# Context is Ubiquitous, but Rarely Changes Judgments: Revisiting Document-Level MT Evaluation

**Ahrii Kim**
AI-Bio Convergence Research Institute
South Korea
ahriikim@gmail.com
trotacodigos/H-FALCON.git

## Abstract

As sentence-level performance in modern Machine Translation (MT) has plateaued, reliable document-level evaluation is increasingly needed. While the recent FALCON framework with pragmatic features offers a promising direction, its reliability and reproducibility are unclear. We address this gap through human evaluation, analyzing sources of low inter-annotator agreement and identifying key factors. Based on these findings, we introduce **H-FALCON**, a **H**uman-centered refinement of FALCON. Our experiments show that, even with limited annotator consensus, H-FALCON achieves correlations comparable to or better than standard sentence-level protocols.

Furthermore, we find that contextual information is inherent in all sentences, challenging the view that only some require it. This suggests that prior estimates such as "n% of sentences require context" may stem from methodological artifacts. At the same time, we show that while context is pervasive, not all of it directly influences human judgment.

## 1 Introduction

The conventional approach to automatic machine translation (MT) evaluation has focused primarily on sentence-level analysis, emphasizing lexical overlap or n-gram similarity, as seen in BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and chrF (Popović, 2015). More recent methods account for semantic similarity through embedding-based metrics such as BERTScore (Zhang et al., 2019) and COMET (Rei et al., 2020), while LLM-based (large language model) metrics, including XCOMET (Guerreiro et al., 2024) and Meta-Metrics (Anugraha et al., 2024), demonstrate improved alignment with human judgments. Despite these advances, their scope remains confined to sentence-level evaluation, failing to capture discourse phenomena such as cohesion, coreference, consistency, and pragmatic adequacy. Document-level metrics have been proposed (Jwalapuram et al. 2021; Zhao et al. 2023; Jiang et al. 2022), but they typically target narrow aspects of discourse and lack comprehensive coverage.

Human evaluation at the document level poses additional challenges due to the complexity of quantifying context-dependent phenomena. Approaches that rely only on overt discourse markers risk underestimating the role of context (Voita et al. 2019; Castilho 2022). Furthermore, protocols vary in context length, annotation granularity, and guideline specificity (Hardmeier et al. 2015; Kocmi et al. 2022). The resulting cognitive burden on evaluators can lead to longer annotation times and reduced inter-annotator agreement (IAA) (Läubli et al., 2018; Bawden et al., 2018; Graham et al., 2017). Collectively, these factors render document-level evaluation both methodologically complex and resource-intensive, limiting its adoption in MT research and practice (Sharma and Sridhar, 2025).

To address this gap, the FALCON framework (**F**unctional **A**ssessment of **L**anguage and **Co**ntextuality in **N**arratives; Kim 2025) integrates pragmatic features into a structured document-level protocol, with LLMs as judges. However, its human evaluation component remains underdeveloped and untested for reproducibility and reliability. We therefore conduct a meta-evaluation of FALCON through human assessments with professional translators, and extend the protocol by introducing H-FALCON, a reproducible and streamlined human evaluation framework. Our contributions are as follows:

- Conduct the first systematic reliability study of FALCON, identifying sources of inter-annotator variation,
- Provide a comprehensive meta-evaluation of FALCON across diverse proprietary models, revealing its limitations,



*Proceedings of the Tenth Conference on Machine Translation*, pages 81–97

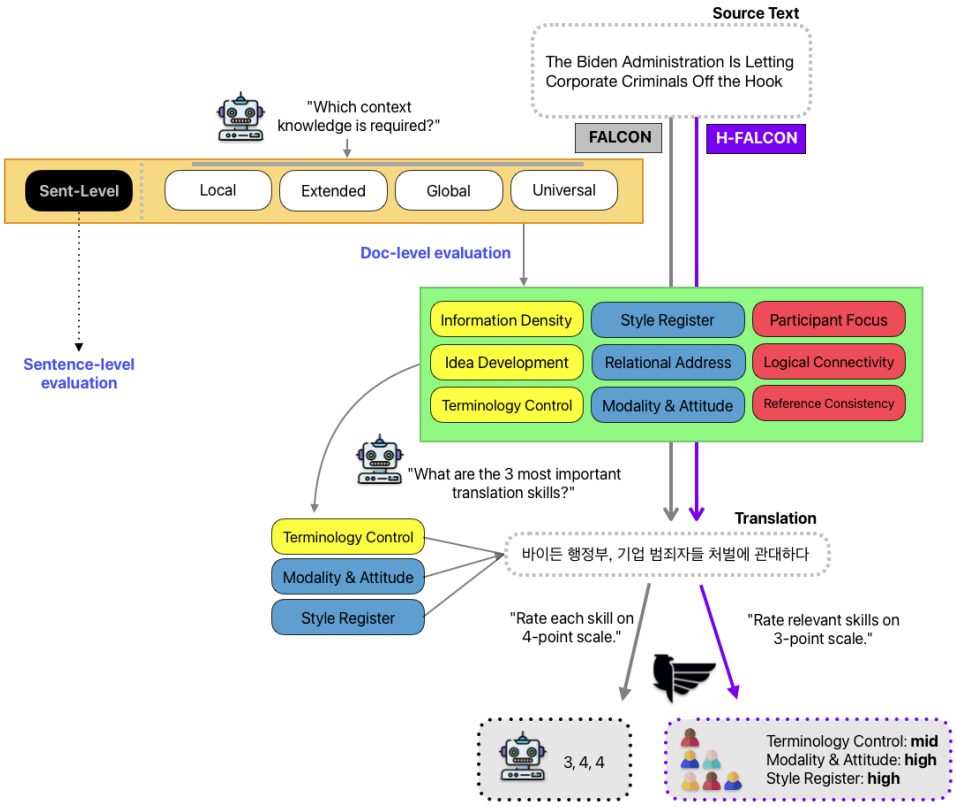



Figure 1: The evaluation process of FALCON consisting of labeling 1) relevant context knowledge and 2) assessment of translation skills, accompanied by 3) rating. This dual-phase process is integrated in H-FALCON by simultaneously conducting labeling and rating for all sentences.

- Introduce H-FALCON, a simplified and reliable protocol for document-level human evaluation,
- Present evidence that contextual information is inherent in all sentences,
- Demonstrate statistically that document-level evaluation contributes 10% to holistic evaluation scores.

## 2 Related Works

Document-level evaluation is not merely a scaled-up version of sentence-level evaluation; it captures translation phenomena that rely on extended context, such as coreference resolution, lexical cohesion, discourse connectives, and pragmatic intent (Thai et al. 2022; Dahan et al. 2024). These features recur across the document, with evidence distributed over multiple segments, shaping a distinctive atmosphere or nuance (Halliday and Matthiessen, 2004). Evaluating such phenomena enables a more accurate assessment of MT systems that appear statistically indistinguishable at the sentence level (Sharma and Sridhar, 2025). This section reviews prior efforts in both manual (§ 2.1) and automatic (§ 2.2) evaluation of document-level phenomena, including FALCON (§ 2.3).

### 2.1 For Manual Evaluation

The most visited sentence-level evaluation frameworks are MQM (Multidimensional Quality Metrics; Lommel et al. 2014) and TAUS DQF (Dynamic Quality Framework; Valli 2015). Their comprehensive error categories encompass some discourse elements such as *Language register* and *Inconsistent use of terminology*,[1], but predominantly focus on textual quality.

