Tell Me Your Story: Evaluating Life Writing Generation from Older Adults' Spoken Narratives

Anonymous ACL submission

Abstract

While life writing is essential for preserving memory, its reliance on literacy creates a barrier for many older adults. Our study addresses this by exploring spoken-to-life writing, which transforms oral narratives into written documentation. Although LLMs facilitate automated generation, progress in this field has been hindered by inadequate evaluation methods. To bridge this gap, we introduce an evaluation framework grounded in narrative theory, measuring both Faithfulness (accuracy to the source) and Situatedness (relevance to the audience). Utilizing a benchmark of 402 annotated oral narratives, our study reveals that current LLMs fail to grasp the nuances of spoken discourse and struggle with the discursive gap between older adults and Gen Z, underscoring the need for targeted future research.

1 Introduction

"We Die Twice: First, When We Cease To Be; Second, When We Are Forgotten."
— Coco

Life writing provides a narrative account of specific life facets, from personal experiences to significant milestones (Howes, 2020; Saunders, 2008; Olney, 1998). As a repository for collective memory, it preserves the socio-cultural fabric of individual lives (de Medeiros et al., 2007). Yet, because traditional documentation relies on formal literacy, it poses a structural barrier for many older adults who, despite their rich storytelling abilities, have limited educational backgrounds (National Bureau of Statistics of China, 2021; Thompson, 2017; Bonnin, 2009). This reliance risks marginalizing their voices and losing invaluable social histories (Arthur, 2009).

While the advancement of LLMs(Ouyang et al., 2022; Yang et al., 2023; Wu et al., 2025) offers unprecedented opportunities to automate the generation of "spoken-to-life-writing", effectively removing traditional literacy barriers, it remains particularly unclear whether current LLMs can reliably reconstruct fragmented, colloquial oral accounts into high-fidelity, readable life writings. The absence of such benchmarks hinders the ability to diagnose model failure modes or iteratively refine modeling strategies for automated life writing.

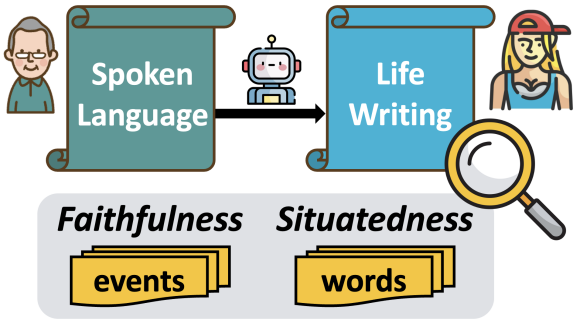

Figure 1: Spoken-to-life-writing task and evaluation framework. Oral narratives from older adults are adapted for a Generation Z audience. The evaluation framework assesses quality via Faithfulness (event-level fidelity) and Situatedness(adaptation of culturally specific terms for the target audience).

Assessing automated life writing is uniquely challenging. Unlike typical open-ended generation, life writing is a fact-constrained and contextually situated genre, as grounded in narrative theory(Piper et al., 2021). *Faithfulness* assesses whether the generated life story remains faithful to the source oral narrative. Meanwhile, life writing is inherently audience-dependent, as its effectiveness depends on whether narratives can be meaningfully interpreted by intended readers. We instantiate *Situatedness* with Generation Z (Gen Z, born between 1997 and 2012 (Dimock, 2019)) as the target audience, motivated by intergenerational memory transmission and the pronounced cultural gap(Giles and Gasiorek, 2011) that makes this setting a stringent stress test for narrative

accessibility (Dolot, 2018; De Medeiros, 2014; Arthur, 2009; Giles and Gasiorek, 2011).

Prior evaluation for narrative generation primarily employs reference-based metrics and the LLM-as-a-Judge paradigm, yet neither effectively generalizes to life writing. Reference-based metrics (Papineni et al., 2002; Lin, 2004; Sellam et al., 2020) are ill-suited for this genre, which allows for vast expressive variation and lacks the "golden label". While LLM-as-a-Judge approaches (Kim et al., 2023; Liu et al., 2023) have gained traction, their reliance on holistic subjective scoring fails to rigorously verify the preservation of granular narrative details (Wu et al., 2025). Furthermore, existing protocols remain predominantly text-centric, focusing on internal attributes such as fluency (Yang et al., 2023) or creativity (Bae and Kim, 2024), while overlooking situatedness, an audience-centered dimension critical for ensuring narrative accessibility to target readers.

To address these limitations, we introduce an automated evaluation framework that integrates Faithfulness and Situatedness. To quantify faithfulness, we develop an event-based comparison mechanism that measures the fidelity of generated life writing against source transcripts. This involves annotating discrete events and their attributes within oral narratives of older adults, which are then converted into a suite of template-based Boolean questions. The faithfulness is assessed by an evaluator's ability to accurately answer these questions using only the generated life writing. To operationalize situatedness, we shift the focus to audience-centered adaptation. Recognizing that intergenerational comprehension barriers primarily arise from cultural specific words, our framework evaluates whether the model proactively clarifies terminology that may be obscure to younger readers.

We construct a dataset comprising 402 oral life narratives collected from older adults, accompanied by fine-grained ground-truth event and vocabulary annotations. We evaluated the life-writing capabilities of various open-source and closed-source models, focusing on two distinct approaches: transcript-centric transformation and structured data-guided generation. Our findings indicate that LLMs lack the capacity to capture nuanced details from spoken language and identify Cultural Specific Items (CSIs) that bridge the expression gap between older adults and GenZ, underscoring a need for targeted future studies. Recognizing that the primary bottleneck lies in processing the oral discourse of seniors, we propose a multi-round QA strategy for information extraction to mitigate information distortion.

The main contributions of this paper are summarized as follows:

- We formally introduce the task of *spoken-to-life-writing* generation, evaluating it as a narrative generation problem that addresses literacy barriers in elderly memory preservation.
- We release the first benchmark for this task, together with a narrative-theory-grounded evaluation framework that jointly measures faithfulness and intergenerational situatedness.
- We conduct a systematic evaluation of state-of-the-art LLMs on spoken-to-life-writing, revealing their limitations and outlining directions for improving generation quality.

2 Related Work

2.1 Narrative Generation

Narrative generation has been investigated in both HCI and NLP, yet life writing remains underaddressed. While HCI research, such as StorySage (Talaei et al., 2025), facilitates memory elicitation for older adults, it emphasizes interaction design over the systematic quality control of generated texts. Within NLP, frameworks like DOC (Yang et al., 2023) and Re^3 (Yang et al., 2022) focus on fictional storytelling, prioritizing artistic creativity and long-range coherence. In contrast, life writing demands factual fidelity and communicability, rendering fictional evaluation protocols insufficient. While we also consider spoken-to-written task, existing approaches remain misaligned with life writing requirements. Sentence-level rewriting (Guo et al., 2023; Kang et al., 2025) focuses on eliminating disfluencies but lacks the discourse-level restructuring necessary for global narrative flow. Conversely, document-level summarization (Chen et al., 2021) prioritizes information compression, often at the expense of the rich narrative detail and personal voice essential to biography.

2.2 Evaluation methodologies

The evaluation landscape for generative writing has undergone several significant paradigm shifts. Early evaluation methodologies primarily relied

on n-gram overlap metrics(Papineni et al., 2002; Lin, 2004). To address their inability to capture semantic nuances, model-driven evaluation metrics (Zhang et al., 2020; Sellam et al., 2020; Yuan et al., 2021)that leverage deep contextual embeddings have been developed to measure semantic similarity beyond literal word matching. Despite their improved correlation with human judgment, these reference-based metrics necessitate "golden labels", which are often unavailable or insufficient for narrative generation.

