AI translation of literary texts is 'fine', but readers *still* prefer human translations

Anonymous ACL submission

Abstract

AI translation of literary works is increasingly common. While the content may be rendered adequately, we do not know enough about how readers experience it in terms of immersiveness and literary effect–aspects not well captured by automatic machine translation metrics or traditional human evaluation targeting fluency and adequacy. We ask 15 avid readers to compare recently published human translations (HT) to those generated with an agentic large language model (LLM)-based pipeline of 15 recent novels originally in French, Polish, and Japanese, and translated into English; each was evaluated twice. Readers evaluated ∼8K-word excerpts in two conditions: *immersive reading* of the whole excerpt and *close reading* of 772 aligned HT-MT chunks, with two readers per book and in alternating order of presentation. Overall, readers find MT "fine", but prefer HT (19/30 excerpts; 522/772 chunks) for its ease, clarity, and immersive nature. Crucially, readers cannot reliably tell the two apart (only 17/30 guess correctly) and tend to prefer the version they believe to be human. We release LAIT (Literary AI Translation), a reader-centered evaluation dataset with 1K reader comments, 2K judgments and preference ratings, and 7.2K span-level annotations, along with our evaluation protocol and supporting interface.

1 Introduction

Machine-translated[1] literature is reaching real readership. Veen Bosch & Keuning, the largest Dutch trade publisher, announced AI-assisted translation (Creamer, 2024); Amazon introduced Kindle Translate for self-publishing authors (Amazon Staff, 2025); and services such as GlobeScribe market AI translation to smaller publishers and self-published authors (GlobeScribe.AI Ltd). A reader picking up a novel translated in 2026 may be confronted with AI translation even if they are not aware of it.

This raises interesting questions: If AI-translated literature is reaching readers, how does it change their reading experience and perception of the author's work compared to a professional human translation? Readers respond to rhythm, emotions, voice, and deliberate word choices (Brems and Pinto, 2013; Walker, 2021), but this experience cannot be captured by current evaluation setups which often concentrate a small units (Karpinska and Iyyer, 2023; Wu et al., 2025; Zhang et al., 2025b) and break the immersion with segment-level ratings (Kocmi et al., 2025) or error annotations (Freitag et al., 2021). To capture the value of literary translation we need to pose this question directly to the *reader* (Carpuat et al., 2025).

To fill this gap, we recruit 15 avid readers and ask them to read and compare 8k-word long excerpts translated from Japanese, Polish and French into English by professional literary translators (HT)[2] and an agentic LLM-based pipeline (MT), reading each version in full and uninterrupted, as they would a published book (*immersive reading*). Then, we ask them to look closer at both texts in 300-word chunks presented side by side and to highlight phrasing they judge as **good** or **poor** before indicating which version they prefer (*close reading*).

After reading each excerpt, readers are asked to comment on anything that stood out to them positively or negatively; after reading both excerpts they are asked to guess which one is MT and explain their guess in writing; and after each preference choice, they are asked to justify it, which provides insight into readers' perceptions beyond simple ratings (see Figure 1).

Overall, readers show slight but consistent preference for HT, which is clearer under *close reading*

[1] We use *AI translation* and *MT* interchangeably for LLM-based translation; questionnaires used *AI* as it is likely the term readers are more familiar with.

[2] We choose new books translated recently (2025–2026) into English to mitigate the training data contamination issue.

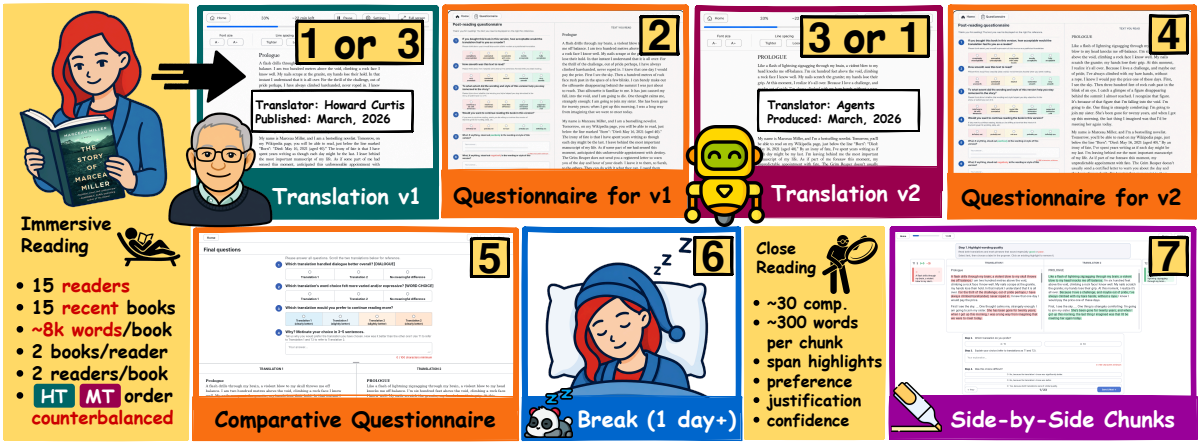


Figure 1: **Evaluation pipeline:** Avid readers of published fiction evaluate two versions of an 8,000-word book excerpt: a **human translation (HT)** and an **AI-generated machine translation (MT)**. Participants **(1)** read the first translation, **(2)** complete a perception questionnaire, **(3)** read the competing translation, **(4)** complete a second questionnaire, **(5)** compare both versions, **(6)** take a one-day break, and **(7)** perform a side-by-side evaluation of 300-word chunks, selecting the preferred version with a justification and confidence score. The presentation order is counterbalanced: each book is evaluated in both HT-first and MT-first conditions, and each reader evaluates two different books, encountering HT first for one and MT first for the other.

(522/772) than under *immersive reading* (19/30). However, readers do not completely reject MT as they often find it readable and sometimes preferred, with MT success varied more by book than the source language. Reader comments suggest that HT is praised more for immersion, naturalness, and literary rendering (e.g., "*having soul*"). MT sometimes win due to readability or local word choices.

Readers also struggle to identify MT reliably. Their accuracy remains close to change even after reading both excerpts (17/30). Their explanations show that they are misled by common AI tells (Russell et al., 2025), such as em-dashes or assumptions about AI not using swearing words. Finally, automatic metrics, including LLM-as-a-judge approaches, fail to recover reader preferences and favor MTs.

To summarize, our contributions are threefold:

1. **DATA:** We introduce LAIT, 15 recent FR/PL/JA→EN novel openings aligned across source, published HT, and MT at the excerpt, chunk, and paragraph levels, with 2K reader ratings and preferences, 1K comments, and 7.2K span annotations. LAIT also includes a 25-book openings devset with 3.4K paragraph-level alignments across 19 MT systems.
2. **ANALYSIS:** We analyze readers' preferences and show that modern literary MT can be readable and hard to detect, but readers still value HT.
3. **METHODOLOGY:** We introduce a reader-centered protocol combining *immersive* and *close* readings, and release our guidelines and annotation software to support reproducibility and future evaluation efforts.

2 Data & Methods

In this section, we first decribe the contruction of LAIT (§2.1). Then, we detail our human evaluation pipeline including *immersive reading* (excerpt-level; 8k words) and *close reading* (chunk-level; 300 words) evaluations (§2.2).

2.1 Constructing LAIT

Book selection. We select recently published, critically acclaimed fiction books originally written in French, Polish, and Japanese, with professional English translations published in 2025–2026, to mitigate data contamination (Jacovi et al., 2023). Overall, we obtain **15 books** for human evaluation,[3] with five books per source language (see Table 4 for the list, and Table 1 for statistics), and reserve *25* books for optimizing the MT pipeline and ablation experiments in § D.3 (see Table 5).[4] See §A for details.

[3]We limit our human evaluation to 15 books due to the prohibitive cost.

[4]All source books were purchased and processed solely for research purposes. To comply with *Fair Dealing* guidelines for research, we have retained no more than 10% of any individual work.

	HT		MT		SRC	
	TOKENS	WORDS	TOKENS	WORDS	TOKENS	WORDS
Books			(*n=15*)			
MEAN	9,680	7,623	9,582	7,676	13,026	4,784
ST. DEV.	535	449	626	529	1,527	3,334
MAX	10,261	7,995	10,500	8,595	15,967	8,244
MIN	8,600	6,747	8,474	6,928	11,155	141
Chunks			(*n=386*)			
MEAN	376	296	372	298	508	186
ST. DEV.	66	49	73	57	122	128
MAX	611	486	644	530	1,139	417
MIN	118	100	121	102	141	1

Table 1: Summary statistics for the evaluation dataset. Words are whitespace-delimited counts. Tokens are computed with `tiktoken` (`o200k_base`).

Source and target languages. We select French, Polish, and Japanese as source languages because they span different language families, cultures, and literary traditions. We choose English as the target language to probe the upper bound of translation capability, since modern language models are exceptionally strong in English (Ahuja et al., 2023). Furthermore, the researchers' familiarity with all the languages facilitated data collection and analysis.

Excerpt extraction. For each book, we select an opening excerpt from the English human translation (HT) of ~8k words ($\mu = 7{,}623$, $SD = 449$) under whitespace tokenization while preserving paragraph boundaries. This length supports *immersive* reading while keeping the task feasible for readers' evaluation. We then manually identify the corresponding source excerpt to obtain source-target mapping with HT.

Preprocessing. Before generating MTs, we manually inspected the excerpts and removed editorial notes and footnotes while preserving the original paragraph structure. Additional normalization, e.g., unifying quotation marks, was applied to MTs and HTs before evaluation to minimize the influence of surface-level features (see § C.4 for detail).

MT pipeline selection. To select a strong MT pipeline for the evaluation, we test five configurations on 17 development books:[5] (P1) full-excerpt translation with post-editing and (P2) chunk-level translation with excerpt-level post-editing, each with GPT-5.4 or Gemini 3.1 Pro; and (P3) an AutoFiction-inspired agentic pipeline (Pham et al., 2026) using Claude Code and Codex.[6] In a blind five-way preference task with six volunteers, most outputs were comparable, except GPT-5.4 (P1), which underperformed. We selected (P3) because it was slightly preferred for more idiomatic vocabulary. See §B for details.

Agentic pipeline. We adapt the AutoFiction long-form fiction generation framework (Pham et al., 2026) to literary translation, deploying two coding agents to produce and validate translations: Claude Code (Claude Opus 4.6) and Codex (GPT-5.4).[7] The pipeline comprises three stages (see Figure 6). In **(1) preprocessing**, Claude Code produces translation guidelines based on the source excerpt, which is then split into 1k-token chunks along paragraph boundaries. At the **(2) chunk-level**, each chunk is translated by Codex, and reviewed for for translation quality (e.g., faithfulness) via LiTransProQA (Zhang et al., 2025b) (Codex) and literary quality (e.g., voice, style) adapted from AutoFiction (Claude Code); an `acceptance gate` either advances the chunk or returns it for revision, capped at three revisions due to the cost and small gains. Finally, at the **(3) excerpt-level**, chunks are merged and reviewed for global literary quality and cross-chunk consistency (Codex), and a `final acceptance gate` triggers Claude Code revision (up to two) if major issues are still present. See §B for details.

Chunking. To prepare the data for the chunk-level *close reading*, we split each HT excerpt into ~300-word chunks ($\mu = 296$, $SD = 49$) preserving the paragraph boundaries. We then align these chunks using Gemini 3.1 Pro to the corresponding chunks from MT and source text. All alignments are verified manually.

2.2 Human Evaluation Protocol

Overview. We evaluate 15 long book excerpts (~8k words) translated by professional human translators (published HT) and by our pipeline (agentic MT). Each excerpt is evaluated by two readers and each reader evaluates two books[8] with a total of

[5] We were unable to process all 25 books as we were restricted by the rate limits of the coding agent-based pipeline.

[6] We chose GPT and Gemini models for their strong WMT25 performance (Kocmi et al., 2025), and coding agents for their capabilities to handle long context and multiple references.

[7] We set reasoning effort to maximum (`max` for Claude Code, `xhigh` for Codex). Both agents were accessed via monthly subscriptions, at a total cost of $400 USD for the experiment.

[8] The evaluations for both books are consecutive, i.e., the reader has to finish one book before proceeding to the next.

Question	Type
***Immersive reading** — isolated-reading*	
Q1: Acceptability as a published translation	5L
Q2: Reading smoothness (fluency)	5L
Q3: Wording/style support for immersion	5L
Q4: Willingness to keep reading this version	5L
Q5: Positive aspects of wording/style	Open
Q6: Negative aspects of wording/style	Open
Q7: Human- vs. machine-translated guess	2C
Q8: Confidence in guess	5L
***Immersive reading** — comparison*	
Q1: Better dialogue handling	3C
Q2: More varied / expressive word choice	3C
Q3: Version to continue reading	4C
Q4: Justification for preferences	Open
Q5: More likely AI-translated	2C
Q6: Confidence in choice	5L
Q7: Justification for the judgment	Open
***Close reading** — per chunk pair*	
Q1: Highlight good / poor phrases	Span
Q2: Preferred translation	2C
Q3: Justify choice	Open
Q4: Difficulty of the choice	3C
***Post-study** — after both assigned books*	
Q1: Prior expectations for AI translation quality	Open
Q2: Surprise at MT-identification accuracy	Open
Q3: Opinion of AI translation after the study	Open
Q4: Other comments	Open

Table 2: Questionnaire items for human evaluation. Scales: 5-point Likert (5L); categorical with 2/3/4 options (2C/3C/4C); good/poor span labels (Span); free text (Open). See §C for the full wording.

60 excerpt-level evaluations (*immersive reading*, 15 books $\times$ 2 readers $\times$ 2 translations). After a one day break, each reader evaluates the same excerpt at the chunk level, with 300-word MT and HT chunks presented side by side for a total of 772 comparisons (*close reading*). See Figure 1 for the pipeline.

Participants. We recruit 15 avid readers through Upwork[9] (reported reading 12–120 books/year). Each reader evaluates two book excerpts and for each excerpt they evaluate both MT and HT (within-subject design). We aim to accommodate their reading preferences by matching each reader with their preferred genre/story type. Participants are paid $110 USD per book with a $5-10 USD bonus upon completion. The task takes approximately 4h resulting in an hourly rate of $27.50 USD. Total cost of the evaluation including bonuses and Upwork fees was **$4k USD**. The protocol was approved by the Institutional Review Board. See §C for details.

Procedure. Prior to the task, each participant provided informed consent, filled in a demographic survey on their AI use and reading habits, and reviewed interface guidelines.[10] They were then provided with the credentials to the interface where they completed an *immersive reading* task followed by a one-day break and a *close reading* task. Both tasks were administered at an interface designed to mimic book readers (See Figure 9 and Figure 10). After finishing both assigned books, participants completed a *post-study* questionnaire.

Immersive reading. Participants first read one translation of entire excerpt and answer eight isolated-reading questions about it, with access only to that version. They then read the second translation and answer the same eight questions, again with access only to that version. Only after completing both readings do they compare the two translations directly and answer seven comparative questions. Reading order is counterbalanced: for each book, one participant reads HT first and the other reads MT first; each participant also reads HT first for one of their two assigned books and MT first for the other. Participants are not told which version is MT.

Close reading. After a one-day break, participants complete a side-by-side close reading task on shorter chunks of about 300 words ($M = 296$, $SD = 49$) with HT and MT aligned. For each chunk, they highlight wording they judge especially good or poor, choose the translation they prefer, explain their choice, and indicate whether this decision was difficult.

Questionnaires. We used four questionnaires. The *immersive reading* task included two: (1) an isolated-reading questionnaire presented immediately after each reading, and (2) a comparative questionnaire presented after both readings. Together, these captured global reading experience. The *close reading* task used (3) a questionnaire for each chunk pair, capturing local preference. Finally, after participants completed both books, we administered (4) a *post-study* questionnaire with four open-ended questions about their thoughts. See Table 2 for questions.

Quality control. To ensure data quality, we require a guideline passphrase to access credentials, screen reading sessions for anomalies like fast scrolling, and flag AI-generated comments using the Pangram AI detector (Emi and Spero, 2024).

[9] upwork.com

[10] The exact wording can be found in §C.

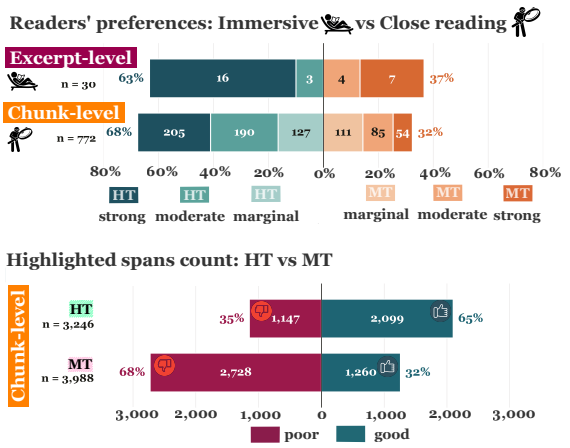


Figure 2: The distribution of readers' preferences between the **human translation** (**HT**) and the **machine translation** (**MT**) after excerpt-level immersive reading and chunk-level close reading. The readers were judging which version they preferred and how strong the preference was. The bottom panel shows the number of **good** and **poor** spans highlighted for each translation type during close reading.

3 Results & Analysis

In this section, we present our evaluation results.

Readers prefer HT, especially in close reading. After reading both translations of an assigned excerpt, participants chose which version they would prefer to continue reading (co-Q3). Across the 30 excerpt-level judgments, 19 favored HT and 11 favored MT, with 16 strongly preferring HT compared to only 7 strongly preferring MT (Figure 2), though this difference was not statistically significant (p=0.219).[11] The preference is clearer under close reading: for the chunk-level preference question (ch-Q1), HT was significantly preferred, with 522 of 772 judgments favoring HT, 205 of them strong preferences, compared to 250 favoring MT, only 54 of them strong (p<.001; Figure 2). Because chunk judgments are nested within readers and books, we also checked the result after aggregating by these clusters: the mean HT-preference rate was 68.4% by book (bootstrap 95% CI: 56.2–79.5) and 68.0% by reader (bootstrap 95% CI: 57.1–78.1). [12]

[11] We note that with more data, the difference between HT and MT might reach significance.

[12] Unless stated otherwise, binary responses are analyzed with logistic models or exact binomial tests; ordinal responses with cumulative-link models. To account for repeated judgments from the same readers and books, we use mixed-effects models with random intercepts for reader and book when estimable; otherwise, we report the corresponding fixed-effects model. We additionally tested for an effect of reading order (HT-first vs. MT-first) and confirmed that it had no significant effect on the results. See §D.3 for details.

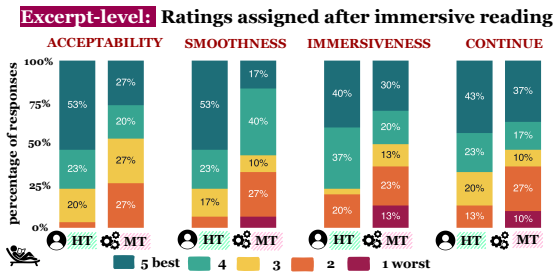


Figure 3: The distribution of readers' ratings after immersive reading of each excerpt **in isolation**. The raters were judging **acceptability** (whether it is acceptable as a published translation), **smoothness**, **immersiveness** (whether it supported immersion), and **continue** (whether they would continue reading). The ratings were provided after each isolated reading.

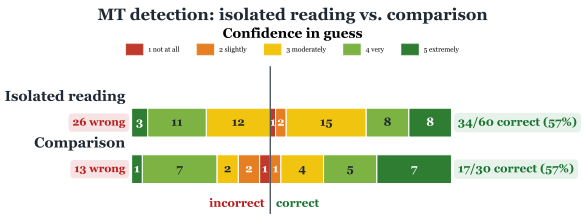


Figure 4: Machine-translation (MT) detection accuracy and confidence in the isolated-reading (n=60) and comparison conditions (n=30). Colors represent confidence levels, from *not at all* to *extremely* confident.

HT is rated higher for quality, especially for smoothness and as a publishable translation. After reading each version, participants rated it on a 5-point scale for acceptability as a published translation, smoothness, support for immersion, and willingness to continue (Figure 3). HT received a large proportion of top ratings (4 or 5) and no rating of 1. The HT ratings were significantly higher for acceptability as a published translation (p=0.0069) and smoothness (p=0.0029), whereas the difference for how well the versions supported immersion and whether readers would continue reading was smaller but still in favor of HT.

HT is preferred on dialogue and word choice. At the excerpt level, readers also compared the two versions on dialogue and word choice (Figure 20). On both attributes HT was preferred more often, though not significantly, than MT (dialogue: 17 vs. 9; word choice: 18 vs. 10), and readers rarely judged the versions equivalent (4 and 2 ties out of 30, respectively).

