Chinese Morph Resolution in E-commerce Live Streaming Scenarios

Jiahao Zhu[1] , Jipeng Qiang[1*], Ran Bai[2], Chenyu Liu[2], Xiaoye Ouyang[2]
[1] School of Information Engineering, Yangzhou University, China
[2] China Academy of Electronic and Information Technology, China
mz120231031@stu.yzu.edu.cn, jpqiang@yzu.edu.cn
{bairan, liuchenyu, ouyangxiaoye }@cetc.com.cn

Abstract

E-commerce live streaming in China, particularly on platforms like Douyin, has become a major sales channel, but hosts often use morphs to evade scrutiny and engage in false advertising. This study introduces the Live Auditory Morph Resolution (LiveAMR) task to detect such violations. Unlike previous morph research focused on text-based evasion in social media and underground industries, LiveAMR targets pronunciation-based evasion in health and medical live streams. We constructed the first LiveAMR dataset with 86,790 samples and developed a method to transform the task into a text-to-text generation problem. By leveraging large language models (LLMs) to generate additional training data, we improved performance and demonstrated that morph resolution significantly enhances live streaming regulation.

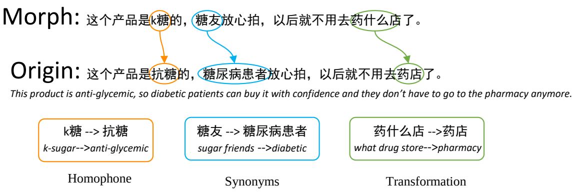

Morph: 这个产品是k糖的，糖友放心拍，以后就不用去药什么店了。
Origin: 这个产品是抗糖的，糖尿病患者放心拍，以后就不用去药店了。
This product is anti-glycemic, so diabetic patients can buy it with confidence and they don’t have to go to the pharmacy anymore.
k糖 --> 抗糖
k-sugar-->anti-glycemic
糖友 --> 糖尿病患者
sugar friends -->diabetic
药什么店 -->药店
what drug store-->pharmacy
Homophone
Synonyms
Transformation

Figure 1: Example of morph used in the live streaming scenarios

1 Introduction

E-commerce live streaming has become an immensely popular and influential sales channel in China. For example, one short video platform Douyin hosted over 9 million live broadcasts each month, selling more than 10 billion items through there sessions (Center, 2022). To increase sales and attract customers, hosts engage in practices such as using morphs to evade scrutiny and conducting false advertising. As shown in the Figure 1, morphs are used in promotional language that suggests the product has medicinal effects in order to evade scrutiny. Detecting violations during the live commerce process is crucial for protecting consumer rights and promoting industry standardization (Xiao, 2024; Xu, 2024).

To detect violations in live commerce, resolving morphs used in the live content is intuitively important. Previous morph research has primarily focused on social media commentary and underground industries (Sha et al., 2017; You et al., 2018; Wang et al., 2024). There are two main differences between their research and this paper.

(1) Different purposes for morphing: Their focus is on making the written text appear different to evade keyword recognition (You et al., 2018; Wang et al., 2024), whereas the live streaming field focuses on differences in pronunciation to evade voice censorship. For example, in visual scenarios, characters with a left-right structure are often split into two words, such as “胡” (*hú*)->“古月”(*gŭ yùe*). In the live streaming field, a very common situation is inserting some meaningless words , like “某”(*mŏu, some*) or “什么”(*shén mē, what*) can help maintain the rhythm of speech without interfering with the listener’s understanding of the information, such as “手术”(*shŏu shù, surgery*)->“手某术”(*shŏu mŏu shù, surgery*)".

(2) Different subjects of interest: Social media commentary focuses on current affairs and politics (You et al., 2018), and underground industries focus on illegal gambling and the sex industry (Wang et al., 2024), while our study focuses on the health and medical industry.

In this paper, we focus on auditory-based morph resolution task in live screaming scenarios, denoted as LiveAMR task. Voice censorship is first processed using automatic speech recognition (ASR) technology (Wang et al., 2023a), which converts speech into text. By observation, we can find that

[*] Corresponding authors.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Industry Track), pages 380–389
April 30, 2025 ©2025 Association for Computational Linguistics

the LiveAMR task is similar to the grammar correction task (Kobayashi et al., 2024). In this way, we can train a text generation model to convert the input text with morph words into normal text. This study produces two main contributions toward the development and evaluation of LiveAMR methods. Our contributions are listed below:

(1) To the best of our knowledge, there is no existing work on LiveAMR. We constructed a LiveAMR dataset containing 86,790 samples, including 2,688 different morphs. In live streaming scenarios, considering the noise in the live environment and the variations in presenters' expressions, the results of different ASR systems vary greatly. We re-annotated the second test set, selecting different live streaming rooms and different ASR methods which includes 400 positive and 400 negative samples. This approach allows us to comprehensively assess the model's performance and adaptability under different conditions.