With the advent of LLMs, the *LLM-as-a-Judge* paradigm has emerged as a predominant framework (Kim et al., 2023; Liu et al., 2023). In creative writing, state-of-the-art benchmarks (Wu et al., 2025; Gómez-Rodríguez and Williams, 2023) demonstrate that LLMs can assess multi-dimensional quality, such as coherence and vividness, achieving high alignment with human experts through direct scoring. However, this paradigm is less applicable to life writing. Unlike fictional narratives, life writing is a fact-constrained medium where relying on holistic, model-based scores often lacks the granularity required to detect subtle factual inconsistencies. Instead of direct scoring, we leverage LLMs to answer factual questions derived from the source transcripts; the narrative's quality is then quantified by the accuracy of these responses. This approach provides a fine-grained assessment of factual fidelity and situatedness, ensuring the generated narrative is both accurate and interpretable.

3 Life Writing

Life writing, as a specialized genre of narrative, requires an evaluation framework that transcends basic linguistic fluency. Drawing upon narrative theory(Piper et al., 2021), we combine between *Classical Narrative Theory*(Genette, 1980), which focuses on internal narrative features (the "what"), and *Post-classical Narrative Theory*(Prince, 2019), which emphasizes the relationship between the text and its audience (the "to whom"). Our evaluation framework decomposes the generated life writing into two distinct dimensions: (1) Faithfulness (the internal narrative structure), which assesses the fidelity of the generated text to the original spoken language of older adults; and (2) Situatedness (the external communicative context), which evaluates the narrative's appropriateness for audience delivery. This dual requirement ensures that historical truth is preserved while communicative barriers are removed.

3.1 Task Formulation

Faithfulness: Narrative Elements. The factual integrity of life writing is defined by the structural consistency of its underlying narrative elements. Drawing upon the narrative schema (Piper et al., 2021), an original narrative N_{orig} is modeled as a structured collection of discrete events $\mathcal{E} = \{E_1, E_2, \ldots, E_n\}$. Each individual event $E \in \mathcal{E}$ is represented as a five-tuple of extracted semantic properties:

$$E = (S, A, L, T, P) \quad (1)$$

where S, A, L, T, and P represent the *Subject* (the primary agent), *Action* (the specific event type), *Location* (the spatial context), *Time* (the temporal coordinates), and *Peripheral properties* (encompassing causal motivations or additional semantic participants), respectively.

Situatedness: Cultural Gap Words. While factual integrity preserves the objective content of a life narrative, the transformation from spoken language to written life writing necessitates situatedness, which is the optimization of delivery for communicative efficacy and memory transmission.

Following the narrative framework (Piper et al., 2021), we characterize situatedness by focusing on the narrative "Receiver", specifically identifying this role as the Gen Z. This choice is motivated by two factors: 1). Gen Z represents the grandchild generation of the older adults, making them the most direct and salient audience for the preservation of life narratives; 2). The generational cohort gap between the silent generation and Gen Z presents the most significant challenge for mutual understanding.

Our empirical analysis suggests that intergenerational friction is not systemic (e.g., syntactic or logical) but is primarily localized within the lexicon. We categorize these lexical obstacles into two distinct types: 1). Semantic Equivalence Items(SEIs): Concepts that exist in both cultures but differ in linguistic realization. 2). Culture-Specific Items(CSIs): Diachronic concepts tied to specific historical-political contexts that are absent from the contemporary Gen Z cultural schema.

We formulate the delivery of these two lexical categories as a translation task, adopting a framework inspired by translation strategies: 1) Domes-

Table 1: Classification of Cultural Gap Words

Category	Source & Transformed Comparison
SEIs	**Source:** 我年纪大了，也**抽抽**了。(I am getting older, and I have **shrunk**.) **Transformed:** 我年龄大了，也**变矮**了。(I am getting older and have **become shorter**.)
CSIs	**Source:** 那个时候都在**生产队**里。(At that time, everyone was in the **production team**.) **Transformed:** 那个时候都在**生产队**(1) 里。 (1) 生产队（1958-1984）：中国农村人民公社时期最基层的农业生产与行政单位。 At that time, everyone was in the **production team**(1). (1) Production Team (1958-1984): The basic agricultural and administrative unit in rural China during the People's Commune era.

tication Strategy(Nida, 1964): For SEIs, we prioritize reading flow by substituting source terms with contemporary equivalents or employing in-text glossing to eliminate comprehension barriers. 2) Foreignization Strategy(Venuti, 2017): For CSIs, we retain the original terminology and use annotations (i.e., footnotes) to preserve the narrative's socio-historical significance, avoiding excessive in-text paraphrasing that would disrupt the narrative rhythm, as illustrated in Table 1.

Consequently, we evaluate situatedness based on the precision and strategic appropriateness of these lexical transformations.

3.2 Data Construction

Data Sources We collected 402 spoken transcriptions of interviewer-older adult dialogues from Bilibili[1], a leading long-form video platform in China. This corpus captures the complexity of spoken language as older adults recount their life experiences. Detailed demographic profiles is provided in Appendix A.1.

Data Annotation Each entry in the dataset is formalized as a structured triplet:

$$\mathcal{D} = \langle T_{spoken}, L_{event}, C_{gap} \rangle \quad (2)$$

where: T_{spoken} represents the raw spoken transcription. L_{event} denotes a structured event list mapped to the narrative elements defined in Section 3.1. C_{gap} contains cultural gap words, including identified target words, their classification (i.e., SEIs or CSIs), and their respective contextual meanings. A data sample is provided in Appendix A.2.

To balance annotation cost and quality, we adopted an LLM-in-the-loop pipeline. The process began with preliminary event extractions and word classifications generated by the DeepSeek-V3.2(DeepSeek-AI, 2025)(hereafter DeepSeek). Detailed prompts are provided in Appendix A.3. To ensure annotation quality and consistency, we conducted a rigorous manual verification phase following the model's output generation. A team of 15 annotators meticulously validated and refined the results under the supervision of researchers. The annotation process are detailed in Appendix A.4.

3.3 Automatic Evaluation Metrics

We define formal metrics to assess the dual objectives of narrative adaptation: factual integrity and communicative efficacy for Gen Z.

3.3.1 Faithfulness Metrics

We propose a Reading Comprehension framework to evaluate factual consistency in life writing. Unlike ROUGE, which misses semantic equivalence, or BLEURT, which is optimized for sentences rather than phrasal event attributes, our framework leverages LLMs' near-human comprehension(OpenAI, 2024; DeepSeek-AI, 2025). We formalize evaluation by converting annotated events into a set of boolean questions $\mathcal{Q}$.

For each spoken transcription, let $G(q)$ be the ground truth answer derived from our annotations, and $M(q)$ be the answer generated by an LLM evaluator when asked the generated narrative and question q. We design two types of question template is provided in Appendix B.1.

Based on this framework, we define three hierarchical metrics to measure Faithfulness:

Event Type Accuracy (Acc_{type}): This metric measures the model's ability to correctly identify the occurrence of specific event types:

$$\text{Acc}_{\text{type}} = \frac{1}{|\mathcal{Q}_{\text{type}}|} \sum_{q \in \mathcal{Q}_{\text{type}}} \mathbb{1}(M(q) = G(q)) \quad (3)$$

where $\mathcal{Q}_{\text{type}}$ denotes the set of existence questions and $\mathbb{1}(\cdot)$ is the indicator function.

