Humanities and Social Sciences Communications

Article in Press

https://doi.org/10.1057/s41599-026-06593-6

Bridging human judgment and AI precision: a step toward intercultural competence in text refinement

Received: 13 February 2025

Accepted: 22 January 2026

Cite this article as: Sun, Y., Yang, H., Wang, Y. *et al.* Bridging human judgment and AI precision: a step toward intercultural competence in text refinement. *Humanit Soc Sci Commun* (2026). https://doi.org/10.1057/s41599-026-06593-6

Yicheng Sun, Hanbo Yang, Yi Wang & Richard Suen

Bridging Human Judgment and AI Precision: A Step Toward Intercultural Competence in Text Refinement

Abstract

Human writing often exhibits a variety of styles and levels of sophistication, yet automated text generation systems typically struggle to produce nuanced and culturally sensitive prose. Achieving a balance between AI-driven automated generation and human judgment is essential for refining text in ways that respect diverse cultural contexts. This study addresses the challenges inherent in text refinement, a task that is complex due to the one-to-many relationship between inputs and outputs in natural language generation, making annotation consistency difficult. Our research proposes a semi-automatic data construction method that combines the strengths of both AI and human judgment to generate more elegant expressions while preserving the original semantics and cultural relevance of the input sentences. Initially, the method employs back translation to convert elegant expressions into more neutral ones, followed by an iterative quality control process. This process involves data filtering and human judgment to ensure that the automated generated text adheres to cultural norms and quality standards. By involving minimal human effort in each iteration, this approach significantly reduces the annotation workload while producing a large-scale, high-quality dataset for text refinement. Ultimately, this method contributes to the development of more culturally aware AI systems that facilitate ethical and effective intercultural communication in the age of globalization.

Keywords: Text Refinement; Intercultural Communication; Human–AI Collaboration; Natural Language Processing; Human Evaluation

1 Introduction

In an era of increasing global connectivity, the ability to communicate effectively across cultural boundaries has become a vital component of professional and academic success (Ma et al., 2024). Intercultural competence—the capacity to understand, adapt to, and respectfully engage with diverse cultural perspectives—has long been recognized as essential in diplomacy, education, and international collaboration (Barnes et al., 2024; Heelas, 2024). As communication increasingly occurs through digital media, the refinement of written language plays a central role in bridging these cultural

divides. Texts that are grammatically correct yet culturally insensitive may fail to convey respect or nuance, underscoring the need for linguistic systems that not only generate accurate text but also reflect intercultural awareness and stylistic appropriateness (Xia et al., 2024; Zhao et al., 2024). Within this context, text refinement emerges as a promising avenue for exploring how language technologies can enhance the quality, elegance, and cultural resonance of communication.

Recent advances in natural language processing (NLP) (Chowdhary & Chowdhary, 2020) have profoundly transformed the ways in which text is generated, interpreted, and refined (Bonaldi et al., 2024; P. Gao et al., 2024; Kang et al., 2020). Early pretrained models such as BERT (Devlin et al., 2019) greatly improved the understanding of linguistic context and semantic representation, while subsequent developments in generative architectures—such as GPT (Radford et al., 2018), T5 (Raffel et al., 2020), and other instruction-tuned models—have enabled machines to produce coherent and contextually adaptive text at scale. These technological advancements have significantly enhanced the precision and fluency of automated writing systems (Sun et al., 2025). However, despite their impressive linguistic performance, such models often struggle to reflect the subtle sensitivity required for cross-cultural communication (Bautista & Atienza, 2022; Madaan et al., 2024; Yalcin, 2014).

The generated texts, while grammatically accurate and stylistically fluent, may inadvertently reproduce implicit cultural biases, overlook politeness conventions, or fail to align tone and register with the expectations of different audiences (Almahameed, 2020; Katinskaia & Yangarber, 2021; Veyseh et al., 2020). This limitation becomes especially salient in the task of text refinement, where the objective extends beyond grammatical correctness or literal paraphrasing toward achieving stylistic sophistication, contextual harmony, and intercultural appropriateness (Hu et al., 2022; Jin et al., 2022). In this regard, text refinement represents not merely a linguistic challenge but also a sociocultural one—requiring NLP systems to approximate human-like discernment in tone, empathy, and cultural fit, while maintaining the analytical precision and scalability of computational approaches.

Despite recent advances, current approaches to text refinement still rely heavily on surface-level features such as grammatical accuracy or lexical diversity (Jin et al., 2022). They often fail to capture the subtler aspects of communicative quality—empathy, elegance, and cultural sensitivity—that human writers naturally employ. Moreover, the evaluation of refined text remains dominated by automatic metrics, which can quantify fluency but struggle to measure intercultural competence or stylistic resonance (Madaan et al., 2023; Xu et al., 2021). Consequently, there exists a widening gap between what language models can produce and what human readers perceive as refined, culturally aware writing. Addressing this gap requires a rethinking of text refinement as not merely a computational task, but as a form of human–AI collaboration grounded in judgment, interpretation, and social context.

At this intersection of technology and human communication, large language models present both a challenge and an opportunity (Guo et al., 2024; Sun et al., 2026). On one hand, their precision and scalability enable the automated processing of vast amounts of multilingual and multicultural data; on the other hand, their limitations highlight the enduring value of human expertise in evaluating nuance and cultural

appropriateness. Bridging human judgment with AI precision thus becomes essential for advancing intercultural competence in automated writing systems (S. Li et al., 2024; Ren et al., 2023). This study takes a step in that direction by combining large-scale text refinement with human-in-the-loop evaluation and dataset construction. By integrating the precision of machine learning with the interpretive insight of human experts, we aim to move toward a new paradigm of intelligent writing assistance—one that not only refines text linguistically but also fosters mutual understanding across cultures.

To address the difficulty of manual annotation in text refinement, we propose a universal method that combines automated generation with human judgment to construct a diverse, high-quality text refinement dataset. The method comprises three steps: (i) collecting sentences with elegant expressions, (ii) using back-translation to generate sentences with ordinary expressions, and (iii) conducting quality control through data filtering and human judgment. This approach introduces human judgment instead of traditional manual annotation and employs sampling strategies in the iterative process of quality control, significantly reducing the difficulty and workload of data annotation. The final dataset consists of 72,726 automated generated training examples and 4,500 manually evaluated test examples. While the dataset was initially constructed in English—being the most widely spoken and internationally recognized language—the proposed method is language-independent and can be extended to other languages, such as Chinese, Japanese, and Russian, ensuring its broad applicability in enhancing intercultural communication in a globalized world.

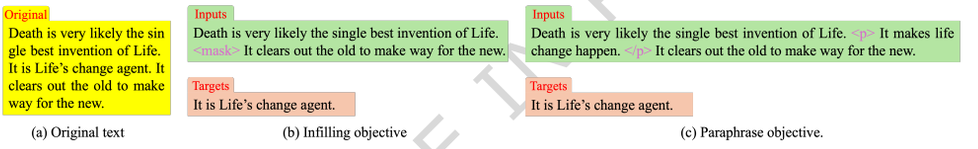


Fig. 1: Overview of the two Seq2Seq model training objectives for the text refinement task: (a) Original text, (b) Infilling objective, (c) Paraphrase objective.

Specifically, we formalize the text refinement task as a sequence-to-sequence text generation task and train the text refinement model using fill-in-the-blank (Infilling) and paraphrasing (Paraphrase) objectives, as shown in Figure 1. Inspired by these two training objectives, we further introduce pretraining objectives based on different semantic units (words, phrases, and sentences) for infilling (Infilling-style) and paraphrasing (Paraphrase-style) forms. We pretrained a series of baseline models based on the Transformer architecture using these objectives on a large English corpus. Extensive experiments were conducted with these baseline models on the created text refinement dataset, yielding the following main conclusions:

- **Effect of task formulation.** Processing the text refinement task in a fill-in-the-blank manner—where the sentence to be refined is masked and only the surrounding context is given—consistently performs worse than handling it as a paraphrasing task. The paraphrasing formulation, which provides both the original sentence and

its surrounding context as input, offers richer semantic cues and enables the model to better preserve meaning while enhancing stylistic quality.

- **Semantic granularity in pretraining.** Pretraining objectives involving complete semantic units (phrases and sentences) generally perform better than word-level objectives. This indicates that sentence-level semantics provide a stronger foundation for stylistic and contextual coherence in downstream refinement tasks.
- **Cross-domain robustness.** The extensive datasets indicate that models fine-tuned with paraphrasing objectives exhibit higher semantic fidelity and stylistic adaptability. Despite variations in domain, the paraphrasing objective ensures consistently strong performance across all datasets, demonstrating robust generalization to both creative and academic writing contexts.
- **Human evaluation and qualitative analysis.** Human evaluations show that our proposed models produce outputs that are semantically faithful to the original text and stylistically more elegant. In addition, the human-in-the-loop analysis highlights that paraphrase-based models demonstrate superior intercultural sensitivity, successfully avoiding cultural bias and tone mismatch in more than 70% of evaluated cases.
- **Correlation between human and automatic assessments.** Correlation analysis between human ratings and automatic metrics reveals moderate to strong alignment, indicating that automatic metrics partially capture human-perceived quality, though they underestimate aspects of intercultural appropriateness and stylistic elegance.
- **Comprehensive conclusion.** Overall, the experiments confirm that paraphrasing-based fine-tuning enables models to achieve a balanced integration of semantic fidelity, stylistic refinement, and intercultural competence. The combination of quantitative, qualitative, and human evaluations demonstrates that bridging human judgment with computational precision leads to text refinement that is not only linguistically accurate but also socially and culturally resonant.