(2) We transform LiveAMR task into a type of text-to-text generation task. By training the T5 model using the constructed morph dataset, we achieved F1 scores of 94% and 82% on Test Set 1 and Test Set 2, surpassing the performance of other models respectively. Considering the efficiency of manual annotation is relatively low, we propose an innovative solution that leverages large language modeling to generate LiveAMR examples, thereby improving the scale of LiveAMR training set. Experimental results show that incorporating the dataset generated by LLM into the training process also improved the performance of LiveAMR methods. Additionally, we investigated the peformance of morph resolution in detecting violations. We also verify that morph resolution can significantly improve the model's accuracy in the live streaming regulation. The dataset and code is available at github [1].

2 Related Work

There has been extensive research on morph resolution across different language backgrounds including English (Ji and Knight, 2018; Li et al., 2022; Wang et al., 2023b; Qiang et al., 2023c), and Chinese (Huang et al., 2017, 2019; Qiang et al., 2023a), etc. In this paper, we only focus on morph resolution in Chinese. Because Chinese is a pictographic language, methods for identifying morph words in other languages cannot be applied to Chinese. Existing research on Chinese morphs primarily focuses on social media and underground industries.

Initially, it was considered a filtering problem, with researchers using statistical and rule-based matching methods to identify problematic text (Wang et al., 2013; Choudhury et al., 2007; Qiang et al., 2023b; Yoon et al., 2010). Subsequently, Sha et al. (Sha et al., 2017) proposed incorporating radicals into Chinese characters to enhance their features and improve morphs resolution. You et al.(You et al., 2018) further extracted actual contextual information and enhanced embedded representations by integrating transformed mentions or target candidates with their relevant context into an AutoEncoder. Recently, addressing the characteristics of morph words in underground industries, Wang et al.(Wang et al., 2024) introduced a morph parsing algorithm based on machine translation models.

However, existing research on morphs mainly focuses on social media and underground industries, with studies on morph resolution in the emerging context of live streaming still being relatively scarce.

3 Task Definition

In the research context of this paper, 'morph' refers to the process where live streamers avoid platform censorship by replacing sensitive or restrictive words during product promotion, while ensuring that the audience can easily understand the original meaning conveyed by the transformation. Here, we formally define the auditory-based morph resolution task in live screaming scenarios as the LiveAMR task. By analyzing thousands of videos, the main types of transformations can be categorized into three major types (transformation, homophones, and synonyms), as shown in Table 1.

Suppose one example is "咱们一些<小糖人>都是一样可以放心去喝，也不用去找<白褂褂>了。" (Some diabetes patients can safely drink without needing to consult a doctor.) with two morphs “小糖人”(*sugar doll*)->“糖尿病患者”(*diabetic*) and “白褂褂”(*people with white*)->“医生”(*doctor*). The correct output by LiveAMR method should be “咱们一些<糖尿病患者>都是一样可以放心去喝，也不用去找<医生>了。”.

4 Dataset Construct

In this section, we describe the whole process of constructing a LiveAMR dataset.

[1] https://github.com/loopback00/LiveAMR

Type	Characteristic	Examples
Transformation	Insert meaningless characters into words, or change the structure while keeping the sound similar to the original words.	某医某院:医院 (hospital) *mǒu yī mǒu yuàn:yī yuàn* 祛什么斑:祛斑 (spot removal) *qū shén mē bān:qū bān* 小问小题:问题 (problem) *xiǎo wèn xiǎo tí:wèn tí*
Homophone	Use symbols to replace Chinese characters	k糖:抗糖 (anti glycemic) *k táng: kàng táng* k老:抗老 (anti aging) *k lǎo: kàng lǎo*
Synonyms	Use words that are highly related or synonymous with the target word	白大褂:医生 (people in white:doctor) 心灵之窗:眼睛 (windows to the soul:eyes)

Table 1: The three types of transformations in LiveAMR. For the two types of morphs, transformation and homophone, we have additionally annotated their pinyin below them.

Data Collection: We crawled videos from four domains in Douyin website [2]: health supplements, pharmaceuticals, medical devices, and cosmetics. These areas are chosen due to their unique risks and challenges in live streaming. As products aimed at improving health, they have a large market size and diverse categories. However, due to their specific nature, consumers often face significant information asymmetry regarding their efficacy and safety. This asymmetry creates opportunities for false advertising and misleading marketing, particularly in the highly interactive and instant-feedback environment of live streaming (Auronen, 2003).

From the four domains, we carefully selected 25 live streaming channels as data sources. These channels are well-known on the platform and have high sales, ensuring they are representative. We crawled a total of 7,812 live video clips, each limited to 60 seconds. This duration ensures sufficient information capture while reducing data processing complexity to some extent, providing rich material for subsequent data annotation.

