Recording for Eyes, Not Echoing to Ears: Contextualized Spoken-to-Written Conversion of ASR Transcripts

Jiaqing Liu[1], **Chong Deng**[1], **Qinglin Zhang**[1], **Shilin Zhou**[2], **Qian Chen**[1], **Hai Yu**[1], **Wen Wang**[1]
[1]Tongyi Lab, Alibaba Group
[2]School of Computer Science and Technology, Soochow University
{mingzhai.ljq,dengchong.d,qinglin.zql,tanqing.cq,yuhai.yu,w.wang}@alibaba-inc.com {slzhou.cs}@outlook.com

Abstract

Automatic Speech Recognition (ASR) transcripts exhibit recognition errors and various spoken language phenomena such as disfluencies, ungrammatical sentences, and incomplete sentences, hence suffering from poor readability. To improve readability, we propose a **Contextualized Spoken-to-Written conversion (CoS2W)** task to address ASR and grammar errors and also transfer the *informal* text into the *formal* style with content preserved, utilizing contexts and auxiliary information. This task naturally matches the in-context learning capabilities of Large Language Models (LLMs). To facilitate comprehensive comparisons of various LLMs, we construct a document-level **S**poken-to-**W**ritten conversion of **A**SR Transcripts **B**enchmark (**SWAB**) dataset. Using SWAB, we study the impact of different granularity levels on the CoS2W performance, and propose methods to exploit contexts and auxiliary information to enhance the outputs. Experimental results reveal that LLMs have the potential to excel in the CoS2W task, particularly in grammaticality and formality, our methods achieve effective understanding of contexts and auxiliary information by LLMs. We further investigate the effectiveness of using LLMs as evaluators and find that LLM evaluators show strong correlations with human evaluations on rankings of faithfulness and formality, which validates the reliability of LLM evaluators for the CoS2W task.

Datasets — https://github.com/alibaba-damo-academy/SpokenNLP/tree/main/swab

Extended version — https://www.arxiv.org/abs/2408.09688

Introduction

As can be seen in Figure 1, since ASR transcripts aim to provide verbatim transcriptions of oral communications, they often exhibit various spoken language phenomena and informal styles, such as filler words, repetitions, repairs, and fragments, and include ASR errors and ungrammatical text. These characteristics lead to poor readability. To enhance the readability of ASR transcripts, we propose the **Contextualized Spoken-to-Written conversion (CoS2W)** task, which aims to correct ASR errors and grammatical errors and transfer the informal style to the written and formal style while preserving the content, that is, the CoS2W task is designed to convert verbatim transcripts ("echoing to ears") into readable documents ("recording for eyes"). Note that different domains vary in their requirements for adjustments on ASR transcripts. Some domains only allow removal of filler words while some domains (e.g., speech analysis for diagnosis) require verbatim transcripts. Nonetheless, the readable documents of CoS2W task are particularly helpful for many domains such as podcasts, education, and meetings (business, project, etc.), where efficiency in information delivery and knowledge acquisition is essential.

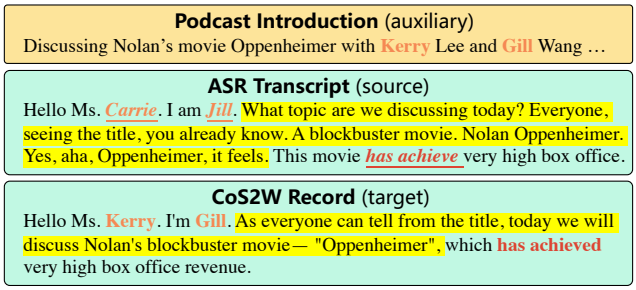

Figure 1: An example of the proposed CoS2W task. CoS2W involves Text Style Transfer across sentences, ASR Error Correction and Grammar Error Correction (GEC).

Moreover, for a wide variety of downstream tasks on ASR transcripts, such as machine translation, summarization, question-answering (Gupta et al. 2021), and sentiment classification (Zhang et al. 2023), even competitive models (including LLMs) perform substantially better on written text than on transcripts, and the performance could benefit substantially from applying CoS2W to ASR transcripts. For example, for machine translation of ASR transcripts, we employ the competitive CSANMT model[1] on the development set of the speech translation BSTC dataset (Zhang et al. 2021b) and find that compared to directly translating ASR transcripts, applying CoS2W (with GPT-4) on ASR transcripts then translating improves the BLEURT (Sellam, Das, and Parikh 2020) score from 57.07 to 61.9.

Although the CoS2W task could improve both readability and performance of downstream tasks of ASR transcripts,

[1]https://www.huggingface.co/modelscope-unofficial/damo-csanmt-zh-en-large

there is a lack of research efforts on this task due to its significant challenges, including *task complexity*, *effective exploitation of contexts*, *understanding the impact of granularity*, *the challenge in evaluations*, and *scarcity of labeled data*, which we will detail below.

Task complexity is challenging because CoS2W combines many subtasks such as ASR error correction, Grammatical Error Correction (GEC), and text style transfer (i.e., *informal* spoken language expressions to the *formal* style). These subtasks are not simply pipelined but are interconnected: all subtasks are based on a complete semantic understanding of the input and have dependencies among them. ASR and grammar errors can alter semantics, thereby affecting the performance of text style transfer. On the other hand, text style transfer could benefit GEC as GEC performs better on formal texts than spoken texts.

Understanding the impact of granularity is necessary. Traditional text style transfer focuses on sentence-level conversion. Conversion at a lower granularity (e.g., paragraph-level) instead of sentence-level, may achieve larger improvements to readability, as shown in Figure 1. However, reducing granularity may introduce faithfulness issues and output limitation as discussed in the results section.