Top labels	Definitions	Representative participant comment
Top 3 reasons HT was chosen	Excerpt comparison	
1. **28.2%** Smoothness 2. **10.3%** Comprehension 3. **7.7%** Word choice: naturalness	**Smoothness**: The prose flowed without friction, without consideration to understandability or clarity. **Comprehension**: The reader could understand what was happening: local sentence-level understanding and global scene/plot-level comprehension. **Word choice: naturalness**: Natural English words use and phrasings.	**Ex.** *From the start of the story, HT was a lot clearer to me*[Comp.] *and I think a lot of that has to do with punctuation and word choice. MT doesn't necessarily use obscure or uncommon words, but the translation doesn't feel quite right*[WC-natur.] *at times. For example, near the end of the first paragraph, MT states "his legs refuse obedience and his eyes no longer come together in seeing." I understand what this is supposed to mean, but HT's translation (when your legs won't do what they're told to and your eyes can't make out the world) is so much easier and natural to read*[Smooth.]. —Wiesław Myśliwski, *Needle's Eye* (Bill Johnson, Trans.)
Top 3 reasons MT was chosen	Excerpt comparison	
1. **14.5%** Smoothness 2. **14.5%** Engagement 3. **9.1%** Enjoyment	**Smoothness**: *same definition as row 1.* **Engagement**: The reader was pulled into the story: losing reading awareness, attention focused on the scene. **Enjoyment**: General reader satisfaction with clear statement: "I liked it", "pleasant to read", "a pleasure", "I enjoyed both".	**Ex.** *The language in MT is much more natural than HT and is clearer to understand. MT pulls the reader in*[Engag.] *with more vivid descriptions and is generally well-written*[Enjoy.]*. HT was boring to read and I felt my mind wandering*[Engag.]*.* —Haruo, *The Ark* (Jim Rion, Trans.)
Top 3 reasons HT was chosen	Chunk comparison	
1. **15.1%** Smoothness 2. **13.1%** Word choice: clarity 3. **12.1%** Word choice: naturalness	**Smoothness**: *same definition as row 1.* **Word choice: clarity**: Clear descriptions, actions and intents due to words and phrasings choices. The reader is not trying to guess what is happening. **Word choice: naturalness**: *same definition as row 1.*	**Ex.** *Despite one of two better-worded sections in MT, like "but Cosima will have to be patient..." instead of the repetitive "Until Cosima grew up...", HT is clearer*[WC-clar.] *and better worded overall*[WC-natur.]*. The part about Cosima refusing help is clearer in HT*[WC-clar.]*. The part about the cook being impressive in MT isn't clear, as it doesn't explain WHAT about her is impressive. HT makes it clear that it's her physique. The Latin phrase "manu militari" seems out of place in MT. "Roughly shoved by a man in uniform" is clearer. The last sentence is also rendered more clearly and flows more smoothly in HT*[Smooth.]*.* —Marc Levy, *Symphony Of Monsters* (Tina Kover, Trans.)
Top 3 reasons MT was chosen	Chunk comparison	
1. **11.8%** Word choice: naturalness 2. **9.8%** Smoothness 3. **9.8%** Word choice: register	**Word choice: naturalness**: *same definition as row 1.* **Smoothness**: *same definition as row 1.* **Word choice: register**: Word choices match the context in terms of formality, period, dialect/local speech, genre, speaker, situation, or emotional context.	**Ex.** *MT sounds slightly more natural*[WC-natur.]*, but it also has phrases that stuck out to me as strange. Both have formatting issues. In MT, I would not have known that first line was a quote if it weren't next to HT. The address on the stationery in MT flows much better*[Smooth.] *in all caps than in HT. I did not know what "a flowery balcony" meant in HT. The final sentence in both is a little long and hard to follow, but MT translates the word as "violin" which makes more sense in context than HT's "fiddle"*[WC-regis.]*.* —Mikołaj Łoziński, *My Name Is Stramer* (Antonia Lloyd-Jones, Trans.)

Table 3: Top labels behind reader preferences for HT or MT in excerpt- and chunk-level comparisons. Percentages are computed separately within each system and task from positive labels assigned to the chosen translation. Participant comments are shown in full with translation labels normalized to HT/MT, and colored highlights correspond to the code chips in the top-label list.

HT preference does *not* imply MT failure. The overall preference for HT does not mean that readers rejected MT across the board. Of the 11 excerpt-level judgments that favored MT, 7 expressed a clear preference (Figure 2). In the isolated-reading ratings, 54% of MT responses indicated moderate-to-strong willingness to continue reading (a rating of 4 or 5), compared with 66% for HT (Figure 3). And in the close-reading evaluation, roughly one-third (32%) of chunk-level choices favored MT.

MT preference varies by book but not by language. The one-third MT share is stable across target languages (French 32%, Japanese 34%, Polish 31%; Figure 15b) but varies widely across books, from 4% to 88% (Figure 15a). Preference is therefore book-specific rather than language-specific. The clearest case is *Hooked: A Novel of Obsession*, preferred in MT 88% of the time and consistently by both readers, suggesting that sometimes HT may fall short rather than MT having issues with one specific language.

Coding comments. We collected 180 comments during *immersive reading* and 772 during *close reading*.[13] Following an inductive approach (Thomas, 2006), we iterated over the comments to develop two coding schemas: one for translation quality (29 categories, Table 42) and one for why a translation was judged AI-generated (10 categories, Table 43).[14] All *immersive reading* comments were manually coded by two researchers, who resolved disagreements. For *close reading*, we validated GPT-5.4 coding against a manually coded 10% sample before coding the rest automatically.[15] See §D.2 for details.

Both HT and MT are praised for smoothness, but HT receives broader literary praise. In the isolated-reading comments, readers praised both versions for basic readability (Table 3, Table 48a). Smoothness was the most frequent positive dimension for both, but it was mentioned more often for HT (19/30 readings, 63%) than MT (11/30, 37%). Comprehension showed a smaller gap (HT 12/30, 40%; MT 9/30, 30%), suggesting that MT often

[13] For *immersive reading*, 120 comments concern the good and bad qualities of each translation, 30 explain a preference choice, and 30 explain why a translation was selected as MT. For *close reading*, all comments explain readers' preferences.

[14] The quality schema groups its 29 categories into four families: (A) language-level, (B) narrative-level, (C) reader experience, and (D) meta-translation.

[15] As comments can receive multiple labels, we measure agreement as average Jaccard overlap between human- and model-annotated sets, scoring 0.727 (high overlap).

reaches a readable quality.[16] HT, however, was praised more for qualities tied to sustained literary experience, including engagement or immersion (12/30 vs. 5/30) and vividness of descriptions (9/30 vs. 4/30), while dialogue naturalness was mentioned equally often (7/30 each). HT was described as "fully immersive," as reading like an "originally-written-in-English book," and even as a text that "has soul." MT praises, by contrast, more often emphasized clear readability: readers described it as "smooth," "clean," "tightly constructed," or easy to follow. For three MT translations readers found nothing positive to mention about the translation, whereas HT always received some positive comments.

Negative comments highlight different issues in HT and MT. Complaints about HT were diffuse and centered on local word-choice: naturalness, idiomaticity, and blandness (each 5/30), followed by grammar, register, smoothness, and AI tells (each 4/30) (Table 48b). Complaints about MT were more directed at reading friction: sentence structure (15/30), smoothness (11/30), and comprehensibility (11/30). Readers described MT as "abrupt and choppy," with "longwinded run-on", or requiring them to "stop and reread parts," and often flagged grammar (8/30) and word naturalness (8/30). In six cases readers found nothing negative about the HT, compared to only two for MT.

Why do readers choose HT? Readers prefer HT at the excerpt level for the global reading experience praising it for smoothness (15/19), comprehensibility (8/19), natural wording and sentence structure (6/19 each) (Table 48c). HT is described as a "seamless reading experience," or as reading more like a story originally written in English. At the chunk level (Table 48d), readers mention smoothness (92/522), but also focus more on word choices such as word clarity (87/522) and naturalness (80/522). They value HT for comprehensibility (64/522), sentence structure (39/522), and vividness of the description (39/522).

Why do readers choose MT? When MT is preferred at the excerpt level, the readers mention enjoyment, smoothness, and immersion (each 5/11) as well as comprehension and word richness (each 4/11). These comments often frame MT as more expressive, for instance, one reader wrote that MT

[16] One of the readers commented on this calling MT *"almost like a translation for dummies."*

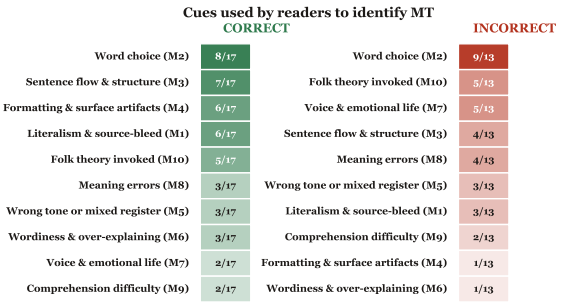


Figure 5: The cues readers mentioned when explaining which translation they believed was machine-translated (MT). Correct judgments identify the actual MT version; incorrect judgments identify the human translation (HT) as MT. Comments could receive multiple cue labels.

had "more intricate and varied" wording and gave a "clearer picture of personalities and relationships." At chunk level, the readers pay more attention to the words being used, with the most common reasons being the natural word choice (35/250) and contextual fit (29/250) followed by smoothness (26/250), rich word choice (23/250), and vivid descriptions (22/250). Typical comments included praises for specific word choices such as "bento containers" over "plastic containers."

Readers can't reliably detect MT. In isolated reading (judging one text), readers identified MT in only 16 and HT in 18 cases, failing to reliably distinguish the two (34/60; p=0.366). This held in the comparison setting, where they correctly identified MT in 17 of 30 cases (p=0.585). Although accuracy was similar across both settings, correct readers were more confident after seeing both texts, suggesting MT was not obvious alone but distinguishable from HT for this group (Figure 4). Regardless of accuracy, most readers preferred the text they though is HT.

What cues do readers use to identify MT? The most common cue for both correct (8/17) and incorrect (9/13) guesses was word choice (Figure 5). For correct choices, the readers were mentioning things like "awkward" word choice, while for incorrect readers the words being "generic" or weird ("people don't speak like that"). The second most common clue that helped the readers was sentence structure (7/17; e.g., MT had run-on sentences), while misleading clue were different types of AI tell (5/13; folk theory) such as em-dashes.

Readers were surprised by the MT quality. In post-study interview most readers said they expected the MT to be worse or easier to detect (12/15) and some stated the MT was surprisingly strong (8/15). The readers also identified the remaining limitations of MT in literary and contextual qualities (8/15; "missing emotional import") (§D.4).

What about translation into languages other than English? Since MT into English exceeded readers' expectations, we ran a case study on two books translated into Spanish, French, Polish, and Japanese using the same setup.[17][18] The contrast between MT and HT was much clearer: only the French reader misidentified MT as HT and initially preferred it, yet at the chunk level *all* readers preferred HT (92.8%), suggesting that even the French MT's flaws surfaced on closer reading. This indicates that MT into languages other than English may not yet have reached a readable quality.

Can the difference between MT and HT be captured with automatic metrics? We compare our chunk-level human judgments with automatic metrics. Since some metrics cannot handle long texts, we use the `par3` aligner (Thai et al., 2022) to align source paragraphs to HT and MT, then map them to chunks, averaging paragraph scores within each chunk.[19] We run METRICX-QE (Juraska et al., 2024), COMETKIWI (Rei et al., 2022), and LITRANSPROQA (Zhang et al., 2025b) with Gemini 3.1 Pro (Google DeepMind, 2026). All metrics prefer MT over HT at the system level (LITRANSPROQA↑ MT: 0.996, HT: 0.920; COMETKIWI↑ MT: 0.760, HT: 0.722; METRICX-QE mean↓ MT: 5.184, HT: 5.671). Kendall τ (Kendall, 1938) correlation with human judgments at chunk-level is weak to negative for all metrics (LITRANSPROQA $\tau = -0.124$, COMETKIWI $\tau = -0.318$, METRICX-QE $\tau = -0.216$), and is identical whether computed over all judgments or only those where annotators agree, since clear human agreement favors HT while metrics favor MT. Thus, automatic metrics fail to recover reader preferences, consistent with prior work (Thai et al., 2022; Zhang et al., 2025b,a).

[17]We choose *The Witcher: Crossroads of Ravens* (PL) for translation into Japanese and *My Grandfather, the Master Detective* (JA) for translation into Spanish, French, and Polish, as both have been recently human-translated into these languages.

[18]For Spanish, French, and Polish the readers were researchers and volunteers who are native speakers; we used two readers for the Polish book. For Japanese we followed the English procedure and hired one reader via Upwork, paid $110 USD.

[19]As one chunk consists of multiple paragraphs, we compute the mean score over all paragraphs in a chunk.

4 Related Work

Automatic evaluation. Automatic evaluation is practical but surface-level (Papineni et al., 2002; Popović, 2015) and trained neural metrics (Rei et al., 2022; Guerreiro et al., 2024; Juraska et al., 2023, 2024) were shown to work poorly on literary machine translation (Thai et al., 2022; Zhang et al., 2025b) and offer weak signal for what humans perceive as literary quality (Gerrits et al., 2026). LLM-as-a-judge approaches based on question answering protocols were developed but tested only on short passages (Zhang et al., 2025b; Shafayat et al., 2024) and its unclear how they handle longer dependencies.

Human evaluation. Human evaluation often relies on error annotations (Lommel et al., 2013; Freitag et al., 2021; Kocmi et al., 2024) or crowd-sourced ratings (Graham et al., 2013; Freitag et al., 2021), which do not capture reading experience (Zhang et al., 2025a). Most evaluation of literary MT uses short passages or local judgments (Wu et al., 2025; Karpinska and Iyyer, 2023; Wang et al., 2023; Kocmi et al., 2025), missing the purpose of literary translation as sustained reading. Other work targets translator-centered post-editing (Castaldo et al., 2025), and few studies scale to longer units but deploy error annotation (Fonteyne et al., 2020) or use old MT systems (Toral et al., 2018; Fonteyne et al., 2020). In contrast, we target readers' experience of long literary MT versus HT in *immersive reading*, and what issues they notice in *close reading*.

5 Discussion & Conclusion

We evaluated literary MT in *immersive* and *close* reading by comparing professional human translations (HT) with a strong agentic LLM pipeline (MT) across 15 recent French, Polish, and Japanese novels translated into English. Our analysis reveals that readers found MT reasonable and were not able to detect it reliably but they consistently preferred HT, especially in the close reading. Preferences varied mostly by book not source language, and readers often preferred the version they believed to be human, while automatic metrics favored MT. We release LAIT and our evaluation pipeline to support future research on literary MT.

Limitations

Excerpt scope. Due to financial limitations we translate and evaluate translations of novel openings, not full books. Literary quality, pacing, and narrative voice may evolve over longer reading, and book-spanning context and background may be lost by LLM translations of full books. These factors may impact judgments over longer translations however we note that 8K words is already at the length of short stories.

Corpus selection. We selected recently published (2025–2026), critically acclaimed fiction with professional English translations available to mitigate training-data contamination. However, this limits the diversity of our books and may not represent the whole set of literature out there.

Languages and directions. The main evaluation is French, Polish, and Japanese into English only (n=5 books per source language). Our multilingual case study covers two additional books and four target languages with a much smaller reader sample and mixed recruitment (Upwork vs. researchers/volunteers); those results are exploratory and should not be generalized. A much larger-scale study is required to determine whether the effects observed in the current work apply to other source-target language pairs.

Sample size and agreement. Each book is read by only two readers recruited via Upwork. This scale was necessary for immersive, long-form evaluation to be financially feasible, but limits statistical power and leaves chunk-level inter-reader agreement modest in places (see §D). We employ a within-subject design, where each reader judges both MT and HT for the same excerpts, reducing individual variability and requiring fewer annotators than a between-subjects design (Allen, 2017).

Ethics Statement

Human participants. This study was approved by our Institutional Review Board. Participants gave informed consent before the evaluation (§C.1), could withdraw with prorated compensation, and were paid fairly for the amount and difficulty of work ($110 USD per book, ∼$27.50/hour for ∼4 hours of work), plus completion bonuses. Participant identities are anonymized by participant codes. Two early participants were excluded (but paid fully for their work) and replaced because their open-ended responses were insufficient for qualitative analysis (§C).

Copyright and data sharing. Source books were purchased for research use. Following fair-dealing / fair-use principles for scholarly research, we retain no more than 10% of any individual work. The dataset will be released **for research purposes only**.

Corpus and societal impact. Our corpus over-represents internationally published, commercially successful fiction. We study AI-assisted literary translation to inform evaluation practice and publishing transparency: **not to advocate replacing human translators or deceiving readers**; in fact, we demonstrate that readers prefer books that they know are human-authored and translated. As LLM-translated fiction enters trade and self-publishing channels, we hope that better reader-centered evaluation can support fair labor practices and help readers make informed choices.

Risks of literary machine translation. Literary MT can look acceptable while still changing the reading experience. As our reader comments show, problems with phrasing, context, and voice can affect the reader's impression of a book and its author, especially when AI-translated fiction reaches readers with little human review. Like other LLM output, MT may also carry biased or harmful wording into the translation.

References

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

Mike Allen. 2017. *The SAGE encyclopedia of communication research methods*. SAGE Publications, Inc, 2455 Teller Road, Thousand Oaks California 91320.

Amazon Staff. 2025. Amazon introduces Kindle Translate, an AI-powered translation service for authors to reach global readers. https://www.aboutamazon.com/news/books-and-authors/amazon-kindle-translate-books-authors. Accessed: 2026-05-22.

Elke Brems and Sara Ramos Pinto. 2013. *Reception and Translation*, chapter Handbook for Translation Studies (vol. IV). John Benjamins.

Marine Carpuat, Omri Asscher, Kalika Bali, Luisa Bentivogli, Frédéric Blain, Lynne Bowker, Monojit Choudhury, Hal Daumé III, Kevin Duh, Ge Gao, Alvin Grissom II, Marzena Karpinska, Elaine C. Khoong, William D. Lewis, André F. T. Martins, Mary Nurminen, Douglas W. Oard, Maja Popovic, Michel Simard, and François Yvon. 2025. An interdisciplinary approach to human-centered machine translation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22859–22879, Suzhou, China. Association for Computational Linguistics.

Antonio Castaldo, Sheila Castilho, Joss Moorkens, and Johanna Monti. 2025. Extending CREAMT: Leveraging large language models for literary translation post-editing. In *Proceedings of Machine Translation Summit XX: Volume 1*, pages 506–515, Geneva, Switzerland. European Association for Machine Translation.

Rune H. B. Christensen. 2023. *ordinal—Regression Models for Ordinal Data*. R package version 2023.12-4.1.

Ella Creamer. 2024. Dutch publisher to use AI to translate limited number of books into English. *The Guardian*.

Bradley Emi and Max Spero. 2024. Technical report on the pangram ai-generated text classifier. *Preprint*, arXiv:2402.14873.

Margot Fonteyne, Arda Tezcan, and Lieve Macken. 2020. Literary machine translation under the magnifying glass: Assessing the quality of an NMT-translated detective novel on document level. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3790–3798, Marseille, France. European Language Resources Association.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Kyo Gerrits, Rik van Noord, and Ana Guerberof Arenas. 2026. Creativity bias: How machine evaluation struggles with creativity in literary translations. *arXiv preprint arXiv:2605.13596*.

GlobeScribe.AI Ltd. GlobeScribe.AI. `https://globescribe.ai/`. Accessed: 2026-05-22.

Google DeepMind. 2026. Gemini 3.1 Pro Model Card. `https://deepmind.google/models/model-cards/gemini-3-1-pro/`. Accessed 2026-05-26.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

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.

Kilem Li Gwet. 2021. *Handbook of inter-rater reliability*. Advanced Analytics.

Alon Jacovi, Avi Caciularu, Omer Goldman, and Yoav Goldberg. 2023. Stop uploading test data in plain text: Practical strategies for mitigating data contamination by evaluation benchmarks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5075–5084, Singapore. Association for Computational Linguistics.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore. Association for Computational Linguistics.

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

M. G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1–2):81–93.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Christof Monz, Kenton Murray, and 10 others. 2025. Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*, pages 355–413, Suzhou, China. Association for Computational Linguistics.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024. Error span annotation: A balanced approach for human evaluation of machine translation. In *Proceedings of the Ninth Conference on Machine*

Translation, pages 1440–1453, Miami, Florida, USA. Association for Computational Linguistics.

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

Arle Richard 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*, London, UK. Aslib.

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

Chau Minh Pham, Yapei Chang, and Mohit Iyyer. 2026. Autofiction pipeline. Research pipeline for long-form AI novel generation.

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.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Jenna Russell, Marzena Karpinska, and Mohit Iyyer. 2025. People who frequently use ChatGPT for writing tasks are accurate and robust detectors of AI-generated text. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5342–5373, Vienna, Austria. Association for Computational Linguistics.

Sheikh Shafayat, Dongkeun Yoon, Woori Jang, Jiwoo Choi, Alice Oh, and Seohyon Jung. 2024. A 2-step framework for automated literary translation evaluation: Its promises and pitfalls. *arXiv preprint arXiv:2412.01340*.

Katherine Thai, Marzena Karpinska, Kalpesh Krishna, Bill Ray, Moira Inghilleri, John Wieting, and Mohit Iyyer. 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 9882–9902, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

David R. Thomas. 2006. A general inductive approach for analyzing qualitative evaluation data. *American Journal of Evaluation*, 27(2):237–246.

Antonio Toral, Martijn Wieling, Sheila Castilho, Joss Moorkens, and Andy Way. 2018. Project PiPeNovel: Pilot on post-editing novels. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, page 385, Alicante, Spain.

Callum Walker. 2021. *An Eye-Tracking Study of Equivalent Effect in Translation: The Reader Experience of Literary Style*. Springer International Publishing.

Longyue Wang, Zhaopeng Tu, Yan Gu, Siyou Liu, Dian Yu, Qingsong Ma, Chenyang Lyu, Liting Zhou, Chao-Hong Liu, Yufeng Ma, Weiyu Chen, Yvette Graham, Bonnie Webber, Philipp Koehn, Andy Way, Yulin Yuan, and Shuming Shi. 2023. Findings of the WMT 2023 shared task on discourse-level literary translation: A fresh orb in the cosmos of LLMs. In *Proceedings of the Eighth Conference on Machine Translation*, pages 55–67, Singapore. Association for Computational Linguistics.

Minghao Wu, Jiahao Xu, Yulin Yuan, Gholamreza Haffari, Longyue Wan, Weihua Luo, and Kaifu Zhang. 2025. (perhaps) beyond human translation: Harnessing multi-agent collaboration for translating ultra-long literary texts. *Transactions of the Association for Computational Linguistics*, 13:901–922.

Ran Zhang, Wei Zhao, and Steffen Eger. 2025a. How good are LLMs for literary translation, really? literary translation evaluation with humans and LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10961–10988, Albuquerque, New Mexico. Association for Computational Linguistics.

Ran Zhang, Wei Zhao, Lieve Macken, and Steffen Eger. 2025b. LiTransProQA: An LLM-based literary translation evaluation metric with professional question answering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29099–29121, Suzhou, China. Association for Computational Linguistics.

A Dataset

Table 4 lists the evaluation books used in the English human evaluation and the multilingual target-language case study. Table 5 lists the development books used to optimize the MT pipeline and ablation setup. Both tables report the source language or language pair, original and translated titles, authors and translators, and publication dates.

