COAS2W: A Chinese Older-Adults Spoken-to-Written Transformation Corpus with Context Awareness

Anonymous ACL submission

Abstract

Spoken language from older adults often deviates from written norms due to omission, disordered syntax, constituent errors, and redundancy, limiting the usefulness of automatic transcripts in downstream tasks. We present COAS2W, a Chinese spoken-to-written corpus of 10,004 utterances from older adults, each paired with a written version, fine-grained error labels, and four-sentence context. Unlike existing resources, COAS2W captures cross-sentence dependencies crucial for resolving ambiguities and recovering missing content. Fine-tuned lightweight open-source models on COAS2W outperform larger closed-source models. Context ablation shows the value of multi-sentence input, and normalization improves performance on downstream translation tasks. COAS2W supports the development of inclusive, context-aware language technologies for older speakers.

1 Introduction

With the rapid advancement of digital technologies, voice-based interaction has become an increasingly important modality for older adults to access and operate electronic devices more intuitively (Pradhan et al., 2020). While automatic speech recognition (ASR) systems can transcribe spoken input into text with reasonable accuracy(Radford et al., 2023), the resulting transcripts—particularly those from older adult' speakers—often reflect informal, fragmented, and structurally divergent language(Liu et al., 2023). However, most downstream systems are trained on well-formed written corpora and expect inputs that conform to standard written conventions(Sun et al., 2021). This mismatch between the linguistic style of older adults' speech and the expectations of existing digital systems limits the effectiveness of voice interaction(Michel and Neubig, 2018), even in the absence of ASR errors. To bridge this gap, it is essential to develop corpora and models that can transform naturally spoken older adults' language into coherent written text, thereby enhancing system compatibility and promoting inclusive human-computer interaction.

Existing work on spoken-to-written transformation can be grouped into two lines of research: document-level modeling, which summarizes an entire dialogue or transcript into a concise written form (Pan et al., 2018; Chen et al., 2021a), and sentence-level modeling, which produces a one-to-one written rendition for each spoken utterance (Guo et al., 2023). Document-level methods often compress information and therefore fail to meet the requirements of downstream tasks that demand complete semantic preservation—e.g., machine translation or voice-command execution. Sentence-level approaches are a closer fit, yet they have largely been developed for general, well-structured speech and do not address the linguistic idiosyncrasies of older adult speakers.

Through an empirical analysis of Chinese older adults' speech (see Table 1), we propose a categorization of four error types that is both exhaustive and mutually exclusive: **(i) Constituent Omission, (ii) Disordered Syntax, (iii) Constituent Errors, and (iv) Constituent Redundancy**. Correcting such errors often requires information beyond the sentence boundary—for example, resolving a missing subject typically depends on cues from surrounding sentences. Sentence-level models that process utterances in isolation are therefore ill-suited for normalizing older adults' speech.

Category	Example	Context
Constituent Omission e.g., subject omission	到了月底的时候，跑到我母亲这儿。 At the end of the month, went to my mother. 到了月底的时候，我家邻居跑到我母亲这儿。 At the end of the month, our neighbor came to my mother.	我们在南市住，有个邻居啊，他生活也够困难的。到了月底的时候，跑到我母亲这儿。 We lived in Nanshi, and there was a neighbor his life was pretty hard too. At the end of the month, came to my mother.
Disordered Syntax e.g., improper clause order	南市住的话，她要是回一趟咸水沽的话，每次回去一趟起码得将近半天差不多。 If living in Nanshi, if she wanted to go back to Xianshuigu, each time going back would take almost half a day, more or less. 从南市到咸水沽，她每次往返需要半天。 From Nanshi to Xianshuigu, each round trip took her half a day.	我们家在南市住，南市平房，现在食品街那儿。南市住的话，她要是回一趟咸水沽的话，每次回去一趟起码得将近半天差不多。 Our family lived in Nanshi. Nanshi had single-storey houses, now it' s where the Food Street is. Living in Nanshi, if she wanted to go back to Xianshuigu, each time going back would take almost half a day, more or less.
Constituent Errors e.g., incorrect use of personal pronouns	我弟妹说让我们把家腾干净了，我们要来住。 My sister-in-law said we should clear out the house, we're coming to live here. 我弟妹说让我们把家腾干净了，他们要来住。 My sister-in-law said we should clear out the house —they're coming to live here.	我弟跟我妈说他们下个月就来。我弟妹说让我们把家腾干净了，我们要来住。 My younger brother told my mom they're coming next month. My sister-in-law said we should clear out the house, we're coming to live here.
Constituent Redundancy e.g., self-repair(speaker restates or corrects)	我们哥五个哥六个觉得一定得这么办。 We five or six brothers felt that this was definitely the way to go. 我们哥六个觉得一定得这么办。 We six brothers felt that this was definitely the way to go.	我们哥五个哥六个觉得一定得这么办。我们认为既然是老太太的房子，那么就应该作为老太太的遗产，咱们哥六个应该共同继承。 The five of us, or six of us, all felt it had to be done this way. We thought since it was the old lady's house, then it should be treated as her inheritance, and the six of us brothers should inherit it together.

Table 1: Four typical categories of linguistic errors in older adults' spoken Chinese. For each category, the first pair of sentences in the Example column presents the original spoken utterance and its English translation. The second pair provides the corrected written form and its corresponding translation. The Context column includes the surrounding spoken context, which is used as a reference for error correction.

To address these challenges, we introduce a context-aware modeling approach that incorporates surrounding-sentence context into the transformation process. Central to this effort is COAS2W, a corpus of 10,004 utterances from Chinese older adults, each paired with context, fine-grained error annotations, and fully normalized written counterparts. By providing explicit context and detailed supervision for all four error types, COAS2W enables models to better preserve meaning and conform to written conventions.

To validate the effectiveness of COAS2W, we conduct three sets of experiments. First, fine-tuning open-source models on COAS2W substantially improves their performance in spoken-to-written transformation, outperforming prior work (CS2W (Guo et al., 2023)) and even surpassing closed-source models such as GPT-4o and Claude-3.7-Sonnet, while remaining more resource-efficient. Second, ablation studies show that compared to full document context, a four-sentence window—with two preceding sentences in normalized form and two following sentences unformatted—offers a more effective and efficient context modeling strategy. Third, we demonstrate that normalization significantly improves downstream translation quality, underscoring the value of spoken-to-written transformation for cross-lingual tasks.

Our contributions are as follows:

1. We conduct an empirical analysis of Chinese older adults' spoken language and propose a categorization of deviations into four error types that are exhaustive and mutually exclusive.

2. We release COAS2W, the context-annotated, sentence-aligned corpus of older adults' spoken-to-written pairs, together with error labels.

3. We demonstrate that context-aware sentence-level modeling, enabled by COAS2W, empowers lightweight models to achieve state-of-the-art performance in spoken-to-written transformation and enhances downstream tasks such as translation, thereby laying the groundwork for inclusive voice-based technologies.

2 Related Work

Linguistic Challenges in Older Adults' Speech. Older adults' spoken language poses unique challenges for NLP, such as syntactic omissions, redundant self-repairs, disordered structure, and topic shifts (Wang et al., 2023; Iida and Wakita, 2021; Barnett and Coldiron, 2021). These deviations stem from both cognitive aging and habitual colloquial use, and persist even in carefully transcribed utterances (Luo et al., 2020; Burke et al., 2024).

