0.8901
Fast-DetectGPT	81.25%	78.47%	<u>86.14%</u>	23.64%	<u>13.86%</u>	0.8867
GPT-Sentinel	83.75%	90.65%	75.26%	<u>7.76%</u>	24.74%	0.9094
Text Fluoroscopy	<u>94.18%</u>	<u>92.51%</u>	<u>96.14%</u>	7.78%	<u>3.86%</u>	<u>0.9795</u>
DeTeCtive	<u>91.20%</u>	<u>95.96%</u>	86.02%	<u>3.62%</u>	13.98%	<u>0.9231</u>
ReMoDetect	**98.69%**	**98.54%**	**98.84%**	**1.46%**	**1.16%**	**0.9986**

For detecting LGAs, we employ state-of-the-art LGT detection methods [11, 16, 21, 22, 26, 33, 40, 52]. To train and test these methods, we first construct the dataset containing both LGAs and HWAs. Then, using this dataset, we implement nine LGT detection methods and evaluate them.

4.1 Constructing Datasets for LGA Detection

We construct a dataset composed of LGAs and HWAs. To collect HWAs, we crawl *Naver Knowledge iN* and collect 50K answers written before November 2022, the release date of ChatGPT.

For LGAs, we collect data instances from two sources: (1) *Naver-AI* answers and (2) answers to questions generated by various LLMs. For the first source, we crawl 5,000 answers generated by *Naver-AI*, a beta AI-bot service launched in August 2024 that automatically posts AI answers on *Naver Knowledge iN*. For the second, we collect 5,000 questions from *Naver Knowledge iN* and prompt LLMs to generate answers. Since LGAs can be generated by various types of LLMs at different times, we select nine LLMs spanning diverse model sizes and release periods. Table 1 provides detailed information about the LLMs, and Appendix A.1 presents the instruction prompts used for generating LGAs.

As a result, we successfully collected 100K answers (i.e., 50K LGAs and 50K HWAs). We split 90% of them into a training dataset and 10% into a testing dataset.

4.2 Evaluating LGA Detection Methods

LGT detection methods. Recall from Section 2.2 that there are two approaches to detecting LGTs: (1) zero-shot methods and (2) training-based methods. We select log-likelihood, rank, log-rank [21, 26], DetectGPT [40], Fast-DetectGPT [11] as zero-shot methods, and GPT-Sentinel [16], Text-fluoroscopy [52], DeTeCtive [22], and ReMoDetect [33] as the training-based methods. We report accuracy, precision, recall, FPR, and FNR to evaluate each method. The optimal threshold is set at the point where TPR - FPR reaches its maximum value. We also report the threshold-independent metric AUC. Detailed settings for implementing the LGT detection methods are described in Appendix A.5.

Table 2 shows the performance of the LGT detection methods. The second to sixth rows correspond to the zero-shot methods, while the seventh to tenth rows correspond to the training-based methods. Overall, the training-based methods show superior performance over the zero-shot methods. Specifically, ReMoDetect achieves the best performance, with accuracy, precision, recall, and AUC of 98.69%, 98.54%, 98.84%, and 0.9986, respectively.

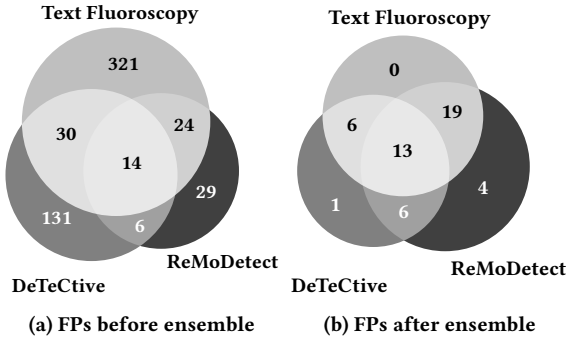
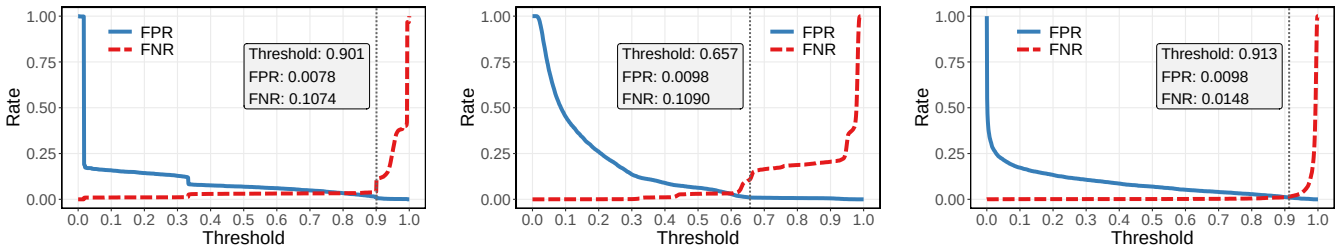


Figure 1: Overlap between FPs by the three LGT detection methods before and after applying stacking ensemble.

Ensemble LGA detection methods. Figure 1a illustrates a Venn diagram of the total number of false positive samples, classified by the three best-performing LGT detection methods: ReMoDetect, Text Fluoroscopy, and DeTeCtive. Although the three methods share some overlap, each method also has exclusive false positive samples. For instance, ReMoDetect yields 29 false positive samples that the other methods correctly classify, demonstrating that relying solely on ReMoDetect is insufficient for accurately identifying LGAs. Furthermore, when we deploy this method for LGA classification, the absence of ground truth makes it difficult to convince whether an answer is an LGT or an HWT.

Therefore, we ensemble multiple LGT detection methods to improve reliability and prevent overfitting to a single method. Specifically, we ensemble ReMoDetect, Text Fluoroscopy, and DeTeCtive using three approaches: (1) hard voting, (2) weighted voting, and (3) stacking.