Pair	Original Title	Author	Translation Title	Translator	Original Date	Translation Date
Main human evaluation: source to English						
FR→EN	*Le Roman de Marceau Miller*	Marceau Miller	*The Story of Marceau Miller*	Howard Curtis	Jan. 2025	Mar. 2026
FR→EN	*La Symphonie des monstres*	Marc Levy	*Symphony of Monsters*	Tina Kover	Oct. 2023	Jan. 2026
FR→EN	*Veiller sur elle*	Jean-Baptiste Andrea	*Watching Over Her*	Frank Wynne	Aug. 2023	Aug. 2025
FR→EN	*Les Yeux de Mona*	Thomas Schlesser	*Mona's Eyes*	Hildegarde Serle	Jan. 2024	Sept. 2025
FR→EN	*Célèbre*	Maud Ventura	*Make Me Famous*	Gretchen Schmid	Aug. 2024	May 2025
JA→EN	*Hakobune*	Haruo Yuki	*The Ark*	Jim Rion	Sept. 2022	Feb. 2026
JA→EN	*Meitantei no Mama de Ite*	Masateru Konishi	*My Grandfather, the Master Detective*	Louise Heal Kawai	Jan. 2023	Mar. 2026
JA→EN	*Kokun, Vol. 1: Nishi kara Kita Shojo*	Nahoko Uehashi	*Kokun: The Girl from the West*	Cathy Hirano	Mar. 2022	Feb. 2026
JA→EN	*Nairu Pachi no Joshikai*	Asako Yuzuki	*Hooked: A Novel of Obsession*	Polly Barton	Mar. 2015	Mar. 2026
JA→EN	*Kiiroi Ie*	Mieko Kawakami	*Sisters in Yellow*	Laurel Taylor and Hitomi Yoshio	Feb. 2023	Mar. 2026
PL→EN	*Rozdroże kruków*	Andrzej Sapkowski	*The Witcher: Crossroads of Ravens*	David French	Oct. 2024	Sept. 2025
PL→EN	*Stacja*	Jakub Szamałek	*Inner Space*	Kasia Beresford	May 2023	July 2025
PL→EN	*Stramer*	Mikołaj Łoziński	*My Name is Stramer*	Antonia Lloyd-Jones	Oct. 2019	Aug. 2025
PL→EN	*Ucho Igielne*	Wiesław Myśliwski	*Needle's Eye*	Bill Johnston	Oct. 2018	Oct. 2025
PL→EN	*Empuzjon. Horror przyrodoleczniczy*	Olga Tokarczuk	*The Empusium: A Health Resort Horror Story*	Antonia Lloyd-Jones	June 2022	Sept. 2024
Multilingual target-language case study						
JA→FR	*Meitantei no Mama de Ite*	Masateru Konishi	*Les Histoires de Kaede*	Mathilde Tamae-Bouhon	Jan. 2023	May 2024
JA→PL	*Meitantei no Mama de Ite*	Masateru Konishi	*Mój dziadek – genialny detektyw*	Dariusz Latoś	Jan. 2023	Sept. 2025
JA→ES	*Meitantei no Mama de Ite*	Masateru Konishi	*La pequeña habitación de los misterios*	Juan Francisco González Sánchez	Jan. 2023	Sept. 2025
PL→JA	*Rozdroże kruków*	Andrzej Sapkowski	*Witcher: Karasu no Jujiro*	Yasuko Kawano and Aya Sugiura	Oct. 2024	Sept. 2025

Table 4: Evaluation books used in the human study and multilingual target-language case study. The table lists the source-target language pair, original title and author, translated publication title and translator, and publication dates for the original and translated editions.

B Machine Translation

Figure 6 shows the agentic pipeline used to generate the machine-translated excerpts evaluated in this study.

- P1 Full-excerpt translation with post-editing by (a) GPT-5.4 or (b) Gemini 3.1 Pro.
- P2 Chunk-level translation with excerpt-level post-editing by (a) GPT-5.4 or (b) Gemini 3.1 Pro.
- P3 An agentic pipeline with Claude Code (Opus 4.6) and Codex (GPT-5.4), inspired by AutoFiction (Pham et al., 2026), an agent-based novel-generation pipeline.

We select GPT-5.4 and Gemini 3.1 Pro because these model families were strong performers in the WMT 2025 General Translation Shared Task (Kocmi et al., 2025). We further explore coding agents because they are optimized for long-context tasks and for handling multiple reference documents. We translate 8k-word excerpts from 17 of 25 development books under all five setups.[20] In a blind 5-way preference task with six volunteers, most outputs were comparable, except GPT-5.4 (P1) which noticeably underperformed. We adopted the agentic pipeline (P3), which was preferred by a small margin owing to its more idiomatic vocabulary. See below for the prompts and pipeline details used for this process.

The complete prompt suite and pipeline diagnostics are provided in §B and §B.

Pipeline prompts. The following tables reproduce the prompt suite used for alignment, translation, review, and revision stages in the MT pipeline.

[20] Because agents make many model calls, running them via API is prohibitively expensive; instead we used subscription services, which caps our translation count.

Lang.	Original Title	Author	Translation Title	Translator	Original Date	Translation Date
FR	*Les guerres précieuses*	Perrine Tripier	*Our Precious Wars*	Alison Anderson	Jan. 2023	Nov. 2025
FR	*Naufrage*	Vincent Delecroix	*Small Boat*	Helen Stevenson	Aug. 2023	Mar. 2025
FR	*Déserter*	Mathias Énard	*The Deserters*	Charlotte Mandell	Aug. 2023	May 2025
FR	*Ilaria ou la conquête de la désobéissance*	Gabriella Zalapi	*Ilaria, or The Conquest of Disobedience*	Adriana Hunter	Aug. 2024	Nov. 2025
FR	*Les villages de Dieu*	Emmelie Prophète	*Cécé*	Aidan Rooney	Nov. 2020	Sept. 2025
FR	*Le vieil incendie*	Elisa Shua Dusapin	*The Old Fire*	Aneesa Abbas Higgins	Aug. 2023	Jan. 2026
FR	*Le Fruit le plus rare*	Gaëlle Bélem	*The Rarest Fruit*	Hildegarde Serle	Aug. 2023	June 2025
FR	*Le rêve du jaguar*	Miguel Bonnefoy	*The Dream of the Jaguar*	Ruth Diver	Aug. 2024	Dec. 2025
JA	*Hennae*	Uketsu Uketsu	*Strange Pictures*	Jim Rion	Oct. 2022	Jan. 2025
JA	*Jakuson Hitori*	Jose Ando	*Jackson Alone*	Kalau Almony	Noc. 2022	Jan. 2026
JA	*Machi to Sono Futashika na Kabe*	Haruki Murakami	*The City and Its Uncertain Walls*	Philip Gabriel	Apr. 2023	Nov. 2024
JA	*Baba Yaga no Yoru*	Akira Otani	*The Night of Baba Yaga*	Sam Bett	Oct. 2020	July 2024
JA	*Kai ni Tsuzuku Bashonite*	Mai Ishizawa	*The Place of Shells*	Polly Barton	July 2021	Mar. 2025
JA	*Tokyo-to Dojo-to*	Rie Qudan	*Tokyo Sympathy Tower*	Jesse Kirkwood	Jan. 2024	Aug. 2025
JA	*Ōkina tori ni sarawarenai yō*	Hiromi Kawakami	*Under the Eye of the Big Bird*	Asa Yoneda	Apr. 2016	Sept. 2024
JA	*Shometsu Sekai*	Sayaka Murata	*Vanishing World*	Ginny Tapley Takemori	Jan. 2018	Apr. 2025
JA	*Kyūkanbi no kanojotachi*	Emi Yagi	*When the Museum Is Closed*	Yuki Tejima	Mar. 2023	Jan. 2026
PL	*Powrót z gwiazd*	Stanisław Lem	*Return from the Stars*	Barbara Marszal	Jan. 1961	June 1989
PL	*Robot*	Adam Wiśniewski-Snerg	*Robot*	Tomasz Mirkowicz	Jan. 1973	May 2022
PL	*Księgi Jakubowe*	Olga Tokarczuk	*The Books of Jacob*	Jennifer Croft	Oct. 2014	Nov. 2021
PL	*Lód*	Jacek Dukaj	*Ice*	Ursula Phillips	Jan. 2007	Nov. 2025
PL	*Harda*	Elżbieta Cherezińska	*The Widow Queen*	Maya Zakrzewska-Pim	June 2016	Apr. 2021
PL	*Mgła*	Kaja Malanowska	*Fog*	Bill Johnston	Jan. 2015	Mar. 2021
PL	*Łakome*	Małgorzata Lebda	*Voracious*	Antonia Lloyd-Jones	Aug. 2023	Jan. 2025
PL	*Poufne*	Mikołaj Grynberg	*Confidential*	Sean Gasper Bye	June 2020	Jan. 2025

Table 5: Development books reserved for optimizing the machine-translation (MT) pipeline and ablation experiments. The table lists the source language, original title and author, english translated publication title and translator, and publication dates for the original and english translated editions.

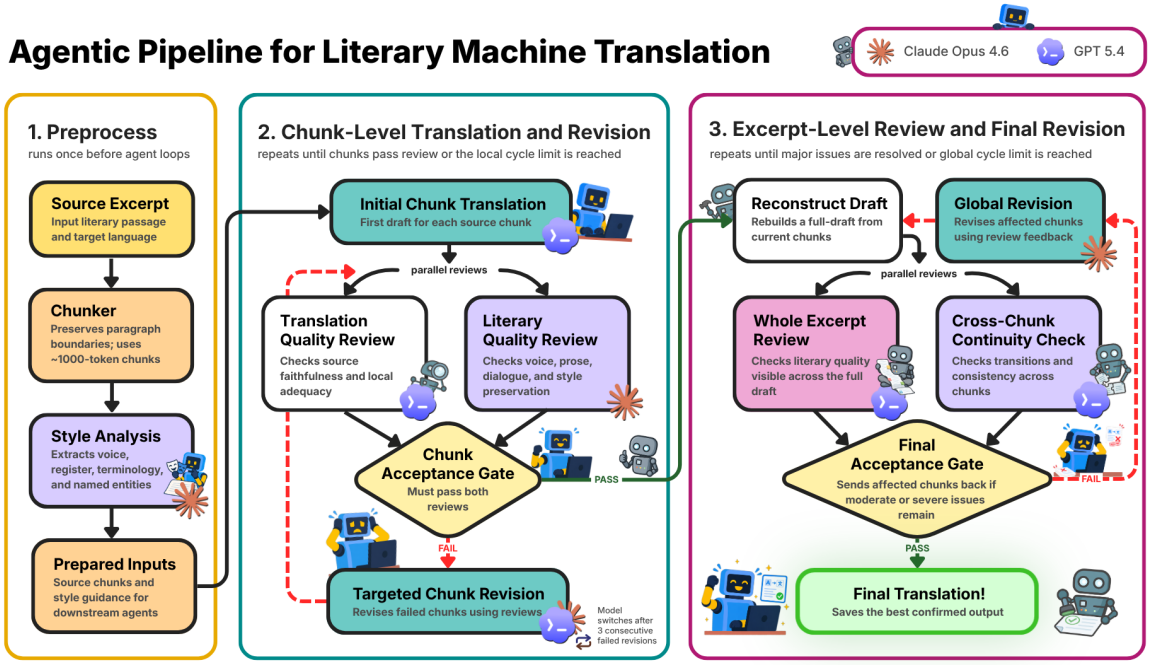


Figure 6: Agentic literary MT pipeline used in this study. Source excerpts are chunked and paired with style guidance, then translated and revised through chunk-level and full-draft review loops.

Chunk Alignment (API) Prompt

```
You are aligning two English translations of the same novel, paragraph by paragraph.

You will receive:
```

- a CANDIDATE CHUNK from one translation, made up of one or more paragraphs separated by blank lines;
- a larger TARGET WINDOW from a second translation.

Your task: pick the contiguous run of whole paragraphs from the TARGET WINDOW whose combined content corresponds — semantically, scene-for-scene — to the CANDIDATE CHUNK.

—CANDIDATE CHUNK—
{source_chunk}
—END CANDIDATE CHUNK—

—TARGET WINDOW—
{target_window}
—END TARGET WINDOW—

Rules:

- The match MUST be a single contiguous span of paragraphs from the TARGET WINDOW. Do not skip paragraphs in the middle.
- Paragraph boundaries are blank lines. Begin at the start of a paragraph and end at the end of a paragraph; do not break mid-paragraph.
- The two translations may differ in paragraph count, phrasing, or minor content — translators sometimes add, omit, or merge sentences. Choose the smallest contiguous run of target paragraphs that covers the same narrative content as the candidate chunk.
- Use the FIRST plausible match in the window — do not jump ahead to a later, similar passage.
- Copy the matched text VERBATIM from the TARGET WINDOW: same characters, same whitespace, same blank lines. Do not paraphrase or add commentary.
- There should always be a match. If you are certain that there is no match, please return "no match".

Return ONLY the matched text itself, copied verbatim from the TARGET WINDOW.
Do not add any explanations or change the text in any way.

Table 6: API prompt used to align whole-paragraph chunks between two English translations.

Chunk Alignment (API) Prompt - Internal Evaluation

You are aligning two versions of the same novel, paragraph by paragraph.

You will receive:
- a CANDIDATE CHUNK in **{source_language}**, made up of one or more paragraphs separated by blank lines;
- a larger TARGET WINDOW in **{target_language}**.

Your task: pick the contiguous run of whole paragraphs from the TARGET WINDOW whose combined content corresponds — semantically, scene-for-scene — to the CANDIDATE CHUNK.

—CANDIDATE CHUNK—
{source_chunk}
—END CANDIDATE CHUNK—

—TARGET WINDOW—
{target_window}
—END TARGET WINDOW—

Rules:

- The match MUST be a single contiguous span of paragraphs from the TARGET WINDOW. Do not skip paragraphs in the middle.
- Paragraph boundaries are blank lines. Begin at the start of a paragraph and end at the end of a paragraph; do not break mid-paragraph.
- The two versions may be in different languages and may differ in paragraph count, phrasing, or minor content — translators sometimes add, omit, or merge sentences. Choose the smallest contiguous run of target paragraphs that covers the same narrative content as the candidate chunk.
- Use the FIRST plausible match in the window — do not jump ahead to a later, similar passage.
- Copy the matched text VERBATIM from the TARGET WINDOW: same characters, same whitespace, same blank lines. Do not paraphrase or add commentary.
- There should always be a match. If you are certain that there is no match, please return "no match".

Return ONLY the matched text itself, copied verbatim from the TARGET WINDOW.
Do not add any explanations or change the text in any way.

Table 7: API prompt used to align whole-paragraph chunks across source and target languages.

Translation Prompt

You are an expert literary translator.

Your task is to translate an excerpt from a novel written in **{source_lang}** into English. Produce a fluent, idiomatic literary translation intended for an English-language readership.

Be faithful to the original meaning, tone, voice, register, rhythm, and stylistic effect, while producing natural, fluent, idiomatic prose in English.

Adapt wording and syntax where needed for naturalness and readability, while preserving distinctive stylistic features wherever possible.

Maintain consistency of narrative perspective, register, and style throughout the text.

Preserve ambiguity, characterization, and differences in voice between narrator and characters wherever possible.

Avoid unjustified omissions or additions.

Output only the final translation.

Source excerpt to translate:

{text}

Table 8: Base machine-translation prompt used to instruct the model to translate literary source text into English.

Post-Editing Prompt

You are given:

1. A literary source text written in **{source_lang}**, which is a fragment of a novel.
2. An English translation of that text produced by a machine translation system.

Your task is to revise the English translation by comparing it carefully with the original **{source_lang}** text.

When revising ensure faithfulness to the original meaning, tone, voice, register, rhythm, and stylistic effect, while producing natural, fluent, idiomatic prose in English.

Adapt wording and syntax where needed for naturalness and readability.

Preserve ambiguity, characterization, and differences in voice between narrator and characters of the original text wherever possible.

Ensure consistency of narrative perspective, register, and style throughout the text.

Keep the existing English where it already works well, but revise where needed. Correct mistranslations, unjustified omissions or additions, and awkward or unnatural phrasing, making the translation more suitable for an English-language readership.

Return only the revised English translation.

Original literary text in **{source_lang}**:
<SOURCE>
{source}
</SOURCE>

English translation to revise:
<TRANSLATION>
{text}
</TRANSLATION>

Table 9: Post-editing prompt used to revise draft translations for faithfulness, omissions, additions, and natural literary English.

Translation Reviewing Prompt

You are an expert literary translation reviewer.

You have been given the full English translation of a **{source_lang}** excerpt, along with the source text split into **{num_chunks}** chunks for traceability and targeted revision.

Read the reconstructed English excerpt at ‘**{book_name}**_en_draft.txt‘ as your primary review text. Read the source chunks at ‘inputs/source_chunk_0000.txt‘, ‘inputs/source_chunk_0001.txt‘, etc. only to verify the source and to map each finding to the chunk(s) that should be revised. Read ‘outputs/style_bible.json‘ before reviewing.

Your task is to review the complete translated excerpt for excerpt-level translation quality issues that weaken the translated work as a whole. This review is inspired by editorial manuscript review, but it is still a translation review: do not reward fluency if it comes from meaning drift, added interpretation, domestication that erases source texture, or smoothing away ambiguity that belongs to the source.

Focus on excerpt-level issues that chunk-level QE or chunk-level literary review may have missed:

- repeated translationese, flattening, over-explanation, sanitization, or generic diction across the excerpt
- mistranslations, omissions, distortions of implication, or loss of ambiguity that become visible only when the excerpt is read as a whole
- narrator or characterization flattening that emerges across the excerpt, not just at one boundary
- prose-level monotony, dead rhythm, or repeated explanatory habits that weaken the translated excerpt as a literary whole

Do not focus on seam repair, chunk-boundary duplication, term consistency, or local boundary continuity unless they are part of a broader excerpt-level literary or translational pattern. Those belong to the separate cross-chunk audit.

Report unresolved ‘MEDIUM+‘ issues. Keep findings genuinely distinct. For each issue found, record:

- ‘chunk_id‘: integer index of the affected chunk (0-based). For issues spanning a boundary, record one finding per affected chunk.
- ‘severity‘: ‘LOW‘, ‘MEDIUM‘, ‘HIGH‘, or ‘CRITICAL‘
- ‘evidence‘: quote the problematic English passage
- ‘problem‘: explain what is wrong and why it is a translation issue
- ‘rewrite_direction‘: give concrete, source-faithful guidance on how to fix it
- ‘acceptance_test‘: give a concrete criterion that would confirm the fix without drifting from the source

Write your complete findings to ‘outputs/book_review.json‘ with this exact JSON structure:
{{"verdict": "PASS", "summary": "<one-paragraph summary>", "findings": [{{"finding_id": "f_001", "chunk_id": 0, "severity": "MEDIUM", "evidence": "<quoted text>", "problem": "<description>", "rewrite_direction": "<guidance>", "acceptance_test": "<criterion>"}}]}}

Set ‘verdict‘ to ‘FAIL‘ if any finding has severity ‘MEDIUM‘, ‘HIGH‘, or ‘CRITICAL‘. Set it to ‘PASS‘ only if all findings are ‘LOW‘ or there are none.

If there are no issues at all, write:
{{"verdict": "PASS", "summary": "No excerpt-level translation issues found across chunks.", "findings": []}}

Table 10: Book-level translation review prompt used to assess a full excerpt for translation quality and literary adequacy.

Chunks Reviewing Prompt (part 1)

You are running CHUNK LITERARY REVIEW for chunk **{chunk_id:04d}**.

You are an expert literary translation reviewer. This is still a translation review, not a free-standing English workshop pass.

Before reviewing, read ‘outputs/style_bible.json‘ for register, terminology, named entities, character voice, and prose-style targets.

Source text (**{source_lang}**):
{source_text}

English translation:
{translation}

Use the adjacent context below only to judge continuity, seam quality, voice carryover, dialogue continuity, and local register at the chunk boundary. The current chunk remains the primary unit under review.

Previous source context:
{prev_source_context}

Next source context:
{next_source_context}

Previous translated context:
{prev_translation_context}

Next translated context:
{next_translation_context}

Task:

1. Review the translation through four lenses: ‘accuracy‘, ‘voice‘, ‘dialogue‘, ‘prose‘.
2. Fail by default unless the translation clearly achieves pass-level quality for the lens being judged.
3. Report all unresolved ‘MEDIUM+‘ issues. Keep findings genuinely distinct.
4. This is a translation, so do not reward fluent paraphrase that changes meaning, shifts ambiguity, modernizes register without warrant, domesticates away source texture, or explains what the source leaves implicit.
5. Prefer findings that identify literary failures caused by translation, such as flattening, over-explanation, dead rhythm, generic diction, stiff dialogue, loss of subtext, or prose that sounds mechanically literal rather than alive.
6. When a line is rough, ask whether it is rough because the source is rough or because the translation is weak. Only flag the latter.

Lens guidance:

- ‘accuracy‘: meaning, implication, ambiguity, figurative force, and source-specific texture are preserved.
- ‘voice‘: narrator stance, character presence, register, and tonal pressure feel equivalent to the source.
- ‘dialogue‘: speech sounds lived-in and character-true in English rather than thesis-like, over-formal, or flattened.
- ‘prose‘: the English has literary vitality, rhythm, precision, and image-force without drifting from the source.

Table 11: First chunk-level literary review prompt, used to identify local issues in style, voice, rhythm, and reader experience.

Chunks Reviewing Prompt (part 2)

Flag issues such as:

- over-explanation that kills subtext
- flattened or generic diction where the source is sharp or strange
- dialogue that is semantically correct but stiff, composed, or unnaturally formal
- rhythm or sentence movement that has gone dead through literalism
- euphemizing or sanitizing difficult material that the source renders directly
- unnecessary upgrading into more elegant or more quotable English than the source warrants

For each finding, include:

- ‘finding_id‘
- ‘source‘: one of ‘accuracy‘, ‘voice‘, ‘dialogue‘, ‘prose‘
- ‘severity‘: ‘MEDIUM‘, ‘HIGH‘, or ‘CRITICAL‘
- ‘chunk_id‘: must equal **{chunk_id}**
- ‘evidence‘: quote the problematic text from the English translation
- ‘problem‘: explain what is wrong and why it is a translation problem, not just a matter of taste
- ‘rewrite_direction‘: give a concrete, source-faithful revision direction
- ‘acceptance_test‘: give a concrete condition that would show the issue is fixed while staying faithful to the source

Write your review to ‘outputs/chunk_literary_review_**{chunk_id:04d}**.json‘ with this exact JSON structure:

```
{{
    "chunk_id": {chunk_id},
    "verdict": "PASS",
    "verdicts": {{
        "accuracy": "PASS",
        "voice": "PASS",
        "dialogue": "PASS",
        "prose": "PASS"
    }},
    "findings": [
        {{
            "finding_id": "clr_001",
            "source": "prose",
            "severity": "MEDIUM",
            "chunk_id": {chunk_id},
            "evidence": "<quoted English passage>",
            "problem": "<what is wrong and why>",
            "rewrite_direction": "<source-faithful fix guidance>",
            "acceptance_test": "<concrete pass/fail condition>"
        }}
    ],
    "summary": "<brief summary>"
}}
```

Set top-level ‘verdict‘ to ‘FAIL‘ if any lens fails or any finding is ‘MEDIUM+‘. Set it to ‘PASS‘ only when the chunk is source-faithful and literarily strong in English.