ASR Process: We first need to convert the audio information into text format. We tested the transcription performance of mainstream ASR tools in this scenario, with FunASR (Gao et al., 2023) achieving the best recognition results, followed by Kaldi (Ravanelli et al., 2019) and Whisper (Radford et al., 2023). We employed this FunASR to perform ASR, converting the spoken content in the crawled videos into text for subsequent morph annotation. A total of 86,750 speech statements were transcribed.

This process of converting video to text not only adds a new modality to the research but also makes the form of morphs more flexible and varied. In the video context, morph words themselves are very difficult to distinguish by ASR. Additionally, other factors such as the host's colloquial expressions, fast speaking pace, and background noise can lead to inaccuracies in ASR recognition results, resulting in a more diverse range of extracted morph forms.

Label Suggestions via LLMs: Recently, LLMs have been widely used for data annotation (Zhang et al., 2023). Despite the challenges posed by the presence of grammatical morphs in the annotation of morphs, LLMs with their powerful contextual learning capabilities, can still identify some standard morphs and provide the correct original terms. Therefore, we provided the annotation suggestions from the LLMs to human annotators as a reference, assisting them in the annotation process to enhance both efficiency and accuracy. Whether some morphs recommended by the LLMs actually exist in the original document, annotators can more quickly locate the variant words. To specifically illustrate the performance of LLMs in LiveAMR task, we selected three representative LLMs as baselines to comparison.

Human Annotation: In order to make it easier for annotators to label, we created a website for annotation. We provided corresponding videos and LLM annotation suggestions as auxiliary information, with video support being essential. When we attempted annotation without referencing the videos, annotators reported that many words could not be clearly understood. We recruited three interns with bachelor's degrees with annotation expe-

[2] https://www.douyin.com/

rience and an understanding of morph characteristics as annotators.

The unique research scenarios required annotators to process multiple modalities of information, enhancing the quality and accuracy of the annotations. Prior to formal annotation, detailed training was provided, including explanations of guidelines and procedures, along with trial annotations to ensure understanding and adherence to the tasks. Each annotator needs to undergo training before starting their annotation work, and they can only begin once they have passed the training. As a result, the annotation process yielded 6,853 positive sentences containing morphs and 90,137 negative sentences without morphs.

Data Filtering: Despite manually annotating morph words, we found that a small number of variant words were still not annotated. Therefore, we further adopted a process of human-machine collaboration for secondary annotation to achieve the goal of constructing a high-quality dataset.

First, we use the corpus manually annotated in the previous step to build a morph resolution model, employing both rule-based method and pre-trained language model based method. Second, we automatically annotate the manually annotated corpus from the previous step using the trained method. Third, we manually verify the correctness of the machine's automatic annotation results, retaining correct annotations and discarding incorrect ones. Finally, the morphs corresponding to each original document are the combination of the results from the previous manual annotation and this step of collaborative annotation.

(1) Rule-based method: Using the corpus manually annotated in the previous step, we constructed a morph dictionary D whic contains 430 original words and their corresponding 2,688 morphs. Each entry in the dictionary contains one original word along with their multiple morph words, where the relationship between original word and morphs is one-to-many.

During the annotated process automatically, we search each instance of the manually annotated corpus to find the morphs in the dictionary. If a match is found, this instance and the identified morph word will undergo further manual verification

(2) Pre-trained language model based method: Using the manually annotated corpus, we fine-tuned the pre-trained language model Mengzi-T5 (Zhang et al., 2021). The details of the method is shown in section 5.1. During the annotated process automatically, each instance is input into the fine-tuned model, and the model's input and output were compared. If the input and output differed, it indicated that there might be omitted morph in the sample. These samples were further examined, and upon confirmation, they were appropriately annotated.

	Positive&Negative	Morph Num
Train	6,236/76,554	7,301
Valid	800/800	1,025
Test1	800/800	1,081
Test2	400/400	548

Table 2: The statistics of the constructed Chinese morph dataset.

Data Analysis: Since the dataset construction is highly dependent on ASR outputs, the same speech input may produce different ASR results when processed by different ASR models. For example, the morph form “白某障”(*bái mǒu zhàng*) for “白内障”(*bái nèi zhàng, cataract*) could be transcribed as “白母障”(*bái mǔ zhàng*), “白某张”(*bái mǒu zhāng*), “白某章”(*bái mǒu zhāng*) by different ASR models.

To conduct a more comprehensive evaluation, We re-annotated the second test set (denoted Test2), selecting both different live streaming rooms and different ASR method. The Test2 includes 400 positive and negative instances.

Following the above process, we constructed a high-quality and comprehensive morph dataset, as shown in Table 2. Dataset consists of 8,236 positive samples and 78,554 negative samples. The dataset includes a total of 431 original words and their corresponding 2,688 morphs forms, in which each word has nearly 7 morph words on average.

5 Methods

LiveAMR method: Existing morph resolution methods generally use non-autoregressive language model MacBERT, a corrective masked language model pre-training task was added to the BERT model (Wang et al., 2024). In the LiveAMR task, since the length of the variant words does not equal the length of the original word, we will use a text-to-text pre-trained model as a backbone, such as BART (Lewis, 2019) and Mengzi-T5 (Zhang et al., 2021). Below are the steps involved in this process.