Effective exploitation of contexts is critical to understand speech transcripts, thus we define CoS2W as contextualized. For example, context is helpful to grasp the meanings of fragments in meetings with frequent speaker interactions. Some domains also provide auxiliary information like introductions of podcasts, which could benefit ASR error correction as it may contain relevant named entities, domain terms, and words with challenging pronunciations. Essentially, the auxiliary information can be considered as *extended context*. Effectively leveraging auxiliary information to enhance the CoS2W performance poses a challenge.

Challenge in evaluations arises due to its multifaceted nature, requiring different evaluation methods for its subtasks. Moreover, tasks such as informal-to-formal style transfer remain highly subjective and predominantly depend on human evaluations despite the existence of objective metrics. LLMs offer a promising opportunity for automatic evaluations. However, leveraging LLMs as reliable evaluators for CoS2W poses many challenges, including designing appropriate prompts, mitigating bias, and ensuring the accuracy, fairness, and reproducibility of the LLM evaluators.

Scarcity of labeled data for CoS2W hinders model evaluations and efforts to tackle the challenges discussed above.

The advent of LLMs offers new prospects for addressing the challenges of CoS2W. Firstly, the contextualized CoS2W task matches well with LLMs' in-context learning abilities. Secondly, the CoS2W task considers an output satisfactory as long as it is free of ASR and grammar errors, faithful to the input in semantics, and formal (i.e., close to written text). The generative nature of the problem definition also aligns with LLMs. Thirdly, LLMs have been explored as evaluators for many tasks (Li et al. 2024; Lai, Toral, and Nissim 2023) and achieved notable performance.

Our contributions can be summarized as follows:

- We propose the Contextualized Spoken-to-Written conversion (CoS2W) task to improve both readability of ASR transcripts and performance of downstream tasks. To promote research in this field, we construct and make available the *document-level* **Spoken2Written of ASR transcripts Benchmark (SWAB)** dataset with manual annotations, covering meeting, podcast, and lecture domains in both Chinese and English languages.
- We investigate various methods for LLMs to utilize contexts and auxiliary information for CoS2W and provide insights into future research directions.
- We find that LLM evaluators show strong correlations with human evaluations on rankings of faithfulness and formality and also analyze their strengths and weaknesses as CoS2W evaluators.

Related Work

ASR Error Correction task aims to correct misrecognitions within ASR transcripts. Recent studies (Min and Wang 2023; Yang et al. 2023) have investigated LLMs' effectiveness in ASR error correction, using varied prompt strategies like Chain of Thought (CoT). LLMs show potential for this task, but their free-generation paradigm often causes unnecessary changes, like paraphrasing error-free text, affecting performance on metrics like Word Error Rates (WER). This problem would also be an issue for the CoS2W task.

In this task, auxiliary information helps resolve challenging pronunciations, such as N-best transcripts (Ma et al. 2023), dialogue history (Mai and Carson-Berndsen 2023), video titles and descriptions (Lakomkin et al. 2024), rare words (He, Yang, and Toda 2023), etc. However, this auxiliary information is often brief, comprising just a few sentences or word lists. And studies combining auxiliary information with LLMs for this task are relatively scarce.

Grammar Error Correction task entails correcting textual grammatical mistakes. Similar to ASR error correction, LLM performance in GEC isn't fully satisfactory (Zhang et al. 2023; Fang et al. 2023). GEC datasets usually follow minimum change principle, clashing with LLMs' relatively free generation paradigm. Disregarding objective metrics, human evaluations confirm LLMs' GEC effectiveness (Li et al. 2023), which suggests current metrics don't evaluate LLMs fairly and reliably. Therefore, Li et al. (2024) aim to refine GEC evaluation methods, using LLMs to categorize edits and calculate metrics like $F_{0.5}$ to evaluate models.

Most GEC task corpora are sentence-level, sourced from second-language learners, and categorized into error-coded and direct rewriting paradigms. Table 1 lists common datasets: For English, FCE (Yannakoudakis, Briscoe, and Medlock 2011) and AESW (Daudaravicius et al. 2016) use error-coded, while JFLEG (Napoles, Sakaguchi, and Tetreault 2017) and WI-LOCNESS (Bryant et al. 2019) use direct rewriting. Chinese datasets include NLPCC18 and Lang-8 (Zhao et al. 2018), CGED (Rao, Yang, and Zhang 2020) from HSK essays, and re-annotated YACLC (Wang et al. 2021) and MuCGEC datasets (Zhang et al. 2022).

Text Style Transfer task aims to modify text to a specific style (e.g., informal to formal, negative to positive) while preserving content. The CoS2W task focuses on informal to formal transfer. Recent studies (Reif et al. 2022; Tao et al. 2024) have employed LLMs to improve text style transfer

Dataset	#Sents	Language & Domain
FCE	34.0K	EN FCE Exam Essay
AESW	1489.2K	EN Journal Articles
JFLEG	1.5K	EN TOFEL Exam
WI-LOCNESS	43.1K	EN Website, Essay
NLPCC18	2.0K	CH Essay
Lang-8	717.0K	CH Language-learning Website
CGED	7.2K	CH HSK Exam
YACLC	32.1K	CH Website
MuCGEC	7.1K	CH Essay, HSK Exam, Website
Japanese S2W	18.2K	JA Conversation, Voicemail
CS2W	7.2K	CH Telephone Conversation
SWAB	29.5K	EN & CH Document-level annotated Podcasts, Meetings, Lectures with Auxiliary

Table 1: Comparison between SWAB and other datasets.

quality and achieve diverse style conversions. In addition, Lai, Toral, and Nissim (2023) employ the LLM as a multi-dimensional evaluator and find that it achieves competitive correlations with human evaluation compared to existing automatic metrics, especially in terms of content preservation.

