Exploring Context Strategies in LLMs for Discourse-Aware Machine Translation

Ritvik Choudhary **Rem Hida** **Masaki Hamada**
Hayato Futami **Toshiyuki Sekiya**
Sony Group Corporation
ritvik.choudhary@sony.com

Abstract

While large language models (LLMs) excel at machine translation (MT), the impact of how LLMs utilize different forms of contextual information on discourse-level phenomena remains underexplored. We systematically investigate how different forms of context such as prior source sentences, models' generated hypotheses, and reference translations influence standard MT metrics and specific discourse phenomena (formality, pronoun selection, and lexical cohesion). Evaluating multiple LLMs across multiple domains and language pairs, our findings consistently show that context boosts both translation and discourse-specific performance. Notably, the context strategy of combining source text with the model's own prior hypotheses effectively improves discourse consistency without gold references, demonstrating effective use of model's own imperfect generations as diverse contextual cues.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable machine translation (MT) capabilities, often producing outputs close to human quality (Hendy et al., 2023; Xu et al., 2024; Zhu et al., 2024). However, accurately rendering discourse-level phenomena such as appropriate formality or consistent lexical cohesion has been a persistent challenge in MT (García and Firat, 2022; Voita et al., 2019; Kim et al., 2023). Discourse phenomenon is vital for producing translations that are not only fluent but also preserve nuanced meaning across segments. While LLMs' extensive context windows offer a promising avenue for improving discourse-aware translation by moving beyond sentence-level limitations (Wang et al., 2023; Wu et al., 2024), how to make most of this potential for improving discourse accuracy remains an issue (Jiang et al., 2023; Gautam et al., 2024).

In addition, much of the current LLM-MT research, while demonstrating impressive gains, often focuses on improving overall translation quality measured by standard metrics like BLEU (Papineni et al., 2002) or COMET (Rei et al., 2020), with less emphasis on a granular analysis of context-strategy effectiveness for specific discourse phenomena (Zhang et al., 2023).

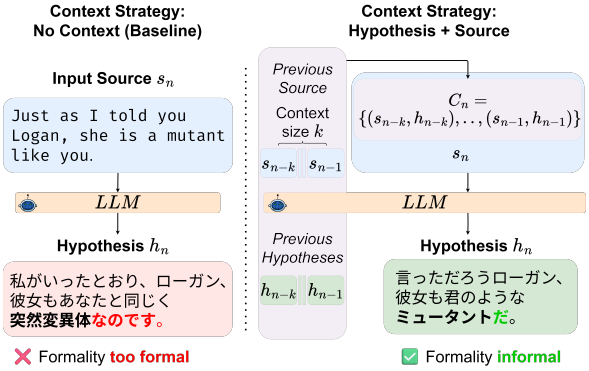

Figure 1: Overview of our context strategy (right) against a no-context baseline (left). Our strategy uses a combination of the past source and the model's own hypotheses as context for improving discourse-level attributes (such as formality).

This paper aims to bridge this gap by analyzing how LLMs leverage diverse forms of context to improve discourse-aware translation. We investigate the impact of various context strategies, such as using source text, model-generated hypotheses, reference translations, and combinations thereof, on the performances of two leading LLMs, GPT-4o (Hurst et al., 2024) and Gemma 2-27B (Riviere et al., 2024). To ensure the potential generalizability of our findings, our evaluation covers multiple language pairs (English-to-Japanese, English-to-French, and English-to-German) and two distinct domains: monologues from the TED Talks corpus, dialogue from an curated OpenSubtitles subset. Also, we move beyond conventional MT evaluation metrics by employing the MuDA framework (Fernandes et al., 2023) to assess discourse phenomena such as formality, pronoun, and lexical cohesion.

Our key findings reveal that: (1) LLMs utilize

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 24382–24391
November 4-9, 2025

context to improve performance compared to a no-context baseline, across both translation and discourse metrics, confirming and extending findings from NMT (Fernandes et al., 2023); (2) Even without gold references as context, LLMs can effectively leverage source-side context and their own prior hypotheses indicating a capacity for self-supervision that enhances discourse consistency.

2 Methodology

This section outlines the methodology used to investigate how context utilization affects discourse phenomena in LLM-based machine translation.

Formally, given a sequence of cohesive source sentences $S = \{s_1, ..., s_{n-1}\}$, the task is to generate a hypothesis translation h_n for the final source sentence s_n, conditioned on s_n and relevant context C_n. This can be represented as generating h_n given (s_n, C_n).