The created dataset consists of source-target pairs (X and Y), where: X is the input text (live stream transcript), Y is the desired output text (the normal text without morph words). The goal of the model is to learn a mapping from X to Y.

The pre-trained model $\mathcal{M}$ is a transformer-based sequence-to-sequence architecture, which is typically structured as: (1) Encoder: Takes the input sequence X and encodes it into hidden states; (2) Decoder: Takes the encoder's hidden states and generates the target sequence Y.

During training, the model aims to minimize the loss, which is typically the Cross-Entropy Loss for text generation tasks. The formula for Cross-Entropy Loss is:

$$\mathcal{L} = -\sum_{i=1}^{T}\sum_{v=1}^{V} \hat{y}_{i,v} \log p(y_{i,v}|X)$$

where T is the length of the target sequence, V is the size of the vocabulary, $\hat{y}_{i,v}$ is a one-hot encoding of the true token at position i in the target sequence, and $p(y_{i,v}|X)$ is the predicted probability of token y_i at position i given the input X.

During training, the model minimizes the loss function $\mathcal{L}$ with respect to the model parameters θ over multiple iterations (epochs):

$$\theta^{\star} = \arg\min_{\theta} \mathbb{E}[\mathcal{L}(X, Y; \theta)]$$

Where $\mathbb{E}$ denotes the expectation over the training data, $\mathcal{L}(X, Y; \theta)$ is the loss function dependent on the input X, the target Y, and the model parameters θ.

After fine-tuning, the model generates new outputs for unseen inputs. This is done by feeding the input X_{input} through the model to obtain the predicted sequence Y_{pred}:

$$Y_{\text{pred}} = \mathcal{M}(X_{\text{input}})$$

Where Y_{pred} is the generated sequence, which can be decoded back into text.

Data Augmentation via LLMs: Some studies suggest that LLMs can be used to generate training datasets (Ding et al., 2023). Although manual annotation can yield morph data from the real world, it comes at a high cost and may contain some redundancy, limiting the scale and diversity of the dataset. Therefore, we aim to leverage LLMs to generate more morph data to supplement manually annotated data and enhance the model's generalization ability.

However, given the complexity of morph forms and the limitations of LLMs in understanding them, we did not directly ask the LLMs to generate sentences containing morphs. To this end, we propose a more reliable construction strategy that combines the annotated morphs lexicon with LLM capabilities. The specific steps are as follows:

(1) We randomly select a positive example from the training set and extract the corresponding morph words WS. There may be one or more morph words.

(2) Based on the morph dictionary D, we obtain the original word WO for WS.

(3) We had the LLM simulate a live commerce scenario to generate 5 different sentences containing WO.

(4) According to the morph dictionary D, we replace the original word WO with different morph words to construct a set of sentences containing different morph words.

Through this approach, we constructed a manually created morph dataset containing 11,280 positive samples and 2,155 negative samples. Additionally, each positive sample generated by the LLM averages 2.87 morphs. This data effectively supplements the manually annotated data, increasing the scale and diversity of the model's training data. In Table 6, show some specific examples.

6 Experiment

6.1 Experimental Setup

Metrics. We expect the model to modify only the morphs in the target sentences without altering any other parts. A strict sentence-level assessment is applied: a positive sample is considered successfully predicted only when all morphs are correctly restored. For negative samples, a negative sample is deemed successfully predicted only if the model makes no modifications at all.

Baselines. The following models were selected as the baseline for comparison:

(1)**LLMs**: To explore the morphs resolution capabilities of LLMs, we chose three representative models in the field of Chinese language understanding: GPT-3.5-turbo [3], Deepseek -V2[4], and GLM4-Plus[5]. We manually selected 8 examples from the training set, including 6 positive samples and 2 negative samples, to be added as context to the prompt. The temperature was uniformly set to 0.7.

[3] https://openai.com/

[4] https://platform.deepseek.com/

[5] https://chatglm.cn/

	Test1				Test2			
Method	Acc	Pre	Recall	F1	Acc	Pre	Recall	F1
GPT	0.405	0.421	0.320	0.364	0.496	0.494	0.441	0.466
Deepseek	0.605	0.660	0.529	0.587	0.677	0.667	0.626	0.646
GLM	0.451	0.484	0.515	0.499	0.532	0.525	0.649	0.580
Kenlm	0.583	0.607	0.372	0.537	0.516	0.515	0.513	0.514
Seq2Edit	0.651	0.968	0.361	0.526	0.702	0.987	0.408	0.588
Convseq2seq	0.740	0.978	0.527	0.685	0.687	0.898	0.421	0.573
BART	0.708	0.701	0.767	0.738	0.656	0.670	0.611	0.639
T5	0.893	**0.989**	0.801	0.888	0.760	**0.968**	0.536	0.690
+Aug	**0.928**	0.937	**0.927**	**0.932**	**0.863**	0.929	**0.787**	**0.852**

Table 3: The results of different methods, where "+Aug" indicates fine-tuned the model using data augmentation via LLM.