Spoken-to-Written task transforms ASR transcripts into formal and readable text, which was initially proposed for Japanese (Ihori, Takashima, and Masumura 2020) and has later been extended to Chinese (Guo et al. 2023).

Different from previous work, the CoS2W task emphasizes the contextualized ability, aiming to convert whole paragraphs across multiple sentences rather than single sentences, with the help of context and auxiliary information. It will further enhance the readability of the results, as illustrated in Figure 1. And it aligns well with the in-context learning capabilities of LLMs. To support this research, we construct the SWAB dataset and report the experiment and evaluation results of LLMs. Additionally, our research spans multiple scenarios and languages as shown in Table 1.

The SWAB Dataset

To conduct experiments and evaluations of the CoS2W task, we construct the **S**poken2**W**ritten of **A**SR Transcripts **B**enchmark (**SWAB**) dataset, including multiple scenarios (i.e., podcasts, meetings, lectures) in both Chinese and English. There are 60 transcripts with auxiliary information, with each subcategory comprising 10 documents. The SWAB provides **document-level** annotations for entire transcripts, offering an annotated target for each paragraph. Furthermore, we provide links to the original audio or video with timestamps to support multi-modal research.

Data Source

We collect Chinese and English data by leveraging the abundant data resources available. And we select three typical scenarios (i.e., meeting, podcast, and lecture) where efficiency in information delivery is essential. Among them, meetings are high-frequency interactive discussions among multiple participants; podcasts typically involve chats and interviews between two or more individuals; and lectures are monologues from a single person. Essentially, these domains differ significantly in interactivity, thus varying from written texts and posing different challenges to the CoS2W task.

Meetings are sourced from open-source meeting corpora with manually annotated information as auxiliary information. Chinese meetings are sampled from the training dataset of the AliMeeting corpus (Yu et al. 2022). We select the title and manually annotated topic titles (i.e., sub-topics) as auxiliary information. English Meetings are sampled from the AMI corpus (Carletta et al. 2005). We choose third-party annotated abstractive summarization as auxiliary information, which includes sections on decisions, action items, etc.

Podcasts are derived from many Chinese and English podcast programs covering a multitude of topics on YouTube[2]. We collect the podcast introduction as auxiliary information, which provides rich background knowledge such as guest names, specialized terminology, and discussion topics.

Lectures are sourced from many Chinese and English individual speeches on YouTube, covering a variety of topics. Both Chinese and English lectures are provided with meta-information of the speeches as auxiliary information, which includes the name and biography of the speaker, as well as the title, category, and description of the video.

Dataset Construction

We design a relatively free annotation paradigm like direct rewriting of GEC. Targets are considered "correct" as long as they fix ASR and grammatical errors, and adopt a written and formal style while faithfully preserving the original content. We also ensure paragraph-level consistency between source and target to enable flexible content division and fair comparison at various granularity levels.

All collected data are first transcribed using a competitive ASR system (Gao et al. 2022)[3], then the ASR 1-bests are processed with punctuation insertion, paragraph segmentation (Zhang et al. 2021a), and speaker diarization (Zheng et al. 2023). This ASR structure is widely utilized and operates at peak performance while preserving the overall generality of the dataset. We utilize GPT-4 (gpt-4-0125-preview)[4] to obtain the initial target of CoS2W for the SWAB dataset. The input is chunk-level to leverage local context and to reduce token consumption. The auxiliary information is also provided to enhance the ASR error correction performance. The prompt for this procedure is in the arXiv version.

The results of GPT-4 still contain some issues (as shown in our analysis). Consequently, we use human annotation for manual revisions. We recruit more than ten college-educated Chinese annotators. Each annotator must undergo comprehensive training to understand the annotation guidelines and meet the accuracy standards on test data before they can officially begin annotating. Given the transcription (source), model results (target), and auxiliary information, each qualified annotator must thoroughly review and annotate the entire transcript from beginning to end, to ensure consistency

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

[3]https://www.github.com/modelscope/FunASR

[4]https://platform.openai.com/docs/models

at the document level. They are responsible for (1) ensuring paragraph-level consistency between source and target, (2) correcting any remaining ASR errors and grammatical errors, and (3) guaranteeing that the content preserves the original content and embodies a written and formal style. During the annotation process, any confusing cases will be discussed, and the annotation guidelines and examples will be continuously updated to ensure clarity and accuracy.

Quality assurance is managed by three senior annotators. They will conduct subjective quality checks on 10% to 20% of the sampled paragraphs and utilize heuristic methods to perform rule-based inspections on all paragraphs. For documents with an error rate of less than 10%, identified and similar mistakes must be corrected. For documents with an error rate greater than 10%, a complete revision is required until the document passes the inspection.

We observe that the CoS2W task shows significant diversity. Due to labeling costs, we currently provide only one target but aim to expand target diversity in future work. Therefore, some objective metrics may not adequately reflect the performance. As a result, we have adopted metrics that focus more on semantics and introduced evaluations by LLMs and humans to mitigate the limitations of a single target.

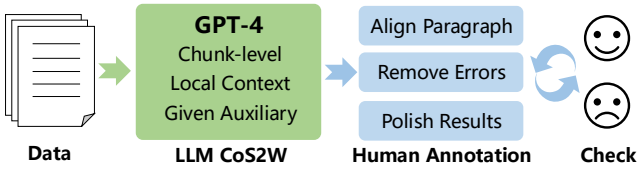

Figure 2: The pipeline of constructing the SWAB dataset.