Document-level evaluation was initially driven by community efforts such as DiscoMT (Workshop on Discourse in Machine Translation; Hardmeier et al. 2015) and WMT (Conference on Machine Translation). Barrault et al. (2019) proposed a document-level scoring protocol (DR+DC), but its effectiveness was limited by low statistical power, often producing tied rankings. As a result, evaluations were shifted to the sentence level, either by considering adjacent segments (SR+DC) (Barrault et al., 2019) or entire documents (SR+FD)

[1] https://themqm.org/the-mqm-full-typology/

(Akhbardeh et al., 2021) to assess cross-sentence dependencies. The SR+DC approach later became standard practice (Kocmi et al. 2022; Kocmi et al. 2023), with Kocmi et al. (2024) extending the context window to ten consecutive sentences. In parallel, new error categories were introduced for discourse-related issues such as *Accuracy/Gender mismatch* and *Style/Archaic or obscure word choice* (Freitag et al., 2024). While these initiatives primarily focus on contextual conveyance, **our work broadens error typology by shifting from textual to discourse-level quality, systematically incorporating pragmatic, referential, and thematic dimensions into a structured protocol.**

### 2.2 For Automatic Evaluation

On the machine side, several automatic metrics have been developed to better capture discourse and context in MT evaluation. DiscoScore (Zhao et al., 2023) explicitly models discourse relations and coreference chains to assess cohesion and coherence. BlonDE (Jiang et al., 2022) integrates lexical, syntactic, semantic, and discourse-level features, making it suitable for narrative and dialogic text. Doc-COMET (Vernikos et al., 2022) extends COMET (Rei et al., 2020) to accept document-level inputs, leveraging contextual embeddings to evaluate translations within their broader discourse environment. While these approaches mark progress toward automated document-level evaluation, they generally emphasize only one or two discourse aspects—such as coherence or coreference—rather than offering a comprehensive, structured assessment of discourse phenomena.

Another line of research has focused on test suites targeting specific discourse elements. These include domain-specific investigations (Vojtěchová et al. 2019; Biçici 2019; Mukherjee and Yadav 2024; Bhattacharjee et al. 2024; Rozanov et al. 2024; Bawden and Sagot 2023), studies examining linguistic features (Avramidis et al. 2019; Popović 2019; Raganato et al. 2019; Zouhar et al. 2020; Macketanz et al. 2021; Manakhimova et al. 2023; Savoldi et al. 2023; Ármannsson et al. 2024; Friðriksdóttir 2024; Manakhimova and Macketanz 2024; Dawkins et al. 2024), and analyses incorporating discourse phenomena (Rysová et al. 2019; Kocmi et al. 2020; Avramidis et al. 2020; Scherrer et al. 2020; Mukherjee and Shrivastava 2023). DiscoBench (Wang et al., 2023) further addresses discourse-sensitive content, detecting pronoun mistranslation, topic drift, and other cross-sentence errors overlooked by sentence-level metrics.

Overall, these benchmarks highlight that document-level evaluation introduces qualitatively distinct challenges and opportunities, necessitating dedicated protocols and models for holistic MT quality assessment.

### 2.3 The FALCON Framework

FALCON (Functional Assessment of Language and Contextuality in Narratives; Kim 2025) proposes a structured protocol for document-level MT evaluation by incorporating pragmatic and discourse-level factors into a unified scoring scheme. It rests on two hypotheses:

(a) Document-level evaluation can be approximated at the sentence level if contextual information is effectively propagated across sentences.

(b) Such information can be inferred solely from the source, independent of the target language.

Discourse phenomena are classified into three meta-categories (MODE, TENOR, FIELD) and nine subcategories (specified in §3 ) collectively termed "translation skills." For each sentence, the judge selects the three most salient skills, with this restriction enhancing scoring stability.

Sentences not requiring context are first excluded through a labeling step, where annotators assign one of five context types (specified in §3 ). In the subsequent rating stage, each selected skill receives a 4-point score, as illustrated in Figure 1. Scores are then aggregated per segment or skill set to yield interpretable document-level indicators.

This protocol has so far been validated only indirectly: human annotators were asked to judge whether the model' s selections were appropriate, yielding an acceptance rate of 80.4% for context labeling and 71.6% for skill selection. However, no direct evaluation from a classification perspective has been conducted, which is the focus of the present study. Additional concerns may arise from the way context is presented, but this issue falls outside the scope of our work.

## 3 Experiment Setup

We conduct direct human evaluation of FALCON across two tasks and assess whether the current experimental design yields reproducible human judgments (§4). Using these gold annotations, we

| Domain | Dataset | #Doc | #Seg | #Sent/Doc | #Sent/Seg |
|---|---|---|---|---|---|
| Canary | Original | 1 | 1 | – | – |
| | Ours | – | – | – | – |
| Literary | Original | 8 | 206 | 74.13 | 2.88 |
| | Ours | 3 | 76 | 27.67 | 1.09 |
| News | Original | 17 | 149 | 19.53 | 2.23 |
| | Ours | 12 | 233 | 16.67 | 1.01 |
| Social | Original | 34 | 531 | 22.76 | 1.46 |
| | Ours | 23 | 500 | 23.26 | 1.07 |
| Speech | Original | 111 | 111 | 6.49 | 6.49 |
| | Ours | – | – | – | – |
| All | Original | 171 | 998 | 30.73 | 3.27 |
| | Ours | 38 | 809 | 22.53 | 1.06 |

Table 1: Comparison of dataset statistics between the original WMT24++ corpus and our filtered dataset. Here, **Seg** denotes a segment (paragraph in WMT24++), and sentence counts are reported per document and segment.

further perform a meta-evaluation to validate the framework's reliability (§5). The tasks are:

**Task I: Context Knowledge** Judges assign one of five levels of contextual knowledge required for translation: Sentence-level, Local, Extended, Global, Universal.

**Task II: Translation Skills** Judges select the three most relevant skills from nine predefined categories: Information Density, Idea Development, Terminology Control, Style Register, Relational Address, Modality & Attitude, Reference Consistency, Participant Focus, Logical Connectivity.

### 3.1 Dataset

The original WMT24++ English–Korean dataset (Deutsch et al., 2025) contains 998 segments from four domains (social, news, speech, and literary) with translations from ten systems. Because many segments span multiple sentences, it is unsuitable for our *sentence-level* design.

We construct a filtered subset while preserving domain balance. The speech domain is removed, as each document corresponds to a single segment without context, and the literary domain is partially pruned due to disproportionate length. Sentences with hyperlinks, hashtags, or timestamps are discarded, while emojis and user tags are retained as they are considered relevant to the evaluation.

The remaining data are re-segmented into individual sentences using NLTK (Bird et al., 2009) for English and KSS[2] for Korean. Source, target, and reference segments are automatically aligned with newline markers and then manually verified. For translation, we select the best-performing system (based on COMET scores), assuming that context-aware translation is unlikely from low-quality systems.

The final evaluation set consists of 809 unique sentences across three domains (social, news, and literary), preserving proportional domain distribution (Table 1). All retained segments preserve document boundaries, with sentence order tracked by custom IDs.