(2)**Seq2seq Model**: We selected two Seq2seq models Convseq2seq (Gehring et al., 2017) and BART (Lewis, 2019) as backbone, and fine-tune the model on the constructed training datset.

(3)**Others**: To better illustrate that seq2seq is more suitable for the morph resolution task, we chose to analyze the statistical language model Kenlm (Heafield, 2011) and BERT-based model Seq2Edit (Omelianchuk et al., 2020).

(4) **Our method**: It is based on T5 (mengzi-T5 (Zhang et al., 2021)). This model adopts the T5 training paradigm and has been retrained on large-scale Chinese corpora.

6.2 Implementation Details

It is based on T5 (mengzi-T5 (Zhang et al., 2021)). The Mengzi T5 model includes an encoder and decoder, where each consisting of 12 layers of Transformer layers. This model adopts the T5 training paradigm and has been retrained on large-scale Chinese corpora.

During the training process, the maximum length of the input sequence is set to 128, and the initial learning rate is set to 1e-4. We train the model for 20 epochs on a 24GB Nvidia 3090Ti GPU with the batch size set to 32. We use the AdamW optimizer, and the model employs a cosine annealing learning rate schedule.

6.3 Experimental Results

The experimental results, presented in Table 3, reveal that character-level correction methods like Seq2Edit and the statistical language model Kenlm are inadequate for addressing morphs in live streaming scenarios. In contrast, Seq2seq models (Convseq2seq, BART, and T5) perform better at managing inconsistencies in output length. Notably, the T5 model achieved the highest F1 score across both test sets, demonstrating its effectiveness for this task.

For T5 method, the results via data augmentation improved the F1 scores of T5 model by 4.95% on Test1; on Test2, the improvements was 23.47%. Our method shows stable performance across different test sets due to its contextual learning capabilities. On Test1, its performance is slightly lower than the baseline model, likely because the baseline excels with data similar to the training set. However, on Test2, which uses data from a different ASR model, the LLM's performance matches that of fine-tuned Seq2seq models, demonstrating its generalization ability with varied data distributions.

6.4 Usefulness of Morph Resolution

To investigate the role of morph resolution in detecting violations in e-commerce live streaming scenarios, we conducted a simple usability experiment.

Setup. We selected 4,641 live streaming clips for ASR processing and annotated the transcription results for each clip. After thorough consultation with market regulators, we have categorized the identification of violations in live-streaming sales videos into three types: compliance, suspected violation, and serious violation. Specifically, the "compliance" category refers to content that fully adheres to relevant regulations and platform rules, without any violation. The "suspected violation" category covers content that may potentially involve violation behaviors but requires further verification, such as suspected acts of inducing irrational consumption. The "serious violation" category pertains to actions that are explicitly prohibited by the platform or regulations, such as promoting healthcare products as drugs.

We annotated a total of 4,447 instances including 2,430 compliances, 1,305 suspected violations, and 712 serious violations. We divided them into a

Table 4: Statistical information on dataset.

	Class	Number
Training set	Compliance	2,250
	Suspected violation	557
	Serious Violation	1,150
Validation Set	Compliance	130
	Suspected violation	130
	Serious Violation	130
Test set	Compliance	50
	Suspected violation	25
	Serious Violation	25

training set, a validation set, and test set. The test set includes 100 samples, and the validation set contains 390 samples. The statistical information of the constructed CLiveSVD dataset is presented in Table 4.

Method	Cat.	Acc	Pre	Recall	F1
Defalut	0	0.81	0.917	0.88	0.89
	1	0.81	0.77	0.68	0.72
	2	0.91	0.66	0.80	0.72
Morph	0	0.90	0.96	0.96	0.96
	1	0.90	0.77	0.84	0.80
	2	0.90	0.91	0.84	0.87

Table 5: Comparison of experimental results. "Default" indicates that the ASR results of the video are not processed. "Morph" refers to the processing of the ASR results for morph resolution. "0" represents compliant categories, "1" indicates suspected violation categories, and "2" denotes serious violation categories.

Implements. It is important to note that in the default method, neither the training set nor the test set undergoes any changes, while in the comparison method, both the training set and the test set are processed with morph resolution. The BERT (Kenton and Toutanova, 2019) model was fine-tuned for classification task.

Results. As shown in Table 5, after resolution morphs in the original ASR results, the F1 scores for the compliant, suspected violation, and serious violation categories increased by approximately 6.91%, 11.76%, and 20.36%, respectively, compared to the unprocessed results. This demonstrates that morph resolution can significantly improve the model's accuracy in detecting v.

