Speaker Identification and Dataset Construction Using LLMs: A Case Study on Japanese Narratives

Seiji Gobara, Hidetaka Kamigaito, Taro Watanabe
Nara Institute of Science and Technology
{gobara.seiji.gt6, kamigaito.h, taro}@is.naist.jp

Abstract

Speaker identification in narrative analysis is a challenging task due to complex dialogues, diverse utterance patterns, and ambiguous character references. Cosly and time-intensive manual annotation limits the scalability of high-quality dataset creation. This study demonstrates a cost-efficient approach of constructing speaker identification datasets by combining small-scale manual annotation with LLM-based labeling. A subset of data is manually annotated and is used to guide LLM predictions with a few-shot approach followed by refinement through minimal human corrections. Our results show that LLMs achieve approximately 90% accuracy on challenging narratives, such as the "Three Kingdoms" dataset, underscoring the importance of targeted human corrections. This approach proves effective for constructing scalable and cost-efficient datasets for Japanese and complex narratives.

1 Introduction

Narrative analysis is essential for understanding cultural values, psychological dynamics, and creative processes. Examining narrative structures and themes provides valuable insights into societal norms and human behavior (Piper et al., 2021). Large language models (LLMs) (Zhao et al., 2023a) have introduced new possibilities in narrative analysis, enabling tasks such as character emotion analysis and plot progression prediction.

Speaker identification, a key task in narrative analysis, involves accurately attributing dialogue to characters and understanding character dynamics within a story. However, constructing high-quality speaker identification datasets is costly and labor-intensive, requiring consistency and attention to paraphrase variations (Elson and McKeown, 2010; He et al., 2013; Muzny et al., 2017; Chen et al., 2019a; Vishnubhotla et al., 2022).

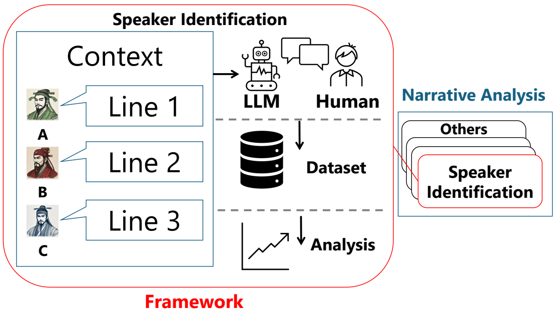

Figure 1: Method for constructing a dataset through collaboration between LLMs and human annotators for speaker identification in narrative analysis.

To address these challenges, we employ a collaborative approach to dataset construction, combining LLM-based initial annotations with targeted manual corrections (Tan et al., 2024). This significantly reduces annotation costs while maintaining quality. Inspired by the PDNC dataset (Vishnubhotla et al., 2022), we annotate both primary speaker names and their paraphrased forms (aliases). This dual annotation improves efficiency and flexibility. Figure 1 outlines our framework: LLM predictions, followed by iterative human correction, encompassing dialogue extraction, speaker labeling, and refinement.

Existing speaker identification datasets have primarily focused on English and Chinese, limiting the scope of research to these languages. To address this, we first constructed a speaker identification dataset for the Japanese narrative "Romance of the Three Kingdoms", a Japanese translation of the original Chinese work, chosen for its complex plot and character interactions, leveraging data from Aozora Bunko[1]. This method demonstrated the feasibility of creating high-quality datasets with reduced annotation costs.

Our results show that LLMs achieve approximately 90% accuracy, even without human corrections, while human intervention further enhances

[1] https://www.aozora.gr.jp/

Proceedings of the The 7th Workshop on Narrative Understanding, pages 97–119
May 4, 2025 ©2025 Association for Computational Linguistics

accuracy. Additionally, this approach significantly lowers the cost of dataset creation, making it scalable for larger and more diverse datasets. We also highlight the critical role of contextual input length in improving LLM performance, providing valuable insights for handling complex narratives.

2 Related Work

2.1 Dataset Construction

Elson and McKeown (2010) annotated speaker names and genders in 11 English narratives from the 19th century. He et al. (2013) treated separated lines in *Pride & Prejudice* as a single utterance for annotation. Muzny et al. (2017) expanded these datasets, creating the QuoteLi3 dataset, which includes annotations for all utterances in three narratives. Chen et al. (2019a) annotated utterances in the Chinese narrative World of Plainness (WP). Vishnubhotla et al. (2022) developed the Project Dialogism Novel Corpus (PDNC), annotating speakers, addressees, quote types, referring expressions, and mentions across 28 English novels, including main names and their variations.

Despite these advancements, existing datasets are primarily limited to English or Chinese, with no publicly available datasets for Japanese. Moreover, since these datasets depend on manual labor for annotation, they are inherently labor-intensive and costly to produce.

2.2 Speaker Identification

Feature-Based Approaches Several studies have employed linguistic features and manually crafted attributes for speaker identification (Elson and McKeown, 2010; He et al., 2013; Bamman et al., 2014; Muzny et al., 2017).

Deep Learning Approaches With the advent of deep learning, more advanced methods for speaker identification have emerged. These include approaches that fine-tune models such as BERT (Bidirectional Encoder Representations from Transformers; (Devlin et al., 2019)), BART (Lewis et al., 2020) for speaker identification tasks (Cuesta-Lazaro et al., 2022; Vishnubhotla et al., 2023), and prompt tuning techniques with models such as GPT-3.5 (Ouyang et al., 2022) which have also demonstrated high accuracy on the Chinese WP dataset (Su et al., 2024).

Despite these advances, limitations remain, particularly regarding the size of the context window. Michel et al. (2024) demonstrated that while LLaMA-3 (Dubey et al., 2024) expanded the context window and improved accuracy on the PDNC, their evaluation was constrained by the range of models and languages, leaving it incomplete.

3 Methods

Task Definition Speaker identification in narrative analysis involves determining which character or entity is responsible for a given utterance. This process requires analyzing both the utterance and its context to accurately attribute it to the correct speaker. In our approach, the set of possible speakers S is not predefined but derived from the context of the input text. Given a set of utterances $U = u_1, u_2, \ldots, u_m$, we establish a mapping function $f : U \rightarrow S$ so that each utterance $u_i \in U$ is correctly attributed to a speaker $s_j \in S$. We annotated two types of speaker names: the 'main name,' representing the most contextually appropriate identifier (e.g., Elizabeth Bennet), and 'candidates,' which include alternative names or alternative forms (e.g., Lizzy, Liz, Elizabeth). This dynamic speaker identification is crucial for capturing the fluid and complex nature of narrative interactions, enabling more accurate analysis of character relationships and narrative structure.

Refining Prompts and Manual Correction To cost-effectively create a high-quality speaker identification dataset, we manually annotated a small development set and refined prompt configurations for the LLM to generate speaker labels, which were then manually corrected. This approach ensured high data quality while minimizing costs. We also employed a specialized chat template[2] with a few-shot approach to enhance LLM performance (see Appendix I).

Robust Evaluation Metrics To ensure a robust evaluation of generation-based speaker identification systems like LLMs, we incorporated additional metrics such as substring match ratio and uncased evaluations. These metrics allow for a more relaxed and accurate assessment of speaker identification performance by accounting for variations in text, thereby improving the reliability of the evaluation results.

4 Dataset Construction

The dataset construction was carried out according to the following steps, as shown in Figure 2.

[2] https://github.com/chujiezheng/chat_templates

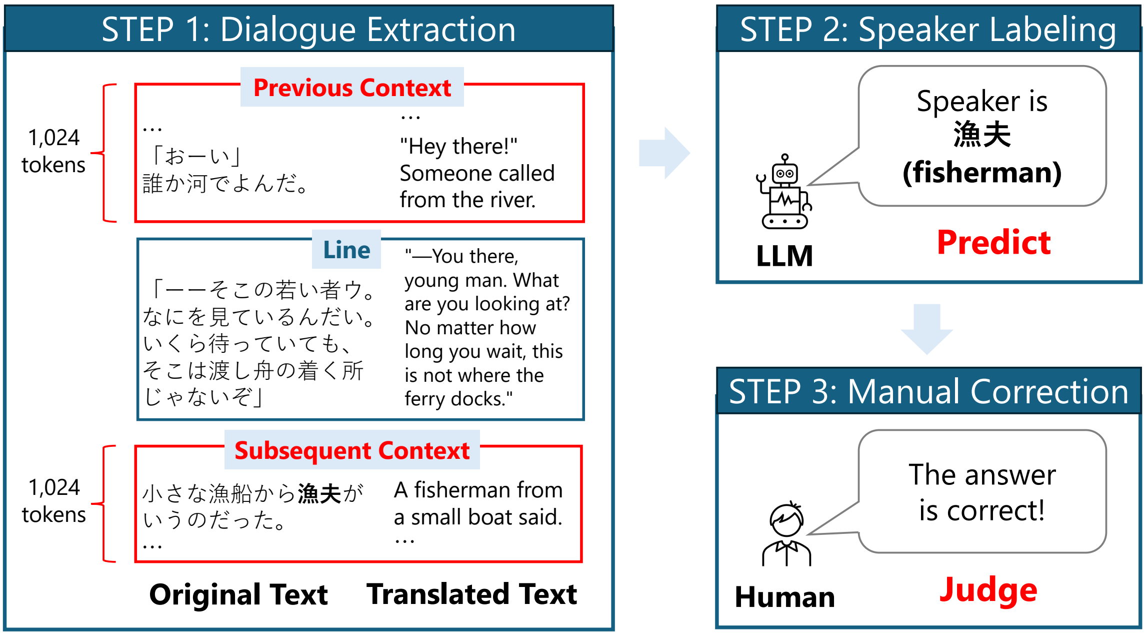
STEP 1: Dialogue Extraction
Previous Context
1,024 tokens
…
「おーい」
誰か河でよんだ。
…
"Hey there!"
Someone called from the river.
Line
「ーーそこの若い者ウ。なにを見ているんだい。いくら待っていても、そこは渡し舟の着く所じゃないぞ」
"—You there, young man. What are you looking at? No matter how long you wait, this is not where the ferry docks."
Subsequent Context
1,024 tokens
小さな漁船から漁夫がいうのだった。
…
A fisherman from a small boat said.
…
Original Text
Translated Text
STEP 2: Speaker Labeling
Speaker is 漁夫 (fisherman)
LLM
Predict
STEP 3: Manual Correction
The answer is correct!
Human
Judge

Figure 2: Workflow for constructing a speaker identification dataset using `LLaMA-3-70B-Instruct`. The process includes three steps: dialogue extraction, LLM-based labeling, and manual correction. LLM-generated labels are reviewed by human annotators—correct labels are retained, while errors are corrected.

