Context-Driven and Reference-Guided Data Augmentation for Subtitle Translation

Hitoshi Ito, Naoto Shirai, Kazutaka Kinugawa, Hideya Mino, Rei Endo, Yoshihiko Kawai
Japan Broadcasting Corp., Tokyo, Japan
{itou.h-ce, shirai.n-hk, kinugawa.k-jg, mino.h-gq, endou.r-mm, kawai.y-lk}@nhk.or.jp

Abstract

Large language models (LLMs) have demonstrated strong performance in translation tasks. Subtitle translation presents unique challenges, such as preserving the original work's worldview and the distinctive speaking styles of its characters. Achieving high-quality translations that reflect these stylistic nuances typically requires bilingual data for a specific movie, which is often scarce or unavailable. Thus, we propose a data augmentation method that uses LLMs to improve translation performance for specific movies, even when only a few hundred bilingual sentence pairs are available. The method expands source-side data by rewriting original subtitles using information that can be extracted from the context, such as character profiles and scene descriptions, to maintain the tone and thematic consistency of the movie. For translation, the augmented sentences are aligned with manually translated originals using structural similarity, which enables style-preserving bilingual data generation via one-shot learning. Experimental results show that data augmented using the proposed method effectively improves BLEU scores for film subtitle translation, and achieves superior stylistic quality in human evaluation.

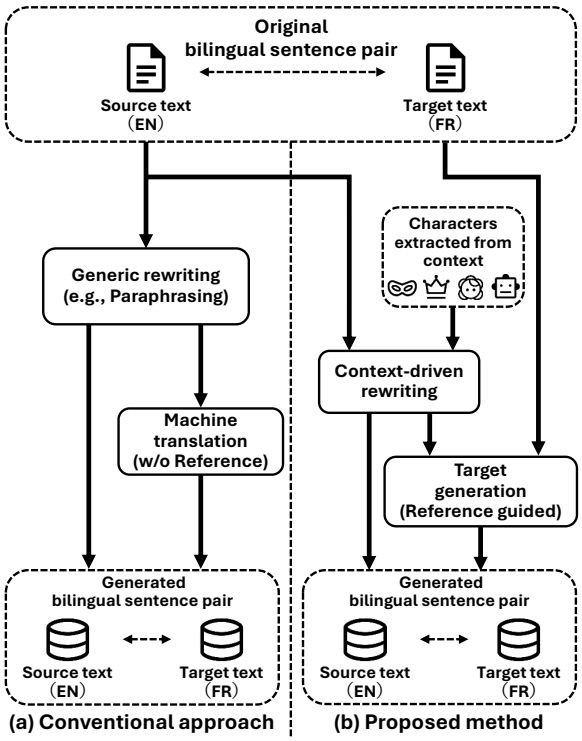

Figure 1: Comparison between the conventional approach and the proposed method. Conventional approaches are not designed to jointly consider the target work's content and style in the generated data.

1 Introduction

Large language models (LLMs) acquire extensive linguistic knowledge and demonstrate strong performance in context understanding, stylistic imitation, and translation (Achiam et al., 2024; Grattafiori et al., 2024; Yang et al., 2025). Instruction tuning is a technique for optimizing LLMs for specific tasks (Wei et al., 2022) using tailored training data to adapt models to particular writing styles and idiomatic expressions. In this study, we focus on instruction tuning in the context of translation tasks. Subtitles have long been used as a cost-effective communication tool that minimizes language barriers for audiences with diverse cultural and linguistic backgrounds (Liao et al., 2020). However, the machine translation of subtitles remains a challenging task. Subtitle translation must preserve colloquial expressions and cultural nuances specific to the work (Gupta et al., 2019). If these nuances are lost, character portrayals become flat, which diminishes viewer immersion and the overall appeal of the work.

When LLMs are applied to subtitle translation tasks, instruction tuning offers a promising approach to address these challenges. However, the application of instruction tuning to subtitle translation requires access to high-quality bilingual data

Findings of the Association for Computational Linguistics: ACL 2026, pages 41381–41394
July 2-7, 2026

for a specific movie. Publicly available bilingual datasets are typically limited to curated, specialized cases that involve specific movies or language pairs. When translating released movies into new language pairs, sufficient bilingual data is often unavailable, which makes the manual creation of new data for effective tuning impractical.

In this paper, we address these constraints by proposing a practical scenario: we prepare bilingual data for only a portion of a movie, such as a few hundred sentences, and expand it to build a high-performance translation model tailored to that work. This setting introduces two technical challenges for generating pseudo-paraphrased subtitle data. The first challenge is maintaining the worldview and thematic scope of the original work during monolingual data expansion. Conventional paraphrasing techniques may be able to preserve meaning (Moslem et al., 2022; Oh et al., 2023); however, subtitle translation requires additional consideration of the speaker's individuality and the unique worldview of the specific movie. Although in-context learning (Brown et al., 2020) has been shown to improve subtitle translation performance by incorporating meta-information such as movie titles (Pramodya et al., 2025), effective strategies in data augmentation for this domain have not been sufficiently investigated.

The second challenge is preserving the writing style during the translation of augmented sentences. The simple translation of augmented source text may result in the loss of the speaker's tone and stylistic nuances. Few-shot learning is useful for controlling inference outputs; however, the criteria for selecting appropriate example sentences that preserve the speaker's individuality and worldview remain unclear. Moreover, providing inappropriate examples can lead to performance degradation (Agrawal et al., 2023).