Dataset Analysis

The target length of the final results is streamlined to 85.40% of the source length. Additionally, annotators adjust 10.43% of the paragraph segmentation results to improve segmentation and ensure paragraph-by-paragraph consistency between the source and target. In total, 20.09% of the paragraphs are identified and corrected for ASR errors. Within the GPT-4 results, 11.29% of the paragraphs still contain ASR errors and are further modified manually by annotators. Only 2% of the paragraphs in the GPT-4 results have grammatical errors to be corrected manually. Annotators further optimize 20.26% paragraphs of GPT-4 to achieve a better written and formal style. Note that further optimization does not imply errors in the formal style of GPT-4 results.

Method

We compare the performance on the CoS2W task of different LLM across various granularity levels, and introduce contexts and auxiliary information to enhance model performance. To evaluate comprehensively, we employ automatic metrics, the LLM Evaluator, and human evaluation.

Different LLMs

To assess different LLMs' abilities in solving this complex task, we choose 4 typical LLMs across both open-source and closed-source models. The open-source LLMs selected are QWen-14B[5] and LLaMA3-8B[6], chosen for their suitable computational demands for future SFT-related work. For closed-source LLMs, we pick GPT-4 (gpt-4-0125-preview)[7] and QWen-Max (qwen-max-0107)[8], to obtain the state-of-the-art performance of LLMs on this task.

Different Granularity Levels

To explore the impact of granularity level, we construct comparisons at the document level, chunk level, and paragraph level, ranging from coarse to fine granularity. Document-level granularity requires the model to rewrite all paragraphs of one entire document, using paragraph indexes to correlate the source and target. At the chunk-level granularity, we divide the document into chunks based on length (1.5K tokens). The model is required to rewrite each paragraph within the chunk on a paragraph-by-paragraph basis. For paragraph-level granularity, we conduct experiments focusing on the revision of individual paragraph content. Furthermore, we set chunk-level granularity as the baseline, as it allows the use of local context while requiring fewer tokens.

Context

Context is essential in text understanding, especially for spoken ASR transcripts. We divide context into local context (neighboring paragraphs) and global context (semantically related but non-adjacent paragraphs) (Han, Soldaini, and Moschitti 2021). The global context is acquired by retrieving top-k relevant paragraphs using BM25 (Robertson and Zaragoza 2009). For chunk-level experiments, paragraphs within a chunk naturally possess local context. To assess the influence of context more accurately, we conduct paragraph-level experiments. By providing 4 local or global paragraphs, we compare the effects of context.

We utilize auxiliary information to enhance ASR error correction by one-step and two-stage methods. The one-step method integrates full or retrieved auxiliary information as a reference part of the context in the prompt. The retrieved part is top-K relevant sentences of auxiliary information for current content based on BM25. For the two-stage approach, we first use LLMs to extract the summary or keywords from the auxiliary information, with the prompts detailed in the arXiv version. We require the summary to retain all key terms as much as possible, especially challenging words for ASR. The keywords should be categorized by their entity types. Next, we use either the summary or keywords as references.

Evaluations

We thoroughly evaluate using objective metrics, LLM Evaluator, and human evaluation. For objective metrics, we calculate BLEU (Papineni et al. 2002), ROUGE (Lin 2004), and BLEURT (Sellam, Das, and Parikh 2020) with the human target as reference. It's important to emphasize that the capabilities of BLEU and ROUGE are limited by only one

[5]https://www.modelscope.cn/models/qwen/Qwen-14B-Chat
[6]https://www.github.com/meta-llama/llama
[7]https://platform.openai.com/docs/models
[8]https://www.alibabacloud.com/help/en/model-studio/getting-started/models

Model	Auxiliary	BLEU↑	ROUGE-L↑	BLEURT↑	CK-Recall↑	S-Faithful↑	S-Formal↑
Source	None	16.22	41.51	55.19	12.89	7.53	4.97
LLaMA3-8B	None	8.94	20.73	40.53	13.18	6.35	7.33
QWen-14B	None	3.41	15.96	30.14	11.46	3.72	6.04
QWen-Max	None	9.47	28.92	50.58	22.35	6.00	7.55
GPT-4	None	**15.51**	**39.66**	**62.15**	**22.49**	**7.32**	**8.20**
GPT-4	+ Origin	15.33	39.53	62.73	43.70	7.38	8.09
	+ RAG	15.22	39.56	62.48	37.11	7.39	8.07
	+ Summary	**15.51**	**39.98**	**62.75**	40.40	7.43	8.11
	+ Keywords	15.08	39.34	62.72	**45.42**	**7.46**	**8.16**

Table 2: The performance comparison between closed-source and open-source LLMs at the chunk-level, with different auxiliary information utilization strategies, including directly providing (Origin), retrieving the top-10 sentences (RAG), based on the summary (Summary), or a list of keywords (Keywords) derived from LLMs. Source refers to the original ASR transcripts.

Granularity	Context	BLEU↑	ROUGE-L↑	BLEURT↑	CK-Recall↑	S-Faithful↑		S-Formal↑	
						Para.	Chunk	Para.	Chunk
Document	None	1.80	4.97	10.64	5.30	2.28	2.44	4.57	3.74
Chunk	None	15.51	**39.66**	62.15	**22.49**	7.32	**7.57**	8.20	**7.81**
Paragraph	None	**16.32**	36.74	**63.32**	19.20	**7.99**	6.60	**8.31**	7.55
Paragraph	+ Local	**16.44**	**39.73**	**67.11**	22.21	**8.19**	7.41	**8.40**	7.76
	+ Global	15.55	38.73	66.00	23.64	7.89	7.05	8.36	7.74
	+ Both	16.38	39.60	67.10	**23.78**	8.18	**7.42**	8.39	**7.80**

Table 3: The performance of GPT-4 at different levels of granularity including document, chunk, and paragraph under different contexts. Additionally, S-Faithful and S-Formal show results for both paragraph-level and chunk-level evaluations.

target of the SWAB dataset. Therefore, we focus more on semantic-oriented automatic metrics such as BLEURT.