### 3.2 Recruitment & Training

We recruited three professional translators, all native Korean speakers with 5–10 years of English translation experience. For confidentiality, they were anonymized as Judge 1, Judge 2, and Judge 3 and are collectively referred to as judges. Based on the reduced segment length relative to the original dataset, we estimated an average throughput of 60 sentences per hour, corresponding to 13.3 hours per task and 27 hours in total per judge across two tasks. Judges were compensated at $30 per hour.

An online orientation was conducted via Google Meet to introduce the evaluation guidelines and demonstrate the platform. During the session, participants performed a preliminary evaluation using the platform. For the main study, judges were given one week to complete their evaluations. Time was tracked per item, and participants were instructed to maintain focus during annotation. They were provided full access to the document and permitted to review and revise their annotations prior to final submission.

### 3.3 Platform & Interface

We used Label Studio[3] as the evaluation platform (see Figure 9 in the Appendix). The interface allowed evaluators to consult label definitions, prior annotations, and relevant domain information throughout the task.

### 3.4 Metrics

We use IAA as the primary metric of reproducibility. For Task I, Cohen's Kappa ($\kappa$) is computed for each judge as in Equation 1, where $P_o$ denotes the observed agreement and $P_e$ the expected agreement under chance.

[2] https://github.com/hyunwoongko/kss
[3] https://labelstud.io

| | ($\kappa$) ↑ | ($J$) ↑ |
|---|---|---|
| Judge 2–Judge 3 | 0.4995 | 0.6098 |
| Judge 1–Judge 2 | 0.3883 | 0.4629 |
| Judge 3–Judge 1 | 0.3646 | 0.4529 |
| **Avg.** | **0.4175** | **0.5085** |

Table 2: Pairwise IAA scores for Task I (Cohen's Kappa) and Task II (Jaccard similarity).

$$\kappa = \frac{P_o - P_e}{1 - P_e} \quad (1)$$

For Task II, which involves multi-label annotation, we use the Jaccard similarity $J$ (Equation 2), where $A$ and $B$ are the label sets from two annotators. For qualitative assessment, we collect participant feedback via Google Sheets and conduct subsequent linguistic analysis.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (2)$$

## 4 Result

### 4.1 Reproducibility

Table 2 reports IAA for the two tasks. Across tasks, judges reach a **fair to moderate level of agreement** according to empirical standards (Landis and Koch 1977; Zhang and Zhou 2014; Rajpurkar et al. 2016), with average scores of $\kappa = 0.42$ and $J = 0.51$, and maximum scores of $\kappa = 0.50$ and $J = 0.61$. Agreement is highest between Judge 2 and Judge 3, suggesting that Judge 1 applied different criteria.

### 4.2 Analysis

We analyze disagreement by computing the proportion of pairwise label mismatches per task. For each judge pair, we identify the labels on which they disagreed and calculate their distribution. As shown in Figure 2, the largest divergence arises in the SENTENCE-LEVEL and LOCAL categories, accounting for 39.7% and 36.4% of disagreements, respectively, between Judge 1 and Judge 2.

To further examine this confusion, we merge related labels and recompute IAA. As shown in Table 3, the primary source of disagreement across judges lies in distinguishing SENTENCE-LEVEL from LOCAL. Merging these categories increases agreement from $\kappa = 0.4995$ to $\kappa = 0.58$.

To better understand this ambiguity, we analyze qualitative feedback on the difficulty of distinguishing context-independent (SENTENCE-LEVEL) from context-dependent labels. A recurring theme is the treatment of pronouns. For example, when the English pronoun "it" is translated into an equivalent pronoun in Korean and judged correct, the label is typically SENTENCE-LEVEL. By contrast, if the same translation is considered inadequate—requiring explicit mention of the referent noun—the label shifts to LOCAL. An illustrative case is shown in Table 4. As Judge 2 noted, "the interpretation of a pronoun's referent also influences verb choice, and thus I categorize the sentence as LOCAL."

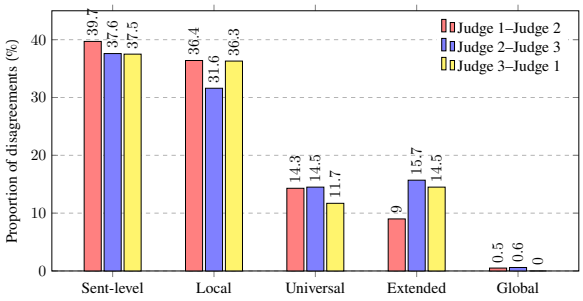



Figure 2: Distribution of Task I label disagreements across judge pairs (%).

| Group | $J_1$–$J_2$ ($\Delta$) | $J_2$–$J_3$ ($\Delta$ ↑) | $J_3$–$J_1$ ($\Delta$) |
|---|---|---|---|
| L + S | **0.482 (+0.093)** | **0.580 (+0.080)** | **0.454 (+0.090)** |
| E + U | 0.411 (+0.022) | 0.568 (+0.069) | 0.411 (+0.046) |
| E + L | 0.414 (+0.026) | 0.500 (+0.001) | 0.397 (+0.033) |
| E + S | 0.372 (–0.016) | 0.480 (–0.020) | 0.340 (–0.025) |
| L + U | 0.372 (–0.017) | 0.463 (–0.036) | 0.343 (–0.022) |
| S + U | 0.298 (–0.091) | 0.418 (–0.081) | 0.264 (–0.100) |

Table 3: Cohen's $\kappa$ after merging two labels. Parentheses indicate the change from the original $\kappa$. The highest agreement per column is shown in bold. Labels are abbreviated as L = Local, S = Sentence-level, E = Extended, and U = Universal. Judges are abbreviated as $J_i$.

| | |
|---|---|
| **SRC** | I bought it like that and couldn't modify it, so I had to design around it. |
| **TGT** | 구매했을 때부터 그런 형태였고, 수정할 수 없어서 그 형태에 맞춰 디자인해야 했어요. |
| **BT** | *It was in that form from the moment I purchased it, and since I couldn' t change it, I had to design everything to fit that shape.* |

Table 4: A notable instance of pronoun provoking frequent misunderstanding between SENTENCE-LEVEL and LOCAL labels. The source (SRC) and target (TGT) segments are exemplified with the help of back-translation (BT).

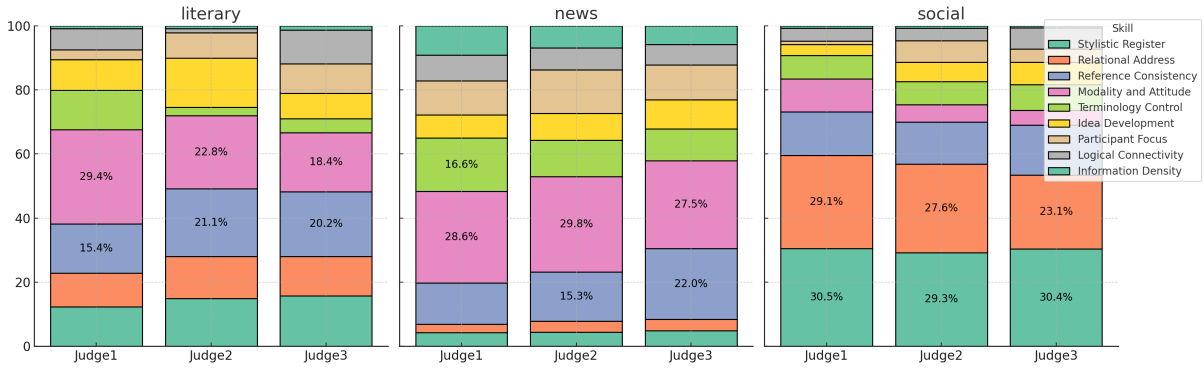



Figure 3: Distribution of Task II label choices across domains and judges. Values are shown for the largest slices.