To address these challenges, we propose a data augmentation method that leverages context-dependent information using LLMs, as illustrated in Figure 1. The method consists of two key components: The first component is source-side data augmentation using context-dependent information. In this process, contextual descriptions such as character profiles are provided to the LLM to guide the rewriting of source sentences. This enables the generation of diverse sentences that reflect the context of the movie. The model is instructed to preserve the original sentence structure to support downstream translation. The second component is bilingual data generation via one-shot learning. To maintain the structure of the augmented source sentences during translation, one-shot learning is used by referencing manually translated versions of the original sentences. This approach aims to reproduce the original style by leveraging structural correspondences.

We evaluate how subtitle translation performance for the selected movie can be improved using only a few hundred bilingual sentence pairs extracted from publicly available data. Although it is realistic to prepare large amounts of source data alongside limited bilingual data, we test our method under more constrained conditions.

The main contributions of this study are as follows:

- We propose a novel data augmentation method for subtitle translation that rewrites original sentences using contextual descriptions such as character profiles, while preserving sentence structure. Unlike conventional paraphrasing, our method is guided by both source-side context and target-side references and is expected to respect the work's content while matching the tone of the target language.
- The proposed approach enables the generation of parallel training data suitable for instruction tuning using only a few hundred consecutive translated subtitles from a specific movie, without relying on metadata such as the movie title. This significantly reduces the human effort required for data preparation.
- Instruction tuning of a general-purpose LLM with the augmented data improved translation performance for the selected movie. Comparative experiments across four movies and four language pairs demonstrated that our method outperformed existing data generation techniques.

2 Related Work

Subtitle translation can be regarded as a low-resource task because of the difficulty to collect bilingual data for a specific movie. To achieve translations that effectively reflect the stylistic features and character-specific language of the original sentences, it is ideal to use parallel data drawn exclusively from the same movie. However, collecting such data can be challenging in practice.

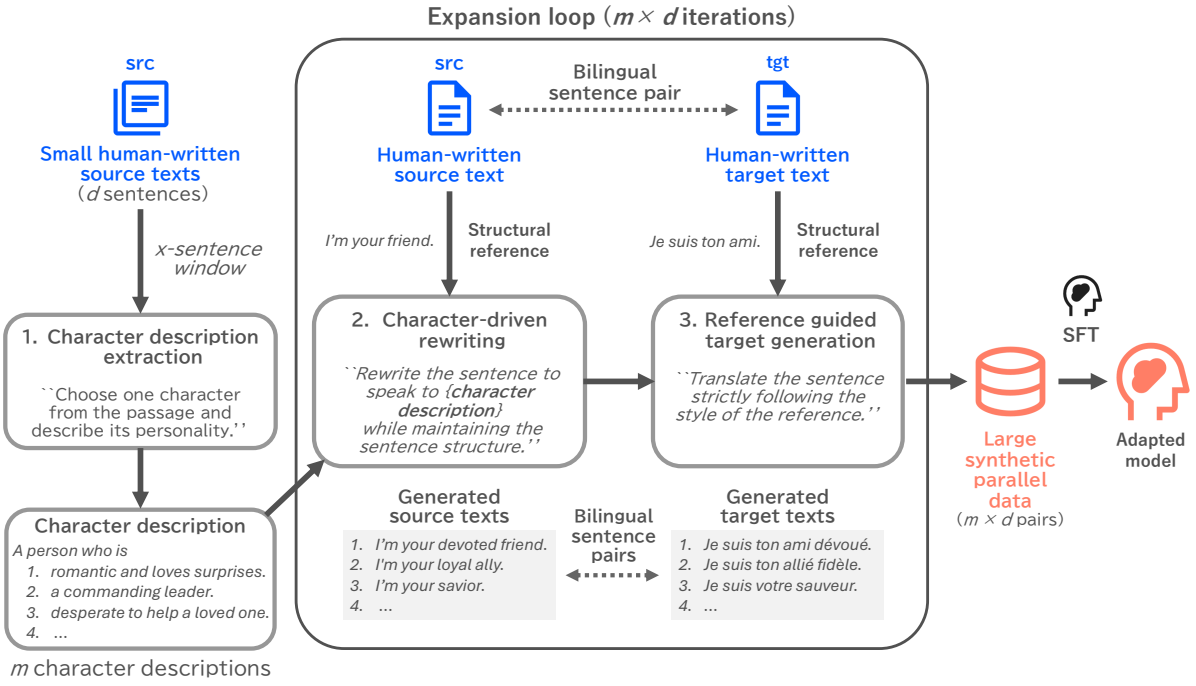

Figure 2: Overview of the proposed data augmentation method. This shows the flow of expanding d sentence-pairs of human-written bilingual texts with m types of character descriptions.

Two primary approaches have been proposed to enhance the translation performance of LLMs in low-resource settings, where bilingual data for the specific movie is limited: instruction tuning and in-context learning.

In-context learning is widely used to improve translation quality by incorporating informative elements into prompts. In previous studies, researchers demonstrated that providing lexical information (Pei et al., 2025) and appropriate example sentences can effectively guide translation output (Zhu et al., 2024). This study focuses on instruction tuning.

In low-resource tasks, data augmentation is often used to compensate for the lack of training data. Recent LLM-based data generation methods include the use of multiple personas to create diverse training data (Ge et al., 2024), text generation and paraphrasing with domain-specific terminology (Pengpun et al., 2024; Santoso et al., 2024), and multilingual synthetic data generation using multiple LLMs (Whitehouse et al., 2023). These methods aim to acquire broad task-specific knowledge but do not target narrow domains such as individual movies.

To address the scarcity of training data in translation tasks, various methods have been explored, including the use of lexical resources (Zheng et al., 2024), data augmentation via word-level paraphrasing (Fadaee et al., 2017). Although these approaches aim to enrich linguistic coverage, they generally do not focus on replicating the stylistic characteristics of specific works such as movies.

