Chinese Live-Streaming E-Commerce Morph Resolution: Datasets and Methods

Jipeng Qiang, Jiahao Zhu, Yi Zhu*, Chaowei Zhang
School Information and Artificial intelligence, Yangzhou University, China
jpqiang@yzu.edu.cn, mz120231031@stu.yzu.edu.cn, {zhuyi, cwzhang}@yzu.edu.cn

Abstract

Live-stream E-commerce faces significant challenges from morphs, deliberate linguistic variants used to evade real-time voice filters and amplify product claims illegally. While critical for regulatory enforcement, Live Auditory Morph Resolution (LiveAMR) research is hindered by limited datasets: prior work relied on narrow, redundant health domain corpora, restricting model robustness. To bridge this gap, we introduce two datasets: (1) HealthAMR, a refined health-domain corpus via deduplication and re-annotation. (2) GeneralAMR, a general domain benchmark with 28K annotated sentences from 77 channels across 7 E-commerce categories. Further, we propose JointMRE, a multi-task framework that jointly resolves morphs and generates structured explanations, transferring grammatical insights from large language models to enhance generalization. Predictions are refined by our Conflict-aware Dual-output Refinement Framework (CDRF), which detects inconsistencies between corrections and explanations. Experiments show CDRF significantly improves morph resolution accuracy and interpretability. Our datasets and code are available [1].

1 Introduction

Morphs are deliberate linguistic variants that substitute standard words with acoustically, visually, or semantically related alternatives, allowing speakers to obscure or disguise sensitive information. Main research on text morphing focused on written domains such as social media or underground forums, where variants like *weather* → *waether* or splitting strategies (e.g., 胡(*hú*) → 古月(*gǔ yuè*)) were used to evade keyword-based filters (Hsiung et al., 2005; Pantel, 2006; Zhang et al., 2014). These studies demonstrated that morphs can effectively bypass content moderation and facilitate covert communication, but largely ignored spoken contexts and the unique challenges they present.

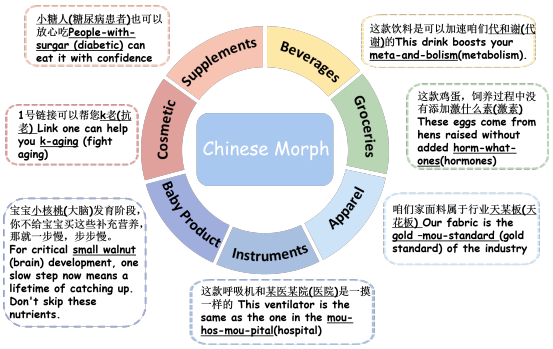

Figure 1: The corresponding morph example cases in different domains. Morphs are marked with an underline, with their corresponding original words provided in parentheses.

With the explosive growth of live-stream E-commerce, where hosts may sell over ten billion items annually in China alone, morphing has migrated into the auditory domain. In this setting, speakers employ phonetic or filler-word tweaks (e.g., "手术" (*shǒu shù*, "surgery") becomes "手某术" (*shǒu mǒu shù*)) to slip past real-time voice filters and make exaggerated product claims without detection (Zhu et al., 2025; Center, 2022). Such practices not only deceive consumers but also undermine regulatory enforcement, making accurate Live Auditory Morph Resolution (LiveAMR) critical for safeguarding E-commerce integrity.

Despite its importance, this area of research for LiveAMR has not received sufficient attention. We only found a single-domain dataset limited to health-related streams (Zhu et al., 2025). Through our analysis, we found that it contains 7,812 clips from only 25 channels, suffers from high redundancy, and exhibits a skewed variant distribution, with 47 base terms accounting for half of all morph forms. This narrow scope hampers model robust-

*Corresponding author
[1] https://github.com/loopback00/Morph-Resolution-Datasets-and-Methods

Findings of the Association for Computational Linguistics: ACL 2026, pages 2632–2645

ness and fails to reflect the breadth of morph usage across diverse product categories and platform conventions. As shown in Figure 1, many livestreams across different domains use morphs. Some of these morphs are shared across domains, while others are specific to individual fields.

To fill this gap, we introduce two complementary contributions. First, we refine and extend the data foundation from this work (Zhu et al., 2025). We apply automated deduplication and two-stage re-annotation to cleanse the original corpus, yielding *HealthAMR*. We then expand to a new multi-domain benchmark, *GeneralAMR*, by collecting 2,752 one-minute clips from 77 channels across three major Chinese platforms (Douyin, Kuaishou, Jingdong) and seven E-commerce domains (groceries, cosmetics, baby products, supplements, instruments, beverages, apparel). The combined corpus comprises over 28,000 sentences (3,092 positive morph instances, 24,949 negatives), covering 567 base words and 1,309 distinct variants. This realistic, high-quality dataset enables rigorous evaluation of generalization induced morph diversity.

Second, we develop novel methods to tackle the heightened generalization challenge. Standard fine-tuning on diverse data often learns superficial correlations rather than the underlying transformation logic (McCoy et al., 2019; Pruksachatkun et al., 2020; Wang et al., 2024). Drawing inspiration from human experts who provide both corrections and their rationales, we propose JointMRE, a multi-task framework that jointly normalizes morphs and generates structured explanations including (morph, morph type, original word) triplets. By explicitly teaching models to "explain" their morph normalization decisions, we transfer rich domain and lexical insights from large language models to fine-tune language models, enhancing both resolution accuracy and interpretability in Chinese live-streaming E-commerce.

Building on this, our Conflict-aware Dual-output Refinement Framework (CDRF) serves as a lightweight second stage: it uses the explanations produced by JointMRE to identify, arbitrate, and apply targeted edits, ensuring consistency between predictions and rationales without full sentence regeneration. This three-step process includes conflict identification, arbitration, and correction, which is executed in a single LLM call to maintain operational efficiency. Comprehensive experiments demonstrate that our methods achieve state-of-the-art results on both datasets. JointMRE improves sentence-level $F_{0.5}$ by 1.93–3.12 points over baselines, while CDRF further boosts gains by 2.8–5.8 points via conflict arbitration.

2 Related Work

Morphing, the intentional modification of text for concealment or evasion, has been observed in many languages, though most research has focused on English. Early English studies examined both character-level variants (e.g., *weather-> waether*) and word-level splits (e.g., *weather-> wea ther*), leveraging resources such as Aspell dictionaries, FastText embeddings (Joulin et al., 2017), and handcrafted orthographic features (van der Goot, 2019). More recent approaches exploit pre-trained language models (PLMs), casting morph resolution as either sequence to tagging (Muller et al., 2019) or sequence to sequence generation (Lourentzou et al., 2019; Samuel and Straka, 2021). However, these studies focus predominantly on alphabetic scripts and on variants arising in written social-media or forum contexts (Eisenstein, 2013; Baldwin et al., 2015).