Table 12: Second chunk-level literary review prompt, used to guide targeted revision after local literary-quality issues are identified.

Cross-Chunks Audit Prompt

You are an expert literary translation reviewer running a CROSS-CHUNK AUDIT.

You have been given the full English translation of a **{source_lang}** excerpt, along with the source text split into **{num_chunks}** chunks for traceability and targeted revision.

Read the reconstructed English excerpt at ‘**{book_name}**_en_draft.txt‘ as your primary audit text. Read the source chunks at ‘inputs/source_chunk_0000.txt‘, ‘inputs/source_chunk_0001.txt‘, etc. only to verify cross-chunk consistency and to map each finding to the chunk(s) that should be revised. Read ‘outputs/style_bible.json‘ before auditing.

Your task is to audit only cross-chunk consistency and boundary quality. This is still a translation audit, not a free-standing English edit. Do not reward fluency if it comes from meaning drift, added interpretation, domestication that erases source texture, or smoothing away ambiguity that belongs to the source.

Focus only on issues such as:

- inconsistent translation of recurring terms, names, places, or culturally marked material across chunks
- narrator, character-voice, or register drift from one chunk to another
- tense or perspective drift across chunk boundaries
- seam problems at chunk boundaries, including dropped context, duplicated content, broken sentence flow, broken dialogue continuity, or paragraph texture mismatch
- repeated cross-chunk translation patterns that create inconsistency, such as one chunk preserving ambiguity while another over-explains a similar source pattern

Do not spend time on broader excerpt-level literary weakness, repeated translationese across the whole excerpt, generalized flattening, or global prose monotony unless they directly create a seam or consistency defect. Those belong to the separate excerpt-level book review.

Keep findings genuinely distinct. For each issue found, record:

- ‘chunk_id‘: integer index of the affected chunk (0-based). For issues spanning a boundary, record one finding per affected chunk.
- ‘severity‘: ‘LOW‘, ‘MEDIUM‘, ‘HIGH‘, or ‘CRITICAL‘
- ‘evidence‘: quote the problematic English passage
- ‘problem‘: explain what is wrong and why it is a cross-chunk translation issue
- ‘rewrite_direction‘: give concrete, source-faithful guidance on how to fix it
- ‘acceptance_test‘: give a concrete criterion that would confirm the fix without drifting from the source

Write your complete findings to ‘outputs/cross_chunk_audit.json‘ with this exact JSON structure:
{{"verdict": "PASS", "summary": "<one-paragraph summary>", "findings": [{{"finding_id": "cca_001", "chunk_id": 0, "severity": "MEDIUM", "evidence": "<quoted text>", "problem": "<description>", "rewrite_direction": "<guidance>", "acceptance_test": "<criterion>"}}]}}

Set ‘verdict‘ to ‘FAIL‘ if any finding has severity ‘MEDIUM‘, ‘HIGH‘, or ‘CRITICAL‘. Set it to ‘PASS‘ only if all findings are ‘LOW‘ or there are none.

If there are no issues at all, write:
{{"verdict": "PASS", "summary": "No cross-chunk consistency or seam issues found.", "findings": []}}

Table 13: Cross-chunk audit prompt used to check continuity, consistency, and repeated translation choices across an excerpt.

Final Revise Prompt

You are an expert literary translator performing a targeted revision.

Before revising, read ‘outputs/style_bible.json‘ for register, terminology, named entities, and character voice guidelines. Apply them consistently throughout your revision.

Excerpt-level review and cross-chunk audit have identified the following grouped issues for chunk **{chunk_id}** of the translation:

{findings}

Use the boundary context only to preserve continuity, seam quality, dialogue carryover, and local register at the chunk boundary. Do not rewrite neighboring chunks from this stage.

Boundary context:

- previous source context: **{prev_source_context}**
- next source context: **{next_source_context}**
- previous translated context: **{prev_translation_context}**
- next translated context: **{next_translation_context}**

Source text (**{source_lang}**):
{source_text}

Current translation:
{draft_translation}

Revise the translation to address all findings listed above. The findings are grouped by review source so you can distinguish excerpt-level literary issues from seam/consistency issues. Preserve the literary quality, voice, and style established in the rest of the excerpt. Make only the changes needed to fix the identified issues. This is still a translation: do not add unsupported information, erase source-specific texture, or rewrite the passage into a different effect from the original.

Write your revised translation to ‘outputs/segment_translation_**{chunk_id:04d}**.json‘ with this exact JSON structure:
{{"chunk_id": **{chunk_id}**, "translation": "<your revised English translation here>"}}

Also write a revision report to ‘outputs/final_revise_**{chunk_id:04d}**.json‘ with this exact JSON structure:
{{"chunk_id": **{chunk_id}**, "findings_addressed": [{{"finding_id": "<id>", "status_after_revision": "FIXED"}}]}}

Valid values for status_after_revision: "FIXED", "PARTIAL", "UNRESOLVED".

Table 14: Final revision prompt used to incorporate review findings into the completed machine-translated excerpt.

Translation Reviewing Prompt (part 1)

You are a professional literary translator with extensive experience. You are reviewing a translation from **{source_lang}** into English for a work of great aesthetic value and cultural significance.

Your task: answer YES, NO, or MAYBE to each of the questions below, judging whether the translation covers the given aspect of translation quality. Be honest and consider every aspect carefully.

Source text:
{source_text}

Translation:
{translation}

Use the adjacent context below only to judge continuity, seam quality, local register, speaker tracking, and whether the current chunk fits naturally with its neighbors. Do not excuse a weak current chunk just because the surrounding context makes it guessable.
Previous source context:
{prev_source_context}

Next source context:
{next_source_context}

Previous translated context:
{prev_translation_context}

Next translated context:
{next_translation_context}

Questions:

Answer every question independently. The groups below are organizational only and do not change the required output format.

Group A: Lexical Accuracy and Meaning

1. Does the translation avoid false friends, misleading cognates, and other lexical choices that distort the source meaning?
2. Do the English word choices preserve the important connotations of the source rather than only its surface dictionary meaning?
3. Does the translation preserve the meaning and force of metaphors, similes, and other figurative language?
4. Does the translation preserve plot-relevant information and avoid omissions, additions, or distortions that could create later inconsistencies?
5. Does the translation preserve the source's degree of ambiguity, explicitness, and interpretive openness?

Group B: Pragmatics and Subtext

6. Does the translation capture pragmatic meaning, including tone, irony, implicature, politeness, and subtext, rather than only literal content?
7. Does the translation preserve the passage's underlying intention and subtext, not just its explicit content?

Table 15: First LiTrans-style review prompt, used to diagnose translation quality issues before revision.

Translation Reviewing Prompt (part 2)

Group C: Cultural Transfer

8. Are idioms, humor, and figurative expressions rendered in a way that preserves their function and effect in English?
9. Are culture-specific references handled appropriately for an English-language literary reader without unnecessary flattening, over-explanation, or loss of source specificity?
10. Are culturally specific emotions, gestures, and forms of expression rendered clearly and appropriately in English?
11. Does the translation avoid excessive domestication that erases the source culture's identity?
12. Does the translation avoid exoticizing or over-marking the source culture in ways not warranted by the original?
13. Are historical, local, and cultural references understood well enough to be rendered accurately?
14. Does the translation preserve the overall cultural atmosphere and emotional world of the passage?

Group D: Voice, Register, and Style

15. Does each character or narrative voice retain the same tone, register, and level of formality as in the source?
16. Is tone and register internally consistent within this passage unless the source itself shifts deliberately?
17. Does the English narrative voice feel equivalent in stance and texture to the source narrative voice?
18. Does the translation preserve the author's stylistic character, whether plain, restrained, ornate, lyrical, ironic, or otherwise distinctive?
19. Would an English-language reader encounter substantially the same narrator or character presence as in the source?

Group E: Narrative and Local Consistency

20. Are narrative point of view and perspective shifts preserved exactly where they matter?
21. Are names, recurring terms, slang, dialectal cues, and other key local details handled consistently within this passage and in line with established translation choices?

Group F: Fluency and Literary Effect

22. Does the translation read as natural, idiomatic English while remaining faithful to the source?
23. Is descriptive imagery rendered vividly, precisely, and without dulling or over-explaining the source?
24. Does the passage work as coherent literary English rather than sounding translated, formulaic, or mechanically literal?
25. Taken as a whole, does this passage achieve faithful equivalence at lexical, syntactic, textual, and pragmatic levels?

Answer all 25 questions. Do not omit any question and do not add extra keys.

Write your answer to the file ‘outputs/litrans_answers_**{chunk_id:04d}**.json‘ as a JSON object where each key is the question number (as a string), and each value is an object with exactly these fields:

- ‘judgment‘: exactly one of ‘"YES"‘, ‘"NO"‘, or ‘"MAYBE"‘
- ‘issue‘: a short statement of the possible issue for this question

If the answer is ‘"YES"‘, the ‘issue‘ field should briefly say that no material issue was detected for that question.

Example format:

```
{{
    "1": {{"judgment": "YES", "issue": "No material issue detected for lexical distortion."}},
    "2": {{"judgment": "NO", "issue": "The English wording flattens an important connotation
in the source."}},
    "3": {{"judgment": "MAYBE", "issue": "The sentence is readable, but one phrase still sounds
slightly translated."}}
}}
```

Table 16: Second LiTrans-style review prompt, used to convert diagnosed issues into concrete revision guidance.

Revise Prompt

You are an expert literary translator.

You have produced a draft translation of the following **{source_lang}** text, and the current chunk gate has flagged it for revision.

Before revising, read ‘outputs/style_bible.json‘ for register, terminology, named entities, and character voice guidelines. Apply them consistently throughout your revision.

LiTransProQA review summary:

- score: **{score:.2f}** / 1.0
- threshold: **{threshold:.2f}**
- verdict: **{litrans_verdict}**
- failed question ids: **{failed_question_ids}**
- failed or non-YES question details:

{failed_questions}

Chunk literary review summary:

- verdict: **{literary_verdict}**
- medium+ findings:

{literary_findings}

Use the boundary context only to preserve continuity, seam quality, dialogue carryover, and local register at the chunk boundary. Do not rewrite neighboring chunks from this stage.

Boundary context:

- previous source context: **{prev_source_context}**
- next source context: **{next_source_context}**
- previous translated context: **{prev_translation_context}**
- next translated context: **{next_translation_context}**

Source text (**{source_lang}**):
{source_text}

Your previous draft translation:
{draft_translation}

Please revise the translation to improve its literary quality while remaining faithful to the source. Address the concrete weaknesses surfaced by the reviews above, especially any meaning drift, flattened voice, dead rhythm, stiff dialogue, over-explanation, or loss of ambiguity.

This is still a translation. Do not add information not supported by the source, do not domesticate away source-specific texture without need, and do not "improve" the passage by rewriting it into a different effect from the original.

Output only the revised translation, no commentary.

Write your revised translation to the file ‘outputs/segment_translation_**{chunk_id:04d}**.json‘ with this exact JSON structure:
{{"chunk_id": **{chunk_id}**, "translation": "<your revised English translation here>"}}

Table 17: Targeted revision prompt used after chunk-level review to repair identified translation and literary-quality issues.

Style Analysis Prompt

You have been given the full source text of a **{source_lang}** novel.

Read the source text at ‘inputs/full_source.txt‘.

Analyse the text for the following, then write a style bible that will guide consistent, high-quality translation of every chunk into English for the English language readership.

Focus on:

- Named characters: For each named character who speaks or is described, note their speech register (formal/informal/dialect), distinctive syntax patterns, and any recurring verbal habits.
- Prose style: Identify the overall register (elevated literary, colloquial, journalistic, etc.), literary domain and period, narrative tense, characteristic sentence rhythm, and diction level.
- Terminology: List recurring domain-specific or culturally-specific terms that require a consistent English rendering. Include untranslatable terms with a recommended translation strategy.
- Named entities: List all proper nouns (characters, places, organisations) with a decision on how each should appear in the English translation (keep source form, translate, romanise, etc.).

Write your analysis to ‘outputs/style_bible.json‘ with this exact JSON structure:

```
{{
    "character_voice_profiles": [
        {{
            "character_id": "<name>",
            "register": "<description>",
            "speech_patterns": "<description>",
            "example_lines":  ["<line from source — CALIBRATION ONLY, never copy into
translation>"]
        }}
    ],
    "prose_style_profile": {{
        "register": "<e.g. elevated literary>",
        "domain": "<e.g. 19th-century French realism>",
        "narrative_tense": "<past tense / present tense>",
        "rhythm_target": "<e.g. long, complex sentences with subordinate clauses>",
        "diction": "<e.g. archaic vocabulary preserved where unambiguous>"
    }}
    "terminology": [
        {{
            "source_term": "<term in {source_lang}>",
            "preferred_translation": "<agreed English rendering>",
            "note": "<optional: rationale or usage note>"
        }}
    ],
    "named_entities":
        {{
            "entity": "<name as in source>",
            "type": "<character | place | organisation | other>",
            "preferred_english": "<how to render in translation>",
            "note": "<optional: keep source form / translate / romanise etc.>"]
        }}
    ],
}}
```

Make sure to include any information important for producing a high quality literary translation. If a section has no entries (e.g. no named characters speak), write an empty array [].

Table 18: Style-analysis prompt used to extract voice, register, terminology, and consistency guidance before translation.

Translation with Context Prompt

You are an expert literary translator.

Your task is to translate an excerpt from a novel written in **{source_lang}** into English. Produce a fluent, idiomatic literary translation intended for an English-language readership.

Before translating, read ‘outputs/style_bible.json‘ for register, terminology, named entities, and character voice guidelines. Apply them consistently throughout your translation.

Use the adjacent source context below only to preserve local continuity, sentence carryover, dialogue flow, and register at the chunk boundary. The current chunk remains the primary unit to translate.

Previous source context:
{prev_source_context}

Next source context:
{next_source_context}

Be faithful to the original meaning, tone, voice, register, rhythm, and stylistic effect, while producing natural, fluent, idiomatic prose in English.

Adapt wording and syntax where needed for naturalness and readability, while preserving distinctive stylistic features wherever possible.

Maintain consistency of narrative perspective, register, and style throughout the text.

Preserve ambiguity, characterization, and differences in voice between narrator and characters wherever possible.

Avoid unjustified omissions or additions.

Output only the final translation.

Source excerpt to translate:

{source_text}

Write your translation to the file ‘outputs/segment_translation_**{chunk_id:04d}**.json‘ with this exact JSON structure:
{{"chunk_id": **{chunk_id}**, "translation": "<your English translation here>"}}

Table 19: Context-aware machine-translation prompt used to translate a chunk with source context and style guidance.

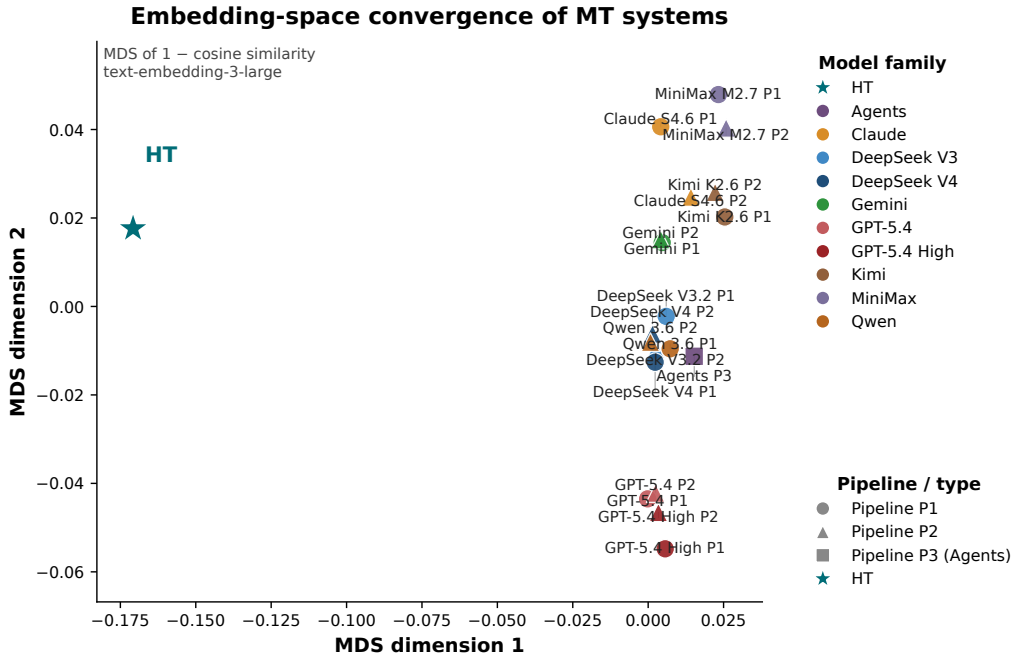


Figure 7: Embedding-space convergence of machine translation systems. Multidimensional scaling (MDS) projection of pairwise distances based on one minus cosine similarity between text embeddings. Human translation (HT) is separated from the machine translation systems, while outputs from the different MT model families and pipelines form a comparatively compact cluster.

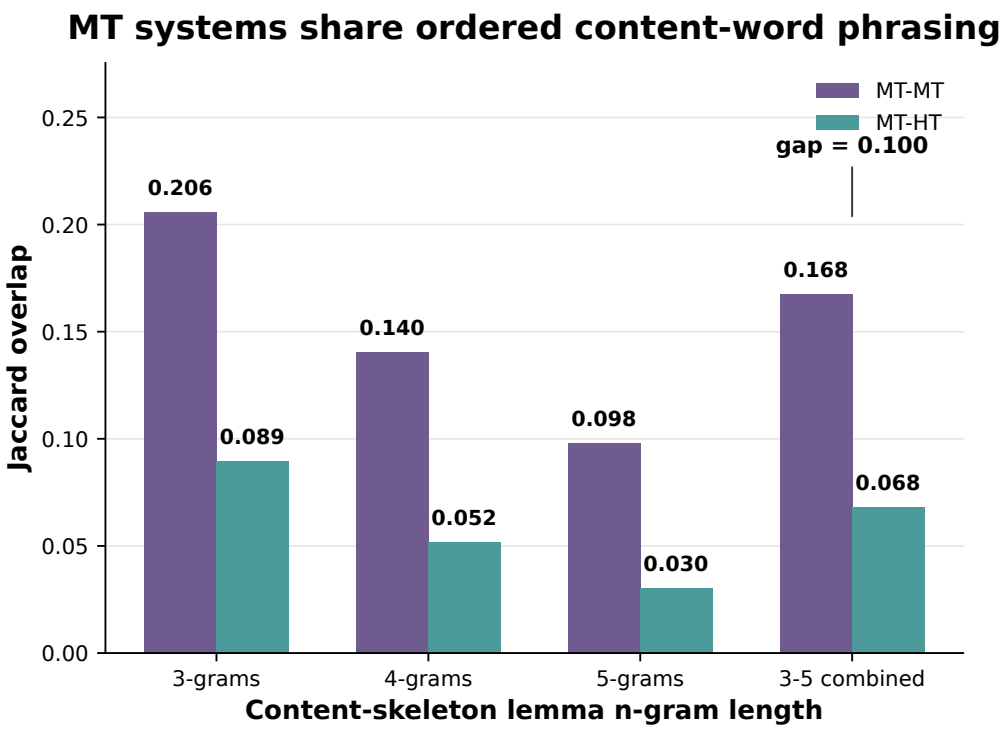


Figure 8: Ordered content-word phrasing overlap across translation outputs. Jaccard overlap of content-skeleton lemma n-grams is consistently higher between machine translation outputs (MT–MT) than between machine and human translations (MT–HT), indicating greater convergence in the phrasing patterns produced by MT systems.

Pipeline diagnostics and corpus statistics. The following tables report model costs, pipeline length statistics, automatic-evaluation results, and corpus/evaluation summary statistics.

Model				**OpenRouter price**		**Observed usage**	
Model	Provider	Params	Context	Input $/1M	Output $/1M	Tokens	Spend ($)
Claude Sonnet 4.6	Anthropic	–	1M	3.00	15.00	4.89M	31.89
Qwen3.6 Plus	Alibaba (Qwen)	∼70B	1M	0.325	1.95	13.3M	21.69
MiniMax M2.7	MiniMax	∼230B	–	0.30	1.20	5.35M	4.83
Kimi K2.6	Moonshot AI	∼1T	–	–	–	10.8M	29.05
DeepSeek V3.2	DeepSeek	∼671B	–	0.27	1.10	4.02M	1.62
DeepSeek V4 Pro	DeepSeek	–	–	–	–	305K	0.87
Total	–	–	–	–	–	**38.665M**	**89.95**

Table 20: OpenRouter model pricing and observed usage totals for MT pipeline runs. Tokens and spend are summed over all recorded calls in the screenshots.