Existing corpora such as CCC (Pope and Davis, 2011), DementiaBank (Lanzi et al., 2023), SeniorTalk (Chen et al., 2025a), and MCGD (Huang and Zhou, 2025) provide valuable speech resources but mainly offer raw transcripts without sentence-aligned rewrites or annotations of syntactic irregularities. This limits their utility in training models for coherent normalization—critical for translation, summarization, or voice command processing.

Spoken-to-Written Normalization. Spoken language often diverges from written norms due to disfluencies, informal phrasing, and incomplete syntax, reducing its effectiveness in downstream tasks (Saini et al.; Wang et al., 2014; Asrifan, 2021). Prior work typically treats normalization as sentence-level rewriting to improve fluency and grammaticality.

For example, CS2W (Guo et al., 2023) constructed a corpus of ASR outputs and formal rewrites for correcting filler words and colloquialisms. DialogSum (Chen et al., 2021b) paired informal dialogue with concise summaries. However, these assume structurally complete inputs and lack mechanisms to address the deeper disruptions common in elderly speech (Liu et al., 2023).

Context-Aware Modeling. Context is essential for rewriting fragmented or ambiguous speech. Prior work shows that multi-sentence input improves ASR post-editing, entity resolution, and discourse coherence (Zhou et al., 2022; Peng et al., 2024). Yet most focus on short, well-structured utterances and overlook complex structural rewrites.

Our work complements these efforts by enabling sentence-level normalization with contextual input and error annotations, addressing omissions, reordering, and reference ambiguities specific to older adults' spoken language.

3 Dataset Construction

This section outlines the construction of COAS2W, which transforms spoken Chinese utterances from older adults into fluent written sentences and labels them with linguistic error types. Figure 1 presents an overview of the annotation workflow.

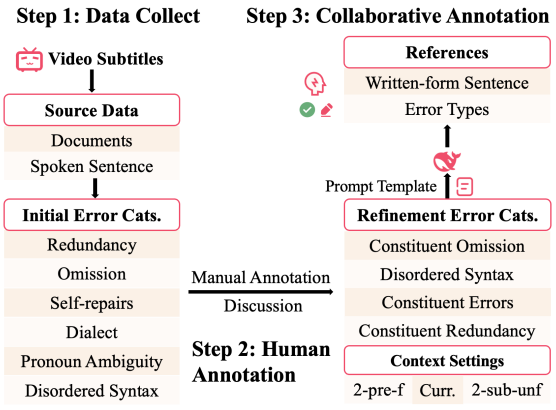

Figure 1: Overview of the annotation workflow for the COAS2W dataset.

3.1 Data Sources

Most publicly available Chinese older adult speech datasets contain only audio and lack aligned transcriptions (Chen et al., 2025b). Using ASR to generate transcripts introduces noise such as homophone errors (e.g., "虹桥"

misrecognized as "红桥") (Fan et al., 2023), which fall outside the scope of our target linguistic phenomena. To avoid this, we manually collected and proofread subtitles from social media videos.

We selected Bilibili[1], a major long-form video platform in China, for its abundance of naturally occurring, unscripted older adult speech. We identified 23 vloggers focused on later-life content and used the `you-get`[2] tool to download 282 relevant videos, filtered by titles containing terms like "老人" (older adults) or "岁" (age). We then extracted hardcoded subtitles from these videos using OCR via the Video-Subtitle-Extractor (VSE)[3], yielding 282 document-level transcripts.

As downstream tasks like translation and voice-command execution operate at the sentence level, we treat sentences as our basic modeling unit. We applied automatic segmentation (Appendix A.1), resulting in **10,004** spoken sentences. Dataset statistics are provided in Section 4. All videos were either publicly licensed for research or approved via direct consent from uploaders. Content was manually screened to ensure no sensitive or personally identifiable information (PII) was included.

3.2 Annotation Rule Induction

To balance annotation accuracy and cost, we adopted a two-stage collaborative framework integrating human expertise and LLM assistance. In Stage 1, two NLP-trained PhD students conducted manual labeling on a data subset to develop the initial guidelines. In Stage 2, these guidelines were used to prompt LLMs for large-scale annotation.

Preliminary Error Analysis and Guideline Drafting. A preliminary linguistic analysis, informed by prior studies (Yan et al., 2024; Wang and Wang, 2024; Hu et al., 2021; Liu et al., 2021), identified six common deviation types in elderly speech, including fillers, omissions, dialectal expressions, ambiguous pronouns, disordered syntax, and self-repairs. Annotators followed a two-step protocol: label error types and produce normalized rewrites. Full definitions and examples are in Appendix A.2.

Subsequent pilot annotation of 300 utterances revealed category overlap, leading to a refined taxonomy of four mutually exclusive syntactic categories: (1) **Constituent Omission**, (2) **Disordered Syntax**, (3) **Constituent Errors**, and (4) **Constituent Redundancy**. A formal proof of the completeness and independence of this taxonomy is provided in Appendix A.3. Context was found crucial, especially for resolving omissions and ambiguous references often dependent on preceding sentences (see Table 1).

Context Design and Evaluation. We evaluated how context configurations affect annotation quality, varying (1) *context length* (none, 4-sentence window, full document) and (2) *context type* (raw vs. normalized). The 4-sentence window was based on working memory research (Cowan, 2001) and includes 2 preceding and 2 following utterances.

Five master's students rewrote 100 utterances under five configurations. In the *partially formatted* 4-sentence setting, the two preceding utterances were rewritten manually, simulating incremental processing where past content is normalized and future content is not. Two PhD annotators rated outputs for semantic completeness and readability.

The partially formatted 4-sentence window yielded the best performance(see Appendix A.4) and was adopted as the default context setting for both annotation and modeling.

3.3 LLM-Assisted Collaborative Annotation

With the annotation schema finalized, we employed DeepSeek-V3[4], a high-performance open-source language model known for its strong performance on Chinese NLP tasks and significantly lower cost compared to commercial alternatives[5].

Based on the finalized guidelines (Section 3.2), we constructed structured prompts

[1] https://www.bilibili.com

[2] https://github.com/soimort/you-get

[3] https://github.com/YaoFANGUK/video-subtitle-extractor

[4] https://github.com/deepseek-ai/DeepSeek-V3

[5] As of May 2025, processing 1M input tokens costs approximately $5.00 with GPT-4o (https://openai.com/api/pricing/) and only $0.27 with DeepSeek-V3 (https://api-docs.deepseek.com/quick_start/pricing).

that included the spoken utterance along with its four-sentence context window. For each input, the model was asked to generate (1) the corresponding error types and (2) a revised written version. The full prompt template is provided in Appendix A.5.

To ensure annotation quality and consistency, we conducted a manual verification phase following model output generation. Five students with NLP training—both graduate and undergraduate—were recruited to review and revise the LLM-generated annotations. They corrected incorrect error labels and refined unnatural, incomplete, or ambiguous written rewrites. All annotators followed a shared annotation protocol, and difficult cases were resolved through group discussion.