To measure the performance of ASR Error Correction, we design the Challenging Keyword Recall (CK-Recall) metric. The challenging keywords are a subset of ASR errors in transcripts, focusing specifically on named entities (e.g., person names, podcast titles). Correcting these errors can significantly improve the reading experience. Furthermore, these words typically cannot be modified or adjusted, making it convenient to directly search in the results. We use the recall calculation formula, where the numerator represents the number of keywords that appear in the results, and the denominator represents the total number of keywords.

In addition, we utilize an LLM (i.e., GPT-4) as the evaluator with the prompts shown in the arXiv version. We request the LLM to evaluate and score the model-generated target across faithfulness (S-Faithful) and formality (S-Formal). Scores range from 1 (worst) to 10 (best) for each criterion. S-Faithful reflects content preservation to retain original meaning, and S-Formal reflects the degree of formality in style. For chunk-level results, we require the outputs to align each paragraph by the paragraph index, so we evaluate them at the paragraph-level granularity. To fairly compare paragraph-level and chunk-level results, we evaluate them both at paragraph and chunk levels. To enhance effectiveness, we sample a maximum of 100 paragraphs per document for evaluation. Furthermore, we sample over 200 paragraphs, whose results of different LLMs are ranked by human evaluators.

Results

Results of Different LLMs

As shown in Table 2, we compare GPT-4 and QWen-Max (closed-source LLMs) with LLaMA3-8B and QWen-14B (open-source LLMs) at the chunk-level granularity. GPT-4 outperforms all other models across all metrics, followed by QWen-Max. Due to the limitations in model size, the performance of open-source models isn't as good. Among them, the LLaMA3-8B model exhibits superior performance compared to the QWen-14B model.

LLMs show potential in solving this complex task, yet still encounter "faithfulness problems". The performance of LLMs on the faithfulness score is unsatisfactory, even for state-of-the-art models like GPT-4 (7.32). We observe some "paragraph drift" situations at the chunk level. For example, the end of one paragraph might be migrated to the beginning of the following one. Even worse, the target and source cannot be aligned by paragraph index. Furthermore, faithfulness problems occur even when alignments are correct. The model frequently prioritizes coherence and fluency over faithfulness, such as fabrications for fragments or the omission of disorganized insertions in spoken scenarios, which compromises the accuracy of some essential information.

Results of Granularity Levels and Contexts

As presented in Table 3, we compare the performance at different levels of granularity based on the GPT-4 model.

Although the SWAB dataset is constructed at the document level, the CoS2W task at the **document level still presents a challenge**. It requires a considerable number of tokens for both input and output (about 8K), which is difficult even for GPT-4 as shown in Table 4. For document-level results, 76.10% of the paragraphs are blank without the corresponding index, compared to 1.60% at the chunk-level granularity. And there is a noticeable decline in all metrics.

Due to "paragraph drift" issues at the chunk level as mentioned before, we present LLM evaluation results (i.e., S-Faithful and S-Formal) for both paragraph and chunk levels to ensure a fair comparison. The chunk-level evaluations show that the CoS2W results perform better at chunk-level (7.57 and 7.81) than at paragraph-level (6.60 and 7.55) granularity. This highlights the **advantages of utilizing local context at the chunk level**. Chunk-level granularity enables LLMs to naturally leverage the local contexts to enhance faithfulness and formality performance.

In addition, we experiment with various contexts at paragraph level explicitly, **confirming the role of context in enhancing text understanding**, as is shown in Table 3. Paragraph-level results with local context show improvements across all metrics compared to results without context, approaching chunk-level performance. This further verifies the supportive role of local context. Local context improves CK-Recall from 19.20% to 22.21%, assisting the model to better understand the current paragraph. Moreover, global context can also yield some benefits, while not as significant as local context. Combining both local and global contexts, further improvements are seen in some metrics.

Results of Auxiliary Information

We employ multiple methods to leverage auxiliary information. Table 2 includes results of directly providing (Origin), retrieving the top-10 most relevant sentences (RAG), based on a summary (Summary), or list of keywords (Keywords) derived from LLMs.

Auxiliary information can enhance ASR error correction, while requiring more effective methods. Based on the GPT-4 model at chunk-level granularity, directly given auxiliary information (Origin) increases CK-Recall from 22.49% to 43.70%. The improvement of RAG is limited, indicating a need for better retrieval methods. The Summary method improves objective metrics but shows limited CK-Recall gains, possibly due to missing keywords in the summary. The Keywords method is most effective, further boosting recall to 45.42%, and enhancing faithfulness scores while maintaining formality scores basically. However, improvements in CK-Recall stay below expectations, suggesting the limited benefits of current methods. Effectively utilizing auxiliary information still needs further research.

Analysis

Instruction Following Capability

We observe that **larger LLMs (e.g., GPT-4) show stronger instruction-following ability** compared to smaller LLMs (e.g., QWen-14B). We instruct the models to organize their responses by aligning each paragraph with the corresponding content using the paragraph index. We refer to paragraphs without aligned paragraph indices as *blank paragraphs*, which do not have corresponding target results. The rate of blank paragraphs can reflect a model's instruction-following capability. The fewer the blank paragraphs, the stronger the instruction-following ability. Thus we compute the rate of blank paragraphs as shown in Table 4. At the chunk-level granularity, GPT-4 demonstrates robust alignment with scarcely any blank paragraphs (1.60%), whereas QWen-14B exhibits more alignment errors (36.56%). Given the auxiliary information, as the number of input tokens increases, QWen-14B's performance further declines to 48.33%, while GPT-4 remains virtually unaffected.