2.1 Context Strategies

The formulation of C_n is crucial for context-aware machine translation. Building upon prior work that explored previous source as context in NMT (Pal et al., 2024), we systematically vary the type and combination of information provided to the LLM, exploring the following context strategies:

(a) **Source Only** (`src`): The context consists solely of the k preceding source sentences. Formally, $C_n = \{s_{n-k}, ..., s_{n-1}\}$.

(b) **Hypothesis Only** (`hyp`): k preceding generated hypothesis translations. Formally, $C_n = \{h_{n-k}, ..., h_{n-1}\}$.

(c) **Hypothesis with Source** (`src+hyp`): Extending prior work, this strategy combines both the preceding source sentences and their corresponding generated hypotheses. Formally, $C_n = \{(s_{n-k}, h_{n-k}), ..., (s_{n-1}, h_{n-1})\}$.

Note that the hypotheses h_n for (b) and (c) are generated sequentially with their respective context strategies C_n. Furthermore, following Scherrer et al. (2019), we also evaluate two strategies that utilize reference (gold) translations, $R = \{r_1, ..., r_{n-1}\}$, the parallel translations corresponding to S. While impractical for real-world inference, these strategies provide access to perfect contextual information thereby serving as an upper bound for context-aware translation performance.

(d) **Reference Only** (`ref`): The context consists of the k preceding reference translations. Formally, $C_n = \{r_{n-k}, ..., r_{n-1}\}$.

(e) **Reference with Source** (`src+ref`): The context combines both the preceding source sentences and their corresponding reference translations. Formally, $C_n = \{(s_{n-k}, r_{n-k}), ..., (s_{n-1}, r_{n-1})\}$.

where k represents the context window size. Figure 1 shows our context strategies with an example.

2.2 Discourse Phenomena

A wide range of linguistic elements contribute to discourse-level phenomena in translation. Our study focuses on three key discourse phenomena, formality, lexical cohesion, and pronoun selection, chosen for their sensitivity to context and their impact on translation quality across our target languages: Japanese, French, and German.

Formality We assess formality through its distinct linguistic realizations: in Japanese, the appropriate use of honorifics (keigo) conveying respect and politeness (Marrese-Taylor et al., 2023); and in French and German, the correct T-V distinction in second-person pronouns reflecting speaker-addressee relationships (Brown et al., 1960).

Lexical Cohesion This refers to the consistent translation of entities or concepts across sentences within a discourse, essential for maintaining coherence (Katori, 2006; Halliday and Matthiessen, 2013).

Pronoun Selection This evaluates context-dependent accuracy in translating pronouns, particularly for determining correct referents and ensuring appropriate agreement (e.g., gender), such as in French and German (Bawden et al., 2018).

3 Experiments

This section outlines the setup used to evaluate the impact of the various context strategies. We describe the dataset and evaluation metrics, followed by the models and experimental configurations.

3.1 Data and Evaluation

Datasets To evaluate and assess generalizability of our strategies across domains and language pairs we use the following datasets. (1) TED Talks corpus (Qi et al., 2018) which consists of speeches (monologues) across multiple

