FALCON: Holistic Framework for Document-Level Machine Translation Evaluation

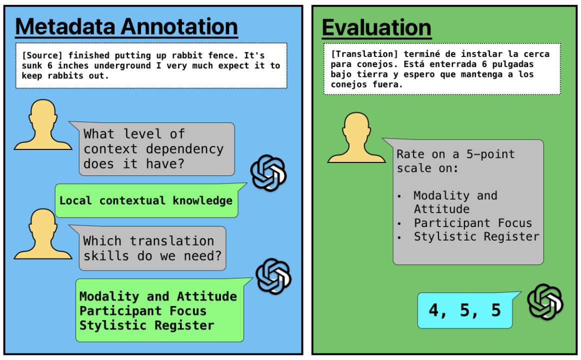
Anonymous ACL submission

Abstract

As per Michael Halliday, language is not just a system of rules, but a tool for meaning-making within sociocultural contexts, whereby language choices shape the functions of a text. We employ Julian House's Translation Quality Assessment model inspired by Halliday's Systemic Functional Linguistics to assess Machine Translation (MT) at the document level, establishing a novel approach titled FALCON (**F**unctional **A**ssessment of **L**anguage and **CO**ntextuality in **N**arratives). It is a skill-specific evaluation framework offering a holistic view of document-level translation phenomena with fine-grained context knowledge annotation. Rather than concentrating on the textual quality, our approach explores the discourse quality of translation by defining a set of core criteria on a sentence basis. To the best of our knowledge, this study represents the inaugural attempt to extend MT evaluation into pragmatics. We revisit WMT 2024 with the English-to-X test set encompassing German, Spanish, and Icelandic, assessing 29 distinct systems in four domains. We present groundbreaking but compelling findings concerning document-level phenomena, which yield conclusions that differ from those established in existing research. Emphasizing the pivotal role of discourse analysis in current MT evaluation, our findings demonstrate a robust correlation with human values, inclusive of the ESA gold scores. [1]

1 Introduction

The demand for document-level evaluation in Machine Translation (MT) has been increasing since the suspicion that system performance exceeded human capabilities, in part at the sentence level (Hassan et al. 2018; Läubli et al. 2018; Toral et al. 2018; Graham et al. 2020a; Graham et al. 2020b).



[Source] finished putting up rabbit fence. It's sunk 6 inches underground I very much expect it to keep rabbits out.

What level of context dependency does it have?

Local contextual knowledge

Which translation skills do we need?

Modality and Attitude
Participant Focus
Stylistic Register

[Translation] terminé de instalar la cerca para conejos. Está enterrada 6 pulgadas bajo tierra y espero que mantenga a los conejos fuera.

Rate on a 5-point scale on:

- Modality and Attitude
- Participant Focus
- Stylistic Register

4, 5, 5

Figure 1: Evaluation process of FALCON.

Despite substantial attempts to establish a workable arrangement, the current methodology continues to assess general ***textual quality*** of translation in a similar manner to sentence-level evaluations (Maruf et al., 2021). Within this context, human scores often yield inflated perfect ratings, failing to distinguish between rapidly advancing MT models as Large Language Models (LLM) capabilities grow (Kocmi et al. 2023; Kocmi et al. 2024a; Freitag et al. 2024). Despite increased efforts last year (Kocmi et al., 2024a), including a more diverse dataset with new domains like speech and user-generated texts and new document-level error labels in MQM, such as ACCURACY/GENDER MISMATCH and STYLE/ARCHAIC OR OBSCURE WORD CHOICE (Freitag et al., 2024), the findings provide little insight into translation quality or system performance at this level. Consequently, the development of a comprehensive framework capable of elucidating document-level phenomena with enhanced clarity and interpretability is urgently required at present.

Given the notable capabilities exhibited by LLMs across a variety of Natural Language Generation (NLG) tasks, including Summarization, Question Answering, Code and Dialogue Generation, and Image Captioning (Chen et al. 2021; Ouyang et al. 2022; Korbak et al. 2023), many studies aim to evaluate model performance in aspects like Co-

[1] Data and code will be available after review.

herence, Correctness, Relevance, and Informativeness (Zheng et al. 2023; Dennstädt et al. 2024; Liang et al. 2025; Sheokand and Sawant 2025). Despite varying task requirements and criteria names, the main aim is to assess the quality of discourse as a communication method (Sai et al., 2022). In this context, the work by Ye et al. (2024) is particularly innovative, as their evaluation methodology integrates many NLG tasks, excluding MT, into a singular platform, termed FLASK. This approach uses 12 detailed criteria to evaluate the response's functionality from both textual and non-textual perspectives, demonstrating a strong alignment with human values.

By leveraging this strategy, our objective is to enhance the current status of MT through the integration of FLASK into our ecosystem. We utilize their detailed skill-specific evaluation method with a score rubric strategy. The primary distinction is that our evaluation process is designed to promote the multilingual capabilities of the LLM for a single task, translation.

We introduce FALCON, an innovative paradigm for document-level MT evaluation, which assesses ***discourse quality*** by implementing a detailed, skill-specific evaluation framework complemented by comprehensive context annotations. Prior to undertaking the evaluation, we assess the context dependency of each sentence to effectively differentiate between sentence-level and document-level errors. Sentences identified with document-level errors are subsequently annotated with pertinent translation skills. Translations are then rated on a 5-point scale based on the annotated skills and a specific scoring rubric (Figure 1). To our knowledge, this introduces a new methodology in MT evaluation, focusing on the translation's pragmatic function.

We revisit the WMT 2024 with a comprehensive English-to-X test set encompassing German, Spanish, and Icelandic, which respectively represent a high-resource language, a language deemed to be the most accessible, and a low-resource language pair. **We demonstrate that liberation from textual quality constraints is essential for comprehending document-level phenomena in MT.** Our experiments show that our results are more in line with ESA gold scores than MQM scores are, even without sentence-level scores, leading to some differing conclusions. Here are some findings:

- Speech is the most context-intensive domain (94.4%), with context mostly captured within documents. It is challenging for models as they struggle more with adjacent context than real-world knowledge.
- `Dubformer`, `Claude-3.5`, and `Unbabel-Tower70B` excelled across language pairs with strong textual abilities, but lacked in non-textual skills, particularly in low-resourced pairs.
- A decoder-only architecture with paragraph-level strategy was optimal for document-level performance.

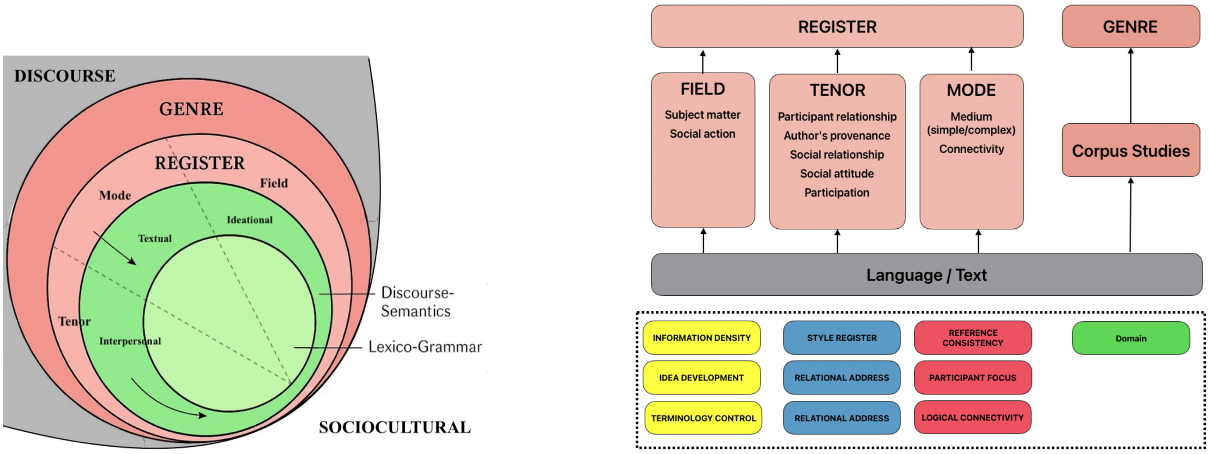
2 Related Works

This section outlines initiatives in WMT conventions to enhance document-level evaluation through varied procedures and tailored test suites. The detailed research has been relegated to Appendix A due to constraints of space. Initially, a single score for a document (DR+DC) was used, but it suffered from low statistical power and frequently generated tie ranks (Barrault et al., 2019). Since then, they have adhered to the sentence-level (or segment-level) scoring method and explored tactics to provide more context. They choose either by using a handful of nearby sentences (SR+DC) (Barrault et al. 2019; Barrault et al. 2020) or the entire document (Akhbardeh et al., 2021). Since Kocmi et al. (2022), it has become standard to use 10 adjacent sentences. Moreover, sentences are organized into paragraphs to allow translators and evaluators flexibility beyond sentence boundaries (Barrault et al. 2019; Kocmi et al. 2023; Kocmi et al. 2024a). Recently, the judgment method shifted from basic scoring to ESA (Kocmi et al., 2024b), which includes error annotation. Yet neither emphasizes document-level features. Our approach, though distinct, also utilizes a scoring method. We calculate a skill-specific score thrice per sentence, using two consecutive sentences for context.

3 FALCON: Functional Assessment of Language and COntextuality in Narratives

3.1 What is a document-level error?

A *document* in document-level MT is a coherent set of sentences, possibly organized into sections or paragraphs (Dahan et al., 2024). Such coherence is influenced by dependencies between sentences, playing a crucial role in evaluation at the document level (Thai et al., 2022). Structurally, it is the extra-sentential context that is vital, focusing on elements whose contextual signals lie beyond the sentence.



a) Halliday's SFL (Eggins, 2004)

b) Julian House's TQA

Figure 2: Our fine-grained evaluation criteria outlined within the dotted box and two discourse language models.

Evaluating a document goes beyond dimension; it involves analyzing how sentences contribute to creating a coherent, structured text, known as *discourse* (Jurafsky and Martin, 2008). The original MQM translation quality encompasses not just textual attributes such as Accuracy and Fluency, but also functional factors like *manufacturing quality* and *user quality*, which cover conformity of products, processes, and projects to expectations (Lommel et al., 2013). However, these attributes do not form the error classification within this framework. In contrast to MQM, we analyze text as discourse, performing a document-level evaluation through a functional lens of discourse analysis.

3.2 Julian House's TQA Model

The discourse analysis gained prominence in translation studies in the 1990s. This field considered translation to be a communicative and social process, advancing beyond simple textual analysis (Munday, 2016). Michael Halliday's Systemic Functional Linguistics (SFL) largely influenced this view, emphasizing that language choices dictated the function of a text (Halliday and Matthiessen, 2004). In the Hallidayan language model, genre shaped "register" of language, which linked linguistic devices to their functions, with influence cascading from the sociocultural environment to discourse and until lexicogrammar, as in Figure 2-a).