Model	#Tokens	BLEURT↑	Blank%↓
GPT-4 (paragraph)	340.50	63.32	0.00
w/ local context	654.47	67.11	0.00
GPT-4 (chunk)	1515.00	62.15	1.60
w/ auxiliary	2006.86	62.73	1.28
GPT-4 (document)	8140.42	10.64	76.10
QWen-14B (paragraph)	340.50	59.87	0.00
w/ local context	654.47	55.64	0.00
QWen-14B (chunk)	1515.00	30.14	36.56
w/ auxiliary	2006.86	22.35	48.33
QWen-14B (document)	8140.42	2.87	91.58

Table 4: The comparison of instruction-following abilities between GPT-4 and QWen-14B. The fewer the blank paragraphs, the stronger the instruction-following ability.

Model	S-Faithful↑		S-Formal↑	
	Total	Non-Blank	Total	Non-Blank
GPT-4	**7.32**	**7.39**	**8.20**	**8.29**
QWen-Max	6.00	6.50	7.55	8.21
QWen-14B	3.72	4.99	6.04	**8.29**
LLaMA-8B	6.35	6.53	7.33	7.56

Table 5: Comparison of different LLMs among total paragraphs with non-blank paragraphs at the chunk-level.

The differences in instruction-following ability also impact the model's performance on the CoS2W task. As shown in Table 5, we compare the results among total paragraphs with non-blank paragraphs. Among non-blank paragraphs, QWen-14B's S-Formal performance is actually quite good (8.29). However, the overall performance (6.04) is relatively low due to the influence of blank paragraphs.

Discussion on LLM Evaluation

To measure the reliability of LLM evaluation, we compare the results with human evaluation. Consequently, we find that **the LLM Evaluator is reliable for both faithfulness and formality evaluation.** Based on more than 200 sampled paragraphs, we engage three annotators to independently evaluate and rank the performance of GPT-4, QWen-Max, and QWen-14B in terms of faithfulness and formality. Then

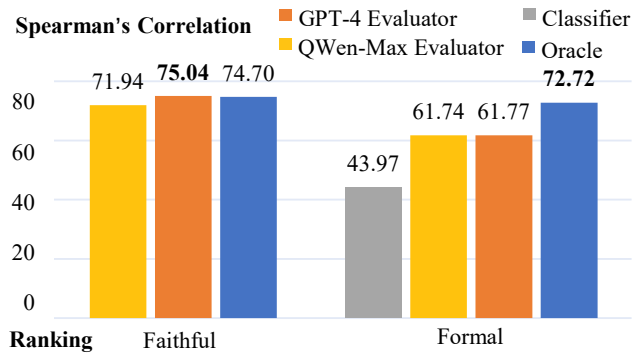

Figure 3: The ranking correlation coefficients between different evaluation methods and human evaluation. Note that "Oracle" shows the average scores of three annotators, while "Classifier" reports only on English paragraphs.

we calculate the Spearman's correlation coefficients (Spearman 1904) between the rankings given by the human evaluation and the LLM evaluation, as shown in Figure 3.

Faithfulness Score: The Spearman's correlation coefficient of the faithfulness rankings between the LLM Evaluator and human evaluation is 0.75, which is comparable to human performance. This demonstrates the reliability of the LLM Evaluator in faithfulness scoring. Additionally, this indicates that the LLM performs well in determining whether the results faithfully preserve the original content.

Formality Score: As shown in Figure 3, the Spearman's correlation coefficient between the LLM evaluation and human evaluation is 0.61. Although there is a gap compared to human performance, there is also a strong correlation, which validates that the formality scores are reliable. Moreover, we compare the commonly used evaluation method. Using a RoBERTa-based classifier[9], we classify English paragraphs as formal or informal and assign the formality probability as the formality score. The LLM evaluator outperforms the classifier in the correlation coefficient, confirming the reliability of LLM evaluation.

As Figure 3 illustrates, GPT-4 Evaluator correlates more strongly with human evaluations than QWen-Max Evaluator, confirming its reliability. In addition, we evaluate GPT-4 results three times. The average standard deviations of faithfulness scores (0.19) and formality scores (0.11) across all samples confirm the robustness. Despite LLM updates rapidly, stable evaluation results can still be achieved with the same model version at a low temperature.

Distribution of Languages and Domains

As indicated in Table 6, LLMs generally excel more in Chinese than English at the chunk level. There are more blank paragraphs in English (2.68% for GPT-4) than in Chinese (0.83%). This may be because LLMs prioritize fluency in English over Chinese, often merging paragraphs for better fluency at the expense of faithfulness.

As shown in Table 7, **the podcast domain, with the longest auxiliary information, shows the most significant improvement in CK-Recall**, especially with the Keywords method. This indicates that plentiful auxiliary information aids ASR error correction improvement. Optimizing the use of long-text auxiliary information for further improvement is an area that warrants deeper exploration.

[9]https://www.huggingface.co/s-nlp/roberta-base-formality-ranker

Model	BLEURT↑		S-Faithful↑	
	Chinese	English	Chinese	English
GPT-4	**66.95**	**55.52**	**7.83**	**6.80**
QWen-Max	57.67	40.79	6.99	5.00
QWen-14B	36.73	21.05	4.77	2.63
LLaMA-8B	36.55	46.03	7.33	5.36

Table 6: The performance across different languages.

Model	CK-Recall↑		
	Podcast	Meeting	Lecture
GPT-4 (chunk)	19.82	30.00	32.81
w/ aux. (Origin)	45.82	30.00	36.72
w/ aux. (RAG)	36.55	35.00	**39.84**
w/ aux. (Summary)	41.27	35.00	37.50
w/ aux. (Keywords)	**47.27**	**45.00**	37.50

Table 7: The CK-Recall performance of GPT-4 with various auxiliary information methods across different domains.