Hard voting aggregates the predictions of the three methods and determines the final prediction by majority vote. For instance, if the methods predict *LGT*, *LGT*, and *HWT*, the final prediction is *LGT*. Weighted voting is similar to hard voting, but each method is assigned a weight based on its performance. We set the weights according to the AUC of each method. In Stacking, the meta-model is trained on the prediction outputs of the base models. The base models are first trained on the training dataset, and their outputs are then used to train the meta-model. We employ three LGT detection methods as the base models and use a logistic regression classifier as the meta-model. However, this approach can lead to overfitting, as the prediction scores are derived from the same training data used to train the base models. Therefore, we perform five-fold cross-validation on the training dataset. In each fold, the base models are trained on four folds and produce out-of-fold predictions for the remaining fold. Repeating this process across all folds yields the prediction results on the five validation folds, which are then used to train the meta-model. Finally, we retrain the base models on the entire training set and use the test dataset for the final evaluation of both the base and meta-models.



(a) Hard voting (b) Weighted voting (c) Stacking

Figure 2: FPR and FNR of the three ensemble methods while varying the threshold.

(a) Text Fluoroscopy (b) DeTeCtive

Figure 3: Confusion Matrix for LLM Attribution.

To ensure fair comparison, we normalize the output prediction scores to the [0, 1] range before applying hard voting and weighted voting. Stacking does not require normalization because the meta-model automatically learns to handle the different scales across the base model outputs.

Figure 2 presents the results of the three ensemble approaches. For each approach, we report FPR and FNR as the threshold varies. We also draw the threshold line where the FPR first falls below 1%, along with the corresponding FNR value. While hard voting and weighted voting show the FNR values above 10% when FPR is below 1%, stacking achieves an FNR of 1.48% with an AUC of 0.9987. Figure 1b shows how many false positive samples are removed after applying stacking. Stacking reduces the number of exclusive false positives by 321, 130, and 25 for Text Fluoroscopy, DeTeCtive, and ReMoDetect, respectively.

Based on these results, we employ the stacking model to identify LGAs and subsequently analyze their impact on *Naver Knowledge iN* in Section 5. Specifically, if the meta-model's prediction score for a given answer is above 0.913, it is classified as an LGA; otherwise, it is classified as an HWA.

LLM Attribution. We have focused on classifying whether a given text is generated by LLMs or written by humans (i.e., binary classification). We are also interested in whether such LGT detection methods can identify which LLM family—such as GPT, Gemini, or Llama—was used to generate the text.

To this end, we prepare 15,000 samples for each class (i.e., Human, GPT, Gemini, and Llama), resulting in a total of 60,000 samples, which are split 9:1 into training and testing sets. Distinct labels are assigned to each class, and the LGT detection model is trained on this dataset. We adopt Text Fluoroscopy and DeTeCtive for training, because ReMoDetect is tailored to binary classification.

Text Fluoroscopy achieves 91.15 accuracy, 91.48 precision, and 91.15 recall, while DeTeCtive achieves 92.58 accuracy, 92.68 precision, and 92.58 recall. Figure 3 shows the corresponding confusion matrix. The attribution models perform well in distinguishing LGAs from HWAs. They also demonstrate reasonable performance in classifying LGAs generated by major LLM services (i.e., GPT, Gemini, and Llama). However, Text Fluoroscopy and DeTeCtive exhibit relatively lower performance when classifying LGAs generated by Gemini and Llama, respectively. We deploy this model to examine which LLM family is most widely used for LGAs (see Section 5.2).

5 Evaluation

In this section, we address the research questions presented in Section 3.2. To facilitate follow-on research, we release our experiment scripts and the LGA classifier at https://github.com/WSP-LAB/LLMs-Killed-QnA-Stars.

5.1 Experimental Setup

All our experiments are conducted on a machine equipped with two Intel Xeon Platinum 8568Y (2.3 GHz) CPUs, one NVIDIA A100 GPU, and 768 GB of DRAM. The machine runs Ubuntu 22.04 (64-bit).

Dataset. We randomly crawl the 886,556 questions and their corresponding 1,457,709 answers from *Naver Knowledge iN*. The detailed process for crawling is described in Appendix A.2

For an in-depth analysis of LGAs on *Naver Knowledge iN*, we extract multiple features from the answers. Table 3 in Appendix A.2 summarizes each feature with its description, type, and average quantity. Additional dataset statistics, including the number of crawled answers by quarter, the distribution of answer lengths, and the distribution of answer topics, are presented in Appendix A.3.

5.2 RQ1: How Prevalent are LGAs on the Online Q&A Platforms?

Deploying LGA detector. We apply our stacking-based ensemble LGA detector in classifying the answers in our dataset. As a result, 75,558 LGAs are identified, accounting for 5.18% of all answers.

To examine how the distribution of LGAs changes over time, we illustrate (1) the number of LGAs and (2) the ratio of LGAs for each quarter from 2020-Q4 to 2024-Q3 in Figure 4. The ratio of LGAs remains below 1.57% until 2022-Q4, then increases rapidly after 2023-Q1, reaching a peak of 26.91% in 2023-Q3. From 2023-Q4 to

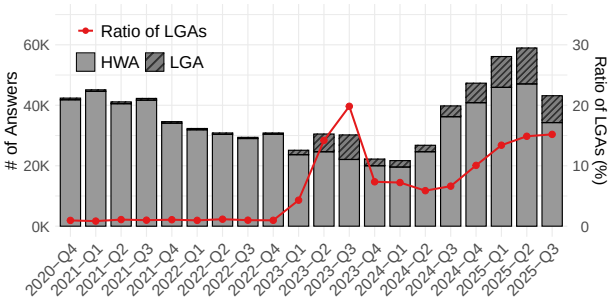
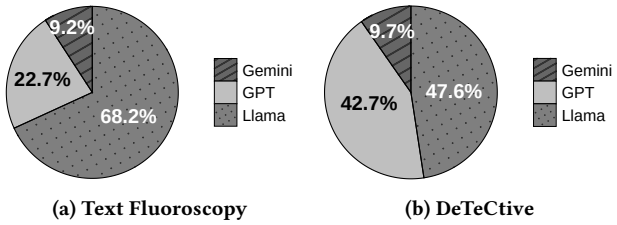


Figure 4: Number and ratio of LGAs aggregated at six-month intervals from 2020-Q4 to 2025-Q3.