Through collaborative annotation process, we obtained a total of 10,004 high-quality annotated instances. A sample instance is provided in Appendix A.6. The resulting dataset will be publicly available following the peer review process.

4 Dataset Analysis

We provide document-level statistics, dataset partition details (training/test splits), and a comprehensive analysis of error type distributions.

4.1 Document-Level Statistics

We manually analyzed each document (i.e., a video interview from an older adult speaker) and summarized key properties as shown in Table 2. Topic definitions are provided in Appendix B. These topic categories indicate that our dataset reflects common everyday themes among older adults, differing from younger-oriented corpora in both content and structure.

Property	Value
#Documents (Videos)	282
Avg. #Sentences per Document	35.5
Avg. Duration per Video (s)	781.4
# Documents per Topic	
Life Experience	222
Family Relations	219
Life in Old Age	102
Social Values	143

Table 2: Document-level statistics of COAS2W.

4.2 Dataset Partitioning and Statistics

We randomly split the 10,004 annotated sentence pairs into training and test sets at an 8:2 ratio. Table 3 presents detailed statistics for each split, including the number of sentences, error type distributions, and the average sentence length, with the observation that a single sentence may contain multiple error types.

Statistic	Train	Test	Total
#Sentences	8003	2001	10,004
Constituent Omission	5780	1445	7225
Disordered Syntax	6077	1505	7582
Constituent Errors	2513	601	3114
Constituent Redundancy	3714	902	4616
#Characters	280842	70028	350870
Avg. #characters	35.09	35.00	35.07

Table 3: Sentence-level statistics of COAS2W.

4.3 Multiple Error Type Analysis

To better understand the complexity of spoken sentences in our dataset, we analyze the distribution of error types per sentence. As shown in Figure 2, only a small fraction (1.33%) of sentences are error-free, while nearly half (44.88%) contain three distinct error types, highlighting a gap between older adults' spoken language and its well-formed written counterpart.

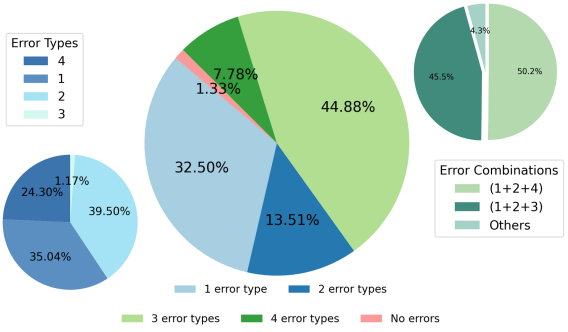

Figure 2: Sentence-level distribution of linguistic error types. The central chart shows the proportion of sentences containing 0–4 error types. The bottom-left chart details the distribution of error types in single-error sentences, while the top-right chart illustrates the most common error combinations among three-error sentences. Error types: 1 = Constituent Omission, 2 = Disordered Syntax, 3 = Constituent Errors, 4 = Constituent Redundancy.

5 Experiments

To assess the effectiveness of COAS2W in enhancing LLMs' ability to process and normalize older adults' spoken Chinese, we design experiments along three axes: **i) Dataset Impact:** We fine-tune four widely used open-source, small-parameter, large language models on the COAS2W dataset and evaluate their improvements in transforming elderly spoken utterances into written form. Performance is compared against existing approaches and closed-source models such as GPT and Claude. **ii) Context Modeling Strategy Effectiveness:** We conduct an ablation study to assess the impact of our proposed context modeling strategy, which incorporates a five-sentence window (two preceding, current, and two following sentences), with the first two sentences presented in normalized form to simulate real-time incremental processing. **iii) Downstream Task Performance:** We examine whether converting spoken text into its written equivalent leads to performance gains in downstream tasks, with a focus on Chinese-to-English translation.

5.1 Dataset

We randomly split the 10,004 annotated instances into training and test sets using an 8:2 ratio, as described in Section 4.2. All experiments were conducted on the test set.

5.2 Model and Baselines

Open-source models. We selected four commonly used open-source large language models with relatively small parameter sizes ($\leq$ 7B): Qwen2.5-7B-Instruct[6], Mistral-7B-Instruct v0.2[7], ChatGLM3-6B[8], and Baichuan2-7B-Chat[9](Hereafter referred to as Qwen, Mistral, ChatGLM, and Baichuan, respectively.) These models were fine-tuned on the COAS2W dataset. Details of the fine-tuning settings are provided in Appendix C.2.

Closed-source models. We evaluate two representative closed-source large language models: GPT-4o[10] and Claude-3.7- Sonnet[11] (hereafter referred to as GPT and Claude, respectively). Their performance is assessed under both 0-shot and 5-shot settings.

Baselines. CS2W (Guo et al., 2023) primarily introduces a dataset for Chinese spoken-to-written transformation. Although no code is released, the paper reports that the best-performing model was CPT-large fine-tuned on their dataset. We reimplemented this setup and adopted the resulting model as a baseline in our experiments.

5.3 Metrics

We evaluate model performance from two perspectives: (1) error type detection accuracy, and (2) spoken-to-written conversion quality. For error detection, we report Joint Accuracy (all gold labels correctly predicted) and Acc-1 (at least one correct label). For generation quality, we use BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), and BLEURT (Sellam et al., 2020) to assess semantic fidelity. Metric definitions and settings are detailed in Appendix C.1.

5.4 Main Results

From the overall results presented in Table 4, we summarize our findings as follows.

COAS2W significantly improves the performance of open-source models through fine-tuning. The results demonstrate consistent improvements across all evaluation metrics after fine-tuning. On average, fine-tuned models exhibit a +0.29 gain in Joint Accuracy and a +0.30 gain in Acc-1. In terms of generation quality, we observe consistent gains in BLEU-1 to BLEU-4 scores (average improvements ranging from +0.13 to +0.19), as well as in ROUGE-L (+0.15) and BLEURT (+0.14), reflecting better semantic alignment with the gold-standard written text (calculation methods are detailed in Appendix C.3). Among the evaluated models, the fine-tuned Mistral achieves the best overall performance, consistently outperforming the others across nearly all metrics, making it particularly well-suited for this task.

[6] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

[7] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

[8] https://huggingface.co/THUDM/chatglm3-6b

[9] https://huggingface.co/baichuan-inc/Baichuan2-7B-Chat

[10] https://platform.openai.com/docs/models/gpt-4o

[11] https://www.anthropic.com/claude/sonnet

Type	Model	Setting	JA	Acc-1	B-1	B-2	B-3	B-4	R-L	BL
Open-source	Qwen	w/ FT	0.3951	0.8893	0.6987	0.4933	0.3638	0.2786	0.5944	0.4142
	ChatGLM		0.2194	**0.9750**	0.6884	0.4519	0.3103	0.2237	0.5414	0.3515
	Mistral		**0.4997**	0.9418	**0.7572**	**0.5709**	**0.4541**	**0.3759**	**0.6481**	**0.4604**
	Baichuan		0.3943	0.9305	0.7423	0.5462	0.4195	0.3345	0.6316	0.4467
	Qwen	w/o FT	0.1299	0.8056	0.6423	0.3974	0.2573	0.1727	0.5094	0.3244
	ChatGLM		0.0422	0.7770	0.5738	0.3070	0.1735	0.1029	0.4312	0.2594
	Mistral		0.0833	0.5785	0.5842	0.3190	0.1837	0.1114	0.4307	0.2416
	Baichuan		0.0765	0.3620	0.5779	0.3138	0.1784	0.1052	0.4375	0.2741
Close-source	GPT	5-shot	0.1084	0.8401	0.7221	0.5170	0.3851	0.2956	0.6091	0.4063
	Claude		0.1713	0.8116	0.6988	0.4816	0.3455	0.2577	0.5846	0.4052
	GPT	0-shot	0.1744	0.7271	0.7029	0.4992	0.3694	0.2807	0.6029	0.4034
	Claude		0.1960	0.7960	0.6790	0.4706	0.3403	0.2551	0.5852	0.4001
Baseline	CS2W	-	N	N	0.6342	0.3483	0.2003	0.1201	0.4599	0.2834