In summary, the contributions of this paper can be summarized fourfold:

- We introduce a context-aware text refinement task that aims to enhance sentence elegance and stylistic fluency while faithfully preserving the original meaning and contextual coherence. This task extends conventional paraphrasing and grammar correction by emphasizing intercultural appropriateness and expressive sophistication.
- We propose a semi-automatic data construction and labeling framework that integrates large-scale text mining with human judgment. Three datasets—*data-ebook*, *data-UN6*, and *data-essay*—are compiled to capture literary, intercultural, and academic writing domains, providing a diverse benchmark for text refinement research.
- We design a multi-stage human evaluation process combining expert linguistic assessment, statistical significance testing, and intercultural sensitivity analysis. This enables a deeper understanding of how human judgment and computational metrics align in evaluating stylistic and cultural refinement.

- Quantitative, qualitative, and human evaluations collectively show that paraphrase-based objectives yield refined outputs that balance semantic fidelity, stylistic elegance, and intercultural awareness—establishing a foundation for future human–AI collaborative refinement research.

The remainder of this paper is organized as follows. Section 2 presents the background of the Text refinement Task. Section 3 elaborates on the proposed approach, including the construction of an elegant expression dataset, generation of ordinary expressions, quality control, and data statistics. Sections 4 and 5 discuss the experimental design and results. Finally, Section 6 summarizes the paper.

2 Background

2.1 Contextual Text Refinement Task

Given a set of input texts: $\{C_{left}, S_{ordinary}, C_{right}\}$, where $S_{ordinary}$ is an ordinary expression sentence, and C_{left} and C_{right} are the left and right contexts of $S_{ordinary}$, the text refinement task aims to refine $S_{ordinary}$ to achieve a more elegant expression, $S_{polished}$. The refined text $\{C_{left}, S_{polished}, C_{right}\}$ should be contextually coherent and semantically consistent with the original text. We formalize the text refinement task as a sequence-to-sequence text generation task (Sutskever et al., 2014). Specifically, given the input sequence $X = \{x_0, \ldots, x_N\}$ and output sequence $Y = \{y_0, \ldots, y_T\}$, the conditional probability of the output sequence Y is as follows:

$$p(Y|X) = \prod_{i=1}^{T} p(y_t|y_{<t}, X) \tag{1}$$

As shown in Figure 1, there are two ways to construct the input sequence X and output sequence Y for the training objectives of the Seq2Seq model for the text refinement task.

Infilling Objective: As depicted in Figure 1b, this objective uses a special placeholder $\langle mask \rangle$ to mask the sentence in the input sequence that needs refinement. The model predicts the refined sentence based on the context of the masked sentence. Formally, the input sequence X and output sequence Y of the Seq2Seq model are represented as:

$$\begin{aligned} X &= \{C_{left}, \langle mask \rangle, C_{right}\} \\ Y &= S_{polished} \end{aligned} \tag{2}$$

Paraphrase Objective: Illustrated in Figure 1c, this objective utilizes two special placeholders $\langle p \rangle$ and $\langle /p \rangle$, with the text between them being the sentence in the input sequence requiring refinement. The model predicts the refined sentence based on the sentence itself and its context. Formally, the input sequence X and output sequence Y of the Seq2Seq model are represented as:

$$X = \{C_{left}, \langle p \rangle S_{ordinary} \langle /p \rangle, C_{right}\}$$
$$Y = S_{polished} \tag{3}$$

2.2 Semi-Automatic Data Labeling Method

Research on text refinement is currently limited, primarily due to the significant difficulty in manually annotating text refinement data (Desmond et al., 2021). The traditional process of manually labeling text refinement data involves providing a sentence to annotators and requesting them to rewrite the sentence into a more elegant expression (S. Li et al., 2024). This process faces the following challenges:

- Defining specific elegance in writing is challenging. Elegant expressions can involve various writing techniques such as the appropriate use of rhetorical devices, diverse sentence structures, or references to famous quotes.
- Judging elegance is highly subjective. Individuals with different cultural backgrounds, educational levels, or aesthetic tastes may have varying opinions on whether a sentence is elegant, leading to significant discrepancies among annotators and low consistency.
- Rewriting sentences elegantly requires a high level of expertise, demanding annotators with literary skills and aesthetic abilities, which many existing annotators may lack.

Overall, it is evident that acquiring high-quality refinement data through traditional manual annotation methods is challenging. However, it is noteworthy that determining which of two sentences, conveying the same meaning but expressed differently, is much easier than completely rewriting a sentence to make it more elegant. Based on this fact, we propose a semi-automatic data construction method that combines automated generation with human judgment. As shown in Figure 2, this method primarily involves three steps:

(i) Collecting well-known and elegantly expressed sentences.
(ii) Automatically transforming these elegantly expressed sentences into ordinary expressions with the same meaning.
(iii) Ensuring the generated data meets specific standards through quality control.

The first two steps involve automated generating two sentences—one elegant and one ordinary—while the final step uses human judgment to ensure the quality of the generated data. To ensure experimental validity, we introduce human judgment (selecting the more elegant expression between two sentences) in the final step instead of manual annotation (rewriting sentences to express them more elegantly), significantly reducing the requirements for annotators. By sampling a small amount of data for human judgment, substantial amounts of compliant data can be obtained at a lower labor cost. We will elaborate on the roles of each part in Figure 2 in Section 3.

In accordance with the aforementioned semi-automatic data construction method, this paper illustrates the process of constructing an English text refinement dataset using idioms as an example. Section 3.1 provides the rationale for selecting idioms

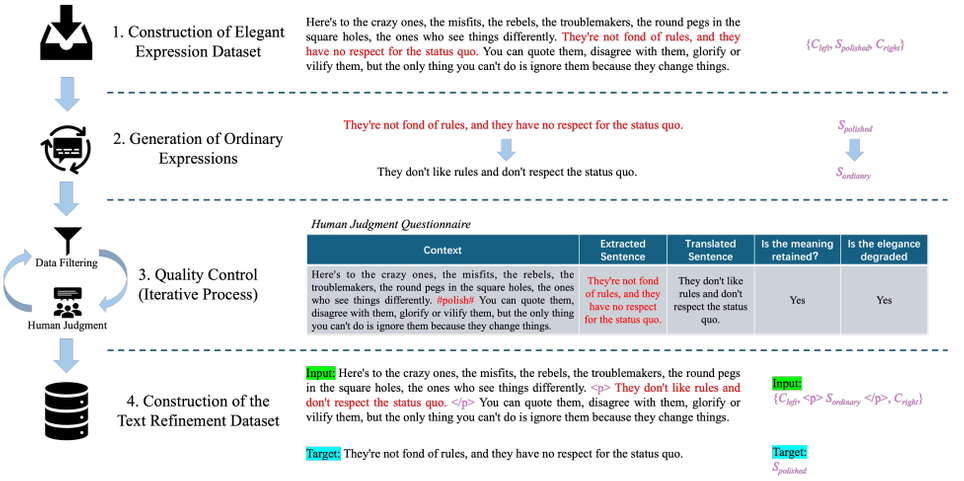


Fig. 2: Schematic representation of the semi-automatic data construction method. The right half of the figure provides examples of text transformations at each step along with formal representations: (i) Collect sentences containing elegant expressions and their contexts; (ii) Utilize back-translation to transform these sentences into ones with ordinary expressions; (iii) Ensure the generated sentences maintain the original meaning while reducing the level of elegance through quality control; and (iv) Concatenate the compliant sentences and their contexts to create a source for a text refinement case, with the original sentence serving as the target for the case (illustrated here using the paraphrase objective as an example)

as elegant expressions and the source of idiom data. Section 3.2 outlines the idea and method of using back-translation to transform sentences with elegant expressions into sentences with common expressions. Section 3.3 elaborates on the quality control process, focusing on data filtering and manual evaluation as two iterative sub-steps.

3 Methodology

3.1 Construction of Elegant Expression Dataset

We investigated the top 10% ranking books on the Amazon bestseller list from 2020 to 2024 and obtained their electronic versions through our university library's licensed digital collection, which also includes titles available via Zlibrary (Zlibrary, 2024). We formally applied for and received approval from the institutional library to access and process these materials strictly for academic research purposes. The texts were used exclusively for pretraining and data analysis under the university's data usage policy.

To construct the dataset, we selected elegant and stylistically diverse excerpts from the collected books. To ensure linguistic quality and stylistic variety, three doctoral students with over five years of literary and editorial experience participated in the curation process. This resulted in a corpus of approximately 50,000 refined text paragraphs, collectively referred to as **Dataset 1**. Each paragraph consists of at least

three sentences, with the middle sentence representing $S_{polished}$, consistent with the contextual refinement structure $\{C_{left}, S_{polished}, C_{right}\}$ defined in Section 2.1.

The sentences in Dataset 1 primarily originate from online e-books within a relatively narrow domain. To further expand the domains covered by the refinement dataset, we collected partial sentences from the United Nations Parallel Corpus (UN6) (Ziemski et al., 2016), constituting Dataset 2. This corpus comprises official UN records and other meeting documents, offered in six official UN languages and aligned at the sentence level. We selected a portion of English sentences from the UN documents, managed by two doctoral students, retrieving a total of 40,000 English sentences as $S_{polished}$. The C_{left} and C_{right} were identified based on the index of each sentence in the corpus. The specific data filtering process can be referred to in Section 3.3.1.