Model	Context Strategy	en-ja					en-fr					en-de				
		MT (↑)		MuDA F1 (↑)			MT (↑)		MuDA F1 (↑)			MT (↑)		MuDA F1 (↑)		
		BLEU	COM.	Form.	Lex.	Pron.	BLEU	COM.	Form.	Lex.	Pron.	BLEU	COM.	Form.	Lex.	Pron.
GPT-4o	no-ctx	15.46	85.49	0.52	**0.53**	**0.41**	32.39	80.25	0.73	0.69	0.60	34.62	79.87	0.58	0.60	0.57
	src	**15.97**	86.24	**0.55**	**0.53**	0.40	**32.87**	**80.48**	0.80	**0.70**	**0.62**	**35.19**	79.92	**0.68**	**0.61**	0.59
	hyp	15.28	**86.30**	**0.55**	0.52	0.38	32.51	80.44	**0.81**	**0.70**	**0.62**	34.61	79.95	0.67	0.60	**0.63**
	src + hyp	15.56	86.16	**0.55**	**0.53**	0.39	32.64	80.43	**0.81**	**0.70**	**0.62**	35.16	**79.97**	**0.68**	**0.61**	**0.63**
	ref	17.06	86.39	0.57	0.61	0.40	35.36	81.14	0.82	0.71	0.64	35.20	79.92	0.73	0.65	0.58
	src + ref	18.38	86.92	0.58	0.62	0.48	46.32	83.03	0.82	0.72	0.66	37.49	80.77	0.75	0.70	0.61
Gemma2	no-ctx	14.14	84.33	0.46	0.53	0.28	28.66	78.95	0.62	0.67	0.55	30.56	78.43	0.48	**0.58**	0.56
	src	**14.32**	84.52	0.49	**0.54**	0.28	27.01	78.37	0.71	0.66	0.57	29.41	77.66	0.52	0.56	0.56
	hyp	13.29	84.67	0.47	0.53	0.27	27.84	78.80	0.70	0.67	0.56	30.29	78.28	**0.55**	**0.58**	**0.59**
	src + hyp	14.15	**84.96**	**0.51**	**0.54**	**0.31**	**30.18**	**79.31**	**0.76**	**0.68**	**0.62**	**31.90**	**78.77**	**0.55**	**0.58**	**0.59**
	ref	14.67	84.83	0.49	0.57	0.29	29.55	79.05	0.76	0.68	0.60	31.62	78.42	0.59	0.59	0.60
	src + ref	16.77	85.45	0.54	0.59	0.39	30.73	79.62	0.78	0.69	0.62	33.11	79.09	0.66	0.63	0.62

Table 1: Evaluation results for GPT-4o and Gemma 2-27B across varying context strategies ($k = 5$) for `en-ja`, `en-fr`, and `en-de` translation using the TED dataset. We report BLEU, COMET (COM.), and MuDA F1 scores for Formality (Form.), Lexical Cohesion (Lex.), and Pronoun Selection (Pron.). **Bold** indicates the best scores without reference-based context; score refer to results with access to (gold) reference translations as context.

language pairs (English-to-Japanese, English-to-French, English-to-German), as employed in prior discourse-aware MT research (Fernandes et al., 2023); (2) Furthermore, to assess performance on conversational dialogue, which is inherently challenging for MT models, we also manually curated English-to-Japanese subset of the OpenSubtitles 2018 corpus (Lison et al., 2018). Further details for both the datasets, including curation process, statistics, and discourse phenomena frequency, are provided in Appendix A.1.

Evaluation Metrics To assess the impact of context strategies on translation quality, we employ both standard and neural evaluation metrics. Specifically, we use BLEU (via SacreBLEU (Post, 2018)) and COMET (`Unbabel/wmt22-comet-da`). For the discourse phenomena detailed in Section 2.2 (formality, lexical cohesion, and pronoun selection), we utilize the MuDA benchmark tagger (Fernandes et al., 2023) and report F1 scores.

3.2 Experimental Setup

Models We evaluate two LLMs, prompting GPT-4o and the open-source Gemma 2-27B-Instruct. GPT-4o was accessed via the OpenAI API (`gpt-4o-2024-11-20`). Gemma 2-27B-Instruct was run locally using the HuggingFace Transformers library (Wolf et al., 2019). More details are covered in Appendix A.2.

Context-Strategy Comparisons We compare the context strategies outlined in Section 2.1 against a sentence-level baseline (`no-ctx`), where $C_n = \varnothing$. Due to computational constraints, the context window size, k, is set to 5 unless specified otherwise. We employ a fixed prompting strategy; details of the prompt are provided in Appendix A.3.

4 Results and Analysis

Table 1 and Table 2 present the results across context strategies, datasets, and language pairs. Our analysis focuses on the following key insights.

4.1 Context Universally Improves Discourse-Aware Translation in LLMs

A primary finding, consistent across both datasets, models, and all tested language pairs, is that all context provision strategies generally improve performance over the `no-ctx` baseline. This holds for both traditional MT metrics and discourse-specific MuDA F1 scores. Moreover, providing gold reference context (`ref` or `src+ref`) typically yields the highest scores, establishing a practical upper bound for context-aware performance. Notably, even providing just the source-side context (`src`) consistently outperforms the baseline, demonstrating that even minimal contextual information can prove to be beneficial.