By contrast, Chinese presents unique challenges: its logographic characters lack clear subword boundaries, and morph variants often involve homophonic or semantic substitutions that cannot be captured by methods designed for alphabetic languages. Main Chinese work has concentrated on social media normalization, e.g., Weibo data (Huang et al., 2013; You et al., 2018), and on industry-specific argot in underground markets (Xu et al., 2021; Wang et al., 2024). To tackle this, early work by (Huang et al., 2013) proposed a similarity-based approach, incorporating surface-level, meta path based semantic, and social relevance features. Subsequently, (You et al., 2018) introduced a method that leverages autoencoders to generate embeddings for morphs and their candidates, enriched by effective contextual information. More recently, (Wang et al., 2024) framed the problem as an alignment task, utilizing translation models to map morphs directly to their target forms.

Live-streaming scenarios introduce further complexity: rapid speech, background noise, and domain-specific jargon result in highly diverse morph phenomena that are not addressed by existing corpora or models. (Zhu et al., 2025)introduced the first health domain AMR dataset, created from ASR transcripts of live-streams, and proposed a

Dataset	HealthAMR	GeneralAMR
Source		
Channels	25	77
Domain		
Videos	7812	2752
Instances	64476	28041

Table 1: A comparison of HealthAMR and GeneralAMR. represents the platforms Kuaishou, Douyin, and Jingdong, respectively. indicates 7 domains: groceries, supplements, cosmetics, baby products, beverages, instruments, and apparel.

data augmentation strategy using LLM-generated synthetic samples. The dataset's quality is unverified due to the lack of a secondary annotation pass, and the synthetic data offers limited improvement to model generalization.

3 Dataset Curation and Expansion

This section details the development of two LiveAMR datasets (HealthAMR and GeneralAMR).

3.1 HealthAMR: Expanded Coverage and Quality Refinement

The previous AMR corpus (Zhu et al., 2025), despite its value for health-domain morph resolution, is constrained by two primary issues. First, sourcing 7,812 video clips from only 25 live-stream channels has produced extensive redundancy in both content and linguistic patterns. Second, its morph-variant lexicon is heavily skewed: 47 base terms account for 1,323 variant forms—half of all observed morphs, thereby narrowing the dataset's representational diversity.

To overcome these limitations, we enacted a two-stage refinement pipeline. In Stage 1, an automated deduplication procedure excised 22,404 duplicate samples across training, validation, and test splits. In Stage 2, annotators performed a detailed manual audit, identifying and correcting 932 previously unannotated positive instances and verifying overall label accuracy. The resulting health domain subset combines broader coverage with significantly reduced redundancy and heightened annotation fidelity, furnishing a more robust foundation for morph resolution research. We refer to this refined collection as *HealthAMR*. Some statistical information of HealthAMR is shown in Tables 1 and 2.

3.2 Building GeneralAMR: Rationale and Construction Pipeline

Hosts on live-stream E-commerce platforms routinely use morphs not only in health supplements but also in domains such as groceries, baby products, cosmetics, and apparel. In order to overcome the narrow focus of the original HealthAMR, which only captures morph variants in the health domain. We have constructed *GeneralAMR*, a more comprehensive LiveAMR dataset. Moreover, different platforms employ distinct monitoring and content moderation strategies, making a multi-platform collection essential for realistic morph resolution research.

Data Collection. We sourced raw video data from three major Chinese E-commerce platforms (Douyin, Kuaishou, and Jingdong) to ensure broad coverage of diverse live-streaming environments. From 72 channels spanning seven high-transaction domains (food, alcoholic beverages, health supplements, medical devices, cosmetics, clothing, and baby products), we extracted 2,752 one-minute clips (Ihmily, 2025). These channels were initially sourced from the top 100 top-selling channels in each target domain. To qualify for inclusion, the selected channels also had to meet three further criteria: (1) over 10,000 followers, (2) more than 50 concurrent viewers, and (3) confirmed use of morphs within a five-minute manual observation. These domains were selected due to their high transaction volumes and frequent use of morphs to promote products or skirt platform rules.

ASR Transcription. Using FunASR's Paraformer model (Gao et al., 2022, 2023), we converted all video audio to text. The variability introduced by host accents, rapid speech, and background noise generates a richer set of morph forms compared to controlled recordings.

Human Annotation. To accurately identify and normalize morphs, we developed a web-based annotation platform that displays each ASR transcript alongside its source clip. We pre-annotated candidates with a fine-tuned model from (Zhu et al., 2025), then allowed annotators to accept, modify, or reject suggestions. To ensure annotation quality, we recruited three graduate students from our team to label the dataset. Before they began, we provided detailed annotation guidelines, and they

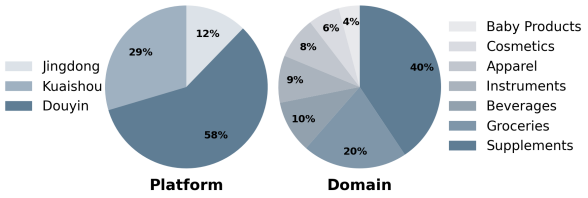

Figure 2: The pie chart on the left illustrates the percentage of data by platform, while the chart on the right displays the percentage of data by domain.

were compensated at a rate of 20 CNY per hour, which exceeds the local minimum wage standard.

Every sample underwent two independent annotation passes. In the second pass, annotators reviewed data labeled by a different annotator. Discrepancies were resolved through consensus meetings with domain experts.

Dataset Statistics. The final GeneralAMR comprises 28,041 samples: 3,092 positive (morph present) and 24,949 negative (no morph), covering 567 base words and 1,309 distinct morph variants. Compared to HealthAMR's single-domain focus, GeneralAMR spans 77 channels across three platforms and seven E-commerce sectors, as shown in Table 1. The pie chart on the left in Figure 2 illustrates the distribution ratio of the data across the platforms, while the pie chart on the right presents the data distribution within each domain.

3.3 Dataset Partitioning and Quality Evaluation

Dataset	Usage	Sents	ASR Model	Ori Num
HealthAMR	Train	4,863&55,619	Paraformer	372
	Valid	805&789	Paraformer	136
	Test1	821&779	Paraformer	142
	Test2	398&402	Kaldi	85
GeneralAMR	Train	2135&23,599	Paraformer	567
	Valid	400&400	Paraformer	126
	Test1	700&700	Paraformer	209
	Test2	250&250	Whisper	87

Table 2: Dataset statistics for HealthAMR and GeneralAMR settings. "Sents" denotes the number of positive & negative samples; "ASR Model" indicates the transcription system used; "Ori Num" is the count of unique original words that are morphed, reflecting lexical diversity.

Dataset Partitioning. We observed that the same audio segment can yield phonetically similar but textually divergent outputs when processed by different ASR systems. This is mainly because there is usually only one correct expression, whereas incorrect expressions can vary widely, requiring the ASR system to identify and interpret a diverse range of possible errors. For example, the cataract morph "白某障" (*bái mǒu zhàng*) may be rendered by different models as "白母障" (*bái mǔ zhàng*), "白某张" (*bái mǒu zhāng*), or "百某章" (*bǎi mǒu zhāng*). To address this situation, constructing an additional evaluation test set can help verify the generalization ability of the method.