Julian House's Translation Quality Assessment (TQA) Model utilizes SFL for discourse analysis of translation by comparing profiles of the source and target texts through the elements of SFL — **FIELD**, **TENOR**, and **MODE** (House, 1997; House, 2015). Unlike MQM, it is function-oriented, evaluating the *metafunctions* of a linguistic device in discourse.

For instance, Nominalization could not only shift discourse focus but alter information density (Halliday and Matthiessen, 2004). Examples (1) and (2) illustrate such phenomena, with sentence (b) as the nominalized construction of sentence (a). In Example (1), the agent (*the committee*) and action (*rejected*) are both highlighted in sentence (a); nominalization then shifts the focus to the action's consequences. In Example (2), three verbs (*examined, tested, reported*) are nominalized in sentence (b), omitting the information of the step-by-step process of the actions from sentence (a).

(1) Shifting focus
a. The committee rejected the proposal.
b. The rejection of the proposal was done by the committee.

(2) Changing information density
a. The scientists examined the samples carefully. They then tested the samples for contamination. Finally, they reported the results to the agency.
b. The scientists' careful examination and testing of the samples for contamination, followed by their reporting of the results to the agency, completed the study.

3.3 Evaluation Criteria

FALCON implements the House's TQA model for LLM-as-judge, organizing **FIELD**, **TENOR**, and **MODE** as meta-categories divided into 9 subcategories. We experimentally select subcriteria for

MT evaluation and name them as illustrated in Figure 2-b). For consistent application of the labels in the evaluation, we ensure that categories have distinct functional characteristics. Here's a brief introduction to our taxonomy, and an overview with linguistic devices and examples in Table 5 in the Appendix:

FIELD is concerned with the content and subject matter of the text; what is being talked about. It includes **INFORMATION DENSITY** (complexity of information aligned with genre norms), **IDEA DEVELOPMENT** (structures and thematic progression), **TERMINOLOGY CONTROL** (correct and consistent use of domain-specific terms).

TENOR addresses the relationship between participants—who is involved in the communication, and how power, stance, and social distance are linguistically expressed. It includes **STYLE REGISTER** (politeness and register appropriate to the context), **RELATIONAL ADDRESS** (author's sociocultural background and relationship with readers), **MODALITY AND ATTITUDE** (author's intent affecting mood and tone).

MODE focuses on the textual organization and the way information is structured—how the message is delivered through sentence linking, cohesion, and rhetorical focus. It includes **REFERENCE CONSISTENCY** (coherent identification of the same entity), **PARTICIPANT FOCUS** (emphasis of the key participants or elements), **LOGICAL CONNECTIVITY** (relationships between ideas).

3.4 Evaluation Protocol

Evaluating a translation presupposes that the source text is complete with contextual information (Smith et al., 2016). In a manner akin to House (2015), we first scrutinize the source profile individually to establish standards for evaluating the target profile. This analysis considers two factors: context dependency and translation skills. This section elaborates on this process.

Profile I: Context Dependency

Each sentence is labeled with its degree of contextual dependency. Dahan et al. (2024) classifies context into local and global based on its range; *Local Context* is located near the current sentence, while *Global Context* covers the entire document. We propose five specific types of context, varying from intra-sentence knowledge to extensive real-world knowledge, enhancing the granularity of this classification: **SENTENCE-LEVEL KNOWLEDGE** (which is confined to the sentence), **LOCAL CONTEXT KNOWLEDGE** (minimal surrounding context), **EXTENDED CONTEXT KNOWLEDGE** (broader scene or paragraph, affecting tone and emotion), **GLOBAL CONTEXT KNOWLEDGE** (entire work such as a novel or TV series), **UNIVERSAL CONTEXT KNOWLEDGE** (history, philosophy, or world event). See Table 4 in the Appendix for a full description. To implement a one-label-per-sentence scheme, it is recommended to select broader knowledge. The labels help choose subsamples for document-level evaluation, since not all sentences need contextual knowledge (Castilho, 2022).

Profile II: Translation Skills

Sentences without SENTENCE-LEVEL are labeled with key translation skills. An evaluator (LLM-as-judge model, in our case) selects 3 primary skills per sentence, which later serve as criteria for the evaluation across any target languages. Domain information is explicitly provided during the annotation due to its importance in the Hallidayan model. See Appendix E for prompt lines. This method significantly contrasts with typical MT error analysis, where neglecting mistakes misleadingly implies flawless quality. FALCON creates a unified assessment framework by defining clear sentential criteria.

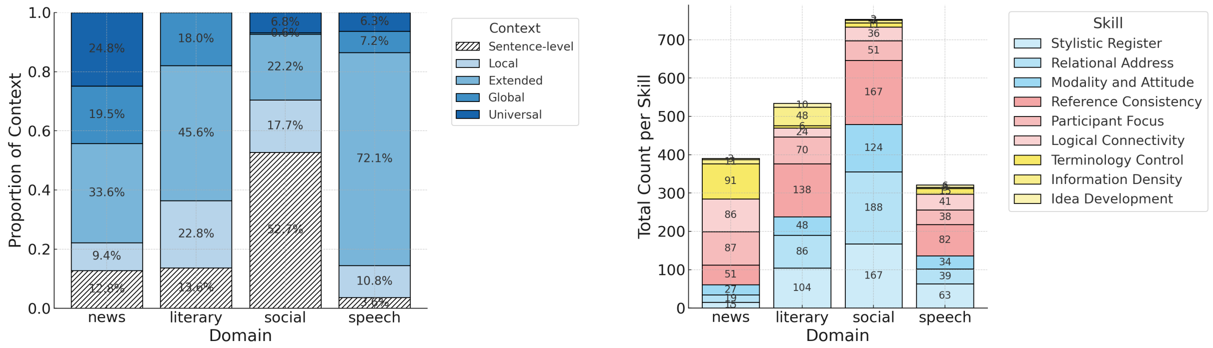
3.5 Evaluation Configuration

Type of Judgment

Given the source and translation along with their previous segments, an evaluator rates each sentence on a 5-point scale per profiled translation skill, ranging from complete failure (1) to full success (5). Although House (2015) recommends a 3-point scale (High-Mid-Low), our pilot study indicates that a 5-point scale offers clearer distinctions. Refer to Appendix E.3 for prompt lines.

Segment Length

Similar to the traditional method, we gather scores for each segment, but our approach assigns three unique scores to each segment and captures various document-level phenomena. This strategy is in



a) Context dependency per domain (unit:%).

b) Translation skills per domain.

Figure 3: Contextual knowledge and translational skills per domain of the WMT24's source test set (En→X).

line with our belief that sentences are fundamental components of a cohesive document.

Amount/Type of Context

Two previous source and reference segments are given as context information for each sentence in the evaluation. We use references instead of hypothesis translations to prevent the accumulation of previous errors affecting current judgments. Our pilot study found that coherence with previous translations leads to current errors being judged as correct.

Scoring Model

The sentence-level score is derived from the average score of three skills. Consequently, correlations with gold scores can be straightforwardly computed at both the sentence and system levels. To enhance statistical power, we prioritize this approach, although correlations at the document level are also feasible. It is important to highlight that sentences with SENTENCE-LEVEL knowledge do not get annotations or scores in this calculation. Thus, focusing on document-level aspects yields scores that differ from official WMT gold scores.

4 Experiment

We perform a document-level assessment using the WMT24 dataset (Kocmi et al., 2024a), noted for its emphasis on discourse features more than earlier iterations, while also reducing the possibility of data leakage due to its freshness. The source text includes 2,383 sentences forming 997 segments with document boundaries. See Table 6 and 7 in the Appendix for key statistics. We study three language pairs: English to Spanish, German, and Icelandic. The English-to-German (`En-De`) pair is the most visited, the English-to-Spanish (`En-Es`) pair is reported to be the easiest (Kocmi et al., 2024a), and the English-to-Icelandic (`En-Is`) pair is the most challenging and low in resources. In total, we have collected 137,862 document-level annotations.

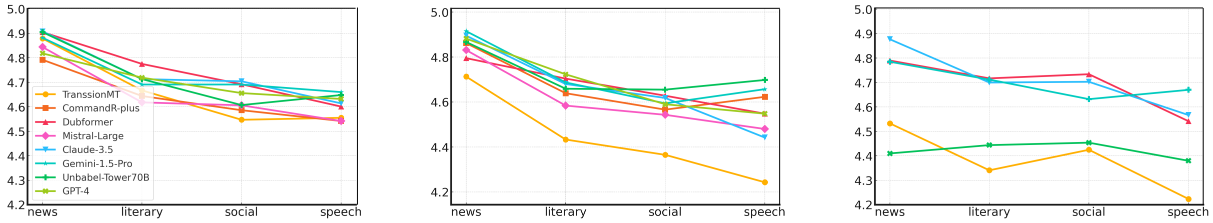
The `GPT-4.1-mini` model (`gpt-4_1-mini-2025-04-14`) serves as the primary evaluator, demonstrating comparable performance to `GPT-4.1` in MT evaluation (Kim, 2025) due to budget limits. It assesses 29 unique systems, and the number of systems is divergent per pair. For reproducibility, the parameters are fixed to `temperature=0, max_tokens=1024`.

4.1 Reliability of FALCON

A subset of 230 segments from the `En-Es` language pair is selected for human evaluation. Three translators and linguists assess the correctness of two profiles: context dependency and translation skills, and rate skills on a 5-point scale. Detailed information regarding the evaluation process is provided in Appendix B. The evaluation results in an acceptance rate of 80.4% for context dependency and 71.6% for translation skills. The overall correlation for the skill score, as presented in Table 8 in the Appendix, is notably high at approximately 0.60. The Inter-annotator Agreement (IAA), quantified by Fleiss' Kappa (Fleiss, 1971), indicates a moderate level with 0.519. The reliability of these findings is further substantiated through its correlation with the public gold ESA score in §5.

4.2 Analysis of Test Set

Unexpectedly, Figure 3-a) reveals that 66.8% of the test set strongly relies on context. Particularly, the speech domain stands out, with 94.4% of its sentences requiring context knowledge, 72.1% of which depends on information found far but within the document. The news domain presents challenges with deep knowledge (UNIVERSAL, 24.8%),



a) En-De

b) En-Es

c) En-Is

Figure 4: Scores of top performers across domains. The systems absent from En-Is are below the range.