4.2 Self-Supervision with `src+hyp` Enhances Discourse Consistency

In real-world scenarios where gold reference translations are unavailable for context, the `src+hyp` strategy of combining prior source sentences with the LLM's own previously generated hypotheses, emerges as a highly effective approach across both datasets and language pairs. Particularly for the Gemma 2-27B, its performance closely approaches

Model	Context Strategy	MT Metrics (↑)		MuDA F1 (↑)		
		BLEU	COM.	Form.	Lex.	Pron.
GPT-4o	no-ctx	5.86	79.04	0.19	0.28	0.20
	src	**7.09**	**79.97**	0.28	0.29	**0.27**
	hyp	6.40	79.83	0.27	**0.30**	0.22
	src + hyp	6.88	79.78	**0.30**	**0.30**	0.22
	ref	7.78	80.36	0.36	0.35	0.32
	src + ref	9.64	80.45	0.38	0.37	0.20
Gemma2	no-ctx	5.51	77.55	0.20	**0.30**	0.15
	src	5.87	**78.20**	0.27	**0.30**	**0.17**
	hyp	5.99	77.60	**0.30**	**0.30**	0.09
	src + hyp	**6.26**	78.05	**0.30**	**0.30**	0.10
	ref	6.61	77.95	0.33	0.34	0.15
	src + ref	8.22	78.27	0.34	0.38	0.17

Table 2: Evaluation results for GPT-4o and Gemma 2-27B across varying context strategies ($k = 5$) for `en-ja` translation using our OpenSubtitles2018 subset. **Bold** are the best scores without reference-based context, meanwhile score refers to results with access to (gold) reference translations as context.

or, in some instances (e.g., `en-fr`, Table 1), even surpasses that of the `ref` strategy. This indicates that LLMs can effectively leverage both semantic information and stylistic attributes from the source and their own prior (potentially imperfect) outputs, suggesting a capacity for self-supervision to improve discourse-level consistency.

Table 3 presents qualitative examples of `src+hyp`'s impact across different phenomena and language pairs. For instance the `no-context` baseline produces an overly formal Japanese translation of the source dialogue (ending in "です"). In contrast, the `src+hyp` strategy, having seen informal preceding context k, generates a more informal output ("だ") consistent with the tone of the context. Similar improvements are noted for selecting correct pronoun in German.

Next, to quantitatively validate this capacity for self-supervision, we analyzed the correlation between the discourse quality of the context batch and that of the subsequently generated sentence for the Gemma 2-27B model on the TED dataset. As shown in Table 4, we found a consistent positive correlation between the F1 scores of the k-sentence context batch and the current translated sentence's F1 score across phenomena and language pairs.

The observed positive correlations (e.g., $r = 0.263$ for formality, `en-ja`) confirm that higher-quality context in the `src+hyp` batch tends to produce higher-quality translations. This further supports the self-supervision insight, as the model generates more accurate and consistent discourse, the subsequent translations are more likely to also be consistent. While the modest strength of the correlations suggest that context quality is one of the several influential factors, alongside immediate source sentence and noise from potentially irrelevant context, the overall positive direction across languages validates the effectiveness of using model's own hypothesis to guide future generations.

Discourse Phenom.	Segment	Text & *Annotation*
Formality (en-ja)	Source (current)	He's like a father figure in my life, you know.
	Ref. (Informal)	親父みたいな存在だ。
	`no-context`	彼は私の生活における父親のような存在です。*(× Too formal)*
	Prev. context k (for `src+hyp`)	S: Me and Sosa been rocking since day one. H: 俺とソーサは最初からずっと仲良しだった。 *(✓ Informal tone set in prior hyp)*
	`src+hyp`	俺の人生で父親のような存在だ。 *(✓ Adapts to context)*
	`src+hyp` *adapts to formality established in its own prior turn's context.*	
Pronoun (en-de)	Source (current)	And **they** create 90 pieces of content each month.
	Reference	Und **sie** erstellt jeden Monat 90 Inhaltselemente.
	`no-context`	Und **er** erstellt jeden Monat 90 Inhaltselemente. *(× Incorrect gender)*
	Prev. context k (for `src+hyp`)	S: Consider average Facebook user; perhaps **she's** a student. H: Denken Sie an die durchschnittliche Facebook-Nutzerin; vielleicht ist **sie** Studentin. *(✓ Gender established in prior hyp)*
	`src+hyp`	Und **sie** erstellt jeden Monat 90 Inhaltselemente. *(✓ Correct gender from context)*
	`src+hyp` *infers gender from own prior turn.*	

Table 3: Qualitative examples from Gemma2-27B across multiple language pairs illustrating the impact of context on discourse phenomena. Key parts bolded.