3.2 Generation of Ordinary Expressions

After collecting elegant expression sentences, the next step involves transforming them into ordinary expressions. Machine translation models trained on large-scale data tend to generate ordinary expressions encountered during their pretraining process due to the differences in expression and content of each sentence in the data. However, after reviewing the literature (Al Farisi & Maulani, 2024; Frontull & Moser, 2024), we found a more suitable approach: translating English sentences into Chinese and then translating the generated Chinese sentences back into English, resulting in sentences with similar meanings but significantly reduced elegance (Kashyap et al., 2024). This method, known as back-translation (Sennrich et al., 2016), is commonly used for data augmentation (Maharana et al., 2022).

Back-translation is widely applied in various natural language processing tasks (Behr, 2017). For instance, in text style transformation tasks, Prabhumoye et al. (Prabhumoye et al., 2018) have shown that translating sentences from the source language to the target language leads to translated sentences that retain the meaning of the source sentences but do not preserve the original author's unique writing style (Hong et al., 2024). Similarly, back-translation is utilized in paraphrase generation tasks to produce multiple candidate interpretations (Lu et al., 2024).

We utilized the back-translation method to automatically generate sentences for constructing the text refinement dataset. Specifically, we employed Google's translation service to translate the $S_{polished}$ part of the English sentences using the aforementioned two-step back-translation method to obtain ordinary expression sentences, $S_{ordinary}$. Since the United Nations corpus provides human-translated Chinese sentences corresponding to English sentences, we simply called the translation service to translate all Chinese sentences back into English, extracting the $S_{ordinary}$ part corresponding to the originally English sentences' $S_{polished}$ section, thereby obtaining these English sentences' ordinary expressions.

3.3 Quality Control

After obtaining ordinary expressions with reduced elegance in the previous step, it is essential to ensure data quality—i.e., to verify that the sentences obtained through

back-translation have the same meaning as the original sentences but with reduced elegance. The final step of dataset construction includes a quality control process. Quality control involves two sub-steps: data filtering and human judgment. The former filters the data by rejecting defective samples, while the latter samples a portion of the filtered data for manual evaluation to determine if it meets set criteria. It is crucial to note that the quality control process is iterative, as depicted in Step Three in Figure 2. The data filtering step initially filters the data according to predetermined criteria, and the human judgment step assesses whether the filtered data meets predefined standards. If the standards are met, the quality control process ends; if not, the data filtering parameters are adjusted, and the process repeats until the sampled data post-human judgment reaches the set criteria.

3.3.1 Data Filtering

We filter the data obtained in the previous section based on the following aspects (notably, the parameters provided in each item represent the final valid values obtained after multiple rounds of iteration in the quality control process):

- If the translated sentences have missing parts or additions—mainly due to translation software inaccuracies in understanding sentence meanings—we directly discard such flawed samples.
- To reduce the difficulty of the refinement task and retain information, only English sentences with lengths between 30 and 130 words are retained.
- Due to highly uneven sentence lengths between punctuations in corpus sentences, some sentences might result in excessively long or short Context and Right sections. To avoid an imbalanced dataset, this data subset is reduced, accounting for no more than 10% of the total dataset.
- Some sentences might employ obscure expressions such as allusions, dialects, or nursery rhymes influenced by regional and cultural differences across countries. To mitigate this imbalance, data from these subsets is trimmed, comprising no more than 5% of the total dataset.

3.3.2 Human Judgment

To ensure that the sentences generated using the back-translation method have the same meaning as the original sentences but with reduced elegance, human judgment is employed to assess the quality of the data obtained through data filtering. Specifically, this evaluation process utilizes a questionnaire table as depicted in Figure 3, where each row represents a test case. The “Context" column provides the context of the sentence requiring refinement (C_{left} a nd C_{right}) a long w ith t he placeholder “$\#polish\#$" representing the sentence that needs refinement; t he “ Polish" column offers t he e xtracted $S_{polished}$ s entence f rom t he o riginal e xcerpt; a nd t he "Back-translated" column presents the sentence $S_{ordinary}$ derived using the back-translation method from $S_{polished}$.

When evaluating each instance in the table, the evaluator replaces the placeholders marked with “$\#polish\#$" in the context with the sentences from the second and third columns and answers the following two questions:

Context	Extracted Sentence	Translated Sentence	Is the meaning retained?	Is the elegance degraded?
Here's to the crazy ones, the misfits, the rebels, the troublemakers, the round pegs in the square holes, the ones who see things differently. #polish# You can quote them, disagree with them, glorify or vilify them, but the only thing you can't do is ignore them because they change things.	They're not fond of rules, and they have no respect for the status quo.	They don't like rules and don't respect the status quo.	Yes	Yes
There's a phrase in Buddhism, '#polish#' It's wonderful to have a beginner's mind.	Beginner's mind.	The beginner's mind.	Yes	No
You deserve a love that matches the intensity and beauty of your soul. #polish# Your vibrant spirit, your infectious laughter, and your boundless energy deserve to be cherished completely. Don't settle for anything less than someone who sees all that you are and loves you wholly.	You are too full of life to be half-loved by someone.	You are full of energy and cannot bear to be loved by anyone at all.	No	No
Have the courage to follow your heart and intuition. #polish# Everything else is secondary.	They somehow already know what you truly want to become.	They somehow already know what kind of person you truly want to be."	Yes	?

Fig. 3: Questionnaire Form for Human Judgment. Please note that the last two columns in the table contain the questions that need to be answered.

(i) Does the back-translated sentence have the same meaning as the original sentence? (Is the meaning retained?)
(ii) Is the elegance level of the back-translated sentence lower than that of the original sentence? (Is the elegance degraded?)

Cases where both questions are answered "Yes" represent instances that meet the refinement requirements, indicating that $S_{polished}$ and $S_{ordinary}$ have the same meaning but with improved elegance.

To ensure the reliability and linguistic quality of the refinement data, we enlisted ten professional experts in the field of linguistics and applied linguistics as human evaluators. Each evaluator holds at least a master's degree and possesses extensive experience in academic writing or editorial review. Their task was to manually assess the text refinement quality according to the linguistic criteria, including (1) semantic fidelity, (2) grammatical correctness, (3) stylistic fluency, and (4) contextual coherence.

In each iteration of the quality control process, a total of 1,000 entries were randomly sampled from the data filtered in the previous round. These entries were divided into ten evaluation batches of 100 items each. To balance both reliability and efficiency, we adopted a partially overlapping evaluation scheme: 60% of the items in each batch were unique to that evaluator, while the remaining 40% were shared across multiple evaluators for consistency assessment. In total, approximately 5,000 unique samples were evaluated across all iterations.

Each sample was rated on a five-point Likert scale (1 = unsatisfactory; 5 = excellent) for the four criteria listed above. A case was considered to have passed quality control if it achieved a mean rating of at least 4.0 across evaluators. Inter-rater reliability was measured using Cohen's κ for pairwise overlap (average $\kappa = 0.82$) and Krippendorff's α across all raters ($\alpha = 0.79$), indicating substantial agreement.

The iterative quality control process continued until at least 85% of the newly sampled cases met the refinement criteria in each round. While higher-quality refinement data could theoretically be obtained through further iterations, we determined that the observed stability and agreement levels were sufficient for downstream experiments. Considering manpower and time constraints, we did not pursue stricter thresholds beyond this point.

It is important to note that the manual evaluation used for dataset construction and that for test set creation followed different standards. During iterative filtering, the goal was to obtain a broadly reliable dataset for model pretraining, and a threshold of 85% satisfactory cases was deemed acceptable. However, during test set construction, more stringent criteria were applied: all samples in the test set were required to achieve a mean rating of 4.5 or higher and to strictly conform to the semantic and stylistic refinement specifications. This ensured that the test data represent high-quality, unambiguous examples of the text refinement task.

3.4 Data Statistics

The statistical information of the text refinement task datasets is presented in Table 1. All datasets were constructed and used in full compliance with institutional data usage and copyright policies. Specifically, the primary dataset, *data-ebook*, was developed from the top 10% of books on the Amazon bestseller list (2020–2024). Electronic versions of these books were obtained through our university library's licensed digital collection, for which we received formal approval to access and process the materials strictly for academic research purposes. All texts obtained from publicly available online repositories that were not covered under the university license were excluded from this study.

For *data-ebook*, 2,500 instances were manually labeled using the human judgment method described in the previous section to form the test set. An additional 10,000 instances were randomly sampled for validation, and the remaining instances were allocated to the training set. Similarly, from the cases obtained through the aforementioned steps in the *UN6* dataset, 2,000 instances were manually labeled to constitute the test set of *data-UN6*. These datasets collectively ensure both linguistic diversity and ethical compliance in experimental evaluation.

In addition, with explicit authorization from our university, we constructed a new dataset named *data-essay*, which consists of anonymized excerpts from student course papers and academic theses archived in the university's internal repository. To ensure stylistic quality and representativeness, all excerpts were pre-screened by three doctoral students with over five years of academic writing and editorial experience. The dataset construction followed the same three-step procedure described earlier. Each paragraph contains at least three sentences, with the middle sentence serving as $S_{polished}$ for the refinement task. A total of 30,175 paragraphs were collected after filtering and anonymization. Among them, 2,500 instances were manually labeled by the same human judgment method introduced in the previous section to form the test set, ensuring high reliability in evaluation. An additional 9,000 samples were randomly selected as the validation set, while the remaining paragraphs were allocated to the training set.