Event Property Accuracy (Acc_{prop}): This metric evaluates the overall correctness of all event at-

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

tributes mentioned in the narrative:

$$\text{Acc}_{\text{prop}} = \frac{1}{|\mathcal{Q}_{\text{prop}}|} \sum_{q \in \mathcal{Q}_{\text{prop}}} \mathbb{1}(M(q) = G(q)) \quad (4)$$

where $\mathcal{Q}_{\text{prop}}$ represents the set of all attribute-related questions.

Mean Event Accuracy (MEA): While Acc_{type} and Acc_{prop} provides a global measure of faithfulness, it is susceptible to bias from events with a disproportionate number of properties. To ensure a balanced assessment of factual integrity at the event level, we define the Mean Event Accuracy (MEA) as:

$$\text{MEA} = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \left(\frac{1}{|\mathcal{Q}_e|} \sum_{q \in \mathcal{Q}_e} \mathbb{1}(M(q) = G(q)) \right) \quad (5)$$

where $\mathcal{E}$ denotes the set of events, and $\mathcal{Q}_e$ contains all boolean questions associated with event E. Additionally, if the predicted event type is incorrect, all associated properties are automatically treated as incorrect.

3.3.2 Situatedness Metrics

To evaluate the model's efficacy in bridging cultural gaps, we introduce *Situatedness Metrics*, which assess both the strategic selection and the semantic quality of cultural adaptations.

Strategy Classification Recall: Let W_c be the set of words that truly belong to category c, and $\hat{W}_c$ be the set of words predicted by the model as category c. The strategy classification recall for category $c \in \{\text{SEIs, CSIs}\}$ is defined as:

$$R_{strat}(c) = \frac{|W_c \cap \hat{W}_c|}{|W_c|} \quad (6)$$

where $|\cdot|$ denotes the cardinality of the set. This metric quantifies the model's ability to correctly identify cultural terms requiring specific strategies.

Explanatory Fidelity: For terms undergoing *foreignization*, the model must generate descriptive footnotes to provide necessary context. Since these explanations are typically sentential, we employ BLEURT to quantify the semantic similarity between the generated footnotes $\mathcal{F}$ and the gold-standard meanings G_w:

$$\text{Score}_{\text{sem}} = \frac{1}{|\mathcal{F}|} \sum_{F \in \mathcal{F}} \text{BLEURT}(F, G_w) \quad (7)$$

where a higher score indicates superior semantic alignment with the authoritative cultural interpretation.

4 Methods

We investigate two distinct paradigms for life writing generation: direct linguistic transformation and structured-data-driven generation.

4.1 Direct Spoken-to-Written Transformation

This paradigm focuses on converting raw oral transcripts into life writing by addressing the inherent disfluencies and fragmented nature of spoken language. We employ a zero-shot approach where the LLM is prompted to rewrite the spoken transcript into a life writing narrative while preserving original spoken language details. The specific prompt template used for this task is provided in Appendix C.1.1.

4.2 Structure-based Generation

To mitigate the semantic noise and structural ambiguity inherent in raw oral transcripts, we introduce a structure-driven paradigm that utilizes explicit event sequences and culture-specific keywords to ground the generation process:

Augmented CoT Generation: This approach employs a Chain-of-Thought (CoT) strategy that requires the LLM to internally extract structured events and cultural keywords prior to synthesis. Appendix C.1.2 details the corresponding prompt.

Multi-turn QA Pipeline: Using a Multi-turn QA strategy, this method systematically extracts events and attributes to guide structured life writing. For each transcription, the model identifies the occurrence of predefined event types, then fills in their specific attributes through iterative questioning, followed by the extraction of cultural keywords. Detailed prompt templates are available in Appendix C.1.3.

Human-in-the-loop Refinement: To assess the performance upper bound of structural grounding and reflect established biographical practices in real world, we incorporate human-verified event sequences and keywords into the generation workflow. These structural constraints are combined with raw oral transcripts to guide the LLM's synthesis of life narratives (see Appendix C.1.4).

Table 2: Experimental results across different generation methods and LLM backbones. All values are percentages (%) except BR_{exp}. Higher scores denote better results across all metrics. ALL: All annotations. HC: Human-Corrected annotations. BR_{exp}: explanation BLEURT score of CSIs. **bolded**: the best results. underline: the second best results.

Method	Model	Setting	Acc_{type}	Acc_{prop}	MEA	R_{CSIs}	BR_{exp}	R_{SEIs}
Spoken Lang. w/ Golden Labels	GPT	All	91.07	82.96	80.90	97.23	91.60	88.02
		HC	89.43	75.41	79.98	94.44	86.00	91.67
	Claude	All	87.12	78.91	75.61	96.13	90.03	92.07
		HC	84.82	70.03	73.62	97.78	89.10	87.50
	Qwen	All	93.13	82.73	81.40	99.17	93.66	51.05
		HC	92.32	78.63	82.48	100.00	91.45	62.50
Spoken Lang.	GPT	All	82.65	72.77	68.17	35.51	13.87	71.86
		HC	80.02	64.01	66.17	32.22	12.12	85.42
	Claude	All	85.27	76.22	72.62	43.81	16.88	57.04
		HC	82.61	66.23	70.19	35.56	10.43	76.04
	Qwen	All	85.90	74.28	71.03	**58.02**	**21.29**	75.45
		HC	84.82	67.12	70.37	**41.67**	**12.20**	90.62
Spoken Lang. w/ CoT	GPT	All	79.96	70.91	65.45	32.02	9.89	**79.19**
		HC	77.33	60.58	63.82	16.67	4.42	89.58
	Claude	All	81.43	72.85	68.10	28.71	10.40	69.31
		HC	80.02	65.86	67.07	14.44	4.32	87.50
	Qwen	All	85.86	73.22	69.39	21.13	8.21	73.95
		HC	85.01	66.39	69.66	13.89	3.95	92.71
Spoken Lang. w/ Multi-turn QA	GPT-4o	All	86.46	74.24	71.03	45.82	14.90	74.56
		HC	**86.31**	68.18	72.92	25.53	8.16	71.43
	Claude	All	**86.93**	**76.67**	**73.48**	48.61	19.05	68.64
		HC	84.79	**69.09**	**74.10**	19.15	6.68	78.57
	Qwen	All	85.20	72.78	70.11	48.61	17.59	72.19
		HC	82.51	65.00	70.16	17.02	3.36	**100.00**

5 Results

This section evaluates LLMs across four distinct methods for life writing generation. Our analysis reveals core challenges in faithfulness and situatedness, supported by both quantitative metrics and qualitative case studies.

5.1 Model Choice

We employ DeepSeek as the evaluator for the reading comprehension task used to assess faithfulness. To mitigate self-bias (Xu et al., 2024), we include other SOTA closed-source models for benchmarking, specifically GPT-4o (OpenAI, 2024) and Claude-3.7-Sonnet (hereafter GPT and Claude). Furthermore, to investigate models with varying parameter scales, we evaluate the open-source model Qwen3-30B-A3B-Instruct-2507 (Qwen, 2025) (hereafter Qwen). Hyperparameters are provided in Appendix C.2.

5.2 Main Results

The results are shown in Table 2. We find that: **In general, providing structured facts is essential for fixing factual inconsistencies in spoken-to-written tasks.** *1)* The superior performance of the "Golden" setting indicates that LLMs struggle to autonomously capture nuanced factual details and cultural gap words when generating life writing from spoken language. Given current LLM capabilities, human-in-the-loop intervention represents a high-potential strategy for life writing. *2)* Notably, models performed worse on the human-corrected subset than on the full annotation set regarding faithfulness and Culture-Specific Items (CSIs). This performance drop suggests that LLMs are insensitive to fine-grained factual refinements and often fail to identify which terms require background explanation for younger readers. *3)* Surprisingly, CoT (Chain-of-Thought) prompting underperformed compared to direct generation across multiple metrics, suggesting that in long-form oral transformation tasks, endogenous rea-

soning chains may introduce additional noise in the absence of external verification. *4)* Despite its smaller parameter scale, Qwen matched or exceeded the performance of larger models across nearly all indicators; this suggests that for Life Writing tasks rooted in specific Chinese historical contexts, the intensity of pre-training on native corpora is more critical than raw model scale.