STEP 1: Dialogue Extraction We gathered and tokenized dialogues from *Aozora Bunko's "Romance of the Three Kingdoms"* and Wikipedia sources by LLaMA-2 tokenizer and then extracting the surrounding 1,024-token contexts for each dialogue. This process resulted in a dataset of 16,423 instances. The dataset is composed of 10 books, with book_id=52410 serving as the development data, and book_id=52411 to 52420 serving as the evaluation data (see Appendix B).

STEP 2: Speaker Labeling We utilized an LLM to identify and label the speakers in the extracted dialogues. As the LLM, we used `LLaMA-3-70B-Instruct` with a few-shot setting, which showed the highest performance on the development dataset (see Appendix B and G).

STEP 3: Manual Correction We manually corrected the speaker names based on the annotation rules (see Appendix F.1) and adjusted approximately 20% of the identified labels. We excluded instances where the context lacked vocabulary corresponding to the speaker's name or involved multiple speakers in a single dialogue. This process removed 1,011 instances and finalized the dataset at 15,412 instances. We used GPU for 200 hours during inference (see Appendix H).

This method significantly reduced the time required to create evaluation data. While annotating 1,500 instances originally took 10 hours, focusing on correction tasks cut this time to 3.5 hours per 1,500 instances. Table 1 summarizes the tokens (LLaMA-2 and LLaMA-3 base models), lines, unique speakers, and skips for each book_id. The annotated speaker names include 856 unique speakers after excluding duplicates.[3]

4.1 Quality Assessment of Annotations

To verify the quality of the annotations, three independent annotators reviewed 100 samples from the evaluation dataset. They labeled the speaker names as "appropriate," "inappropriate," or "neutral," and we calculated the agreement rates for the "appropriate" labels. The results showed high consistency, with two annotators achieving an agreement rate of 0.97 and one annotator achieving an agreement rate of 0.96 (see Appendix F.2).

A comprehensive human evaluation under the exact same conditions as model inference would be prohibitively expensive. Manually reading the entire text, identifying the position of each input utterance, and determining the corresponding speaker are time-intensive and impractical at scale. In contrast, verifying whether a predicted speaker name is appropriate is relatively more manageable

[3] The datasets are available at https://huggingface.co/datasets/satoshi-2000/romance_of_the_three_kingdoms/.

		tokens	tokens			unique
book_id	Title	(Llama-2)	(Llama-3)	lines	skip	speakers
Excluded Data						
052409	Introduction	1,866	1,129	0	2	0
Development (dev) Data: Fully Human-Annotated						
052410	Oath of the Peach Garden	195,226	124,143	1,686	70	113
Evaluation (eval) Data: LLM-Labeled + Manual Correction						
052411	Stars of Destiny	195,589	124,772	1,662	108	157
052412	Heroes from the Grasslands	193,973	124,364	1,649	129	136
052413	The Way of the Minister	201,042	129,000	1,616	82	123
052414	Zhuge Liang	205,799	131,796	1,461	89	159
052415	The Battle of Red Cliffs	209,759	133,797	1,532	88	117
052416	Longing for Shu	204,514	130,989	1,598	83	153
052417	Plans for the South	222,992	143,735	1,433	95	171
052418	The Expedition	249,258	159,547	1,426	96	186
052419	The Battle of Wuzhang Plains	223,710	143,901	1,308	130	122
052420	Additional Records	27,050	16,968	40	40	26
Total		2,130,778	1,364,141	15,411	1,012	1,463

Table 1: Number of Tokens and Speakers by Dataset. The dataset was extracted and aligned based on token counts measured with the Llama-2 tokenizer, using 1,024 tokens as the standard segment length. `book_id=052409` represents the introductory chapter, setting the stage for the epic narrative of *Romance of the Three Kingdoms*. From the Oath of the Peach Garden (`book_id=052410`) to the final records of the Three Kingdoms (`book_id=052420`), the dataset follows the chronological progression of the story. `book_id=052410` served as development (dev) data, fully annotated by humans, while `book_id=052411-052420` were used as evaluation (eval) data, where initial LLM-generated labels were refined manually.

and can be done in a realistic timeframe. Therefore, we adopted this evaluation approach for human assessment, ensuring both feasibility and reliability while maintaining high annotation quality.

5 Experiment

To assess LLM capability in speaker identification and, simultaneously, to validate the quality of our constructed dataset, we conduct a series of experiments evaluating LLM performance. A primary aim of these experiments is to identify the characteristics of LLMs that facilitate efficient and effective dataset construction, allowing for the identification of optimal model features for similar tasks.

5.1 Prompt

As shown in Table 2, our approach employs a chat-based template to guide LLMs through the speaker identification task. By providing a few-shot prompt and assigning the LLM a system role, we effectively direct it through the necessary steps in a conversational format (see Appendix I).

5.2 Model

To compare model performance using LLMs, we selected LLaMA-3 (Dubey et al., 2024), a standard in LLM comparisons, along with Swallow-3 (Fujii, 2024), ELYZA-JP-8B (Hirakawa et al., 2024), and LLaMA-3-youko-8B (Mitsuda et al.), all based on LLaMA-3 with additional Japanese training. For broader model evaluation, we included Mistral 7B (Jiang et al., 2023) and RakutenAI-7B (Group et al., 2024), which, like Mistral 7B, are trained on Japanese data. To assess the impact of training data composition on accuracy, we selected CALM-3-22B (Ishigami, 2024), primarily trained on Japanese data, and Karakuri-8x7B (Inc., 2024), which uses the Mixture of Experts technique (Jiang et al., 2024) (see Appendix G).

5.3 Evaluation Metrics

We evaluated speaker attribution accuracy using the gold labels in the datasets of both languages:

Exact Match Ratio This metric, commonly used in prior research (Vishnubhotla et al., 2023; Michel et al., 2024), measures the percentage of exact matches between the speakers identified in the generated text and those in the annotations.

Substring Match Ratio Given the variations in texts generated by LLMs, this metric recognizes partial matches in key elements of the speaker names (see Appendix A).

BERTScore (Zhang* et al., 2020) This metric as-

Role	Content
user	Please guess who is speaking each line of dialogue in the following story (# Example Story) and provide only the speaker's name.
assistant	Understood. I will provide answers based on the story and dialogues below.
user	# Example Story `{Example Context}`
assistant	I have reviewed the story. Now, I will identify the speaker for each line of dialogue.
user	Who said the following line?
assistant	Please provide the line of dialogue.
user	Hey there!
assistant	Fisherman
user	— You there, young man. What are you looking at? No matter how long you wait, this is not where the ferry docks.
assistant	Fisherman
user	Thank you,
assistant	Young Man
user	Hey, hey, traveler.
assistant	Farmer
user	— What have you been waiting for since this morning in a place like this? Lately, there have been bandits called the Yellow Turbans around. The officials will get suspicious of you.
assistant	Farmer
user	Similarly, guess who is speaking each line of dialogue in the following story (# Target Story) and provide only the speaker's name.
assistant	Understood. I will provide answers based on the story and dialogues below.
user	# Target Story `{Context}`
assistant	I have reviewed the story. Now, I will identify the speaker for each line of dialogue.
user	Who said the following line?
assistant	Please provide the line of dialogue.
user	`{Line}`

Table 2: Prompts for Speaker Identification (Translated one). This table represents prompts designed for application in chat templates. The `{Context}` section contains the story content, while the `{Line}` section specifies the dialogue for which the speaker is to be identified. Appendix I shows the original Japanese text.

sesses similarity based on embeddings, capturing cases where surface expressions differ but the underlying meaning remains the same.

Edit Distance (Levenshtein et al., 1966) Edit distance calculates similarity by counting character insertions, deletions, and substitutions to transform one string into another.

5.4 Results

Overall Performance Table 3 shows the speaker identification accuracy for each model. Across both the dev (book_id=052410) and eval (book_id=052411–052420) phases, accuracy of approximately 90%, the models demonstrated robust performance in speaker identification (see Appendix B). The highest accuracy was achieved by a model that underwent continued pre-training on Japanese data using the base LLaMA-3 model, followed by instruction tuning. This combination proved particularly effective for speaker identification. The original LLaMA-3 model ranked second.

Additionally, `Swallow-3-8B-Instruct` showed a 5% improvement over `Swallow-3-8B`, highlighting the benefits of instruction tuning.

The results highlight the importance of combining high-quality datasets with large-scale models (e.g., 70B parameters) to achieve accurate speaker identification. Continued pre-training on Japanese data and instruction tuning not only ensure high accuracy but also reduce the cost of human corrections. This efficient and scalable method underscores the importance of leveraging well-trained large-scale models to balance accuracy and cost efficiency.

Accuracy by Book We analyzed the substring match ratio for each book_id to evaluate model accuracy, focusing on `LLaMA-3-70B-Instruct` as an example. This model consistently achieved approximately 0.9 accuracy across book_ids, as shown in Table 3, demonstrating robust performance in speaker identification.

In book_id=052419, the character "Sima Yi Zhongda" was labeled variably as "Sima Yi" or "Zhongda." Annotation rules prioritized the given name when present, leading to frequent use of "Zhongda." As a result, instances labeled as "Sima Yi" reflect the same individual, potentially skew-