Pivot translation (Cohn and Lapata, 2007) and back-translation (Sennrich et al., 2016; Sälevä and Lignos, 2024) are widely used methods for generating synthetic bilingual data from monolingual sources. In contrast to these approaches, we propose a novel bilingual data augmentation method that expands the dataset by referencing both source-language and target-language data simultaneously.

Visual information has been used to guide character-specific translations in domains such as manga translation (Lippmann et al., 2025). We focus exclusively on textual information.

3 Data Augmentation with Context-Dependent Information

Because of the limited availability of bilingual corpora, we propose a data augmentation method that leverages contextual information for subtitle translation. Figure 2 illustrates the overall workflow, where each original sentence is expanded into multiple stylistically varied utterances using contextual cues, and then translated into the target language while preserving structural alignment. The proce-

dure consists of the following steps:

1. Divide the source corpus into segments of fixed consecutive sentences, and extract phrases by using LLMs, referred to as contextual descriptions, such as character profiles for each segment (Section 3.1).

2. Augment each source sentence using contextual descriptions by using LLMs (Section 3.2).

3. Translate each augmented sentence into the target language using one-shot learning, referencing the human translation of the original source sentence (Section 3.3).

3.1 Extracting Contextual Descriptions

We define the source corpus D as:

$$D = (D_1, D_2, \ldots, D_d), \tag{1}$$

where each D_i $(1 \leq i \leq d)$ is a manually created sentence, and d is the total number of sentences in the corpus. To extract contextual descriptions, we partition D into segments of x consecutive sentences, yielding:

$$G = (G_1, G_2, \ldots, G_m), \tag{2}$$

where each G_j $(1 \leq j \leq m)$ consists of x consecutive sentences. For each segment G_j, we extract character description F_j. In Appendix A, we provide the prompt used for extracting these features.

3.2 Augmenting Source Sentences Using Contextual Descriptions

After F_j is estimated for each segment G_j, the LLM generates augmented source data by rewriting each original sentence D_i using randomly selected combinations of sentences and contextual features. Each sentence D_i is expanded into a set of m augmented sentences:

$$S_i = (s_1^{(i)}, s_2^{(i)}, \ldots, s_m^{(i)}). \tag{3}$$

The complete augmented corpus is defined as:

$$S = (S_1, S_2, \ldots, S_d). \tag{4}$$

During augmentation, each original sentence D_i is rewritten into a new utterance conditioned on the given feature F_j. The prompt used for rewriting is provided in Appendix B. To preserve the speaker's tone, which is an essential aspect of subtitle translation, we instruct the model to maintain the original sentence structure. When character descriptions are used, modifying the speaker may alter the structure; therefore, instead, we instruct the model to change the recipient of the utterance.

3.3 One-Shot Translation Leveraging Structural Similarity

Each augmented sentence $s_k^{(i)} \in S_i$ $(1 \leq k \leq m)$ is translated using a one-shot prompting approach. We denote by $t_k^{(i)}$ the target-language translation of the augmented source sentence $s_k^{(i)}$, produced by the LLM using a one-shot prompt that references the human translation H_i of D_i (see Appendix C). We define P_i as the ordered sequence of aligned pairs:

$$P_i = \big(\langle s_1^{(i)}, t_1^{(i)}\rangle, \langle s_2^{(i)}, t_2^{(i)}\rangle, \ldots, \langle s_m^{(i)}, t_m^{(i)}\rangle\big). \tag{5}$$

Collecting these per-sentence sequences yields the complete bilingual parallel corpus P:

$$P = \big(P_1, P_2, \ldots, P_d\big). \tag{6}$$

An important aspect of this design is the rationale for referencing H_i. For subtitle translation, the goal is not merely to produce a literal translation of the rewritten source sentence, but to maintain the worldview and character-specific speaking style of the original work. Explicitly defining such stylistic fidelity in instructions is challenging; however, providing H_i as a one-shot example offers a strong contextual signal for preserving tone and character identity. Although $s_k^{(i)}$ differs from D_i, the rewriting process preserves structural similarity with D_i, ensuring that the one-shot reference remains effective.

4 Experiments

4.1 Experimental Settings

We conducted translation experiments using synthetic subtitle data generated by our proposed method. Specifically, we performed instruction tuning on the *Llama-3-8B-Instruct* model[1] using synthetic data produced by *Llama-3.3-70B-Instruct*[2].

4.1.1 Data

We used the *SubtitleMetaData* dataset[3], which contains multilingual subtitle data and metadata for various movies. Table 10 in Appendix D summarizes the genres and number of subtitle sentences used in our experiments. "Elio" is a work that was

[1] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

[2] https://huggingface.co/meta-llama/Meta-Llama-3.3-70B-Instruct

[3] https://huggingface.co/datasets/Ash96/SubtitleMetaData

Instruction tuning	Data augmentation	Reference guidance	Method	Spider-Man 3		The Duchess		The Princess and the Frog		Elio	
				BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
–	–	–	Base model	37.44	0.8412	42.76	0.8354	29.81	0.7628	21.79	0.7347
–	–	–	In-context learning (Pramodya et al., 2025)	37.98	0.8423	43.44	0.8351	29.28	0.7649	20.17	0.7327
✓	–	–	Original parallel data only	36.72	0.8388	42.54	0.8354	27.66	0.7558	21.22	0.7351
✓	✓	–	Paraphrase	37.99	0.8344	46.75	0.8447	31.57	0.7685	22.78	0.7385
✓	✓	–	Storytelling (Oh et al., 2023)	40.01	0.8501	47.76	0.8510	31.60	0.7713	21.41	0.7361
✓	✓	–	Persona (Ge et al., 2024)	39.05	0.8460	47.89	0.8508	31.83	0.7769	22.39	0.7373
✓	✓	✓	Proposed	**43.91**	**0.8561**	**49.11**	**0.8539**	**33.97**	**0.7822**	**24.57**	**0.7451**

Table 1: Overall results.

published after the release date of the model used for data generation. We selected it to evaluate the effectiveness of our proposed method on unseen works. The bilingual data for "Elio" was collected using the same procedure as for the other movies in this study. For each movie, the data were evenly split into training and test sets. The training set was first expanded 100-fold using our proposed method, and then 10,000 sentences were randomly selected for instruction tuning. During data augmentation, each original sentence was expanded by generating 100 variants using randomly applied contextual descriptions extracted from the source text. Each generated sentence was translated to generate parallel translation data.