Regarding faithfulness, the Multi-turn QA paradigm effectively secures factual anchors. In autonomous generation scenarios without human intervention, the Multi-turn QA paradigm yielded the best results. This demonstrates that using fine-grained information extraction as a preprocessing step effectively mitigates noise from oral narratives and provides models with structured factual anchors. These findings underscore the critical importance of high-precision factual extraction methods in life writing.

In terms of situatedness, LLMs excel at linguistic conversion but struggle with cross-generational cultural mediation. *1)* R_{SEIs} generally exceeded R_{CSIs}, suggesting that LLMs have attained adequate proficiency in semantic equivalence transformations, effectively bridging the linguistic gap between older adults and Gen Z through tasks such as dialect-to-standard language conversion. *2)* Exceptionally low $BLEURT_{exp}$ scores in non-Golden settings indicate that autonomous LLMs often fail to provide essential cultural footnotes. Such omissions underscore a deficiency in the situational awareness vital for cross-generational communication, marking a critical gap for future research.

5.3 Detailed Analysis

Validity of LLM-based Evaluation. To validate the reliability of using LLMs as evaluators for event information extraction in life writing, we conducted a verification experiment. We utilized different models to generate life writing narratives based solely on "golden" structured data (events and keywords). DeepSeek was then employed as an evaluator to answer boolean questions derived from this same structured data. As illustrated in Table 3, the models achieved nearly 100% accuracy in Acc_{type} and consistently scored around 90% in Acc_{prop} and MEA. These results demonstrate that LLMs can achieve near-lossless information extraction and accurately identify event details, confirming that using an LLM as a reliable evaluator for this task is both valid and effective. Notably, Qwen exhibits exceptional performance in life writing tasks, consistently matching or exceeding larger models across several metrics. This underscores the fact that given high-quality structured data, even relatively smaller models can achieve superior results with significantly lower computational overhead.

Table 3: Evaluation of DeepSeek on factual recognition via boolean question answering.

Model	Setting	Acc_{type}	Acc_{prop}	MEA
GPT	All	95.94	89.13	88.23
	HC	97.12	89.34	89.76
Claude	All	95.13	**89.61**	88.28
	HC	94.33	85.38	87.59
Qwen	All	**98.97**	89.32	**89.25**
	HC	**98.66**	**90.40**	**92.17**

Structured Data as a Prerequisite. To isolate the primary challenge in life-writing generation, we conducted a verification experiment where DeepSeek was tasked with answering boolean questions directly from raw oral transcriptions. As indicated in Table 4, the performance on raw transcriptions is suboptimal, suggesting that the inherent disfluencies and fragmented nature of elderly speech pose substantial hurdles for LLMs' comprehension.

We further tested whether converting these transcriptions into a written style using a SOTA sentence-level method (Kang et al., 2025) could bridge this gap. Surprisingly, accuracy declined further. This stems from the limitation of sentence-level processing in addressing discourse-level issues, such as anaphora resolution. For instance, when an elderly man said, "went to the countryside in the Northeast, and later her sister was in Gansu" (where "she" refers to the speaker' s wife), it was incorrectly transcribed as "I went to the countryside... and later my sister was in Gansu." Such errors lead to significant semantic distortion and misrepresent the original familial relationships. These findings confirm that structured data is not merely a convenience, but an essential prerequisite for high-fidelity life writing generation.

5.4 Case Study

To demonstrate how structured event data enhances life writing generation, we conduct a case study comparing two distinct strategy applied to the same spoken language transcription: Direct Strategy and Multi-turn QA Guided Strategy.

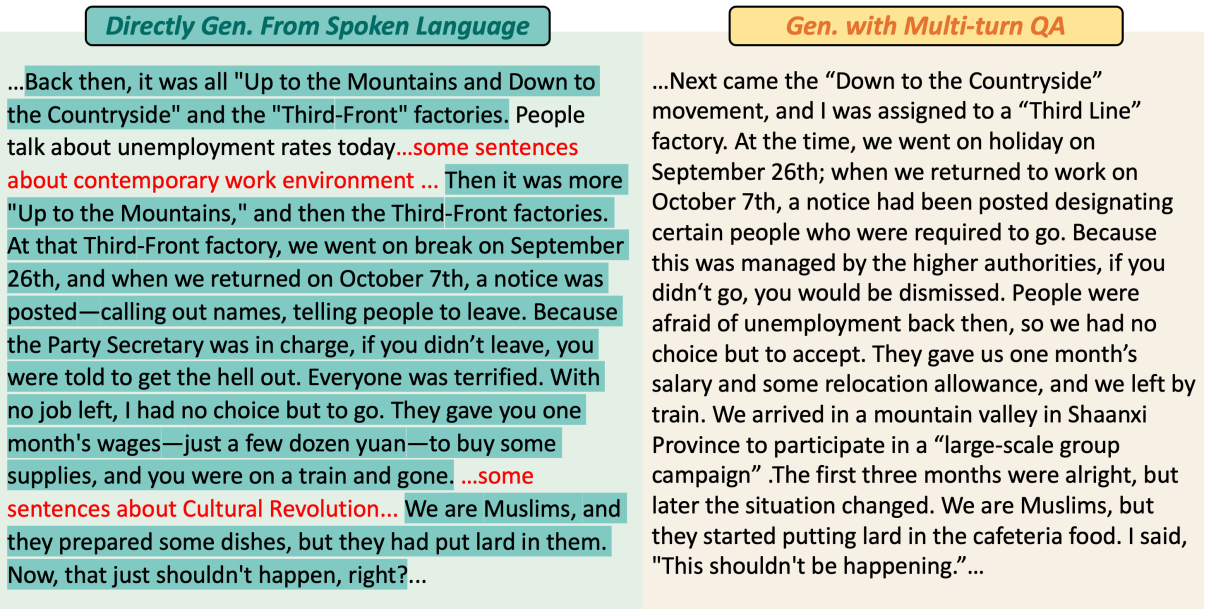

Figure 2: Case study of life writing excerpts regarding the "Third Front Construction". The left panel shows text generated directly from spoken language by GPT, while the right panel illustrates the results using a Multi-turn QA strategy.

Table 4: Impact of source discourse mode on faithfulness: Written Text V.S. Spoken Language.

Source	Setting	Acc_{type}	Acc_{prop}	MEA
Spoken Lang.	All	85.12	78.88	75.53
	HC	90.62	66.14	74.14
Written Text	All	73.90	61.71	56.77
	HC	71.09	55.36	56.05

Logical Cohesion vs. Spatiotemporal Fragmentation. The spoken transcript exhibits frequent shifts across time and context: the narrator recounts the "Third Front Construction" experience but intermittently digresses to other life stages, resulting in an under-specified temporal order and a fragmented narrative flow (Figure 2, left). Direct transcription improves surface fluency (e.g., removing disfluencies and grammatical errors) but preserves the spatiotemporal fragmentation in the original spoken language (see Appendix C.3), yielding a narrative that remains difficult to follow. In contrast, Multi-turn QA Guided Reconstruction first identifies key event nodes (who/what/when/where) and uses them as a planning scaffold during rewriting. This produces a clearer event progression and a tighter narrative arc, improving readability while keeping the story grounded in the speaker's original experience.