Table 4: Performance of Different Models on the Speech Error Recognition and Correction Task. JA = Joint Accuracy; Acc-1 = At-least-one Accuracy; B-1 to B-4 = BLEU scores with 1–4 grams; R-L = ROUGE-L; BL = BLEURT (Bilingual Evaluation Understudy with Representations from Transformers). w/ FT = with fine-tuning; w/o FT = without fine-tuning.

Compared to closed-source models, fine-tuned open-source models offer competitive performance with better resource efficiency. Closed-source models (GPT and Claude) perform better under the 5-shot setting than 0-shot, but still underperform compared to fine-tuned open-source models. Given their larger sizes and higher inference costs, Mistral fine-tuned on COAS2W remains the most practical option.

Compared to previous work, our approach achieves significantly better results across all metrics. We implemented the best-performing model described in CS2W and evaluated it on our test set. Across all evaluation metrics, it underperforms compared to any of our fine-tuned models. This suggests that prior spoken-to-written systems failed to adequately capture linguistic phenomena specific to elderly speech, such as disorganized syntax and missing constituents.

In summary, COAS2W improves model performance on older adults' spoken language transformation, while serving as a feasible solution in terms of cost and efficiency.

5.5 Ablation Experiments

We randomly sampled 1,000 instances from the 2,000-item test set to evaluate GPT's performance under four context settings: (i) no context, (ii) unformatted context (±2 sentences), (iii) partially formatted context (2 formatted preceding + 2 unformatted following), and (iv) fully document context. This experiment assesses the effectiveness of our context design, with results presented in Figure 3. Our key findings are as follows:

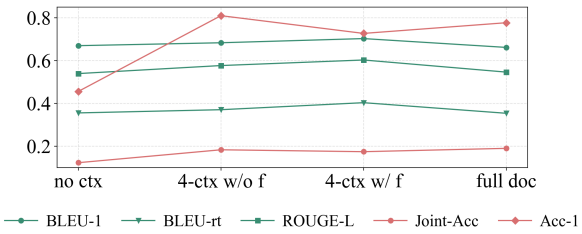

Figure 3: Performance of GPT under different context settings. The horizontal axis represents different input settings: no ctx = no context; 4-ctx w/o f = unformatted context (±2 sentences); 4-ctx w/ f = partially formatted context (2 formatted preceding + 2 unformatted following); full doc = full-document context. The vertical axis indicates the values of different evaluation metrics.

Incorporating context enhances performance. GPT equipped with contextual information consistently outperforms single-sentence baselines across all evaluation metrics. This highlights the necessity of context for accurately interpreting and transforming spoken utterances.

4-sentence context is both effective and efficient. The 4-sentence context achieves comparable or even superior performance to full-document context, while significantly reducing token consumption. In con-

trast, full-document inputs introduce irrelevant or noisy information (e.g., topic shifts or digressions), which can degrade model performance. For example, in the following case:

> Spoken Utterance: 还有这老师傅吗？坐轮椅的啊，那俩也是，都是，都是航天人人才。(Is that senior master still here? The one in the wheelchair? Those two as well, they' re all aerospace talents.)
>
> Reference: 这位坐轮椅的老师傅和那两位都是航天人才。(This senior master in the wheelchair and the other two are all aerospace talents.)
>
> GPT (full-document context): 还有那位坐轮椅的老师傅，他们也都是航天人才。(And that senior master in the wheelchair, they are all aerospace talents.)
>
> GPT (4-sentence context with formatted preceding): 还有这位坐轮椅的老师傅，以及那两位，他们都是航天领域的人才。(There is also this senior master in the wheelchair, and those two as well—they are all talents in the aerospace field.)

Here, the full-document model omits the explicit mention of "那两位 (those two)" and instead merges all referents into a generic group "他们 (they)," resulting in a less faithful rendering of the original utterance.

5.6 Downstream Transfer Experiments

In real-world scenarios such as international travel or cross-lingual medical consultations, older adults often require accurate English translations of their speech. To assess whether converting speech to written form improves translation, we conducted a downstream experiment using 100 COAS2W test samples. Human-annotated written sentences and their English translations served as references. Six input types—including original spoken text, CS2W output, GPT (5-shot), Claude (5-shot), and fine-tuned outputs from Baichuan and Mistral—were translated via the iFLYTEK API[12] and evaluated with BLEU-1/2/4 scores (Figure 4).

Converting spoken language to written form significantly enhances translation performance. The results indicate that translating normalized text yields substantially higher BLEU scores than directly translating spoken input. For example, Mistral with fine-tuning achieves relative improvements of 35.1%, 71.5%, and 146.3% on BLEU-1, BLEU-2, and BLEU-4, respectively, compared to the spoken input, demonstrating that normalization enables translation models to better capture semantic content.

[12] https://www.xfyun.cn/doc/nlp/xftrans/API.html

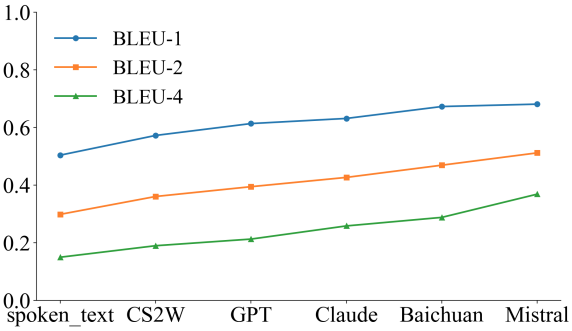

Figure 4: BLEU scores for English translations under different input normalization settings.

Higher-quality transformation leads to better downstream translation. Fine-tuned Mistral produces the best translation results among all models, outperforming even closed-source systems. This aligns with its superior performance in the normalization task (Table 4) and highlights the practical value of high-quality upstream processing for cross-lingual applications.

6 Conclusion

In this paper, we introduce COAS2W, a large-scale, context-rich corpus for transforming Chinese spoken language from older adults into written form. By analyzing linguistic deviations in the spoken language of older adults and annotating 10,004 utterances with corresponding written rewrites and error labels, we provide the first resource tailored to the structural irregularities commonly observed in this demographic. Experimental results show that lightweight models fine-tuned on COAS2W achieve competitive or superior performance compared to closed-source models, and that incorporating 4-sentence context significantly improves normalization quality. Moreover, spoken-to-written transformation enhances performance on downstream translation tasks. Our work lays a practical foundation for age-aware language technologies and underscores the importance of context-aware modeling for real-world spoken language processing.