Book ID	Swallow-3				Karakuri-8x7B	Mistral-7B	RakutenAI-7B	ELYZA-JP-8B	llama-3-youko-8B	LLaMA-3		CALM-3-22B
	8B	8B-Instruct	70B	70B-Instruct						8B-Instruct	70B-Instruct	
Exact Match Ratio												
052410	0.219	0.465	0.803	0.802	0.658	0.000	0.138	0.483	0.345	0.537	0.781	0.580
052411	0.222	0.582	0.835	0.829	0.687	0.000	0.108	0.540	0.310	0.537	0.824	0.507
052412	0.234	0.588	0.861	0.876	0.718	0.000	0.111	0.526	0.301	0.570	0.864	0.542
052413	0.240	0.621	0.887	0.892	0.744	0.000	0.126	0.593	0.313	0.593	0.849	0.547
052414	0.229	0.608	0.882	0.884	0.744	0.000	0.114	0.571	0.317	0.611	0.859	0.520
052415	0.238	0.582	0.873	0.871	0.706	0.000	0.139	0.536	0.343	0.555	0.839	0.543
052416	0.219	0.541	0.842	0.835	0.658	0.000	0.133	0.509	0.283	0.514	0.810	0.495
052417	0.228	0.584	0.866	0.871	0.719	0.000	0.109	0.537	0.278	0.603	0.865	0.505
052418	0.225	0.554	0.825	0.802	0.681	0.000	0.121	0.501	0.293	0.565	0.822	0.546
052419	0.193	0.476	0.735	0.727	0.617	0.000	0.098	0.469	0.239	0.499	0.728	0.426
052420	0.325	0.675	0.800	0.800	0.600	0.000	0.250	0.550	0.350	0.475	0.775	0.400
Substring Match Ratio												
052410	0.520	0.794	0.864	0.895	0.735	0.469	0.725	0.530	0.563	0.648	0.863	0.664
052411	0.536	0.795	0.892	0.918	0.745	0.510	0.705	0.589	0.555	0.649	0.916	0.610
052412	0.585	0.817	0.894	0.926	0.750	0.535	0.739	0.552	0.566	0.648	0.911	0.598
052413	0.582	0.827	0.906	0.925	0.759	0.502	0.728	0.618	0.546	0.666	0.880	0.605
052414	0.554	0.797	0.906	0.916	0.762	0.466	0.700	0.598	0.546	0.678	0.900	0.600
052415	0.567	0.790	0.891	0.896	0.717	0.456	0.698	0.555	0.519	0.623	0.866	0.589
052416	0.516	0.750	0.880	0.887	0.689	0.428	0.669	0.539	0.496	0.594	0.870	0.581
052417	0.549	0.792	0.897	0.912	0.739	0.486	0.721	0.569	0.539	0.687	0.914	0.572
052418	0.547	0.797	0.893	0.907	0.738	0.468	0.687	0.564	0.505	0.684	0.914	0.660
052419	0.479	0.684	0.797	0.806	0.664	0.417	0.635	0.518	0.455	0.609	0.808	0.539
052420	0.575	0.925	0.900	0.975	0.750	0.350	0.775	0.700	0.525	0.700	1.000	0.700
Edit Distance												
052410	7.751	1.543	0.446	0.476	0.845	10.423	6.837	1.432	5.852	2.705	0.620	4.240
052411	7.552	1.220	0.395	0.430	0.745	10.563	6.842	1.261	5.816	2.601	0.449	5.732
052412	7.155	1.178	0.321	0.301	0.191	11.091	6.735	1.421	6.127	2.646	0.320	5.179
052413	7.970	1.134	0.237	0.241	0.610	11.704	6.498	1.225	7.323	2.097	0.351	4.851
052414	7.949	1.162	0.265	0.277	0.704	11.260	6.903	1.386	6.602	2.086	0.369	5.307
052415	7.989	1.183	0.263	0.290	0.855	11.497	6.765	1.314	6.809	2.796	0.379	3.692
052416	8.243	1.377	0.362	0.406	0.885	11.538	7.342	1.406	6.869	2.857	0.489	5.267
052417	8.045	1.230	0.301	0.293	0.723	11.193	6.731	1.387	6.915	2.439	0.322	3.773
052418	7.735	1.262	0.431	0.531	0.893	11.250	6.608	1.426	6.996	2.705	0.500	4.211
052419	7.973	1.489	0.661	0.716	1.061	11.502	7.119	1.517	7.402	2.731	0.687	4.570
052420	8.925	1.025	0.475	0.475	1.225	11.150	4.375	1.300	5.150	3.500	0.525	5.475
BERTScore F1												
052410	0.792	0.888	0.959	0.958	0.923	0.676	0.772	0.706	0.812	0.877	0.950	0.879
052411	0.797	0.914	0.964	0.962	0.928	0.675	0.765	0.741	0.800	0.881	0.962	0.850
052412	0.809	0.918	0.970	0.974	0.936	0.675	0.768	0.699	0.797	0.886	0.972	0.864
052413	0.808	0.925	0.977	0.979	0.944	0.675	0.773	0.769	0.792	0.898	0.969	0.871
052414	0.810	0.924	0.976	0.976	0.944	0.682	0.770	0.764	0.803	0.904	0.971	0.861
052415	0.811	0.920	0.975	0.974	0.939	0.677	0.778	0.744	0.805	0.887	0.968	0.885
052416	0.794	0.906	0.967	0.966	0.926	0.671	0.762	0.744	0.789	0.875	0.960	0.856
052417	0.800	0.915	0.971	0.973	0.939	0.682	0.771	0.731	0.789	0.899	0.972	0.870
052418	0.813	0.917	0.965	0.961	0.932	0.685	0.776	0.732	0.794	0.893	0.965	0.875
052419	0.797	0.897	0.946	0.944	0.920	0.680	0.765	0.737	0.778	0.881	0.945	0.848
052420	0.809	0.939	0.956	0.960	0.908	0.664	0.825	0.853	0.817	0.860	0.960	0.825

Table 3: Performance metrics for all models (Exact Match Ratio, Substring Match Ratio, Edit Distance, BERTScore F1) evaluated across different books, highlighting variations by model category. The scores presented in the table are averaged values across the dataset. The background color gradient represents performance: darker red indicates higher performance, while darker blue indicates lower performance.

ing the evaluation for this book_id.

Relaxed Evaluation by Candidate Sets Using candidate sets for best matching enabled relaxed evaluation, enhancing accuracy. In book_id=52419, "Sima Yi Zhongda" appeared under various names, such as "Sima Yi" and "Zhongda." Per annotation rules, "Zhongda" was used when present in context, and "Sima Yi" otherwise. Both names could serve as main identifiers. Following PDNC (Vishnubhotla et al., 2023), we prepared interchangeable candidate sets for "Zhongda," including "Zhongda," "Sima Yi," "Sima Yi Zhongda," and "Sima Zhongda."

We then evaluated the predictions by matching them to the most corresponding name from these candidate sets. Compared to strict substring matching, this approach allowed for a more relaxed evaluation. For book_id=52419, the substring match ratio increased from 80.8% (without candidates) to 89.3% (with candidates), an improvement of 8.5%. This suggests that a relaxed strictness in the representation of speaker names leads to a more accurate and consistent evaluation (see Appendix K for details).

5.5 Analysis

Table 4 presents case study examples.

Case Study A: Long-Turn Dialogues The model generally identifies speakers accurately, even when relevant information is at the edges of the context. In Case A, although the model correctly

Case	Line	Excerpt Context	Pred	True
A	Hahaha.	Yang Biao, harboring his secret plan, returned to his residence. As soon as he arrived, he went into his wife’s room and said, "So, how is it these days? Do you often meet with Lady Guo? I hear you ladies frequently have various gatherings." Placing his hands gently on his wife’s shoulders, he spoke with an unusual tenderness. Yang Biao’s wife, puzzled, teased him, "What’s gotten into you today? You’ re never this sweet to me." "What’s the matter?" "Well, it’s just that you never act this way towards me normally." "Hahaha." "It actually makes me feel uneasy." "Is that so?"	Yang Biao	Yang Biao
B	Land of Jiangdong,	Wu is known as the "Land of Jiangdong," situated along the flow of the Great River.	Narration	Unknown
C	……	Diaochan, without showing any signs of agitation, immediately responded, "Yes. If it is the will of my lord, I am ready to give my life at any time." Wang Yun straightened his posture and said, "Then, I have something I wish to ask of you, trusting in your sincerity." "What is it?" "Dong Zhuo must be killed." "······" "If he is not removed, it will be as if the Han Emperor does not exist." "······"	Diaochan	Diaochan
D	The pleasures of life culminate here,	In the evening, a grand banquet was held with the slaughtering of cattle and horses for a feast. "The pleasures of life culminate here," said Guan Yu and Zhang Fei. "How could it end here? This is just the beginning," replied Xuande.	Guan Yu and Zhang Fei	Unknown
E	Lord Xuande, it is the fervent wish of both of us. Will you not consider it?	"It would be best." "Lord Xuande, it is the fervent wish of both of us. Will you not consider it?" From both sides,	Guan Yu	Guan Yu

Table 4: Case Study: ’Pred’ indicates the predicted speaker, ’True’ indicates the annotated speaker. Examples are translated into English; the original text is available in Appendix 5. Results are based on LLaMA-3-70B-Instruct, with unnecessary text removed via regular expressions.

attributed ’Hahaha.’ to Yang Biao, it erroneously attributed the subsequent line, ’Is that so?’, to his wife. This highlights the increased likelihood of errors in long-turn dialogues.

Case Study B: Narrator Identification We observed that the model correctly identifies the speaker as the narrator.

Case Study C: Silent Utterance Identification We confirmed the model demonstrated the ability to infer speaker names in implicit dialogues, “......” highlighting its contextual reasoning capabilities.

Case Study D: Multiple Speaker Identification The model successfully identified the speaker even in instances involving multiple speakers within the same utterance.

Case Study E: Data Leak We analyzed potential data leakage by comparing ELYZA-JP-8B and LLaMA-3-70B-Instruct predictions with an 8-context length. While LLaMA-3-70B-Instruct inferred speaker names from the context, ELYZA-JP-8B correctly predicted speakers not explicitly mentioned. For example, ELYZA-JP-8B mistakenly identified “Guan Yu” as a speaker, likely due to reliance on prior knowledge triggered by the mention of “Xuande”.

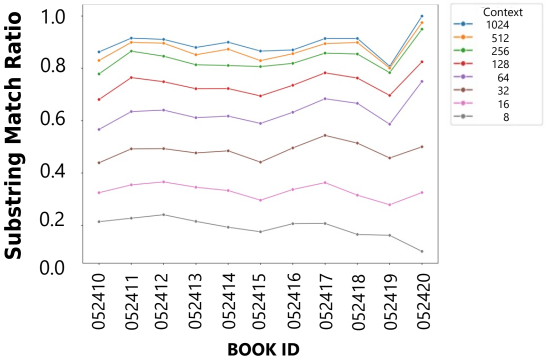

Figure 3: Variation in Substring Match Ratio by Context Length. This figure shows how the substring match ratio changes with different context lengths.

Impact of Varying Context Lengths As shown in Figure 3, the LLaMA-3-70B-Instruct model’s accuracy improves with longer context lengths but plateaus between 512 and 1,024 tokens. Models with smaller parameter sizes (8B or less) peaked at 512 tokens (see Appendix J).

This suggests that optimal context length depends on the model’s parameter size, reflecting its computational capacity and design. Selecting an appropriate context length is essential to maximize performance, especially in resource-limited settings (see Appendix B).

Impact of Context Masking We evaluated the effect of masking tokens within a 1,024 token context window on speaker identification accuracy. We tested the `LLaMA-3-70B-Instruct` model with mask ratios from 0% to 100% in 10% increments, replacing tokens with '<unk>'.