### 4.3 Analysis

While sentence-level agreement on selected translation skills is limited, we analyze the distribution of skill choices per domain and per judge. Figure 3 shows that the three judges assign broadly similar proportions of skill labels across domains, suggesting that individual-level disagreements in exact label sets do not obscure shared evaluative priorities.

Closer inspection reveals domain-specific emphases. In the social domain, judges consistently highlight Style Register (avg. 30.1%) and Relational Address (26.6%), reflecting the importance of interpersonal stance in user-generated content. In the news domain, Modality & Attitude (28.6%) and Reference Consistency (16.7%) dominate, consistent with the demands for precision and coherence in reporting—a tendency also observed in literary text, where Modality & Attitude (23.5%) and Reference Consistency (18.9%) are most frequent. This indicates that low pairwise agreement does not necessarily reflect fundamental divergence, but rather differences in specific label selection. At the same time, the results point to a limitation of the current protocol: constraining annotators to exactly three skills per segment may not capture the full range of relevant judgments.

## 5 Meta-Evaluation of FALCON

The highest human IAA in our configuration is $\kappa = 0.50$ for Task I and $J = 0.61$ for Task II. Using these gold scores as reference, we evaluate the reliability of FALCON as an LLM-as-judge framework. As baselines, we test multiple proprietary models—OpenAI's `gpt-o3`, `o4-mini`, and the baseline from Kim (2025), `4.1-mini`. Model performance is assessed using the same reproducibility metrics defined in §3.4, complemented by accuracy for Task I, where the output is a single categorical label, and Micro-F1 for Task II, where multiple labels must be selected simultaneously.

| Group | Pair | acc (%)↑ | $\kappa$ |
|---|---|---|---|
| vs. | **$J_2$, $J_3$** | **70.09** | **0.4995** |
| | $J_1$, $J_2$ | 66.25 | 0.3883 |
| | $J_1$, $J_3$ | 62.92 | 0.3646 |
| vs. | **$J_3$, `o4-mini`** | **53.89** | **0.2535** |
| | $J_1$, `o4-mini` | 52.29 | 0.1788 |
| | $J_2$, `o4-mini` | 51.67 | 0.1891 |
| | $J_3$, `o3` | 51.17 | 0.2059 |
| | $J_1$, `o3` | 50.31 | 0.1484 |
| | $J_2$, `o3` | 49.57 | 0.1591 |
| | $J_1$, `4.1-mini` | 42.77 | 0.0802 |
| | $J_2$, `4.1-mini` | 39.80 | 0.0478 |
| | $J_3$, `4.1-mini` | 39.68 | 0.0750 |
| vs. | **`o3`, `o4-mini`** | **71.69** | **0.5239** |
| | `4.1-mini`, `o4-mini` | 47.22 | 0.2046 |
| | `o3`, `4.1-mini` | 40.30 | 0.1068 |

Table 5: Pairwise accuracy and Cohen's Kappa $\kappa$ by human () and model () groups for Task I.

### 5.1 Reliability of context knowledge

Table 5 reports pairwise accuracy and IAA across human–human, human–model, and model–model comparisons. The best-performing model, `o4-mini`, achieves 53.89% accuracy, which falls short of even the weakest human pair (Judge 1–-Judge 3, 62.92%). No model approaches the agreement level of the strongest human pair (Judge 2–-Judge 3). The concordance with human annotations remains at most **fair** ($\kappa = 0.25$ for `o4-mini`), underscoring the limited ability of current LLMs to reliably distinguish context categories at a human-comparable level.

To better understand this gap, we analyze which labels drive model–human discrepancies. Figure 4

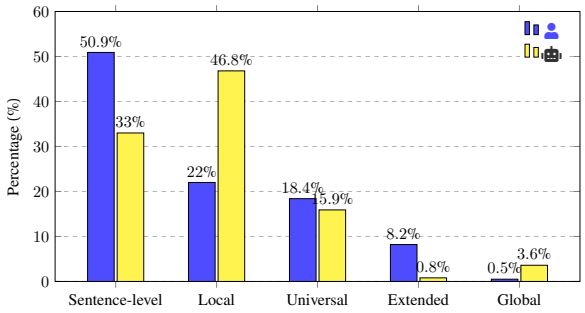



Figure 4: Task I label distribution of disagreements between Judge 2 ( ) and `o4-mini` ( ), identified as the most aligned human–model pair.

illustrates the disagreement distribution between Judge 2 and `o4-mini`, the pair with the highest human–model consensus. The largest share of divergence arises from SENTENCE-LEVEL (50.9%) from the human part and from LOCAL (46.8%) from the machine part. These categories mirror the main sources of confusion among human annotators, suggesting that while models replicate human-like weaknesses, they lack the robustness to resolve such ambiguities consistently.

### 5.2 Reliability of translation skill

Table 6 shows that the strongest human–model agreement is attained with `o4-mini` ($J = 0.406$), substantially lower than both human–human and model–model levels. Model precision reaches 53.6%, comparable to the earlier task, but still insufficient to approximate human reliability. Interestingly, model–model agreement is relatively high, reaching up to $J = 0.597$, on par with the stronger human–human pairs.

These findings suggest that models produce consistent predictions across systems, yet this consistency reflects shared internal heuristics rather than alignment with human reasoning. While human annotators converge through pragmatic interpretation, models seem to exploit surface-level patterns that do not fully capture evaluative criteria. Closing this gap demands not just higher accuracy, but agreement with humans based on human-like reasoning.

### 5.3 Summary

The central hypothesis of FALCON—that document-level evaluation can be approximated at the sentence level—requires caution. Our results show that judges often confuse adjacent levels of context, underscoring the need for clearer definitions

| Group | Pair | avg. $J\uparrow$ | f1 |
|---|---|---|---|
| vs. | $J_2$–$J_3$ | **0.6098** | **0.7183** |
| | $J_1$–$J_2$ | 0.4629 | 0.5915 |
| | $J_1$–$J_3$ | 0.4529 | 0.5737 |
| vs. | $J_2$, `o4-mini` | **0.4067** | **0.5360** |
| | $J_3$, `o4-mini` | 0.3976 | 0.5272 |
| | $J_2$, `o3` | 0.3970 | 0.5231 |
| | $J_1$, `o4-mini` | 0.3912 | 0.5196 |
| | $J_1$, `o3` | 0.3829 | 0.5099 |
| | $J_2$, `4.1-mini` | 0.3704 | 0.4931 |
| | $J_3$, `4.1-mini` | 0.3683 | 0.4893 |
| | $J_1$, `4.1-mini` | 0.3660 | 0.4871 |
| | $J_3$, `o3` | 0.3625 | 0.4854 |
| vs. | `o3`, `o4-mini` | **0.5972** | **0.7082** |
| | `4.1-mini`, `o4-mini` | 0.4665 | 0.5948 |
| | `4.1-mini`, `o3` | 0.4250 | 0.5554 |

Table 6: Average pairwise Jaccard Similarity $J$ and Micro F1 between human ( ) and model ( ) groups for Task II.

of "context." Furthermore, the low agreement in Task II suggests that identifying universal translation skills solely from the source text risks poor reproducibility of gold judgments.