Limitations

While our approach demonstrates improvements in spoken-to-written transformation for speech from older adults, several limitations remain. First, our evaluation primarily focuses on sentence-level accuracy metrics such as BLEU. It does not fully capture the coherence and readability of the output in long-form or conversational contexts. Future work could incorporate human evaluation and discourse-level quality assessment.

Second, although LLMs show promise in text normalization, they still fall short of human-level performance, especially in cases involving structural reordering or contextual inference. LLMs struggle to resolve long-range dependencies and to reconstruct omitted or disordered sentence elements, which are common in the speech of older adults. Additional methods may be needed to handle these structural phenomena more effectively.

Third, our evaluation does not include larger open-source models such as Qwen-14B or DeepSeek-67 B. This is partly due to our emphasis on efficient modeling and practical deployment for aging-friendly applications. Future work can explore whether even greater gains can be achieved with large-scale models. However, our current results already demonstrate that high-quality contextual data like COAS2W can enable strong performance without the need for excessive model scaling.

Finally, while we demonstrate improvements in downstream translation quality, further exploration is required to assess how spoken-to-written normalization impacts higher-level tasks such as narrative generation, summarization, or command understanding. We leave the extension to story-level or task-specific rewriting as future work.

Potential Risks Although the dataset is constructed from publicly available subtitle content with explicit author consent, there remains a potential risk of unintended model biases. Since the speech style reflects a specific demographic (Chinese older adults featured in online videos), models trained on COAS2W may internalize structural or pragmatic patterns that are not representative of the broader population. Additionally, future misuse could arise if normalization models are applied to marginalized speakers or speech data from other sociolinguistic groups without consideration of context and appropriateness. We encourage responsible use and careful downstream deployment.

Ethics Statement

All source videos were publicly available on the Bilibili platform, and we only included content where the video creators (uploaders) explicitly stated that their videos could be reused for research or non-commercial purposes. In cases where such statements were not found, we contacted the video creators via private messages on Bilibili and obtained their written consent before using their content.

All personally identifiable information (e.g., real names, contact details, geographic locations) was anonymized during preprocessing. While some utterances include potentially identifying content such as surnames or family structure (e.g., “my surname is Su” or “I am the second of six siblings”), these references do not enable identification of any individual speaker, especially as the dataset release excludes all audio, visual content, and uploader metadata. We manually screened all data to ensure no offensive or discriminatory content was included. The study did not involve direct human subject interaction and therefore did not require IRB approval.

Annotation was conducted by two PhD students and five graduate students in linguistics and NLP, who participated voluntarily and were not financially compensated. While we anonymized speaker identities, the dataset may reflect linguistic biases from Bilibili’s user demographics (predominantly urban Mandarin speakers). Future work should include rural and dialectal speech. Additionally, we used GPT-4o to assist with prompt formulation and phrasing refinement during the annotation workflow, and acknowledge its contribution accordingly.

All data used in this study are freely available to the public.

References

Andi Asrifan. 2021. The differences between written and spoken language.

Michael D. Barnett and A. Coldiron. 2021. Off-topic verbosity: Relationships between verbal abilities and speech characteristics among young and older adults. *Applied Neuropsychology: Adult*, 29:1362 – 1368.

Erin Burke, Karlee Patrick, Phillip Hamrick, and John Gunstad. 2024. Effects of normal cognitive aging on spoken word frequency: Older adults exhibit higher function word frequency and lower content word frequency than young adults. *The Open Psychology Journal*.

Yang Chen, Hui Wang, Shiyao Wang, Junyang Chen, Jiabei He, Jiaming Zhou, Xi Yang, Yequan Wang, Yonghua Lin, and Yong Qin. 2025a. Seniortalk: A chinese conversation dataset with rich annotations for super-aged seniors. *arXiv preprint arXiv:2503.16578*.

Yang Chen, Hui Wang, Shiyao Wang, Junyang Chen, Jiabei He, Jiaming Zhou, Xi Yang, Yequan Wang, Yonghua Lin, and Yong Qin. 2025b. Seniortalk: A chinese conversation dataset with rich annotations for super-aged seniors. *Preprint*, arXiv:2503.16578.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021a. DialogSum: A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074, Online. Association for Computational Linguistics.

Yulong Chen, Yang Liu, and Yue Zhang. 2021b. Dialogsum challenge: Summarizing real-life scenario dialogues. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 308–313.

Nelson Cowan. 2001. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24(1):87–114.

Jiaxin Fan, Yong Zhang, Hanzhang Li, Jianzong Wang, Zhitao Li, Sheng Ouyang, Ning Cheng, and Jing Xiao. 2023. Boosting chinese ASR error correction with dynamic error scaling mechanism. In *Proceedings of Interspeech 2023*, pages 2173–2177, Dublin, Ireland.

Zishan Guo, Linhao Yu, Minghui Xu, Renren Jin, and Deyi Xiong. 2023. Cs2w: A chinese spoken-to-written style conversion dataset with multiple conversion types. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3962–3979.

Yuechan Hu, Qianxi Lv, Esther Pascual, Junying Liang, and F. Huettig. 2021. Syntactic priming in illiterate and literate older chinese adults. *Journal of Cultural Cognitive Science*, 5:267 – 286.

Lihe Huang and Deyu Zhou. 2025. Multimodal corpus of gerontic discourse (mcgd). Developed at Tongji University; available upon request.

Youtarou Iida and Yumi Wakita. 2021. Topic-shift characteristics of japanese casual conversations between elderlies and between youths. pages 418–427.

Alyssa M Lanzi, Anna K Saylor, Davida Fromm, Houjun Liu, Brian MacWhinney, and Matthew L Cohen. 2023. Dementiabank: Theoretical rationale, protocol, and illustrative analyses. *American Journal of Speech-Language Pathology*, 32(2):426–438.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Na Liu, Quanlin Pu, Yan Shi, Shengtai Zhang, and Luyi Qiu. 2023. Older Adults' Interaction with Intelligent Virtual Assistants: The Role of Information Modality and Feedback. *International Journal of Human Computer Interaction*, 39(5):1162–1183.

Pingping Liu, Q. Lu, Zhen Zhang, Jie Tang, and B. Han. 2021. Age-related differences in affective norms for chinese words (aanc). *Frontiers in Psychology*, 12.

Minxia Luo, Rudolf Debelak, Gerold Schneider, Mike Martin, and Burcu Demiray. 2020. With a little help from familiar interlocutors: real-world language use in young and older adults. *Aging & Mental Health*, 25:2310 – 2319.

Paul Michel and Graham Neubig. 2018. MTNT: A testbed for machine translation of noisy text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 543–553, Brussels, Belgium. Association for Computational Linguistics.

Haojie Pan, Junpei Zhou, Zhou Zhao, Yan Liu, Deng Cai, and Min Yang. 2018. Dial2desc: End-to-end dialogue description generation. *CoRR*, abs/1811.00185.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Letian Peng, Zuchao Li, and Hai Zhao. 2024. Fast and accurate incomplete utterance rewriting. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Charlene Pope and Boyd H Davis. 2011. Finding a balance: The carolinas conversation collection.