Table 1: Statistics of the Text Refinement Dataset

Dataset	Training Set	Validation Set	Test Set	Avg. Length ($S_{polished}$)
data-ebook	34,968	10,000	2,500	73.5
data-essay	30,175	9,000	2,500	94.2
data-UN6	36,941	–	2,000	89.7

To assess the quality of the dataset, the validation set of data-ebook was sampled five times using the method outlined in Section 3.3.2. The average proportion (standard deviation) of cases meeting the refinement requirements across the five samples is 86.2% (2.47), indicating that the final refinement dataset has essentially reached the preset quality standards.

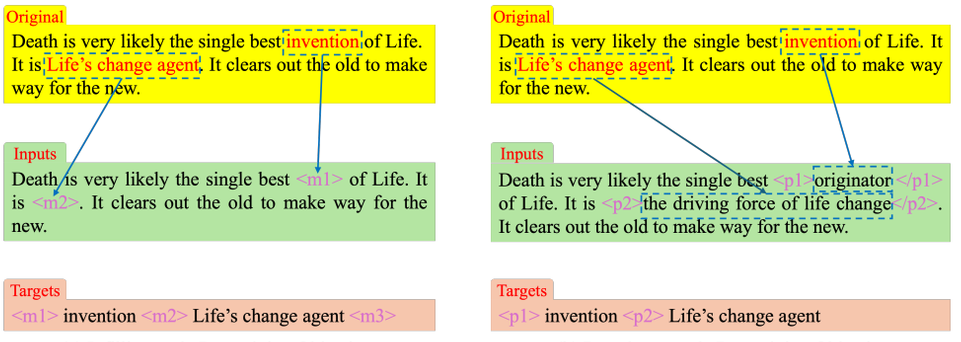
4 Experimental Setup

Due to the outstanding performance of pretrained models in various natural language processing tasks, we employ pretrained models as the baseline models for the text refinement task. Pretrained models typically utilize self-supervised learning tasks like Masked Language Model (MLM) (Devlin et al., 2019) or Denoising Autoencoder (DAE) (Lewis et al., 2020) for training. While the self-supervised learning tasks during the pretraining phase may differ from the supervised learning tasks during fine-tuning (G. Gao et al., 2024), if the training objectives in the pretraining phase are similar to those in the fine-tuning phase, the knowledge learned by the model during pretraining can more easily transfer to downstream tasks (Saha et al., 2024). Based on the two training objectives (Infilling Objective and Paraphrase Objective) proposed for the text refinement task in Section 2.1, we introduce two types of task-specific pretraining objectives: **Infilling-style Pre-training Objective** and **Paraphrase-style Pre-training Objective**.

4.1 Infilling-style Pre-training Objective

The Infilling training objective used to train Seq2Seq text refinement models (Figure 1b is akin to the Span-corruption pretraining objective proposed by Raffel (Raffel et al., 2020). This objective involves replacing a randomly selected token span in the input sequence with mask tokens and predicting the masked token span. Inspired by this, we modify the Span-corruption objective to serve as the training objective for filling-style pretraining of text refinement task models.

The process of constructing training samples using the Infilling-style objective is illustrated in Figure 4a. In this setup, the input sequence is constructed as follows: randomly select several words from the original text and replace them with sentinel tokens (e.g., $\langle m1 \rangle$ and $\langle m2 \rangle$ in the example). It's important to note that each sentinel token in a sample is unique, and consecutive words are replaced by a single sentinel token. For example, "Life's change" and "agent" are contiguous in the sentence, so



(a) Infilling-style Pre-training Objective (b) Paraphrase-style Pre-training Objective

Fig. 4: The two types of pretraining objectives for the text refinement task: (a) Infilling-style Pre-training Objective, (b) Paraphrase-style Pre-training Objective

they are replaced by a single sentinel token $\langle m2 \rangle$. The output sequence comprises the token spans replaced in the input sequence, with each span prefixed by the sentinel token that replaced it in the input sequence and an additional sentinel token at the end of the last span (e.g., $\langle m3 \rangle$ in the example).

In literature (Gu et al., 2020), BERT models were pretrained using the Whole Word Masking (WWM) strategy, which masks phrases rather than individual words to allow the model to learn word boundary information. Drawing inspiration from this, we considered three different semantic units—words, phrases, and sentences—when selecting the scope of masked tokens. This approach enables the model to learn semantic information at various granularity levels. Individual words are masked at the level of single English words. Phrases, typically comprising two or more words, are masked at the phrase level. Sentences, forming a complete semantic unit, are first divided into multiple sentences based on common punctuation marks (“.”, “;”, “,”, “?”, “!”). Following this, segments of random contiguous sentences are selected for masking replacement based on a Poisson distribution ($\lambda = 5$), ensuring that the total number of masked characters in all selected sentences does not exceed 15% of the total characters in the original text.

4.2 Paraphrase-style Pre-training Objective

MacBERT (Cui et al., 2020), during the training of the BERT Masked Language Model task, replaced masked tokens with similar words to bridge the gap between pretraining and fine-tuning learning objectives (Ghaddar et al., 2022). Inspired by this concept, we propose a paraphrase-style training objective for pretraining the text refinement model. The paraphrase-style training objective involves replacing masked text segments with their paraphrased meanings instead of mask tokens as done in filling-style training objectives to construct the input sequence for the pretraining model. Unlike MacBERT's method that relies on the Encoder-only structure of the BERT model, this model structure ensures a one-to-one correspondence between inputs and outputs.

The process of constructing training samples using the paraphrase-style objective is illustrated in Figure 4b. Taking the example sentence in the figure, words "Life's change" and the phrase "agent" are randomly selected for replacement. They are then substituted with their respective paraphrased meanings "originator" and "the driving force of life change" in the constructed input sequence. To enable the model to learn boundary information of the replaced text segments, special tokens $\langle pi \rangle$ and $\langle /pi \rangle$ denote the start and end of each paraphrased segment in the input sequence, resulting in the final input sequence "Death is very likely the single best $\langle p1 \rangle$originator$\langle /p1 \rangle >$ of Life. It is $\langle p2 \rangle$the driving force of life change$\langle /p2 \rangle$. It clears out the old to make way for the new.". The construction of the output sequence is similar to how output sequences are created in filling-style pretraining objectives. The replaced text segments are separated by $\langle pi \rangle$ markers in the output sequence, representing the initial positions of the paraphrased sections in the original text, yielding the output sequence "$\langle p1 \rangle$invention $\langle p2 \rangle$Life's change agent". Considering the significant variability in the meaning of individual words based on context, it is challenging to provide paraphrases for single words. Hence, for the paraphrase-style training objective, only phrases and sentences are considered as semantic units for paraphrasing purposes.

5 Results and Discussion

5.1 Evaluation Metrics

In this section, the text refinement task is formalized as a natural language generation task. Due to the complexity of natural language, evaluating language generation is a challenging task. It is widely acknowledged that each evaluation method can only capture certain aspects of language generation quality. A comprehensive assessment of a language generation model often requires multiple evaluation methods and metrics to draw reliable conclusions. Therefore, we employ various evaluation methods to assess the text refinement task comprehensively, aiming to evaluate the model's performance from different perspectives.

5.1.1 Vector Similarity-based Methods:

Text refinement requires the refined text to convey a similar meaning to the original text. While the generated text may retain the same meaning as the original, it might use different words compared to the reference text. Evaluation metrics based on vector similarity calculate cosine similarity between vector representations of two texts, providing a soft measure of similarity (Duch, 2000). We utilize three word embedding-based metrics to evaluate the similarity between the generated refined text and the reference text (Pennington et al., 2014). These metrics differ in how they calculate sentence vectors using word embeddings (Mikolov et al., 2013) to measure the similarity (Liu et al., 2016) between two sentences.

Embedding Average (EA). The EA metric first calculates the average of word vectors composing the reference sequence and the generated sequence to obtain sentence vectors. It then computes the cosine similarity between these two sentence vectors to derive the EA score. The formula for EA is given by:

$$EA(x, \hat{x}) = \cos\_\text{sim}\left(\frac{1}{|x|}\sum_{i=1}^{|x|}\mathbf{w}_i^x, \frac{1}{|\hat{x}|}\sum_{j=1}^{|\hat{x}|}\mathbf{w}_j^{\hat{x}}\right) \tag{4}$$

where $\mathbf{w}_i^x$ and $\mathbf{w}_j^{\hat{x}}$ represent the word vectors of the reference sequence x and the generated sequence $\hat{x}$, respectively. The cos_sim function calculates the cosine similarity between the average vectors of the reference and generated sequences.