6.5 Ablation Study

We explored the impact of data augmentation quantity on model performance. As shown in Section 5, we controlled the data augmentation by setting the number of sentences generated for each original word. The sentence counts were set to 1, 2, 3, 4, 5, and 6, resulting in data volumes of 2,693, 5,373, 8,058, 10,744, 14,405, and 16,116, respectively.

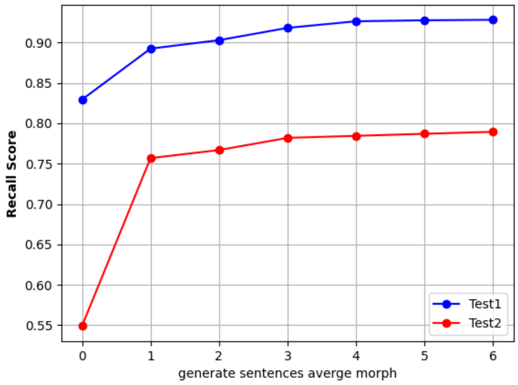


Figure 2: Performance with different number of training samples.

In Figure 2, the experimental results show that data augmentation has a significant positive impact on model performance. At the same time, when the variable is set to 5, the number of augmented samples reaches 14,405, and the model's performance tends to stabilize.

7 Conclusion

This study introduces the task of morph resolution in live streaming scenarios, termed LiveAMR. A LiveAMR dataset was created through human-LLM collaboration, comprising 7,836 positive and 91,119 negative samples. The study analyzed task characteristics and utilized a text-to-text model architecture for morph resolution. Given the impracticality of manually constructing large-scale training corpora, an efficient data augmentation method based on LLMs was proposed, leveraging existing annotated data. Experimental results show that this augmentation method enhances model performance compared to baselines. The findings also indicate that morph resolution can contribute positively to streaming regulation.

Limitations

We only annotated the live streaming domain where morphs are frequently used to evade censorship, without covering all topics in the live streaming field. Additionally, we validated the effectiveness of our proposed data augmentation method on only three models. In the future, we plan to expand this dataset and continue exploring the linguistic phenomenon of morphs.

Ethics Statement

All data was collected from publicly available sources on the Douyin platform, ensuring no violation of privacy or data protection laws. Our aim is to address false advertising in health and medical live streams, contributing to consumer protection and industry standardization. Furthermore, this work serves the dual purposes of addressing moral concerns and navigating political censorship.

Human annotation was conducted by trained annotators who followed ethical guidelines, and we used large language models to enhance annotation accuracy. No personal or sensitive information was used, and all data was anonymized to prevent misuse.

Our findings support the development of tools to combat deceptive practices in e-commerce live streaming, ultimately benefiting consumers. The dataset and code will be made publicly available following ethical guidelines to encourage further research.

Acknowledgement

This research is partially supported by the National Language Commission of China (ZDI145-71), the National Natural Science Foundation of China (62076217), the Blue Project of Jiangsu and Yangzhou University, and the Top-level Talents Support Program of Yangzhou University,

References

Lauri Auronen. 2003. Asymmetric information: theory and applications. In *Seminar of Strategy and International Business as Helsinki University of Technology*, volume 167, pages 14–18. Citeseer.

CINI Center. 2022. The 50th statistical report on china's internet development. *Beijing2022*.

Monojit Choudhury, Rahul Saraf, Vijit Jain, Animesh Mukherjee, Sudeshna Sarkar, and Anupam Basu. 2007. Investigation and modeling of the structure of texting language. *International Journal of Document Analysis and Recognition (IJDAR)*, 10:157–174.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3029–3051.

Zhifu Gao, Zerui Li, Jiaming Wang, Haoneng Luo, Xian Shi, Mengzhe Chen, Yabin Li, Lingyun Zuo, Zhihao Du, Zhangyu Xiao, and Shiliang Zhang. 2023. Funasr: A fundamental end-to-end speech recognition toolkit. In *INTERSPEECH*.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *International conference on machine learning*, pages 1243–1252. PMLR.

Kenneth Heafield. 2011. Kenlm: Faster and smaller language model queries. In *Proceedings of the sixth workshop on statistical machine translation*, pages 187–197.

Longtao Huang, Ting Ma, Junyu Lin, Jizhong Han, and Songlin Hu. 2019. A multimodal text matching model for obfuscated language identification in adversarial communication? In *The World Wide Web Conference*, pages 2844–2850.

Longtao Huang, Lin Zhao, Shangwen Lv, Fangzhou Lu, Yue Zhai, and Songlin Hu. 2017. Kiem: a knowledge graph based method to identify entity morphs. In *Proceedings of the 2017 ACM on conference on information and knowledge management*, pages 2111–2114.

Heng Ji and Kevin Knight. 2018. Creative language encoding under censorship. In *Proceedings of the First Workshop on Natural Language Processing for Internet Freedom*, pages 23–33.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2. Minneapolis, Minnesota.

Masamune Kobayashi, Masato Mita, and Mamoru Komachi. 2024. Revisiting meta-evaluation for grammatical error correction. *Transactions of the Association for Computational Linguistics*, 12:837–855.