Alisha Pradhan, Amanda Lazar, and Leah Findlater. 2020. Use of intelligent voice assistants by older adults with low technology use. *ACM Trans. Comput.-Hum. Interact.*, 27(4).

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Nikhil Saini, Preethi Jyothi, and Pushpak Bhattacharyya. Survey: Exploring disfluencies for speech to text machine translation.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Zhewei Sun, Richard S. Zemel, and Yang Xu. 2021. A computational framework for slang generation. *CoRR*, abs/2102.01826.

Huan Wang, Zhonggen Yu, and Xiaohui Wang. 2023. Expertise differences in cognitive interpreting: A meta-analysis of eye tracking studies across four decades. *Wiley interdisciplinary reviews. Cognitive science*, page e1667.

Minli Wang and Min Wang. 2024. Age-related differences in the interplay of fluency and complexity in chinese-speaking seniors' oral narratives. *International journal of language & communication disorders.*

Xuancong Wang, Khe Chai Sim, and Hwee Tou Ng. 2014. Combining punctuation and disfluency prediction: An empirical study. pages 121–130.

Zhengxu Yan, Victoria Dube, Judith Heselton, Kate Johnson, Changmin Yan, Valerie Jones, Julie Blaskewicz Boron, and Marcia Shade. 2024. Understanding Older People's Voice Interactions With Smart Voice Assistants: A New Modified Rule-based Natural Language Processing Model with Human Input. *Frontiers in Digital Health*, 6:1329910.

X. Zhou, Ruying Bao, and W. Campbell. 2022. Phonetic embedding for asr robustness in entity resolution. pages 3268–3272.

A Data Collection

A.1 Automatic Sentence Segmentation

Subtitle Text	After Segmentation
93 还开车呢 Still driving at 93 开法拉利 Driving a Ferrari	93 还开车呢，开法拉利。 Still driving at 93, driving a Ferrari.
99 走了 99 passed away 少中苦不算苦 Suffering when young doesn't count as suffering 老来贫不算贫 Being poor when old doesn't count as being poor	99 走了，少中苦不算苦，老来贫不算贫。 99 passed away. Suffering when young doesn't count as suffering; being poor when old doesn't count as being poor.

Table 5: Examples of subtitle text before and after sentence segmentation. Chinese utterances are annotated with English glosses.

The raw subtitle transcripts collected from older adults' interview videos are originally unpunctuated. Each line represents a prosodically coherent short utterance, but typically does not form a complete sentence. As illustrated in Table 5, we preprocess these raw utterances by inserting appropriate punctuation and merging lines based on semantic coherence and prosodic continuity.

This segmentation process relies on contextual understanding of meaning. While each original line is internally coherent, some adjacent lines share tight semantic and prosodic connections and should be merged into a single sentence. To ensure consistency and quality in downstream training, we constrain sentence length to avoid overly long or uninformative segments.

Naive segmentation based on character count may lead to semantically incoherent groupings or unnatural splits. Therefore, we leverage a state-of-the-art large language model, DeepSeek-V3[13], to automatically segment multi-line, unpunctuated text into well-formed sentences. The model is prompted with the following instruction:

> **Prompt:** You are a linguistic annotator. Given a list of short, unpunctuated utterances transcribed from spoken Chinese, please insert appropriate punctuation and merge them into complete, well-formed written sentences. Preserve semantic coherence and keep sentence lengths reasonable.

This automatic segmentation constitutes the foundation of our sentence-level spoken-to-written dataset, resulting in a total of 10,004 older adults' utterances.

A.2 Error Categories and Manual Annotation Guidelines

A.2.1 Error Categories in Older Adults' Spoken Language

We identify six major types of spoken-language errors in our dataset of older adults' speech. These categories are derived from empirical observations during manual annotation and form the foundation of our normalization guidelines. Each category is defined below with illustrative examples.

Redundancy, including Fillers and Repetition. Redundancy in spoken language refers to the use of excessive or superfluous expressions that do not contribute new information. It primarily includes two forms: fillers, i.e., meaningless discourse markers (e.g., "嗯" (um)) used to fill pauses or hesitations in speech; and repetition, i.e., the unnecessary reiteration of words or phrases that add no semantic value.

e.g. 嗯，那个，就是吧，我觉得这个事情呢，挺好的。(Um, well, I mean, I think this thing, um, is quite good.)

This sentence contains multiple fillers that add no meaning.

Omission and Simplification. This type of error involves the omission of key grammatical constituents such as subjects, verbs, or objects, making the sentence rely heavily on contextual inference. While common in spontaneous speech, such omissions often lead to ambiguity or incompleteness in written language.

e.g. 6 个 4 个党员。(Six, four were Party members.)

This utterance omits the full noun phrase "6个兄弟姐妹里" (among the six siblings). The intended meaning is "我们六个兄弟姐妹里有四个是党员。" (Among the six siblings, four were Party members.)

[13] https://github.com/deepseek-ai/DeepSeek-V3

Colloquial and Dialectal Expressions. This category includes informal, region-specific, or generational terms that are commonly used in everyday spoken language but are inappropriate for formal written expression. These expressions often reflect local dialects or age-group idiosyncrasies and may hinder comprehension for readers unfamiliar with the speaker's background.

e.g. 我来说吧，因为是我行二。(Let me speak, because I'm ranked second.)

"行二" is a colloquial way of indicating birth order among siblings and should be expressed more clearly in writing, e.g., "我是家中排行第二" (I'm the second-born in the family).

Ambiguous or Inconsistent Pronoun Use. This category refers to the unclear or inconsistent use of personal pronouns such as "他" (he), "她" (she), or "我们" (we) without identifiable antecedents or with shifting referents in the same sentence. Such usage can confuse the listener or reader, making it difficult to determine who is being referred to.

e.g. 我弟妹说让我们把家腾干净了，我们要来住了。(My sister-in-law said we should clear out the house—we' re going to move in.)

The second instance of "我们" (we) should actually refer to "他们" (they), but the pronoun is incorrectly used, leading to confusion.

Disorganized Syntax. This category includes structurally incomplete or overly convoluted sentences that impair comprehension. Common issues include missing core elements (e.g., subject, verb, or object), unclear syntactic dependencies, and excessively long or disjointed constructions. These errors are especially prevalent in spontaneous spoken language and require restructuring for clarity in written form.

e.g. 三班倒，所以这个时间基本上一个星期两个星期，所谓有个大大公休吧，那阵倒班，回家一次。(She was working in rotating shifts, so she could usually only go home once every one or two weeks when there happened to be a long public break.)

This sentence aims to convey that she worked on a three-shift rotation and could only return home once every one to two weeks, typically during extended rest periods ("大公休"). However, the original utterance is fragmented and includes multiple vague or redundant expressions, which obscure the intended meaning and make the structure difficult to follow.

Self-Repair. This category refers to speech disfluencies or self-corrections that occur spontaneously during verbal expression. These include slips of the tongue, mid-sentence revisions, and other forms of unintended speech errors. While natural in conversation, such phenomena can introduce redundancy, ambiguity, or grammatical inconsistencies when transcribed directly.

e.g. 我们哥五个哥六个觉得一定得这么办。(Our five—or six brothers think we must do it this way.)