Mitigating Fact Distortion and Hallucination. We also observe a common faithfulness failure in direct prompting: *unsupported specificity*. In the transcript, the narrator attributes going to the factory to a vague authority ("those above were in charge"). Direct transcription may concretize this into a specific role (e.g., "the Party Secretary was in charge"), which is not supported by the source. Such unauthorized detail injection can undermine narrator trust and violates the faithfulness requirement of Life Writing. By conditioning generation on extracted event structures, Multi-turn QA Guided Reconstruction constrains the model to retain the original level of specificity, using structured events as *factual anchors* that reduce hallucinated or altered details.

6 Conclusion

In this work, we introduce an evaluation framework for life writing centered on faithfulness and situatedness. We propose corresponding metrics and construct a dataset comprising 402 oral life narratives from older adults, accompanied by fine-grained annotations to serve as a benchmark. Our findings indicate that LLMs lack the capacity to capture nuanced details from spoken language and identify Cultural Specific Items (CSIs) that bridge the expression gap between older adults and GenZ, underscoring a need for targeted future studies. Recognizing that the primary bottleneck lies in processing the oral discourse of seniors, we propose a Multi-turn QA strategy for information extraction to mitigate information distortion.

Limitations

First, our work focuses on episodic life narratives rather than full-length life histories. While this setting enables controlled evaluation of factual fidelity and audience adaptation, extending the framework to long-form biographical generation would require addressing additional challenges. Second, our factuality evaluation centers on explicit event-level attributes (e.g., time, event type, and location). Although effective for capturing core factual preservation, this formulation does not account for higher-order relations such as causality between events, which we leave for future work. Finally, our dataset is limited in scale and scope, consisting of 402 narrative pairs from a single cultural context. While sufficient for validating the proposed metrics, broader generalization would benefit from larger and more diverse datasets, as well as additional target audiences beyond Gen Z.

Ethics Statement

We acknowledge that all authors of this work are aware of and comply with the ACL Code of Ethics.

Use of Human Data and Annotations This study uses spoken narratives collected from publicly available interview videos. No new data were collected directly from human subjects by the authors. Human annotation is employed to verify and correct automatically annotation. Annotators were recruited by the authors' institution, and were compensated for their work. All annotations were conducted solely for research purposes, with details provided in the appendix A.4. 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, these references do not enable identi- fication of any individual speaker.

This study received approval from the institutional ethics review board. The ethical approval number will be provided upon acceptance of the manuscript to maintain anonymity during the peer-review process. All data used in this study are freely available to the public.

Risks Life narratives may contain personal or sensitive experiences, and automated life writing may risk factual distortion or misrepresentation. To mitigate these risks, our task formulation and evaluation explicitly emphasize factual fidelity and responsible audience adaptation. The data used in this work are derived from publicly accessible sources, and we do not include personal identifiers or attempt to identify individuals. We use Gemini(DeepMind, 2025) to correct grammatical errors in this paper.

References

Paul Longley Arthur. 2009. Saving lives: Digital biography and life writing. In *Save as⋯Digital memories*, pages 44–59. Springer.

Minwook Bae and Hyounghun Kim. 2024. Collective critics for creative story generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18784–18819, Miami, Florida, USA. Association for Computational Linguistics.

Michel Bonnin. 2009. 失落的一代: 中國的上山下鄉運動（*The Lost Generation: The Rustication Movement in China.*）, *1968-1980*. Chinese University Press.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. 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.

Kate De Medeiros. 2014. *Narrative gerontology in research and practice*. Springer Publishing Company.

Kate de Medeiros, Quinn Kennedy, Thomas Cole, Rosemary Lindley, and Ruth O'Hara. 2007. The impact of autobiographic writing on memory performance in older adults: A preliminary investigation. *The American Journal of Geriatric Psychiatry*, 15(3):257–261.

DeepMind. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

DeepSeek-AI. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Michael Dimock. 2019. Defining generations: Where millennials end and generation z begins. *Pew research center*, 17(1):1–7.

Anna Dolot. 2018. The characteristics of generation z. *E-mentor*, 74(2):44–50.

Gérard Genette. 1980. *Narrative discourse: An essay in method*, volume 3. Cornell University Press.

Howard Giles and Jessica Gasiorek. 2011. Intergenerational communication practices. In *Handbook of the psychology of aging*, pages 233–247. Elsevier.

Carlos Gómez-Rodríguez and Paul Williams. 2023. A confederacy of models: a comprehensive evaluation of LLMs on creative writing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14504–14528, Singapore. Association for Computational Linguistics.

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, Singapore. Association for Computational Linguistics.

Craig Howes. 2020. Life writing.

Chun Kang, Zhigu Qian, Zhen Fu, Jiaojiao Fu, and Yangfan Zhou. 2025. COAS2W: A Chinese older-adults spoken-to-written transformation corpus with context awareness. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17887–17906, Suzhou, China. Association for Computational Linguistics.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and 1 others. 2023. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*.

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.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

National Bureau of Statistics of China. 2021. 第七次全国人口普查公报（第六号）：人口受教育情况 (Main Data of the Seventh National Population Census (No. 6): Educational Attainment.). Accessed: 28 December 2025.

Eugene Albert Nida. 1964. *Toward a science of translating: with special reference to principles and procedures involved in Bible translating*. Brill Archive.

James Olney. 1998. *Memory and narrative: The weave of life-writing*. University of Chicago Press, Washington, DC.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Andrew Piper, Richard Jean So, and David Bamman. 2021. Narrative theory for computational narrative understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 298–311, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Gerald Prince. 2019. Narratology.

Qwen. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Max Saunders. 2008. Life-writing, cultural memory, and literary studies. *Cultural memory studies: An international and interdisciplinary handbook*, pages 321–331.

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.

Shayan Talaei, Meijin Li, Kanu Grover, James Kent Hippler, Diyi Yang, and Amin Saberi. 2025. Storysage: Conversational autobiography writing powered by a multi-agent framework. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*, UIST '25, New York, NY, USA. Association for Computing Machinery.

Paul Thompson. 2017. *The voice of the past: Oral history*. Oxford university press.

Lawrence Venuti. 2017. *The translator's invisibility: A history of translation*. Routledge.

Yuning Wu, Jiahao Mei, Ming Yan, Chenliang Li, Shaopeng Lai, Yuran Ren, Zijia Wang, Ji Zhang, Mengyue Wu, Qin Jin, and 1 others. 2025. Writingbench: A comprehensive benchmark for generative writing. *arXiv preprint arXiv:2503.05244*.

Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, and William Wang. 2024. Pride and prejudice: LLM amplifies self-bias in self-refinement. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15474–15492, Bangkok, Thailand. Association for Computational Linguistics.

Kevin Yang, Dan Klein, Nanyun Peng, and Yuandong Tian. 2023. DOC: Improving long story coherence with detailed outline control. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3378–3465, Toronto, Canada. Association for Computational Linguistics.

Kevin Yang, Yuandong Tian, Nanyun Peng, and Dan Klein. 2022. Re3: Generating longer stories with recursive reprompting and revision. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4393–4479, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: evaluating generated text as text generation. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA. Curran Associates Inc.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Data Construction

A.1 Demographic Characteristics of Older Adults

Category	Subcategory	Value
#Participants	–	402
Avg. age	–	82.44
Age distribution	65–74	47
	75–84	125
	85–94	141
	95+	3
	Missing	86
Gender	Male	74
	Female	56
	Missing	272
Region	Northern	270
	Southern	63
	Missing	69

Table 5: Demographic characteristics of older adult speakers.