4.1.2 Compared Methods

We compare the following systems:

- **Base model**: LLM without any task-specific adaptation.
- **Original parallel data only**: Instruction tuning on the original parallel set only; *no* synthetic sentence generation and *no* structural guidance from target references.
- **In-context learning** (Pramodya et al., 2025): Title and the four preceding English subtitle sentences are provided at inference to guide translation via prompting.
- **Paraphrase**: Source-side paraphrasing without contents control.
- **Storytelling** (Oh et al., 2023): Narrative-style rewriting used to expand source sentences.
- **Persona** (Ge et al., 2024): Persona-based rewriting with randomly sampled persona prompts.
- **Proposed**: Source-side augmentation with character descriptions, followed by one-shot translation that references the human translation of the original sentence.

'Data augmentation' indicates whether source-side augmentation was applied. 'Reference guidance' indicates whether human-translated target sentences were referenced during synthetic bilingual generation.

Our approach uses contextual descriptions extracted from five-sentence segments. Table 11 in Appendix F shows examples of extracted contextual descriptions. For expanding the source language data, we set the LLM's temperature value to 0.9, and for generating the target language data, we set it to 0.

4.1.3 Training and Evaluation

We performed instruction tuning using QLoRA (Dettmers et al., 2023), with the hyperparameters described in Appendix E. We performed optimization using AdamW (Loshchilov and Hutter, 2019).

For the evaluation, we translated English test data into French (FR) using zero-shot inference and measured performance using BLEU (Papineni et al., 2002) and COMET (Rei et al., 2020). We computed BLEU scores using SacreBLEU (Post, 2018) and obtained COMET scores using the *Unbabel/wmt22-comet-da*[4] model. To assess statistical significance, we performed a pairwise approximate randomization test (Riezler and Maxwell, 2005) with 10,000 trials and a p-value threshold of 0.05, using the SacreBLEU Python package.

4.2 Overall Results

Table 1 shows the overall translation performance across multiple works. Across all four movies, our proposed approach consistently improved scores compared with conventional methods, showing statistically significant differences. This indicates enhanced semantic translation quality and better consideration of stylistic nuances and tone. Notably,

[4] https://huggingface.co/Unbabel/wmt22-comet-da

English source	Paraphrase	Proposed
This is exactly what I need to ***scoop*** *Parker.* *Black suit Spider-Man.*	*C'est exactement ce que j'ai besoin pour* ***décrocher*** *Parker.* *Spider-Man est un* ***homme en noir****.*	*C'est exactement ce dont j'ai besoin pour* ***coincer*** *Parker.* ***Costume noir*** *Spider-Man.*

Table 2: Comparison of translation results.

Instruction tuning	Data augmentation	Reference guidance	Method	EN→FR		EN→FI		EN→ES		EN→DE	
				BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
–	–	–	Base model	37.44	0.8412	24.72	0.7970	32.63	0.8229	30.84	0.8068
–	–	–	In-context learning (Pramodya et al., 2025)	37.98	0.8423	23.37	0.7971	32.97	0.8191	30.17	0.7966
✓	–	–	Original parallel data only	36.72	0.8388	23.67	0.7809	32.63	0.8212	30.92	0.8065
✓	✓	–	Paraphrase	37.99	0.8344	27.29	0.8162	33.28	0.8191	33.28	0.8024
✓	✓	–	Storytelling (Oh et al., 2023)	40.01	0.8501	28.94	0.8163	32.63	0.8201	33.31	0.8126
✓	✓	–	Persona (Ge et al., 2024)	39.05	0.8460	28.15	0.8260	33.15	0.8224	31.91	0.8162
✓	✓	✓	Proposed	**43.91**	**0.8561**	**30.11**	**0.8295**	**34.78**	**0.8311**	**34.98**	**0.8170**

Table 3: Results of English-to-multilingual translation.

BLEU scores increased while COMET scores remained stable, suggesting that the method successfully aligns surface-level expressions with the stylistic characteristics of each work without sacrificing semantic adequacy. Furthermore, compared with conventional augmentation techniques, our approach consistently achieved higher BLEU scores, indicating its ability to produce translations that better reflect work-specific nuances. We also observed improved performance on "Elio", which was released after the model's training cutoff. This indicates that our method improves translation quality even for works with limited prior exposure to the model.

We also conducted experiments using *Qwen2.5-72B-Instruct*[5] for synthetic data generation and *Qwen2.5-7B-Instruct*[6] for instruction tuning; results are reported in Table 12 in Appendix G.

Table 2 shows an example of the translation outputs. Our method correctly interprets the idiomatic *scoop* in a newsroom rivalry context and preserves the canonical *Black Suit Spider-Man*, while conventional systems yield literal or out-of-context translations.

4.3 Cross-Language Robustness

To evaluate the cross-lingual applicability of our method, we conducted experiments using the movie "Spider-Man 3" with Finnish (FI), Spanish (ES), and German (DE) as target languages. As shown in Table 3, across all four language pairs, our proposed approach consistently outperformed conventional approaches, with statistically significant improvements. These results indicate the potential effectiveness of our one-shot translation and context-based augmentation approach in multilingual settings.