Level	Book	Preferred	Participant comment
Excerpt	*The Witcher: Crossroads of Ravens*	MT	MT also made language choices (particularly the use of an older style of English) that are generally more consistent with this style of fantasy e.g. [MT] "Do not take offence" vs. [HT] "Don't be cross".
Excerpt	*Kokun: The Girl from the West*	HT	HT had no formatting issues that I noticed, for one, and also had a richness/uniqueness to it that pulled me in. The dialogue flowed well and, while complicated at times, is not unusual for an epic fantasy book of this nature.
Chunk	*Sisters in Yellow*, chunk 4	MT	I much preferred "bento containers" over "plastic containers" as it not only falls in line with the Japanese culture, it also doesn't sound cheap.
Chunk	*The Empusium*, chunk 18	MT	The older-style language of MT better matches the period the story is set in. HT is a little too modern, using words like "glasses" instead of "spectacles" and "frames" instead of "rims."
Chunk	*The Witcher: Crossroads of Ravens*, chunk 14	MT	I preferred the chief digger's dialogue in MT, it seemed to fit his character better. It also contrasted with the style of Preston Holt better, while in HT both characters had more similar dialogue patterns.
Chunk	*Mona's Eyes*, chunk 21	MT	The dialogue in both feels clumsy in parts, especially the way Mona speaks, but I'm guessing that might just be how the original book was written.
Chunk	*Kokun: The Girl from the West*, chunk 20	HT	A lot of the sentences were far too long, and this, combined with word choice, made the paragraphs difficult to follow (this could've just been a poorly written original scene, though, and not necessarily poorly translated).

Table 21: Examples of contextual reader judgments in excerpt-level and chunk-level preference explanations. These examples illustrate how participants evaluated translation choices against perceived story context, including genre, period, cultural setting, character voice, and reading effort. Participants saw anonymized T1/T2 labels during evaluation; we replace those labels with HT and MT here for readability.

	Pipeline	Model	TOTAL	MEAN	MEDIAN	ST. DEV.	MIN	MAX
WORDS	Pipeline 1	Claude Sonnet 4.6	248,887	6,222.2	6,816	2,222.4	2,493	11,045
		DeepSeek V3.2	229,460	5,736.5	6,142	2,039.8	2,239	10,277
		DeepSeek V4 Pro	240,700	6,017.5	6,602	2,139.8	2,352	10,736
		Gemini	246,552	6,163.8	6,908	2,145.9	2,508	10,594
		GPT-5.4	244,215	6,105.4	6,686	2,139.4	2,481	10,533
		GPT-5.4 High	246,164	6,154.1	6,850	2,156.6	2,498	10,646
		Kimi K2.6	226,903	5,672.6	5,878	1,999.0	2,168	9,369
		MiniMax M2.7	232,398	5,809.9	6,238	2,064.6	2,264	9,861
		Qwen 3.6 Plus	228,808	5,720.2	6,333	1,960.6	2,438	10,062
	Pipeline 2	Claude Sonnet 4.6	261,040	6,526.0	7,442	2,314.9	2,608	11,299
		DeepSeek V3.2	236,953	5,923.8	6,638	2,189.3	2,307	10,462
		DeepSeek V4 Pro	243,425	6,085.6	6,776	2,158.2	2,455	10,720
		Gemini	251,316	6,282.9	7,155	2,196.4	2,534	10,732
		GPT-5.4	250,702	6,267.6	7,014	2,228.8	2,500	10,908
		GPT-5.4 High	247,964	6,199.1	6,916	2,178.7	2,518	10,721
		Kimi K2.6	237,576	5,939.4	6,444	2,119.3	2,406	10,355
		MiniMax M2.7	242,830	6,070.8	6,874	2,128.9	2,476	10,432
		Qwen 3.6 Plus	249,990	6,249.8	6,984	2,369.3	2,459	12,497
	Pipeline 3	Agents	250,835	6,270.9	7,081	2,242.1	2,476	11,188
TOKENS	Pipeline 1	Claude Sonnet 4.6	311,206	7,780.1	8,672	2,784.2	3,134	12,992
		DeepSeek V3.2	296,859	7,421.5	8,016	2,639.1	2,932	12,655
		DeepSeek V4 Pro	307,589	7,689.7	8,545	2,733.1	3,047	12,927
		Gemini	312,214	7,805.4	8,738	2,757.0	3,144	12,951
		GPT-5.4	307,500	7,687.5	8,486	2,702.0	3,177	12,683
		GPT-5.4 High	309,885	7,747.1	8,638	2,727.8	3,221	12,861
		Kimi K2.6	289,842	7,246.1	7,470	2,565.0	2,777	11,399
		MiniMax M2.7	297,587	7,439.7	7,827	2,760.7	2,790	12,558
		Qwen 3.6 Plus	296,151	7,403.8	8,252	2,567.6	3,121	12,404
	Pipeline 2	Claude Sonnet 4.6	323,789	8,094.7	9,196	2,882.4	3,291	13,359
		DeepSeek V3.2	305,499	7,637.5	8,719	2,823.6	2,897	12,926
		DeepSeek V4 Pro	311,793	7,794.8	8,791	2,772.8	3,156	13,085
		Gemini	318,929	7,973.2	9,051	2,826.0	3,199	13,209
		GPT-5.4	315,982	7,899.6	8,870	2,825.0	3,193	13,254
		GPT-5.4 High	312,653	7,816.3	8,771	2,764.0	3,204	13,065
		Kimi K2.6	302,627	7,565.7	8,385	2,721.1	3,010	12,671
		MiniMax M2.7	306,735	7,668.4	8,800	2,714.1	3,050	12,729
		Qwen 3.6 Plus	322,424	8,060.6	8,921	3,070.4	3,181	16,145
	Pipeline 3	Agents	312,689	7,817.2	8,922	2,803.4	3,143	13,341

Table 22: Per-book word and token count statistics for each pipeline/model system, split into a **WORDS** sub-table (whitespace-delimited word counts) and a **TOKENS** sub-table (`tiktoken o200k_base` token counts). TOTAL is the corpus sum across all 16 books; MEAN, MEDIAN, ST. DEV., MIN, and MAX summarise the per-book distribution. Within each sub-table, rows are grouped by pipeline (Pipeline 1, Pipeline 2, Pipeline 3).

Group	Value	Runs
Run group	Evaluation books	15
Run group	Development books	25
Run group	Inner evaluation books	4
Source language	French	13
Source language	Japanese	17
Source language	Polish	14
Target language	English	40
Target language	French	1
Target language	Japanese	1
Target language	Polish	1
Target language	Spanish	1
Final Acceptance Gate	**FAIL**	28
Final Acceptance Gate	**PASS**	16
Max chunk cycles	2	25
Max chunk cycles	3	19

Pipeline step	Agent backend	Checkpoint	Runs
Style Analysis	Claude	claude-opus-4-6	44
Initial Chunk Translation	Claude	claude-opus-4-6	4
Initial Chunk Translation	GPT	gpt-5.4	40
Translation Quality Review	GPT	gpt-5.4	44
Literary Quality Review	Claude	claude-opus-4-6	44
Targeted Chunk Revision	Claude	claude-opus-4-6	4
Targeted Chunk Revision	GPT	gpt-5.4	40
Whole Excerpt Review	GPT	gpt-5.4	44
Cross-Chunk Continuity Check	GPT	gpt-5.4	44
Global Revision	Claude	claude-opus-4-6	44

Table 23: Model configuration for the agentic machine-translation pipeline. The table shows which agent backend and checkpoint were used for each translation, review, and revision step, along with the number of runs using that configuration.

Agents Pipeline Runs Summary

This appendix summarizes the finished agentic translation runs used in the study. It reports run metadata and aggregate review outcomes; generated source text, translation text, and long review evidence are intentionally excluded.

Corpus and Run Overview

Pipeline Step and Model Configuration

Artifact Footprint

The agentic translation pipeline completed 44 runs: 15 evaluation-book runs, 25 development-book runs, and 4 inner-evaluation runs. Across those runs, the pipeline processed 550 source-language chunks. Counting each completed chunk cycle as a generated chunk-level artifact yields 1,307 chunk-cycle translation or revision artifacts. Job-level logs were available for 40 of the 44 runs and record 3,743 agent jobs, including 19 failed jobs; four development-book runs had no job-metric log. At the final acceptance gate, 16 runs passed and 28 failed, with 105 chunks flagged for final-gate issues. More details can be found in Table 24.

Artifact or log unit	Count	Scope
Completed pipeline runs	44	All evaluation, development, and inner-evaluation runs.
Source-language chunks processed	550	All chunks entering the agentic pipeline.
Chunk-cycle translation/revision artifacts	1,307	Sum of source chunks multiplied by completed chunk cycles.
Runs with job-metric logs	40	Excludes four development-book runs with unavailable metrics.
Logged agent jobs	3,743	All recorded style, translation, review, revision, and continuity jobs.
Failed agent jobs	19	Logged failed jobs across runs with metrics.
Final-gate flagged chunks	105	Chunks flagged by the final acceptance gate.

Table 24: Artifact counts generated by all agentic pipeline runs.

SYSTEM			AUTOMATIC METRIC			
Model	Pipeline	Human eval.	COMET-22 ↑	COMETKiwi ↑	MetricX ↓	**MetricX-QE ↓**
Gemini	P2		**0.805**	**0.773**	**5.182**	**4.758**
Gemini	P1		0.804	0.772	5.243	4.813
Qwen 3	P2		0.798	0.768	5.428	4.890
DeepSeek V4 Pro	P2		0.800	0.770	5.402	4.926
DeepSeek V3	P2		0.797	0.769	5.468	4.927
DeepSeek V4 Pro	P1		0.801	0.771	5.414	4.933
GPT-5.4	P2		0.801	0.770	5.415	4.933
DeepSeek V3	P1		0.795	0.771	5.528	4.968
GPT-5.4	P1		0.800	0.769	5.446	4.978
Qwen 3	P1		0.797	0.769	5.488	4.987
MiniMax M2	P2		0.793	0.767	5.638	4.997
GPT-5.4 High	P2		0.800	0.768	5.482	5.020
Kimi K2	P2		0.794	0.767	5.618	5.022
Kimi K2	P1		0.792	0.768	5.640	5.043
GPT-5.4 High	P1		0.798	0.766	5.542	5.073
Agents	P3	✔	0.793	0.766	5.656	5.109
Claude Sonnet 4	P2		0.791	0.759	5.680	5.124
MiniMax M2	P1		0.785	0.761	5.873	5.191
Claude Sonnet 4	P1		0.792	0.758	5.757	5.234
HT	–	✔	–	0.726	–	5.726

Table 25: Paragraph-level automatic evaluation results on the full par3 development and evaluation corpus. Rows are sorted by MetricX-QE, ascending. Row colors distinguish P1, P2, P3, and HT systems. Scores are unweighted means over book-level system scores across both splits. Higher is better for COMET-22 and COMETKiwi; lower is better for MetricX and MetricX-QE. Bold indicates the best MetricX-QE score. The Human eval. column marks systems used in the human evaluation. Missing HT reference-based scores are shown as "–".

Metric	Chunks	HT mean	MT mean	Oriented HT–MT delta	95% CI
LitTransProQA	386	0.920	0.996	-0.076	± 0.017
MetricX-QE	385	8.657	8.347	-0.309	± 0.128
COMETKiwi (paragraph mean)	377	0.722	0.760	-0.038	± 0.004
MetricX-QE (paragraph mean)	385	5.671	5.184	-0.486	± 0.103

Table 26: Automatic evaluation of the aligned 300-word human-evaluation chunks. The oriented delta column follows the existing analysis convention: positive values favor HT, while negative values favor MT. Paragraph-mean rows map paragraph-level par3 scores onto the reviewed chunk boundaries.

			Books (n=15)	Chunks (n=386)		Participant Comments (n=60)	(n=60)	(n=30)	(n=30)	(n=772)
					#/BK	SINGLE Q5	SINGLE Q6	COMPAR. Q4	COMPAR. Q7	CHUNK JUSTIF.
WORDS	MEAN	HT	7,623.3	296.2	25.7	35.6	49.2	76.9	69.2	47.9
		MT	7,676.5	298.3						
		SRC	4,784.1	186.4						
	MEDIAN	HT	7,877	307	26	29	41.5	72.5	57	38
		MT	7,623	308						
		SRC	6,273	240						
	ST. DEV.	HT	449.3	49.2	2.2	23.3	32.8	35.9	65.1	29.6
		MT	529.2	57.4						
		SRC	3,334.1	128.4						
	MAX	HT	7,995	486	29	93	176	184	380	215
		MT	8,595	530						
		SRC	8,244	417						
	MIN	HT	6,747	100	22	9	7	24	21	16
		MT	6,928	102						
		SRC	141	1						
	TOTAL	HT	114,350	114,350	386	2,136	2,955	2,306	2,077	36,964
		MT	115,147	115,147						
		SRC	71,762	71,762						
TOKENS	MEAN	HT	9,680.3	376.2	25.7	41.3	60.5	95.2	86.2	61.3
		MT	9,582.1	372.4						
		SRC	13,026.4	507.5						
	MEDIAN	HT	9,832	390	26	34	52	86.5	68.5	48
		MT	9,648	384						
		SRC	13,023	514						
	ST. DEV.	HT	535.0	65.7	2.2	27.8	43.9	46.3	82.5	38.6
		MT	625.7	73.5						
		SRC	1,527.0	121.7						
	MAX	HT	10,261	611	29	113	254	231	482	272
		MT	10,500	644						
		SRC	15,967	1,139						
	MIN	HT	8,600	118	22	11	8	33	23	18
		MT	8,474	121						
		SRC	11,155	141						
	TOTAL	HT	145,204	145,204	386	2,480	3,633	2,855	2,586	47,326
		MT	143,732	143,732						
		SRC	195,396	195,396						

Table 27: Summary statistics for evaluation books, chunks, and participant free-text comments, split into a **WORDS** sub-table (whitespace-delimited word counts) and a **TOKENS** sub-table (`tiktoken o200k_base` token counts). Books and chunks are reported separately for human-translated (HT), machine-translated (MT), and source-language (SRC) versions. The #/BK (chunks per book) column does not depend on the metric and is identical across both sub-tables (and across HT / MT / SRC, since chunks are aligned across versions). PARTICIPANT COMMENTS include: *Single Q5/Q6* are free-text responses from the single-reading questionnaires; *Compar. Q4/Q7* are free-text responses from the comparison questionnaire; *Chunk Justif.* are per-chunk preference justifications.

Question / subset	n	Exact (%)	≤ 1 (%)	$M\|d\|$	Gwet's AC2
Q1–Q4 overall	120	35.0	64.2	1.16	0.426
HT only	60	38.3	70.0	0.98	0.616
MT only	60	31.7	58.3	1.33	0.216
Q1 Acceptability	30	36.7	73.3	0.93	0.632
HT only	15	40.0	80.0	0.80	0.765
MT only	15	33.3	66.7	1.07	0.468
Q2 Smoothness	30	36.7	70.0	1.03	0.529
HT only	15	40.0	73.3	0.87	0.725
MT only	15	33.3	66.7	1.20	0.290
Q3 Immersion	30	30.0	53.3	1.40	0.229
HT only	15	33.3	53.3	1.27	0.405
MT only	15	26.7	53.3	1.53	0.057
Q4 Continue reading	30	36.7	60.0	1.27	0.317
HT only	15	40.0	73.3	1.00	0.545
MT only	15	33.3	46.7	1.53	0.089

Table 29: Inter-annotator agreement for isolated-reading ratings by question and source type. Exact is exact percentage agreement; ≤ 1 is off-by-one percentage agreement; $M|d|$ is the mean absolute difference between raters' scores. Because the ratings are ordinal and category distributions were skewed, we report Gwet's AC2 (Gwet, 2021).

Note. HT = human-translated; MT = machine-translated; n is the number of paired ratings. AC2 colors use Landis and Koch's descriptive benchmark bands (Landis and Koch, 1977); descriptors are heuristic.

slight	0.00–0.20
fair	> 0.20–0.40
moderate	> 0.40–0.60
substantial	> 0.60–0.80
almost perfect	> 0.80–1.00

C Human Evaluation

This section documents the consent materials, entry questionnaire, participant-facing guidelines, interface screenshots, and evaluation questions used in the human study. Two participants, p009 and p012, were excluded and replaced before the final dataset was analyzed because their open-ended responses lacked sufficient detail for the qualitative analyses. All reported participant-, assignment-, excerpt-, and chunk-level analyses use the final replacement sample.

C.1 Consent Form

Table 36 displays the instructions and questions for the consent form participants signed before evaluations. It was administered online through Microsoft Forms.

C.2 Entry Questionnaire

Table 38 displays the instructions and questions for the entry questionnaire participants filled out to provide demographic details before the evaluation. It was administered online through Microsoft Forms.

C.3 Eval Questionaire

C.4 Evaluation Interface

Figure 9 shows the immersive reading interface where participants read and evaluated HT and MT texts in isolation, with reading order being counterbalanced across all participants. Figure 10a shows the comparison interface where participants answered questions about the holistic differences between the two translations. Figure 10b shows the close reading interface where readers evaluated chunks of the two translations side-by-side to detect local differences; display order was randomized for each chunk pair.

C.4.1 Display normalization

We applied lightweight display normalization to reduce typographic inconsistencies across texts, including normalizing quotation marks, guillemets, apostrophes, and ellipses. This normalization was used for presentation and annotation consistency, and was applied directly in the evaluation interface without affecting the source files. The same displayed text representation was used when recording span highlights, so participant offsets refer to the normalized interface text.

Book	Lang	Pairs	Agr.	AC1	Pairs excl. unsure (#p1/#p2)	Agr. excl. unsure	AC1 excl. unsure
Make Me Famous	FR	27	44.4%	0.023	$13_{(5/11)}$	46.2%	0.112
Mona's Eyes	FR	29	48.3%	-0.015	$8_{(21/5)}$	50.0%	0.059
Symphony Of Monsters	FR	26	92.3%	0.917	$22_{(4/0)}$	95.5%	0.952
The Story Of Marceau Miller	FR	24	45.8%	0.050	$9_{(10/8)}$	55.6%	0.200
Watching Over Her	FR	28	50.0%	0.113	$11_{(2/15)}$	54.5%	0.197
Hooked: A Novel Of Obsession	JP	29	75.9%	0.694	$19_{(9/4)}$	73.7%	0.659
Kokun: The Girl From The West	JP	24	91.7%	0.909	$21_{(3/0)}$	90.5%	0.895
My Grandfather, The Master Detective	JP	22	95.5%	0.948	$15_{(6/2)}$	93.3%	0.929
Sisters In Yellow	JP	26	76.9%	0.710	$14_{(9/5)}$	71.4%	0.622
The Ark	JP	23	69.6%	0.391	$11_{(11/2)}$	63.6%	0.323
Inner Space	PL	26	42.3%	-0.076	$8_{(13/14)}$	75.0%	0.600
My Name Is Stramer	PL	26	65.4%	0.354	$5_{(19/12)}$	60.0%	0.412
Needle's Eye	PL	25	88.0%	0.865	$17_{(7/2)}$	88.2%	0.868
The Empusium: A Health Resort Horror Story	PL	28	50.0%	0.200	$17_{(11/0)}$	41.2%	-0.006
The Witcher: Crossroads Of Ravens	PL	23	26.1%	-0.467	$4_{(16/12)}$	75.0%	0.529

Table 28: Chunk-level preference agreement by book.

D Human Evaluation Results

This section reports additional agreement analyses, dataset statistics, coding schemes, automatic metrics, and detailed human-evaluation figures that support the main results. Detailed result figures are grouped in §D.3 and §D.3; participant-profile tables are grouped in §D.3.

D.1 Agreement

Table 29 reports inter-annotator agreement for each isolated reading question, calculated using Gwet's AC2 (Gwet, 2021). Table 33 reports chunk-level preference agreement between annotators, including all chunks, and only chunks where annotators were sure of which chunk they preferred. Table 28 reports chunk-level agreement by book.

Measure	Pairs	Agreement	Gwet AC1
All chunks	386	63.7%	0.355
Excl. unsure	194	72.2%	0.546

Table 33: Binary inter-annotator agreement for chunk-level translation preference.

Table 34: Codebook-based reader-comment classification prompt used to label stated reasons for preferring one translation over another. The variable **{criteria_codebook}** contains the allowed codebook labels and definitions, **{few_shot_examples}** is the optional example block, **{preferred_translation}** identifies whether T1 or T2 was chosen by the reader, and **{comment}** contains the reader's explanation. We did not supply few-shot examples in the reported annotation run.

Reader-Comment Codebook Prompt

```
You are an expert annotator performing codebook-based classification of reader comments about two
translations.

Each comment explains why a reader preferred one translation over the other. The translations are
referred to as T1 and T2.

Your task is to assign labels from the provided codebook to the reader's stated reasons.

Important definitions:
  - The preferred translation is the translation chosen by the reader.
  - The non-preferred translation is the translation not chosen by the reader.
  - POS labels describe explicitly stated strengths of the preferred translation.
  - NEG labels describe explicitly stated weaknesses of the non-preferred translation.

General rules:
  1. Read the reader's comment carefully.
  2. Assign POS labels only for positive qualities explicitly attributed to the preferred translation.
  3. Assign NEG labels only for negative qualities explicitly attributed to the non-preferred
translation.
  4. Do not infer, assume, or add reasons that are not clearly stated in the comment.
```

5. Do not automatically mirror labels across POS and NEG.
   - For example, if the reader says the non-preferred translation is “awkward,” assign the relevant NEG label.
   - Do not also assign a POS label unless the reader explicitly says the preferred translation is fluent, natural, smooth, or similar.
6. Comparative comments should be coded according to what is explicitly stated.
   - If the reader says “T1 is more natural than T2,” and T1 is preferred, assign a POS label that applies to this from the codebook.
7. Use only labels that appear in the allowed-label list for the corresponding output column.
8. Place each label under the correct POS/NEG category and broader category according to the codebook.
9. For every label you assign, include:
   - the label name,
   - a brief reason explaining why the label applies,
   - an exact quote from the reader comment that supports the label.
10. The evidence must be copied directly from the reader comment.
11. If no label applies to a category, use an empty array [].
12. Output valid JSON only. Do not include markdown, commentary, or explanations outside the JSON.