The datasets (HealthAMR and GeneralAMR) consist of Training, Validation, Test1, and Test2 splits. Test1 uses the same ASR model as training, while Test2 employs a different ASR model to evaluate cross-model generalization and robustness on the morph resolution task. Incorporating Test2, transcribed entirely with an alternative ASR allows us to rigorously assess our model's ability to handle such variability. The detailed dataset composition is summarized in Table 2.

F1-Score (A1-A2)	F1-Score (A1-A3)	F1-Score (A2-A3)	Average
0.86	0.92	0.88	0.88

Table 3: F1-Score agreement scores for pairs of annotators and average score for three annotators.

Data Quality Analysis. To verify the quality of the constructed dataset, we conducted an inter-annotator agreement (IAA) (Artstein, 2017) test to ensure the reliability and consistency of our annotation results. To evaluate the consistency among three annotators, we computed pairwise F1-score by treating each annotator's labels as predictions against others, as shown in Table 3. This statistic underscores a substantial level of agreement among our human annotators.

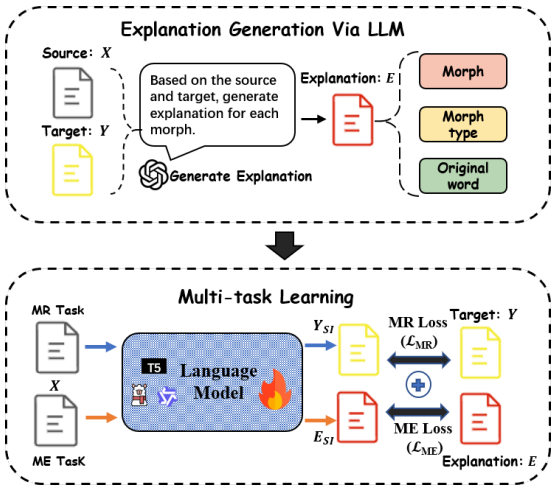

Figure 3: An overview of our proposed method, JointMRE. MR and ME are morph resolution and morph explanation tasks, respectively.

4 Methods: JointMRE and CDRF

Below, we will provide a detailed introduction to the two methods we propose: JointMRE and CDRF.

4.1 JointMRE: Expert-Inspired Morph Explanation for LiveAMR

First, consider how humans correct unfamiliar morphs: they consult an expert, who not only provides the correct form but also explains the reasoning behind the change. Inspired by this process, we leverage large language models (LLMs) as "explainers" that encode rich knowledge. By extracting morph explanations from LLMs and using them to guide small or large models during training, we transfer sophisticated transformation rules that improve both accuracy and generalization. The method is denoted as JointMRE, shown in Figure 3.

Morph Resolution (MR). Given an input sentence $X = [x_1, \ldots, x_n]$ containing morphs, the goal of MR is to generate the corrected sentence $Y = [y_1, \ldots, y_m]$.

Morph Explanation (ME) Task. Given an input sentence X, the goal of ME is to produce a structured explanation $E = [e_1, \ldots, e_k]$, which details all triples including:

- **Morph:** non-standard surface form,
- **Type:** morph category (homophone, synonym, transformation),
- **Original Word:** the standard term represented.

For example, in "白褂褂也会推荐你吃这个" ("The people in white will also recommend this"), the explanation is *"people in white"* | *synonym* | *"doctor"*.

Gold Explanation Construction. As shown in Figure 3, we first need to generate gold explanations for the ME task. Here, we prompt an LLM with each source sentence (X) and its target (Y) to generate candidate explanations, then manually verify them to ensure high-quality supervision that captures the LLM's transformation reasoning and morphing knowledge. The prompt template is shown in Appendix A.4.

JointMRE Multi-Task Learning. JointMRE trains a single model on two complementary tasks (Morph Resolution and Morph Explanation) using tailored strategies for PLMs and LLMs, respectively.

(1) PLM Fine-Tuning. We fine-tune a sequence-to-sequence PLM (e.g., T5) by casting both tasks as text generation problems with task-specific prefixes:

- `<resolution>` X: generate the normalized sentence Y.
- `<explanation>` X: generate the structured explanation sequence E (or output "no morphs" if none are present).

(2) LLM LoRA Fine-Tuning. To efficiently adapt LLMs (e.g., Llama3.1, Qwen2.5) to both tasks, we employ Low Rank Adaptation (LoRA). We fine-tune the model with two distinct prompt templates (shown in Appendix A.2). Each LoRA augmented pass updates a small subset of the model's parameters, ensuring efficient training while preserving the LLM's general knowledge.

Training Objective. The combined loss encourages accurate resolution and faithful explanation:

$$\begin{aligned}\mathcal{L}_{\text{JointMRE}} &= \mathcal{L}_{\text{MR}} + \mathcal{L}_{\text{ME}} \\ &= -\sum_{t=1}^{m} \log P(y_t \mid Y_{<t}, X) \\ &\quad -\sum_{j=1}^{k} \log P(e_j \mid E_{<j}, X) \end{aligned} \quad (1)$$

By emulating expert behavior and harnessing LLM explanations, JointMRE not only produces a corrected sentence Y_{SI} but also provides human-interpretable rationales E_{SI}, thereby equipping small models with the grammatical reasoning needed for robust morph resolution.

4.2 CDRF: Conflict-aware Dual-output Refinement Framework

While JointMRE delivers both a preliminary resolved sentence Y_{SI} and its explanation E_{SI}, these dual outputs may occasionally disagree, undermining reliability. Inspired by expert reviewers who cross-check corrections against their reasoning, we propose a Conflict-aware Dual-output Refinement Framework (CDRF) to detect and resolve inconsistencies between Y_{SI} and E_{SI}, yielding a more trustworthy final output. Crucially, all three refinement steps include conflict identification, arbitration, and targeted correction. These are executed in a single