This sentence includes a mid-sentence correction: the speaker first says "哥五个" (five brothers) and immediately corrects it to "哥六个" (six brothers). This kind of self-repair is common in spontaneous speech but should be edited for clarity in written form.

A.2.2 Manual Annotation Guidelines

To convert spoken utterances into coherent written form, annotators follow a two-step procedure grounded in the six identified categories of spoken-language errors shown in A.2.1. As multiple error types may co-occur within a single utterance, the annotation includes two components: 1) Error Labeling, where each detected error is annotated with its corresponding type and a brief description of its manifestation; 2) Written-text Correction, where the utterance is revised into its well-formed written counterpart.

To ensure consistency, annotators follow a standardized two-step workflow:

Step 1: Error Identification. Identify which of the six error types are present in the utterance and provide a brief description of each.

Step 2: Targeted Revision. For each identified error, revise the utterance accordingly to produce a fluent, complete, and stylistically appropriate written counterpart.

A.3 Formal Proof

Notation. Let the gold (well-formed) sentence be $S = \langle c_1, \ldots, c_n \rangle$ and the observed (ill-formed) sentence be $\widetilde{S} = \langle d_1, \ldots, d_m \rangle$, where each c_i or d_j is a *sentence constituent* (e.g., subject, predicate, object).

[Atomic Operations]

$$\mathcal{O} = \{ \underbrace{\Delta(i)}_{\text{Deletion}}, \underbrace{\mathrm{I}(x,j)}_{\text{Insertion}}, \underbrace{\Sigma(y,i)}_{\text{Substitution}}, \underbrace{\tau(i,k)}_{\text{Permutation}} \}.$$

Each operation acts on one constituent of S:

- $\Delta(i)$ removes the constituent c_i;
- $\mathrm{I}(x,j)$ inserts a new constituent x at position j;
- $\Sigma(y,i)$ replaces c_i with a different constituent $y \neq c_i$;
- $\tau(i,k)$ swaps the constituents at positions i and k (equivalently, applies a permutation π to their indices).

[Completeness] For any finite sentences $S = \langle c_1, \ldots, c_n \rangle$ and $\widetilde{S} = \langle d_1, \ldots, d_m \rangle$, There exists a finite sequence of operations $E = (o_1, \ldots, o_r) \subseteq \mathcal{O}$ such that $E(S) = \widetilde{S}$.

Let $L = \mathrm{LCS}(S, \widetilde{S})$ be the longest common subsequence with respect to constituents. Construct E in four stages:

1. For every constituent $c_i \in S \setminus L$, apply $\Delta(i)$.
2. For every constituent $d_j \in \widetilde{S} \setminus L$, apply $\mathrm{I}(d_j, j)$ at its target position.
3. Let π be the minimal permutation that aligns the current sequence with the order in $\widetilde{S}$; realise π using a sequence of $\tau(i,k)$ operations.
4. After alignment, any residual mismatch of constituents (identical position but different content) is fixed by $\Sigma(d_j, i)$.

Because each step draws solely from $\mathcal{O}$, the composed transformation E maps S to $\widetilde{S}$.

[Independence] Assume each atomic operation in $\mathcal{O}$ costs 1. Then no single operation can be *exactly* simulated by any multiset of the remaining three at a total cost ≤ 1.

For a sentence T, let

$$\big(|T|,\ \mathrm{bag}(T),\ \mathrm{order}(T)\big)$$

denote, respectively,

(i) its number of constituents, (ii) the unordered multiset (bag) of those constituents, and (iii) their left-to-right order.

1. If $o = \Delta$, then the target effect is $|T| \mapsto |T| - 1$. None of the remaining operations decreases $|T|$.
2. If $o = \mathrm{I}$, the argument is symmetric.
3. If $o = \Sigma$, the target effect is to leave $|T|$ and $\mathrm{order}(T)$ unchanged, while modifying $\mathrm{bag}(T)$ at one position.
4. If $o = \tau$, we must permute two constituents while leaving the bag intact. Without τ, the only way to change order is to perform *two* substitutions $c_i \mapsto c_k$ and $c_k \mapsto c_i$. Again the cost is ≥ 2.

Hence, no operation can be simulated by the others at equal (or lower) cost, establishing mutual independence.

Conclusion. The four error types—Deletion, Insertion, Substitution, and Permutation—constitute a complete and mutually irreducible basis for sentence-level error classification.

A.4 Context Experiment Result

Table 6 presents the evaluation results of different context settings as assessed by two PhD students.

A.5 LLM Annotation Prompt

You are a documentation editor at an older adult service organization. Your task is to accurately and clearly transform oral narratives from older adult individuals into written texts in a natural, everyday written style. You should also identify the types of errors present in the spoken utterance.
First, read and analyze the contextual content surrounding the oral sentence. Summarize the main idea of the paragraph in Chinese to ensure you have understood the discourse structure and the core message. Then, for the target spoken sentence, follow the four steps below in sequence to detect and correct four types of errors, ensuring semantic consistency with the original meaning.
For each step, first determine whether the sentence contains this type of error. If so, list the corresponding error type number under "Error Type" and correct the sentence accordingly. If no error is detected, do not output the number.
1. Constituent Omission. The original sentence lacks essential components.
`Correction:` Supplement the missing parts based on the intended meaning to make the sentence clear and easy to understand.
`Examples:`
母亲一直跟我在一起。(Missing "生活") My mother has always been with me. (Missing the

Context Length	Context Type	4-Sent. (Written)	4-Sent. (Spoken)	Full (Written)	Full (Spoken)	Single Sentence
Semantic	Completeness	42/34	24/20	13/20	14/11	7/15
	Reading Fluency	42/30	24/22	11/21	15/10	8/17
Total Selection Rate		74.00%	45.00%	32.50%	25.00%	23.50%

Table 6: Evaluation of Transformation Results under Different Context Lengths and Types. Each cell shows the number of utterances selected by two PhD annotators.

verb “live”)
母亲的影响尤其我们日后都退了休了又跟他在一起生活，影响越来越深刻了。(Missing “对我们的影响”) My mother’s influence became increasingly profound, especially after we both retired and started living with her again. (Missing “her influence on us”)
2. Disordered Syntax. The sentence contains disordered syntax or excessive parentheticals, making it difficult to follow.
`Correction:` Extract the main message, adjust the word order, and simplify or remove extraneous elements.
`Examples:`
因为他们也是三班倒，那阵都是工人嘛。Because they were also working in three shifts—everyone was a laborer back then.
三班倒，所以这个时间基本上一个星期两个星期，所谓有个大大公休吧，那阵倒班，回家一次。Because of the three-shift system, we basically had time off only once every one or two weeks—what they called a ”major rest day” back then. Due to the rotating shifts, we could only go home occasionally.
3. Constituent Errors. Some expressions are inappropriate, such as wrong pronouns.
`Correction:` Fix incorrect expressions and resolve ambiguous pronoun references.
`Examples:`
我弟妹说让我们把家腾干净了，我们要来住了。(“我们” → “他们”) My sister-in-law said, “You need to clear out the house —we’re going to move in.” (“we” should be “they”)
4. Constituent Redundancy. The sentence includes meaningless repetitions or corrections.
`Correction:` Remove redundant or self-repaired parts.
`Examples:`
我们哥五个哥六个觉得一定得这么办。(删除 “哥五个”) The six of us brothers felt that this was something we absolutely had to do. (Delete “哥五个”)
嗯，那个，就是吧，我觉得这个事情呢，挺好的。(去除填充词) Yeah, well, I think this thing is pretty good. (Remove fillers)
After correcting all types of errors, refine the overall sentence style. Replace informal expressions with moderately formal written ones. Maintain a natural and fluent tone, and convert region-specific or generational expressions into contemporary standard Mandarin. For polysemous colloquial words, choose the most natural and unambiguous interpretation based on context.
`Examples:`
她虚岁 99 岁走的。(“走的” → “去世”) She passed away at the nominal age of 99. (Replace “走的” with “去世”)
我们老头那阵有咱说实在的，倒退 30 年都求着我们来都知道吗？(“老头” → “丈夫”) Back then, my husband—honestly—people were begging us to come even 30 years ago, you know? (Replace “老头” with “丈夫”)
Original Sentence: `{oral_sentence}`
Context (for understanding only, do not translate): `{context}`
Only output the translation result and error type number for the original sentence. Do not output reasoning or explanation. Use the following format:
`Translation Result:`
`Error Type:`