M Lewis. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Gengsong Li, Hongmei Li, Yu Pan, Xiang Li, Yi Liu, Qibin Zheng, and Xingchun Diao. 2022. Name disambiguation based on entity relationship graph in big data. In *International Conference on Data Mining and Big Data*, pages 319–329. Springer.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. 2020. Gector–grammatical error correction: Tag, not rewrite. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–170.

Jipeng Qiang, Yang Li, Chaowei Zhang, Yun Li, Yi Zhu, Yunhao Yuan, and Xindong Wu. 2023a. Chinese idiom paraphrasing. *Transactions of the Association for Computational Linguistics*, 11:740–754.

Jipeng Qiang, Kang Liu, Ying Li, Yun Li, Yi Zhu, Yun-Hao Yuan, Xiaocheng Hu, and Xiaoye Ouyang. 2023b. Chinese lexical substitution: Dataset and method. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 29–42.

Jipeng Qiang, Kang Liu, Yun Li, Yunhao Yuan, and Yi Zhu. 2023c. Parals: Lexical substitution via pretrained paraphraser. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3731–3746.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Mirco Ravanelli, Titouan Parcollet, and Yoshua Bengio. 2019. The pytorch-kaldi speech recognition toolkit. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6465–6469. IEEE.

Ying Sha, Zhenhui Shi, Rui Li, Qi Liang, and Bin Wang. 2017. Resolving entity morphs based on character-word embedding. *Procedia Computer Science*, 108:48–57.

Aobo Wang, Min-Yen Kan, Daniel Andrade, Takashi Onishi, and Kai Ishikawa. 2013. Chinese informal word normalization: an experimental study. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 127–135.

Nannan Wang, Cheng Huang, Junren Chen, and Lingzi Li. 2024. Cmright: Chinese morph resolution based on end-to-end model combined with enhancement algorithms. *Expert Systems with Applications*, page 124294.

Qingyu Wang, Tielin Zhang, Minglun Han, Yi Wang, Duzhen Zhang, and Bo Xu. 2023a. Complex dynamic neurons improved spiking transformer network for efficient automatic speech recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 102–109.

Wenxuan Wang, Jen-tse Huang, Weibin Wu, Jianping Zhang, Yizhan Huang, Shuqing Li, Pinjia He, and Michael R Lyu. 2023b. Mttm: Metamorphic testing for textual content moderation software. In *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*, pages 2387–2399. IEEE.

Pinghui Xiao. 2024. The rise of livestreaming e-commerce in china and challenges for regulation: A critical examination of a landmark case occurring during covid-19 pandemic. *Computer Law & Security Review*, 52:105955.

Ying Xu. 2024. Research on legal regulation of false propaganda behavior in online live streaming sales in china. *Open Journal of Legal Science*, 12:3338.

Taijin Yoon, Sun-Young Park, and Hwan-Gue Cho. 2010. A smart filtering system for newly coined profanities by using approximate string alignment. In *2010 10th IEEE International Conference on Computer and Information Technology*, pages 643–650. IEEE.

Jirong You, Ying Sha, Qi Liang, and Bin Wang. 2018. Morph resolution based on autoencoders combined with effective context information. In *Computational Science–ICCS 2018: 18th International Conference, Wuxi, China, June 11–13, 2018 Proceedings, Part III 18*, pages 487–498. Springer.

Ruoyu Zhang, Yanzeng Li, Yongliang Ma, Ming Zhou, and Lei Zou. 2023. LLMaAA: Making large language models as active annotators. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13088–13103, Singapore. Association for Computational Linguistics.

Zhuosheng Zhang, Hanqing Zhang, Keming Chen, Yuhang Guo, Jingyun Hua, Yulong Wang, and Ming Zhou. 2021. Mengzi: Towards lightweight yet ingenious pre-trained models for chinese. *arXiv preprint arXiv:2110.06696*.

A The annotation Website

We have built a website based on Vue+FastAPI for annotators' labeling work, as shown in Figure 3. Due to the unique nature of the research scenarios, the annotators needed to process multiple modalities of information, which enhanced the quality and accuracy of the annotation results. At the same time, this is a time-consuming task, and we extend our sincerest gratitude to the annotators for their efforts.

B Prompt templates in this paper

ChatGPT-Generate Sentences. The prompting template of ChatGPT-Generate sentences include targets words is shown in Figure 4.

C More Examples

Here, we randomly some samples from morph dataset in Table 6.