## 6 Refined Protocol: H-FALCON

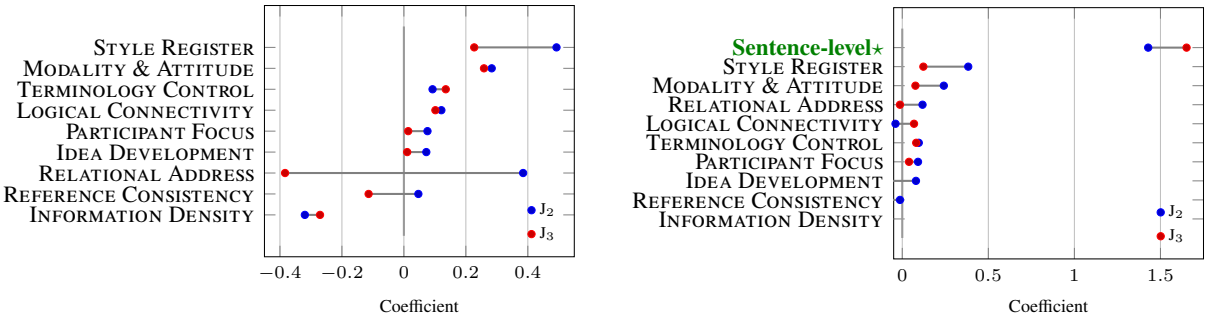
The current protocol of FALCON suffers from ambiguous definitions of context and limited reproducibility in skill selection, calling into question its central hypotheses. Building on these findings, we identify three structural limitations of FALCON: unclear translation objectives for human evaluators, the rigid requirement to assign exactly three skills per sentence, and the lack of adaptability to the domain and language pair.

To address them, we propose H-FALCON (**H**uman-centered FALCON), grounded in two revised hypotheses: (i) every sentence is influenced by context, and (ii) judges should flexibly decide the number of translation skills.

### 6.1 Design

Given these assumptions, H-FALCON removes Task I, since all sentences are subject to evaluation. For Task II, rather than selecting a fixed set of relevant skills, judges directly evaluate the pertinence of each skill, thereby unifying annotation and rating into a single step (Figure 1).

To support this protocol, every skill is initialized as NOT RELEVANT. Judges then assign one of three




(a) Doc Features

(b) Doc + Sent Features

Figure 5: Linear regression coefficients for Judge 2 ($J_2$) and Judge 3 ($J_3$) with (b) and without (a) **sentence-level** score. Features with scores near 0 have minimal influence on the holistic score.

ratings—High, Medium, or Low—following House's theoretical framework (House, 2015). This triadic scale replaces the 4-point scheme of Kim (2025), to represent preliminary feedback from our evaluators that three levels suffice, as discourse phenomena often lend themselves to relatively clear judgments.

### 6.2 Experiment

To verify the reproducibility of the refined H-FALCON protocol, we sample 300 new instances from WMT24++ (Deutsch et al., 2025) that are not included in the earlier experiments. Human evaluation is conducted by Judge 2 and Judge 3, the pair with the highest agreement in prior tasks.

The evaluation environment remains unchanged, using the same platform as in Figure 10. In this setting, judges simultaneously select and rate relevant skills, eliminating the separation of annotation and scoring. To provide additional baselines, we also collect MQM-style sentence-level error annotations on a 4-point scale and holistic quality scores (sentence + document level) on a 10-point scale. These parallel evaluations allow us to establish a benchmark IAA threshold for H-FALCON and to examine relationships among the three metrics. All ratings are obtained at the sentence level, and scale variation is deliberately employed to minimize task confusion.

The reliability of skill selection is measured by excluding Not Relevant labels and computing Jaccard similarity between the two judges. Correlations between evaluation metrics are quantified using Pearson, Spearman, and Kendall's tau coefficients.

### 6.3 Reproducibility of H-FALCON

The Jaccard similarity for overlapping translation skills between the two judges is 0.532, remaining consistently low and consistent with the earlier experiment. This highlights the inherent difficulty of achieving consensus, regardless of the method of label collection.

To further examine how the judges weigh each skill when assigning holistic scores, we fit a linear regression model for each judge, using the holistic score as the dependent variable and the individual label scores as predictors (with an intercept). This analysis quantifies the relative contribution of each skill while controlling for the others. As shown in Figure 5 (a), the judges diverge most clearly on Relational Address: Judge 2 associates higher holistic scores with stronger performance in this skill, whereas Judge 3 tends to assign lower scores. A similar but weaker divergence is observed for Reference Consistency. Importantly, these opposite directions remain significant within 95% confidence intervals, underscoring that the divergence reflects genuine differences in evaluative criteria rather than statistical noise. These divergent patterns suggest that the guidelines for the labels may require refinement and additional evaluator training to ensure consistent application.

### 6.4 Further Analysis

#### H-FALCON score as a proxy measure

We examine whether the obtained labels can serve as proxies for document-level scoring. Each annotation is assigned a numerical value (High=3, Medium=2, Low=1, Not Relevant=0), and scores are computed either by aggregating values ("sum") or by counting non-zero labels. Correlation between the two judges across sentence-, document-, and holistic-level scores (Table 7) indicates that the document-level scheme achieves agreement comparable to sentence-level evaluation ($\rho = 0.55$ vs.

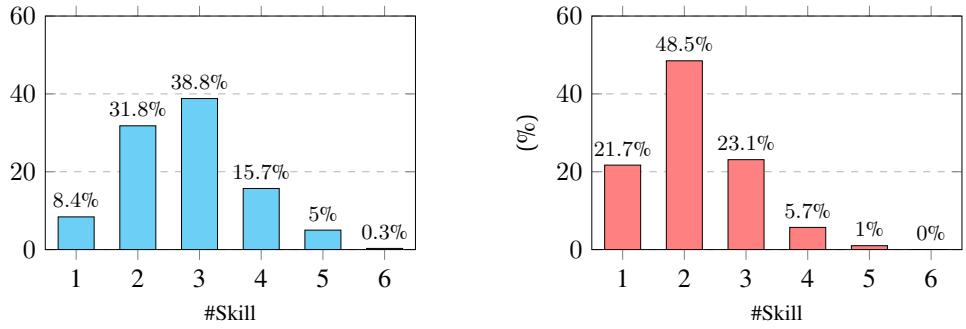



Figure 6: Distribution of the number of selected skills per sentence for each judge (left: Judge 2, right: Judge 3).

0.44). Notably, the counting method yields slightly higher consensus (0.55 vs. 0.48), highlighting its potential as an effective approach for annotating document-level quality.

| Type | Pearson | Spearman | Kendall |
|---|---|---|---|
| Sentence-level | 0.494 | 0.441 | 0.413 |
| H-FALCON (sum) | 0.499 | 0.483 | 0.378 |
| H-FALCON (count) | 0.562 | 0.545 | 0.486 |
| Holistic | 0.650 | 0.587 | 0.502 |

Table 7: Correlations between two raters across sentence-level, document-level (using two aggregation styles), and holistic scores.