(a) Text Fluoroscopy **(b) DeTeCtive**

Figure 5: LLM attribution results.

2025-Q3, the ratio fluctuates between 7.97% and 20.61%. Considering that ChatGPT was released in November 2022, this result indicates that people began using LLMs to post answers on the Q&A platform after the release of LLM services.

LLM attribution in LGAs. Figure 5 shows the proportion of LLM families within the LGAs. Text Fluoroscopy classifies 68.2% of LGAs as Llama and 22.7% as GPT, while DeTeCtive classifies 47.6% as Llama and 42.7% as GPT. These differences stem from the limited classification ability to distinguish between Llama and GPT, as shown in the confusion matrix in Figure 3. Based on this discrepancy, we argue that there are technical difficulties in accurately identifying LLM families, as noted in prior work [9, 25]. In contrast, both models consistently assign the lowest proportion to Gemini, with 9.2% for Text Fluoroscopy and 9.7% for DeTeCtive.

Finding 1: In *Naver Knowledge iN*, 5.18% of answers are generated by LLMs. Following 2023-Q1, the ratio of LGAs increased rapidly, reaching its peak of 26.91% in 2023-Q3.

5.3 RQ2: What are the Characteristics of LGAs?

To characterize LGAs compared to HWAs on the Q&A platform, we analyze both groups in two aspects: linguistic and profile analysis. Linguistic analysis examines the linguistic patterns of answers, such as their length and lexical diversity. Profile analysis explores the profiles of users who wrote the answers. We focus on the anonymity of their profile and the expert badges that they have earned.

Linguistic analysis. Figure 6a and Figure 6b illustrate the number of words and the frequency of punctuation marks in the answers. The punctuation marks include periods, commas, question marks, exclamation marks, brackets, and parentheses.

As a result, LGAs are longer and use more punctuation marks than HWAs. This is because LLMs tend to provide answers with richer context and additional information, often using more punctuation marks, whereas humans usually respond to the question more concisely.

For lexical diversity, we employ the measure of textual lexical diversity (MTLD) [37]. Conventional metrics such as the type–token ratio (TTR) [44] are highly sensitive to text length and thus unsuitable for our dataset since LGAs are generally longer than HWAs. MTLD sequentially tracks the TTR and starts a new segment whenever it falls below a predefined threshold of 0.72, and the MTLD score is computed as the total number of tokens divided by the number of segments. Higher MTLD scores indicate greater lexical diversity. Figure 6c demonstrates that LGAs have higher MTLD scores, indicating that they tend to use a more diverse vocabulary than HWAs. In contrast, HWAs show lower MTLD and a wider interquartile range. These results suggest that LGAs consistently use richer wording when answering questions.

Profile analysis. When posting an answer, users can choose whether to anonymize their profiles. Figure 6d shows the proportion of anonymized profiles for LGAs and HWAs. Among users who wrote LGAs, 19.6% anonymize their profiles, compared to 23.4% of users who wrote HWAs. This indicates that users posting LGAs do not necessarily conceal their identities on the platform when compared to users posting HWAs.

In *Naver Knowledge iN*, users can earn *eXpert* badges upon meeting certain criteria. These users may offer paid consultations or educational services as *eXpert*. They often promote themselves by posting high-quality answers to questions within their areas of expertise. We analyze how often the *eXpert* badge appears in the profiles of users who authored LGAs and HWAs. Figure 6e shows that 14% of LGA authors have the *eXpert* badges, compared to 10% of HWA authors, indicating that users with the *eXpert* badge are more likely to generate LGAs than HWAs.

Finding 2: LGAs tend to generate longer passages with more punctuation marks, exhibiting greater lexical diversity compared to HWAs. Authors of LGAs are also more likely to reveal their profile and to hold an *eXpert* badge than the authors of HWAs.

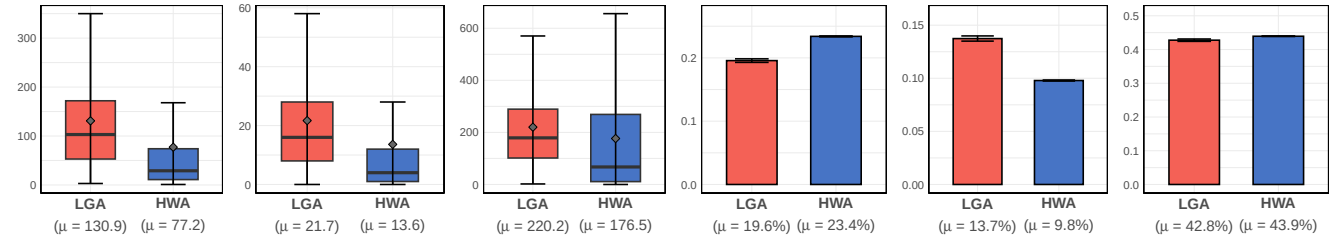
5.4 RQ3: How Do Users React to LGAs?

On the Q&A platform, users may react to answers by giving upvotes, downvotes, or by leaving comments. In addition, questioners can select the best answer they consider most valuable from among all the answers to their question.

We analyze how these reactions vary between LGAs and HWAs. To this end, we examine differences in the selection ratio, upvote counts, and downvote counts across the two groups. Furthermore, we perform sentiment analysis on the comments to answers using KoBERT [47], a BERT model pre-trained on Korean text.