Methods	HealthAMR Test1						HealthAMR Test2					
	C-P	C-R	C-$F_{0.5}$	S-P	S-R	S-$F_{0.5}$	C-P	C-R	C-$F_{0.5}$	S-P	S-R	S-$F_{0.5}$
LLMs												
GPT-4.1-Mini	37.21	58.22	40.10	55.52	42.36	55.27	33.30	49.00	35.58	43.40	33.83	41.08
Gemini-2.5-Flash	40.45	61.40	43.42	59.08	44.07	55.13	43.92	63.19	46.78	53.06	39.19	49.55
DouBao-1.6	41.32	54.49	43.42	68.08	35.07	57.30	49.37	60.12	51.20	69.95	40.95	61.27
Deepseek-V3.2-Exp	42.12	57.13	44.45	69.14	38.48	59.64	50.26	59.05	51.80	72.76	38.94	62.00
Fine-Tuning T5												
T5 (Zhang et al., 2021)	90.70	80.12	88.37	94.57	76.67	90.35	59.39	47.69	56.61	67.17	33.08	55.69
DataAug (Zhu et al., 2025)	90.78	81.64	88.79	94.76	77.41	90.70	59.43	48.00	56.73	66.41	33.94	55.75
EXAM (Li et al., 2025)	90.09	81.16	88.79	94.81	78.14	90.93	59.29	48.81	56.70	67.11	34.56	56.05
EditTask (Bout et al., 2023)	90.93	81.16	88.79	95.11	78.19	91.16	59.70	48.31	57.01	68.11	34.34	56.85
JointMRE	**91.77**	81.51	**89.52**	96.01	79.17	92.09	**65.40**	48.76	60.72	72.44	**40.95**	57.75
CDRF	90.28	**85.00**	89.17	**96.49**	**80.31**	**92.75**	64.90	**54.45**	**62.50**	**80.69**	**40.95**	**67.57**
Fine-Tuning LLM												
Llama3.1	86.95	89.17	87.39	93.42	84.09	91.39	79.14	65.18	75.89	96.56	55.38	84.03
JointMRE (Llama3.1)	**87.28**	90.58	87.92	95.90	84.98	93.50	79.89	68.87	77.41	95.96	59.64	85.54
CDRF (Llama3.1)	86.83	**96.30**	**88.57**	**95.96**	**89.76**	**94.65**	**87.48**	**96.47**	**89.14**	**96.94**	**87.68**	**94.94**
Qwen2.5	87.12	91.11	87.89	95.62	84.13	93.08	78.51	71.17	76.92	95.34	61.65	85.95
JointMRE (Qwen2.5)	**87.22**	92.52	88.23	96.84	84.16	94.05	77.91	71.93	76.63	95.48	63.65	86.80
CDRF (Qwen2.5)	86.09	**98.06**	**88.24**	**96.89**	**87.33**	**94.81**	**83.72**	**94.63**	**85.69**	**96.44**	**81.90**	**93.14**

Table 4: Performance of all methods on HealthAMR. Best results are in **bold** and second-best are underlined. Key evaluation scores $F_{0.5}$ are highlighted in gray.

LLM invocation to maintain practical efficiency. The used prompt is shown in Appendix A.4.

Conflict Identification. First, we assess whether Y_{SI} and E_{SI} are consistent. We input the source sentence X, Y_{SI}, and E_{SI} into an LLM module that outputs a set of edit tuples:

$$\mathcal{D} = \text{LLM}_{\text{identify}}(X, Y_{\text{SI}}, E_{\text{SI}}),$$

where each $d = \{m, t, o_y, o_e\}$ records a morph m, its type t, and the two conflicting original words o_y (from Y_{SI}) and o_e (from E_{SI}). If $\mathcal{D}$ is empty, Y_{SI} is accepted as the final output; otherwise, we proceed to arbitration.

Conflict Arbitration. For each conflict $d \in \mathcal{D}$, we form two candidate triplets:

$$d_y = \{m, t, o_y\}, \quad d_e = \{m, t, o_e\}.$$

An LLM arbitrator evaluates which triplet exhibits greater self-consistency, meaning any two elements logically imply the third:

$$\mathcal{P}_{\text{CA}} = \text{LLM}_{\text{arbitrate}}(X, d_y, d_e),$$

where $\mathcal{P}_{\text{CA}} = 0$ selects d_y and otherwise selects d_e. Aggregating these choices yields the resolved edit set $\mathcal{D}'$.

Targeted Correction. Finally, we apply each validated edit $d' \in \mathcal{D}'$ directly to Y_{SI} by instructing the LLM to substitute the morph with its chosen standard form. This targeted approach preserves the remainder of the sentence and avoids unwanted changes, producing the refined output Y_{SII}.

5 Experiments

5.1 Experimental Setup

Evaluation Metrics. We adopt both sentence-level and character-level metrics to evaluate LiveAMR performance.

(1) Sentence-Level Evaluation. Measures whether an entire sentence has been correctly normalized without introducing erroneous changes. We report precision (S-P), recall (S-R), and a combined $F_{0.5}$ score (S-$F_{0.5}$), which weights precision twice as heavily as recall to penalize overcorrection.

(2) Character-Level Evaluation. Captures more granular performance by comparing individual characters in the predicted output against the gold standard. We similarly report character-level precision (C-P), recall (C-R), and $F_{0.5}$ (C-$F_{0.5}$).

Comparison Methods. We organize baselines into three categories:

(1) Few-Shot LLMs. We evaluate the performance of GPT-4.1-mini, Gemini-2.5-Flash, Doubao-1.6, and Deepseek-v3.2-Exp in an 8-shot setting. Prompts are detailed in Appendix A.4.

(2) Fine-Tuned PLMs. T5 base model (Zhang et al., 2021), fine-tuned directly on the LiveAMR datasets.

DataAug (Zhu et al., 2025), which augments training data via LLM generation.

EXAM (Li et al., 2025), which prepends LLM-generated rationales to the input.

Methods	GeneralAMR Test1						GeneralAMR Test2					
	C-P	C-R	C-$F_{0.5}$	S-P	S-R	S-$F_{0.5}$	C-P	C-R	C-$F_{0.5}$	S-P	S-R	S-$F_{0.5}$
LLMs												
GPT-4.1-Mini	30.42	39.42	31.88	50.00	29.42	43.86	23.72	37.22	25.57	32.90	21.98	29.92
Gemini-2.5-Flash	43.02	42.15	42.84	67.60	31.00	54.68	30.22	45.83	32.43	46.15	33.62	42.95
DouBao-1.6	37.78	41.33	38.44	66.78	27.85	52.19	32.77	38.42	33.76	58.41	25.99	46.75
Deepseek-V3.2-Exp	42.21	41.79	42.13	72.24	30.29	55.65	35.66	37.57	36.02	59.61	27.31	48.21
Fine-Tuning T5												
T5 (Zhang et al., 2021)	75.05	65.33	72.88	98.51	56.71	85.85	17.97	50.28	30.62	47.39	35.34	44.37
DataAug (Zhu et al., 2025)	79.12	63.32	75.37	98.51	57.28	86.72	17.97	50.28	30.62	47.39	34.94	44.14
EXAM (Li et al., 2025)	79.37	62.67	75.36	99.49	56.81	86.50	18.00	50.31	30.66	47.59	35.00	44.27
EditTask (Bout et al., 2023)	73.71	63.96	71.53	99.49	56.14	86.18	20.16	48.16	29.81	45.83	35.30	43.48
JointMRE	**80.27**	64.96	**76.66**	99.52	59.28	87.62	27.13	50.96	33.62	46.92	40.42	47.56
CDRF	77.93	**67.61**	76.58	**99.54**	**62.85**	**88.97**	**57.30**	**57.78**	**57.40**	**72.22**	**50.43**	**66.47**
Fine-Tuning LLM												
Llama3.1	73.29	67.61	72.08	93.83	60.85	84.65	64.87	56.94	63.12	88.28	48.49	75.83
JointMRE (Llama3.1)	75.15	70.89	74.25	**94.75**	64.57	86.65	66.96	63.06	66.14	86.61	52.78	76.77
CDRF (Llama3.1)	**87.30**	**91.61**	**88.13**	90.85	**86.71**	**92.87**	**76.44**	**88.33**	**78.56**	**88.88**	**79.31**	**86.79**
Qwen2.5	73.37	69.83	72.63	94.18	62.33	85.45	66.37	62.50	65.56	87.85	52.78	77.53
JointMRE (Qwen2.5)	75.57	72.26	74.89	94.93	64.19	86.63	67.89	66.94	67.70	83.97	56.23	76.42
CDRF (Qwen2.5)	**83.77**	**90.88**	**85.10**	**94.84**	**81.42**	**91.81**	**81.38**	**88.61**	**82.73**	**94.24**	**77.58**	**90.36**

Table 5: Performance of all methods on GeneralAMR. Best results are in **bold** and second-best are underlined. Key evaluation scores $F_{0.5}$ are highlighted in gray.