System	En-De ↑	En-Es	En-Is
Dubformer	4.74★	4.67	4.71★
Claude-3.5	4.73	4.66	4.71★
Gemini-1.5-Pro	4.72	4.69	-
GPT-4	4.70	4.68	4.43
Unbabel-Tower70B	4.70	4.70★	4.69
ONLINE-B	4.65	4.41	4.37
Mistral-Large	4.65	4.60	2.69
TranssionMT	4.65	4.43	4.39
CommandR-plus	4.63	4.65	2.70
ONLINE-W	4.59	4.54	-
IOL-Research	4.56	4.57	4.10
Llama3-70B	4.49	4.54	3.56
Aya23	4.48	4.44	1.75
ONLINE-A	4.43	4.46	4.16
IKUN	4.29	4.36	4.36
ONLINE-G	4.28	4.28	3.74
Phi-3-Medium	4.25	4.36	1.57
IKUN-C	4.13	4.24	4.27
CUNI-NL	4.03	-	-
Occiglot	3.72	3.69	-
NVIDIA-NeMo	3.60	4.10	-
AIST-AIRC	3.46	-	-
TSU-HITs	2.72	2.18	1.61
MSLC	2.68	3.30	-
CycleL	1.16	1.04	1.10
CycleL2	1.16	-	-
AMI	-	-	4.39

Table 1: General performance of all participating systems in WMT24 for the three language pairs. Winning models are shown in darker colors, with the absolute best marked by ★.

raising questions about its classification as the most basic genre in translation. In the social domain, half of the sentences are clear without context.

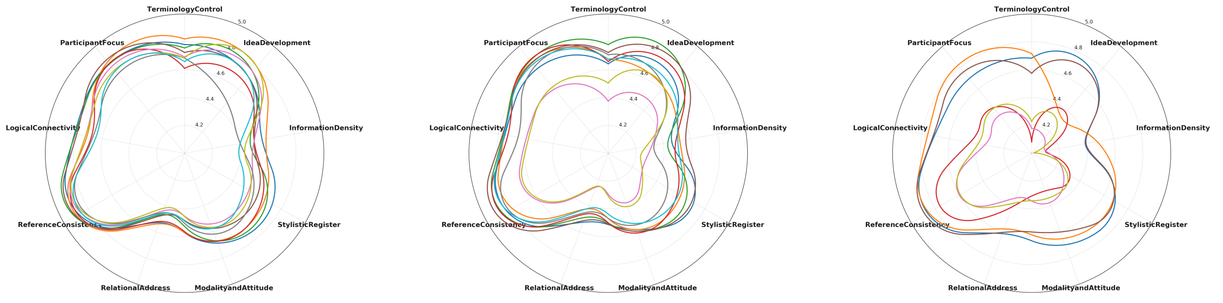
Analyzing requiring skills per domain in Figure 3-b) REFERENCE CONSISTENCY is essential across domains, especially when LOCAL and EXTENDED contexts are required. TERMINOLOGY CONTROL is crucial for news, ensuring accurate domain-specific translations. The social domain emphasizes interpersonal tone (RELATIONAL ADDRESS, STYLISTIC REGISTER). These insights into the test set are crucial for evaluation organizers to establish a targeted objective.

4.3 System Performance

Table 1 displays the overall performance of the systems in three different languages. Dubformer, Claude-3.5, and Unbabel-Tower70B consistently score around 4.70 across all language pairs, which is remarkable, especially considering the inclusion of low-resource languages such as Icelandic. It demonstrates their document-level multilingual capability. Patterns differ between high and low-resource groups. High-resource pairs show competitive top model performance, which creates some *good enough* threshold of around mid-4.50s. In En-Is, however, top performers are rare, with ongoing competition among the mid or low tiers. Due to spatial constraints, systems below 4.6 will be exempt from the main discussion. See Appendix D for full results.

Speech stands as a challenging domain.

Figure 4 indicates that some domains pose greater challenges, making it tougher for systems to excel, in contrast to domains like literary or news, which may feature more predictable language or on which the models are better trained. All systems obtained their highest scores in the news domain, notably with Claude-3.5 scoring 4.9 in both German and Spanish. However, aside from Unbabel-Tower70B, scores significantly decline in the speech domain. This domain effectively tests the systems' robustness at the document level, where even good systems have room to grow. Certainly, this trend is not applicable to En-Is. The graph indicates that models lack generalization across domain types.



a) En-De b) En-Es c) En-Is

Figure 5: Translation skills of nine top performers, colored as `Dubformer`, `Claude-3.5`, `Gemini-1.5-Pro`, `GPT-4`, `Unbabel-Tower70B`, `Online-B`, `Mistral-Large`, `TranssionMT`, `CommandR-plus` for clear distinction. Focusing the range to 4 - 5 has omitted some systems from the chart.

Adjacent contexts are underestimated.

`Claude-3.5` and `Dubformer` lead En-De with 4.76, while `Unbabel-Tower70B` tops En-Es with 4.71, indicating these models are well-tuned across different context types. Such a trend is also witnessed in En-Is. It is notable that Table 10 in the Appendix indicates that difficulties are more pronounced with adjacent context (LOCAL) compared to broader context (GLOBAL). This hints at why the performance has declined in the social domain (Figure 4) rich in local context, countering the belief that broader context is harder to translate. However, there is also a possibility that the referential context in our evaluation setup (the prior two sentences) might be insufficient for accurate judgment on LOCAL context. However, our ablation study proves this is not the case (Appendix C). Contextual understanding is quite inadequate in the En-Is scenario.

Models lack enhanced relational abilities.

Figure 5 shows systems' varied strengths by skill. Top performers have balanced skills, excelling in textual abilities (REFERENCE CONSISTENCY, PARTICIPANT FOCUS) but lacking interpersonal nuances such as RELATIONAL ADDRESS, MODALITY AND ATTITUDE especially in Spanish translation compared to German. This complexity challenges lower-end models more, while top systems manage better. The low-resource language shows a different pattern, with notably weak Tenor- and Field-related skills. In all languages, REFERENCE CONSISTENCY is well captured.

Decoder-only models, trained at the paragraph level, generally yield positive results.

Some systems do not disclose their architectures or context strategies, but for those that do in Kocmi et al. (2024a), we classify them as LLM, Online, and Custom, and assess their average scores by their type, architecture, and strategy. Customized models outperform significantly, as shown in Table 2. Though conclusions are challenging, a decoder-only architecture with paragraph-level training and translation strategy appears to enhance document-level performance. Providing model details will help clarify more uncertainties.

Type	Architecture	Strategy	En-De	En-Es	En-Is	#Sys
Custom	enc-dec	?	4.65★	4.45	4.39★	1
Custom	dec	para	4.28	4.47	4.39★	3
Custom	dec	sent	4.21	4.33	4.31	3
LLM	dec	para	4.58	4.59★	3.06	8
Online	?	?	4.49	4.44	3.32	4
Custom	enc-dec	para	2.68	3.36	-	1

Table 2: Performance categorized by type, architecture, and strategy. Two types of architecture —decoder-only (dec) and sequence-to-sequence (enc-dec) —and two strategies —paragraph-level (para) and sentence-level (sent) —are considered. Systems evaluated for performance are shown in *#Sys*.

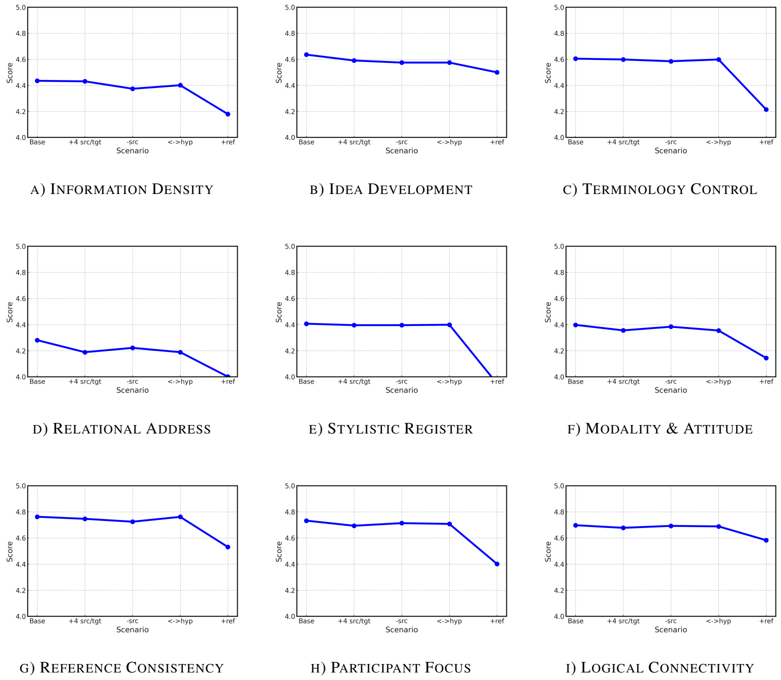
5 Ablation Study

Context is crucial for decision-making in the document-level evaluation. We, thus, review our setup by hypothesizing four scenarios to address the research questions:

+4 src/tgt: Do earlier sentences provide assistance? ⇒ Context is provided with four prior source and reference segments.

↔ **hyp:** Should previous target segments be translation or reference? ⇒ Two prior translations provide context.

− **src:** Are the earlier source sentences required? ⇒ Context is provided with two prior reference segments.



A) Information Density B) Idea Development C) Terminology Control

D) Relational Address E) Stylistic Register F) Modality & Attitude

G) Reference Consistency H) Participant Focus I) Logical Connectivity

Figure 6: Skill score variations for four scenarios compared to the baseline setup.

	Sentence level			System level		
	r	p	τ	r	p	τ
Base	0.447	0.282	0.226	0.843	0.843	0.734
↔ **hyp**	0.451 ↑	0.286 ↑	0.228 ↑	0.918 ↑	0.918 ↑	0.782 ↑
+ **ref**	0.445	0.292 ↑	0.225	0.836	0.836	0.709
+4 **src/tgt**	0.444	0.274	0.219	0.863 ↑	0.864 ↑	0.745 ↑
− **src**	0.444	0.264	0.210	0.745	0.745	0.636
MQM	0.346	0.255	0.210	0.718	0.718	0.564

Table 3: The system- and sentence-level Pearson (r), Spearman (p), and Kendall-Tau (τ) correlations for four scenarios with the ESA gold score. MQM results are also compared. ↑ shows positive change.

+ **ref** Does judgment get more accurate with reference? ⇒ The current reference segment is provided alongside two previous source and reference segments.

We examine the sentence- and system-level correlation with ESA and MQM scores and calculate FALCON scores for various scenarios. After discarding the empty annotations from the full dataset of En-Es, we retain 2,526 sentences. Table 3 demonstrates that our baseline consistently outperforms the MQM gold score, reaffirming the reliability of our framework and its promising potential when incorporating sentence-level scores.

Scenarios generally have better correlation than MQM. While the changes are minor, using translations over references (↔ hyp) modestly enhances correlation in part due to the re-computation of Relational Address (in Figure 6). Note that excessive context information (+4 **src/tgt**) or lack of source context (− **src**) can worsen judgments, contradicting the claim that target context alone suffices for evaluation (Castilho, 2022). Skill scores remain similar except when reference translation is added (+ **ref**). **Thus, we ascertain that the two prior sentences offer ample context, ideally extracted from both the source and target side, but hypothesis can yield more reliable results.**

6 Conclusion

This study introduces an innovative framework for document-level MT evaluation, taking a functional perspective on the text. Our re-assessment of WMT24 experimentally suggests that sophisticated document-level evaluation should integrate neighboring context frequently found in speech or social domains and assess non-textual attributes such as Field and Tenor. It is also advisable to compare performance across diverse language pairs. By adopting this approach, the evaluation is anticipated to acquire enhanced discriminative capacity.