Phenomenon	Lang.	Context Corr. (r)
Formality	`en-ja`	0.263
	`en-fr`	0.258
	`en-de`	0.247
Lexical Cohesion	`en-ja`	0.260
	`en-fr`	0.213
	`en-de`	0.265

Table 4: Pearson Correlation (r) between Context Batch F1 and Current Sentence F1 (`src+hyp` strategy, Gemma 2-27B, $k = 5$) evaluated on the TED dataset.

4.3 Practical Considerations: Optimal Context Length

While our results underscore the benefits of context, its practical application benefits from identifying an optimal length.

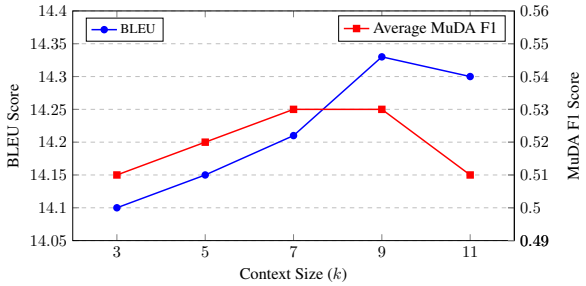

Figure 2: BLEU and average MuDA F1 scores (across formality and lexical cohesion) for different context sizes ($k = 3, 5, 7, 9, 11$) using the `src+hyp` strategy, for the Gemma 2-27B model on the TED `en-ja` dataset.

We further analyze the effect of varying the context window size, $k \in \{3, 5, 7, 9, 11\}$, focusing on the Gemma 2-27B model and the `src+hyp` strategy due to computational constraints. Figure 2 illustrates this relationship for the TED `en-ja` task, as measured by BLEU and the average MuDA F1 score (across formality and lexical cohesion). Initially, both metrics exhibit an upward trend with increasing k from 3 to 7, indicating that larger context windows generally improve both overall translation quality and discourse-level appropriateness (captured by MuDA). However, both metrics show a plateauing effect, or even a slight decrease, at higher k values. This indicates diminishing returns, suggesting there exists an optimal window, beyond which, further increases in k offer minimal to no improvement and can even slightly degrade quality, potentially due to the introduction of less relevant or noisy information.

5 Related Work

Context-Aware Machine Translation The importance of context for coherence and disambiguation in MT is well-established (Jin et al., 2023; Stahlberg, 2019). While traditional NMT systems faced challenges in incorporating information beyond the current sentence (Lopes et al., 2020), the advent of LLMs, with their longer context windows (Tsirmpas et al., 2024), has opened new avenues for leveraging context in MT (Karpinska and Iyyer, 2023). Prior work with LLMs has explored in-context learning (Brown et al., 2020; Hendy et al., 2023), and various other prompting strategies to guide translation (Zhang et al., 2023; Lippmann et al., 2025; He et al., 2024). While these studies demonstrate the general contextual capabilities of LLMs, our work differs in its specific focus and methodology. We systematically compare several practical context formation strategies for prompting LLMs. This contrasts with approaches centered on architectural modifications for context integration or fine-tuning for document-level translation. Furthermore, our investigation provides a broader empirical validation of these strategies across multiple language pairs and datasets, moving beyond single-setting evaluations.

Discourse Phenomena in Machine Translation Accurately translating discourse phenomena (e.g., formality, lexical cohesion etc.) is a significant challenge in MT (Jin et al., 2023; Gautam et al., 2024). Evaluating these nuanced aspects often requires specialized metrics and datasets beyond standard sentence-level MT evaluation. Fernandes et al. (2023) introduced MuDA, a tagger designed to identify the translation of words involved in discourse-level ambiguities, which we utilize in this work. Meanwhile recent research on LLMs have examined their handling of long-range dependencies and contextual information (Mohammed et al., 2024; Wu et al., 2024; Wang et al., 2024), with some studies exploring discourse in domains like literary translation (Jiang et al., 2023). However, the impact of different context strategies on fine-grained discourse phenomena, particularly in a multi-lingual LLM prompting setup remains unexplored. Our work addresses this gap by evaluating how various context strategies influence formality, lexical cohesion, and pronoun selection, thereby offering insights into improving discourse consistency with LLM-based MT.