EditTask (Bout et al., 2023), which introduces auxiliary edit-based tasks during training.

(3) Fine-Tuned Open-Source LLMs. Llama3.1-8B (Grattafiori et al., 2024) and Qwen2.5-7B (Yang et al., 2024), both fine-tuned on the same LiveAMR data under identical hyperparameters (see Appendix A.2).

(4) JointMRE and CDRF: They are offered in three versions, differing only in the choice of the language model for JointMRE. Specifically, we provide implementations based on a PLM (T5) and two LLMs (Llama3.1-8B and Qwen2.5-7B). GPT-4.1-Mini in CDRF is used to generate conflict-aware refinement. All results are obtained from a single run under the same experimental setting.

5.2 Results and Analysis

We first present results on the HealthAMR benchmark (Table 4), then on GeneralAMR (Table 5).

HealthAMR Performance. On HealthAMR Test1 and Test2, fine-tuned PLMs (T5 and its enhancements) substantially outperform zero-/few-shot LLMs, underscoring the value of task-specific training. Among PLM baselines, JointMRE sets a new high for sentence-level $F_{0.5}$ (92.09% on Test1, 57.75% on Test2), demonstrating that explanation supervision boosts generalization even within a single domain.

Introducing CDRF yields further gains: on Test1, CDRF raises sentence-level $F_{0.5}$ to 92.75%, driven by improved recall (80.31%) without sacrificing precision. On Test2, where ASR variability is greater, CDRF maintains robustness, achieving 67.57% $F_{0.5}$ and outperforming all other methods. This indicates that conflict-aware refinement effectively leverages explanations to correct residual errors. Fine-tuned LLMs follow the same pattern.

GeneralAMR Performance. GeneralAMR poses a more challenging, multi-domain evaluation. On Test1, PLM baselines drop substantially (T5 yields 85.85% sentence-level $F_{0.5}$), whereas JointMRE recovers much of this gap, achieving 87.62%. CDRF dramatically boosts performance to 88.97%, a 3.12-point improvement over JointMRE, driven by both higher precision (99.54%) and recall (62.85%). This demonstrates CDRF's ability to harness explanations to resolve domain-specific ambiguities.

On GeneralAMR Test2, which uses alternative ASR transcripts, PLMs see further degradation, but CDRF again proves most robust (66.47% sentence-level $F_{0.5}$), surpassing all baselines by a wide margin. Similarly, fine-tuned LLMs benefit from our methods.

5.3 Ablation Study

Impact of Explanation Components. Table 6 shows that each explanation element (morph, morph type, and original word) contributes to improved morph resolution. Including only the morph or type yields solid gains, but adding the original word information further enhances character-

Methods			Metrics					
m	**t**	**o**	C-P	C-R	C-$F_{0.5}$	S-P	S-R	S-$F_{0.5}$
HealthAMR								
✓			91.72	80.99	89.36	96.22	77.58	91.81
	✓		**92.31**	81.34	89.88	96.23	77.83	91.88
✓	✓		91.74	81.16	89.41	95.62	77.22	91.27
✓		✓	92.20	**82.22**	**90.02**	95.34	**79.90**	91.79
✓	✓	✓	91.77	81.50	89.52	**96.49**	79.17	**92.09**
GeneralAMR								
✓			**80.58**	63.23	76.39	99.50	57.85	86.98
	✓		79.52	63.41	75.68	99.49	56.57	86.38
✓	✓		79.16	63.41	75.41	99.50	57.42	86.78
✓		✓	78.84	63.24	75.13	99.50	57.85	86.98
✓	✓	✓	80.27	**64.96**	**77.66**	**99.54**	**62.76**	**87.62**

Table 6: Impact of explanation components in the JointMRE. **m**, **t**, and **o** represent morph, morph type, and original word, respectively. The experiment was performed on the T5 model in Test1.

Model	Metrics			
	M-F_1	T-F_1	O-F_1	E-F_1
HealthAMR				
T5	70.16	86.45	85.08	41.08
Llama3.1	**74.98**	**87.41**	**85.14**	**46.99**
Qwen2.5	72.80	87.14	83.28	44.54
GeneralAMR				
T5	73.67	86.92	72.62	41.10
Llama3.1	72.99	85.34	75.41	42.29
Qwen2.5	**75.57**	**87.04**	**76.31**	**44.01**

Table 7: Evaluation of explanation quality on Test1. M-F_1, T-F_1, and O-F_1 measure the performance of identifying the morph, morph type, and original word individually. E-F_1 evaluates the accuracy of the entire triplet.

level accuracy and sentence-level precision. The full triplet configuration consistently achieves the best balance of precision and recall across both HealthAMR and GeneralAMR, confirming that comprehensive explanations are crucial for robust model generalization.

Impact of Explanation Quality on CDRF Gains. To understand why small and large models yield similar performance for JointMRE yet diverge when used in CDRF, we first evaluate their ability to generate accurate explanations and then observe how CDRF leverages those explanations.

For the sentences in Test1, we employed human experts to manually annotate the golden explanations. Then, we split each explanation into its *morph*, *morph type*, and *original word* components. For each component and for the full triplet, we perform binary classification: does the model's predicted element match the gold? We report F_1 scores for each (M-F_1, T-F_1, O-F_1) and for the entire explanation triplet (E-F_1) on Table 7.

Larger LLMs produce noticeably more accurate explanations (higher E-F_1) than T5. When these explanations feed into CDRF, LLMs yield substantial refinement gains, whereas T5, whose explanations are less precise, shows minimal improvement. This contrast confirms that CDRF effectively uses explanation quality as a supervisory signal: the more reliable the explanation, the more CDRF can enhance final resolution results.

Models	Metrics					
	C-P	C-R	C-$F_{0.5}$	S-P	S-R	S-$F_{0.5}$
HealthAMR						
No Model	77.91	71.93	76.63	95.48	63.65	86.80
Qwen-7b	82.50	95.42	84.49	94.63	83.80	92.24
Qwen-14b	84.0	97.54	86.40	96.44	85.99	94.15
Qwen-32b	83.90	97.71	86.34	96.58	86.23	94.32
Gemini-2.5-Flash	85.90	95.42	87.65	95.83	89.64	94.52
Grok4-Fast	89.53	98.47	91.20	96.30	92.98	94.61
GPT4.1-mini ⋆	86.09	98.06	88.24	96.89	87.33	94.81
GeneralAMR						
No Model	67.89	66.94	67.70	83.97	56.23	76.42
Qwen-7b	80.11	90.42	81.98	93.29	79.57	90.18
Qwen-14b	82.79	93.52	84.74	94.34	83.42	91.91
Qwen-32b	81.15	90.33	82.83	94.65	81.00	91.56
Gemini-2.5-Flash	86.91	87.86	87.10	95.19	82.14	92.26
Grok4-Fast	90.87	97.34	92.09	95.61	92.39	94.95
GPT4.1-mini ⋆	83.77	90.88	85.10	94.84	81.42	91.81

Table 8: Performance of different LLMs employed in CDRF. "No Model" represents the result of using only the output from JointMRE. ⋆ indicates the model used for the reasoning stage in our CDRF experiments.