FALCON holds potential utility for 1) evaluation organizers in designing a targeted evaluation environment with a holistic dataset profile, 2) developers in testing model architectures, and 3) lay users seeking general information, all due to its profound interpretability and reliability.

Limitation & Future Works

The research does not include all language pairs from WMT24. Our results suggest that some pairs might yield fascinating outcomes, due either to distinct linguistic characteristics or model training methods. Future studies should investigate En-to-X, X-to-En, and non-English pairs.

The evaluation employs a proprietary model, making the results timely. It is crucial to guarantee the reproducibility of the evaluation. Similarly, we opt for the less robust model (GPT-4.1-mini) rather than identifying the optimal one or utilizing GPT-4.1. While our results demonstrate reliability, it is anticipated that further enhancements will be pursued.

The scope of our human evaluation is confined to a limited cohort of professionals, exhibiting moderate but not robust IAA. Considering the novelty of this intent, further investigations are imperative to ensure a reliable human evaluation. Participants have specifically noted that the size of certain segments are predominantly too large for coherent assessment. Since our evaluation supports sentence-level annotation, our objective is to ensure that the evaluation remains straightforward, efficient, and robust. Furthermore, we recognize that achieving comprehensive consistency throughout the document would benefit from alternative methods of providing contextual information to the evaluator.

References

Farhad Akhbardeh, Andrey Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Christian Federmann, Yvette Graham, Barry Haddow, Kenneth Heafield, Philipp Koehn, Christof Monz, and Others. 2021. Findings of the 2021 conference on machine translation (wmt21). In *Proceedings of the Sixth Conference on Machine Translation (WMT21)*, pages 1–88. Association for Computational Linguistics.

Eleftherios Avramidis, Vivien Macketanz, Aljoscha Burchardt, and et al. 2020. Fine-grained linguistic evaluation for state-of-the-art machine translation. *arXiv preprint arXiv:2010.06359*.

Eleftherios Avramidis, Vivien Macketanz, Ulrich Strohriegel, and Aljoscha Burchardt. 2019. Linguistic evaluation of german-english machine translation using a test suite. *arXiv preprint arXiv:1910.07457*.

Loic Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Barry Haddow, Philipp Koehn, Christof Monz, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (wmt20). In *Proceedings of the Fifth Conference on Machine Translation (WMT20)*. Association for Computational Linguistics.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Barry Haddow, Chris Hokamp, Philipp Koehn, Shervin Malmasi, Christof Monz, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (wmt19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Rachel Bawden and Benoît Sagot. 2023. RoCS-MT: Robustness challenge set for machine translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 198–216, Singapore. Association for Computational Linguistics.

Soham Bhattacharjee, Biswajit Gain, and Asif Ekbal. 2024. Domain dynamics: Evaluating large language models in english-hindi translation. In *Proceedings of the Ninth Conference on Machine Translation (WMT24)*. Association for Computational Linguistics.

Ergun Biçici. 2019. Machine translation with parfda, moses, kenlm, nplm, and pro. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 66–73. Association for Computational Linguistics.

Sheila Castilho. 2022. How much context span is enough? examining context-related issues for document-level mt. In *Proceedings of the 13th Language Resources and Evaluation Conference (LREC 2022)*. Available at: https://doras.dcu.ie/27009/1/How_Much_Context_Span_is_Enough_Castilho_LREC_2022.pdf.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, and 1 others. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Nicolas Dahan, Rachel Bawden, and François Yvon. 2024. Survey of automatic metrics for evaluating machine translation at the document level. Technical report, HAL Open Science. Available at HAL Open Science.

Hillary Dawkins, Isar Nejadgholi, and Chi-Kiu Lo. 2024. WMT24 test suite: Gender resolution in speaker-listener dialogue roles. In *Proceedings of the Ninth Conference on Machine Translation*, pages 307–326, Miami, Florida, USA. Association for Computational Linguistics.

Florian Dennstädt, Michael Gusenbauer, Lisa Langnickel, Alexander Spangher, Albert Barque-Duran, Holden Thorp, Lutz Bornmann, and Maximilian Röglinger. 2024. Title and abstract screening for literature reviews using large language models: a systematic review. *Systematic Reviews*, 13(1):74.

Suzanne Eggins. 2004. *An introduction to systemic functional linguistics*, 2nd edition. London: Continuum.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.

Sigríður Rut Friðriksdóttir. 2024. The genderqueer test suite. In *Proceedings of the Ninth Conference on Machine Translation (WMT24)*, pages 265–273. Association for Computational Linguistics.

Yvette Graham, Christian Federmann, Maria Eskevich, Niko Jojic, and Alexandra Birch. 2020a. Assessing human-parity in machine translation on the segment level. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4197–4211. Association for Computational Linguistics.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2020b. Statistical power and translationese in machine translation evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 72–81. Association for Computational Linguistics.

Michael Alexander Kirkwood Halliday and Christian Matthias Ingemar Martin Matthiessen. 2004. *An Introduction to Functional Grammar*, 3rd edition. Hodder Arnold.

Hany Hassan, Aurelien Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, and et al. 2018. Achieving human parity on automatic chinese to english news translation. *arXiv preprint arXiv:1803.05567*.

Juliane House. 1997. *Translation Quality Assessment: A Model Revisited*. Gunter Narr Verlag, Tübingen.

Juliane House. 2015. *Translation Quality Assessment: Past and Present*. Routledge, London and New York.

Daniel Jurafsky and James H. Martin. 2008. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 2nd edition. Prentice Hall, Upper Saddle River, NJ.

Ahrii Kim. 2025. Straightforward meta-evaluation of LLMs-as-judges in machine translation and DR-100, the LLM-tailored assessment metric. In *ACL 2025 Industry Track*.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, and 3 others. 2024a. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, and 3 others. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Markus Freitag, Yvette Graham, Barry Haddow, Kenneth Heafield, Keisuke Hirasawa, Antonio Jimeno Yepes, Philipp Koehn, Anoop Kunchukuttan, Qingsong Liu, André F. T. Martins, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Graham Neubig, and 3 others. 2022. Findings of the 2022 conference on machine translation (wmt22). In *Proceedings of the Seventh Conference on Machine Translation (WMT22)*, pages 1–172, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tom Kocmi, Tomasz Limisiewicz, and Gabriel Stanovsky. 2020. Gender coreference and bias evaluation at wmt 2020. *arXiv preprint arXiv:2010.06018*.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Sheila Castilho, Barry Haddow, and Lucia Specia. 2024b. Error span annotation: A balanced approach for human evaluation of machine translation. *arXiv preprint arXiv:2406.11580*.

Tomasz Korbak, Kevin Shi, Alice Chen, Rohit Vyas Bhalerao, and 1 others. 2023. Pretraining language models with human preferences. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 17332–17354.

Tao Liang, Zhen Wang, Kaibo Yu, Jiani Chen, Zizheng Wang, Jingyang Zhang, Zhou Yu, Jie Zhou, and Caiming Xiong. 2025. When "yes" meets "but": Can large models comprehend contradictory humor? *arXiv preprint arXiv:2503.23137*.

Arle Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2013. Multidimensional quality metrics: A flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*. AsLing.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has neural machine translation achieved human parity? a case for document-level evaluation. *arXiv preprint arXiv:1808.07048*.

Vivien Macketanz, Eleftherios Avramidis, and Aljoscha Burchardt. 2021. Linguistic evaluation for the 2021 state-of-the-art machine translation systems for german to english and english to german. In *Proceedings of the Sixth Conference on Machine Translation (WMT21)*, pages 1122–1137. Association for Computational Linguistics.

Sabina Manakhimova, Eleftherios Avramidis, and Vivien Macketanz. 2023. Linguistically motivated evaluation of the 2023 state-of-the-art machine translation: Can chatgpt outperform nmt? In *Proceedings of the Eighth Conference on Machine Translation (WMT23)*. Association for Computational Linguistics.

Sabina Manakhimova and Vivien Macketanz. 2024. Investigating the linguistic performance of large language models in machine translation. In *Proceedings of the Ninth Conference on Machine Translation (WMT24)*. Association for Computational Linguistics.

Sajjad Maruf, Faisal Saleh, and Gholamreza Haffari. 2021. A survey on document-level neural machine translation: Methods and evaluation. *ACM Computing Surveys (CSUR)*, 54(2):1–36.

Anwesha Mukherjee and Manish Shrivastava. 2023. Iiit hyd's submission for wmt23 test-suite task. In *Proceedings of the Eighth Conference on Machine Translation (WMT23)*. Association for Computational Linguistics.

Anwesha Mukherjee and Shruti Yadav. 2024. Cost of breaking the llms. In *Proceedings of the Ninth Conference on Machine Translation (WMT24)*. Association for Computational Linguistics.

Jeremy Munday. 2016. Discourse and register analysis approaches. In *Introducing Translation Studies: Theories and Applications*, 4th edition, chapter 6. Routledge.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katrin Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. `https://openai.com/research/instruction-following`. OpenAI technical report.

Sofia Picinini and Sheila Castilho. 2025. Context-aware monolingual human evaluation of machine translation. *arXiv preprint arXiv:2504.07685*.

Maja Popović. 2019. Evaluating conjunction disambiguation on english-to-german and french-to-german wmt 2019 translation hypotheses. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 597–602. Association for Computational Linguistics.

Alessandro Raganato, Yves Scherrer, and Jörg Tiedemann. 2019. The mucow test suite at wmt 2019: Automatically harvested multilingual contrastive word sense disambiguation test sets for machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 603–611. Association for Computational Linguistics.

Nikita Rozanov, Vladislav Pankov, and Danila Mukhutdinov. 2024. Isochronometer: A simple and effective isochronic translation evaluation metric. *arXiv preprint arXiv:2410.11127*.

Kateřina Rysová, Magdaléna Rysová, Tomáš Musil, Lucie Poláková, and Ondřej Bojar. 2019. A test suite and manual evaluation of document-level NMT at WMT19. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 455–463, Florence, Italy. Association for Computational Linguistics.

Abhijeet Bhandari Sai, Abhijit Krishnan Mohankumar, and Mitesh M. Khapra. 2022. A survey of evaluation metrics used for nlg systems. *ACM Computing Surveys (CSUR)*, 55(2):1–35.

Beatrice Savoldi, Marco Gaido, Matteo Negri, and Luisa Bentivogli. 2023. Test suites task: Evaluation of gender fairness in mt with must-she and ines. *arXiv preprint arXiv:2310.19345*.