6 Conclusion

This work evaluated how LLMs leverage various contextual cues for discourse-aware machine translation across diverse settings. We found that all explored context strategies improved both standard MT metrics and discourse phenomena like formality and pronoun selection compared to no-context baselines. Notably, combining prior source text with model-generated hypotheses proved to be effective for improving discourse consistency without gold references, hinting at potential for self-supervision. Future work includes extending this analysis to more typologically diverse languages and a wider range of discourse phenomena.

Limitations

While this work offers insights into context use in LLM-based, discourse-aware machine translation, several limitations should be acknowledged. Although the TED Talks dataset covered additional language pairs (en-fr, en-de), analysis for more typologically diverse languages, domains, and other larger-scale datasets warrants further investigation. Our experiments within the OpenSubtitles movie dialogue domain focused mainly on English-to-Japanese translation due to the expertise of manual curation required for high-quality, genre-specific parallel data.

Furthermore, the opacity of closed-source models like GPT-4o, regarding their specific architecture and training data, limits our ability to fully interpret the observed behavior or to guarantee completely fair comparisons with open-source models.

Regarding evaluation, MuDA's reliance on surface-form matching for F1-score calculation may underestimate performance by penalizing valid alternative phrasing. Finally for our experiments, we employed a single, fixed prompting strategy. Different prompting approaches could potentially yield different results, particularly for aspects like formality. Exploring the sensitivity of the results to variations in prompting is an important direction for future work.

Ethics Statement

This research adheres to the ACL Ethics Policy. We acknowledge several key ethical considerations: the potential misuse of LLMs to generate misleading or harmful translations; the risk of perpetuating societal biases embedded in LLM training data, potentially leading to unfair or discriminatory outputs; the inherent presence of potentially offensive or inappropriate content within the TED data or the OpenSubtitles film dialogue dataset, despite our curation; the environmental impact of large-scale LLM research due to high computational demands; and the limitations on accessibility and reproducibility caused by reliance on proprietary models like GPT-4o. While we employed instruct models and curated our dataset, these broader ethical concerns require ongoing attention in the development and responsible deployment of LLM-based MT systems.

References

Reina Akama, Sho Yokoi, Jun Suzuki, and Kentaro Inui. 2020. Filtering noisy dialogue corpora by connectivity and content relatedness. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 941–958, Online. Association for Computational Linguistics.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.

Roger Brown, Albert Gilman, et al. 1960. *The pronouns of power and solidarity*, pages 253–276. Bobbs-Merrill.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Patrick Fernandes, Kayo Yin, Emmy Liu, André Martins, and Graham Neubig. 2023. When does translation require context? a data-driven, multilingual exploration. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 606–626, Toronto, Canada. Association for Computational Linguistics.

Xavier García and Orhan Firat. 2022. Using natural language prompts for machine translation. *ArXiv*, abs/2202.11822.

Vagrant Gautam, Eileen Bingert, Dawei Zhu, Anne Lauscher, and Dietrich Klakow. 2024. Robust pronoun fidelity with English LLMs: Are they reasoning, repeating, or just biased? *Transactions of the Association for Computational Linguistics*, 12:1755–1779.

Michael Alexander Kirkwood Halliday and Christian MIM Matthiessen. 2013. *Halliday's introduction to functional grammar*. Routledge.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2024. Exploring human-like translation strategy with large language models. *Transactions of the Association for Computational Linguistics*, 12:229–246.

Amr Hendy, Mohamed Gomaa Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *ArXiv*, abs/2302.09210.

Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, et al. 2024. Gpt-4o system card. *ArXiv*, abs/2410.21276.

Yuchen Eleanor Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Mrinmaya Sachan, and Ryan Cotterell. 2023. Discourse-centric evaluation of document-level machine translation with a new densely annotated parallel corpus of novels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7853–7872, Toronto, Canada. Association for Computational Linguistics.

Linghao Jin, Jacqueline He, Jonathan May, and Xuezhe Ma. 2023. Challenges in context-aware neural machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15246–15263, Singapore. Association for Computational Linguistics.

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. In *Proceedings of the Eighth Conference on Machine Translation*, pages 419–451, Singapore. Association for Computational Linguistics.

Yoshikazu Katori. 2006. Translating cohesion in journalistic texts, between japanese and english. *Interpretation studies: The Journal of the Japan Association for Interpretation Studies*, 6:69–89.