Impact of Reasoning Model Selection in CDRF. In our CDRF, we adopt GPT-4.1-mini as the dedicated reasoning model for conflict identification and arbitration. Table 8 illustrates the results using different open-source LLMs. The results reveal that swapping in different models produces only minor variations in both character- and sentence-level metrics. This consistency demonstrates that CDRF's targeted refinement mechanism is largely agnostic to the specific LLM used, underscoring the framework's robustness and practical flexibility.

Efficiency Study. We evaluated the deployment efficiency of the selected models by benchmarking performance over 5 rounds at 100 concurrency. Specifically, we used `vLLM v0.11.0` for the Qwen and Llama models, and `Mosec v0.9.5` for the T5 model. The results, including average throughput, queries per second (QPS), latency, and GPU memory usage, are summarized in Table 9.

Effectiveness of Reasoning Strategies in CDRF. Table 10 demonstrates that our conflict-aware refinement consistently outperforms Direct Answer, CoT, and CoT+RAG across two datasets. In HealthAMR, we achieve 88.57%

Model	Task	Avg. Throughput	QPS	Latency (s)		GPU Memory
		(tokens/s)	(r/s)	Single Request	Avg. Concurrent	(GB)
Qwen2.5-7b-MRE	Correct	1221.4	22.22	1.07	4.64	44.8
	Explain	1462.3	59.41	0.37	1.62	44.8
	Arbitrate	962.4	8.15	2.585	12.14	43.7
Qwen2.5-14b	Arbitrate	779.8	5.00	5.553	20.01	45.3
Llama3.1-8b-MRE	Correct	1103.2	15.21	1.44	6.46	45.2
	Explain	1394.3	50.47	0.52	1.89	45.2
T5-MRE	Correct	1328.6	25.34	1.12	4.24	1.61
	Explain	1294.2	48.59	0.45	1.83	1.45
Gpt4.1-mini	Arbitrate	/	/	6.05	/	/
Gemini2.5-Flash	Arbitrate	/	/	5.03	/	/
Grok4-Fast	Arbitrate	/	/	3.34	/	/

Table 9: Deployment Efficiency of Models Benchmarked at 100 Concurrency. Performance metrics for closed models are marked with / as it was accessed via an external API service.

Methods	Metrics					
	C-P	C-R	C-$F_{0.5}$	S-P	S-R	S-$F_{0.5}$
HealthAMR						
Direct Answer	81.65	90.85	83.33	94.12	81.97	91.41
COT	80.02	90.93	81.98	93.70	81.60	91.00
COT+RAG	84.85	91.73	86.14	95.42	83.80	92.84
Ours	**86.83**	**96.30**	**88.57**	**95.96**	**89.76**	**94.65**
GeneralAMR						
Direct Answer	69.22	70.99	69.56	89.33	63.42	82.58
COT	69.03	71.99	69.60	90.32	62.71	83.01
COT+RAG	71.37	73.45	71.77	**93.88**	63.57	85.70
Ours	**87.30**	**91.61**	**88.13**	90.85	**86.71**	**92.87**

Table 10: Comparison of different reasoning strategies in CDRF based on Llama3.1 output. **Direct Answer** refers to direct generation, **CoT** denotes Chain of Thought prompting, and **CoT+RAG** represents BM25 Retrieval Augmented Chain of Thought.

character-level $F_{0.5}$ and 94.65% sentence-level $F_{0.5}$, showing clearer, more precise corrections. In GeneralAMR, we gain are even larger—88.13% C-$F_{0.5}$ and 92.87% S-$F_{0.5}$, highlighting robustness in diverse, multi-domain settings. These results confirm that our targeted edit arbitration outperforms both standard and retrieval-augmented reasoning prompts by enforcing consistency between resolution and explanation.

Cross-Lingual Generalization. To evaluate the cross-lingual potential of our method, we conducted experiments on the Vietnamese lexical normalization dataset, ViLexNorm (Nguyen et al., 2024) (consisting of 84k training samples, with 1k samples each for development and testing). Notably, the ViLexNorm test set comprises exclusively positive samples, which explains the high precision scores due to the absence of negative samples. We employed settings aligned with our interpretable framework, classifying error types into phonetic, visual, and semantic categories. In CDRF, GPT-4.1-mini serves as the arbitrator, while the RAG strategy utilizes BM25 retrieval with $k = 8$. The results demonstrate that our CDRF method exhibits strong cross-lingual generalization capabilities as shown in Table 11.

Methods	Metrics					
	C-P	C-R	C-$F_{0.5}$	S-P	S-R	S-$F_{0.5}$
ViLexNorm						
Gpt4.1-mini(RAG)	88.69	92.13	89.36	100.0	82.24	95.86
ViT5	80.98	77.38	80.24	**100**	62.66	89.35
JointMRE (viT5)	81.64	83.03	81.92	**100**	66.24	90.75
CDRF (viT5)	**96.8**	**98.52**	**97.14**	**100**	**97.50**	**99.49**
Qwen2.5	81.75	83.44	81.37	**100**	69.12	91.80
JointMRE (Qwen2.5)	82.03	83.98	82.40	**100**	71.90	92.75
CDRF (Qwen2.5)	**96.80**	**98.52**	**97.14**	**100**	**95.10**	**98.98**

Table 11: Cross-lingual performance comparison of different methods on the Vietnamese ViLexNorm dataset.

6 Conclusions

We introduce HealthAMR and GeneralAMR, two high-quality LiveAMR benchmarks covering health and seven E-commerce domains, totaling over 92K instances. To improve generalization, we propose JointMRE, which jointly resolves morphs and generates structured explanations, and CDRF, a lightweight refinement stage that uses those explanations to correct inconsistencies. Experiments show our methods outperform strong PLM and LLM baselines on two datasets. Future work will explore real-time deployment and cross-lingual extensions.

7 Ethics Statement

We strictly adhere to ethical guidelines and legal regulations regarding data collection. The live streams utilized in this study are publicly accessible broadcasts intended for commercial sales. Under Article 17 of the E-Commerce Law of the People Republic of China, such commercial activities are subject to public supervision. Furthermore, in accordance with the Regulations for the Implementation of the Copyright Law, the promotional rhetoric contained therein is not categorized as a protected "work", thereby permitting its use for non-commercial academic research.

To ensure privacy protection, all transcribed texts have undergone rigorous anonymization: we have removed all user IDs, contact information, and personally identifiable information (PII), retaining only product-related discourse essential for lexical analysis.