### Limited explanatory power of document-level score

To further assess the relative impact of sentence- and document-level features on holistic judgments, we extend the regression model by adding the sentence-level score as an independent variable. As shown in Figure 5 (b), the sentence-level score is the strongest predictor of holistic quality, with coefficients of 1.43 (95% CI: 1.22–1.63) for Judge 2 and 1.65 (95% CI: 1.49–1.82) for Judge 3.

Table 8 reports the explanatory power ($R^2$) of models with and without the sentence-level score. Document-level scores alone account for little variance in holistic judgments ($R^2 = 0.11$), explaining only 11% of the variance in holistic judgments. However, incorporating the sentence-level score increases explanatory power to 0.54 and reduces the intercept from 7.11 to 2.29. These results confirm that sentence-level quality is the primary driver of holistic assessments.

### At least one discourse feature per sentence

We calculate the number of translation skills annotated per judge. Figure 6 shows that every sentence is annotated with at least one skill, most frequently with three to four skills (38.8% and 48.5% for Judge 2 and Judge 3, respectively). This finding challenges the claim that only a subset of sentences requires contextual information (Castilho, 2022). On the contrary, we emphasize that contextual information can influence translation in all cases—even for simple utterances such as "hi." However, as shown in the previous section, its impact on the holistic score is relatively limited. Still, this does not diminish the importance of document-level evaluation, which remains a key factor for distinguishing higher-performing models.

| | Doc | | | Doc + Sent | | |
|---|---|---|---|---|---|---|
| | $J_2$ | $J_3$ | Avg | $J_2$ | $J_3$ | Avg |
| $R^2$ | 0.12 | 0.09 | 0.11 | 0.47↑ | 0.61↑ | 0.54↑ |
| Intercept | 6.46 | 7.76 | 7.11 | 2.10↓ | 2.48↓ | 2.29↓ |

Table 8: The explanatory power ($R^2$) of models with document-level score (**Doc**) and with document- and sentence-level scores (**Doc+Sent**). Doc+Sent results are highlighted.

## 7 Conclusion

Our findings challenge prevailing assumptions in MT evaluation by demonstrating that contextual information, though modest in magnitude, is both universal and consequential for human judgment. Operationalizing this insight, H-FALCON provides a reproducible, context-aware evaluation protocol that aligns as closely with human preferences as traditional sentence-level approaches. These results underscore the need to move beyond narrow, sentence-bounded metrics toward richer document-level assessments that capture the pragmatic realities of translation quality. As MT performance converges at the sentence level, such holistic, context-sensitive evaluation will be essential for driving the next phase of progress in the field.

## 8 Limitation

Our study is limited to a single mid-resourced language pair. While this is acceptable given our focus on the human evaluation setting—which is largely consistent across languages—the reproducibility and reliability of FALCON may be underestimated. For the same reason, we did not experiment with other open-weight models such as LLaMA or Mistral.

On the human side, only three annotators were engaged, one of whom showed notably divergent behavior. In addition, even under the refined protocol, the consensus on translation skills remained low (§ 6.3). These issues highlight the need for more proactive calibration sessions among annotators.

Finally, we did not investigate how context should be presented or which types of context were most informative on the target side for FALCON. We leave this as an avenue for future work.

## 9 Acknowledgment

This research was supported by G-LAMP Program of the National Research Foundation of Korea (NRF) grant funded by the Ministry of Education (No. RS-2025-25441317).

## 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.

David Anugraha, Garry Kuwanto, Lucky Susanto, Derry Tanti Wijaya, and Genta Winata. 2024. MetaMetrics-MT: Tuning meta-metrics for machine translation via human preference calibration. In *Proceedings of the Ninth Conference on Machine Translation*, pages 459–469, Miami, Florida, USA. 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*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. 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.

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.

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.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python*: *Analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

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 Thirteenth Language Resources and Evaluation Conference*, pages 3017–3025, Marseille, France. European Language Resources Association.

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.

Daniel Deutsch, Eleni Briakou, Isaac Caswell, Max Finkelstein, Roni Galor, and 1 others. 2025. WMT24++: Expanding the language coverage of wmt24 to 55 languages & dialects. *arXiv preprint arXiv:2502.12404*.

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, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2017. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23(1):3–30.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

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

Christian Hardmeier, Liane Guillou, Pierre Lison, and Jörg Tiedemann. 2015. Report on the discomt 2015 shared task on discourse translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 1–16. ACL.

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

Zheng Jiang, Yang Yu, Yang Feng, Bing Qin, and Ting Liu. 2022. Blonde: An automatic evaluation metric for document-level natural language generation. In *Proceedings of NAACL*, pages 1679–1698.

Prathyusha Jwalapuram, Barbara Rychalska, Shafiq Joty, and Dominika Basaj. 2021. Dip benchmark tests: Evaluation benchmarks for discourse phenomena in {mt}.

Ahrii Kim. 2025. Falcon: Holistic framework for document-level machine translation evaluation. *TechRxiv*.

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. 2024. 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, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, Maja Popović, and Mariya Shmatova. 2022. Findings of the 2022 conference on machine translation (wmt22). In *Proceedings of the Seventh Conference on Machine Translation*, pages 1–45, Abu Dhabi. 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*.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Arle Richard Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2014. Multidimensional quality metrics (mqm): A framework for defining translation quality. In *Proceedings of the Eighth Conference on International Language Resources and Evaluation (LREC'14)*, pages 1285–1291. European Language Resources Association (ELRA).

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.

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.

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. ACL.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

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.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pages 2383–2392. 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.

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.

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*).

Himanshu Sharma and Bharat Ram Sridhar. 2025. Document-level machine translation through discourse modelling: A survey. *CFILT - IITB*.

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.

Paola Valli. 2015. The TAUS quality dashboard. In *Proceedings of the 37th Conference Translating and the Computer*, pages 127—136, London, UK. AsLing.

Giorgos Vernikos, Brian Thompson, Prashant Mathur, and Marcello Federico. 2022. Embarrassingly easy document-level MT metrics: How to convert any pretrained metric into a document-level metric. In *Proceedings of the Seventh Conference on Machine Translation* (*WMT*), pages 118–128, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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.

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

Longyue Wang, Zefeng Du, Donghuai Liu, Deng Cai, Dian Yu, Haiyun Jiang, Yan Wang, Leyang Cui, Shuming Shi, and Zhaopeng Tu. 2023. Disco-bench: A discourse-aware evaluation benchmark for language modelling. *Preprint*, arXiv:2307.08074.

Min-Ling Zhang and Zhi-Hua Zhou. 2014. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26 (8):1819–1837.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675.

Wei Zhao, Michael Strube, and Steffen Eger. 2023. DiscoScore: Evaluating text generation with BERT and discourse coherence. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3865–3883, Dubrovnik, Croatia. Association for Computational Linguistics.

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*.