A.6 Data Example

An example from the dataset is shown in Listing 1.

Listing 1: An example instance from the COAS2W dataset

```
{
  "id": 12,
  "file_id": 59,
  "spoken_text": "就我这个还活,按属性说啊,我属
      虎是6个老虎啊,死了5个呦,最后就剩我一个了
      嗯。(As for me, I am still alive.
      According to my attributes, there are
      6 tigers in the Tiger genus, and 5
      have died. In the end, I am the only
      one left.)",
  "context": "我姓苏，苏家是个大家族，如今我在
      兄弟中排行第二。我的兄长们都已去世，我们
      家兄弟不多，也就十五六个，兄长们都已离
      世。就我这个还活,按属性说啊,我属虎是6个
      老虎啊,死了5个呦,最后就剩我一个了嗯。年
      轻时候啊没少吃苦啊,吃不上饭,受累是这个。
      为什么没有文化呀？啊那念书念不紧,那能挣
      多少钱赶马车,那就挣多少咱不管,就咱就图这
      5毛钱啊。(My surname is Su. The Su
      family is a big family, and now I am
      the second among my brothers. My older
       brothers have all passed away. There
      are only fifteen or sixteen brothers
      in our family, and all of them have
      passed away. As for me, I am still
      alive. According to my attributes,
      there are 6 tigers in the Tiger genus,
       and 5 have died. In the end, I am the
       only one left. When I was young, I
      suffered a lot. I couldn't afford to
```

```
        eat, and that's why I was burdened.
        Why is there no culture? Ah, if you
        can't study hard, then you can earn as
         much money as you want to drive a
        carriage. We don't care how much you
        earn, we just want this 50 cents.)",
    "written_text": "按生肖来说，我属虎，原本有
        六个属虎的兄弟，如今五人已去世，只剩下我
        一个了。(According to the zodiac sign,
        I belong to the tiger. Originally, I
        had six brothers born in the year of
        the tiger, but now five of them have
        passed away, leaving only me.)",
    "error_type": [
      1,
      2,
      4
    ]
  }
```


B Topics of Old Adults

To better characterize the content of older adults' speech, we categorize utterances into four high-level thematic topics. These categories are derived from empirical observations and manual analysis during corpus construction. Table 7 provides definitions and representative examples for each topic.

C Evaluate

C.1 Metric Definitions

C.1.1 Error Type Detection

We evaluate whether the predicted error labels match the annotated labels for each sentence.

- **Joint Accuracy** = $\frac{1}{N}\sum_{i=1}^{N} 1[\hat{Y}_i = Y_i]$: the prediction is correct only if all gold labels are exactly matched.
- **Acc-1** = $\frac{1}{N}\sum_{i=1}^{N} 1[\hat{Y}_i \cap Y_i \neq \emptyset]$: the prediction is correct if at least one gold label is identified.

C.1.2 Spoken-to-Written Generation Quality

We assess the quality of generated written text using:

- **BLEU** (Papineni et al., 2002): Measures n-gram precision with a brevity penalty, reflecting surface-level fluency.
- **ROUGE-L** (Lin, 2004): Based on the longest common subsequence (LCS), evaluating content recall.
- **BLEURT** (Sellam et al., 2020): A pretrained semantic metric that captures meaning similarity beyond lexical overlap.

C.2 Fine-tuning Settings

Table 8 summarizes the LoRA fine-tuning hyperparameters used for different models in our experiments. All experiments were conducted on an NVIDIA RTX 4090 GPU with 24 GB of VRAM. The software environment includes Python 3.10.12, PyTorch 2.6.0 with CUDA 12.4, Transformers 4.51.3, and Ubuntu 22.04 as the operating system.

C.3 Calculation of Average Performance Gains

To quantify the performance improvements brought by fine-tuning on the COAS2W dataset, we report the **average absolute gains** across models for each evaluation metric. Specifically, for each metric

$$M \in \{\,\text{JA, Acc-1, BLEU-1, BLEU-2, BLEU-3, BLEU-4, ROUGE-L, BLEURT}\,\}$$

, and for each model i, we compute the absolute gain as:

$$\text{Gain}_M^{(i)} = M_{\text{w/FT}}^{(i)} - M_{\text{w/oFT}}^{(i)}$$

where $M_{\text{w/FT}}^{(i)}$ and $M_{\text{w/oFT}}^{(i)}$ denote the values of metric M for model i under the fine-tuned and zero-shot settings, respectively.

The **average gain** for metric M is then obtained by averaging across all $N = 4$ models:

$$\text{Average Gain}_M = \frac{1}{N}\sum_{i=1}^{N} \text{Gain}_M^{(i)}$$

This procedure ensures a fair and model-agnostic quantification of fine-tuning benefits and allows for direct comparison of improvement magnitudes across different evaluation dimensions.

Topic	Description	Example
Life Experience	Early-life recollections, career experiences, and reflections derived from personal history	Born in Dezhou, Shandong; studied at a specialized school; moved to Heilongjiang; war experiences; assigned housing after demobilization
Family Relations	Friends, spouse, children, kinship structures, and family changes	Two children; helping daughter care for grandchildren; spouse passed away
Life in Old Age	Retirement, healthcare; physical conditions; hardship or well-being in old age	Singing opera; cooking; caring for grandchildren; shopping difficulties; pension, healthcare, illness
Social Values	Perceptions of social change; evaluations of social events; life attitudes	"We used to starve; now we can eat our fill"; gratitude; distrust in children

Table 7: Topic Definitions and Examples in the Older Adults' Speech Dataset.

Model	Epochs	Batch	Grad Acc.	LR	Rank	Alpha	Scheduler	Dropout	Time(h)
Qwen	3	4	4	5e-5	8	32	cosine	0.1	1.79
Mistral	3	2	4	5e-5	8	32	cosine	0.1	2.49
ChatGLM	3	2	8	5e-5	8	32	cosine	0.1	1.85
Baichuan	3	4	4	5e-5	8	32	cosine	0.1	1.51

Table 8: LoRA Fine-tuning Hyperparameters for Different Models.