Criteria codebook:
{criteria_codebook}

Few-shot examples:
{few_shot_examples}

Preferred translation:
{preferred_translation}

Reader comment:
{comment}

Return your answer in exactly this JSON format:

```
{
  "POS A. Language-level features": [
    {
      "label": "",
      "reason": "",
      "evidence": ""
    }
  ],
  "POS B. Narrative-level features": [
    {
      "label": "",
      "reason": "",
      "evidence": ""
    }
  ],
  "POS C. Reader experience": [
    {
      "label": "",
      "reason": "",
      "evidence": ""
    }
  ],
  "POS D. Meta-translation": [
    {
      "label": "",
      "reason": "",
      "evidence": ""
    }
  ],
  "NEG A. Language-level features": [
    {
      "label": "",
      "reason": "",
      "evidence": ""
    }
```

```
  ],
  "NEG B. Narrative-level features": [
    {
      "label": "",
      "reason": "",
      "evidence": ""
    }
  ],
  "NEG C. Reader experience": [
    {
      "label": "",
      "reason": "",
      "evidence": ""
    }
  ],
  "NEG D. Meta-translation": [
    {
      "label": "",
      "reason": "",
      "evidence": ""
    }
  ]
}
```

D.2 Annotation schema results

Table 44 shows the top positive labels assigned to comments by readers who preferred HT in the immersive reading comparison questionnaire to indicate why they preferred it. Table 45 shows the top positive labels for comments from readers who preferred MT in that questionnaire. Table 46 shows the top positive labels assigned to comments by readers who preferred HT in the close reading per-chunk questionnaire to indicate why they preferred it. Table 47 shows the top positive labels for comments from readers who preferred MT in that questionnaire.

D.3 Statistical Test Results

For ordinal Acceptability, Smoothness, Immersion, and Would Continue Reading ratings (Likert 1-5 scale), we fit cumulative-link models in R. We use cumulative link mixed models when the random-effects structure is estimable and cumulative link models when the mixed model is singular. Table 35 reports the fitted R formula for each outcome and the two focal effects: translation type (`typeMT`, with HT as the reference) and presentation order (`orderMT-first`). HT is **4.0 times**[21] more likely to be rated higher for Acceptability than MT ($p = 0.0069 < 0.05$) and **4.3 times**[22] more likely to be rated higher for Smoothness ($p = 0.0029 < 0.05$). HT is also rated higher for Immersion and Continue Reading, but this effect is not statistically significant.

[21] OR (odds ratio) for MT is 0.25, meaning that MT is $1 \div 0.25 = 4.0$ times less likely to be rated higher than MT.

[22] OR for MT is 0.23; $1 \div 0.23 = 4.3$

#	Question	Response Options
1	**How acceptable did this translation feel overall as a published version of the book?** *Please think about your overall impression of this version as a published translation.*	completely unacceptable somewhat unacceptable acceptable with some issues mostly acceptable fully acceptable
2	**How smooth was this text to read?** *Please think about how naturally and clearly the sentences flowed while you were reading.*	very unsmooth somewhat unsmooth neutral somewhat smooth very smooth
3	**To what extent did the wording and style help you stay immersed in the story?** *Please think about whether the wording and style helped you stay absorbed in the story, or pulled you out of it.*	interfered a lot with immersion interfered somewhat made little difference supported immersion somewhat supported immersion a lot
4	**Would you want to continue reading the book in this version?** *If you were to continue reading, would you be willing to continue this version given its wording and style?*	definitely not probably not not sure probably yes definitely yes
5	**What, if anything, stood out positively in the wording or style of this version?**	Open-ended response
6	**What, if anything, stood out negatively in the wording or style of this version?**	Open-ended response
7	**Do you think this version was…**	human-translated machine-translated
8	**How confident are you in that guess?** *(Regarding Question 7)*	not at all confident slightly confident moderately confident very confident extremely confident

Table 30: Immersive-reading questionnaire shown after each isolated excerpt. Questions Q1–Q4 collect ordinal ratings of acceptability, smoothness, immersion, and willingness to continue reading; Q5–Q8 collect free-text explanations, origin guesses, and confidence.

#	Question	Response Options
1	**Which translation handled dialogue better overall?** *[Dialogue]*	Translation 1 Translation 2 No meaningful difference
2	**Which translation's word choice felt more varied and/or expressive?** *[Word choice]*	Translation 1 Translation 2 No meaningful difference
3	**Which translation would you prefer to continue reading more?**	Translation 1 (clearly better) Translation 1 (slightly better) Translation 2 (slightly better) Translation 2 (clearly better)
4	**Why? Please motivate your choice in 3–5 sentences.** *Tell us why you would prefer the translation you selected. How was it better than the other one? Use T1 to refer to Translation 1 and T2 to refer to Translation 2.*	Open-ended response
5	**Which version is more likely to be AI-translated?**	Translation 1 Translation 2
6	**How confident are you in that choice?** *(Regarding Question 5)*	not at all confident slightly confident moderately confident very confident extremely confident
7	**Why do you think it is more likely to be AI-translated?** *What are the indicators that make you think so?*	Open-ended response

Table 31: Comparison questionnaire shown after participants read the paired HT and MT excerpts for the same book. The questions collect translation preference, preference strength, reasoning, perceived machine-translation origin, and confidence.

#	Question	Response Options
1	**Step 1.** Highlight wording quality *Read both translations and mark phrases that sound especially **good** or **poor**. Select text, then choose a label in the popover. Click an existing highlight to remove it.*	Open-ended reading and highlighting
2	**Step 2.** Which translation do you prefer?	T1 T2
3	**Step 3.** Explain your choice (refer to translations as T1 and T2).	Open-ended response
4	**Step 4.** Was this choice difficult?	No, because the translation I chose was significantly better. No, because the translation I chose was better. Yes, because both translations were of similar quality.

Table 32: Close-reading per-chunk questionnaire. Participants choose the preferred translation for each aligned chunk, report preference strength, optionally highlight supporting spans, and explain the judgment.

Outcome	Predictor	Estimate	SE	z	p	OR (odds ratio)	95% CI (confidence interval)
Acceptability	`clmm(q1 ~ type + order + (1|reader) + (1|book), data = model_data, link = "logit")`						
	`typeMT`	-1.38	0.51	-2.70	0.0069	0.25	[0.09, 0.68]
	`orderMT-first`	0.41	0.49	0.85	0.3978	1.51	[0.58, 3.90]
Smoothness	`clm(q2 ~ type + order, data = model_data, link = "logit")`						
	`typeMT`	-1.49	0.50	-2.98	0.0029	0.23	[0.08, 0.59]
	`orderMT-first`	0.30	0.48	0.63	0.5279	1.36	[0.53, 3.51]
Immersion	`clm(q3 ~ type + order, data = model_data, link = "logit")`						
	`typeMT`	-0.83	0.48	-1.74	0.0815	0.44	[0.17, 1.10]
	`orderMT-first`	0.27	0.47	0.56	0.5724	1.30	[0.52, 3.30]
Continue reading	`clmm(q4 ~ type + order + (1|reader) + (1|book), data = model_data, link = "logit")`						
	`typeMT`	-0.79	0.50	-1.58	0.1150	0.45	[0.17, 1.21]
	`orderMT-first`	0.07	0.49	0.14	0.8860	1.07	[0.41, 2.79]

Table 35: Cumulative-link model results (Christensen, 2023) for immersive-reading ratings ($n = 60$ observations per outcome).

Note. HT is the reference level for `type`, so odds ratios below 1 indicate lower odds of receiving a higher rating for MT than HT.

Consent Form

Consent Form – Evaluating machine translation of literary texts (Study #[REDACTED])

Thank you for your interest in our study!

Who is conducting the study? The co-principal investigators (co-PIs) are [REDACTED] and [REDACTED]. Another researcher involved in data collection is [REDACTED], a student at [REDACTED].

Contact: [REDACTED] or [REDACTED]. You can also reach out to us via the Upwork messaging tool.

Who is funding this study? [REDACTED] awarded to [REDACTED].

Why are we doing this study? We want to investigate the quality of literary machine translation (i.e., Can AI translate literary texts as well as humans)

Who can participate in this study? Participants are adults who are comfortable reading long literary texts in English (up to 8,000 words) and who read fiction in English regularly.

Your participation is voluntary. You have the right to refuse to participate in this study. Refusal to participate or withdrawal/dropout after agreeing to participate and signing this consent form will not have any adverse effects or consequences.

How is the study done? This study takes place online and will last about 4 to 4.5 hours, divided into 2 sessions over a period of a few days. If you agree to take part: (1) You will first **read this consent form** and indicate whether you agree to participate. (2) You will use your own device to complete the study online. **A personal computer is recommended.** (3) You will complete a **short questionnaire about your demographic**, language background, and reading habits (~5min). **In Part 1 (about 1.5 hours),** you will read a long excerpt from a novel in translation (~8,000 words in English), answer a small number of questions about it, provide a written comment, and indicate whether you think it was translated by a person or produced using artificial intelligence. You will then read a second translation of the same excerpt and complete the same set of questions. At the end of Part 1, you will choose which version you preferred overall and explain why. You are encouraged to take a break between the readings. **In Part 2 (about 2.5 hours),** which will take place a few days later, you will compare the two translations side by side in shorter chunks. For each chunk, you will highlight good and bad phrasing, choose the version you prefer, briefly explain why, and indicate whether the choice felt easy or difficult. You may take breaks during the study whenever needed. We may ask you for additional clarification if your comments are unclear.

Risks, benefits, and your privacy

What are the risks of participating? This study poses a minimal risk: none of the tasks involved in the experiment will have any direct or lasting impacts on participants. Possible inconveniences include fatigue or boredom, and we encourage taking breaks to prevent these.

What are the benefits of participating? We do not think that taking part in this study will have direct benefits for you. However, in the future, others may benefit from what we learn from this study. Specifically, we think that this study might help us understand more about the quality of translation and how automatic machine translation works.

Will you be paid for taking part in this research study? You will receive payment of $110 USD for the entire task (part 1 and part 2). You will receive the money within one week of submitting your answers via the Upwork payment system. If you withdraw before completing the study, compensation will be prorated based on the amount of the study completed.

How will your privacy be maintained? Your confidentiality will be respected. Information that discloses your identity will not be released without your consent.

- All participants will be identified only by a unique code number.
- Participants will not be identified by name in any reports of the completed study. Name will be collected and stored by the PIs in a separate spreadsheet, and only for the purposes of matching name and code in case you decide to withdraw.
- The study involves a questionnaire. The initial demographic questionnaire will be collected via Microsoft Forms, which stores data on Canadian servers. This data will be stored in a secure [REDACTED]-hosted server to ensure your information remains within our local, protected environment. Consent forms and study responses will initially be collected through secure online tools used for this study. After collection, the data will be downloaded and stored on a password-protected [REDACTED] researcher's computer.
- Access to study data (identifiable only through participant codes) will be limited to the PI and research assistants, all of whom have completed the required ethics training.

What if you decide to withdraw your consent to participate? You may withdraw from this study at any time without giving any reasons, up to two weeks from the date next to your signature in the consent form. If you choose to enter the study and withdraw later, all data collected about you will be destroyed. Please contact us through email ([REDACTED] or [REDACTED]) and state that you wish to withdraw your data from this study. Please include the date of your participation. After two weeks have passed from the date of collection, we will not be able to accommodate a withdrawal request, as study data will already have been used in presentations and publications.

Results, contact information, and future data usage

Study results. The main findings will be reported in academic presentations and publications.

Contact for Complaints. If you have any concerns about your rights as a research participant and/or your experiences while participating in this study, please contact the Director, Research Ethics at [REDACTED] or [REDACTED].

Continued on next page

Consent Form (continued)	
Future use of participant data. The data will be coded and stored on [REDACTED] servers. No identifying information will be shared outside of our research group. All data that we collect will be used for research purposes only. De-identified study data will be made available to other researchers for future, unspecified use in academic publications and open-access servers. While limited study data will be accessible to others once published (e.g., age, gender, language background), **we will never share identifying or confidential information.**	
Participant consent and signature. Taking part in this study is entirely up to you: You have the right to refuse to participate. If you decide to take part, you may pull out of the study at any time without giving a reason and without any negative impact by sending an email to [REDACTED] or [REDACTED]. Your continuation on the online experiment platform indicates that you received a copy of this consent form for your own records and that you consent to participate in the study. By consenting, you acknowledge that you have not waived any rights to legal recourse in the event of research-related harm.	
Participant consent and signature. Taking part in this study is entirely up to you: You have the right to refuse to participate. If you decide to take part, you may pull out of the study at any time without giving a reason and without any negative impact by sending an email to [REDACTED] or [REDACTED]. Your continuation on the online experiment platform indicates that you received a copy of this consent form for your own records and that you consent to participate in the study. By consenting, you acknowledge that you have not waived any rights to legal recourse in the event of research-related harm.	
Primary Study Participation (*Required*)	**YES** – I have read and understood this Consent Form and agree to participate in the study **NO** – I do not wish to participate in the study.
Study Results (*Optional*)	**YES** – I would like to be notified of the results or findings of this study (e.g., via a summary report or infographic). I understand that the researchers will contact me through the secure Upwork messaging app. **NO** – I do not wish to be contacted with the study results.
Please type your full name (*Required*)	Open-ended response
Date (*Required*)	Date input field

Table 36: Consent form presented to the participants prior to the evaluation.

Book	*Inner Space*, chunk 23	
	Participant A	**Participant B**
Preference	HT	MT
Comment	They were of similar quality again, but I preferred T2's wording in general. The double hyphen in T1 is unmissable, and the first sentence was also a bit awkwardly phrased when compared to T2.	Might just be personal choice, but I like *puke* versus *spew*. In the context of them being in space, *spew* just felt like the wrong word choice. T2 has editing issues, but the word choice wins.
Coding labels	POS: A3 Word choice: naturalness NEG: A3 Word choice: naturalness NEG: D4 MT/AI tell named	POS: A5 Word choice: register/contextual fit NEG: A5 Word choice: register/contextual fit
HT text	Even though she had been here before, though she had trained on hundreds of parabolic flights to accustom her body to weightlessness, Lucy felt dizzy and her stomach clenched. Up, down—suddenly, these words lost all meaning and were useless. [A]She felt as if she were inside an optical illusion where nothing was where it should be and the longer you looked, the more befuddled your senses became. [A]Lucy could feel acid in her gullet. Don't spew, she thought. You really mustn't spew. Somebody was already hugging her and patting her back. She almost hit her head on the camera lens. They were in the Russian segment, in the Rassvet module, so it was cramped. [A]It would have been tight for two people, let alone nine. They teemed like fish in a net. It would have been easier for Lucy to plan a move to checkmate in three-dimensional chess than to say who was where at any moment, or whose hand or leg she could see.	
MT text	Although she had been here before, [A]although she had taken hundreds of parabolic flights, getting her body used to weightlessness, her head still spun and her stomach clenched. Up, down – suddenly the words had lost their meaning, [B]become useless. She felt as though she had found herself inside an optical illusion in which nothing was where it ought to be, one that [B]tangled the senses more and more the longer you looked at it. Lucy tasted sour saliva at the back of her throat. [A]Don't puke, she thought, just don't puke. Someone was already hugging her, slapping her on the back. She almost hit her head on a camera lens. They were in the Russian segment, in the Rassvet module, so it was cramped, too cramped for two people, let alone nine. [A]They writhed like fish in a net.	
Legend	positive span · positive overlap · negative span · negative overlap · positive/negative overlap	

Table 37: Illustrative close-reading disagreement from *Inner Space*. The table preserves only the information needed to read the case: participant comments, preference labels, qualitative coding labels, and the highlighted evidence spans in the aligned human translation (HT) and machine translation (MT). Superscript span tags identify participant A, participant B, and same- or cross-valence overlaps.

Entry Questionnaire

`Entry Questionnaire`

`Thank you for your interest in participating in our study!`

`This short questionnaire helps us understand your background, language experience, and reading habits. It should take only` **`~5 minutes`** `to complete.`

`Your responses will be used solely for research purposes and to help us ensure that you meet the requirements of this study.`

`Thank you for your time and we appreciate your help!`

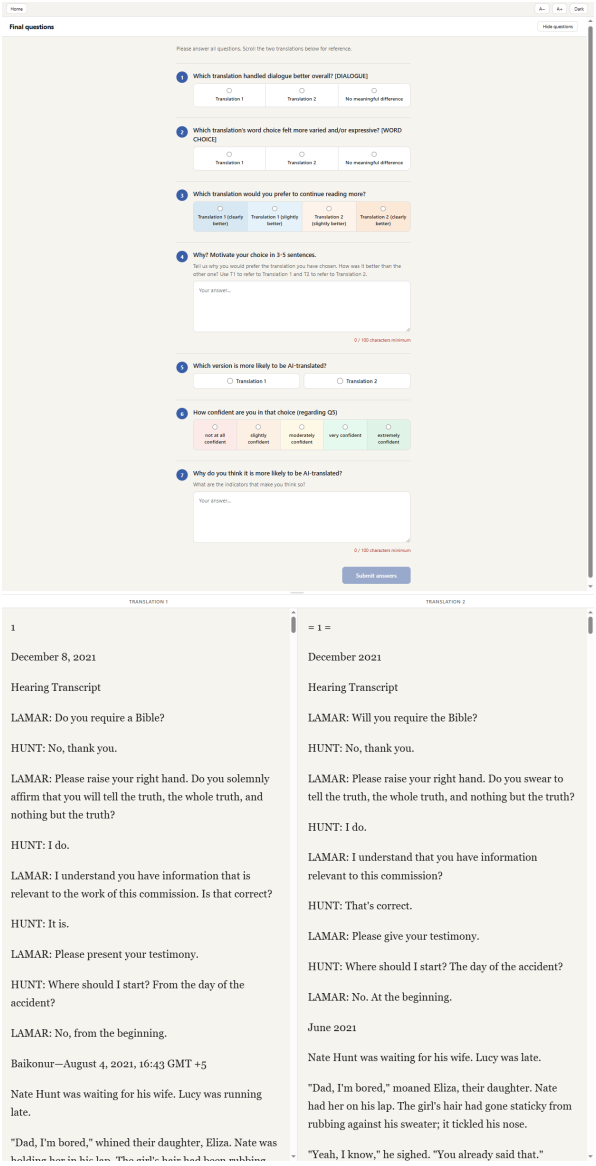
#	Question	Response Options	#	Question	Response Options
1	**Your ID**	Open-ended response	8	**How do you use AI models?** (if at all)	Open-ended response
2	**What is your age range?**	18-25; 25-34; 35-44; 45+; Prefer not to say	9	**What is your native language?**	English; Other (open-ended response)
3	**What is your gender?**	Woman; Man; Non-binary; Prefer not to say; Other (open-ended response)	10	**Do you primarily read in English?**	Yes; No
4	**What is the highest level of education you are currently in, or have completed?**	High school or equivalent; College / university undergraduate; Master's degree; Doctoral degree; Prefer not to say; Other (open-ended response)	11	**How comfortable are you reading long literary texts in English?**	Very uncomfortable; Somewhat uncomfortable; Neither comfortable nor uncomfortable; Somewhat comfortable; Very comfortable
5	**What is your nationality?**	Open-ended response	12	**Do you know any of these languages (pre-intermediate level or above)?** Select all that apply.	Japanese; Polish; French; None of these
6	**How often do you use AI models** (such as ChatGPT or Gemini)?	Never; Rarely; Sometimes; Often; Always	13	**How often do you read literary fiction in English?**	Almost never; A few times a year; About once a month; A few times a month; Weekly or more
7	**What AI models do you use most often?** (if any)	Open-ended response	14	**How often do you read translated fiction in English?**	Almost never; A few times a year; About once a month; A few times a month; Weekly or more; Not sure

Table 38: Entry questionnaire presented to the participants prior to the evaluation.

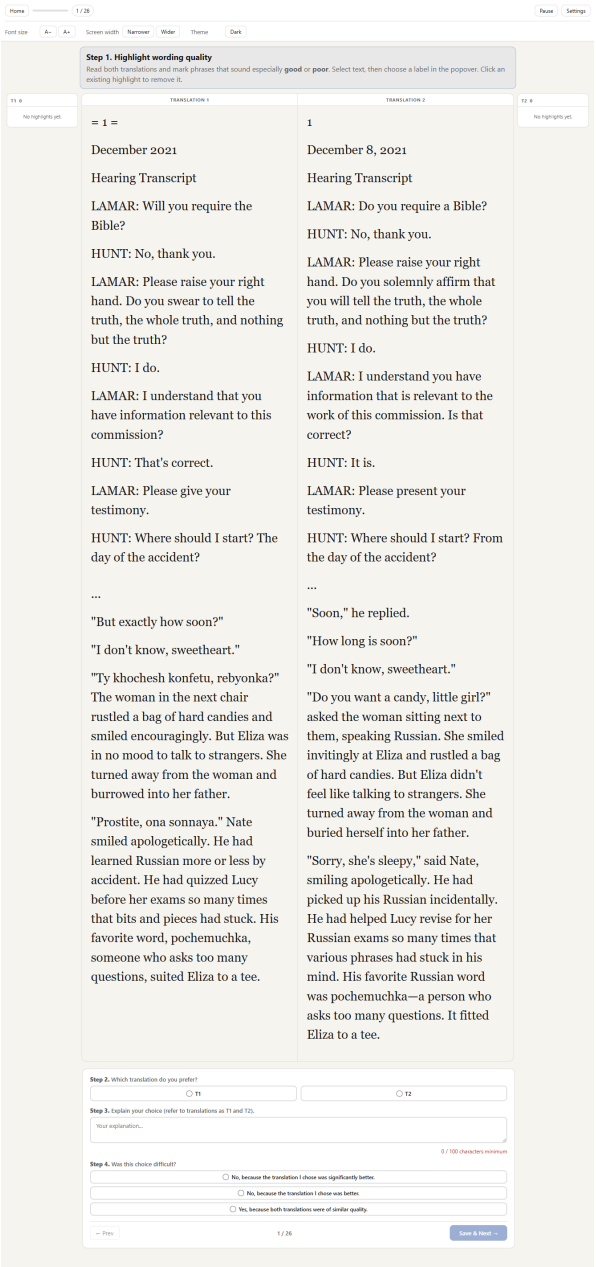
#	Question
1	What were your expectations for the quality of AI translation before starting this experiment?
2	You **{correctly/incorrectly}** identified the AI-translated version of Book 1, and **{correctly/incorrectly}** identified the AI-translated version of Book 2. Are you surprised by your accuracy?
3	What do you think about the quality of AI translation now, after reading the translation?
4	Any other comments?

Table 39: Post-study questionnaire. All responses are free text.

Figure 9: Isolated-reading questionnaire shown after each single translation. Participants rated fluency, literary quality, immersion, willingness to continue reading, and whether the passage seemed AI-generated.



(a) Post-reading comparison questionnaire



(b) Per-chunk close-reading questionnaire

Figure 10: Human-evaluation questionnaire interfaces used after paired readings. The comparison form records whole-excerpt preferences; the per-chunk form records local preference judgments and reasons.