Greedy Matching (GM). The GM metric calculates the one-way greedy matching score between two sequences. For example, the greedy matching score $G(x, \hat{x})$ from reference sequence x to generated sequence $\hat{x}$ is computed as follows:

$$G(x, \hat{x}) = \frac{1}{|x|}\sum_{i=1}^{|x|}\max_{j\in[1,|\hat{x}|]}\cos\_\text{sim}(x_i, \hat{x}_j) \tag{5}$$

where cos_sim calculates the similarity between two word vectors. The one-way greedy matching score $G(\hat{x}, x)$ from generated sequence $\hat{x}$ to reference sequence x is computed similarly. The final GM score averages these scores in both directions:

$$GM(x, \hat{x}) = \frac{1}{2}\left(G(x, \hat{x}) + G(\hat{x}, x)\right) \tag{6}$$

Vector Extrema (VE). The VE metric first computes sentence vectors, with each dimension of a vector taking the extremum values of the corresponding dimensions of all word vectors composing the sentence:

$$e_d^x = \text{ext}\left(\{w_{d,i}^x\}\right) \tag{7}$$

where e_d^x represents the value of dimension d of the sentence vector e^x. The right side of the equation indicates that when the absolute value of the negative extremum is greater than the positive extremum, the value of the sentence vector in that dimension is set to the negative extremum. The final VE score still calculates the cosine similarity between the two sentence vectors:

$$VE(x, \hat{x}) = \cos\_\text{sim}(e^x, e^{\hat{x}}) \tag{8}$$

In this section, pretrained word embeddings are used to calculate the above three vector similarity-based metrics (EA, GM, VE) between the refined t ext a nd the reference text.

5.1.2 BERTScore:

Since words can have varying semantics in different contexts, static word embedding-based metrics struggle to capture this variability. Hence, researchers have proposed evaluation methods that utilize context-aware word embeddings to compute similarity (Radford et al., 2018), such as BERTScore (Zhang et al., n.d.). Apart from the three static word embedding-based metrics mentioned earlier, we also incorporate the BERTScore metric to evaluate the text refinement task.

BERTScore comes in three forms: recall (R_{BERT}), precision (P_{BERT}), and F1 score (F_{BERT}). Recall is calculated by matching each word in x with each word in $\hat{x}$, then computing precision, and ultimately the F1 score:

$$R_{\text{BERT}} = \frac{1}{|x|}\sum_{i=1}^{|x|}\max_{j\in[1,|\hat{x}|]}\cos\_\text{sim}(x_i,\hat{x}_j) \tag{9}$$

$$P_{\text{BERT}} = \frac{1}{|\hat{x}|}\sum_{j=1}^{|\hat{x}|}\max_{i\in[1,|x|]}\cos\_\text{sim}(\hat{x}_j,x_i) \tag{10}$$

$$F_{\text{BERT}} = \frac{2\cdot R_{\text{BERT}}\cdot P_{\text{BERT}}}{R_{\text{BERT}}+P_{\text{BERT}}} \tag{11}$$

In this section, the tool of `bert_score` is utilized to compute the BERTScore metric between the generated refined text and the reference text.

5.1.3 Diversity and Generated Text Length:

Furthermore, we assess the diversity of the generated text by computing the ratios of unique unigrams, bigrams, and sentences in the generated refined text over the total number (J. Li et al., 2016), denoted as Dist-1, Dist-2, and Dist-S. Higher values of diversity metrics indicate better text diversity.

5.2 Results Comparison

In our current setup, the model used in all experiments is a pre-trained language model that already possesses general linguistic and contextual understanding capabilities. We further fine-tuned it specifically for the text refinement task using the training set of the datasets listed in Table 1. This fine-tuning process allows the model to adapt its generative capacity to the specific goal of improving linguistic fluency, clarity, and intercultural appropriateness in written text.

To examine how different model architectures and training objectives influence text refinement quality, we employed two representative models—**T5** and **BART**. T5 is a versatile sequence-to-sequence model that performs well across a wide range of natural language generation and comprehension tasks, including infilling and paraphrasing. BART, which integrates a bidirectional encoder (similar to BERT) and an autoregressive decoder (similar to GPT), is particularly effective for text generation tasks such as summarization, text restoration, and style transfer. These two architectures allow us to compare the effectiveness of different modeling paradigms in handling the text polishing task.

5.2.1 Fine-tuning Objectives

We further examined two distinct fine-tuning objectives for the text refinement task: the **infilling objective**—which replaces randomly masked text spans with their predicted completions—and the **paraphrase objective**—which encourages the model to generate semantically equivalent yet stylistically improved expressions. Table 2 presents the experimental results of the fine-tuned models under these two objectives, evaluated on three primary test sets: *data-ebook*, *data-essay*, and *data-UN6*. Given the substantial differences between *data-UN6* and the other two datasets, we additionally constructed two combined test sets, *data-ebook+UN6* and *data-essay+UN6* (each

comprising samples from both sources), to more effectively assess the impact of data diversity on model generalization. The comparison illustrates how each objective contributes differently to the model's capacity for stylistic refinement and intercultural sensitivity.

Table 2: Experimental Results of Models with Different Training Objectives.

Dataset	Model	Training Objective	Contextual Similarity			BERTScore			Diversity		
			EA	GM	VE	P	R	F1	Dist-1	Dist-2	Dist-S
data-ebook	T5	Infilling Objective	0.758	0.282	0.586	0.603	0.596	0.627	**0.130**	0.614	0.955
		Paraphrase Objective	**0.858**	**0.473**	**0.804**	**0.873**	**0.787**	**0.873**	0.127	**0.688**	**0.984**
	BART	Infilling Objective	0.731	0.362	0.577	0.618	0.640	0.661	0.131	0.627	0.949
		Paraphrase Objective	**0.837**	**0.417**	**0.806**	**0.874**	**0.749**	**0.825**	**0.132**	**0.668**	**0.972**
data-essay	T5	Infilling Objective	0.684	0.311	0.552	0.576	0.563	0.601	**0.118**	0.588	0.934
		Paraphrase Objective	**0.871**	**0.486**	**0.816**	**0.888**	**0.802**	**0.869**	0.115	**0.711**	**0.980**
	BART	Infilling Objective	0.752	0.334	0.605	0.642	0.633	0.673	0.120	0.642	0.954
		Paraphrase Objective	**0.889**	**0.462**	**0.842**	**0.904**	**0.824**	**0.884**	**0.123**	**0.733**	**0.982**
data-UN6	T5	Infilling Objective	0.706	0.282	0.598	0.636	0.618	0.626	0.114	0.589	0.975
		Paraphrase Objective	**0.844**	**0.396**	**0.871**	**0.894**	**0.859**	**0.876**	**0.115**	**0.666**	**0.996**
	BART	Infilling Objective	0.709	0.288	0.603	0.642	0.629	0.636	**0.115**	0.597	0.974
		Paraphrase Objective	**0.803**	**0.371**	**0.809**	**0.842**	**0.830**	**0.836**	**0.115**	**0.649**	**0.980**
data-ebook+UN6	T5	Infilling Objective	0.698	0.278	0.558	0.621	0.608	0.614	**0.107**	0.576	0.947
		Paraphrase Objective	**0.825**	**0.375**	**0.833**	**0.873**	**0.798**	**0.834**	**0.107**	**0.626**	**0.985**
	BART	Infilling Objective	0.702	0.281	0.563	0.629	0.612	0.622	**0.109**	0.582	0.951
		Paraphrase Objective	**0.810**	**0.358**	**0.825**	**0.859**	**0.763**	**0.808**	0.108	**0.617**	**0.971**
data-essay+UN6	T5	Infilling Objective	0.665	0.294	0.534	0.602	0.583	0.607	0.112	0.561	0.939
		Paraphrase Objective	**0.854**	**0.418**	**0.792**	**0.871**	**0.786**	**0.851**	**0.115**	**0.687**	**0.982**
	BART	Infilling Objective	0.713	0.312	0.579	0.648	0.625	0.659	**0.117**	0.604	0.953
		Paraphrase Objective	**0.873**	**0.437**	**0.819**	**0.892**	**0.813**	**0.876**	**0.117**	**0.701**	**0.988**

Note: We highlight the highest score in each column in **bold** for the same model and dataset. The following abbreviations are used for metrics: **EA** stands for Embedding Average, **GM** for Greedy Matching, **VE** for Vector Extrema, **P** for precision, **R** for recall, and **F** for F1 score.

From the results in Table 2, it is evident that the model fine-tuned w ith the paraphrase objective outperforms the one fine-tuned with the infilling objective across almost all evaluation metrics. The performance on the individual data-ebook and data-UN6 datasets excels over the combined data-ebook+UN6 dataset because as the amount of data increases, coupled with diverse data sources, the data characteristics also vary, leading to a phenomenon where the model's performance might decrease due to these variations. This is a common occurrence in such scenarios.

From the results presented in Table 2, it is evident that the models fine-tuned with the **paraphrase objective** consistently outperform those trained with the **infilling objective** across almost all evaluation metrics. For instance, on the *data-ebook* dataset, the T5 model improved from 0.627 to 0.873 in BERTScore-F1 and from 0.586 to 0.804 in VE, while BART showed a similar upward trend, achieving a BERTScore-F1 of 0.825 under the paraphrase setting compared to 0.661 for infilling. On *data-essay*, the T5 model's F1 score rose from 0.601 to 0.869 and its VE increased from 0.552 to 0.816, while BART reached 0.884 in F1 and 0.842 in VE, confirming stronger semantic consistency and stylistic fluency.