[5] https://huggingface.co/Qwen/Qwen2.5-72B-Instruct

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

Contextual augmentation	Reference guidance	BLEU	COMET
–	–	37.99	0.8344
✓	–	38.70	0.8423
–	✓	42.52	0.8441
✓	✓	**43.91**	**0.8561**

Table 4: Ablation study of proposed components.

4.4 Analysis

Unless otherwise noted, analyses in this section are primarily conducted on "Spider-Man 3" (EN→FR).

4.4.1 Impact of Contextual Augmentation and Reference Guidance

The proposed method consists of two key components: contextual augmentation and reference guidance. To evaluate the individual and combined effects of these components, we conducted an ablation study with four configurations: (1) no contextual augmentation and no reference guidance, (2) contextual augmentation only, (3) reference guidance only, and (4) both components enabled. Table 4 summarizes the results.

The findings indicate that both procedures contribute positively to translation performance. When combined, these components yield the highest BLEU and COMET scores, confirming their complementary roles in enhancing translation quality.

Source type	Constraint	BLEU	COMET
None	None	42.52	0.8441
Metadata	Genre	41.24	0.8458
Metadata	Title	41.91	0.8433
General	Persona	43.04	0.8548
Context	Scene	42.52	0.8441
Context	Character	**43.91**	**0.8561**
Context	Scene + Character	42.90	0.8548

Table 5: Translation performance under different generation constraints.

Original	I'd like to sing on stage for the rest of my life with you in the first row.
Constraint	Title ("Spider-Man 3")
Output	I'd love to swing through Manhattan for the rest of my life with you watching from the ground below.
Constraint	Scene (A scene where a woman reminisces about a past romantic encounter with her uncle's planning of a proposal.)
Output	I'd love to see him propose to me in a garden for the rest of my life with our family in the front seats.
Constraint	Character (A person who is fiercely loving and takes their time in relationships, as seen in the character of M.J.)
Output	I'd love to grow old together with you, hand in hand, and cherish every moment we share.

Table 6: Comparison of generated outputs with different constraints.

4.4.2 Comparison of Different Contextual Information Sources

We conducted an ablation study to examine which types of constraint information are most effective when guiding data augmentation. We compared (i) metadata (genre and title), (ii) pre-defined personas extracted from Ge et al. (2024) , and (iii) contextual cues obtained directly from the subtitles (character profiles, scene descriptions, and a random combination of both characters and scenes).

Table 5 shows the results. Overall, adding metadata did not consistently outperform the no-keyword setting, indicating that metadata alone is not a reliable driver of quality gains. In contrast, using character information extracted by our proposed method yielded higher performance than using general personas, suggesting that contextual constraints are more effective for subtitle translation. Furthermore, the variant of our method that uses scene descriptions did not surpass the character-only setting.

Table 6 analyzes generated texts. Across Title, Scene, and Character constraints, the rewriting largely preserved the original sentence structure. However, scene-based augmentation tended to over-condition on scene cues, yielding semantically incongruent sentences and likely contributing to lower BLEU than character-based augmentation.

Reference guidance	Method	Human evaluation		
		Accuracy	Fluency	Style
–	Paraphrase	3.79	4.08	3.56
✓	Metadata (Title)	3.97	**4.21**	3.80
✓	Proposed	**4.02**	4.17	**3.95**

Table 7: Human evaluation results.

4.4.3 Human Evaluation

We conducted a human evaluation following the MQM–DQF framework (Accuracy, Fluency, Style; 5-point scales) with three native French raters. For Style, they judged preservation of character-specific traits as well as overall tone and narrative consistency. We assessed inter-rater reliability via the intraclass correlation coefficient $\mathrm{ICC}(2, k)$ (McGraw and Wong, 1996), obtaining values of 0.737 for Accuracy, 0.710 for Fluency, and 0.728 for Style. The full rater instructions, recruitment details, and payment information are provided in Appendix H.

Table 7 reports the average scores across evaluators for each method. The results show that our proposed method achieved the highest scores in Accuracy and Style. This suggests that our approach is particularly effective in maintaining speaker individuality and narrative tone, which are critical for subtitle translation tasks.

4.4.4 Multi-Work Instruction Tuning

To examine whether instruction tuning on multiple works improves generalization, we compared two settings: (i) a model trained using data from a single work (Individual) and (ii) a model trained using combined data from four works (Multi). Table 8 summarizes the results.

Although not statistically significant, Multi consistently achieved higher COMET scores than Indi-

Model	Spider-Man 3		The Duchess		The Princess and the Frog		Elio	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
Individual	**43.91**	0.8561	49.11	0.8539	33.97	0.7822	**24.57**	0.7451
Multi	43.17	**0.8586**	**49.90**	**0.8579**	**34.24**	**0.7882**	24.25	**0.7453**

Table 8: Translation performance: Individual (trained on 1 work) vs Multi (trained on 4 works).

Method	Number of parallel data samples (Data augmentation rate)									
	10k (25.40)		20k (50.90)		40k (101.78)		60k (152.67)		80k (203.56)	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
Paraphrase	37.54	0.8412	**38.95**	**0.8430**	37.99	0.8405	38.50	0.8429	38.30	0.8386
Proposed	39.04	0.8411	41.16	0.8499	42.90	0.8554	**44.40**	**0.8602**	43.90	0.8579

Table 9: Translation performance across various expansion conditions.