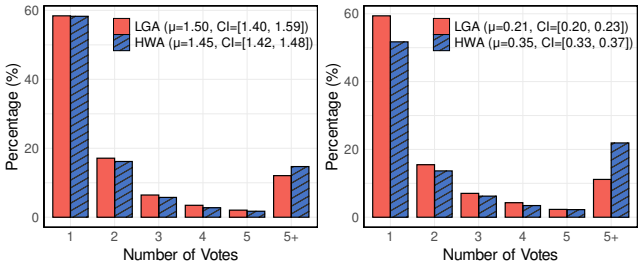
Selected answers and votes. The questioners for our collected answers selected HWAs slightly more often than LGAs, as shown in Figure 6f. However, the difference is marginal, with 42.8% for LGAs and 43.9% for HWAs. This result suggests that questioners do not exhibit a distinct preference between LGAs and HWAs.

Similarly, the numbers of upvotes and downvotes show little difference. On average, LGAs receive 1.50 upvotes and 0.21 downvotes,



(a) Number of words **(b) Punctuation marks** **(c) MTLD** **(d) Anonymize** **(e) *eXpert* badges** **(f) Selected answers**

Figure 6: Comparison of LGA and HWA characteristics. We use box plots to visualize the metrics in Figures 6a, 6b, and 6c. Each box spans the interquartile range (i.e., from the 25th to the 75th percentile). The line inside the box marks the median, the diamond denotes the mean, and the whiskers extend to the most extreme points within 1.5× interquartile range of the quartiles. For Figures 6d, 6e, and 6f, we plot the mean values with 95% confidence intervals (CI).



(a) Upvotes **(b) Downvotes**

Figure 7: Histogram of the number of upvotes and downvotes.

while HWAs receive 1.45 upvotes and 0.35 downvotes. From the perspective of vote distribution, Figure 7 shows the distribution of upvotes and downvotes among answers that received at least one vote, with counts separated into one to five votes and more than five votes. LGAs are concentrated in the lower range (i.e., one to five), whereas HWAs are more frequently found in the higher range (i.e., more than five). This indicates that HWAs are more likely to elicit stronger reactions—either positive or negative—from multiple users than LGAs.

Sentiment analysis. For sentiment analysis in Korean, we fine-tune KoBERT on the Naver Sentiment Movie Corpus [42] which contains 100K positive and 100K negative movie reviews. We split the dataset into training and testing sets with a 9:1 ratio. The fine-tuned model takes text as input and produces prediction scores ranging from 0 to 1. A higher score indicates that the text is more likely to express a positive sentiment, while a lower score indicates a negative sentiment. The fine-tuned model achieves an accuracy of 89.45%, a precision of 89.46%, and a recall of 89.45%.

We collect 8,912 comments on LGAs and 361,056 comments on HWAs, and then apply the sentiment analysis model to both groups. The average score is 0.5595 for comments on LGAs and 0.4932 for comments on HWAs, indicating that users leave more positive comments on LGAs than on HWAs.

To further investigate negative comments, we manually reviewed the 100 lowest-scoring comments in each group. 16 comments on LGAs claim the answers were LLM-generated. In contrast, none of the comments on HWAs refer to LLMs. In summary, only a small number of users recognize that answers are LLM-generated and express negative sentiment. However, on average, users show a more positive stance toward LGAs.

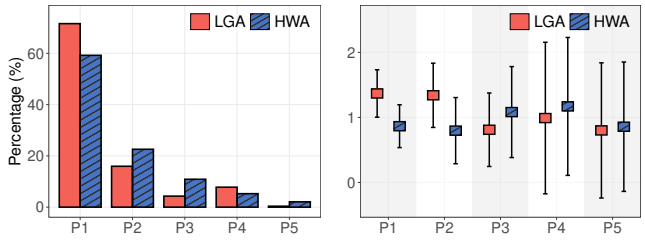
Finding 3: There is no clear difference in the number of upvotes, downvotes, or selection ratios between LGAs and HWAs. User comments on LGAs show more positive sentiment than those on HWAs. However, a manual review shows that 16 out of the 100 most negative LGA comments claim the answers were LLM-generated.

5.5 RQ4: What is the Purpose of LGAs?

We classify the purposes of LGAs and HWAs to investigate why LLMs are used to post answers by leveraging LLMs tasked with classifying their purposes.

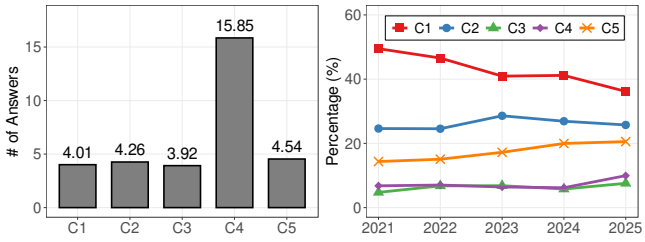
Answer purposes. It is challenging to manually identify the purpose of answers, as manual identification is not scalable and inherently subjective. To address this, we prepared five predefined answer purpose categories and instructed the LLM to select among them. The categories are: *(P1) Knowledge sharing*: sharing objective information or explanations, *(P2) Advertising/Marketing*: promoting products, services, or external links, *(P3) Personal experience*: sharing personal experiences, reviews, know-how or specific cases, *(P4) Emotional support*: providing comfort, empathy, or encouragement rather than to share information or knowledge, and *(P5) Irrelevant/Chitchat*: giving answer unrelated to the question, make jokes, or provide meaningless answers. The LLM can choose multiple categories at once, and we employ *gpt-5-mini*. The detailed instruction prompt and the human validation of the LLM-based classification are provided in Appendix A.1 and A.4, respectively.

Figure 8 presents the distribution of answer purposes for 1,000 randomly sampled LGAs and 1,000 HWAs. Figure 8a represent the proportion of each purpose, while Figure 8b represent the preference for each purpose. Preferences are calculated as the mean of (upvotes - downvotes) across answers. The two most frequent purposes in both groups are *(P1) Knowledge sharing* and *(P2) Advertising/Marketing*. Specifically, LGAs focus more on *Knowledge sharing* (71.59%) than HWAs (59.22%), while HWAs concentrate more on *Advertising/Marketing* (21.58%) than LGAs (15.95%). The only difference in ranking lies between *(P3) Personal experience* and *(P4) Emotional support*. Notably, *Personal experiences* account for



(a) Proportions (b) Preferences

Figure 8: (a) Proportions and (b) preferences (95% CIs) for LGA and HWA purposes.