Figure 4 shows that the accuracy decreases as the Mask ratio increases. At 0% Mask, the model achieved 1.9% accuracy, which decreased as the Mask ratio increased. The `LLaMA-3-70B-Instruct` model's accuracy decreased with higher Mask ratios but still identified some speakers correctly. In contrast, the `ELYZA-JP-8B` model performed better at a 20% Mask ratio, indicating superior context retention. However, accuracy declined with excessive Masking due to reduced context. At 100% Mask, the `ELYZA-JP-8B` model achieved a 2.7% match rate, surpassing the `LLaMA-3-70B-Instruct` model's 1.9%. This suggests that the `ELYZA-JP-8B` model retains valuable contextual information even with full Masking (see Appendix E.2).

Extending Applicability Across Narratives To evaluate the applicability of our approach to different narratives and languages, we constructed a bi-lingual dataset comprising 14 diverse stories in Japanese and English. This dataset, sourced from Wikisource and Aozora Bunko, enabled us to analyze the `LLaMA-3-70B-Instruct` model' s performance across languages and cultural contexts.

Our analysis revealed that the model achieved higher accuracy on Japanese datasets, likely due to fewer variations in referring terms compared to English, which often includes synonyms for the same entity (e.g., Mother" and Woman"). This suggests the importance of designing candidate sets for consistent name recognition across languages. For further details on dataset construction and results, see Appendix C.

6 Conclusion

We collaborated with LLMs to create a speaker labeling dataset by annotating "Romance of the Three Kingdoms" from Aozora Bunko in Japanese. The dataset included 15,412 entries.

Using LLMs like LLaMA-3, we achieved a substring match ratio of approximately 90%. To handle multiple potential speakers, we developed a paraphrase dataset to improve evaluation accuracy.

Instead of manually annotating the entire dataset, we adopted an approach where LLMs performed the initial labeling, and human annotators focused on correcting the generated labels. This shift significantly reduced human labor costs while maintaining high annotation quality.

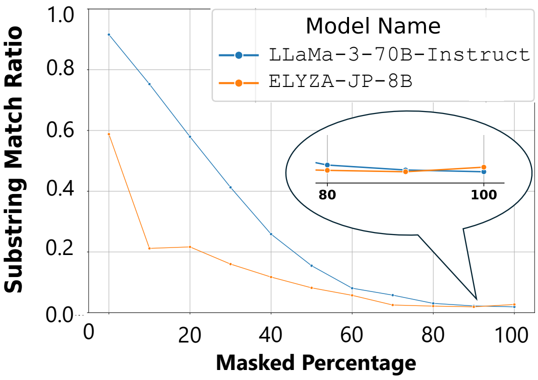

Figure 4: Substring Match Ratio by Mask Ratios for `LLaMA-3-70B-Instruct`. This figure shows how the substring match ratio changes as the proportion of masked tokens increases. The model demonstrates a gradual decline in accuracy with higher mask ratios, reflecting its dependency on contextual information.

Our findings demonstrate the potential of scalable, LLM-assisted methods for narrative analysis, offering a cost-effective solution for speaker identification in complex texts.

7 Future Plans

We will expand our datasets with advanced translation techniques and enhanced annotations, including Addressees and Quote Types, following the PDNC approach (Vishnubhotla et al., 2022). We also plan to refine speaker labeling methods and extend our analysis to complex stories with extensive character lists, improving LLMs' capacity for handling intricate narratives.

Our datasets also offer potential applications beyond speaker identification:

- **Character Interaction Analysis**: Exploring power dynamics, alliances, and conflicts in narratives.
- **Sentiment and Emotion Attribution**: Studying emotional tones associated with characters or interactions.
- **Cross-Cultural Studies**: Comparing storytelling across languages and cultures.
- **Education and Language Learning**: Teaching narrative structures and cultural contexts.

These applications highlight the versatility of our dataset, supporting both academic research and practical applications.

8 Limitations

Supported Languages This study primarily focuses on Japanese, with additional experiments conducted on a small-scale Japanese-English bilingual dataset. The English dataset was limited in size and scope, constraining the generalizability of the findings. While speaker identification performance in Japanese was strong, direct comparison with English posed challenges due to linguistic differences.

English narratives, with their diverse synonyms and alternative expressions, introduce variability that complicates direct comparisons to the contextually uniform nature of Japanese texts. Future work should expand datasets to address these linguistic differences. These differences may have influenced the results, underscoring the need for caution when evaluating bi-lingual performance. Future work should expand the dataset to include larger and more diverse bi-lingual samples, enabling more robust and comprehensive evaluations.

Models One of the objectives of this study is to demonstrate how high-quality datasets can be collaboratively created at a low cost using local LLMs without relying on APIs. While this approach highlights the potential of local models, the experiments were limited to models with a maximum size of 70 billion parameters. Comparisons with state-of-the-art models, such as GPT-4 (Achiam et al., 2023), which are accessible through APIs, remain unexplored.

Future work should include evaluations using more powerful models like GPT-4 to better understand the upper bounds of performance in speaker identification tasks. Additionally, it is worth noting that for Japanese tasks, certain models like ELYZA-JP-8B and Swallow-3 have been reported to perform at levels comparable to GPT-4 in specific scenarios, suggesting that sufficiently high-performance models are available for meaningful comparisons. However, given the steady improvement in the performance of local LLMs, we believe that our evaluations provide a reasonably comprehensive assessment within the scope of this study.

Translation In this study, we created a dataset translated using GPT-4o-mini for the purpose of bi-lingual evaluations. However, we only performed format checks on the translations (see Appendix D). To further enhance the quality of the dataset, human evaluation is deemed necessary.

Vulnerability to Tokenizer Limitations During dataset creation, some words may not be tokenized effectively, potentially impacting the quality of the extracted contextual information. To address this vulnerability to tokenizer limitations, future work could explore using alternative, more comprehensive tokenizers with larger vocabularies. This approach could mitigate the risk of data omissions stemming from inadequate tokenization, leading to more complete and reliable contextual representations within the dataset.

9 Assurance of Research Ethics

Explanation to Annotators We ensured adherence to research ethics by providing comprehensive explanations to the annotators about the study. Additionally, once the annotation was completed, we anonymized the collected data and paid careful attention to protecting personal information.

Licenses and Approvals Furthermore, we verified the licenses for the artifacts, obtained the necessary approvals, and confirmed that our usage complies with the intended purposes.

Potential Misuse Risks and Mitigation While our study focuses on the development of speaker identification datasets for narrative analysis, we acknowledge the potential risks associated with misuse of the generated datasets or data generation approach. For instance, speaker identification systems could be misused to monitor conversations or infringe on individual privacy if applied inappropriately. To mitigate such risks, we emphasize that our research is intended solely for academic purposes and large-scale narrative analysis, and not for surveillance or other unethical applications.

Transparency and Accountability Additionally, the datasets and methodologies are designed with transparency and accountability in mind, ensuring that their usage aligns with ethical standards.

Content Warning for Violent Expressions This dataset contains stories written several decades ago, during a period when violent expressions and provocative language, including depictions of murder and aggressive behavior, were more commonplace. Users are advised to exercise caution and be mindful of the potentially disturbing content when utilizing this dataset.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

David Bamman, Ted Underwood, and Noah A. Smith. 2014. A Bayesian mixed effects model of literary character. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 370–379, Baltimore, Maryland. Association for Computational Linguistics.

Lasse Bergroth, Harri Hakonen, and Timo Raita. 2000. A survey of longest common subsequence algorithms. In *Proceedings Seventh International Symposium on String Processing and Information Retrieval. SPIRE 2000*, pages 39–48. IEEE.

Jia-Xiang Chen, Zhen-Hua Ling, and Li-Rong Dai. 2019a. A Chinese Dataset for Identifying Speakers in Novels. In *Proc. Interspeech 2019*, pages 1561–1565.

Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Wang, and Jia-Bin Huang. 2019b. A closer look at few-shot classification. In *International Conference on Learning Representations*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Carolina Cuesta-Lazaro, Animesh Prasad, and Trevor Wood. 2022. What does the sea say to the shore? a BERT based DST style approach for speaker to dialogue attribution in novels. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5820–5829, Dublin, Ireland. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

David Elson and Kathleen McKeown. 2010. Automatic attribution of quoted speech in literary narrative. *Proceedings of the AAAI Conference on Artificial Intelligence*, 24(1):1013–1019.

Kazuki Fujii. 2024. Llama-3-swallow.

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual pre-training for cross-lingual llm adaptation: Enhancing japanese language capabilities. *Preprint*, arXiv:2404.17790.

Ulrich Germann. 2003. Greedy decoding for statistical machine translation in almost linear time. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, page 1 – 8, USA. Association for Computational Linguistics.

Rakuten Group, Aaron Levine, Connie Huang, Chenguang Wang, Eduardo Batista, Ewa Szymanska, Hongyi Ding, Hou Wei Chou, Jean-François Pessiot, Johanes Effendi, Justin Chiu, Kai Torben

Ohlhus, Karan Chopra, Keiji Shinzato, Koji Murakami, Lee Xiong, Lei Chen, Maki Kubota, Maksim Tkachenko, Miroku Lee, Naoki Takahashi, Prathyusha Jwalapuram, Ryutaro Tatsushima, Saurabh Jain, Sunil Kumar Yadav, Ting Cai, Wei-Te Chen, Yandi Xia, Yuki Nakayama, and Yutaka Higashiyama. 2024. Rakutenai-7b: Extending large language models for japanese. *Preprint*, arXiv:2403.15484.

Hua He, Denilson Barbosa, and Grzegorz Kondrak. 2013. Identification of speakers in novels. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1312–1320, Sofia, Bulgaria. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Masato Hirakawa, Shintaro Horie, Tomoaki Nakamura, Daisuke Oba, Sam Passaglia, and Akira Sasaki. 2024. elyza/llama-3-elyza-jp-8b.

KARAKURI Inc. 2024. KARAKURI LM 8x7B Instruct v0.1.

Ryosuke Ishigami. 2024. cyberagent/calm3-22b-chat.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre–training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Gaspard Michel, Elena V. Epure, Romain Hennequin, and Christophe Cerisara. 2024. A realistic evaluation of llms for quotation attribution in literary texts: A case study of llama3. *Preprint*, arXiv:2406.11380.

Koh Mitsuda, Xinqi Chen, Toshiaki Wakatsuki, and Kei Sawada. rinna/llama-3-youko-8b.

Grace Muzny, Michael Fang, Angel Chang, and Dan Jurafsky. 2017. A two-stage sieve approach for quote attribution. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 460–470, Valencia, Spain. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

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

Zhenlin Su, Liyan Xu, Jin Xu, Jiangnan Li, and Mingdu Huangfu. 2024. Sig: Speaker identification in literature via prompt-based generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19035–19043.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation: A survey. *Preprint*, arXiv:2402.13446.