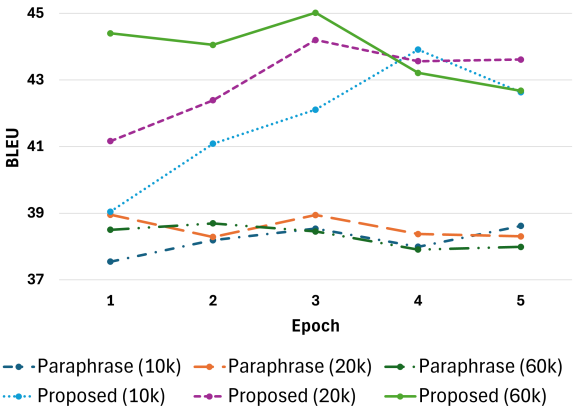

Figure 3: Comparison of BLEU scores across training epochs as the number of augmented sentences.

vidual. This trend suggests potential generalization benefits from incorporating diverse writing styles and contextual information.

4.4.5 Expansion Rate Analysis

To evaluate the scalability of our proposed method using character profiles, we conducted a series of experiments comparing translation performance across various expansion rates. We used Paraphrase as the baseline. For all expansion rates, we used the same sets of contextual descriptions, and only increased the number of generated variations per sentence. Because of the variation in the number of training samples caused by different expansion rates, it was difficult to compare performance using a fixed number of training epochs. We adopted two evaluations: translation performance after one epoch of training for each expansion rate (Table 9) and performance trends across multiple epochs using each dataset (Figure 3).

In the first evaluation, we observed that the baseline method plateaued in performance beyond 20,000 samples, whereas the proposed method continued to improve up to 60,000 samples. This indicates that our method possesses superior scalability and generalization capability.

In the second evaluation, we analyzed performance trends across multiple epochs. The baseline method showed little improvement, even as the number of training epochs increased. In contrast, the proposed methods using 10,000 and 20,000 samples did not reach peak performance in the first epoch but continued to improve with additional training. Although the optimal number of epochs varied with the expansion rate, performance consistently surpassed that of the first epoch under all conditions. These findings indicate that maximizing the utility of expanded samples can enhance the effectiveness of limited bilingual data in resource-constrained settings.

5 Conclusion

We proposed a parallel text data augmentation method for improving subtitle translation using LLMs. Our approach leverages contextual information, such as character profiles, extracted from the original source text, rather than relying on metadata such as the title. Experimental results show that combining source-side augmentation using contextual descriptions with bilingual data generation via one-shot learning improves translation performance while preserving the work's tone.

These improvements were consistent across various model architectures and language pairs, which confirms the robustness of our approach. It also showed effectiveness on works released after the model's training cutoff. Future work includes extending our approach to broader domains, such as multi-episode television dramas, where maintaining tone across episodes is critical.

Limitations

This study has several limitations.

First, the approach is tailored to dialogue-driven media and has not been verified for other formats such as narration-based content, including documentaries or educational materials.

Second, the models used for data generation and instruction tuning differ in parameter size. This discrepancy was due to hardware limitations, which made instruction tuning with the larger model infeasible. Moreover, for the same reason, all experiments were conducted only once per condition, without averaging across multiple runs or random seeds.

Finally, like other LLM-based approaches, our method carries a risk of hallucination. Since the training data is synthetically generated using LLMs, there is a possibility that hallucinated or contextually inaccurate sentences are included in the instruction tuning corpus. This may lead to translation outputs that contain hallucinated content.

Ethical Considerations

SubtitleMetaData dataset was obtained from publicly released films that have undergone a formal review process by experts. Based on this review process and the public availability of the films, we believe the risk of including personally identifiable information or offensive content is minimal.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, and 7 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. In-context examples selection for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8857–8873, Toronto, Canada. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, and Tom Henighan et al. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Trevor Cohn and Mirella Lapata. 2007. Machine translation by triangulation: Making effective use of multi-parallel corpora. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 728–735. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 567–573. Association for Computational Linguistics.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *Preprint*, arXiv:2406.20094.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 9 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Prabhakar Gupta, Mayank Sharma, Kartik Pitale, and Keshav Kumar. 2019. Problems with automating translation of movie/tv show subtitles. *Preprint*, arXiv:1909.05362.

Sixin Liao, Jan Louis Kruger, and Stephen Doherty. 2020. The impact of monolingual and bilingual subtitles on visual attention, cognitive load, and comprehension. *Journal of specialised translation*, (33):70–98.

Philip Lippmann, Konrad Skublicki, Joshua Tanner, Shonosuke Ishiwatari, and Jie Yang. 2025. Context-informed machine translation of manga using multimodal large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3444–3464, Abu Dhabi, UAE. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

Kenneth O. McGraw and Seok P. Wong. 1996. Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1(1):30–46.

Yasmin Moslem, Rejwanul Haque, John Kelleher, and Andy Way. 2022. Domain-specific text generation for machine translation. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 14–30. Association for Machine Translation in the Americas.

Seokjin Oh, Su Ah Lee, and Woohwan Jung. 2023. Data augmentation for neural machine translation using generative language model. *Preprint*, arXiv:2307.16833.

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. Association for Computational Linguistics.

Renhao Pei, Yihong Liu, Peiqin Lin, François Yvon, and Hinrich Schuetze. 2025. Understanding in-context machine translation for low-resource languages: A case study on Manchu. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8767–8788, Vienna, Austria. Association for Computational Linguistics.

Parinthapat Pengpun, Can Udomcharoenchaikit, Weerayut Buaphet, and Peerat Limkonchotiwat. 2024. Seed-free synthetic data generation framework for instruction-tuning LLMs: A case study in Thai. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 445–464. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191. Association for Computational Linguistics.

Ashmari Pramodya, Yusuke Sakai, Justin Vasselli, Hidetaka Kamigaito, and Taro Watanabe. 2025. Translating movie subtitles by large language models using movie-meta information. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 315–330, Vienna, Austria. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702. Association for Computational Linguistics.

Stefan Riezler and John T. Maxwell. 2005. On some pitfalls in automatic evaluation and significance testing for MT. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 57–64. Association for Computational Linguistics.