A.2 Data Example

{
“file_id”:8，
“text”: Elderly Person：83 了，应该 50 退休，后来我 52 退休的，退休又干了 10 年，干到 62，12 个人，就剩我一个人了。(I’m 83 now. I was supposed to retire at 50, but I ended up retiring at 52. Then I worked for another 10 years after retiring, until I was 62. Out of the 12 of us, I’m the only one left now.)
Interviewer：那您有没有什么长寿的方法？(So, do you have any secrets to a long life?)
Elderly Person：我没有，我就是不爱在家这么懒坐着，就是想活动啊。我从 62 岁啊，就就就退休不上了，我原来是得退休以后又上来着，后来人单位有下岗的了，我就说我不上了，他说不上不行，你这还有好多事呢，我说那不合适吧，哈，后来我才说那你先上吧，以后再说，就这么答复的我。(I don’t, really. I just don’t like sitting around at home being lazy; I just want to keep moving. From the age of 62, I... I... I stopped working for good. I was actually supposed to stay on after retiring, but then people at the unit were getting laid off, so I said I wouldn’t stay. They said, ”You can’t leave, there are still so many things to do.” I said, ”That wouldn’t be right, would it?” Haha. Later I told them, ”Fine, let me finish this up first, then we’ll see.” That’s how I answered them.)
Interviewer：您比较重要还是？(So you were quite important there?)
Elderly Person：咳是呢，又干到了 62 啊，正好我接着，我接着那一笔呀，我接着给他就归是好喽，弄好喽我就不上了。(Oh, you could say that. I worked until I was 62. I just wanted to finish that last batch—I sorted everything out and got it all done, and then I stopped.)
Interviewer：您责任心还是挺强啊。(You really have a strong sense of responsibility.)
Elderly Person：就在天坛了，真棒就在天坛来玩来了。(Now I’m at the Temple of Heaven. It’s great, just coming here to spend time.)
Interviewer：那以前做什么工作呀？(What kind of work did you do before?)
Elderly Person：我最早的时候了是钳工啊，后来就管的材料管理啊。(In the very beginning, I was a fitter. Later on, I was in charge of materials management.)
Interviewer：这领导对您信任。(The management must have really trusted you.)
Elderly Person：我是我那个钳工也特别重要的，因为有些事牵扯好多人的事，我这我这个给那个件零件上划线啊，应该那么着做，他就按我那线走，后来我就退不下来，赶紧就把我这一批活干完喽，我就不干了，应该 50 退休，对后来我 52 退休的，退休又干了 10 年干到 62。(My job as a fitter was very important too, because it involved many other people’s work. I had to scribe the lines on the parts—it had to be done a certain way, and they would follow the lines I made. That’s why I couldn’t just quit. I had to hurry up and finish that batch of work before I left. I should have retired at 50, but I retired at 52, and then worked another 10 years until I was 62.)
Interviewer：那等于 62 岁一直在天坛。(So since 62, you’ve been at the Temple of Heaven all the time?)
Elderly Person：在天坛练着弄来，后来慢慢我们就人就多了，就打太极拳，打完太极拳以后啊，就走大圈，就走东西南边都都都走完了，就玩扑克，就半天就下来了，就这样，后来 12 个人就赶着闹闹，那个疫情嘛不是，后来就慢慢，慢慢就，就有走着的，有来不了的了，有比我大的，就是这样啊，就剩我一个人了，还有一个人了，是 86 的有一个，她也是有时候来有时候不来，因为她她儿子也没了，是老伴也没了，跟儿媳妇过吧，挺不顺心的，所以她就，啊她就挺没意思的，我老劝她就就凑合着过吧，你看她儿子没了，儿媳妇不是心情也不好吗，对是不是，我就这么着过吧，后来也碰不见她，她那边拆迁了啊，她那平房卖了，她孙女让她卖了平房，跟她妈一块过去，她 86 了，身体要说也算凑合，没有什么大毛病。(Practicing and hanging out here, yeah. Gradually, more and more people joined us. We’d do Tai Chi, and after that, we’d walk big laps—all the way around the east, west, and south sides. Then we’d play cards. Half the day just passes like that. Later, there were 12 of us hanging out together, but then the pandemic happened, you know? Slowly, some passed away, some couldn’t make it anymore, some were older than me... that’s how it went. Now I’m the only one left. Well, there’s one more lady, she’s 86. She comes sometimes and sometimes she doesn’t. Her son passed away, and her husband too. She lives with her daughter-in-law, but it’s not a very happy situation, so she... she feels like life is a bit meaningless. I always tell her to just make the best of it. See, her son is gone, so the daughter-in-law isn’t in a good mood either, right? So I just live my life. Later I stopped running into her. Her place was demolished for relocation—

her granddaughter made her sell the old house and move in with her mom. She’s 86; her health is alright, no major issues.)

Interviewer：看您这身体一点毛病没有吧？(It looks like you don’t have any health issues at all?)

Elderly Person：我也有血脂高，吃每天晚上吃一片降血脂的药，后来我血压也高了，血压呢早晨起来吃一片降血压的药，我就一天吃两片药，我就这么着挺好的。(I have high blood lipids. I take one pill for that every night. Later my blood pressure got high too, so I take a pill for that in the morning. I just take two pills a day, and I feel quite good.)

Interviewer：那还行，还可以，比大部分人都强多了，什么反应都没有，那您您老伴呢？(That’s not bad at all. Better than most people—no side effects or anything. What about your spouse?)

Elderly Person：不跟您走，我老伴走了哦，我老伴是好有十来年了呢，他是心脏，我儿子他们都单过，我女儿单过，他们都让我跟他们在一起过，我谁都不跟他们在一起过，你看我上午来一次吧，我下午还来一次，我从西门走到东门，我就这么走。(My spouse passed away about ten years ago—it was the heart. My son lives on his own, and my daughter does too. They all want me to live with them, but I won’t live with any of them. You see, I come here once in the morning and once in the afternoon. I walk from the West Gate to the East Gate. I just keep walking like that.)

Interviewer：那您这一天得走多少啊？(How much do you walk in a day?)

Elderly Person：这一次两个多小时，我原来打太极拳来着，身子有底子底子好，现在有时候呢，就打就打一套拳，别给儿女添麻烦就行，国家也挺照顾我们的，对也不收票，也不收钱，还给100块钱一个月。(Each time takes over two hours. I used to do Tai Chi, so my body has a good foundation. Now sometimes I just do one set of the form. As long as I don’t cause trouble for my kids, it’s fine. The state takes good care of us, too. No ticket fees, no charges, and they even give us 100 yuan a month.)

Interviewer：什么钱？(What money is that?)

Elderly Person：80啊80岁就给100块钱哦，打到您老年卡里头，您这老年卡就随身带，您买什么东西你就用老年卡买就行。(When you turn 80, they give you 100 yuan. It goes into your Senior Citizen Card. You carry the card with you, and whatever you need to buy, you just use the card.)

Interviewer：您现在退休金高吗？(Is your pension high?)

Elderly Person：我退休金不高，我就说6000块钱吧。(My pension isn’t high. Let’s just say it’s about 6,000 yuan.)

Interviewer：还可以了，够花吗？(That’s actually okay. Is it enough to spend?)

Elderly Person：够花不了，你干嘛花那么多钱呢，我们这年龄什么都不买了，就买点吃的就行了，你说那个药费给你拿多少，国家给你拿多少，我就这两片药，一天能花多少钱呢，你说是不是，挺好的。(More than enough! Why would you need to spend that much? At our age, we don’t buy anything anymore, just some food. And think about the medical costs—the state covers so much for you. I only take these two pills; how much could that cost in a day? Don’t you think? It’s pretty good.)

Interviewer：就像您锻炼好身体就都好了。(It’s like you said, as long as you exercise and stay healthy, everything is good.)