Yves Scherrer, Alessandro Raganato, and Jörg Tiedemann. 2020. The mucow word sense disambiguation test suite at wmt 2020. In *Proceedings of the Fifth Conference on Machine Translation (WMT20)*.

Maitreya Sheokand and Parth Sawant. 2025. Codemixbench: Evaluating large language models on code generation with code-mixed prompts. *arXiv preprint arXiv:2505.05063*.

Kenny S. Smith, Wilker Aziz, and Lucia Specia. 2016. The trouble with machine translation coherence. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, pages 178–190.

Katherine Thai, Magdalena Karpinska, Kalpesh Krishna, Baishakhi Ray, Kathleen McKeown, Ron Artstein, and Benjamin Van Durme. 2022. Exploring document-level literary machine translation with parallel paragraphs from world literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1256–1274, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. *arXiv preprint arXiv:1808.10432*.

Tereza Vojtěchová, Matúš Novák, Matěj Klouček, and Ondřej Bojar. 2019. Sao wmt19 test suite: Machine translation of audit reports. *arXiv preprint arXiv:1909.01701*.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. 2024. Flask: Fine-grained language model evaluation based on alignment skill sets. *Preprint*, arXiv:2307.10928.

Liang Zheng, Winston L Chiang, Yuhui Sheng, Zhuohan Xu, Rohan Taori, Yan Zhang, Guyu Hu, Tianyi Zhao, Xinying Wang, Noam Shinn, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*.

Vilém Zouhar, Tereza Vojtěchová, and Ondřej Bojar. 2020. Wmt20 document-level markable error exploration. In *Proceedings of the Fifth Conference on Machine Translation (WMT20)*, pages 347–356. Association for Computational Linguistics.

Björn Ármannsson, Hrafn Hafsteinsson, and Atli Jasonarson. 2024. Killing two flies with one stone: An attempt to break llms using english→icelandic idioms and proper names. *arXiv preprint arXiv:2410.03394*.

Type	Description
SENTENCE-LEVEL KNOWLEDGE	The sentence can be fully understood and translated without any outside information. All necessary meaning is present within the sentence itself — vocabulary, grammar, and semantics are straightforward. **Example:** The cat is sleeping on the couch.
LOCAL CONTEXTUAL KNOWLEDGE	Understanding requires minimal surrounding context — maybe the previous or next sentence — but nothing broader. Without it, pronouns, references, or logical connectors might be confusing. **Example:** "She picked it up carefully." (Needs to know who 'she' is and what 'it' is, but usually just from nearby sentences.)
EXTENDED CONTEXTUAL KNOWLEDGE	Grasping the meaning requires understanding the broader scene, paragraph, or emotional flow. Cultural nuance, emotional undertones, or evolving character perspectives start to matter. **Example:** "He knew it was the only way to save them." (Without knowing the stakes or characters, the meaning could shift a lot.)
GLOBAL CONTEXTUAL KNOWLEDGE	The sentence depends on knowledge of the entire work (novel, article, movie) or even multiple entries (book series, TV seasons). Important world-building, character arcs, fictional history, or long-term motifs influence meaning. **Example:** "Winter is coming." (In A Song of Ice and Fire, it's loaded with symbolic and political meaning; outside that, it just sounds like a weather report.)
UNIVERSAL CONTEXTUAL KNOWLEDGE	Understanding draws on extensive external knowledge — history, philosophy, science, mythology, social structures, or famous world events. Without that shared knowledge, translation risks misfiring badly. **Example:** "Opening Pandora's box." (Without knowing Greek mythology, this could be meaningless or totally misinterpreted.)

Table 4: Five types of context dependency, ranging from the superficial (SENTENCE-LEVEL) that do not necessitate document-level analysis, to the world-knowledge (UNIVERSAL).

Table 5: Detailed overview of 9 translation skills and scoring rubric, organized by meta-category of FIELD, TENOR, and MODE, as used in prompt lines.

	Skill	**Description & Score Rubric**
FIELD	INFORMATION DENSITY	Does the sentence compress information into abstract or complex structures required by the genre or audience? Important linguistic devices are nominalization, complex noun phrases, embedded clauses, compounding, metaphors, analogies, symbolic imagery, etc. 1: The translation fails to maintain information density; complex ideas are flattened or left out entirely. 2: Few complex structures are preserved; much of the density is lost through simplification or omission. 3: About half of the compressed or abstract structures are retained; others are overly simplified or miss key nuances. 4: Most complex information is well-preserved, with only minor simplifications or occasional under-representation of density. 5: All compressed, abstract, or complex information is fully and effectively rendered, preserving the dense style required by the genre.
	IDEA DEVELOPMENT	Do some elements in the sentence influence the development of the central theme and the rhetorical structure expected by the genre? Important linguistic devices are discourse markers, schematic structures (e.g., introduction-body-conclusion), paragraph transitions, etc. 1: No discernible preservation of idea development; the translation is fragmented or lacks expected rhetorical structure. 2: The translation weakly maintains discourse flow, with frequent lapses in logical progression and inadequate use of structural devices. 3: Some key elements of discourse structure are present, but notable gaps or awkward transitions hinder smooth idea development. 4: Most discourse markers and structural elements are preserved, with clear development of ideas, though minor lapses in flow or cohesion occur. 5: The translation fully preserves discourse structure and logical flow, ensuring all ideas develop coherently and appropriately within the genre's framework.
	TERMINOLOGY CONTROL	Does the sentence have technical or domain-specific vocabulary that requires accurate and consistent use across an entire text? Important linguistic devices are technical nouns, specialized terminology, standard collocations, fixed expressions, etc. 1: No evidence of accurate terminology control; technical terms are mistranslated, omitted, or inconsistent. 2: Few technical terms are translated accurately; inconsistencies and inaccuracies weaken clarity and precision. 3: About half of the technical terms are handled correctly, while others are inconsistent, overly generic, or slightly inaccurate. 4: Most technical terms are translated accurately and consistently, with only minor inconsistencies or less precise choices. 5: All technical and domain-specific terms are used with precise, consistent, and contextually accurate translations throughout.
TENOR	RELATIONAL ADDRESS	Does the sentence rely on an understanding of the author's cultural, historical, or social background that affects his/her voice, intent, and the nuanced relationships with listener/reader? Important linguistic devices are gendered forms, titles and vocatives, pronoun, honorifics, relational expressions, sociolect, etc. 1: No preservation of relational address; cultural or social subtleties are entirely lost. 2: Weak attention to relational address; many cultural or social cues are mistranslated or ignored. 3: Some relational elements are maintained, but notable omissions or misrepresentations affect reader understanding of nuance. 4: Most relational nuances are well-preserved, with only minor cultural or social context misalignments. 5: All cultural, historical, and social nuances are accurately rendered, fully respecting the relationships and voice of the original text.
	STYLISTIC REGISTER	Do some elements in the sentence require a degree of linguistic politeness and stylistic appropriateness suited to the context and purpose of the text? Important linguistic devices are lexical choice, pronoun usage, verb conjugation, discourse markers, euphemisms, idiomatic expressions, etc. 1: No control of stylistic register; tone and style are inappropriate or absent throughout. 2: Frequent issues with tone, politeness, or formality; the translation regularly strays from the expected register. 3: Mixed success in maintaining register; some passages reflect correct style, while others shift inappropriately. 4: Generally good stylistic control with only occasional mismatches in tone, formality, or politeness. 5: The translation consistently reflects the appropriate stylistic tone and politeness level, matching context and purpose precisely.
	MODALITY AND ATTITUDE	Do some elements in the sentence express possibility, obligation, certainty, or speaker/writer's stance that convey the text's mood and tone? Important linguistic devices are modal verbs and auxiliaries (e.g., must, might), evaluative adjectives (e.g., important, unfortunate), stance adverbs (e.g., perhaps, clearly, surprisingly), emotionally charged expressions, subjunctive or conditional constructions, etc. 1: The translation fails to capture modality or attitude; meaning and tone are significantly altered or obscured. 2: Few modal aspects are retained; much of the nuance in attitude and stance is lost. 3: Some modality is conveyed well, but other elements are flattened, omitted, or mistranslated, affecting tone. 4: Most modal elements are accurately translated, with only minor shifts or losses in the expression of attitude or tone. 5: All nuances of modality and attitude (e.g., certainty, obligation, possibility) are precisely conveyed, preserving the source text's tone.
MODE	REFERENCE CONSISTENCY	Does the sentence contain elements that refer to the same entity within the text? The consistent use of such elements creates connections and coherence and ensures clear identification of participants, objects, and ideas throughout the text. Important linguistic devices are reference, substitution of clause, gender/tense/number agreement, deixis, ellipsis, repetition, synonyms, etc. 1: References are not handled consistently; the translation fails to maintain clarity in tracking entities and ideas. 2: Frequent inconsistency in references leads to unclear or ambiguous connections within the text. 3: Mixed consistency; some references are clear, while others confuse or disrupt textual coherence. 4: Most references are consistent, with minor lapses that do not heavily impact coherence. 5: All references are handled with precision, ensuring consistent, coherent connections across sentences and within the wider text.
	PARTICIPANT FOCUS	Should the emphasis of the sentence on key participants or elements (such as people, places, or objects) be preserved to convey the original meaning across a text? Important linguistic devices are subject-specific terminology, transitivity structures (verb types, selection of active/passive, selection of grammatical subject, use of nominalization instead of verb), etc. 1: Participant focus is not preserved; the translation obscures or neglects important actors or elements. 2: Frequent shifts or losses of focus on key participants reduce clarity and fidelity. 3: Some participants are emphasized correctly, but others are downplayed or misplaced, affecting meaning. 4: Most participant focus is preserved, with only minor deviations in emphasis or clarity. 5: All key participants or elements are clearly and accurately emphasized, fully reflecting their prominence in the source text.
	LOGICAL CONNECTIVITY	Does the sentence have connectors or structures that require clear expression of relationships — such as cause, contrast, or sequence — between ideas? Important linguistic devices are logical connectors (e.g., however, therefore), adversatives, causal linkers, etc. 1: Logical connectivity is lost; relationships between ideas are unclear or missing. 2: Logical links are often unclear or mistranslated, causing confusion in idea relationships. 3: Some connectors are accurately translated, but noticeable gaps or errors weaken logical clarity. 4: Most logical relationships are well-maintained, though minor lapses or slightly awkward expressions appear. 5: All logical connectors and relationships (cause, contrast, sequence) are precisely expressed, preserving the original's clarity of logic.

A Detailed related works

Annotators in Barrault et al. (2019) had access to the surrounding sentences while evaluating sentential scores (SR+DC), as single document scores (DR+DC) often lacked statistical power and increased ties between systems (Graham et al., 2020a). Test suites comprised domain-specific features (Vojtěchová et al. 2019; Biçici 2019), linguistic features (Avramidis et al. 2019; Popović 2019; (Raganato et al., 2019)), and discourse linguistic phenomena Rysová et al. (2019). In 2020, the evaluation methodology covered more language pairs (Barrault et al., 2020) and included expanded or new test suites, such as those focusing on terminology Zouhar et al. (2020), linguistic phenomena Avramidis et al. (2020), Scherrer et al. (2020), as well as coreference and gender bias Kocmi et al. (2020).