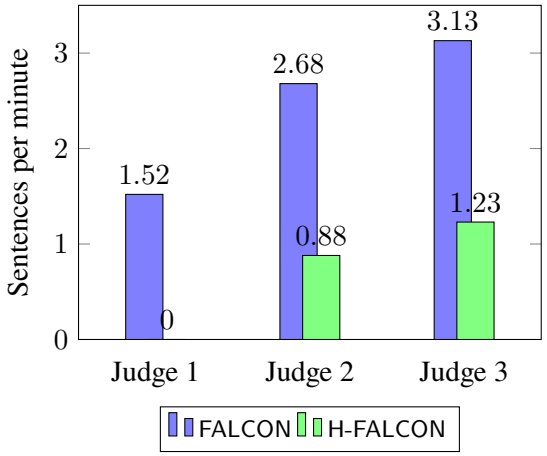



Figure 7: Average throughput per judge in FALCON vs. H-FALCON. Judge 1 was not hired for H-FALCON.

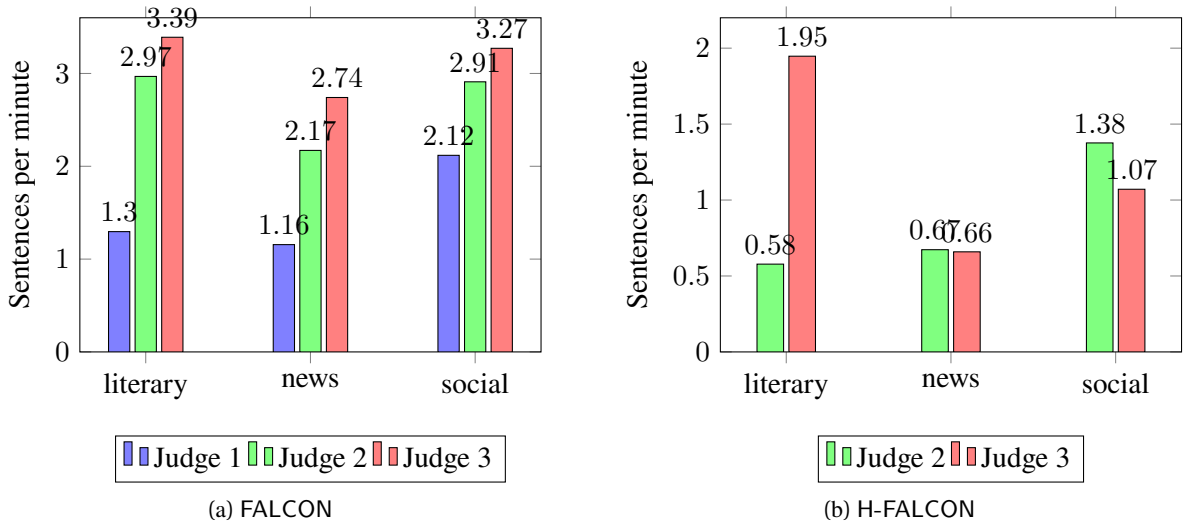
# Appendix

## A Evaluation Throughput

We calculate throughput per judge under two different frameworks: FALCON and H-FALCON. The key distinction is that the H-FALCON setting requires both label annotation and rating, which introduces additional cognitive load and time, whereas the FALCON condition measures throughput without the rating phase.

Figure 7 shows that throughput values are consistently lower in H-FALCON than in FALCON, reflecting the extra annotation steps. For example, the average throughput per judge decreases from 1.52–3.13 sent/min in FALCON to 0.88–1.23 sent/min in H-FALCON. This suggests that rating is the most time-consuming component of the evaluation: despite H-FALCON consolidating the task into a single step, throughput falls to less than half of FALCON, indicating that the rating phase dominates the overall processing time.

When examining domain-level performance in Figure 8, consistent patterns emerge across both setups. Social texts yield the highest throughput, reflecting their relatively simple and conversational style, while literary texts slow down judges the most, likely due to complex syntax and stylistic density. News texts fall in between, with moderate difficulty and processing speed. This ordering is preserved in both FALCON and H-FALCON, though absolute throughput values are lower in the latter due to the added annotation and rating tasks. These results confirm that genre characteristics strongly shape translation throughput, and that such effects remain robust even under heavier annotation requirements.




(a) FALCON (b) H-FALCON

Figure 8: Throughput by domain and judge across FALCON and H-FALCON setups. Higher values indicate faster processing.

## B Descriptions of Context Levels

**Sentence-level** 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.

**Local** Understanding requires minimal surrounding context — maybe the previous or next sentence — but nothing broader. Without it, pronouns, references, or logical connectors might be confusing.

**Extended** Grasping the meaning requires understanding the broader scene, paragraph, or emotional flow. Cultural nuance, emotional undertones, or evolving character perspectives start to matter.

**Global** 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.

**Universal** Understanding draws on extensive external knowledge — history, philosophy, science, mythology, social structures, or famous world events. Without that shared knowledge, translation risks misfiring badly.

## C Descriptions of Translation Skills

**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.

**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.

**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.

**Style 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.

**Reference Consistency** Does the sentence contain elements that refer to the same entity within

(a) Task I

| Label | $J_1$ | $J_2$ | $J_3$ | Avg↑ |
|---|---|---|---|---|
| Sent-level | 63.16 | 60.57 | 54.14 | 59.29 |
| Local | 23.11 | 21.76 | 26.33 | 23.73 |
| Universal | 10.88 | 13.23 | 10.01 | 11.37 |
| Extended | 2.84 | 4.08 | 9.52 | 5.48 |
| Global | 0.00 | 0.37 | 0.00 | 0.12 |

(b) Task II

| Label | $J_1$ | $J_2$ | $J_3$ | Avg↑ |
|---|---|---|---|---|
| Style Register | 21.26 | 20.77 | 21.67 | 21.23 |
| Relational Address | 19.70 | 19.28 | 16.44 | 18.47 |
| Reference Consistency | 13.51 | 14.50 | 17.84 | 15.28 |
| Modality and Attitude | 17.39 | 14.05 | 12.48 | 14.64 |
| Terminology Control | 10.47 | 7.95 | 8.24 | 8.89 |
| Idea Development | 5.07 | 7.62 | 7.70 | 6.80 |
| Participant Focus | 4.04 | 8.82 | 6.55 | 6.47 |
| Logical Connectivity | 5.40 | 4.49 | 6.84 | 5.58 |
| Information Density | 3.17 | 2.51 | 2.22 | 2.63 |

Table 9: Proportion of Task I, II labels annotated by three judges (%).

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.

**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.

**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.

**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.

**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.

## D Analysis of Collected Data

Table 9-(a) reports the number of annotations per context type, indicating broadly consistent distributions across judges. Roughly 60% of sentences were judged as translatable without additional context, though the exact subset of sentences varied considerably by annotator. Among context-dependent categories, Local was the most frequent, averaging 24%. By contrast, Global was almost never selected, suggesting that this type of context is difficult to capture reliably at the sentence level.

Turning to translation skills in Table 9-(b), Style Register (21.23%) and Relational Address (18.47%) emerged as the most frequently required skills, aligning with qualitative feedback that highlights their importance in context-sensitive translation. Conversely, Information Density was rarely chosen (2.6%), which may reflect either limited judge awareness or the relatively low salience of this feature in the dataset. These observations underscore the need for further clarification of certain skill definitions to improve annotation reliability.