Jonne Sälevä and Constantine Lignos. 2024. Language model priors and data augmentation strategies for low-resource machine translation: A case study using Finnish to Northern Sámi. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12949–12956. Association for Computational Linguistics.

Joan Santoso, Patrick Sutanto, Billy Cahyadi, and Esther Setiawan. 2024. Pushing the limits of low-resource NER using LLM artificial data generation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9652–9667. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. *Preprint*, arXiv:2109.01652.

Chenxi Whitehouse, Monojit Choudhury, and Alham Fikri Aji. 2023. LLM-powered data augmentation for enhanced cross-lingual performance. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 671–686, Singapore. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Francis Zheng, Edison Marrese-Taylor, and Yutaka Matsuo. 2024. Improving low-resource machine translation for formosan languages using bilingual lexical resources. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11248–11259. Association for Computational Linguistics.

Shaolin Zhu, Menglong Cui, and Deyi Xiong. 2024. Towards robust in-context learning for machine translation with large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16619–16629, Torino, Italia. ELRA and ICCL.

A Prompt for Contextual Cues Extraction

A.1 Character Feature Extraction

The following passage is a subtitle for a film. Choose one character from the passage and describe its personality in one line using the form "A person who..."
Input: `{segmented sentences}`
Output:

A.2 Scene Feature Extraction

The following sentences are subtitles for a movie.
Please explain in one line which scene these sentences are about, using the form "A scene where..."
Input: `{segmented sentences}`
Output:

B Prompt for Context-Driven Rewriting

`{character}` and `{scene}` refer to the contextual descriptions, which were extracted in Appendix A.

B.1 Character-Based

You're a talented movie subtitler. Please follow the steps below.
The following English text is a subtitle for a movie. Rewrite the entire content so that it sounds like it's addressing someone described as '`{character}`'.
Maintain the speaker's tone and personality as faithfully as possible. Focus on creating a sentence that fits naturally within the new context, even if its meaning differs from the original. You may substitute the subject, verb, and other elements to suit the new character, but keep the overall sentence structure, flow, and phrasing intact. If the sentence cannot be naturally adapted to the given context, it is acceptable to leave it unchanged.
Output only the rewritten sentence. Do not add any annotations or additional information.
Input: `{src}`
Output:

B.2 Scene-Based

You're a talented movie subtitler. Please follow the steps below.
The following English text is a subtitle for a movie. Rewrite the entire content so that it sounds like it would be spoken in the context of "`{scene}`".
Maintain the speaker's tone and personality as faithfully as possible. Focus on creating a sentence that fits naturally within the new context, even if its meaning differs from the original. You may substitute the subject, verb, and other elements to suit the new scene, but keep the overall sentence structure, flow, and phrasing intact. If the sentence cannot be naturally adapted to the given context, it is acceptable to leave it unchanged.
Output only the rewritten sentence. Do not add any annotations or additional information.
Input: `{src}`
Output:

B.3 Scene- and Character-Based

You're a talented movie subtitler. Please follow the steps below.
The following English text is a subtitle for a movie. Rewrite the entire content so that it sounds like it would be spoken in the context of "`{scene}`", and directed toward a person described as: "`{character}`" Maintain the speaker's tone and personality as faithfully as possible. Focus on creating a sentence that fits naturally within the new context, even if its meaning differs from the original. You may substitute the subject, verb, and other elements to suit the new scene, but keep the overall sentence structure, flow, and phrasing intact. If the sentence cannot be naturally adapted to the given context, it is acceptable to leave it unchanged.
Output only the rewritten sentence. Do not add any annotations or additional information.
Input: `{src}`
Output:

Movie title	Genres	Subtitle count
Spider-Man 3	Action, Adventure, Sci-Fi	787
The Duchess	Biography, Drama, History, Romance, Thriller	880
The Princess and the Frog	Animation, Adventure, Comedy, Family, Fantasy, Musical, Romance	968
Elio	Family, Comedy, Adventure, Animation, Science Fiction	724

Table 10: Genres and subtitle count for each movie.

Movie title	Scene	Character
Spider-Man 3	**Input:** I've come a long way from being the boy who was bit by a spider. Back then, nothing seemed to go right for me. Now That's okay. The city is safe and sound. **Extracted:** A scene where the main character, likely Spider-Man, reflects on his past and how his life has improved since gaining his powers and becoming a hero.	**Input:** I've come a long way from being the boy who was bit by a spider. Back then, nothing seemed to go right for me. Now that's okay. The city is safe and sound. **Extracted:** A person who has overcome adversity and become confident in their abilities.
The Duchess	**Input:** Lord Robert. Lord Walter. You'd better not let me down, Charles Grey. I've got 20 guineas riding on you. Come on, ladies. **Extracted:** A scene where a high-stakes horse racing event is about to take place with the characters placing bets and cheering on their favorites.	**Input:** Of course, Lady Spencer. She's an accomplished lady of quality and devoted to her duties. She's fluent in French, Italian, Latin and fully versed in horsemanship and dancing and Yes, I'm aware of all that. She's a credit to you. **Extracted:** A person who is accomplished, devoted, and well-versed in various skills, as exemplified by Lady Spencer.
The Princess and the Frog	**Input:** Here comes my favorite part. Then, the frog was transformed into a handsome prince. Say good night, Tiana. Is that so? Well, here's your Prince Charming, Tia. **Extracted:** A scene where the main character is watching a romantic movie and the story within the movie unfolds with a magical transformation and a romantic encounter.	**Input:** Kiss him, kiss him, kiss him! I won't, I won't, I won't! I would kiss a hundred frogs if I could marry a prince and be a princess. You girls, stop tormenting that poor little kitty. Poor little thing. **Extracted:** A person who is compassionate and caring towards animals, as evidenced by their concern for the well-being of the kitty.
Elio	**Input:** Is life really out there, Tía Olga? What? Come on. You can't run away like that, Elio. Okay. **Extracted:** A scene where a character is confused or surprised and asks for clarification.	**Input:** Do you remember what your mom used to order for you? I'm trying to get him to eat. I heard about your brother and his wife. I'm so sorry. But it's– But it's been great Right, buddy? **Extracted:** A person who is nostalgic and sentimental about their childhood memories.