(a) Avg. number of answers (b) Proportions of questions

Figure 9: (a) Average number of answers per question category and (b) Proportions of question categories by years.

more than twice the proportion in HWAs (10.88%) compared to LGAs (4.30%).

User preference. User preferences show distinct differences between LGAs and HWAs. For LGAs, users prefer *(P1) Knowledge Sharing* and *(P2) Advertising/Marketing*, which require clear, informative, and practical content. In contrast, for HWAs, users tend to prefer *(P3) Personal Experience* and *(P4) Emotional support*, which require experience-based narratives and empathetic expression.

Finding 4: The purposes of LGAs are more concentrated on *Knowledge Sharing* than those of HWAs. Users prefer *Knowledge sharing* and *Advertising/Marketing* purposes in LGAs, whereas they prefer *Personal experience* and *Emotional support* purposes in HWAs.

5.6 RQ5: Why Do Users Still Use Q&A Platforms Instead of LLM Services?

To investigate why users continue to use Q&A platforms rather than LLM services, we analyze the most-viewed questions before and after the release of the LLM services by characterizing the shared attributes of these questions. Specifically, for each year from 2021 to 2025, we select the 1,000 most-viewed questions from our dataset, compiling a total of 5,000 questions. These questions receive an average of 13,781 views, ranging from 2,502 to 614,953.

Question category. After reviewing the collected questions, we defined five question categories and instructed the LLM to choose among them: *(C1) Basic information*: requests for simple, objective facts or definitions. *(C2) Recommendation*: asks which option to choose under specific needs or constraints. *(C3) Step-by-step guide*: asks instructions to accomplish a task. *(C4) Opinion/Experience*: asks for personal views, reviews, or experiences. *(C5) Context-specific troubleshooting*: asks to diagnose and fix a problem in a specific context. All other setups for instructing LLM are the same as in RQ4, and the detailed instruction prompt is provided in Section A.1.

Figure 9a indicates the average number of answers per question by category. Notably, the *(C4) opinion/experience* category attains 15.85 answers, while the other categories attain about five answers. This suggests that users engage more with questions asking for personal opinion or experience, which LLMs generally do not offer.

We further analyze how category trends vary over time by computing the proportion of each category from 2021 to 2025. Figure 9b represents the proportion of question categories from 2021 to 2025. The ranking of categories in 2025 remains the same as in 2021: *(C1) basic information* accounts for the largest share, while the *(C3) step-by-step guide* accounts for the smallest. However, the proportion of some categories changed significantly. Specifically, *(C1) basic information* decreased from 49.49% in 2021 to 36.18% in 2025, a drop of 13.31%p. In contrast, *(C5) context-specific troubleshooting* increased from 14.35% in 2021 to 20.56% in 2025, a rise of 6.21%p. The *(C4) opinion/experience* category showed the second-largest increase, growing by 3.16%p.

We believe that these shifts suggest association in which categories well served by LLMs (i.e., *(C1) basic information*) have declined, while the categories requiring complex context (i.e., *(C5) context-specific troubleshooting*) and those better addressed by humans (i.e., *(C4) opinion/experience*) have grown over time. However, we note that these trends are correlational, and unobserved external factors may also influence the observed patterns.

Finding 5: Users engage far more with *opinion/experience* questions about three times as many answers as other categories. Over time, *Basic information* questions, which LLMs handle well, decreased by 13.31%p, while *context-specific troubleshooting* and *opinion/experience* questions, which LLMs handle poorly, increased by 6.21%p and 3.15%p, respectively.

6 Conclusion

In this paper, we empirically analyzed the impact of LGAs on *Naver Knowledge iN*. To this end, we evaluate recent LGT detection methods and ensemble them to develop a robust LGA detector. Using this detector, we identified 75,558 LGAs and characterized their shared properties. We further find only minimal differences in user reactions between LGAs and HWAs. We also reveal that the primary purpose of LGAs is knowledge sharing and that recent question categories have shifted from basic information-seeking to more context-specific, and experience- or opinion-oriented queries. We hope this work provides background knowledge on how prevalent LGAs are on the online Q&A platforms and outlines pathways for coexistence with LLM services.

Acknowledgments

We thank the anonymous reviewers for their constructive feedback. This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2020-II200153).