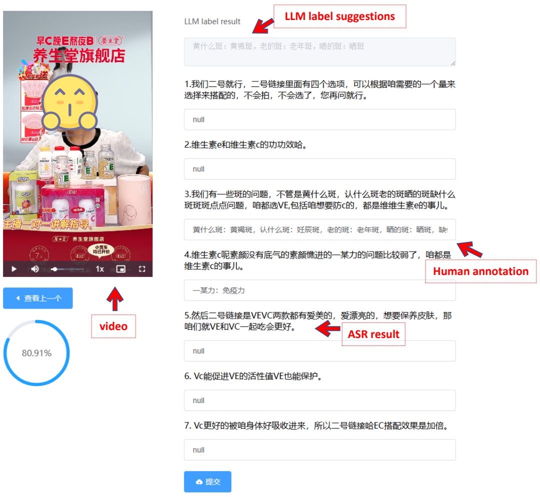
Method	Sentence
Real	BC组合在三号选项三宝贝那维生c呢孩子，我们自己老年人免某粒特别弱，经常被其他人连带，经常阿秋阿秋的。 The BC combination in option three significantly impacts children. Older adults have particularly weak immunity and often catch colds from others. 免某粒(*miǎn mǒu lì:Free of certain pills*):免疫力(*miǎn yì lì, immunity*) 阿秋阿秋(*ā qiū ā qiū,Aqiu Aqiu*):感冒(*gǎn mào,catarrh*)
	都知道用小蓝帽什么意思吧，对不对? You all know what the little blue hat means, right? 小蓝帽(*xiǎo lán mào,small blue hat*):保健食品标志(*bǎo jiàn shí pǐn biāo zhì,Health Supplement Approval Mark*)
	我们一号链接三百一十八米，两桶。 Our link number one is 318 yuan, for two barrels. 米(*mǐ,rice*)元(*yuán,yuan*)

LLM	想要改某善身体某平某衡？试试我们的新品，今天下单有特别优惠，立减50米！
	Want to improve your balance? Try our new product, order today for a special discount of 50 yuan off!
	改某善(*gǎi mǒu shàn,improvement*):改善(*gǎi shàn,improvement*)
	某平某衡(*mǒu píng mǒu héng,balance*):平衡(*píng héng,balance*)
	米(*mǐ,rice*)元(*yuán,yuan*)

	我们的产品专为孕妈妈设计，能够帮助控制糖高，减轻身体猛副某用，让孕期更加轻松。
	Our products are designed specifically for pregnant women to help control hyperglycemia and relieve certain body effects, making pregnancy easier.
	孕妈妈(*yùn mā mā,Pregnant mother*):孕妇(*yùn fù,pregnant*)
	糖高(*táng gāo,high in sugar*):高血糖(*gāo xuè táng,hyperglycemia*)
	猛副某用(*měng fù mǒu yòng,side effect*):副作用(*fù zuò yòng,side effect*)

	运和动不仅有助于心血管健康，还能减少某血某栓形成的风险，百大褂也经常强调这一点。
	Exercise not only helps cardiovascular health, but also reduces the risk of thrombus, which doctors often emphasize.
	运和动(*yùn hé dòng,movement and motion*):运动(*yùn dòng,exercise*)
	某血某栓(*mǒu xuè mǒu shuān,thrombus*)：血栓(*xuè shuān,thrombus*)
	百大褂(*bǎi dà guà,people in white*):医生(*yī shēng,doctor*)

Table 6: Morph sample display: The first row contains sentences with morphs, the second row is the translation, and the third row shows the morph annotation results. "Real" indicates that the data source is real data, not synthetic data. "LLM" indicates data synthesized using an LLM-based method, shown in 5.



LLM label result

黄什么斑：黄褐斑，老的斑：老年斑，晒的斑：晒斑

1.我们二号就行，二号链接里面有四个选项，可以根据咱需要的一个量来选择来搭配的，不会拍，不会选了，您再问就行。

null

2.维生素e和维生素c的功功效哈。

null

3.我们有一些斑的问题，不管是黄什么斑，认什么斑老的斑晒的斑缺什么斑斑斑点点问题，咱都选VE,包括咱想要防c的，都是维维生素e的事儿。

黄什么斑：黄褐斑，认什么斑：妊辰斑，老的斑：老年斑，晒的斑：晒斑，缺

4.维生素c呢素颜没有底气的素颜憔进的一某力的问题比较弱了，咱都是维生素c的事儿。

一某力：免疫力

5.然后二号链接是VEVC两款都有爱美的，爱漂亮的，想要保养皮肤，那咱们就VE和VC一起吃会更好。

null

6. Vc能促进VE的活性值VE也能保护。

null

7. Vc更好的被咱身体好吸收进来，所以二号链接哈EC搭配效果是加倍。

null

Figure 3: Screenshot of an annotation example on the annotation Website. The red text indicates added comments.

Your role is that of a live-streaming host promoting products. You need to generate five promotional sentences that include the target words. Here are some real promotional sentences for you to mimic. The sentences should not have repeated meanings. The target word should remain unchanged. The length of the sentences should be as consistent as possible with the examples provided.
Target Words:
[*Target Words*]
Examples:
[*Examples*]
Generated Sentences:

Figure 4: The prompting template of generating sentences. Generate context-appropriate sentences that contain the specified vocabulary and meet the required quantity.