Krishnapriya Vishnubhotla, Adam Hammond, and Graeme Hirst. 2022. The project dialogism novel corpus: A dataset for quotation attribution in literary texts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5838–5848, Marseille, France. European Language Resources Association.

Krishnapriya Vishnubhotla, Frank Rudzicz, Graeme Hirst, and Adam Hammond. 2023. Improving automatic quotation attribution in literary novels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 737–746, Toronto, Canada. Association for Computational Linguistics.

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

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023a. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Zirui Zhao, Wee Sun Lee, and David Hsu. 2023b. Large language models as commonsense knowledge for large-scale task planning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

A Substring Match Ratio Evaluation Method

The substring match ratio evaluates whether the true speaker name, as annotated, exists as a substring within the predicted speaker name. This evaluation metric is mathematically formalized as follows:

Definitions In a given dialogue dataset, we define the speaker names as follows:

- P_i: Predicted speaker name
- T_i: Annotated true speaker name

We define the match function M as:

$$M(P_i, T_i) = \begin{cases} 1 & \text{if there exists an integer } j \\ & \quad \text{such that } 0 \le j \le |P_i| - |T_i| \\ & \quad \text{and } P_i[j : j + |T_i|] = T_i \\ 0 & \text{otherwise} \end{cases}$$

Calculation of Substring Match Ratio The substring match ratio for the entire dataset is calculated as the proportion of dialogues where the true speaker name is a substring of the predicted speaker name. Formally, it is defined as:

$$r_s = \frac{1}{n} \sum_{i=1}^{n} M(P_i, T_i)$$

where $n \in \mathbb{N}$ is the total number of lines.

Calculation Steps

1. For each dialogue i, check if the true speaker name T_i is a substring of the predicted speaker name P_i.
2. Assign $M(P_i, T_i) = 1$ if T_i is a substring of P_i; otherwise, assign $M(P_i, T_i) = 0$.
3. Calculate the sum of all $M(P_i, T_i)$ values and divide by the total number of dialogues n.

Example Consider three dialogues with the following predicted and true speaker names:

- $P_1 =$ "John Smith", $T_1 =$ "John"
- $P_2 =$ "Alice", $T_2 =$ "Bob"
- $P_3 =$ "Charlie Brown", $T_3 =$ "Charlie"

The substring matches are calculated as follows:

$$M(P_1, T_1) = 1,$$
$$M(P_2, T_2) = 0,$$
$$M(P_3, T_3) = 1$$

Thus, the substring match ratio is calculated as:

$$r_s = \frac{1}{3}(1 + 0 + 1) = \frac{2}{3} \approx 0.67$$

Using the substring match ratio, we can evaluate how accurately the predicted speaker names contain the true speaker names as substrings.

Particularly, LLMs often generate unnecessary texts, such as special tokens like "[INST]" and unrelated tokens.

B Detailed Dataset Construction Process

Data Extraction The data was meticulously extracted from *Aozora Bunko's "Romance of the Three Kingdoms"* using the Huggingface datasets[4] library. This curated dataset includes furigana and metadata, and was selected for its extensive character list and the potential to extract complex relationships.

Development and Evaluation Sets The dataset was split into development and evaluation sets as follows:

- Volume 02: Peach Garden Oath (Shinjitai, Book ID: 52410) served as the development set.
- Volume 03: Among the Stars (Shinjitai, Book ID: 52411) to Volume 11: Wuzhang Plains (Shinjitai, Book ID: 52419) constituted the evaluation set.

[4] https://huggingface.co/datasets/globis-university/aozorabunko-clean

Token Count Variations Figure 5 shows the maximum input token count per book_id, confirming that the actual number of input tokens in this study falls within 8,192 tokens when converted using the LLaMA 3 Tokenizer. As illustrated in Figure 5, this study employed the LLaMA 2 Tokenizer to extract the preceding and following 1,024 tokens, thereby creating context tokens. Among the tokenizers used in the comparative models, the most commonly utilized base tokenizer was the LLaMA 3 Tokenizer.

Furthermore, Figure 6 demonstrates the variation in token count per index for book_id=052415, which had the highest number of input tokens. Excluding a few exceptionally long dialogue examples, almost all token counts were distributed around 2,250 tokens using the LLaMA 2 Tokenizer and around 1,500 tokens using the LLaMA 3 Tokenizer.

Reducing the length of the input context or randomly masking it was confirmed to significantly decrease identification accuracy (see Section 5.5 and Section 5.5). Therefore, to solve this task with high accuracy, it is necessary to process a sufficiently long context of at least 1,500 tokens using the LLaMA 3 Tokenizer.

This indicates that the number of tokens handled is extremely large compared to the methods used for evaluating the performance of existing LLMs, such as MMLU (Hendrycks et al., 2021) and Commonsense (Zhao et al., 2023b). By addressing this task, it is believed that we can measure the inference performance of LLMs with respect to long contexts.

Additionally, in this study, the dataset length was set to fit within the maximum input token count of 8,192 tokens, which is the limit for the models used in comparison. For identification tasks using similar methods, simply increasing the length of the input context or simultaneously targeting multiple lines for speaker identification could easily extend the evaluation to tasks requiring longer contexts, such as those involving 100,000 tokens.

Number of Tokens and Speakers Table 9 summarizes the number of tokens, utterances, and characters for each story.

In this table, “Tokens (LLaMA-3, JA)” and “Tokens (LLaMA-3, EN)” indicate the number of tokens in the Japanese and English versions of each story, respectively. Similarly, “Lines (JA)” and “Lines (EN)” represent the number of utterances in Japanese and English, respectively.

C Constructing a Bi-lingual Dataset via Crawling

Bi-lingual Dataset Creation To explore the applicability of this approach to other stories and languages, we expanded our research to include bilingual datasets developed from Wikisource[5] and Aozora Bunko, covering 14 diverse narratives in two languages. This approach offers a flexible and scalable framework for narrative analysis across various languages and cultural contexts, enhancing speaker identification by capturing the complexity of character references.

Bi-lingual Performance Figure 7 shows the substring match ratio for speaker identification using the `LLaMA-3-70B-Instruct` model on Japanese and English datasets. The model achieved higher accuracy on Japanese data, likely due to fewer label variations compared to English.

The Japanese dataset, composed mainly of simple folktales, exhibits fewer variations in referring terms. In contrast, the English dataset includes multiple synonyms for the same names, affecting the results. For example, the Japanese term “お母さん” in “matsuyama_kagami” is translated into various English terms, such as “Woman,” “Mother,” and “Wife”.

This suggests that, as noted in Section 5.4, preparing candidate sets for main names could reduce discrepancies. Additionally, to address case sensitivity issues in English, we introduced an Uncased Exact Match approach for more accurate evaluation (see Appendix L).

D Constructing a Bi-lingual Dataset via Translation

To broaden the applicability of our dataset and facilitate bilingual analysis, we translated the Japanese portions of *Romance of the Three Kingdoms* into English using the GPT-4o-mini model,[6] significantly reducing the time and cost associated with manual annotation.

This distinction clarifies that the bi-lingual datasets from Wikisource and Aozora Bunko use professional translations, while the "Romance of

[5] `https://wikisource.org/wiki/Main_Page`

[6] `https://platform.openai.com/docs/models` A smaller variant of GPT-4 with reduced computational requirements.

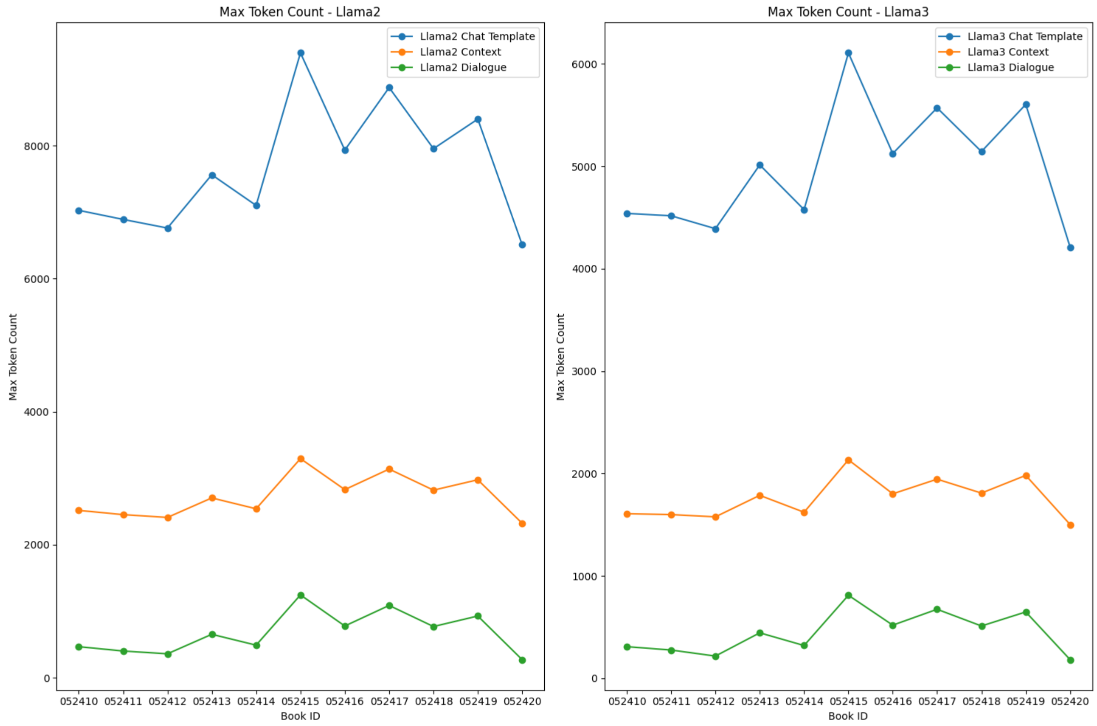

Figure 5: The Chat Template indicates the maximum token count when including tokens that control few-shots and prompt format. Context shows the maximum token count when inferring speaker names and combining the target dialogue with the preceding and following 1,024 tokens. Dialogue shows the maximum token count for the dialogue itself.

the Three Kingdoms" dataset relies on machine-translated content for exploratory purposes.

D.1 Translation Process and Quality Assurance

We followed a translation strategy similar to that used for speaker identification, employing few-shot prompts and incorporating failure cases for robustness (see Table 12). The translation covered 3,348 instances (book_id=052410, 052411), producing 1,574 entries for `book_id=052410` and 1,528 entries for `book_id=052411`.