Detailed qualitative result figures. The figures below collect supplementary visualizations for positive and negative aspect codes, preference-rationale codes, close-reading preference outcomes, and book-level origin-identification and preference summaries.

Why are we collecting this data?

We are a research lab from **[REDACTED]** in [REDACTED]. This group is supervised by [REDACTED] and [REDACTED]. We are trying to understand how well current AI systems can translate novels into English and whether these remain enjoyable for **you** (the reader).

Please do NOT hesitate to reach out to us via the Upwork messaging tool. You can also contact us via email listed in the consent form.

Please read these guidelines to avoid redoing the task.

What is TASK 1?

In this task, you will be given a **link** to two competing translations of the same book over the Upwork messaging tool, along with a **username** and a **password.** We would like to ask you to do the following:

1. Please log in to the tool and **read the entire first excerpt (translation).**
2. After completing the reading, please **fill out the questionnaire about the first excerpt.**
3. Take a **break** as you require!
4. Open the second translation and **read the entire second excerpt (translation).**
5. After completing the reading, please **fill out the questionnaire about the second excerpt.**
6. After reading both translations and filling out both questionnaires, please fill in the questionnaire about **both excerpts.**

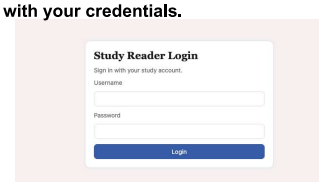
We estimate that all these steps will take you about 1.5h. On average, reading one translation alone would likely take you about 30-40 minutes. You can also take breaks as needed.

Below, we provide screenshots with explanations. Please complete this task on a laptop or large screen.

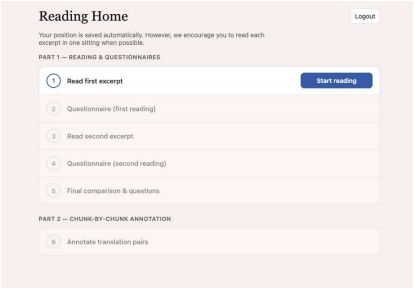
(a) Guidelines: Page 1

TASK 1: Guidelines

1. Go to the website shared with you via UpWork and log in with your credentials.

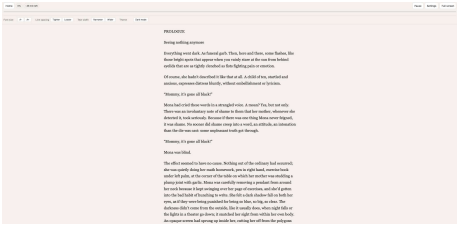


2. You will see the entire experiment flow from your panel.



(b) Guidelines: Page 2

3. Start reading the first excerpt (first translation)

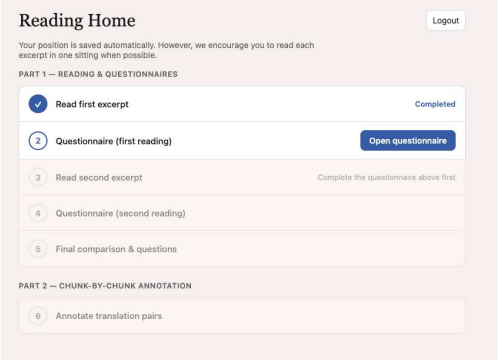


You can **adjust** the reading screen by changing font size, etc., on the left upper side of the panel:

IMPORTANT: If you need a break **during reading,** please press the **PAUSE** button located on the upper-left part of the screen:

(c) Guidelines: Page 3

4. After completing reading the first excerpt, go to the first questionnaire.

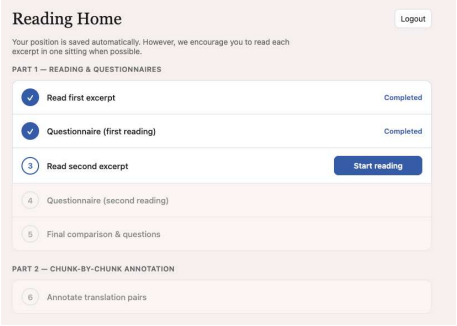


(d) Guidelines: Page 4

You will be asked the following **8 questions**, and you will also have access to the **entire translation** on the **right side of the screen**. Please read the questions now.

(e) Guidelines: Page 5

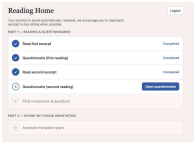
5. After you complete this questionnaire, you will be able to access the second translation. Please take a break before reading the second translation.



You should read the second translation. Just as with the first translation, you can adjust the screen and take breaks by clicking the **PAUSE** button.

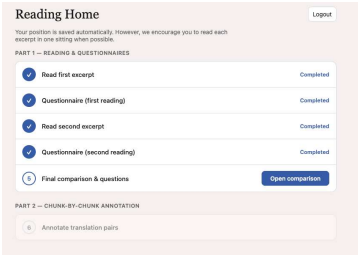
(f) Guidelines: Page 6

6. After reading the second excerpt, you will gain access to the second questionnaire.



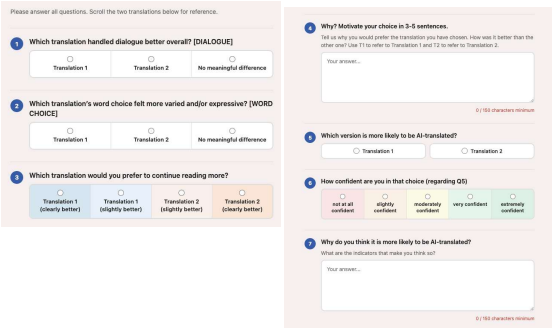
You will see **the same 8 questions** as in the first questionnaire.

7. After both readings and both questionnaires, we will ask you to compare the two translations.



(g) Guidelines: Page 7

In this questionnaire, you will be asked to answer **7 questions** that compare both translations. You will have access to both translations at the bottom of the page. Please refer to translation 1 as **T1**, and to translation 2 as **T2.**



(h) Guidelines: Page 8

What is TASK 2?

In task 2, we will ask you to **compare smaller chunks of both translations** (about 300-400 words each). The goal of this task is to look closer at the translation quality of both versions.

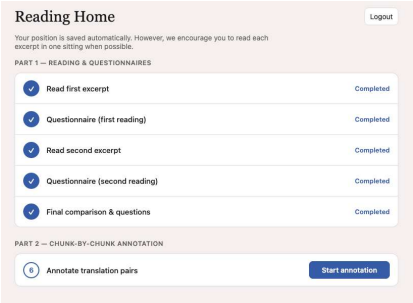
You will have to:

- **Read** both fragments.
- **Highlight** good and poor phrasing.
- **Choose** the translation you like better and tell us why.
- Depending on the chunk length, you will see about 25-30 pairs.

This part will become available after **20h from finishing the first task.** Please reach out via UpWork if you have any questions or need us to unlock it earlier.

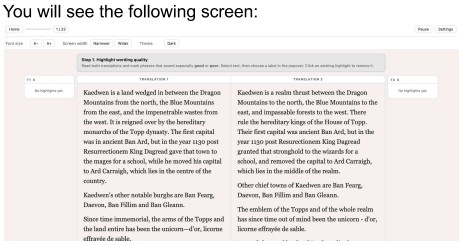
TASK 2: Guidelines

Once the task becomes available, you can start the annotations by clicking on the "Start annotation" button.



(i) Guidelines: Page 9

You will see the following screen:



STEP 1 [important]: Please **highlight good wording** and **poor wording**/potentially problematic wording by dragging your cursor over the text and choosing the correct label:

We mostly want to know what you like and what you don't like. What phrasing seems like a good translation, and what does not?

(j) Guidelines: Page 10

You don't have to highlight everything. But we want you to **use this functionality.**

Highlight **words, phrases, or longer chunks of text** that appear to be well or poorly written.

Poor wording: This may include phrases that sound weird, out of place, flat, mechanistic, etc. Or other wording that you don't like or just reads poorly.

Good wording: This may include phrases that are really good, well-written, natural, vivid, memorable, stylistically strong, emotionally effective (not flat), etc. Or other wording that captures the atmosphere, humor, tension, or helps to imagine the scene better.

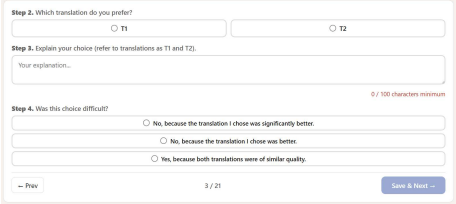
You will see it added on the side of the text (left for the text on the left and right for the text on the right):

(k) Guidelines: Page 11

STEP 2: Choose the translation you prefer.

STEP 3: Explain your choice. Please refer to translation 1 (left side) as **T1** and to translation 2 (right side) as **T2.**

STEP 4: Choose how difficult this choice was (i.e., easy choice if the difference is clear or hard choice if both are similar).



Press "Save & Next" to go to the next chunk.

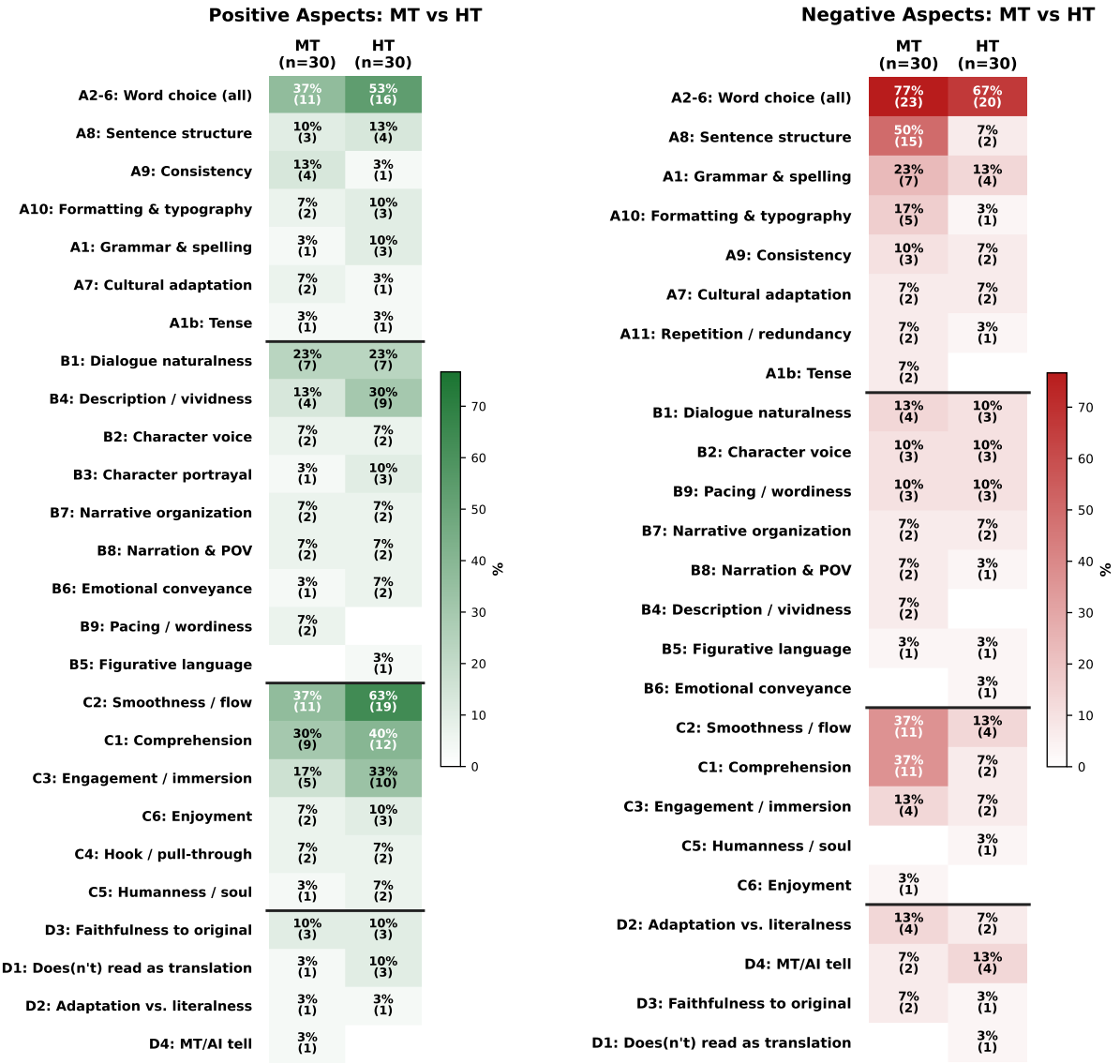
----------END----------

Thank you for reading! Please **send us this code** to get your credentials and access to the interface: **[REDACTED]**

F

(l) Guidelines: Page 12

Figure 11: Guidelines provided to the participants for the evaluation and annotation tasks



(a) Positive aspects

(b) Negative aspects

Figure 12: Positive and negative aspects mentioned by participants when comparing machine translation (MT) and human translation (HT). Each heatmap reports the percentage and count of comments associated with each code for MT and HT, grouped by linguistic quality, narrative qualities, reader experience, and meta-translation judgments.

Reasons readers gave for preferring one translation over the other (excerpt-level)

Category	MT	HT
A. Language-level features		
Word choice: naturalness (A3)	3 (27.3%)	6 (31.6%)
Word choice: clarity (A2)	3 (27.3%)	5 (26.3%)
Word choice: richness vs. blandness (A6)	4 (36.4%)	3 (15.8%)
Sentence structure (A8)	0 (0.0%)	6 (31.6%)
Word choice: register & contextual fit (A5)	3 (27.3%)	2 (10.5%)
Cultural adaptation (A7)	2 (18.2%)	1 (5.3%)
Formatting & typography (A10)	0 (0.0%)	2 (10.5%)
Word choice: idiomaticity (A4)	1 (9.1%)	0 (0.0%)
B. Narrative-level features		
Narrative organization (B6)	2 (18.2%)	5 (26.3%)
Character voice & portrayal (B2)	3 (27.3%)	1 (5.3%)
Description & vividness (B3)	1 (9.1%)	2 (10.5%)
Emotional conveyance (B5)	2 (18.2%)	1 (5.3%)
Dialogue: naturalness (B1)	0 (0.0%)	2 (10.5%)
Figurative language (B4)	1 (9.1%)	0 (0.0%)

Category	MT	HT
C. Reader experience		
Smoothness & reading effort (C2)	5 (45.5%)	15 (78.9%)
Comprehension & ease of following (C1)	4 (36.4%)	8 (42.1%)
Engagement & immersion (C3)	5 (45.5%)	4 (21.1%)
Enjoyment & overall positive affect (C5)	5 (45.5%)	1 (5.3%)
Humanness & soul (C4)	0 (0.0%)	2 (10.5%)
D. Meta-translation		
Adaptation vs. literalness (D2)	1 (9.1%)	2 (10.5%)
Doesn't read as translated (D1)	0 (0.0%)	2 (10.5%)
Faithfulness to original (D3)	2 (18.2%)	0 (0.0%)
AI tell named (D4)	1 (9.1%)	1 (5.3%)

(a) Excerpt-level preferences

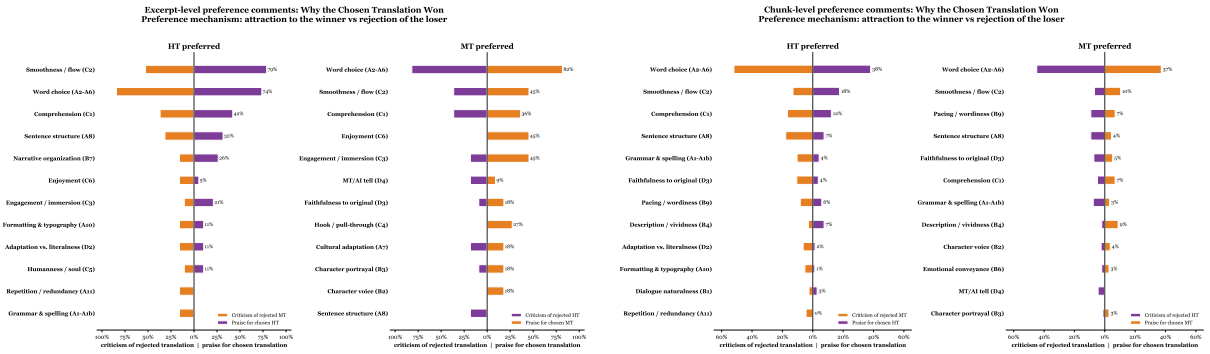
Reasons readers gave for preferring one translation over the other (chunk-level)

Category	MT	HT
A. Language-level features		
Word choice: naturalness (A3)	35 (14.0%)	80 (15.3%)
Word choice: clarity (A2)	16 (6.4%)	87 (16.7%)
Word choice: register & contextual fit (A5)	29 (11.6%)	26 (5.0%)
Sentence structure (A8)	11 (4.4%)	39 (7.5%)
Word choice: richness vs. blandness (A6)	23 (9.2%)	24 (4.6%)
Grammar & spelling (A1)	8 (3.2%)	22 (4.2%)
Word choice: idiomaticity (A4)	3 (1.2%)	15 (2.9%)
Cultural adaptation (A7)	9 (3.6%)	6 (1.1%)
Formatting & typography (A10)	3 (1.2%)	7 (1.3%)
Consistency (A9)	2 (0.8%)	4 (0.8%)
Repetition or redundancy (A11)	0 (0.0%)	1 (0.2%)
B. Narrative-level features		
Description & vividness (B3)	22 (8.8%)	39 (7.5%)
Pacing & information delivery (B8)	17 (6.8%)	31 (5.9%)
Narrative organization (B6)	6 (2.4%)	17 (3.3%)
Character voice & portrayal (B2)	16 (6.4%)	6 (1.1%)
Dialogue: naturalness (B1)	6 (2.4%)	15 (2.9%)
Emotional conveyance (B5)	7 (2.8%)	8 (1.5%)
Figurative language (B4)	7 (2.8%)	5 (1.0%)
Narration & POV effects (B7)	8 (3.2%)	4 (0.8%)

Category	MT	HT
C. Reader experience		
Smoothness & reading effort (C2)	26 (10.4%)	92 (17.6%)
Comprehension & ease of following (C1)	17 (6.8%)	64 (12.3%)
Enjoyment & overall positive affect (C5)	4 (1.6%)	14 (2.7%)
Engagement & immersion (C3)	3 (1.2%)	9 (1.7%)
Humanness & soul (C4)	2 (0.8%)	5 (1.0%)
D. Meta-translation		
Faithfulness to original (D3)	13 (5.2%)	19 (3.6%)
Adaptation vs. literalness (D2)	1 (0.4%)	8 (1.5%)
AI tell named (D4)	0 (0.0%)	4 (0.8%)
Doesn't read as translated (D1)	0 (0.0%)	3 (0.6%)

(b) Chunk-level preferences

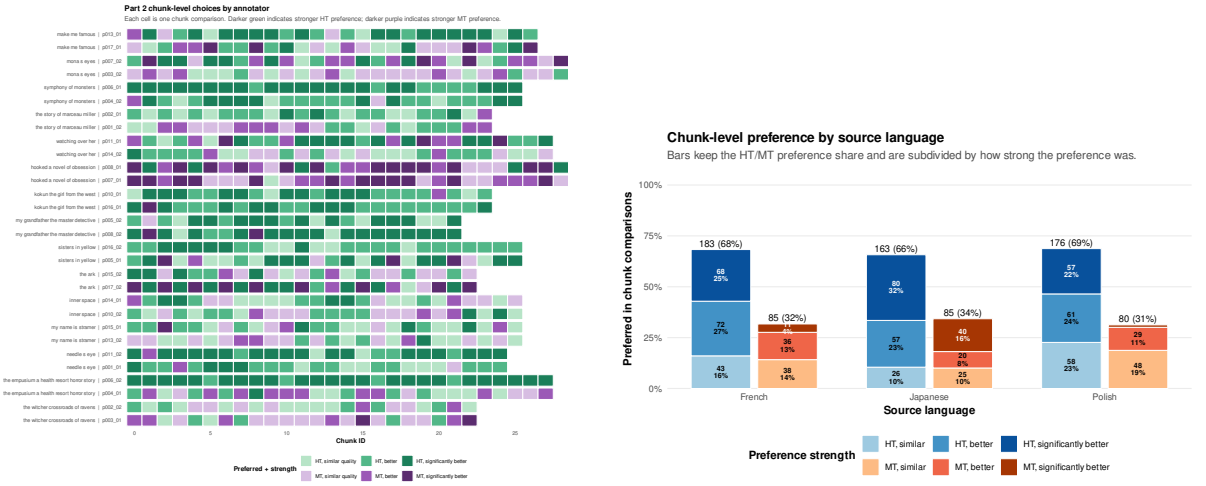
Figure 13: Positive reasons readers reported for preferring one translation over the other, separated by preferred system. Counts and percentages are calculated within each preferred-system group.



(a) Excerpt-level preferences

(b) Chunk-level preferences

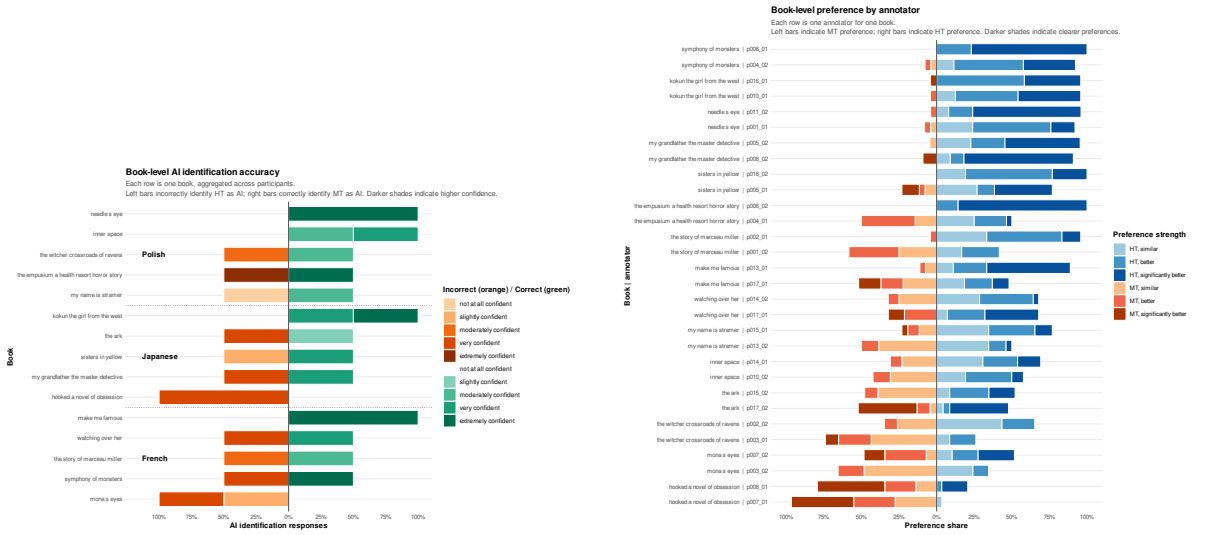
Figure 14: Preference mechanisms explaining why the chosen translation was preferred. Diverging bars contrast attraction to the winning translation with rejection of the losing translation, separately for HT-preferred and MT-preferred cases.