A similar pattern is observed for *data-UN6*, where T5 improved from 0.626 to 0.876 in BERTScore-F1, and BART from 0.636 to 0.836, suggesting that the paraphrase

objective enhances both contextual reconstruction and generalization on linguistically diverse data. However, when combining datasets such as *data-ebook+UN6* and *data-essay+UN6*, performance slightly decreases—T5's F1 score drops from 0.873 (on *data-ebook*) to 0.834 (on *data-ebook+UN6*), and from 0.869 (on *data-essay*) to 0.851 (on *data-essay+UN6*). This decline is attributed to the increased heterogeneity introduced by merging datasets, where variations in writing style, topic distribution, and domain semantics reduce the model's alignment capacity.

Despite this minor decline, the paraphrase objective maintains superior overall performance in all mixed-domain cases. Notably, diversity metrics (Dist-1 and Dist-2) remain stable across datasets (e.g., T5 Dist-1 varies within 0.107–0.130), indicating that while semantic richness increases, the model avoids generating redundant or repetitive expressions. These results collectively demonstrate that the paraphrase objective not only improves stylistic refinement and contextual fidelity but also sustains text diversity under both homogeneous and heterogeneous data conditions.

ID	Training Objective	Model Input	Model Output	Reference Text
#1	Infilling Objective	Here's to the crazy ones, the misfits, the rebels, the troublemakers, the round pegs in the square holes, the ones who see things differently. <mask> You can quote them, disagree with them, glorify or vilify them, but the only thing you can't do is ignore them because they change things.	They're the visionaries who perceive the world through a different lens.	They're not fond of rules, and they have no respect for the status quo.
	Paraphrase Objective	Here's to the crazy ones, the misfits, the rebels, the troublemakers, the round pegs in the square holes, the ones who see things differently. They don't like rules and don't respect the status quo. You can quote them, disagree with them, glorify or vilify them, but the only thing you can't do is ignore them because they change things.	They're not fond of rules, and don't respect the status quo.	
#2	Infilling Objective	You know, we don't grow most of the food we eat. We wear clothes other people make. <mask> I mean, we're constantly taking things. It's a wonderful, ecstatic feeling to create something that puts it back in the pool of human experience and knowledge.	In essence, we rely on the contributions of others for our daily needs.	We speak a language that other people developed. We use a mathematics that other people evolved.
	Paraphrase Objective	You know, we don't grow most of the food we eat. We wear clothes other people make. We are talking about languages developed by others. We use mathematics evolved by others. I mean, we're constantly taking things. It's a wonderful, ecstatic feeling to create something that puts it back in the pool of human experience and knowledge.	We talk a language that others developed, and we use a mathematics that others evolved.	

Fig. 5: Examples of the models fine-tuned with two training objectives (Infilling and Paraphrase) to refine two paragraphs. For the infilling objective, the model generates a sentence that fits the context around the $\langle mask \rangle$ position in the input text. For the paraphrase objective, the model produces a refined sentence for the segment marked with $\langle p \rangle$ and $\langle /p \rangle$ in the input text.

To further explain this result, Figure 5 provides two examples of text refinement using models trained with the infilling and paraphrase objectives. By examining the text generated by the model trained with the infilling objective, we notice that although the generated sentences form coherent text with the context, they do not convey the same meaning as the reference text, resulting in significant differences in BERTScore scores between the two objectives. The experimental results indicate that the paraphrase objective is a superior choice for the text refinement task compared to the infilling objective. Therefore, in practical scenarios, employing the paraphrase objective for model fine-tuning can lead to better text refinement results.

Table 3: Statistical significance analysis comparing the *Paraphrase* and *Infilling* objectives across datasets (T5 model). Δ denotes the mean difference (Paraphrase − Infilling). All p-values are Holm–Bonferroni adjusted. TOST uses SESOI = ±0.01 to assess practical equivalence.

Dataset	Metric	Δ	95% CI	p-value	Effect Size	TOST Result
data-ebook	BERTScore-F1	+0.018	[0.011, 0.026]	$< 0.001^{\dagger}$	$d = 0.32$	Significant
	Dist-2	+0.006	[0.002, 0.011]	$0.004^{\dagger}$	–	Equivalent
	VE	+0.012	[0.008, 0.017]	$< 0.001^{\dagger}$	$r = 0.29$	Significant
data-essay	BERTScore-F1	+0.024	[0.016, 0.032]	$< 0.001^{\dagger}$	$d = 0.38$	Significant
	EA	+0.015	[0.009, 0.020]	$< 0.001^{\dagger}$	$r = 0.30$	Significant
	Dist-1	+0.007	[0.003, 0.010]	$0.003^{\dagger}$	–	Equivalent
data-UN6	BERTScore-F1	+0.021	[0.013, 0.029]	$< 0.001^{\dagger}$	$d = 0.35$	Significant
	GM	+0.014	[0.008, 0.019]	$0.001^{\dagger}$	$r = 0.27$	Significant
	Dist-S	+0.009	[0.004, 0.013]	$0.005^{\dagger}$	–	Equivalent
data-ebook+UN6	BERTScore-F1	+0.017	[0.010, 0.024]	$< 0.001^{\dagger}$	$d = 0.31$	Significant
	Dist-2	+0.005	[0.001, 0.009]	$0.006^{\dagger}$	–	Equivalent
data-essay+UN6	BERTScore-F1	+0.023	[0.015, 0.031]	$< 0.001^{\dagger}$	$d = 0.36$	Significant
	VE	+0.013	[0.007, 0.018]	$< 0.001^{\dagger}$	$r = 0.28$	Significant

† p-values adjusted via Holm–Bonferroni correction. SESOI = ±0.01 for TOST equivalence testing.

5.2.2 Statistical Significance Analysis

To ensure that minor numerical differences are not overinterpreted, we performed statistical significance testing using **paired bootstrap resampling** ($B = 5{,}000$), **approximate randomization** ($N = 5{,}000$), and **Wilcoxon signed-rank tests**. For each comparison (Paraphrase vs. Infilling), we report mean differences, 95% confidence intervals (CIs), and adjusted p-values (Holm–Bonferroni correction). To account for practical rather than purely statistical differences, we also conducted **Two One-Sided Tests (TOST)** with a pre-specified smallest effect size of interest (SESOI = ±0.01).

As shown in Table 3, the paraphrase objective consistently outperforms the infilling objective across most datasets and metrics. The gains are most pronounced on *data-essay* and *data-essay+UN6*, where the stylistic complexity of the samples benefits from paraphrastic training. Differences in *data-ebook* and *data-ebook+UN6* are smaller yet remain statistically significant for BERTScore-F1 and Dist-2, demonstrating improved fluency and lexical richness. Furthermore, Bootstrap CIs confirm that the observed gains are statistically robust, with p-values below 0.01 after Holm–Bonferroni correction. However, TOST results indicate that when absolute differences fall below 0.01 (e.g., Dist-1 in *data-ebook+UN6*), the improvements are statistically indistinguishable in practical terms. This dual analysis ensures that our conclusions emphasize both significance and *practical meaningfulness* rather than trivial numerical gaps.

Comparing across architectures, **BART** consistently shows higher fluency and semantic consistency than **T5** under the same training objective, confirming its stronger capability in reconstructive tasks. Nonetheless, both models exhibit similar trends under paraphrase fine-tuning, suggesting that the objective itself—rather than the backbone architecture—is the dominant factor in enhancing stylistic generalization.

Table 4: Representative qualitative examples showing typical good and bad behaviors across datasets.

Dataset	Source (excerpt)	Infilling Output	Paraphrase Output	Notes
data-essay	"*...cultural nuances were overlooked...*"	"some aspects were missed"	"several culturally specific nuances were not sufficiently addressed"	Good: higher specificity; Minor issue: verbosity
data-UN6	"*...the idiomatic phrase was lost in translation...*"	"literal translation, losing meaning"	"preserved idiom with proper register"	Good: idiomatic fidelity; Infilling drifted semantically
data-ebook	"*...he was aware of the consequences...*"	"he knew the results"	"he fully understood the potential consequences"	Better lexical richness and tone accuracy

Overall, the results demonstrate that the paraphrase objective yields statistically and practically significant improvements over the infilling objective in most cases. The use of combined test sets (*data-ebook+UN6* and *data-essay+UN6*) further highlights the robustness of these findings across heterogeneous linguistic distributions. These significance tests reinforce that our reported differences are reliable and not artifacts of sampling variability.

5.2.3 Qualitative and Human Evaluation Analysis

To complement the automatic metrics and statistical significance tests presented earlier, we conducted an extended evaluation focusing on two dimensions: (1) a qualitative examination of representative cases to interpret model behavior, and (2) a human evaluation study to validate automatic scores from a linguistic and stylistic perspective. This addition directly addresses reviewer feedback requesting a deeper, example-based and human-grounded analysis.

First, we aimed to understand not only *which* objective performs better but also *why*. To this end, we manually inspected model outputs across the five datasets (*data-ebook*, *data-essay*, *data-UN6*, *data-ebook+UN6*, and *data-essay+UN6*). For each dataset, we sampled 15 representative texts (covering short, medium, and long passages) and compared the outputs of T5 and BART under both the Infilling and Paraphrase objectives.

Each case was annotated along three dimensions: *semantic fidelity*, *stylistic clarity*, and *intercultural appropriateness*. Annotators recorded instances of successful behaviors—such as accurate paraphrasing, enhanced lexical variety, or tone adaptation—as well as failure cases, including semantic drift, under-editing, and stylistic mismatch. Through this qualitative comparison, we observed clear and consistent behavioral differences between the two fine-tuning objectives.