Table 11: Extracted scene and character features from input subtitles.

Instruction tuning	Data augmentation	Reference guidance	Method	Llama3 series		Qwen2.5 series	
				BLEU	COMET	BLEU	COMET
–	–	–	Base model	37.44	0.8412	33.50	0.8080
✓	–	–	Original parallel data only	36.72	0.8388	33.85	0.8139
✓	✓	–	Paraphrase	37.99	0.8344	36.79	0.8253
✓	✓	–	Storytelling (Oh et al., 2023)	40.01	0.8501	36.76	0.8307
✓	✓	–	Persona (Ge et al., 2024)	39.05	0.8460	37.27	0.8290
✓	✓	✓	Proposed	**43.91**	**0.8561**	**38.36**	**0.8309**

Table 12: Robustness across model families.

C Prompt for Generating Target Language Texts Using One-Shot Learning

You are an experienced film subtitle translator. Please follow the instructions below.
The following English text is a subtitle for a movie. Your task is to translate the following English sentence into French, strictly following the style of the reference example provided below.
The translation must:

- Faithfully reproduce the tone, structure, and phrasing of the reference example.
- Avoid any additional commentary, explanation, or formatting.
- Output only the translated sentence, and nothing else.

English Sentence: `{augmented source sentences}`
Reference Example: `{human-written target text}`
Translated Sentence:

D Data Details

Table 10 summarizes the genre and number of subtitles for each movie included in the study. For our experiments, the subtitles of each movie were evenly split into two halves, with the first half used for training and the second half for evaluation. When the total number of subtitle sentences is odd, the training set is assigned one more sentence than the evaluation set.

E Hyperparameters of Instruction Tuning

We used the following hyperparameters for instruction tuning with QLoRA: LoRA rank = 8, LoRA alpha = 16, batch size = 32, learning rate = 2e-4, and 4 training epochs.

F Extracted Contextual Descriptions

Table 11 shows examples of contextual descriptions extracted from subtitle texts. contextual descriptions extracted without an explicitly provided title appear to reflect latent knowledge about the work.

G Model Robustness

We conducted the same set of experiments in Section 4.2 with the Qwen2.5 series, using *Qwen2.5-72B-Instruct* for synthetic data generation and *Qwen2.5-7B-Instruct* for instruction tuning. The prompts, augmentation pipeline, and training hyperparameters followed the Llama-based setup.

Table 12 reports the Qwen results of "Spider-Man 3" (EN→FR). Overall, the trends were consistent with the Llama series: the proposed context-guided source augmentation combined with one-shot structural referencing yielded higher BLEU while maintaining COMET, outperforming baselines across the evaluated films.

H Human Evaluation Protocol

H.1 Instructions Given to Raters

Task. Human evaluation of movie subtitle translations (English → French).

Items. 393 English source sentences, each with 3 machine translation outputs.

Procedure. For each English→French translation, provide *absolute ratings* based on the DQF–

MQM framework[7] on:

(1) **Accuracy**: 5-point Likert scale (1 = worst, 5 = best).

(2) **Fluency**: 5-point Likert scale (1 = worst, 5 = best).

(3) **Style (character-specificity)**: 5-point Likert scale (1 = worst, 5 = best).

Notes.

- Provide numeric ratings only; you do not need to create or correct translations.
- You may consult the provided reference translation to interpret the data.
- For item (3) Style, please add best-effort notes on what was good or bad for each item.

H.2 Rater Recruitment and Payment

H.2.1 Recruitment

Raters were recruited and coordinated via a *contracted professional annotation company*. The vendor selected native French speakers with experience in language evaluation according to their internal qualification policy.

H.2.2 Payment

A lump-sum payment was made to the vendor. The authors do not have access to individual annotators' compensation amounts. The vendor confirmed that compensation to annotators complies with local labor regulations and the vendor's internal wage policies for the country of operation.

H.2.3 Privacy

No personally identifying information was shared with the authors; only anonymized ratings and notes were delivered. This study posed minimal risk and did not collect sensitive personal data.

I Computational Resources

The total computation time was approximately 2,870 hours in total across all GPUs.

I.1 Training Data Generation

We generated synthetic training data on NVIDIA A100 PCIe 40 GB GPUs, each with a thermal design power (TDP) of 250 W. The generation process required approximately 20 hours of computation time per 10,000 augmented sentences.

[7] https://globalization.co.jp/resource/taus-dqf/dqf-error-typology/

I.2 Instruction Tuning

Our training process required approximately 2 hours of computation time per 10,000 augmented sentence pairs on the same A100 PCIe hardware.

J License

In our study, we used the *SubtitleMetaData* dataset. This dataset is constructed from publicly available sources: OpenSubtitles (licensed under GNU General Public License v3.0, GPLv3), Wikipedia summaries (licensed under Creative Commons Attribution-ShareAlike 3.0, CC BY-SA 3.0), and IMDb metadata, which was accessed for non-commercial research purposes in accordance with IMDb's Terms of Use.

We used the *Llama-3.3-70B-Instruct* and *Llama-3-8B-Instruct* models provided by Meta. We also used the *Qwen2.5-72B-Instruct* and *Qwen2.5-7B-Instruct* models released by Alibaba Cloud. These models were used solely for academic research, in compliance with their respective licenses.