Dohee Kim, Yujin Baek, Soyoung Yang, and Jaegul Choo. 2023. Towards formality-aware neural machine translation by leveraging context information. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7384–7392, Singapore. Association for Computational Linguistics.

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.

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

António Lopes, M. Amin Farajian, Rachel Bawden, Michael Zhang, and André F. T. Martins. 2020. Document-level neural MT: A systematic comparison. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 225–234, Lisboa, Portugal. European Association for Machine Translation.

Edison Marrese-Taylor, Pin Chen Wang, and Yutaka Matsuo. 2023. Towards better evaluation for formality-controlled English-Japanese machine translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 551–560, Singapore. Association for Computational Linguistics.

Wafaa Mohammed, Sweta Agrawal, Amin Farajian, Vera Cabarrão, Bryan Eikema, Ana C Farinha, and José G. C. De Souza. 2024. Findings of the WMT 2024 shared task on chat translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 701–714, Miami, Florida, USA. Association for Computational Linguistics.

Proyag Pal, Alexandra Birch, and Kenneth Heafield. 2024. Document-level machine translation with large-scale public parallel corpora. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13185–13197, Bangkok, Thailand. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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, Brussels, Belgium. Association for Computational Linguistics.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535, New Orleans, Louisiana. 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, Online. Association for Computational Linguistics.

Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, L'eonard Hussenot, Thomas Mesnard, Bobak Shahriari, et al. 2024. Gemma 2: Improving open language models at a practical size. *ArXiv*, abs/2408.00118.

Yves Scherrer, Jörg Tiedemann, and Sharid Loáiciga. 2019. Analysing concatenation approaches to document-level nmt in two different domains. In *The Fourth Workshop on Discourse in Machine Translation*, pages 51–61, United States. The Association for

Computational Linguistics. Workshop on Discourse in Machine Translation ; Conference date: 03-11-2019 Through 03-11-2019.

Felix Stahlberg. 2019. Neural machine translation: A review. *J. Artif. Intell. Res.*, 69:343–418.

Dimitrios Tsirmpas, Ioannis Gkionis, Georgios Th. Papadopoulos, and Ioannis Mademlis. 2024. Neural natural language processing for long texts: A survey on classification and summarization. *Eng. Appl. Artif. Intell.*, 133(PC).

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.

Longyue Wang, Zefeng Du, Wenxiang Jiao, Chenyang Lyu, Jianhui Pang, Leyang Cui, Kaiqiang Song, Derek Wong, Shuming Shi, and Zhaopeng Tu. 2024. Benchmarking and improving long-text translation with large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7175–7187, Bangkok, Thailand. Association for Computational Linguistics.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661, Singapore. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Minghao Wu, Thuy-Trang Vu, Lizhen Qu, George Foster, and Gholamreza Haffari. 2024. Adapting large language models for document-level machine translation. Abs/2401.06468.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024. A paradigm shift in machine translation: Boosting translation performance of large language models. In *The Twelfth International Conference on Learning Representations*.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 41092–41110. PMLR.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

A Appendix

A.1 Dataset Details

This appendix provides further details on the datasets used in our experiments.

A.1.1 TED Talks (Multilingual Monologues)

To assess generalizability to a different genre and additional language pairs, we used the TED Talks corpus, specifically the IWSLT section as prepared by Qi et al. (2018). This corpus consists of transcribed and translated prepared speeches (monologues). For our experiments, we utilized the following TED talks data, as seen in prior work in discourse-aware MT (Fernandes et al., 2023):

- **English-to-Japanese (en-ja):** Test set created from IWSLT17 comprising of 5565 sentences.
- **English-to-French (en-fr):** Test set created from IWSLT17 comprising of 4866 sentences.
- **English-to-German (en-de):** Test set created from IWSLT17 comprising of 4491 sentences.

The documents within these test sets were processed sequentially to maintain discourse context for our experiments. Frequency of the tagged discourse phenomenon across this dataset is summarized in Table 6.

A.1.2 OpenSubtitles Subset (English-to-Japanese Dialogue)

For evaluating performance on conversational dialogue, we utilized a manually curated subset from the OpenSubtitles2018 corpus (Lison et al., 2018). OpenSubtitles2018 is a large, publicly available multilingual parallel corpus of movie and TV show subtitles, licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 International License. Its conversational style and inherent presence of diverse discourse phenomena make it a relevant resource. However, the raw corpus is known to contain significant noise, including misalignments and translation inaccuracies (Akama et al., 2020). Table 5 provides a genre breakdown of our curated OpenSubtitles en-ja subset.