The collected dataset is released under the CC BY-NC 4.0 license to support reproducibility and further academic research. Any commercial use of this dataset is strictly prohibited to respect the interests of the original content creators. Under this protocol, there is a potential risk that the data we provide could be reverse-engineered and used to develop methods for evading live-streaming detection.

Limitations

Our work has several limitations. First, while our new dataset significantly expands domain coverage, it still primarily focuses on Chinese e-commerce, and the framework's applicability to other languages or contexts (e.g., political discourse) remains to be explored. Second, the Stage-II correction process, which relies on a large language model, introduces additional computational overhead compared to single-stage methods. Finally, our approach currently processes text-only inputs at present.

Acknowledgement

This research is partially supported by the National Language Commission of China (ZDI145-71), the National Natural Science Foundation of China (62403412 and 62076217), the Blue Project of Jiangsu province, and the Top-level Talents Support Program of Yangzhou University.

References

Ron Artstein. 2017. Inter-annotator agreement. *Handbook of linguistic annotation*, pages 297–313.

Timothy Baldwin, Marie-Catherine De Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, and Wei Xu. 2015. Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. In *Proceedings of the workshop on noisy user-generated text*, pages 126–135.

Andrey Bout, Alexander Podolskiy, Sergey Nikolenko, and Irina Piontkovskaya. 2023. Efficient grammatical error correction via multi-task training and optimized training schedule. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5800–5816.

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

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of the 2013 conference of the North American Chapter of the association for computational linguistics: Human language technologies*, pages 359–369.

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*.

Zhifu Gao, Shiliang Zhang, Ian McLoughlin, and Zhijie Yan. 2022. Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Paul Hsiung, Andrew Moore, Daniel Neill, and Jeff Schneider. 2005. Alias detection in link data sets. In *Proceedings of the International Conference on Intelligence Analysis*, volume 4.

Hongzhao Huang, Zhen Wen, Dian Yu, Heng Ji, Yizhou Sun, Jiawei Han, and He Li. 2013. Resolving entity morphs in censored data. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1083–1093.

Ihmily. 2025. Douyinliverecorder: A simple live broadcast recording tool with loopable monitoring. https://github.com/ihmily/DouyinLiveRecorder.

Armand Joulin, Edouard Grave, and Piotr Bojanowski Tomas Mikolov. 2017. Bag of tricks for efficient text classification. *EACL 2017*, page 427.

Yinghui Li, Shang Qin, Jingheng Ye, Haojing Huang, Yangning Li, Shu-Yu Guo, Libo Qin, Xuming Hu, Wenhao Jiang, Hai-Tao Zheng, and Philip S. Yu. 2025. Rethinking the roles of large language models in Chinese grammatical error correction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 553–567, Vienna, Austria. Association for Computational Linguistics.

Ismini Lourentzou, Kabir Manghnani, and ChengXiang Zhai. 2019. Adapting sequence to sequence models for text normalization in social media. In *Proceedings of the international AAAI conference on web and social media*, volume 13, pages 335–345.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Benjamin Muller, Benoit Sagot, and Djamé Seddah. 2019. Enhancing bert for lexical normalization. In *The 5th workshop on noisy user-generated text (W-NUT)*.

Thanh-Nhi Nguyen, Thanh-Phong Le, and Kiet Nguyen. 2024. Vilexnorm: a lexical normalization corpus for vietnamese social media text. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1421–1437.

Patrick Pantel. 2006. Alias detection in malicious environments. In *AAAI Fall Symposium: Capturing and Using Patterns for Evidence Detection*, pages 14–20.

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. Intermediate-task transfer learning with pretrained language models: When and why does it work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5231–5247, Online. Association for Computational Linguistics.

David Samuel and Milan Straka. 2021. \'ufal at multilexnorm 2021: Improving multilingual lexical normalization by fine-tuning byt5. *arXiv preprint arXiv:2110.15248*.

Rob van der Goot. 2019. MoNoise: A multi-lingual and easy-to-use lexical normalization tool. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 201–206, Florence, Italy. Association for Computational Linguistics.

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*, 254:124294.

Canwen Xu, Wangchunshu Zhou, Tao Ge, Ke Xu, Julian McAuley, and Furu Wei. 2021. Blow the dog whistle: A chinese dataset for cant understanding with common sense and world knowledge. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2139–2145.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv e-prints*, pages arXiv–2412.

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.

Boliang Zhang, Hongzhao Huang, Xiaoman Pan, Heng Ji, Kevin Knight, Zhen Wen, Yizhou Sun, Jiawei Han, and Bulent Yener. 2014. Be appropriate and funny: Automatic entity morph encoding. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 706–711.

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*.

Jiahao Zhu, Jipeng Qiang, Ran Bai, Chenyu Liu, and Xiaoye Ouyang. 2025. Chinese morph resolution in E-commerce live streaming scenarios. In *NAACL(Volume 3: Industry Track)*, pages 380–389.

A Appendix

A.1 Data Annotation Website

Figure 4: Screenshot of an annotation example on the annotation Website. The red text indicates added comments.

To facilitate the annotation process, we developed a dedicated web-based platform using Vue and Flask, as depicted in Figure 4. The platform

was designed to present annotators with multimodal information simultaneously, a feature crucial for enhancing the quality and accuracy of the final annotations. We acknowledge that this was a time-consuming task and extend our sincere gratitude to the annotators for their diligent efforts.

Before formal annotation, annotators were provided with written guidelines and examples, and were instructed to identify morphs, annotate their type and normalized form, and revise or reject model suggestions when necessary. Annotators were also informed that the source material might contain exaggerated or sensitive promotional expressions from real-world live-stream e-commerce broadcasts.

A.2 Experiment Detail

In this section, we detail the specific configurations for fine-tuning our models and the prompts used.

Fine-tuning PLM		**Training LLM**		
Parameter	**T5**	**Parameter**	**Qwen**	**Llama**
Finetuning type	full	Finetuning type	lora	lora
Batch size	16	Batch size	1	1
Learning rate	1e-4	Learning rate	1e–4	1e-4
Epochs	10	Epochs	2	2
Lr scheduler	linear	Lr scheduler	cosine	cosine
Max length	512	Max length	1024	1024
Optimization	Adam	Optimization	Adam	Adam
Mixed precision	fp16	Mixed precision	bf16	bf16
-	-	Lora rank	8	8

Table 12: Details of training configuration. The left block shows parameter values used in PLM contrastive learning, while the right block lists training parameter names used in LLM-based classification.