Elderly Person：啊对对对，就这么锻炼。(Yes, exactly. Just keep exercising like this.)

Interviewer：祝您健康长寿。(I wish you health and a long life.)

Elderly Person：好，谢谢啊。(Alright, thank you!)

“events”:[
{
“events_index”:1,
“event_type”:“Task Allocation”,
“role”:“The undersigned”,
“time”:“”,
“location”:“”,
“additional_properties”:{
“Receiving Organization”:“”,
“Occupational Category”:“Fitter”,
“Administrative Allocation”:“”
},
...
]
“words”:[
{
“words_index”:1,
“word”:Redundancy”,
“type”:“B”,
“meaning”:“The term refers to a phenomenon during China’s State-Owned Enterprise (SOE) reforms from the 1990s to the early 2000s, in which a significant number of employees lost their positions due to workforce downsizing or corporate restructuring.”
}
...
]
}

A.3 Annotation Prompt

You are a professional editor of elderly oral history. You need to extract event information and ”cultural gap words” from the following dialogue text.

Event Type List:
{{LIFE_EVENTS}}

Cultural Gap Words Types:
- SEIs: Semantic correspondence exists between the source and target cultures, but the form of expression differs (e.g., dialects). Example: “抽抽 (Chouchou)” (dialect) converted to “height reduction” (standard language).
- CSIs: Concepts exclusive to a specific historical time and space that are missing in the current (Gen Z) cultural context. Examples: “生产队 (Production Team)”, “工分 (Work Points)”.

Dialogue Text:
{{text}}

Output Format
Please output strictly in the following JSON format:
{{
“events”: [
{{
“event type”: “choose from event list”, (Ensure type-/attributes match the list)
“someone”: “I or others with relationship”,
“time”: “”,
“location”: “”,
“additional properties”: {{
“Receiving Unit”: “an example of a property”,
“Job Title/Position”: “an example of a property”,
“Assignment Form”: “an example of a property”
}} (Fill attributes based on text; output empty if info is missing),
}}
],
“words”: [
{{
“word”: “e.g., Production Team”,
“type”: “A(SEIs) or B(CSIs)”,
“meaning”: “Explanation of the term”
}}
]
}}

Ensure the explanation style matches the type:
- A(SEIs): Replace directly with modern general expressions.
- B(CSIs): Act as a footnote to explain its historical meaning.

Ensure all event types and attributes come from the provided list; do not add new ones!

A.4 Annotation Process

We pre-annotated 402 data entries according to the predetermined data format (A.2) using DeepSeek-V3. Subsequently, we recruited 15 annotators, all of whom belong to Gen Z and hold either bachelor’s or master’s degrees in History, Chinese, or Computational Linguistics. They were provided with the following annotation guideline:

Data Annotation Manual: Automated Life Writing and LLM Capability Evaluation
Thank you for participating in this annotation work! We are a research group currently conducting research on Automated Life Writing, aiming to help older adults transform their orally narrated life experiences into written stories suitable for reading. To achieve this, we need to evaluate and optimize the ability of Large Language Models (LLMs) to recognize the oral narratives of the elderly. Your Task: From the perspective of a human reader, correct the LLM’s understanding biases regarding the elderly’s oral content and help eliminate intergenerational reading barriers.
Detailed Annotation Tasks
The annotation work consists of two main parts: Event Annotation and Vocabulary Explanation Annotation.
Part I: Event Annotation
Goal: To ensure that events in oral narratives are accurately extracted, which serves as the critical foundation for narrative transformation.
Operational Instructions
Check and Correct Existing Annotations:Review every event extracted by the LLM and verify the following information:
Event Type: What happened(refer to the Event List for details)?
Participants: Who was involved ? (The speaker, other individuals, or a collective ? If others/collective, specify their relationship to the speaker).
Time: When did it happen ?
Location: Where did it happen ?
Other Attributes:Specific details unique to the event(e.g., “reason for going to the countryside”, “pension benefits”).
Format Requirement: Define according to the format: “Event Type, Participants, Time, Location, Other Attributes”.
Reward Mechanism
Correction/Modification: For each modification (changing attributes, deleting events, or adding defined events) confirmed as correct, a reward of 0.5 RMB is provided.
New Event Type: For each newly discovered and valid event type confirmed by the supervisor, a reward of 1.5 RMB is provided.
Part II: Intergenerational Vocabulary Explanation
Goal: To eliminate the reading gap between the oral expressions of the elderly and Generation Z (readers born between 1997 and 2012).
Vocabulary Categories
Please check words requiring explanation (including dialects and era-specific terms), focusing on two categories:
SEIs(Dialects/Oral Habits): Expressions that have a modern standard written equivalent but were spoken using dialect or oral habits.
Example: “I am old and‘chouchou’(shrunk) now”, where “chouchou” is explained as“ reduction in height/becoming shorter”.
CSIs(Historical Concepts): Proper nouns exclusive to a specific historical time and space that lack a direct modern linguistic equivalent, requiring historical/cultural background to understand.
Example: “Production Team”(Sheng chan dui), “Work Points” (Gong fen).
Annotation Tasks
Check and Correct LLM Explanations: Verify if an explanation is needed (common words do not need them), if the type (A or B) is correct, and if the meaning is accurate (use Baidu Baike or Wikipedia for verification).
Supplement Missing Vocabulary: If Type A or Type B words are used in the text but not explained by the LLM, please point them out and add explanations. Reward Mechanism
Correction: Each confirmed correction of an error earns 0.5 RMB.
Suggested Annotation Steps
Read the Full Text: Read the entire oral narra-

tive first to establish a holistic understanding of the life story and linguistic style.
Verify Events: Check the extracted event list one by one.Correct errors first, then supplement missing events.
Vocabulary"Literacy": Focus on words that might cause reading barriers for Gen Z, check LLM explanations, and supplement omissions.
Submit for Review: Review the results once more to ensure accuracy.
Confidentiality and Contact
Confidentiality: This project is in the research stage. Please strictly abide by confidentiality principles and do not share any data with third parties.
Feedback: If you encounter any ambiguities, please ask questions.

We randomly assigned two annotators to label the same document; specifically, annotators 1–14 each labeled 54 documents, while annotator 15 labeled 48 documents. We employed Cohen's Kappa to calculate the inter-annotator agreement for event type labeling.

The consistency result was $\kappa = 0.68$. The annotator discrepancies were primarily concentrated on the determination of event boundaries. Regarding subtle events with extremely brief descriptions (consisting of only a few words), some annotators tended to retain all mentions, whereas others opted to ignore them due to the lack of subsequent attribute descriptions. Ultimately, we adopted a strategy of retaining events with one or more attributes.

Each annotator was compensated between 200 and 300 RMB (including a base payment and bonuses, at 60+ RMB per hour, the pay rate is higher than the local minimum wage), with the exact amount depending on their workload and performance.

B Faithfulness Metrics

B.1 Question Template

- **Event Existence Questions:** "Does the text mention the event *E*?"

- **Event Attribute Questions:** "In the context of event *E*, is the value of property *P* equal to *V*?"