Conclusion

To improve the readability of ASR transcripts, we propose the Contextualized Spoken-to-Written conversion (CoS2W) task, construct and make available the document-level and multi-domain Spoken2Written of ASR transcripts Benchmark (SWAB) dataset. Based on the SWAB dataset, we compare the performance of different LLMs at various granularity levels, verify the beneficial roles of contexts and auxiliary information, and find that it is worth further exploring how LLMs maintain faithfulness and utilize auxiliary information to enhance ASR Error Correction. Compared with human evaluation, we find that the LLM Evaluator performs well in faithfulness and formality ranking with a good correlation. In the future, we plan to expand the dataset scale to better support research on supervised training for this task.

Limitations

The study's limitation lies in the SWAB dataset having only a single target, which may inadequately reflect model performance through many automatic metrics. Future efforts will focus on offering diverse targets and developing more effective evaluation methods. Note that the source data of SWAB are owned by the copyright holder. We only provide ASR transcripts and annotation targets, along with the corresponding links of audios and videos. The license of SWAB will be for research purposes only. Additionally, it is inevitable that evaluations are influenced by the biases of GPT-4, given that we obtain initial results through GPT-4 and also use it for experiments and evaluation.

Ethics Statement

The SWAB dataset used in this research is strictly for academic and non-commercial purposes. We implemented several measures to ensure compliance with ethical standards, as follows.

- **Data Transparency and Anonymization.** The Chinese meetings in the SWAB are sourced from the training set of the publicly available AliMeeting dataset, and the English meetings come from the publicly available AMI meeting dataset. For our collected podcasts and lectures, we only provide ASR transcripts after rigorous text anonymization processes and our annotations, to ensure transparency regarding the data sources and their usage while maintaining anonymity.
- **Data Access Compliance.** To further ensure the ethical use of the dataset, we require researchers to contact us via email to confirm their compliance with ethical guidelines and the conditions outlined in our data usage declaration, before granting them access to the dataset. This procedure includes ensuring that they are aware of and adhere to the Personal Information Protection Law (PIPL) and any relevant legal frameworks regarding personal data usage.
- **Authorization.** Any personal data should be used only with express authorization, ensuring lawful and fair processing in accordance with applicable laws.

Acknowledgments

We would like to express gratitude to the anonymous reviewers for their valuable feedback.

References

Bryant, C.; Felice, M.; Andersen, Ø. E.; and Briscoe, T. 2019. The BEA-2019 Shared Task on Grammatical Error Correction. In Yannakoudakis, H.; Kochmar, E.; Leacock, C.; Madnani, N.; Pilán, I.; and Zesch, T., eds., *BEA@ACL 2019*, 52–75. Association for Computational Linguistics.

Carletta, J.; Ashby, S.; Bourban, S.; Flynn, M.; Guillemot, M.; Hain, T.; Kadlec, J.; Karaiskos, V.; Kraaij, W.; Kronenthal, M.; et al. 2005. The AMI meeting corpus: A pre-announcement. In *International workshop on machine learning for multimodal interaction*, 28–39. Springer.

Daudaravicius, V.; Banchs, R. E.; Volodina, E.; and Napoles, C. 2016. A Report on the Automatic Evaluation of Scientific Writing Shared Task. In Tetreault, J. R.; Burstein, J.; Leacock, C.; and Yannakoudakis, H., eds., *BEA@NAACL-HLT 2016*, 53–62. The Association for Computer Linguistics.

Fang, T.; Yang, S.; Lan, K.; Wong, D. F.; Hu, J.; Chao, L. S.; and Zhang, Y. 2023. Is ChatGPT a Highly Fluent Grammatical Error Correction System? A Comprehensive Evaluation. *CoRR*, abs/2304.01746.

Gao, Z.; Zhang, S.; McLoughlin, I.; and Yan, Z. 2022. Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition. *arXiv preprint arXiv:2206.08317*.

Guo, Z.; Yu, L.; Xu, M.; Jin, R.; and Xiong, D. 2023. CS2W: A Chinese Spoken-to-Written Style Conversion Dataset with Multiple Conversion Types. In Bouamor, H.; Pino, J.; and Bali, K., eds., *EMNLP 2023*, 3962–3979. Association for Computational Linguistics.

Gupta, A.; Xu, J.; Upadhyay, S.; Yang, D.; and Faruqui, M. 2021. Disfl-QA: A benchmark dataset for understanding disfluencies in question answering. *arXiv preprint arXiv:2106.04016*.

Han, R.; Soldaini, L.; and Moschitti, A. 2021. Modeling Context in Answer Sentence Selection Systems on a Latency Budget. In *EACL 2021*, 3005–3010.

He, J.; Yang, Z.; and Toda, T. 2023. ed-cec: improving rare word recognition using asr postprocessing based on error detection and context-aware error correction. In *ASRU 2023*, 1–6. IEEE.

Ihori, M.; Takashima, A.; and Masumura, R. 2020. Parallel Corpus for Japanese Spoken-to-Written Style Conversion. In Calzolari, N.; Béchet, F.; Blache, P.; Choukri, K.; Cieri, C.; Declerck, T.; Goggi, S.; Isahara, H.; Maegaard, B.; Mariani, J.; Mazo, H.; Moreno, A.; Odijk, J.; and Piperidis, S., eds., *LREC 2020*, 6346–6353. European Language Resources Association.

Lai, H.; Toral, A.; and Nissim, M. 2023. Multidimensional Evaluation for Text Style Transfer Using ChatGPT. *CoRR*, abs/2304.13462.

Lakomkin, E.; Wu, C.; Fathullah, Y.; Kalinli, O.; Seltzer, M. L.; and Fuegen, C. 2024. End-to-end speech recognition contextualization with large language models. In *ICASSP 2024*, 12406–12410. IEEE.