Genre Category in Source Film	Turns
Documentary	984
Action	1945
Drama (American)	427
Drama (Japanese)	605
Total Test Turns	**3961**

Table 5: Genre breakdown of the curated OpenSubtitles en-ja test subset.

To address these quality concerns and create a reliable test bed for English-to-Japanese discourse phenomena, we undertook a careful manual curation process:

- **Film Selection:** We selected 7 films known for their high-quality official Japanese subtitles and diverse dialogue styles, aiming for a mix of genres (see Table 5).
- **Turn Selection and Cleaning:** From these films, we manually reviewed and selected 3961 dialogue turns. This involved:
  - Verifying sentence alignment between English source and Japanese target.
  - Assessing the naturalness and accuracy of the Japanese translations.
  - Filtering out overly short, incomplete, or non-translatable segments (e.g., onomatopoeia without clear equivalents, fragmented dialogue over multiple segments).
  - Ensuring each selected turn was part of a coherent dialogue sequence to preserve contextual dependencies.

  This curated set of 3961 dialogue turns serves as our test set for the OpenSubtitles en-ja experiments.
- **Discourse Phenomena Frequency:** Frequency of the tagged discourse phenomenon across this dataset is summarized in Table 6.

A.2 Model Details and Hyperparameters

We evaluate two large language models on the aforementioned OpenSubtitles subset:

GPT-4o GPT-4o (Hurst et al., 2024) is a closed-source, state-of-the-art LLM developed by OpenAI. We accessed GPT-4o via the OpenAI API [1], specifically using the `gpt-4o-2024-11-20` version. As a closed-source model, details about its architecture and training data are not publicly available. The model is subject to OpenAI's terms of use [2]. For inference, we set the temperature to 0.7 and the maximum response length to 1024 tokens. Other parameters were left at their default API settings.

Gemma 2-27B-Instruct Gemma 2-27B-Instruct (Riviere et al., 2024) is an open-source LLM developed by Google, licensed under the Gemma Terms of Use and an Apache License 2.0. We used the HuggingFace Transformers library (Wolf et al., 2019) to run Gemma 2-27B-Instruct locally [3]. We ran the model inference on two NVIDIA A6000 GPUs. The maximum output length was set to 1024 tokens, using the default parameters provided via the HuggingFace Transformers model hub [4].

A.3 Context Prompt and Example

Figure 3 shows the general prompt structure used for all experiments. The prompt first provides the general task instructions, followed by the specific context for the given strategy, and finally the source sentence to be translated.

Prompt: You are an expert translator from English to Japanese.

Context: `<Context>`

Source sentence to translate: `<Source Sentence>`

Please translate the source sentence from English to Japanese using the above context and return the final translation with the tag `<translation></translation>`.

Translation:

Figure 3: General prompt structure. `<Context>` is replaced with the appropriate previous context for each strategy: `src`, `hyp`, `ref`, or combinations thereof, and `<Source Sentence>` is replaced with the current English sentence to be translated.

Example Here's an example of the prompt for the `src+hyp` strategy with $k = 3$:

[1] https://openai.com/api/

[2] https://openai.com/policies/row-terms-of-use/

[3] https://huggingface.co/google/gemma-2-27b-it

[4] https://github.com/huggingface/transformers

Dataset	Language Pair	Total Sentences	Tagged Phenomena		
			Formality	Lexical	Pronoun
OpenSubtitles	en-ja	3961	2112	1027	108
TED Talks	en-ja	5565	5652	1943	231
	en-fr	4866	1459	2368	1541
	en-de	4491	1383	1374	400

Table 6: Frequency of MuDA-Tagged Discourse Phenomena in evaluation datasets.

```
You are an expert translator from
English to Japanese.

Context:

Source: She heals! Move!
Hypothesis: 彼女が回復している！動け！

Source: Alright move out.
Hypothesis: よし、出発だ。

Source: Come on, hold her.
Hypothesis: さあ、彼女を押さえろ。

Source sentence to translate:
As I told you Logan, she is a
mutant like you.

Please translate the source sentence
from English to Japanese using the
above context and return the final
translation with the tag
<translation></translation>.

Translation:
```

The model would then generate the Japanese translation for "As I told you Logan, She is a mutant like you."