C Life Writing Generation

C.1 Prompts

C.1.1 Direct Transcription Prompt

You are a life writing editor. Your task is to convert interview transcripts into first-person life writing narratives.
Requirements:
1. Narrate in the first person using "I".
2. Retain all events mentioned in the original transcript.
3. Categorize vocabulary requiring explanation into two types:
- Type A (Dialects/Colloquialisms): If there is a corresponding standard written expression today, please replace the term directly with the standard written expression in the transcribed text.
- Type B (Historical Concepts): For terms specific to a historical context that lack modern equivalents and require cultural or historical background to understand, please keep the terms and provide their explanations in the annotations.
Place all annotations after the transcribed text using the following format:
Transcribed Text:
[Your transcription here]
Annotations:
1. Term 1: Explanation.
2. Term 2: Explanation.
Original Interview Transcript:
text

C.1.2 CoT Prompt

You are an oral history editor. Please perform the following analytical steps internally, but do not present the analysis process in the final output. Output only the JSON result that meets the requirements.
Strictly follow these tasks in order:
Task Steps
Step 1: Information Organization
From the following dialogue text, organize:
1. Events information:
- Event types and attributes must strictly belong to the provided event list.
- For each event, fill in only the attributes explicitly mentioned in the original text. If an attribute is not present, leave it empty.
- Do not supplement information that does not appear in the text.
2. Cultural gap words (two categories):
- Type A: Semantic correspondence exists between source and target cultures, but the expression differs (e.g., dialects). For example, "Choucho" (dialect) converted to "height reduction/getting shorter" (standard language).
- Type B: Concepts exclusive to a specific historical time and space that are missing in the current (Gen Z) cultural context. E.g., "Production Team", "Work Points".
Step 2: Oral History Transcription
Based on the events and words organized in Step 1, transcribe the original dialogue into an oral history narrative. Requirements:
1. Narrate in the first person using "I".
2. Retain all events and their attributes from the

extraction, integrating them naturally.
3. Handling cultural gap words:
- Type A: Directly replace with modern standard expressions.
- Type B: Keep the original term in the text and provide annotations at the end.
Event Type List:
{json.dumps(LIFE_EVENTS, ensure_ascii=False, indent=2)}
Dialogue Text:
{text}
Output Format
Strictly output in the following JSON format:
{
"events": [
{
"event_type": "Job Assignment/Recruitment", (Ensure type/attributes match the list)
"role": "I or others with relationship",
"time": "",
"location": "",
"additional_properties": {
"Receiving Unit": "Military Band",
"Job/Position": "Musician (later French Horn)",
"Assignment Form": "Examination"
} (Fill pre-defined attributes; leave empty if information is missing)
}
],
"words": [
{
"word": "Production Team",
"type": "A or B",
"meaning": "Explanation of the term"
}
],
"oral_history": "The transcribed oral history text. Format as follows:
Transcribed Text:
[Text here]
Annotations:
1. Term 1: Explanation.
}
Ensure all event types and attributes come from the provided list! Ensure the transcription is based on the extracted events and gap words, rather than a simple direct transcription of the original text!

C.1.3 Multi-round QA Prompt

Please analyze the occurrence of events in the text and answer in the following JSON format:
{
"exists": true/false,
"count": number,
"trigger": ["trigger sentence 1", "trigger sentence 2", ...]
}
Requirements:
1. Determine whether the event occurred.
2. Count the number of occurrences.
3. Find the trigger sentence for each event occurrence(a sentence that explicitly describes the event; return only one trigger sentence per event).
Dialogue Text: {text}
Has the event "{event_type}" occurred in the text? Please answer strictly in the JSON format specified above.

Please extract the attributes of the event in the text.
Requirements:
1. Refers to the specific event triggered by the trigger sentence.
2. Focus mainly on the trigger sentence and its surrounding context.
3. If the attribute is not explicitly mentioned in the trigger sentence, respond with "Not mentioned" .
4. Provide only the attribute value; do not output any extra content.
5. Keep it concise; use the shortest possible wording.
Dialogue Text: {text}
In the dialogue text, the "{property_name}" attribute of the "{event_type}" event (trigger sentence: "{trigger_sentence}") is:

You are a life writing editor. Your task is to convert interview transcripts into first-person life writing narratives.
Requirements:
1. Narrate in the first person using "I".
2. Retain all events mentioned in the original transcript.
3. Categorize vocabulary requiring explanation into two types:
- Type A (Dialects/Colloquialisms): If there is a corresponding standard written expression today, please replace the term directly with the standard written expression in the transcribed text.
- Type B (Historical Concepts): For terms specific to a historical context that lack modern equivalents and require cultural or historical background to understand, please keep the terms and provide their explanations in the annotations.
Place all annotations after the transcribed text using the following format:
Transcribed Text:
[Your transcription here]
Annotations:
1. Term 1: Explanation.
2. Term 2: Explanation.
Dialogue Text: {text}
Reference Events:{events}

C.1.4 Human In The Loop Prompt

You are an oral history editor. Your task is to strictly transcribe a first-person oral history narrative based on the provided reference events and gap words.
You must ensure no events or attributes are omitted, and no unmentioned information is added.
Requirements:
1. Narrate in the first person using "I".
2. Retain the timeline and all attributes within the "events" list.
3. Handling Cultural Gap Words:
- Type A: Directly replace the term with a modern standard expression based on the provided definition.
- Type B: Keep the original term in the text and provide the definition in the annotations.
Place all annotations at the end of the transcription using the following format:

Transcribed Text:
[Your transcription here]
Annotations:
1. Term 1: Explanation.
2. Term 2: Explanation.
Reference Events: {events}
Reference Gap Words: {words}

C.2 Hyperparameter Selection

The generation parameters were configured as follows: based on empirical observations, the max_tokens limit was set to 4096 for multi-turn QA strategies and 2048 for other generation strategies. A decoding temperature of 0.7 was employed across all tasks to balance generation characteristics.

C.3 Case Study

那时候上山下乡，三线工厂。你老现在失业多少多少，其实那时候就存在这个问题，是不是啊？说老实话，60 年代那时候，三年自然灾害接着文革，你只要是积极分子，心就是有什么说什么的人，到了后来晚期，都是受牵连，后期就弄你了。你那时候到自由市场，你都买根大葱，都是你资本主义，都得写检查，知道吗？接着上山下乡，接着就三线工厂。那时候三线工厂，9 月 26 号放假，10 月 7 号一上班，这贴出布告来了，谁谁谁你得走。因为是上头管着，你不走，不走滚蛋。那时候人都害怕，那我没工作了，走吧。给你一个月工资，给你几十块钱，让你买东西乱七八糟的，坐火车就走了。到那一去，真是真不错。那时候接见我们谁，陕西省的大型集团会战 3 个月。再再去不是那么回事了。我们是回民，就这弄这么些菜，里边告诉回民怎么着吧，给搁猪油，这不应该啊。

Back then, it was all “Up to the Mountains and Down to the Countryside” and those “Third-line Factories”. People talk about unemployment rates today, but honestly, that problem existed back then too, didn’t it? Let’s be real: starting in the 60s, you had the “Three Years of Natural Disasters” followed by the Cultural Revolution. As long as you were an activist—the kind of person who spoke their mind—by the later stages, you’d get dragged into it. They’d come after you. Back then, if you went to a free market just to buy a single green onion, they’d label you a “Capitalist” and force you to write a self-criticism. You know that?

Then came the “Up to the Mountains and Down to the Countryside” movement, followed by the “Third-line Factories”. At those factories, I remember we went on break on September 26th. When we came back to work on October 7th, a notice was posted: “So-and-so, you have to leave”. Because the higher-ups were in charge, if you didn’t go, you were kicked out. People were terrified back then—if I don’t go, I lose my job. So, you go. They’d give you one month’s salary, maybe a few dozen yuan to buy some supplies, and you’d hop on the train and leave.

When we first arrived, it actually seemed alright. I remember who received us—it was some “Three-Month Great Campaign” for a large-scale industrial group in Shaanxi Province. But after that, things weren’t what they seemed anymore. We are Muslim, and they’d prepare all these dishes, but then tell us, “Here’s how it is for the Muslim”—and they’d put lard in the food. That was just wrong.