Li, Y.; Huang, H.; Ma, S.; Jiang, Y.; Li, Y.; Zhou, F.; Zheng, H.-T.; and Zhou, Q. 2023. On the (in) effectiveness of large language models for chinese text correction. *arXiv preprint arXiv:2307.09007*.

Li, Y.; Qin, S.; Ye, J.; Ma, S.; Li, Y.; Qin, L.; Hu, X.; Jiang, W.; Zheng, H.; and Yu, P. S. 2024. Rethinking the Roles of Large Language Models in Chinese Grammatical Error Correction. *CoRR*, abs/2402.11420.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Ma, R.; Qian, M.; Manakul, P.; Gales, M.; and Knill, K. 2023. Can generative large language models perform asr error correction? *arXiv preprint arXiv:2307.04172*.

Mai, L.; and Carson-Berndsen, J. 2023. Enhancing conversational quality in language learning chatbots: An evaluation of GPT4 for ASR error correction. *arXiv e-prints*, arXiv–2307.

Min, Z.; and Wang, J. 2023. Exploring the integration of large language models into automatic speech recognition systems: An empirical study. In *ICONIP*, 69–84. Springer.

Napoles, C.; Sakaguchi, K.; and Tetreault, J. R. 2017. JFLEG: A Fluency Corpus and Benchmark for Grammatical Error Correction. In Lapata, M.; Blunsom, P.; and Koller, A., eds., *EACL 2017*, 229–234. Association for Computational Linguistics.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *ACL 2002*, 311–318. ACL.

Rao, G.; Yang, E.; and Zhang, B. 2020. Overview of NLPTEA-2020 shared task for Chinese grammatical error diagnosis. In *NLPTEA 2020*, 25–35.

Reif, E.; Ippolito, D.; Yuan, A.; Coenen, A.; Callison-Burch, C.; and Wei, J. 2022. A Recipe for Arbitrary Text Style Transfer with Large Language Models. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *ACL 2022*, 837–848. Association for Computational Linguistics.

Robertson, S. E.; and Zaragoza, H. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.*, 3(4): 333–389.

Sellam, T.; Das, D.; and Parikh, A. P. 2020. BLEURT: Learning Robust Metrics for Text Generation. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J. R., eds., *ACL 2020*, 7881–7892. Association for Computational Linguistics.

Spearman, C. 1904. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, 15(1): 72–101.

Tao, Z.; Xi, D.; Li, Z.; Tang, L.; and Xu, W. 2024. CAT-LLM: Prompting Large Language Models with Text Style Definition for Chinese Article-style Transfer. *CoRR*, abs/2401.05707.

Wang, Y.; Kong, C.; Yang, L.; Wang, Y.; Lu, X.; Hu, R.; He, S.; Liu, Z.; Chen, Y.; Yang, E.; and Sun, M. 2021. YACLC: A Chinese Learner Corpus with Multidimensional Annotation. *CoRR*, abs/2112.15043.

Yang, C.-H. H.; Gu, Y.; Liu, Y.-C.; Ghosh, S.; Bulyko, I.; and Stolcke, A. 2023. Generative speech recognition error correction with large language models and task-activating prompting. In *ASRU 2023*, 1–8. IEEE.

Yannakoudakis, H.; Briscoe, T.; and Medlock, B. 2011. A New Dataset and Method for Automatically Grading ESOL Texts. In Lin, D.; Matsumoto, Y.; and Mihalcea, R., eds., *ACL-HLT 2011*, 180–189. The Association for Computer Linguistics.

Yu, F.; Zhang, S.; Fu, Y.; Xie, L.; Zheng, S.; Du, Z.; Huang, W.; Guo, P.; Yan, Z.; Ma, B.; Xu, X.; and Bu, H. 2022. M2MeT: The ICASSP 2022 Multi-Channel Multi-Party Meeting Transcription Challenge. In *Proc. ICASSP*. IEEE.

Zhang, Q.; Chen, Q.; Li, Y.; Liu, J.; and Wang, W. 2021a. Sequence model with self-adaptive sliding window for efficient spoken document segmentation. In *ASRU 2021*, 411–418. IEEE.

Zhang, R.; Wang, X.; Zhang, C.; He, Z.; Wu, H.; Li, Z.; Wang, H.; Chen, Y.; and Li, Q. 2021b. Bstc: A large-scale chinese-english speech translation dataset. *arXiv preprint arXiv:2104.03575*.

Zhang, X.; Zhang, X.; Yang, C.; Yan, H.; and Qiu, X. 2023. Does Correction Remain A Problem For Large Language Models? *CoRR*, abs/2308.01776.

Zhang, Y.; Li, Z.; Bao, Z.; Li, J.; Zhang, B.; Li, C.; Huang, F.; and Zhang, M. 2022. MuCGEC: a Multi-Reference Multi-Source Evaluation Dataset for Chinese Grammatical Error Correction. In Carpuat, M.; de Marneffe, M.; and Ruíz, I. V. M., eds., *NAACL-HLT 2022*, 3118–3130. Association for Computational Linguistics.

Zhao, Y.; Jiang, N.; Sun, W.; and Wan, X. 2018. Overview of the NLPCC 2018 Shared Task: Grammatical Error Correction. In Zhang, M.; Ng, V.; Zhao, D.; Li, S.; and Zan, H., eds., *NLPCC 2018*, volume 11109 of *Lecture Notes in Computer Science*, 439–445. Springer.

Zheng, S.; Cheng, L.; Chen, Y.; Wang, H.; and Chen, Q. 2023. 3D-Speaker: A Large-Scale Multi-Device, Multi-Distance, and Multi-Dialect Corpus for Speech Representation Disentanglement. *arXiv preprint arXiv:2306.15354*.