We applied three main quality checks:

- **Language Accuracy:** Ensuring the translated text was correctly in English.
- **Dialogue Inclusion:** Confirming that each translated dialogue was present within the translated context.
- **Speaker Name Inclusion:** Verifying that translated speaker names appeared correctly in the translated context.

If any criterion was not met, we allowed up to five retries. Cases where the model responded with an inability message (e.g., “I’m sorry, but I can’t...”) were discarded. Additionally, for dialogues not found in the translated context, we employed the longest common subsequence algorithm (Bergroth et al., 2000) to match them with the closest translation. Only entries passing all checks were retained in the final dataset.

E Case Studies and Challenging Examples

E.1 Original Japanese Text of Case Study

Table 5 presents the original Japanese text of the case study discussed (see Section 5.5).

E.2 Further Case Study

Table 6 shows that `ELYZA-JP-8B` had already read these datasets during the training steps.

This finding indicates that the `ELYZA-JP-8B` model may have leveraged learned patterns or relationships to make accurate predictions even when the context is heavily Masked.

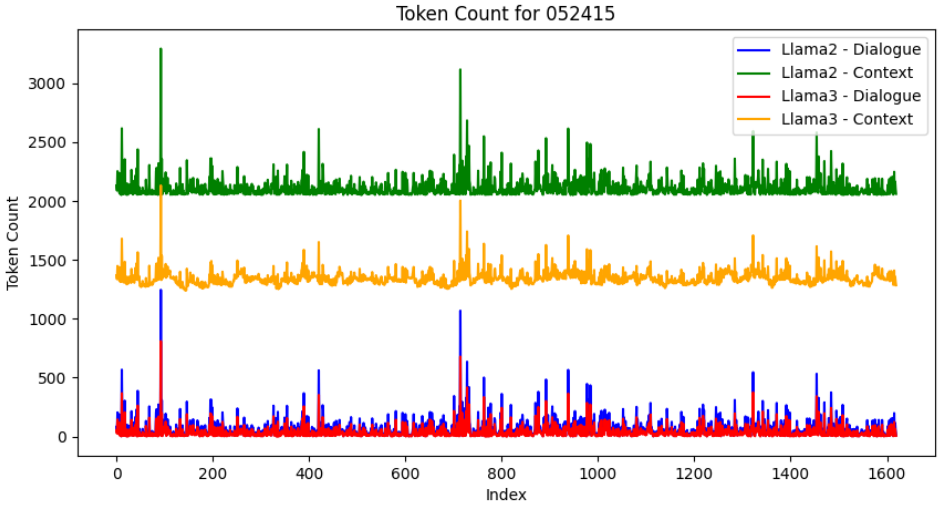

Figure 6: Variation in token count per index for book_id=052415. Excluding exceptionally long dialogues, most token counts are distributed around 2,250 tokens based on the LLaMA 2 Tokenizer and around 1,500 tokens based on the LLaMA 3 Tokenizer.

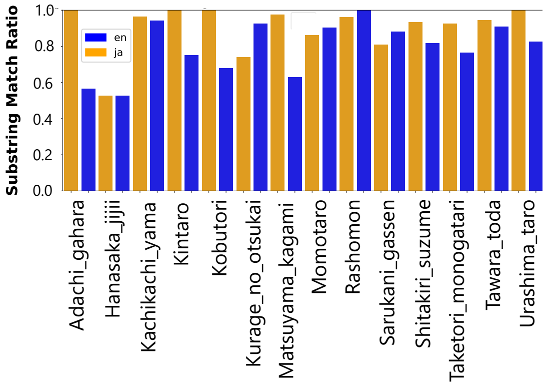

Figure 7: Substring match ratio comparison across stories in Japanese and English datasets, based on results from the `LLaMA-3-70B-Instruct` model.

F Annotation Settings

F.1 Annotation Rules

The following annotation rules were applied for label assignment:

1. As a general principle, the smallest constituent part of a character's name used in the narrative text is considered the correct label. (Example: For "劉備玄徳", "玄徳" is the correct label.)

2. When multiple candidates exist, the given name is preferred if it is present in the context.

3. If the text is not a dialogue, label it as 'Unknown'. (Examples: characters, narrator, book titles)

4. If multiple speakers are indicated for a single utterance, label it as 'Unknown'. (Examples: Guan Yu, Zhao Yun, Liu Bei)

5. Due to the high preparation cost, dynamic generation based on reading the context is preferred, as annotators had prior access to speaker information.

6. Each utterance, along with the preceding and following 1,024 tokens, is set as the context. Only the names found within this context are subject to annotation. The number of tokens is calculated based on the LLaMA-2 Tokenizer[7].

7. If multiple names representing a single person appear in the context, the most appropriate one is labeled as the "main name," while other possible names are labeled as "candidates."

8. List candidates for each main name in a dictionary format. Include various expressions, such as courtesy names or official titles, in the candidates list.

[7] https://huggingface.co/meta-llama/Llama-2-7b-hf

Case	line	excerpt context	pred	true
A	あははは	楊彪は秘策を胸にねりながら、わが邸へ帰って行った。帰るとすぐ、彼は妻の室へはいって、「どうだな。この頃は、郭汜の令夫人とも、時々お目にかかるかね。……おまえたち奥さん連ばかりで、よく色々な会があるとのことだが」と、両手を妻の肩にのせながら、いつになく優しい良人になって云った。 二 楊彪の妻は怪しんで、良人を揶揄した。「あなた。どうしたんですか、いったい今日は」「なにが？」「だって、常には、私に対して、こんなに機嫌をとるあなたではありませんもの」「あははは」「かえって、気味が悪い」「そうかい」	楊彪	楊彪
B	江東の地	呉は、大江の流れに沿うて、「江東の地」と称われている。	不明（ナレーション）	Unknown
C	…………	貂蝉は、さわぐ色もなく、すぐ答えた。「はい。大人のおたのみなら、いつでもこの生命は捧げます」 王允は、座を正して、「では、おまえの真心を見込んで頼みたいことがあるが」 「なんですか」「董卓を殺さねばならん」「…………」「彼を除かなければ、漢室の天子はあってもないのと同じだ」「…………」	貂蝉	貂蝉
D	人生の快、ここに尽くる	夜は、牛馬を宰して、聚議の大歓宴が設けられた。「人生の快、ここに尽くる」 関羽、張飛がいうと、「何でこれに尽きよう。これからである」と、玄徳はいった。	関羽、張飛	Unknown
E	玄徳様、ふたりの熱望です。ご承知くださるまいか	たほうがよい」 「玄徳様、ふたりの熱望です。ご承知くださるまいか」 左右から	関羽	関羽

Table 5: Original Case Study in Japanese. ‘pred‘ indicates the predicted speaker label, and ‘true‘ indicates the annotated speaker label.

id	line	excerpt context	pred	true
1869	ですから、父上のお顔で、富豪を紹介して下さい。曹家は、財産こそないが、遠くは夏侯氏の流れを汲み、漢の丞相曹参の末流です。この名門の名を利用して、富豪から金を出させて下さい		曹操	曹操

Table 6: Correct Identification of an Absent Name： ELYZA-JP-8B accurately predicts the name “曹操,” despite it not being present in the context.

For each main name, the presence of candidates in the context is checked, and a set of potential names is automatically generated.

F.2 Detailed Quality Assessment of Annotations

In this study, all annotations were independently performed by the first author, making it impossible to directly evaluate inter-annotator agreement. To verify the quality of the created annotations, we randomly selected 100 samples from the evaluation dataset and asked three independent annotators to review them.

The annotators were tasked with evaluating the labeled speaker names as “appropriate,” “inappropriate,” or “cannot judge”. We assigned weights to these evaluations: 3 points for “appropriate,” 2 points for “cannot judge,” and 1 point for “inappropriate”. The agreement was calculated based on these weighted scores using a three-point Likert scale.

The results showed that two annotators had an agreement rate of 0.97, and one annotator had an agreement rate of 0.96, indicating a very high level of consistency. This suggests that the dataset constructed in this study is of high quality.

Typically, Cohen’s kappa coefficient (Cohen, 1960) is used to evaluate inter-annotator agreement. However, in this case, the agreement rates were so high that setting the original data labels to 3 when calculating the kappa coefficient could lead to undefined values. Therefore, we report only the agreement rate and its variance (see Appendix F.3 for details).

Additionally, the annotation task required an average of 2 hours per annotator, with a compensation rate set at 1,000 yen per hour. The annotations were performed by three native Japanese graduate students, selected for their advanced language proficiency, further contributing to the reliability and accuracy of the data.

Metric	Annotator ID		
	A	B	C
Agreement Rate	0.97	0.97	0.96
Count (3)	97	97	96
Count (2)	3	2	3
Count (1)	0	1	1
Total	100	100	100
Weighted Average Score	2.97	2.96	2.95

Table 7: Annotation agreement and evaluation distribution by annotator. The "Agreement Rate" represents the proportion of cases where independent evaluators marked the data as "appropriate" (3) when the author had labeled it as 3 in the dataset. The "Count (x)" rows indicate the number of times each annotator selected "appropriate" (3), "neutral" (2), or "inappropriate" (1). The "Total" row indicates that each annotator evaluated 100 cases. The "Weighted Average Score" reflects the average score calculated by assigning weights of 3, 2, and 1 to the respective categories.

F.3 Challenging Cases in Annotation Judgment

Table 8 presents examples where annotation decisions were particularly challenging.

Examining the final portion of the context in Table A, it is evident that the character “張飛” strongly asserts that “呂布” must be defeated. This suggests that the preceding conversation was primarily conducted by “玄徳” and “張飛”. Therefore, considering the immediate context, it is highly likely that the line in question was spoken by “張飛”.

However, reading the previous tokens reveals that the line “何事を曹操からいってよこしたのですか” could be attributed to both “張飛” and “関羽”. Consequently, there is a slight possibility that “関羽” could have responded to “玄徳”　s statement, “まあ、これを見るがいい”.

Two of the independent annotators employed to assess annotation quality provided feedback suggesting that the possibility of “関羽” being the speaker could not be entirely ruled out. Such cases, where reaching a consensus on the speaker annotation was extremely difficult, were reported by the annotators three or four times per 100 cases.

G Model Description

The selection criteria for each model aim to comprehensively evaluate performance across various languages and tasks, adaptation to Japanese data, and differences between architectures. This allows for a multifaceted assessment of LLM performance.

In this study, we selected 12 models for comparison, organized into six categories. Below is a description of each model and the rationale for its selection.