Our fine-tuning configurations, detailed in Table 12, were tailored for two distinct model classes. For the Pre-trained Language Models such as T5, we performed full model fine-tuning. In contrast, for the much larger LLMs (Qwen and Llama), we employed Parameter Efficient Fine-Tuning using the LoRA methodology to ensure training efficiency. We fine-tuned the Large Language Models (LLMs) on two tasks using distinct instructional prompts. For the Morph Resolution task, the model was given the instruction: *Please resolve the morphs in the following sentence. Important: Only correct the morphs and keep all other content, including any potential ASR errors, unchanged.* For the Morph Explanation task, the corresponding instruction was: *Please provide explanations for the morphs in the following sentence. Important: Only explain the identified morphs; do not explain or modify any other content, including potential ASR errors.*

A.3 Case Study

Example1:EXIST CONFLICT
X: 这款软糖可以改和善我们心灵之窗。
(This gummy can change-and-kind our windows of soul.)
Y: 这款软糖可以改和善我们眼睛。
(This gummy can change-and-kind our eyes.)
JointMRE-Y_{SI}: 这款软糖可以改和善我们眼睛。
(This gummy can change-and-kind our eyes.)
Y_{EI}: “改和善”是结构变体词，原词是改善
“心灵之窗”是语义变体词，原词是眼睛
("change and kind" is a Transformation morph of the word "improve";)
("windows of the soul" is a Synonym morph of the word "eyes".)
CDRF-Y_{SII}: 这款软糖可以改善我们眼睛。
(This gummy can improve our eyes.)

Figure 5: One example for CDRF illustrating the conflict-aware correction mechanism. It shows a case with a conflict: JointMRE's prediction (Y_{SI}) is inconsistent with its explanation (E_{SI}), prompting CDRF to generate a refined output (Y_{SII}).

Figure 5 illustrates how CDRF resolves multi-faceted conflicts. The framework is presented with two challenges: a lexical error in the source text, requiring a choice between 该和善 (should be kind) and 改善 (improve), and a factually flawed explanation that incorrectly restores 心灵之窗 (window of the soul) to 眼睛 (eyes). CDRF successfully disentangles these issues, rectifying the source text to 改善 (improve) while using its analysis to discard the flawed explanation and adopt the correct restoration, 眼睛 (eyes). In the 'no conflict' case, the framework simply verifies that the initial prediction is consistent with the explanation and preserves it without intervention, thus avoiding the introduction of new errors.

A.4 Experimental Prompts

This section presents the three complete prompts used in our experiments: one for evaluating baseline LLMs in a fewshot setting, one for extracting gold explanations, and one for guiding the reasoning process of our CDRF framework.

Example2:NO CONFLICT
X: 咱们家这款鱼油尽早吃。等到水龙头被堵住了，再去某医某院花钱去做手术就来不及了。
(Take this fish oil as soon as possible. It will be too late to go to mou-hos-mou-pital and spend money on surgery when the tap is blocked.) Y: 咱们家这款鱼油尽早吃。等到血管被堵住了，再去医院花钱去做手术就来不及了。
(Take this fish oil as soon as possible. It will be too late to go to hospital and spend money on surgery when the blood vessels are blocked.)
JointMRE-Y_{SI}: 咱们家这款鱼油尽早吃。等到血管被堵住了，再去医院花钱去做手术就来不及了。
(Take this fish oil as soon as possible. It will be too late to go to hospital and spend money on surgery when the blood vessels are blocked.)
Y_{EI}: “水龙头”是语义变体词，原词是血管
“某医某院”是结构变体词，原词是医院
("tapl" is a Synonym morph of the word "blood vessels".)
(" mou-hos-mou-pital" is a Transformation morph of the word " mouhospital";)

CDRF-Y_{SII}=Y_{SI}

Figure 6: One example for CDRF illustrating the conflict-aware correction mechanism. It shows a case with no conflict, where the prediction and explanation are consistent, so CDRF confirms the initial prediction ($Y_{\text{SII}} = Y_{\text{SI}}$).

LLM Morph Resolution Prompt

Morph resolution is the task of restoring all morphs in a sentence to their original forms, while leaving non-morph words unchanged. There are three main types of morphs:
1.Transformation Morphs: These are typically formed by inserting meaningless filler words into standard words. Examples: improvechange and kindproblemsmall-prob-small-lem,
2.Synonym Morphs: These use semantically related expressions instead of the original word.
Examples: eyswindow of the soul-wine8+1doctorpeople in white
3.Homophone: These substitute standard characters with phonetically similar ones, relying on auditory resemblance to represent the original word.
Task Instruction:
You need to perform morph resolution on the user's input sentence. If no morphs exist, output the sentence unchanged.The final output must be enclosed in <target></target> tags.
Input: [user_input]
Input: [fewshot][optional]
Output:[prediction]

CDRF Reasoning Prompt

You are an expert in Chinese Morph Resolution, acting as a meticulous verifier and corrector. You will be given a source sentence, an initial prediction called Prediction 1, and its corresponding explanation.
Your task is to identify and resolve all inconsistencies between the prediction and the explanation by following three precise steps.
Step 1: Conflict Identification First, based only on the explanation, reconstruct a corrected sentence and call it Prediction 2. Then, compare Prediction 1 with Prediction 2 to find all discrepancies. For each discrepancy, create a conflict pair of two triplets. The first triplet, Triplet A, is from Prediction 1, and the second, Triplet B, is from Prediction 2. A triplet is formatted as morph, type, resolved_word. If no conflicts are found, the final answer is simply Prediction 1.
Step 2: Conflict Arbitration For each conflict pair you identified, you must arbitrate which triplet is more plausible. Analyze your decision based on these principles: - Semantic Consistency: Which resolved word better fits the context of the source sentence? - Type Consistency: Which transformation is more consistent with the given morph type? - Overall Logic: Which triplet represents a more logical and necessary correction?
You must briefly state your reasoning for each arbitration decision.
Step 3: Targeted Correction and Final Output Based on your arbitration decisions, apply the winning corrections to Prediction 1 to create the final, corrected sentence. If you chose Triplet B for a conflict, apply its correction. If you choose Triplet A, no change is needed for that specific part.
Your final corrected sentence must be the only content enclosed in <final_answer></final_answer> tags. Show your reasoning steps before the final answer.
Input: [prediction from JointMRE]
Input: [explanation from JointMRE]
Output:[final prediction]

LLM Generate Explanation Prompt

You are an expert in Chinese Morph Resolution. Your primary task is to identify morphs in a given sentence, provide a structured explanation for them, and output the corrected sentence.
1. Morph Definitions and Types:
Morph resolution is the task of restoring all morphs in a sentence to their original forms, while leaving non-morph words unchanged. There are three main types of morphs Examples: ...
2. Task Instructions: For the given user input "source" sentence, you must perform the following actions and generate three distinct outputs, each enclosed in specific XML tags. A. Identify Morph Type(s): Determine the type(s) of morphs present in the sentence. The possible types are: ‘transformation‘, ‘synonym‘‘homophone‘ If multiple morphs exist, list their types in the order of their appearance. Enclose the final type string in <type></type> tags. B. Generate Explanation: For each identified morph, create a structured explanation. The format must be: ‘"<morph>" is a <morph_type>, the original word is "<original_word>"‘. If multiple explanations exist, separate them with a semicolon. Enclose the final explanation string in <explanation></explanation> tags. C. Generate Corrected Sentence (Target): Restore all identified morphs to their original words to create the corrected sentence. Enclose the final corrected sentence in <target></target> tags. 3. Handling Sentences with No Morphs:** If the source sentence contains no morphs, the explanation output must be: <explanation>no morph</explanation>, and the target output must be the original sentence enclosed in <target></target> tags.
Input: [source sentence]
Input: [target sentence]
Output:[explanation]