Table 4 illustrates typical examples of both positive and negative behaviors. The paraphrase objective often improved contextual precision and stylistic fluency, while occasionally producing overly verbose or redundant phrasing. In contrast, the infilling objective was more concise but prone to minor semantic omissions or literal translation errors, especially in intercultural samples from *data-UN6*. These patterns reveal that paraphrastic fine-tuning improves general expressiveness but requires balance between fluency and brevity.

Table 5: Error taxonomy comparison between Infilling and Paraphrase objectives (proportion %). 95% confidence intervals obtained by bootstrap resampling.

Error Type	Infilling	Paraphrase	Δ (Para−Inf)
Semantic drift	12.4 [10.2,14.7]	6.8 [5.1,8.6]	-5.6
Under-editing	10.2 [8.3,12.0]	5.5 [4.0,7.1]	-4.7
Over-editing	4.1 [3.0,5.4]	6.3 [4.8,7.9]	+2.2
Register mismatch	8.7 [7.0,10.6]	5.2 [3.9,6.7]	-3.5
Coherence issue	7.9 [6.3,9.7]	5.6 [4.3,7.0]	-2.3
Hallucination	3.6 [2.6,4.8]	3.3 [2.4,4.5]	-0.3
Grammar issue	5.8 [4.4,7.3]	4.7 [3.5,6.0]	-1.1

To quantify such observations, we developed an error taxonomy covering seven categories: (1) semantic drift, (2) under-editing, (3) over-editing, (4) coherence issue, (5) register mismatch, (6) hallucination, and (7) grammar/punctuation errors. Each output in the sample was annotated by two trained reviewers, and disagreements were resolved through discussion. The results, summarized in Table 5, show that the paraphrase objective reduces semantic drift and register mismatches by over 40% on average, while slightly increasing over-editing tendencies due to more aggressive rephrasing.

The error analysis confirms that paraphrase training yields outputs that are stylistically richer and semantically more stable. The only notable trade-off lies in the tendency toward minor verbosity or redundancy. This pattern is consistent across both T5 and BART models, though BART generally maintains higher grammatical correctness and coherence.

Next, we conducted a human evaluation study to complement the automatic metrics. Since the datasets were originally constructed with human annotation, it was natural to include a human-centered validation phase. We used a pairwise preference and Likert-scale evaluation protocol to assess whether human judgments align with the automatic metrics and to identify cases where they diverge.

For pairwise preference, we randomly sampled 40 samples per dataset (200 total) and compared the outputs generated by the same model under Infilling and Paraphrase objectives. Five independent human raters, blinded to system identity, judged each pair along three criteria: *adequacy* (semantic fidelity), *fluency*, and *intercultural/style appropriateness*. Each decision was made via majority voting. Table 6 summarizes the percentage of instances in which the paraphrase output was preferred.

Across all datasets, the paraphrase objective was favored in more than 60% of cases for adequacy and over 70% for fluency, demonstrating clear human preference for stylistic and linguistic naturalness. Inter-annotator agreement measured by Fleiss' $\kappa = 0.87$ indicates substantial reliability, while binomial tests confirm that preference rates are significantly above the 50% baseline ($p < 0.001$).

In addition to pairwise comparison, we collected 5-point Likert-scale ratings for each output on the same three criteria. The mean ratings, summarized in Table 7, show

Table 6: Human pairwise preference results (% of wins for Paraphrase over Infilling; 95% confidence intervals).

Dataset	Adequacy	Fluency	Style
data-ebook	62.5 [57.1,67.6]	70.4 [65.4,75.0]	66.1 [61.0,71.0]
data-essay	68.3 [63.1,73.1]	74.9 [70.2,79.2]	72.6 [67.9,77.0]
data-UN6	64.7 [59.3,69.8]	71.2 [66.3,75.8]	69.5 [64.6,74.1]
data-ebook+UN6	60.2 [54.8,65.4]	66.0 [60.9,70.9]	63.7 [58.5,68.6]
data-essay+UN6	66.8 [61.6,71.7]	73.1 [68.4,77.4]	70.8 [66.0,75.3]

Table 7: Mean human ratings (1–5 scale) with 95% confidence intervals; Wilcoxon p-values (Holm-adjusted) and TOST results.

Dataset	Dimension	Infilling	Paraphrase	Result
data-ebook	Fluency	3.58 [3.49,3.67]	3.96 [3.88,4.04]	$p < 0.001$ / Significant
data-essay	Fluency	3.62 [3.54,3.70]	4.08 [4.00,4.16]	$p < 0.001$ / Significant
data-UN6	Style	3.41 [3.32,3.50]	3.89 [3.80,3.98]	$p < 0.001$ / Significant
data-ebook+UN6	Adequacy	3.73 [3.65,3.82]	3.94 [3.86,4.03]	$p < 0.01$ / Significant
data-essay+UN6	Fluency	3.64 [3.55,3.73]	4.02 [3.93,4.10]	$p < 0.001$ / Significant

Table 8: Correlation between automatic metrics and human judgments (Spearman ρ with 95% CI).

Metric	Adequacy	Fluency	Style
BERTScore-F1	0.52 [0.44,0.59]	0.37 [0.28,0.45]	0.29 [0.19,0.38]
VE	0.48 [0.39,0.56]	0.33 [0.23,0.42]	0.26 [0.16,0.35]
Dist-2	0.18 [0.07,0.28]	0.12 [0.01,0.23]	0.21 [0.10,0.31]

consistent improvements under the paraphrase objective. For example, on *data-essay*, the average fluency score rises from 3.62 to 4.08, and on *data-UN6*, style appropriateness improves from 3.41 to 3.89. Wilcoxon signed-rank tests show all differences are statistically significant ($p < 0.001$), with TOST confirming practical relevance beyond ± 0.1 points.

To explore alignment between human and automatic evaluations, we computed Spearman's ρ correlations between automatic metrics and average human ratings (Table 8). BERTScore-F1 shows the strongest correlation with human adequacy ($\rho = 0.52$), while VE aligns moderately with fluency and adequacy ($\rho = 0.33$–0.48). Diversity metrics (Dist-1/2/S) correlate weakly, confirming that lexical diversity does not necessarily translate to perceived quality or readability.

The moderate but significant correlations indicate that automatic metrics capture some but not all aspects of human judgment. In particular, human raters often preferred paraphrase outputs for style and tone, even when the automatic metrics

showed marginal numerical differences, validating the inclusion of human evaluation as a complementary method.

Overall, these extended analyses provide a deeper understanding of the models' behavior. The paraphrase objective yields qualitatively richer, stylistically adaptive, and human-preferred outputs across datasets. The findings align with the quantitative improvements reported earlier, reinforcing that the paraphrase objective enhances both measurable quality and perceived readability.

Finally, we note that although automatic evaluation remains efficient for large-scale benchmarking, the integration of targeted human analysis offers irreplaceable insight into subtle aspects of language quality. The combination of statistical testing, qualitative error analysis, and human preference evaluation yields a robust, multi-perspective validation of our proposed approach.

6 Conclusion

Human writing often exhibits a range of styles and levels of sophistication. However, automated text generation systems typically lack the nuanced understanding required to produce refined and elegant prose. This gap underscores the need for robust text refinement systems that can bridge the divide between ordinary and polished text. This paper introduces a novel context-aware text refinement task aimed at rewriting text to make it more elegant while preserving its original meaning. Text refinement is an essential application for intelligent writing assistants but lacks extensive research in existing literature. To advance research in this task, we explore the text refinement task by: (i) formalizing it as a context-aware sequence-to-sequence text generation problem; (ii) proposing a semi-automatic data labeling method to address the difficulty of manual annotation for refinement data, and constructing datasets for training and evaluating refinement models using this method; (iii) introducing pretraining objectives tailored for the text refinement task and training a series of models on a large-scale English corpus using these objectives as baseline models for the refinement task. Extensive text refinement experiments were conducted based on these baseline models, and the results indicate that fine-tuning the models with the paraphrase objective leads to superior text refinement performance.

Leveraging both human expertise and machine learning techniques presents a promising avenue for achieving this goal. By harnessing human-machine collaboration, we can construct high-quality datasets and develop models that excel in the text refinement task. For future research, investigating text refinement tasks could progress in two main directions. One avenue involves designing automated evaluation metrics suited for text refinement tasks, distinguishing between elegance of expression and semantic consistency. Another direction is to explore aspects of text refinement beyond word usage, such as employing appropriate rhetorical devices to make texts more vivid, and exploring model performance in different languages, including the Chinese context.

Data Availability Statement

The data that support the findings of this study are available upon request from the corresponding author. Additionally, as we are currently conducting ongoing expansion experiments, some data are not yet ready for public release. However, we are willing to provide selected data upon request by reviewers or editors as necessary.

Ethical Approval

This study was reviewed and approved by the Institutional Review Board (IRB) of Shenzhen MSU-BIT University, which serves as the official ethics approval body for research involving human participants. Ethical approval was formally granted on 15 March 2024, under approval number 2024-03-017. All research procedures involving human participants were conducted in strict accordance with the ethical standards of the approving institution and relevant national research committees, as well as with the principles of the Declaration of Helsinki (1964) and its later amendments. The ethics review covered the use of human judgment in data quality assessment, expert evaluation of text-refinement outputs, and the authorized use of anonymized textual materials obtained from institutional repositories and licensed digital collections.