LLaMA-3 (Dubey et al., 2024) LLaMA-3 is an LLM that considers human preferences, demonstrating high performance in various tasks such as bi-lingual support, coding, and mathematics. It is also used as a base model for many other models, making it suitable for comparative validation.

Swallow-3 (Fujii et al., 2024) Swallow-3 is a model based on LLaMA-3 that has undergone continual pretraining and instruction tuning with Japanese data. It was selected to analyze changes in Japanese performance and potential performance degradation in English data relative to LLaMA-3.

ELYZA-JP-8B (Hirakawa et al., 2024) ELYZA-JP-8B is a model based on LLaMA-3 that has undergone continual pretraining and instruction tuning with Japanese data. We selected this model to evaluate whether instruction tuning leads to differences when compared to Swallow-3.

llama-3-youko-8B (Mitsuda et al.) llama-3-youko-8B is a model based on LLaMA-3 that has undergone continual pretraining using a mixture of Japanese and English datasets.

Mistral-7B (Jiang et al., 2023) Mistral-7B, like LLaMA-3, is frequently used for comparisons with other models and is known for its high performance despite its smaller size. It was selected to compare a model from a different lineage to LLaMA-3.

RakutenAI-7B (Group et al., 2024) RakutenAI-7B is a model fine-tuned with Japanese data based on Mistral 7B. It was selected to compare the performance of models fine-tuned with Japanese data, similar to Swallow-3.

CALM-3-22B (Ishigami, 2024) CALM-3-22B is an LLM primarily trained on proprietary Japanese data. It was selected to compare the performance of models that mainly handle Japanese data with those that support multiple languages, primarily focusing on English.

id	line	excerpt context	true	corr	incor	neu
3818	呂布を殺せという密命ですな	何度も、繰返し繰返し読み直していると、後ろに立っていた張飛、関羽のふたりが、「何事を曹操からいってよこしたのですか」と、訊ねた。 「まあ、これを見るがいい」 **「呂布を殺せという密命ですな」** 「そうじゃ」 「呂布は、兇勇のみで、もともと義も欠けている人間ですから、曹操のさしずをよい機として、この際、殺してしまうがよいでしょう」 「いや、彼はたのむ所がなくて、わが懐に投じてきた窮鳥だ。それを殺すは、飼禽を縊るようなもの。玄徳こそ、義のない人間といわれよう」 「――が、不義の漢を生かしておけば、ろくなことはしませんぞ。国に及ぼす害は、誰が責めを負いますか」 「次第に、義に富む人間となるように、温情をもって導いてゆく」 「そうやすやす、善人になれるものですか」 張飛は、あくまでも、呂布討つべしと主張したが、玄徳は、従う色もなかった。	張飛	1	0	2

Table 8: Challenging Annotation Example. ‘true‘ indicates the predicted speaker label. ‘corr‘ indicates the number of annotators who judged the annotated label to be correct, ‘incor‘ indicates those who judged it to be incorrect, and ‘neu‘ indicates those who judged it to be neutral. This example illustrates a difficult case where the three independent annotators had differing opinions, highlighting the complexity and subjectivity involved in the annotation process.

Karakuri-8x7B (Inc., 2024) Karakuri-8x7B is a model that uses a Mixture of Experts (MoE) approach by combining multiple models for more effective inference, specifically Mixtral-8x7B (Jiang et al., 2024), and has undergone continual pretraining and fine-tuning with Japanese data. It was selected to compare MoE models with other LLMs.

H Inference and Evaluation Setup

In this study, we set the random seed at 42 and performed 4-bit quantization for model inference. We used the Greedy Decoding Algorithm (Germann, 2003) for decoding. Inference was conducted using an A6000 GPU, with a total inference time of approximately 200 hours.

During evaluation, unnecessary strings, such as special tokens [INST] generated by the LLM, were removed using regular expressions wherever possible.

Additionally, various libraries were utilized for inference, evaluation, and visualization. For example, we employed scikit-learn[8], transformers[9], beautifulsoup4[10], tiktoken[11], openai[12], evaluate[13], accelerate[14], torch[15], datasets[16], and matplotlib[17].

I Prompt Configuration

Predict Quoted Utterance Table 10 shows the prompts used for speaker identification (original version). As shown in this table, we provide several few-shot examples in a chat format. The prompt consists of text extracted from the beginning of book_id=052410 included in Aozora Bunko. In Table 10, few-shot examples (Chen et al., 2019b) related to the story, along with the target story (`{Context}`) and are provided the utterance line (`{Line}`) for speaker identification.

Using these prompts, we constructed a dataset to evaluate the accuracy of speaker identification and conducted speaker identification based on this dataset.

In addition, Table 11 shows an example story used for prompts. This example was inserted into the `Context` sections of Tables 2 and 10 as part of the few-shot learning examples.

J Impact of Varying Context Lengths with Other Models

Figures 8–9 illustrate the accuracy of substring matches when varying the input context length

[8] https://scikit-learn.org/
[9] https://github.com/huggingface/transformers
[10] https://beautiful-soup-4.readthedocs.io/
[11] https://github.com/openai/tiktoken
[12] https://github.com/openai/openai-python
[13] https://github.com/huggingface/evaluate
[14] https://github.com/huggingface/accelerate
[15] https://github.com/pytorch/pytorch
[16] https://github.com/huggingface/datasets
[17] https://matplotlib.org/

Story	Tokens (Llama-3)		Lines		Skip	
	JA	EN	JA	EN	JA	EN
Shita-kiri Suzume	2,838	3,256	46	22	1	2
Tawara Toda	2,035	2,823	18	11	0	1
Urashima Taro	4,036	5,272	36	69	0	3
Kachikachi Yama	3,175	2,842	58	17	1	0
Kintaro	2,816	3,920	30	52	1	6
Taketori Monogatari	5,452	6,680	27	17	0	0
Matsuyama Kagami	2,839	6,219	40	46	0	0
Adachigahara	2,479	2,083	17	23	0	0
Hanasaka Jijii	2,237	3,339	19	19	2	2
Kurage no Otsukai	2,837	3,728	58	67	0	0
Saru Kani Kassen	2,498	3,256	42	17	0	0
Momotaro	4,031	5,361	58	83	9	1
Rashomon	2,176	2,730	26	32	4	0
Kubu-tori	3,539	2,579	42	25	0	0
Total	42,988	54,088	517	500	18	15

Table 9: Summary of token and utterance counts for both Japanese (JA) and English (EN) versions of each story. Annotation was performed on the main names of characters, following the methodology used in constructing the dataset for the Japanese version of "Romance of the Three Kingdoms" (see Section 4).

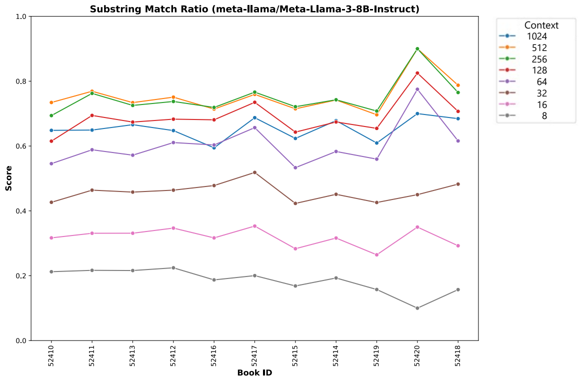

Figure 8: Variation in Substring Match Ratio by Context Length for `LLaMA-8B-Instruct`. This figure shows how the substring match ratio changes with different context lengths.

across different models.

As shown in these figures, models with approximately 70B parameters exhibited improved speaker identification accuracy as the context length increased. Conversely, for models with 8B parameters or fewer, accuracy plateaued when the context length was extended from 256 to 512 tokens. Beyond this point, providing additional context resulted in a performance decline due to the introduction of noise, with the extent of the decline varying across models.

These observations suggest that the effective context length for input varies depending on the model's parameter size and training methodology.

K Candidate Sets for Relaxed Speaker Name Matching

During the evaluation, we matched the predicted speaker names with the most corresponding name from the candidate sets. As shown in Figure 10, the substring match ratio using these sets was higher than in the initial evaluation. For book_id=52419, the evaluation became more consistent with the candidate sets.

Comparison of Paraphrase Set Acquisition with Wikipedia Redirects The Romance of the Three Kingdoms is well-known, resulting in many of its characters having dedicated Wikipedia pages. Therefore, using Wikipedia Redirects [18] to acquire paraphrase expressions is conceivable. However, our attempts revealed that paraphrase expressions could only be acquired for some characters.

Specifically, excluding the names we extracted as Main Name from our created paraphrase sets, only 1.83% of paraphrase candidates could be obtained using Wikipedia Redirects. Notably, expressions corresponding to "劉備" as "青年" or "應德" as "旅人" could not be obtained.

[18] `https://en.wikipedia.org/wiki/Wikipedia:Redirect`

Role	Content
user	次の物語（# Example Story）中のセリフに対して、誰が発言したのかを推測して、話者の名前のみを答えてください。
assistant	了解しました。以下の物語とセリフに基づいて答えを提供します。
user	# Example Story {Example Context}
assistant	物語を確認しました。では、セリフごとに誰が発言したのかを答えます。
user	次の発話は誰が発言しましたか？
assistant	セリフを教えてください。
user	おーい
assistant	漁夫
user	——そこの若い者ウ。なにを見ているんだい。いくら待っていても、そこは渡し舟の着く所じゃないぞ
assistant	漁夫
user	ありがとう
assistant	青年
user	おい、おい、旅の者
assistant	百姓
user	——そんな所で、今朝からなにを待っているんだね。このごろは、黄巧賊とかいう悪徒が立ち廻るからな。役人衆に怪しまれるぞよ
assistant	百姓
user	同様にして、次の物語（# Target Story）中のセリフに対して、誰が発言したのかを推測して、話者の名前のみを答えてください。
assistant	了解しました。以下の物語とセリフに基づいて答えを提供します。
user	# Target Story {Context}
assistant	物語を確認しました。では、セリフごとに誰が発言したのかを答えます。
user	次の発話は誰が発言しましたか？
assistant	セリフを教えてください。
user	{Line}

Table 10: Prompts for Speaker Identification (Original Version). This table represents the chat template prompts in Japanese. The {Example Context} and {Context} sections contain the story content, while the {Line} section specifies the dialogue for which the speaker is to be identified.

These results indicate the limitations of using Wikipedia Redirects for acquiring paraphrase expressions. Hence, combining other methods and data sources is essential for comprehensive paraphrase collection.