Access to context expanded from nearby sentences to the entire document in 2021 (SR+FD) for all language pairs (Akhbardeh et al., 2021). There were no notable developments with test suites, but additional linguistic elements, like idioms, were added to the current set (Macketanz et al., 2021).

In 2022, human evaluation reverted to the SR+DC method by presenting 10 consecutive sentences for context (Kocmi et al., 2022). In that year, rather than creating test suites, they manually identified issue types using 24 linguistic features in the English-to-Croatian translation direction. Although this method was novel, its application is restricted to text analysis.

In 2023, maintaining the conventional approach, they tested paragraph-level evaluation for the English-German pair (Kocmi et al., 2023). No new discoveries were reported, and the differences between the systems were minimal. Test suites were expanded with diverse linguistic features (Manakhimova et al. 2023; Savoldi et al. 2023) and included discourse features related to textual intent (Mukherjee and Shrivastava, 2023).

In 2024, significant progress was made in document-level evaluation, with efforts concentrated on diversifying the test set through new domains (medicine, patents, social, etc.) and text types like speech-to-text and user-generated content (Kocmi et al., 2024a). Some new labels were introduced to capture document-level phenomena, such as "Accuracy/Gender mismatch" or "Style/Archaic or obscure word choice". Nevertheless, the human evaluations often culminated in perfect scores of 100 out of 100, which undermined the efficacy in distinguishing between different systems. The test suites were more robust than ever, with 11 submissions covering additional languages. Some emphasized linguistic traits (Ármannsson et al. (2024) and Friðriksdóttir (2024) for Icelandic, Manakhimova and Macketanz (2024) for German and Russian), while others targeted domain specificity (Mukherjee and Yadav 2024; Bhattacharjee et al. 2024; Rozanov et al. 2024; Bawden and Sagot 2023), or addressed both aspects (Dawkins et al., 2024) in Spanish, Czech, and Icelandic.

	News	Literary	Speech	Social	Total
#docs	17	8	107	34	166
#segs					
all-level	149	206	111	531	997
doc-level	130	178	107	251	666
#sents	332	593	684	774	**2383**
#judgments					
context	447	618	333	1,593	2,991
skill	390	534	321	753	1,998

Table 6: Key statistics of the test set and judgments collected from our experiment. Due to a document boundary, each segment consists of 1-14 sentences based on the domain. Document-level sentences ('doc-level') are selected for the main evaluation from the complete set of sentences ('all-level').

	En-De	En-Es	En-Is	Total
#systems	26	23	20	29 (unique)
#judgments	51,948	45,954	39,960	137,862

Table 7: Systems per language pair and judgments collected in our experiment. Most systems participated across these pairs, with some exceptions.

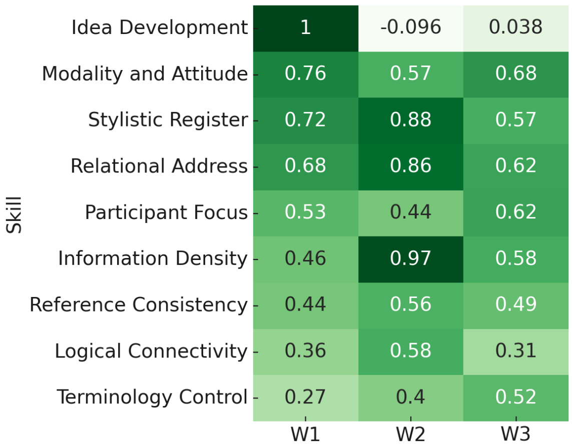


Figure 7: Enter Caption

B Human Evaluation

The En-Es dataset sample was extracted by varying systems and domains. We employed two translators and a linguist for three tasks: 1) verify the accuracy of labeled context dependency, 2) validate labeled skill sets, and 3) rate each skill from 1 to 5. Communication occurred via Zoom. Following a one-hour meeting, we provided them with an Excel sheet to complete the evaluation within a day. They could freely consider context and revise if needed. The reference translation was withheld to avoid potential bias (Picinini and Castilho, 2025). Participants received appropriate compensation after completing the evaluation. The result displayed that our framework was moderately aligned with the professionals. When analyzed by skill criteria in Figure 7, it was noted that participants W2 and W3 demonstrated especially weak comprehension in Idea Development.

	r	p	τ
W1	0.579	0.660	0.634
W2	0.633	0.736	0.720
W3	0.569	0.676	0.648
Overall	0.593	0.690	0.667

Table 8: GPT's Pearson (r), Spearman (p), and Kendall-Tau (τ) correlations of skill scores to professionals.

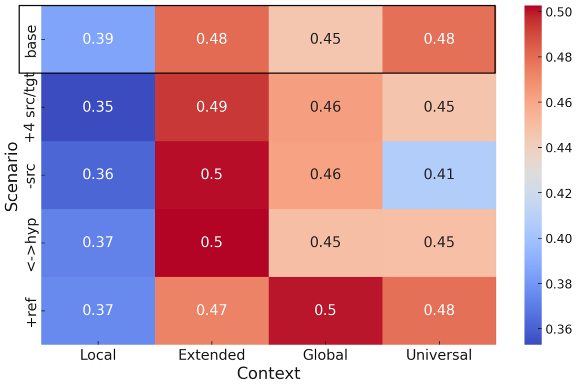


Figure 8: Pearson correlation of various context types to the ESA score.

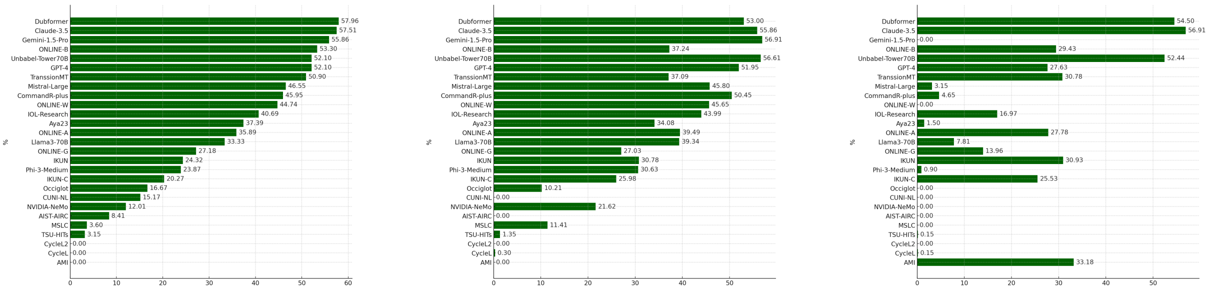
C Additional Results

C.1 FALCON provides sufficient nearby context cues.

Our evaluation framework is criticized for perhaps not sufficiently communicating context, particularly concerning LOCAL CONTEXTUAL KNOWLEDGE. To remedy this, Pearson correlation coefficients for different context types relative to the ESA gold score are calculated. Figure 8 shows that our baseline excels in LOCAL CONTEXTUAL KNOWLEDGE against other versions, with minimal differences. Importantly, UNIVERSAL CONTEXTUAL KNOWLEDGE shows significant fluctuations as context types change, with reliability dropping when source context is omitted.

C.2 Perfect sentences are rare in En-Is.

We evaluate sentences with an average score of 5 of 5 per system for different language pairs. Figure 9 shows that many systems excel in German and Spanish, with mid-tier models showing notable strength in Spanish. Conversely, achieving this level of success is uncommon for Icelandic, particularly among mid-level models, which underscores the difficulty of En-Is translation except by top-tier models at the document level.



a) En-Ee

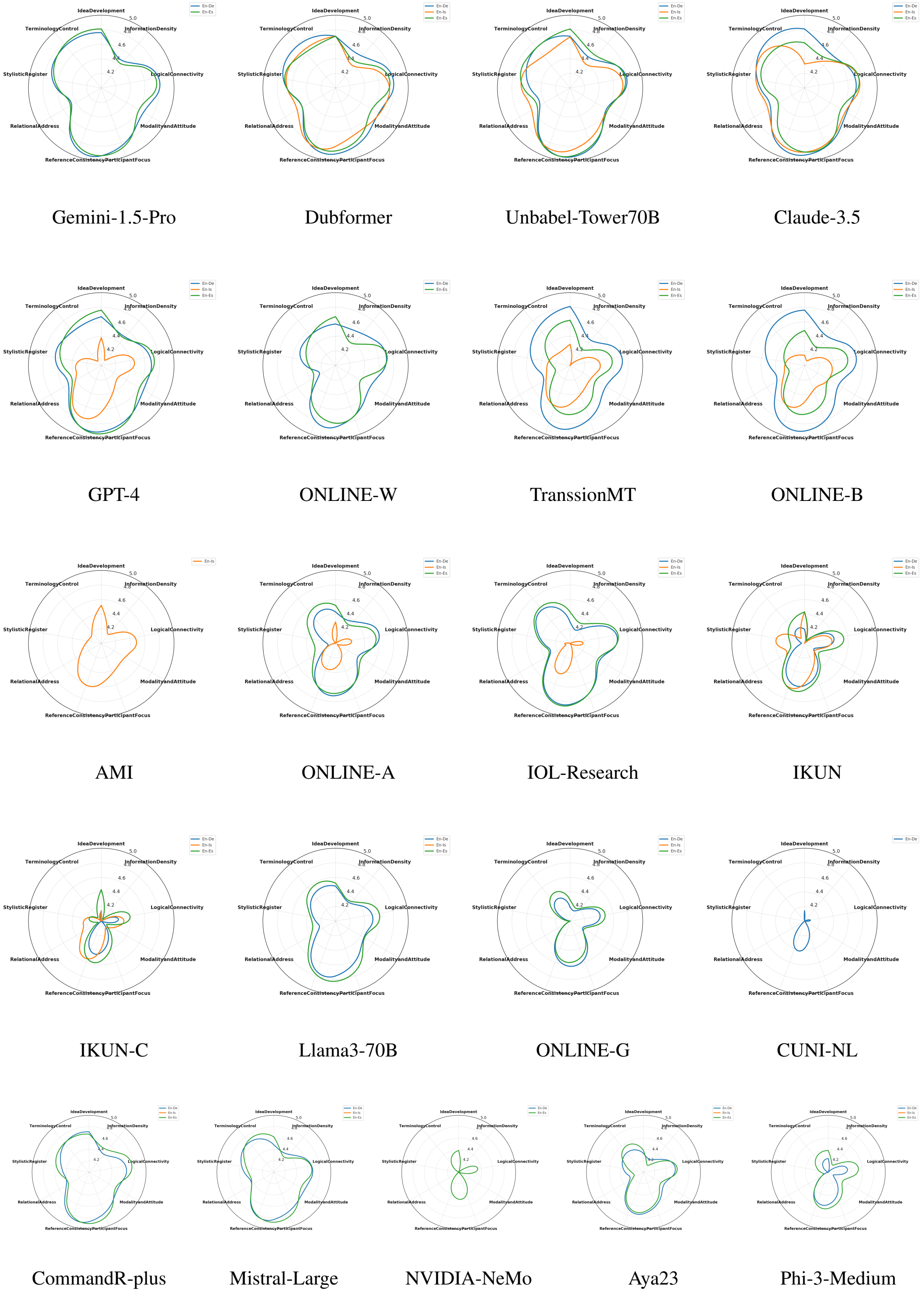
b) En-Es

c) En-Is

Figure 9: Proportion of perfect sentences (average score of 5.0) across language pairs, with three graphs sharing the y-axis for clarity (unit: %).