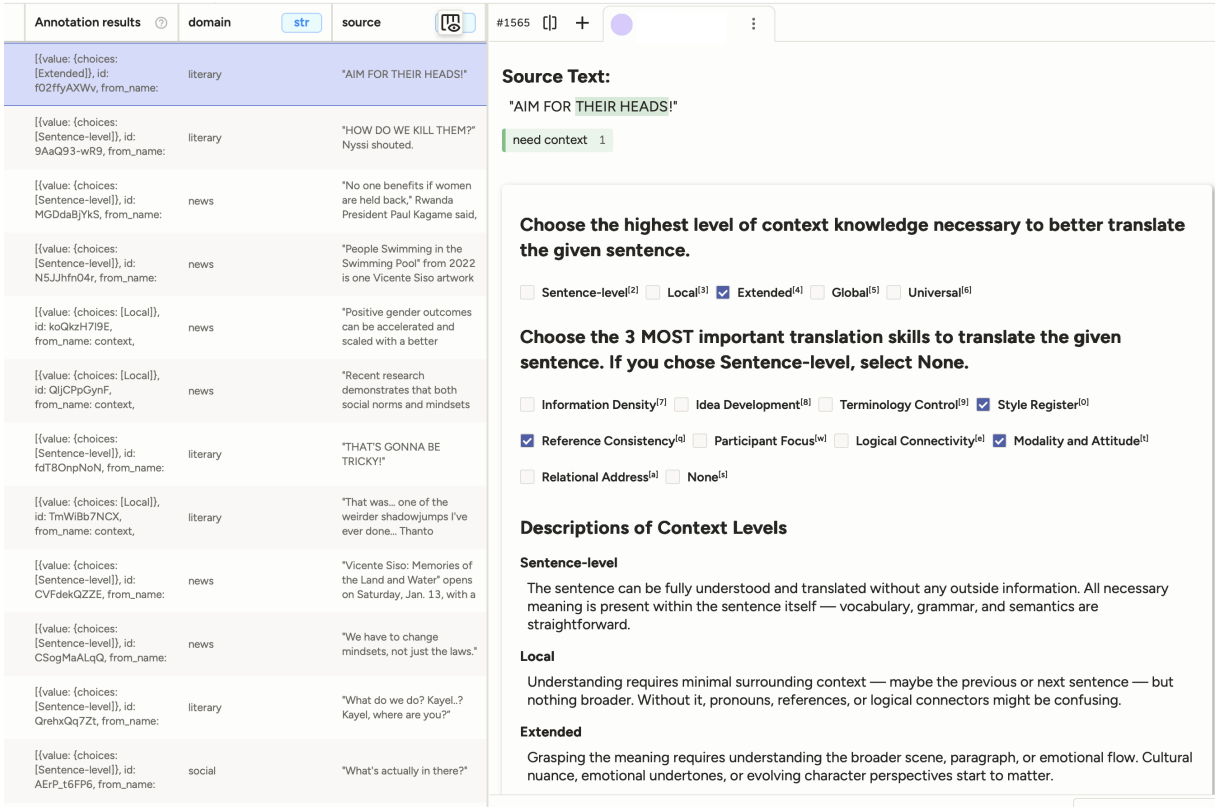



Figure 9: Label Studio interface for human evaluation in FALCON, showing labels of Task I and II. Expanded views provide consistent explanations for each category.

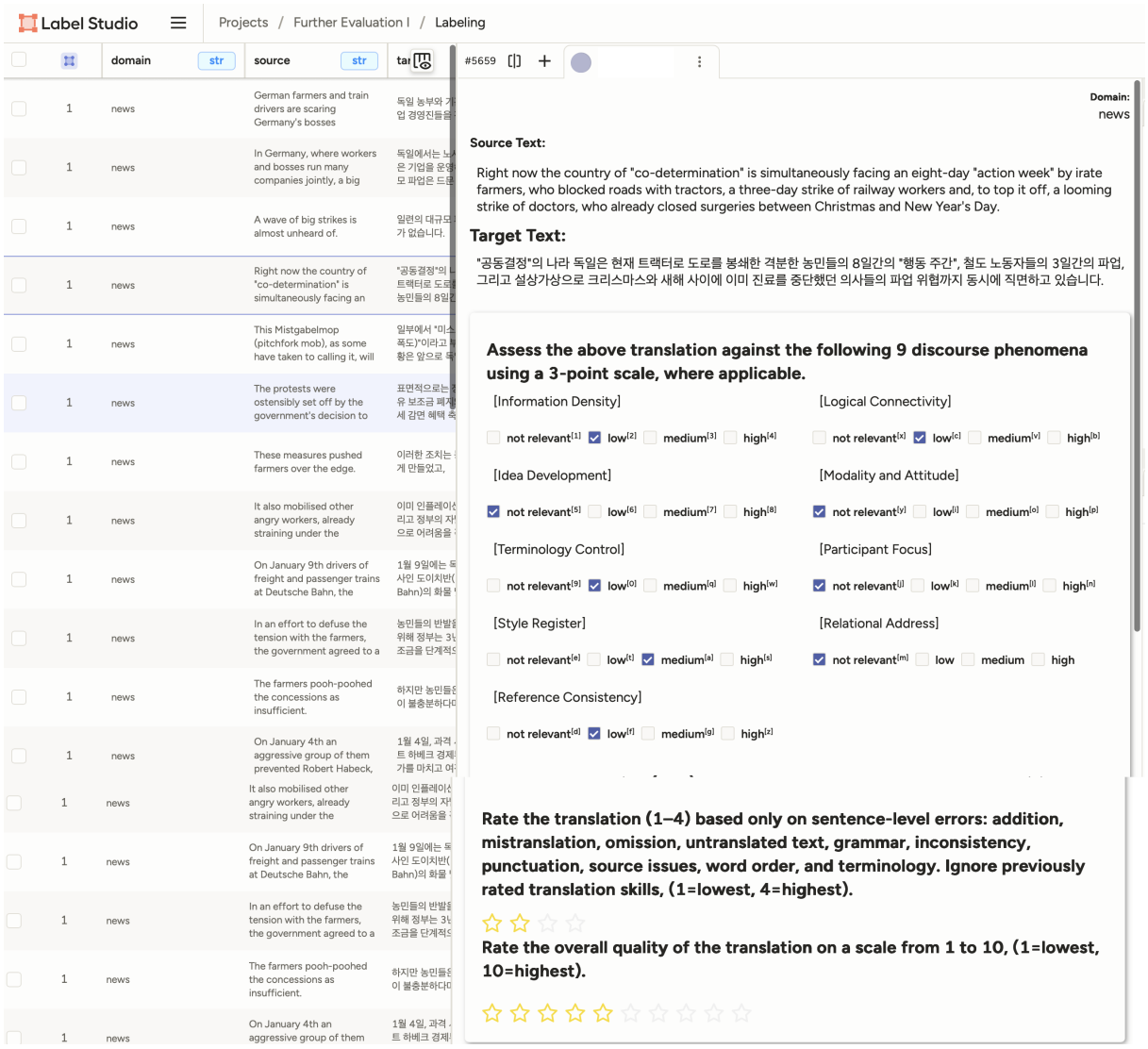



Figure 10: Interface of Label Studio for the human evaluation in H-FALCON. All translation skills are set to “not relevant” by default, and both sentence-level and holistic scores are collected concurrently.