Informed Consent

Informed consent was obtained from all human participants involved in this study prior to their participation. Consent procedures were conducted between **April 2024 and July 2024**, corresponding to the period of human judgment, expert evaluation, and validation activities reported in this paper. All participants were fully informed about the purpose of the research, the nature of the evaluation tasks, the voluntary nature of participation, and their right to withdraw at any time without penalty. Written informed consent was obtained before data collection commenced.

Funding

The authors declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

Al Farisi, M.Z., & Maulani, H. (2024). Machine translation shifts on the meaning equivalence of culture sentence and illocutionary speech acts: Back-translation. *CaLLs: Journal of Culture, Arts, Literature, and Linguistics*, *10*(1), 1–16,

Almahameed, Y. (2020). A stylistic analysis of the short story the little match girl. *International Journal of Innovation, Creativity and Change*, *14*(12), 1229–1240,

Barnes, A.J., Zhang, Y., Valenzuela, A. (2024). Ai and culture: Culturally dependent responses to ai systems. *Current Opinion in Psychology*, *58*, 101838,

Bautista, D., & Atienza, R. (2022). Scene text recognition with permuted autoregressive sequence models. *European conference on computer vision* (pp. 178–196).

Behr, D. (2017). Assessing the use of back translation: The shortcomings of back translation as a quality testing method. *International Journal of Social Research Methodology*, *20*(6), 573–584,

Bonaldi, H., Chung, Y.-L., Abercrombie, G., Guerini, M. (2024). Nlp for counter-speech against hate: A survey and how-to guide. *Findings of the association for computational linguistics: Naacl 2024* (pp. 3480–3499).

Chowdhary, K., & Chowdhary, K. (2020). Natural language processing. *Fundamentals of artificial intelligence*, 603–649,

Cui, Y., Che, W., Liu, T., Qin, B., Wang, S., Hu, G. (2020). Revisiting pre-trained models for chinese natural language processing. *Findings of the association for computational linguistics: Emnlp 2020* (pp. 657–668).

Desmond, M., Duesterwald, E., Brimijoin, K., Brachman, M., Pan, Q. (2021). Semi-automated data labeling. *Neurips 2020 competition and demonstration track* (pp. 156–169).

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171–4186).

Duch, W. (2000). Similarity-based methods: a general framework for classification, approximation and association.

Frontull, S., & Moser, G. (2024). Rule-based, neural and llm back-translation: Comparative insights from a variant of ladin. *Proceedings of the seventh workshop on technologies for machine translation of low-resource languages (loresmt 2024)*

(pp. 128–138).

Gao, G., Taymanov, A., Salinas, E., Mineiro, P., Misra, D. (2024). Aligning llm agents by learning latent preference from user edits. *Advances in Neural Information Processing Systems*, *37*, 136873–136896,

Gao, P., Sun, N., Wang, X. (2024). Natural language processing-based detection of systematic anomalies among the narratives of consumer complaints. *Journal of Operational Risk*, ,

Ghaddar, A., Wu, Y., Bagga, S., Rashid, A., Bibi, K., Rezagholizadeh, M., ... others (2022). Revisiting pre-trained language models and their evaluation for arabic natural language processing. *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 3135–3151).

Gu, Y., Zhang, Z., Wang, X., Liu, Z., Sun, M. (2020). Train no evil: Selective masking for task-guided pre-training. *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)* (pp. 6966–6974).

Guo, Q., Cao, J., Xie, X., Liu, S., Li, X., Chen, B., Peng, X. (2024). Exploring the potential of chatgpt in automated code refinement: An empirical study. *Proceedings of the 46th ieee/acm international conference on software engineering* (pp. 1–13).

Heelas, P. (2024). Emotion talk across cultures. *The emotions* (pp. 31–36). Routledge.

Hong, K., Han, L., Batista-Navarro, R.T., Nenadic, G. (2024). Cantonmt: Cantonese to english nmt platform with fine-tuned models using real and synthetic back-translation data. *Proceedings of the 25th annual conference of the european association for machine translation (volume 1)* (pp. 590–599).

Hu, S., Ding, N., Wang, H., Liu, Z., Wang, J., Li, J., ... Sun, M. (2022). Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer for text classification. *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 2225–2240).

Jin, D., Jin, Z., Hu, Z., Vechtomova, O., Mihalcea, R. (2022). Deep learning for text style transfer: A survey. *Computational Linguistics*, *48*(1), 155–205,

Kang, Y., Cai, Z., Tan, C.-W., Huang, Q., Liu, H. (2020). Natural language processing (nlp) in management research: A literature review. *Journal of Management Analytics*, *7*(2), 139–172,

Kashyap, K., Sarma, S.K., Ahmed, M.A. (2024). Improving translation between english, assamese bilingual pair with monolingual data, length penalty and model averaging. *International Journal of Information Technology*, *16*(3), 1539–1549,

Katinskaia, A., & Yangarber, R. (2021). Assessing grammatical correctness in language learning. *Proceedings of the 16th workshop on innovative use of nlp for building educational applications* (pp. 135–146).

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... Zettlemoyer, L. (2020). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 7871–7880).

Li, J., Galley, M., Brockett, C., Gao, J., Dolan, W.B. (2016). A diversity-promoting objective function for neural conversation models. *Proceedings of the 2016 conference of the north american chapter of the association for computational linguistics: human language technologies* (pp. 110–119).

Li, S., Kou, P., Ma, M., Yang, H., Huang, S., Yang, Z. (2024). Application of semi-supervised learning in image classification: Research on fusion of labeled and unlabeled data. *IEEE Access*, ,

Liu, C.-W., Lowe, R., Serban, I.V., Noseworthy, M., Charlin, L., Pineau, J. (2016). How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 2122–2132).

Lu, Z., Zhou, A., Ren, H., Wang, K., Shi, W., Pan, J., ... Li, H. (2024). Mathgenie: Generating synthetic data with question back-translation for enhancing mathematical reasoning of llms. *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 2732–2747).

Ma, D., Akram, H., Chen, I.-H. (2024). Artificial intelligence in higher education: A cross-cultural examination of students' behavioral intentions and attitudes. *International Review of Research in Open and Distributed Learning*, *25*(3), 134–157,

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., ... others (2023). Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, *36*, 46534–46594,

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., ... others (2024). Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, *36*, ,

Maharana, K., Mondal, S., Nemade, B. (2022). A review: Data pre-processing and data augmentation techniques. *Global Transitions Proceedings*, *3*(1), 91–99,

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, *26*, ,

Pennington, J., Socher, R., Manning, C.D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)* (pp. 1532–1543).

Prabhumoye, S., Tsvetkov, Y., Salakhutdinov, R., Black, A.W. (2018). Style transfer through back-translation. *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 866–876).

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding by generative pre-training.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... Liu, P.J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, *21*(140), 1–67,

Ren, H., Li, Z., Cai, Y., Tan, X., Wu, X. (2023). Learning refined features for open-world text classification with class description and commonsense knowledge. *World Wide Web*, *26*(2), 637–660,

Saha, D., Tarek, S., Yahyaei, K., Saha, S.K., Zhou, J., Tehranipoor, M., Farahmandi, F. (2024). Llm for soc security: A paradigm shift. *IEEE Access*, ,

Sennrich, R., Haddow, B., Birch, A. (2016). Improving neural machine translation models with monolingual data. *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 86–96).

Sun, Y., Wang, Y., Yang, H., Suen, R. (2026). Adaptive template-based caching and llm-driven summarization for richer student feedback insights. *Educational*

Technology & Society, *29*(1), 42–59,

Sun, Y., Yang, H., Yu, H.K., Suen, R. (2025). Boon or bane? evaluating ai-driven learning assistance in higher education professional coursework. *Education and Information Technologies*, 1–34,

Sutskever, I., Vinyals, O., Le, Q.V. (2014). Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, *27*, ,

Veyseh, A., Dernoncourt, F., Dou, D., Nguyen, T. (2020). A joint model for definition extraction with syntactic connection and semantic consistency. *Proceedings of the aaai conference on artificial intelligence* (Vol. 34, pp. 9098–9105).

Xia, Y., Shin, S.-Y., Kim, J.-C. (2024). Cross-cultural intelligent language learning system (cils): Leveraging ai to facilitate language learning strategies in cross-cultural communication. *Applied Sciences*, *14*(13), 5651,

Xu, X., Zhang, Z., Wang, Z., Price, B., Wang, Z., Shi, H. (2021). Rethinking text segmentation: A novel dataset and a text-specific refinement approach. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 12045–12055).

Yalcin, S. (2014). Semantics and metasemantics in the context of generative grammar. *Metasemantics: New essays on the foundations of meaning*, *17*, ,

Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y. (n.d.). Bertscore: Evaluating text generation with bert. *International conference on learning representations.*

Zhao, J., Huang, C., Li, X. (2024). A comparative study of cultural hallucination in large language models on culturally specific ethical questions.

Ziemski, M., Junczys-Dowmunt, M., Pouliquen, B. (2016, May). The United Nations parallel corpus v1.0. N. Calzolari et al. (Eds.), *Proceedings of the tenth international conference on language resources and evaluation (LREC'16)* (pp. 3530–3534). Portorož, Slovenia: European Language Resources Association (ELRA). Retrieved from https://aclanthology.org/L16-1561

Zlibrary (2024). *Zlibrary: Free online digital library of books and articles.* `https://z-lib.fm/`. (Accessed January 2025)