D Full Results

This section delineates the performance of all systems encompassed within the dataset. The performances are systematically categorized according to domain, context, and skill sets of meta-categories and sub-categories.



Gemini-1.5-Pro

Dubformer

Unbabel-Tower70B

Claude-3.5

GPT-4

ONLINE-W

TranssionMT

ONLINE-B

AMI

ONLINE-A

IOL-Research

IKUN

IKUN-C

Llama3-70B

ONLINE-G

CUNI-NL

CommandR-plus

Mistral-Large

NVIDIA-NeMo

Aya23

Phi-3-Medium

Figure 10: Performance of 21 systems across three language pairs (En-De, En-Es, and En-Is). Systems below the range of 4.2-5.0 are exempt.

Table 9: System performance by domain.

System	En-De					En-Es					En-Is				
	Literature	News	Social	Speech	*Avg.* ↑	Literature	News	Social	Speech	*Avg.*	Literature	News	Social	Speech	*Avg.*
Dubformer	4.7753	4.9051	4.6919	4.6012	4.7434	4.7041	4.7949	4.6282	4.5483	4.6689	4.7172	4.7897	4.7344	4.5421	4.6959
Claude-3.5	4.7135	4.9077	4.7039	4.6137	4.7347	4.6798	4.8949	4.6175	4.4424	4.6586	4.7004	4.8769	4.7025	4.5670	4.7117
Gemini-1.5-Pro	4.6910	4.8821	4.6906	4.6604	4.7310	4.6873	4.9154	4.5936	4.6573	4.7134	-	-	-	-	-
Unbabel-Tower70B	4.7135	4.9051	4.6069	4.6480	4.7184	4.6592	4.8667	4.6547	4.6978	4.7196	4.7116	4.7846	4.6321	4.6698	4.6995
GPT-4	4.7191	4.8179	4.6560	4.6324	4.7064	4.7228	4.8821	4.5896	4.5483	4.6857	4.4438	4.4103	4.4542	4.3801	4.4221
ONLINE-B	4.7004	4.8821	4.5405	4.5607	4.6709	4.4494	4.6410	4.3652	4.1620	4.4044	4.3539	4.4872	4.3997	4.1869	4.3569
TranssionMT	4.6667	4.8795	4.5471	4.5545	4.6620	4.4326	4.7128	4.3652	4.2430	4.4384	4.3408	4.5333	4.4250	4.2243	4.3809
Mistral-Large	4.6180	4.8436	4.6056	4.5421	4.6523	4.5843	4.8308	4.5432	4.4798	4.6095	2.8483	2.0897	2.9429	2.5483	2.6073
CommandR-plus	4.6442	4.7923	4.5857	4.5421	4.6411	4.6386	4.8615	4.5657	4.6231	4.6722	2.7210	2.0667	3.0890	2.4922	2.5922
ONLINE-W	4.6479	4.8128	4.4794	4.4704	4.6026	4.5899	4.8667	4.4595	4.2492	4.5413	-	-	-	-	-
IOL-Research	4.5581	4.8359	4.4635	4.4330	4.5726	4.5618	4.8333	4.4914	4.4642	4.5877	4.0000	4.1615	4.1740	4.0156	4.0878
Llama3-70B	4.5262	4.7205	4.4011	4.3458	4.4984	4.5824	4.7615	4.4648	4.3676	4.5441	3.4738	3.6231	3.6282	3.4953	3.5551
Aya23	4.5187	4.7359	4.4050	4.3115	4.4928	4.4064	4.7641	4.3453	4.3115	4.4568	1.7416	1.2026	2.0876	1.6262	1.6645
ONLINE-A	4.4981	4.8154	4.2948	4.1464	4.4387	4.4270	4.8128	4.3958	4.2274	4.4657	4.2940	4.4128	3.9575	4.0841	4.1871
IKUN	4.2472	4.5692	4.2058	4.2243	4.3116	4.2996	4.6821	4.2869	4.2212	4.3724	4.3502	-	4.3240	4.4174	4.3639
ONLINE-G	4.3614	4.6872	4.1780	3.9097	4.2841	4.2041	4.7128	4.1819	4.1028	4.3004	3.6966	3.6359	3.8486	3.6854	3.7166
Phi-3-Medium	4.1667	4.5333	4.2125	4.1215	4.2585	4.3727	4.6641	4.2590	4.2212	4.3792	1.5768	1.1051	1.8406	1.4735	1.4990
IKUN-C	4.1330	4.4154	4.0385	3.9875	4.1436	4.1929	4.5897	4.1859	4.0031	4.2429	4.1442	4.3897	4.2762	4.2928	4.2758
CUNI-NL	4.0243	4.1590	4.0226	3.9065	4.0281	-	-	-	-	-	-	-	-	-	-
Occiglot	3.7809	4.0231	3.5166	3.7103	3.7577	3.7416	3.7795	3.6813	3.4922	3.6736	-	-	-	-	-
NVIDIA-NeMo	3.7884	3.8231	3.4090	3.4735	3.6235	4.2247	4.5872	3.8353	3.8941	4.1353	-	-	-	-	-
AIST-AIRC	3.3296	3.8590	3.4502	3.2430	3.4704	-	-	-	-	-	-	-	-	-	-
MSLC	2.2397	3.3667	2.7251	2.4673	2.6997	3.0337	4.1821	3.1527	3.0125	3.3452	-	-	-	-	-
TSU-HITs	2.8071	2.4718	2.9920	2.2617	2.6332	2.0449	1.9359	2.4104	2.1433	2.1336	1.5936	1.6667	1.5511	1.7165	1.6320
CycleL	1.0693	1.0308	1.3015	1.1371	1.1346	1.0187	1.0000	1.0863	1.0000	1.0263	1.0449	1.0179	1.2191	1.0218	1.0760
CycleL2	1.0824	1.0308	1.3015	1.1121	1.1317	-	-	-	-	-	-	-	-	-	-

Table 10: System performance by context dependency.

System	**En-De**					**En-Es**					**En-Is**				
	Local	Extended	Global	Universal	*Avg.* ↑	Local	Extended	Global	Universal	*Avg.*	Local	Extended	Global	Universal	*Avg.*
Claude-3.5	4.69	4.71	4.83	4.81	4.76	4.64	4.63	4.76	4.75	4.69	4.69	4.70	4.73	4.81	4.73
Dubformer	4.74	4.72	4.79	4.78	4.76	4.68	4.63	4.77	4.70	4.70	4.78	4.69	4.65	4.72	4.71
Gemini-1.5-Pro	4.75	4.70	4.75	4.75	4.74	4.69	4.68	4.70	4.72	4.70	-	-	-	-	-
Unbabel-Tower70B	4.64	4.69	4.75	4.82	4.72	4.68	4.70	4.74	4.74	4.72	4.70	4.71	4.68	4.70	4.70
GPT-4	4.67	4.69	4.69	4.80	4.71	4.68	4.66	4.74	4.68	4.69	4.49	4.41	4.31	4.52	4.43
ONLINE-B	4.62	4.62	4.74	4.75	4.69	4.38	4.34	4.49	4.67	4.47	4.40	4.34	4.35	4.43	4.38
TranssionMT	4.62	4.62	4.73	4.74	4.68	4.40	4.38	4.45	4.70	4.48	4.42	4.37	4.36	4.45	4.40
CommandR-plus	4.59	4.62	4.69	4.75	4.66	4.63	4.61	4.77	4.78	4.70	3.21	2.55	2.17	2.76	2.67
Mistral-Large	4.61	4.65	4.61	4.75	4.65	4.59	4.57	4.65	4.73	4.63	3.12	2.52	2.35	2.82	2.70
ONLINE-W	4.48	4.58	4.75	4.68	4.62	4.50	4.51	4.65	4.66	4.58	3.81	3.74	3.45	3.88	3.72
IOL-Research	4.54	4.53	4.62	4.66	4.59	4.59	4.51	4.70	4.69	4.62	4.19	4.04	3.99	4.27	4.12
Llama3-70B	4.45	4.46	4.58	4.59	4.52	4.51	4.49	4.65	4.71	4.59	3.75	3.46	3.49	3.69	3.60
Aya23	4.47	4.46	4.51	4.62	4.51	4.42	4.39	4.48	4.63	4.48	2.16	1.59	1.37	1.94	1.76
ONLINE-A	4.34	4.38	4.62	4.64	4.49	4.44	4.40	4.49	4.73	4.51	4.15	4.13	4.22	4.21	4.18
ONLINE-G	4.23	4.22	4.44	4.52	4.35	4.21	4.25	4.33	4.51	4.32	3.70	3.64	3.85	3.69	3.72
Phi-3-Medium	4.24	4.18	4.26	4.54	4.31	4.40	4.31	4.44	4.42	4.39	1.87	1.44	1.26	1.78	1.59
IKUN	4.29	4.28	4.22	4.40	4.30	4.28	4.36	4.30	4.57	4.38	4.33	4.40	4.36	4.25	4.34
IKUN-C	4.13	4.08	4.21	4.27	4.17	4.26	4.15	4.33	4.46	4.30	4.31	4.25	4.16	4.36	4.27
CUNI-NL	4.06	3.99	3.91	4.25	4.05	-	-	-	-	-	-	-	-	-	-
Occiglot	3.41	3.85	3.87	3.65	3.69	3.74	3.69	3.56	3.69	3.67	-	-	-	-	-
NVIDIA-NeMo	3.61	3.61	3.63	3.50	3.59	4.07	4.07	4.31	4.04	4.12	-	-	-	-	-
AIST-AIRC	3.44	3.42	3.60	3.59	3.51	-	-	-	-	-	-	-	-	-	-
MSLC	2.66	2.53	2.84	3.20	2.81	3.23	3.17	3.52	3.78	3.42	-	-	-	-	-
TSU-HITs	3.15	2.54	2.36	2.96	2.75	2.34	2.16	1.64	2.41	2.14	1.70	1.57	1.56	1.68	1.63
CycleL	1.24	1.14	1.06	1.16	1.15	1.05	1.01	1.00	1.16	1.06	1.26	1.04	1.00	1.12	1.11
CycleL2	1.26	1.14	1.02	1.17	1.15	-	-	-	-	-	-	-	-	-	-

Table 11: Performance per skill for En-De.