L Uncased Exact Match Evaluation

This section addresses evaluation variations arising from case sensitivity in English data. To mitigate such issues, we employ an Uncased Exact Match metric, normalizing generated text to be case-insensitive. As a result, mentions like “Old Woman” and “old woman” are treated as equivalent, ensuring a fairer comparison. Note that this adjustment is only applied to English datasets.

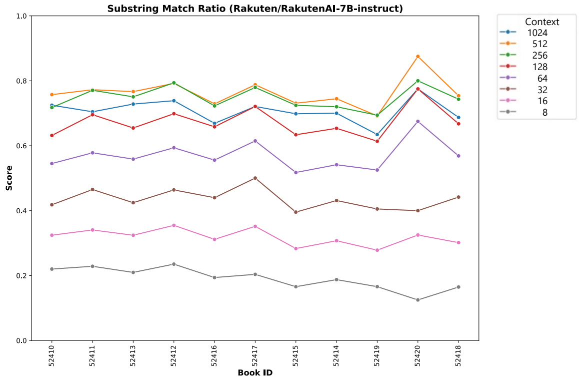

Figure 9: Variation in Substring Match Ratio by Context Length for RakutenAI-7B-Instruct. This figure shows how the substring match ratio changes with different context lengths.

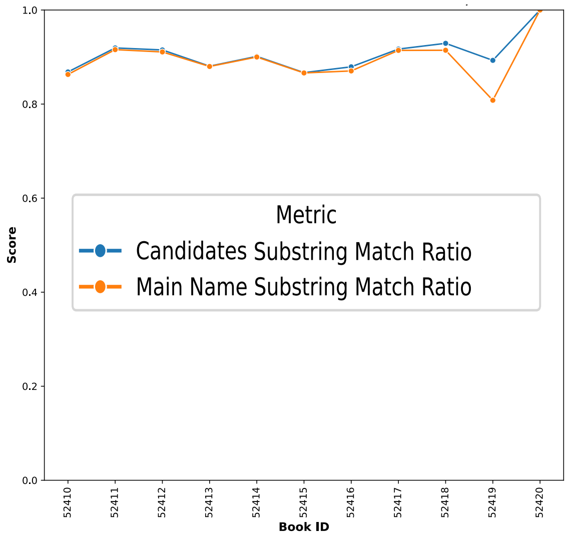

Figure 10: Comparison of the main name and its alternative candidates annotated through substring matching.

Figure 11 illustrates the impact of case sensitivity on evaluation by comparing the uncased substring match ratios for the English and Japanese versions of the story “Kintaro.” Introducing uncased matching consistently improves accuracy. For instance, models such as calm3-22b-chat and LLaMA-3-70B-Instruct benefit notably from this approach. Additionally, the performance of Swallow-70B-Instruct aligns more closely with Swallow-70B, indicating that addressing case-related discrepancies reduces format-driven variance. Overall, uncased evaluation enhances the robustness and reliability of speaker identification metrics.

type	prompt
Japanese Example Story	後漢の建寧元年のころ。今から約千七百八十年ほど前のことである。一人の旅人があった。腰に、一剣を佩いているほか、身なりはいたって見すぼらしいが、眉は秀で、唇は紅く、とりわけ聡明そうな眸や、豊かな頬をしていて、つねにどこかに微笑をふくみ、総じて賤しげな容子がなかった。年の頃は二十四、五。草むらの中に、ぽつねんと坐って、膝をかかえこんでいた。悠久と水は行く——微風は爽やかに鬢をなでる。涼秋の八月だ。そしてそこは、黄河の畔の——黄土層の低い断り岸であった。「おーい」誰か河でよんだ。「——そこの若い者ウ。なにを見ているんだい。いくら待っていても、そこは渡し舟の着く所じゃないぞ」小さな漁船から漁夫がいうのだった。青年は笑くぼを送って、「ありがとう」と、少し頭を下げた。漁船は、下流へ流れ去った。けれど青年は、同じ所に、同じ姿をしていた。膝をかかえて坐ったまま遠心的な眼をうごかさなかった。「おい、おい、旅の者」こんどは、後ろを通った人間が呼びかけた。近村の百姓であろう。ひとりは鶏の足をつかんでさげ、ひとりは農具をかついでいた。「——そんな所で、今朝からなにを待っているんだね。このごろは、黄巾賊とかいう悪徒が立ち廻るからな。役人衆に怪しまれるぞよ」青年は、振りかえって、「はい、どうも」おとなしい会釈をかえした。
English Example Story	In the first year of the Jianning era of the Later Han Dynasty. This was about one thousand seven hundred and eighty years ago. There was a traveler. Apart from wearing a sword at his waist, his appearance was quite shabby. However, he had prominent eyebrows, red lips, especially intelligent-looking eyes, and full cheeks that always seemed to hold a smile, overall giving him an air that was not at all lowly. He appeared to be around twenty-four or twenty-five years old. He was sitting alone in a patch of grass, hugging his knees. Time flows like the eternal river—A gentle breeze brushed his sideburns. It was August, a cool autumn month. And this was the bank of the Yellow River—on a low clay cliff. "Hey there!" Someone called from the river. "—You there, young man. What are you looking at? No matter how long you wait, this is not where the ferry docks." A fisherman from a small boat said. The young man smiled and, "Thank you," he said with a slight nod. The fishing boat drifted downstream. But the young man stayed in the same spot, in the same posture, his eyes still looking into the distance. "Hey, hey, traveler." This time, someone passing by from behind called out. It seemed to be a farmer from a nearby village. One was holding a chicken by its feet, and the other was carrying farming tools. "—What have you been waiting for since this morning in a place like this? Lately, there have been bandits called the Yellow Turbans around. The officials will get suspicious of you." The young man turned and, "Yes, thank you," he replied with a gentle nod.

Table 11: Example Stories

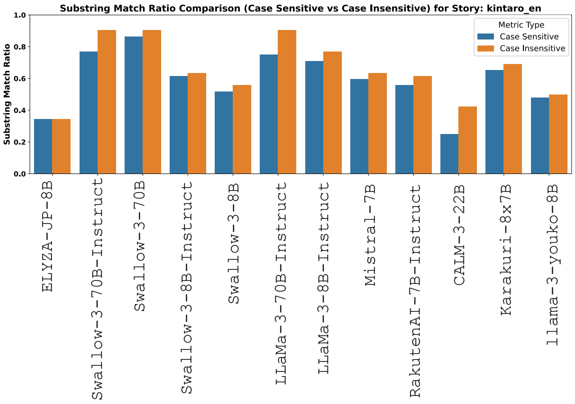

Figure 11: Comparison of Uncased Substring Match Ratio for story: kintaro_en.

Results Figure 12 compares substring match ratios across various models on the English-translated dataset. The English version achieves a substring match ratio of about 70%, approximately 20% lower than the performance on the Japanese data. We attribute this decrease to additional adjectives and extraneous terms introduced in English, which complicate identifying the core speaker references.

These results highlight the importance of translation quality and linguistic nuance when extending datasets to multilingual contexts. Although automated translation accelerates dataset construction, careful consideration of language-specific variations is crucial for maintaining annotation accuracy.

Expenses for Translation Conducting multiple checks and retries for format adherence and correctness increased the total number of tokens processed. The GPT-4o-mini model consumed about 30 million tokens, including retries, resulting in a total translation cost of $6.0. This demonstrates that even with thorough quality controls, automated translation remains a cost-effective strategy for building bilingual datasets.

M Use of AI Tools in Writing and Coding

We used AI tools to assist in the writing and coding processes for this project. Specifically, we employed ChatGPT[19] to help draft and refine the text, and we utilized GitHub Copilot[20] for code completion and suggestions during the coding tasks. These tools were incorporated into our workflow to support the efficient completion of the project.

[19] https://openai.com/chatgpt/

[20] https://docs.github.com/en/copilot

type	prompt
Speaker	Translate the following speaker's name into English, using terms that appear in the translated context. Provide the translation only: Example 1: Translated context: "The farmer walked through his fields, greeting the old man sitting by the road." Output: old man Example 2: Translated context: "In the small village, the young woman was known for her kindness." Output: young woman Example 3: Translated context: "The wise elder spoke to the gathered crowd with great wisdom." Output: wise elder
Dialogue	Extract the entire line that is most similar to this dialogue: 'original_dialogue', excluding the quotation marks. Ensure to extract the full sentence from the start to the end. Example 1: Original dialogue: "これからどうする？" Translated context: "They looked at each other, wondering about the next steps. One of them asked, 'What are we going to do now?' Another responded, 'We need to think carefully.'" Extracted line: What are we going to do now? Example 2: Original dialogue: "何を言えばいいかわからない。" Translated context: "He scratched his head, lost for words. He finally said, 'I have no idea what to say.' Another person nodded in agreement, 'It's a tough situation.'" Extracted line: I have no idea what to say. Failure Example 1: Original dialogue: "こっちへ行こう。" Translated context: "They were considering their options. One said, 'Let's go this way.' Another said, 'I think we should stay here.'" Extracted line: I think we should stay here. # The extracted line is incorrect as it does not match the original dialogue's intent to move.
Context	Translate the following context into English, ensuring consistency and that the provided dialogue is included. The translation should maintain a coherent narrative flow. Provide the translation only: Example 1: Original context: "彼は暗闇の中で独り、静かな夜の音を聞いていた。その時、彼は『おい、誰かいるのか？』と呼びかけた。" Translated dialogue: "Hey, is anyone there?" Translated context: "He sat alone in the darkness, listening to the quiet sounds of the night. At that moment, he called out, 'Hey, is anyone there?'" Example 2: Original context: "彼女は辺りを見回し、そして『ここに何があるの？』と尋ねた。周りには何もないようだった。" Translated dialogue: "What's here?" Translated context: "She looked around and then asked, 'What's here?' There seemed to be nothing around."

Table 12: Prompts for translation

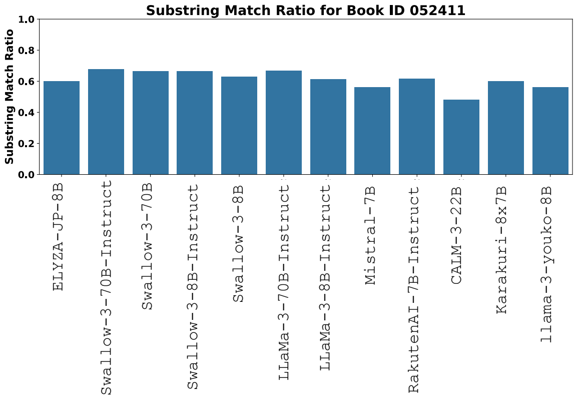

Figure 12: Substring match ratio comparison across models for GPT-4o-mini translated data.