References

[1] [n. d.]. ask Ubuntu. https://askubuntu.com/.
[2] [n. d.]. Chegg. https://www.chegg.com/.
[3] [n. d.]. Naver Knowledge iN. https://kin.naver.com/.
[4] [n. d.]. quora. https://www.quora.com/.
[5] [n. d.]. ResearchGate. https://www.researchgate.net/.
[6] [n. d.]. Stack Overflow. https://stackoverflow.com.
[7] [n. d.]. What is this site's policy on content generated by generative artificial intelligence tools? https://stackoverflow.com/help/gen-ai-policy.
[8] Khalifa Afane, Wenqi Wei, Ying Mao, Junaid Farooq, and Juntao Chen. 2024. Next-Generation Phishing: How LLM Agents Empower Cyber Attackers. In *Proceedings of the IEEE International Conference on Big Data*. 2558–2567.
[9] Wissam Antoun, Benoît Sagot, and Djamé Seddah. 2024. From Text to Source: Results in Detecting Large Language Model-Generated Content. In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation*. 7531–7543.
[10] Navid Ayoobi, Sadat Shahriar, and Arjun Mukherjee. 2023. The Looming Threat of Fake and LLM-generated LinkedIn Profiles: Challenges and Opportunities for Detection and Prevention. In *Proceedings of the ACM Conference on Hypertext and Social Media*.
[11] Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2024. Fast-DetectGPT: Efficient Zero-Shot Detection of Machine-Generated Text via Conditional Probability Curvature. In *Proceedings of the International Conference on Learning Representations*.
[12] Amrita Bhattacharjee, Tharindu Kumarage, Raha Moraffah, and Huan Liu. 2023. ConDA: Contrastive Domain Adaptation for AI-generated Text Detection. In *Proceedings of the International Joint Conference on Artificial Intelligence*. 598–610.
[13] Kyle Bittle and Omar El-Gayar. 2025. Generative AI and Academic Integrity in Higher Education: A Systematic Review and Research Agenda. *Information* 16, 4 (2025).
[14] Gordon Burtch, Kokyun Lee, and Zhichen Chen. 2024. The consequences of generative AI for online knowledge communities. *Nature Scientific Reports* 14, 1 (2024).
[15] Canyu Chen and Kai Shu. 2024. Can LLM-Generated Misinformation Be Detected?. In *Proceedings of the International Conference on Learning Representations*. 16178–16187.
[16] Yutian Chen, Hao Kang, Vivian Zhai, Liangze Li, Rita Singh, and Bhiksha Raj. 2023. GPT-Sentinel: Distinguishing Human and ChatGPT Generated Content. *CoRR* abs/2305.07969 (2023).
[17] Federal Trade Comission. 2024. Federal Trade Commission Announces Final Rule Banning Fake Reviews and Testimonials. https://www.ftc.gov/news-events/news/press-releases/2024/08/federal-trade-commission-announces-final-rule-banning-fake-reviews-testimonials.
[18] Ella Creamer. 2023. Amazon restricts authors from self-publishing more than three books a day after AI concerns. https://www.theguardian.com/books/2023/sep/20/amazon-restricts-authors-from-self-publishing-more-than-three-books-a-day-after-ai-concerns.
[19] Carlos Dantas, Adriano Rocha, and Marcelo Maia. 2023. Assessing the Readability of ChatGPT Code Snippet Recommendations: A Comparative Study. In *Proceedings of the Brazilian Symposium on Software Engineering*.
[20] R. Maria del Rio-Chanona, Nadzeya Laurentsyeva, and Johannes Wachs. 2024. Large language models reduce public knowledge sharing on online Q&A platforms. *PNAS Nexus* 3, 9 (2024).
[21] Sebastian Gehrmann, Hendrik Strobelt, and Alexander M. Rush. 2019. GLTR: Statistical Detection and Visualization of Generated Text. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 111–116.
[22] Xun Guo, Shan Zhang, Yongxin He, Ting Zhang, Wanquan Feng, Haibin Huang, and Chongyang Ma. 2018. DeTeCtive: Detecting AI-generated Text via Multi-Level Contrastive Learning. In *Proceedings of the Advances in Neural Information Processing Systems*. 88320–88347.
[23] Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Spotting LLMs With Binoculars: Zero-Shot Detection of Machine-Generated Text. In *Proceedings of the International Conference on Machine Learning*.
[24] Beizhe Hu, Qiang Sheng, Juan Cao, Yang Li, and Danding Wang. 2025. LLM-Generated Fake News Induces Truth Decay in News Ecosystem: A Case Study on Neural News Recommendation. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 435–445.
[25] Baixiang Huang, Canyu Chen, and Kai Shu. 2025. Authorship Attribution in the Era of LLMs: Problems, Methodologies, and Challenges. *ACM SIGKDD Explorations Newsletter* 26, 2 (2025), 21–43.
[26] Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. Automatic Detection of Generated Text is Easiest when Humans are Fooled. In *Proceedings of the Annual Meeting on Association for Computational Linguistics*. 1808–1822.
[27] Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot Learning with Retrieval Augmented Language Models. *Journal of Machine Learning Research* 24, 251 (2023), 1–43.
[28] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Delong Chen, Wenliang Dai, Ho Shu Chan, Andrea Madotto, and Pascale Fung. 2020. Survey of Hallucination in Natural Language Generation. *ACM SIGKDD Explanations Newsletter* 55, 12 (2020), 1–38.
[29] Zhouyu Jiang, Mengshu Sun, Lei Liang, and Zhiqiang Zhang. 2025. Retrieve, Summarize, Plan: Advancing Multi-hop Question Answering with an Iterative Approach. In *Companion Proceedings of the ACM Web Conference 2024*. 1677–1686.
[30] Samia Kabir, David N. Udo-Imeh, Bonan Kou, and Tianyi Zhang. 2015. Is Stack Overflow Obsolete? An Empirical Study of the Characteristics of ChatGPT Answers to Stack Overflow Questions. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–17.
[31] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 6769–6781.
[32] Jiayi Kuang, Jingyou Xie, Haohao Luo, Ronghao Li, Zhe Xu, Xianfeng Cheng, Yinghui Li, Xika Lin, and Ying Shen. 2025. Natural Language Understanding and Inference with MLLM in Visual Question Answering: A Survey. *Comput. Surveys* 57, 8 (2025).
[33] Hyunseok Lee, Jihoon Tack, and Jinwoo Shin. 2024. ReMoDetect: Reward Models Recognize Aligned LLM's Generations. In *Proceedings of the Advances in Neural Information Processing Systems*. 2886–2913.
[34] Jooyoung Lee, Thai Le, Jinghui Chen, and Dongwon Lee. 2023. Do Language Models Plagiarize?. In *Proceedings of the Web Conference*. 3637–3647.
[35] Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. 2024. Data Augmentation Using Large Language Model for Fake Review Identification. In *Proceedings of The Knowledge and Systems Sciences*.
[36] Chengzhi Mao, Carl Vondrick, Hao Wang, and Junfeng Yang. 2024. Raidar: geneRative AI Detection viA Rewriting. In *Proceedings of the International Conference on Learning Representations*.
[37] Philip M. McCarthy and Scott Jarvis. 2010. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods* 42 (2010), 381–392.
[38] Lorenz Mindner, Tim Schlippe, and Kristina Schaaff. 2023. Classification of Human- and AI-Generated Texts: Investigating Features for ChatGPT. In *Proceedings of the International conference on artificial intelligence in education technology*. 152–170.
[39] Yisroel Mirsky, Ambra Demontis, Jaidip Kotak, Ram Shankar, Deng Gelei, Liu Yang, Xiangyu Zhang, Wenke Lee, Yuval Elovici, and Battista Biggio. 2023. The Threat of Offensive AI to Organizations. *Computers & Security* 124 (2023).
[40] Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature. In *Proceedings of the International Conference on Machine Learning*. 24950–24962.
[41] Sahrima Jannat Oishwee, Natalia Stakhanova, and Zadia Codabux. 2024. Large Language Model vs. Stack Overflow in Addressing Android Permission Related Challenges. In *Proceedings of the International Conference on Mining Software Repositories*. 373–383.
[42] Lucy Park. 2016. Naver sentiment movie corpus v1.0. https://github.com/e9t/nsmc.
[43] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2012. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 21, 140 (2012), 1–67.
[44] Brian Richards. 1987. Type/token ratios: What do they really tell us? *Journal of child language* 14, 2 (1987), 201–209.
[45] Sayak Saha Roy, Poojitha Thota, Krishna Vamsi Naragam, and Shirin Nilizadeh. 2024. From Chatbots to Phishbots?: Phishing Scam Generation in Commercial Large Language Models. In *Proceedings of the IEEE Symposium on Security and Privacy*. 36–54.
[46] Leuson Da Silva, Jordan Samhi, and Foutse Khomh. 2025. LLMs and Stack Overflow discussions: Reliability, impact, and challenges. 230 (2025).
[47] SK telecom. 2025. KoBERT: Korean BERT pre-trained cased. https://sktelecom.github.io/project/kobert/.
[48] Eduard Tulchinskii, Kristian Kuznetsov, Laida Kushnareva, Daniil Cherniavskii, Sergey Nikolenko, Evgeny Burnaev, Serguei Barannikov, and Irina Piontkovskaya. 2023. Intrinsic Dimension Estimation for Robust Detection of AI-Generated Texts. In *Proceedings of the Advances in Neural Information Processing Systems*. 39257–39276.
[49] Pénelope Forcioli Vijini Liyanage, Davide Buscaldi. 2024. Detecting AI-enhanced Opinion Spambots: a study on LLM-generated Hotel Reviews. In *Proceedings of The Workshop on e-Commerce and NLP*. 74–78.
[50] Ivan Vykopal, Matúš Pikuliak, Ivan Srba, Robert Moro, Dominik Macko, and Maria Bielikova. 2024. Disinformation Capabilities of Large Language Models.