Skill	Information Density	Idea Development	Terminology Control	Relational Address	Stylistic Register	Modality & Attitude	Reference Consistency	Participant Focus	Logical Connectivity	*Avg.* ↑
System										
Claude-3.5	4.59	4.81	4.82	4.54	4.66	4.67	4.89	4.86	4.71	4.73
Dubformer	4.57	4.71	4.78	4.53	4.73	4.69	4.88	4.84	4.80	4.73
Gemini-1.5-Pro	4.48	4.76	4.76	4.47	4.68	4.67	4.91	4.84	4.80	4.71
Unbabel-Tower70B	4.51	4.71	4.72	4.48	4.60	4.58	4.90	4.87	4.79	4.68
GPT-4	4.51	4.67	4.61	4.52	4.64	4.67	4.88	4.84	4.68	4.67
TranssionMT	4.45	4.81	4.68	4.42	4.54	4.57	4.84	4.78	4.72	4.65
ONLINE-B	4.51	4.76	4.69	4.44	4.53	4.54	4.86	4.81	4.72	4.65
CommandR-plus	4.41	4.71	4.66	4.46	4.49	4.58	4.86	4.77	4.65	4.62
Mistral-Large	4.41	4.48	4.68	4.45	4.57	4.62	4.83	4.74	4.67	4.61
ONLINE-W	4.57	4.57	4.54	4.26	4.48	4.56	4.81	4.74	4.71	4.58
IOL-Research	4.26	4.29	4.60	4.32	4.42	4.47	4.80	4.74	4.65	4.51
Llama3-70B	4.32	4.48	4.47	4.28	4.31	4.45	4.74	4.65	4.51	4.47
Aya23	4.29	4.33	4.45	4.28	4.36	4.42	4.71	4.63	4.56	4.45
ONLINE-A	4.39	4.43	4.45	4.18	4.16	4.36	4.66	4.64	4.61	4.43
ONLINE-G	4.20	4.19	4.33	4.04	4.01	4.20	4.53	4.56	4.42	4.28
IKUN	4.07	4.19	4.20	4.16	4.05	4.15	4.57	4.46	4.41	4.25
Phi-3-Medium	4.01	4.24	4.15	4.13	3.96	4.10	4.56	4.44	4.34	4.21
IKUN-C	3.74	4.14	4.01	3.93	3.99	3.98	4.43	4.28	4.23	4.08
CUNI-NL	3.78	4.14	3.67	3.86	3.88	3.86	4.39	4.20	4.08	3.98
Occiglot	3.78	4.00	3.67	3.47	3.45	3.67	3.94	3.88	3.95	3.76
NVIDIA-NeMo	3.45	3.38	3.59	3.41	3.32	3.46	3.89	3.92	3.64	3.56
AIST-AIRC	3.19	3.38	3.30	3.32	3.11	3.33	3.81	3.65	3.71	3.42
MSLC	2.29	2.52	2.68	2.60	2.32	2.46	2.89	2.96	3.07	2.64
TSU-HITs	2.35	2.14	2.24	2.86	2.65	2.72	3.06	2.74	2.33	2.57
CycleL	1.07	1.10	1.02	1.21	1.20	1.18	1.21	1.07	1.11	1.13
CycleL2	1.04	1.05	1.01	1.22	1.22	1.17	1.19	1.08	1.12	1.12

Table 12: Performance per skill for En-Es.

Skill	Information Density	Idea Development	Terminology Control	Relational Address	Stylistic Register	Modality & Attitude	Reference Consistency	Participant Focus	Logical Connectivity	*Avg.* ↑
System										
Unbabel-Tower70B	4.57	4.81	4.72	4.52	4.69	4.60	4.91	4.85	4.77	4.72
Gemini-1.5-Pro	4.43	4.81	4.78	4.49	4.65	4.60	4.87	4.87	4.76	4.70
Dubformer	4.46	4.71	4.64	4.53	4.71	4.58	4.84	4.79	4.75	4.67
GPT-4	4.49	4.76	4.66	4.46	4.57	4.61	4.88	4.87	4.73	4.67
Claude-3.5	4.52	4.62	4.67	4.42	4.60	4.58	4.79	4.85	4.76	4.65
CommandR-plus	4.51	4.67	4.67	4.42	4.58	4.54	4.83	4.84	4.76	4.65
Mistral-Large	4.33	4.62	4.71	4.45	4.49	4.55	4.82	4.82	4.66	4.61
ONLINE-W	4.39	4.67	4.53	4.35	4.40	4.40	4.71	4.74	4.72	4.55
IOL-Research	4.35	4.43	4.64	4.34	4.45	4.44	4.82	4.74	4.67	4.54
Llama3-70B	4.35	4.52	4.55	4.39	4.35	4.46	4.78	4.75	4.61	4.53
ONLINE-A	4.32	4.52	4.55	4.34	4.33	4.41	4.65	4.63	4.56	4.48
TranssionMT	4.28	4.62	4.50	4.25	4.28	4.41	4.62	4.62	4.58	4.46
ONLINE-B	4.32	4.48	4.37	4.26	4.28	4.38	4.63	4.60	4.60	4.44
Aya23	4.17	4.38	4.54	4.22	4.26	4.36	4.68	4.59	4.59	4.42
IKUN	4.06	4.43	4.32	4.26	4.21	4.24	4.65	4.51	4.54	4.36
Phi-3-Medium	4.17	4.38	4.33	4.13	4.17	4.30	4.60	4.54	4.53	4.35
ONLINE-G	4.23	4.29	4.42	4.04	4.07	4.22	4.50	4.50	4.49	4.31
IKUN-C	4.04	4.43	4.07	4.07	4.17	4.09	4.55	4.39	4.40	4.25
NVIDIA-NeMo	3.96	4.38	4.19	3.88	3.83	3.93	4.36	4.41	4.34	4.14
Occiglot	3.59	3.71	3.32	3.55	3.59	3.63	3.97	3.83	3.84	3.67
MSLC	2.97	3.81	3.61	3.11	2.89	3.13	3.62	3.70	3.86	3.41
TSU-HITs	1.80	1.62	1.84	2.26	2.47	2.04	2.44	2.06	1.69	2.02
CycleL	1.00	1.00	1.00	1.07	1.05	1.03	1.05	1.02	1.00	1.02

Table 13: Performance per skill for En-Is.

Skill	Information Density	Idea Development	Terminology Control	Relational Address	Stylistic Register	Modality & Attitude	Reference Consistency	Participant Focus	Logical Connectivity	*Avg.* ↑
System										
Dubformer	4.39	4.71	4.68	4.61	4.69	4.70	4.86	4.70	4.71	4.67
Unbabel-Tower70B	4.40	4.70	4.57	4.61	4.68	4.60	4.88	4.76	4.72	4.66
Claude-3.5	4.48	4.33	4.72	4.58	4.68	4.65	4.85	4.83	4.76	4.65
AMI	4.16	4.52	4.20	4.34	4.19	4.38	4.61	4.44	4.47	4.37
GPT-4	4.10	4.38	4.08	4.41	4.32	4.29	4.76	4.43	4.45	4.36
TranssionMT	4.01	4.29	4.23	4.38	4.30	4.34	4.59	4.42	4.39	4.33
IKUN	3.97	4.43	4.09	4.28	4.39	4.18	4.64	4.35	4.37	4.30
ONLINE-B	4.10	4.14	4.18	4.34	4.26	4.38	4.60	4.36	4.34	4.30
IKUN-C	3.80	4.14	3.97	4.30	4.21	4.18	4.55	4.16	4.28	4.18
ONLINE-A	3.93	4.29	4.02	4.17	3.95	4.03	4.37	4.23	4.22	4.13
IOL-Research	3.52	4.00	3.58	4.10	4.02	4.02	4.44	4.08	4.14	3.99
ONLINE-G	3.75	3.57	3.13	3.72	3.59	3.88	3.98	3.64	3.87	3.68
Llama3-70B	3.14	3.62	2.93	3.65	3.38	3.40	3.99	3.56	3.53	3.47
Mistral-Large	2.45	2.48	1.76	2.92	2.70	2.46	3.09	2.54	2.51	2.55
CommandR-plus	2.26	2.67	1.57	2.95	2.72	2.68	3.11	2.53	2.37	2.54
Aya23	1.39	1.43	1.14	2.01	1.81	1.57	2.05	1.58	1.49	1.61
TSU-HITs	1.23	1.43	1.55	1.69	1.74	1.39	1.77	1.55	1.41	1.53
Phi-3-Medium	1.28	1.24	1.09	1.78	1.67	1.44	1.77	1.43	1.32	1.45
CycleL	1.01	1.00	1.00	1.15	1.17	1.11	1.11	1.03	1.07	1.07
ONLINE-empty	1.00	1.00	1.00	1.04	1.02	1.02	1.01	1.00	1.00	1.01

E Prompt Lines

E.1 Context Dependency

```
We would like you to label the context dependency of the
following sentence in the {domain} domain. You should classify
the external knowledge needed to translate the sentence
into sentence-level knowledge, local contextual knowledge,
extended contextual knowledge, global contextual knowledge, and
universal contextual knowledge. You must write only one class
without any explanation.

{definition}

{src_lang} Sentence: {sentence}
```

E.2 Translation Skills

```
You are given the following 9 translation skills.

[Skill Options]
{skill}
What are 3 core skills required to translate the following
sentence into a coherent piece of discourse? Especially, select
the primary skills uniquely required to translate into any
languages within the {domain} domain, rather than skills that
could be applied to ordinary sentences.

[{src_lang} Sentence]
{sentence}

Select and write the index of the 3 most primary skills. Also,
write a brief description of how the skill should be applied
when translating within 1-2 sentences for each selected skill.
Finally, after generating two newlines, return a Python list
object that includes each index of 3 skills, arranged in
descending order of importance, from the most important to
the least.

[System]
```

E.3 Main Evaluation

```
We would like to request your feedback on the discourse-level
quality of the translation within the {domain} domain. In your
feedback, please rate the current translation enclosed with
backticks in the following 3 categories, by referring to the
preceding segments and following each scoring rubric.

[Score rubric]
{skills}

[Preceding segments]
{src_lang} source: {prev_src}
{tgt_lang} target: {prev_tgt}

[Current segments]
{src_lang} source: {src_seg}
{tgt_lang} translation: “‘{tgt_seg}“‘

Please give feedback on the translation. Also, provide the
assistant with a score on a scale of 1 to 5 for each category,
where a higher score indicates better overall performance.
Make sure to give feedback or comments for each category
first, followed by the corresponding score for each category.
Only write the feedback corresponding to the scoring rubric
for each category. The scores for each category should be
orthogonal, indicating that ’register’ should not be considered
for ’modality’ category, for example.

[System]
```