In *Proceedings of the Annual Meeting on Association for Computational Linguistics*. 14830–14847.
[51] Xianjun Yang, Wei Cheng, Yue Wu, Linda Petzold, William Yang Wang, and Haifeng Chen. 2024. DNA-GPT: Divergent N-Gram Analysis for Training-Free Detection of GPT-Generated Text. In *Proceedings of the International Conference on Learning Representations*. 225–240.
[52] Xiao Yu, Kejiang Chen, Qi Yang, Weiming Zhang, and Nenghai Yu. 2024. Text Fluoroscopy: Detecting LLM-Generated Text through Intrinsic Features. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 15838–15846.
[53] Biru Zhu, Lifan Yuan, Ganqu Cui, Yangyi Chen, Chong Fu, Bingxiang He, Yangdong Deng, Zhiyuan Liu, Maosong Sun, and Ming Gu. 2023. Beat LLMs at Their Own Game: Zero-Shot LLM-Generated Text Detection via Querying ChatGPT. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 7470–7483.
[54] Jiajun Zhu, Ye Liu, Meikai Bao, Kai Zhang, Yanghai Zhang, and Qi Liu. 2025. Self-Reflective Planning with Knowledge Graphs: Enhancing LLM Reasoning Reliability for Question Answering. *CoRR* abs/2505.19410 (2025).

A Appendix

A.1 Instruction Prompts

Figures 10, 11, and 12 illustrate the system and user prompts used for generating LGAs for training, classifying answer purposes, and classifying question categories, respectively.

[System Prompt]

```
You are a Q&A system. Generate an answer to the question.
Always generate the answer in Korean.
Output only the answer text. Do not include prefixes like
“Response:” or “Answer:”.
```

[User Prompt]

```
Generate a answer to the following question.

Question:
Title: {QUESTION_TITLE}
Body: {QUESTION_BODY}
```

Figure 10: Instruction prompt for generating LGAs (translated).

A.2 Crawling Process

Questions on *Naver Knowledge iN* are associated with tags that are either provided by the questioner or automatically assigned by the platform. By clicking on a tag, users can browse all questions associated with that tag.

We first collect 3,150 tags by traversing popular questions displayed on the *Naver Knowledge iN* main page. From this tag set, we then randomly sampled tags and retrieved all questions and their corresponding answers associated with the sampled tags. Table 3 summarizes key features of the dataset, including their description, data types, and average quantities.

A.3 Dataset Statistics

Figure 13 shows the number of answers aggregated by quarter from 2020-Q4 to 2025-Q3. The results indicate that we successfully crawled more than 20K answers in each quarter, both before and after the release of the LLM services.

[System Prompt]

```
You are an expert at reading answers on a Q&A platform and
classifying their purposes.
The allowed purpose labels (single or multiple) are exactly the
following five:

1.Knowledge Sharing: Objective information or explanations
intended to be shared
2.Advertising/Marketing: Promoting products, services, or external
links
3.Personal Experience: Sharing personal experiences, reviews,
know-how, or case examples
4.Emotional Support: Providing comfort, empathy, or encouragement
rather than information/knowledge sharing
5.Irrelevant/Chit-chat: Responses unrelated to the question, jokes,
or meaningless answers
```

[User Prompt]

```
Read the question and answer below, and classify the purpose of
the answer.

Question:
Title: {QUESTION_TITLE}
Body: {QUESTION_BODY}

Answer:
{ANSWER_TEXT}

Output format (JSON):
{
"Purpose": ["<one or more of the five above>"],
"Reason": "<a brief explanation for the classification>"
}
```

Figure 11: Instruction prompt for classifying answer purposes (translated).

Figure 14 presents the distribution of answers by word length, ranging from 0 to 600. The answers are concentrated in shorter passages: about 48% contain fewer than 30 words, and only 1.46% exceed 600 words. Previous research [11, 33] demonstrates that LGT detection methods perform better on longer passages. Nevertheless, we confirm that our ensemble-based method achieves near-perfect performance on our short-passage dataset.

Table 4 presents the counts and proportions of answer topics in our crawled dataset. The top three topics are *Lifestyle*, *Education & Study*, and *Health*, which together account for over half of the dataset. By contrast, *Personal Concerns*, *Regions & Places*, and *Gaming* are among the smallest categories.

Table 3: Key fields extracted from answers.

Field	Type	Description	Avg.
question id	int	Identifier of the corresponding question	–
text	string	The content of the answer	–
date	date	The date when the answer was written	–
upvote	int	The number of upvotes received by the answer	1.44
downvote	int	The number of downvotes received by the answer	0.34
comment	int	The number of comments on the answer	0.34
selected	bool	Whether the answer is selected by questioner	0.44

[System Prompt]

You are an expert at reading question text on a Q&A platform and classifying the category of question.
The allowed category labels (single or multiple) are exactly the following five:

1. **Basic information**: Questions asking for simple, objective facts/definitions
2. **Recommendation**: Questions seeking advice on selecting/comparing options under specific needs/constraints
3. **Step-by-step guide**: Questions asking for procedures/steps to perform a task
4. **Opinion/Experience**: Questions requesting personal opinions, reviews, and experiences
5. **Context-specific troubleshooting**: Questions diagnosing/solving errors or symptoms in a specific environment

---

[User Prompt]

Read the question below and classify its category.

Question:
Title: {QUESTION_TITLE}
Body: {QUESTION_BODY}

Output format (JSON):
{
"Category": ["<one or more of the five above>"],
"Reason": "<a brief explanation for the classification>"
}

Figure 12: Instruction prompt for classifying question categories (translated).

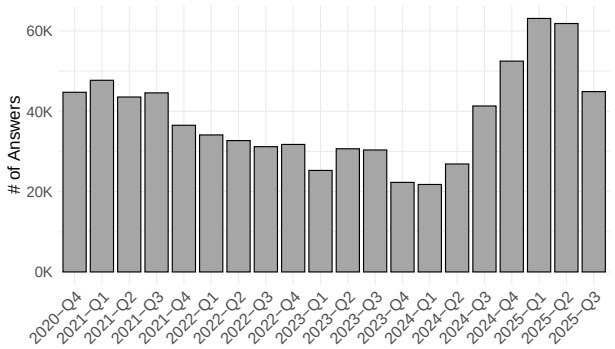


Figure 13: Distribution of the number of answers aggregated at six-month intervals from 2020-Q4 to 2025-Q3.

A.4 Human Validation of LLM-Based Classification

To validate the reliability of the LLM-based classification, we conducted a human evaluation on a 10% random sample of the LLM classification results. Specifically, we reviewed 700 samples in total, covering both the classification of answer purposes (in Section 5.5) and question categories (in Section 5.6), by carefully examining the predicted labels.

The evaluation revealed zero misclassified samples for the classification of answer purposes (in Section 5.5) and six misclassified samples for question categories (in Section 5.6), indicating that the LLM-based classification shows an accuracy of 99.1%, which is highly reliable.

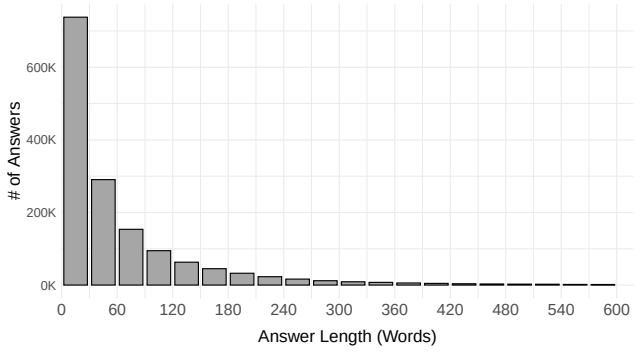


Figure 14: Distribution of the number of answers by word length.

Table 4: Number and proportion of answer topics in our crawled dataset.

Ranking	Topic	# of answers	Proportion
1	Lifestyle	355,091	24.36%
2	Education & Study	245,692	16.85%
3	Health	178,304	12.23%
4	Entertainment & Arts	129,421	8.88%
5	Shopping	120,079	8.24%
6	Society & Political	93,819	6.44%
7	IT & Technology	91,293	6.26%
8	Economy	75,369	5.17%
9	Travel	46,920	3.22%
10	Sports & Leisure	44,835	3.08%
11	Kids	31,104	2.13%
12	Gaming	26,303	1.80%
13	Regions & Places	17,640	1.21%
14	Personal concerns	1,839	0.13%

A.5 Implementation Details for LGT Detection

For the Log-Likelihood, Log-Rank, and Rank methods, we use GPT-Neo 1.3B as the language model f_θ. For DetectGPT and Fast-DetectGPT, we use KoGPT, which performs well on Korean text, as the scoring model and T5-small as the mask-filling model that generates the perturbed texts.

We adopt T5-Sentinel for GPT-Sentinel, as it outperforms RoBERTa-Sentinel. Following the original setups, we use gte-Qwen as the encoder for Text Fluoroscopy and SimCSE-RoBERTa for DeTeCtive. We also set K = 50 when running the K-NN algorithm in DeTeCtive. To generate the LGT–HWT mixed dataset for ReMoDetect, we use Llama-4-Maverick-17B-128E-Instruct-FP8 to paraphrase the training data.