(a) Chunk-level preference matrix

(b) Preferences by source language

Figure 15: Close-reading preference outcomes. The heatmap summarizes participant choices across aligned chunks; the language panel aggregates preference direction and strength by source language.



(a) AI-identification judgments

(b) Translation preferences

Figure 16: Book-level human-evaluation outcomes. Panels compare translation-origin identification judgments with participant preferences for HT or MT, including preference strength.

Participant profile tables. The following tables summarize participant demographics, reading profiles, AI-use patterns, and MT-identification accuracy by AI-use frequency.

Panel A: Demographics and reading profile

ID	Age	G	Ed.	Nat.	L1	Other	Eng. lit.	Trans. lit.	Books/yr
1	25–34	W	UG	European/SE Asian	English	None	Weekly+	Almost never	100
2	25–34	W	UG	American	English	None	Weekly+	Few/yr	40
3	18–25	M	UG	South African	English	None	Few/mo	Not sure	12
4	45+	W	UG	South African	English	None	Monthly	Few/yr	12
5	45+	W	HS	American	English	None	Weekly+	Almost never	12
6	45+	W	MA	South African	English	None	Weekly+	Few/yr	50
7	35–44	W	PhD	Canadian	English	French	Monthly	Almost never	24
8	45+	W	UG	American	English	None	Weekly+	Few/mo	75
9	25–34	W	UG	South African	English	None	Weekly+	Few/yr	120
10	25–34	W	UG	American	Spanish	None	Weekly+	Few/yr	35
11	25–34	W	UG	American	English	None	Weekly+	Few/yr	100
12	45+	W	HS	American	English	None	Weekly+	Almost never	103
13	45+	W	UG	British	English	French	Few/mo	Monthly	20
14	35–44	W	MA	Polish	Polish	Polish; French	Weekly+	Weekly+	80
15	45+	W	PhD	American	English	None	Weekly+	Monthly	100

Panel B: AI model use

ID	Freq.	Models used most often	Reported uses
1	Sometimes	ChatGPT	Deep Research & Summarization
2	Rarely	Google	Does Not Use AI
3	Often	ChatGPT; Grok Imagine; Nano Banana	Deep Research & Summarization; Job Search & Career; Image Generation
4	Often	ChatGPT	Writing & Editing
5	Often	ChatGPT	Web Search; Personal Advice; Job Search & Career
6	Often	ChatGPT; Grok	Web Search; Image Generation
7	Often	Claude Opus; Claude Sonnet	Deep Research & Summarization; Writing & Editing
8	Often	Claude; ChatGPT; Grok	Deep Research & Summarization; Writing & Editing; Translation
9	Always	ChatGPT; Claude; Kiro	Deep Research & Summarization; Writing & Editing
10	Sometimes	Gemini	Deep Research & Summarization; Writing & Editing
11	Never	None	Does Not Use AI
12	Always	Claude Sonnet 4.6; ChatGPT 5.3	Deep Research & Summarization; Writing & Editing
13	Never	None	Does Not Use AI
14	Rarely	ChatGPT	Web Search
15	Always	ChatGPT; Claude; DeepSeek	Deep Research & Summarization; Writing & Editing

Table 40: Participant demographics, reading profile, and AI model use. Originally 17 participants were hired, but two did not provide sufficient details in comments, so results and demographics for 15 are presented.

Note. All participants reported that they primarily read in English and were very comfortable reading long literary texts in English. G = gender; W = woman; M = man. Ed. = education; UG = college/university undergraduate; HS = high school or equivalent; MA = master's degree; PhD = doctoral degree. L1 = native language. Eng. lit. = frequency of reading literary fiction in English; Trans. lit. = frequency of reading translated fiction in English. Weekly+ = weekly or more; Monthly = about once a month; Few/mo = a few times a month; Few/yr = a few times a year. Nationality entries and free-text AI-use responses are shortened for display.

AI-use frequency	Correct / total	Accuracy
Never	2 / 4	50%
Rarely	4 / 4	100%
Sometimes	2 / 4	50%
Often	3 / 12	25%
Always	6 / 6	100%

Table 41: Accuracy for identifying the MT version in co-Q5, grouped by participants' self-reported AI-use frequency. Participant-level association between AI-use frequency and MT-identification accuracy was close to zero (Pearson $r = 0.013$; Spearman $\rho = 0.121$).

Category	Code	Aspect	Definition	Pol.
A. Language-level features	**A1**	Grammar & spelling	Issues related to grammar such as verb-subject agreement, article use, or misspellings.	+/−
	A2	Word choice: clarity	Whether individual word and phrase choices are clear or keep the reader guessing when it was not inteded.	+/−
	A3	Word choice: naturalness	Whether the choice of specific words or phrases is natural in English at the local level. Covers words/phrases that are grammatically correct and even undertandable, but sound off (e.g., awkward) in English.	+/−
	A4	Word choice: idiomaticity	Correct use of English idioms, set phrases, conventional collocations, and figure-of-speech precision; the negative side covers both misused/garbled idioms and over-reliance on clichéd or trite expressions.	+/−
	A5	Word choice: register & contextual fit	Whether word choices match the context in aspects such as formality, period (e.g., historical vs contemporary), dialect/local speech, genre (e.g., sci-fi), speaker/situation (e.g., emotional context).	+/−
	A6	Word choice: richness vs. blandness	Lexical interest of the vocabulary used, such as unusual, varied, atmospheric words and turns of phrase vs. generic, common, sterile, "dumbed-down" choices. Independent of B8 (a passage can be rich-and-concise or rich-and-wordy).	+/−
	A7	Cultural adaptation	Translator's localization decisions in both directions: localizing/explicating foreign concepts for the target reader (e.g., glossing terms, Americanizing place names) vs. preserving foreignness (leaving names/terms in source language, retaining cultural markers).	+/−
	A8	Sentence structure	Sentence length, complexity, word order, run-ons, fragments, choppiness, and variety vs. monotony across sentences.	+/−
	A9	Consistency	Whether the translation remains the same choices across the text in terms of vocabulary, voice, register, etc. (e.g., whether character names are translated consistently).	+/−
	A10	Formatting & typography	The text's visual and typographic conventions, such as the typesetter-and-stylistic-choice layer that sits between grammar and content. It includes paragraph structure (breaks, dividers), dialogue formatting (quotation vs em-dashes).	+/−
	A11	Repetition or redundancy	Overuse of the same words, phrases, structures whether variation or compression would be better. This includes unnecessary lexical repetition, redundant content, etc.	−
	A12	Untranslated content	Source material (words or phrases) that appear untranslated in the target text and are perceived by the reader as translation failures not deliberate choices to retain the source word for culture preservation.	−
B. Narrative-level features	**B1**	Dialogue: naturalness	Whether the dialogue sounds like how people actually speak in the specific situation, context, emotional moment. It covers cases such as word plausability in speech (of this character), belivable phrasing etc.	+/−
	B2	Character voice & potrayal	Whether characters come through as distinct, identifiable, and well-matched to who they are. B2 evaluates the voice of characters and narrators: both in dialogue and in prose attributed to a character's perspective. character distinctness, attribution & character fit	+/−
	B3	Description / vividness / visualizability	Whether the reader can picture the scene and whether word choices conjure clear mental images even when both options are technically correct; includes the craft distinction of showing vs. telling (scene-rendering that creates imagery vs. bare informational delivery).	+/−
	B4	Figurative language	Use and effectiveness of metaphors, similes, imagery, and atmospheric or evocative phrasings; fires positively for evocative figures and negatively for misfiring ones that don't match the referent.	+/−

Continued on next page

Table 42 – continued from previous page

Category	Code	Aspect	Definition	Pol.
	B5	Emotional conveyance	Effectiveness of conveying the character's or scene's emotions to the reader (distinct from C4: B5 is about how well the reader can feel the story's emotional elements, while C4 is about whether the text itself feels alive or soulful).	$+/-$
	B6	Narrative organization	Story structure: scene transitions, paragraph / section flow, balance of exposition vs. action, story cohesion / motif recall (themes and descriptions that call back to previous text in the story).	$+/-$
	B7	Narration & POV effects	Narrator stance: interior monologue, free indirect discourse (blending character's thoughts into narration), narrator's distance from the action, narrator addressing the reader directly; point-of-view handling like POV shifts.	$+/-$
	B8	Pacing / information delivery / wordiness	Narrative-level pacing: info-dumps, lengthy descriptions that interrupt action; phrase-level wordiness (e.g. one word vs. whole phrase to describe the same thing). Independent of A6.	$+/-$
C. Reader experience	**C1**	Comprehension / ease of following	Whether the reader could understand what was happening: local sentence-level understanding and global scene/plot-level comprehension (these can fire independently).	$+/-$
	C2	Smoothness / reading effort	Whether the prose flowed or the reader kept getting stuck; i.e. how much friction was in the reading experience, separate from whether they understood it.	$+/-$
	C3	Engagement / immersion	Whether the reader was pulled into the story: being in the world, losing awareness of reading, focused attention on the scene.	$+/-$
	C4	Humanness / soul	Holistic judgement on whether the prose feels alive and human vs. robotic or mechanical (distinct from A3: individual word choices or sentences can feel natural, while the whole story can overall feel soulless).	$+/-$
	C5	Enjoyment / overall positive affect	General reader satisfaction with the text: "I liked it", "pleasant to read", "a pleasure", "I enjoyed both"; broader affective stance not related to a specific dimension, can be used as a catch-all for positive sentiment when no specific details given.	$+/-$
D. Meta-translation	**D1**	Doesn't read as translated	Holistic judgment on whether the prose reads as if originally written in English vs. feels like it's been translated.	$+/-$
	D2	Adaptation vs. literalness	Whether the translator adapted to English idioms and general audience expectations (positive) vs. translated word-for-word from the source language (negative); can fire with D4 when literalness is interpreted as a cue for MT.	$+/-$
	D3	Faithfulness to original	How well the text preserves the source meaning, voice, emotion, structure, subtle implications, connotations, and specific facts; usually perceived at the chunk-level, and often inferred by comparing small detail inconsistencies between two presented translations.	$+/-$
	D4	MT/AI tell named	Holistic reader feeling that the translation feels machine-generated or human, but without naming specific cues. Reader explicitly articulates the cue that triggers the MT/AI impression: e.g., em-dash overuse, bullet-point styling, "would" overuse, US English defaults, anthropomorphizing concepts, and other details that are percieved as tells for AI-generated text.	$+/-$

Table 42: Annotation scheme for participant comments on Single Q5, Q6, Compar. Q4, and Chunk Justif. Only positive codes are applied to comments for Single Q5, and only negative codes are applied for Single Q6. For Compar. Q4 and Chunk Justif., positive codes are applied to match the comments' positive aspects for the preferred translation, and negative codes applied to match the negative aspects of the non-preferred translation.

Code	Aspect	Definition (AI-pointing)	Human-pointing examples
M1	Literalism / source-bleed	Translation feels word-for-word or dictionary-like; source-language structure, idiom, or convention bleeds through; the method feels direct and unedited rather than adapted for English.	translated in a way that was corrected to make sense for English
M2	Word choice	A specific word or short phrase reads as off; wrong, stilted, archaic, jargon, bland, generic, or conversely vivid and colloquial. The reader points at the word itself.	in a frenzy' more accurate than 'with a fever'
M3	Sentence flow & connection	How sentences are shaped and how they connect. Choppy, truncated, run-on, abrupt, copy-paste structure, sentences that don't link. Not grammar (use M4).	more variation in sentence lengths
M4	Grammar / tense / agreement	Tense inconsistency, verb agreement errors, modal misuse, overuse of grammatical forms (perfect tense, conditional "would"). Grammar mechanics that aren't sentence shape (M3) or meaning (M10).	(rare — usually only flagged when wrong)
M5	Formatting & surface artifacts	Visible page-level mechanics not folk-theory framed: punctuation, scene dividers, quotation marks, dialogue formatting, encoding. Em-dashes here only when cited as plain observation; em-dashes as "AI sign" → M13.	slightly better formatting with proper nouns
M6	Wrong tone for the work	Whole passage is in a wrong tone for a novel, i.e., manual, textbook, technical, Wikipedia-mode; "not how a novel would read." Dominant mode is inappropriate.	older style consistent with fantasy
M7	Mixed / inconsistent registers	Within a single passage, register switches or mixes, e.g., US/UK slang together, source-culture and target-culture mixing, formal-informal swings. Internal inconsistency, not a dominant wrong key.	(rare)
M8	Wordiness / over-explaining	Text uses more words than needed, i.e., padding, extra words, info-dumping, lack of economy. Different from M6; wordiness can occur in any mode.	economical; tight prose
M9	Voice & emotional life	Whether prose is alive, e.g., engages, immerses, evokes mood, conveys emotion, gives characters distinct voices. Fires on flatness, mechanical feel, characters sounding alike, failure to evoke scene, inability to picture/feel. Immersion lives here.	made me feel like a sassy idol
M10	Meaning errors	Text gets meaning, logic, embodied physics, or factual content wrong, e.g., impossible images, lost nuance, dumbed-down content. Error is in the text, not the reader's parsing.	HT treats speaking underwater as hypothetical rather than given

Continued on next page

Table 43 – continued from previous page

Code	Aspect	Definition (AI-pointing)	Human-pointing examples
M11	Creative success	Translation makes a non-literal creative move the reader treats as a human signature, e.g., rhyme rendering, wordplay, creative password/name, right idiom for cultural concept. AI-pointing only by absence-by-contrast.	password 'Make_Me_Famous75!' — wouldn't think AI capable of that leap
M12	Comprehension difficulty	Reader couldn't follow, i.e., had to reread, needed the other version, confused by a specific passage. About reader experience, not text defect.	version of the sentence clearer in HT
M13	Folk theory invoked	Reader explicitly references what AI does, e.g., named stereotype, prior experience, capability claim. Triggers on language like "classic sign of AI," "AI tends to," "I'd expect from AI," "AI texts I've worked with," "AI doesn't [X]," "beyond what AI is capable of."	Human translators manage fluidity AI doesn't match

Table 43: Annotation scheme for participant comments on Compar. Q7 (justifying decision for which translation is perceived as machine translation).

Additional diagnostic figures. The figures below provide supplemental diagnostics for presentation order, origin-guess confidence, per-book ratings, close-reading preferences by book and language, and highlighted-span distributions.

Additional results Figure 20 shows readers preferences for word choices and dialogue in the *immersive* reading task. Table 48 shows distribution of top labels for readers comments.

Rank	POSITIVE label assigned to HT	Count	Example
1	C2. Smoothness / reading effort	14	"HT was much easier for me to read than MT. The wording flowed better and made it a seamless reading experience, whereas in MT I felt like I had to reread sentences now and then to make sure I was understanding right."
2	C1. Comprehension / ease of following	8	"HT was far easier to follow in terms of flow and word choice. MT had at times, bizarre punctuation choices and was very choppy. MT made it difficult to determine who was speaking."
3	A9. Cadence / rhythm / tempo	8	"HT is more cohesive, the text flow is very smooth and there is nothing to distract me from reading. MT has some poor wording choices, the sentences are sometimes awkward and that prevents me from getting immersed in the story."
4	A3. Word choice: naturalness	6	"The HT translation is simply more human. MT translation is overall proper with small errors, but there is no emotional impact on the reader. HT flows better, the sentences are more real for lack of a better word."
5	A8. Sentence structure	6	"The wording in both translations was good, but where HT surpassed MT was in sentence structure, avoiding too many long sentences, and formatting, dividers between scenes."
6	A2. Word choice: clarity	5	"From the start of the story, HT was a lot clearer to me and I think a lot of that has to do with punctuation and word choice. MT does not necessarily use obscure or uncommon words, but the translation does not feel quite right at times."
7	B7. Narrative organization	5	"HT flowed better. Text was broken up and summarized. The dialogue tags were more natural. I like the shorter sentences in the first translation."
8	C3. Engagement / immersion	4	"HT had no formatting issues that I noticed, for one, and also had a richness/uniqueness to it that pulled me in. The dialogue flowed well and, while complicated at times, is not unusual for an epic fantasy book of this nature."
9	A6. Word choice: richness vs. blandness	3	"The descriptions were more 3 dimensional in HT. They interacted with one another and recalled past phrasing. They drew pictures more vividly and expressed thought in a more cohesive manner. The flow had a more nuanced cadence and was less flat than MT."
10	C5. Humanness / soul	2	"HT stands out to me as clearly more human translated. Translations, but also literary fiction in particular, can get away with a bit of a weird/removed style of writing that made MT not seem strange to me, but HT was more natural."

Table 44: Top Q3 POSITIVE labels assigned to HT in the Part 1 comparison questionnaire. Counts are label occurrences.

Rank	POSITIVE label assigned to MT	Count	Example
1	C6. Enjoyment / overall positive affect	5	"I enjoyed both translations immensely and found them easy to read. I prefer MT's word choices because they come off as more intricate and varied to me. It is more formal than HT's phrasing in some parts, without feeling unnatural."
2	C2. Smoothness / reading effort	5	"MT has a better flow and word choice overall compared to HT, which had me rereading sentences to grasp their meaning more than once."
3	C3. Engagement / immersion	5	"The language in MT is much more natural than HT and is clearer to understand. MT pulls the reader in with more vivid descriptions and is generally well-written. HT was boring to read and I felt my mind wandering."
4	A6. Word choice: richness vs. blandness	4	"MT uses richer language with more unusual turns of phrase and wording. This creates a better atmosphere and more interest. HT uses regular language that seems somehow sterile and more common."
5	C1. Comprehension / ease of following	4	"MT made more of an effort to explain foreign words than HT, like 'the Pentateuch' in HT is translated as 'the Five Books of Moses' in MT. I have no idea if that is an accurate translation, but it was easier for me to understand."
6	A5. Word choice: register & contextual fit	3	"MT also made language choices, particularly the use of an older style of English, that are generally more consistent with this style of fantasy, e.g. [MT] 'Do not take offence' vs. [HT] 'Don't be cross'."
7	A3. Word choice: naturalness	3	"The language in MT is much more natural than HT and is clearer to understand. MT pulls the reader in with more vivid descriptions and is generally well-written."
8	A2. Word choice: clarity	3	"When reading HT the last sentence confused me initially, 'But what if he didn't when he could see it,' while in MT it seemed much clearer: 'But what did it matter that he did not know, when he saw.'"
9	C4. Hook / pull-through	3	"MT just felt more refined, despite the couple areas in the beginning with the news article that were a little tricky. I really became engrossed while reading it, I was upset when it stopped."
10	A9. Cadence / rhythm / tempo	3	"MT has a better flow and word choice overall compared to HT, which had me rereading sentences to grasp their meaning more than once. MT helps to drive the author's point and voice across with more common similes and imagery."

Table 45: Top Q3 POSITIVE labels assigned to MT in the Part 1 comparison questionnaire. Counts are label occurrences.

Rank	POSITIVE label assigned to HT	Count	Example
1	A2. Word choice: clarity	87	"HT is more succinct, more precise, and more idiomatic. MT is unnecessarily wordy, e.g., 'frightened her terribly' instead of 'terrified her.'"
2	A3. Word choice: naturalness	80	"HT has more natural phrasing while MT sounds more like a literal translation. Better word choice makes HT easier to read and makes the text flow better when reading."
3	C1. Comprehension / ease of following	64	"I found myself having to reread sentences in MT to follow what it was saying. It was just far easier to understand in HT."
4	C2. Smoothness / reading effort	64	"HT simplified a lot of overcomplicated sentences we see in MT, which makes text in HT flow much better. Sentences in MT are too long and overly complicated."
5	A8. Sentence structure	39	"HT had a better choice of splitting it up into smaller sentences which make it easier to read."
6	B4. Description / vividness / visualizability	39	"HT has a much clearer narrative picture and reads beautifully. The descriptive passages hit very well and it is just very enjoyable to read."
7	A9. Cadence / rhythm / tempo	36	"HT continues the fine narrative flow and maintains conversational stability, where MT feels a little clunky and formal."
8	B9. Pacing / information delivery / wordiness	31	"HT has a better narrative and conversational flow, as well as tighter descriptions and a better arrangement of the action."
9	A5. Word choice: register & contextual fit	26	"Some sentences in MT, while technically correct, are not how a person would say it in English. 'We'll be happy to comply,' from HT is more fitting."
10	A6. Word choice: richness vs. blandness	24	"HT has more unusual and interesting word choice, such as 'omnipresent water' instead of just 'all that water' and 'a sorrowful note appeared in his voice' rather than 'sadness entered his voice.'"

Table 46: Top Q5 POSITIVE labels assigned to HT in the chunk-level comparison task. Counts are label occurrences.

Rank	POSITIVE label assigned to MT	Count	Example
1	A3. Word choice: naturalness	35	"Overall I preferred the wording used in MT. HT felt slightly simple at times, e.g., 'being beaten with sticks' felt like over-explaining."
2	A5. Word choice: register & contextual fit	29	"I slightly preferred the more formal/older English style of MT but both translations seemed to sufficiently convey the message."
3	A6. Word choice: richness vs. blandness	23	"MT uses more interesting, unusual words like 'pellet of paper' instead of 'ball of paper' and 'loathsome' instead of 'hideous.'"
4	B4. Description / vividness / visualizability	22	"I only picked MT as my preference because the descriptions, like of their home, gave me a fuller picture of the scene."
5	A9. Cadence / rhythm / tempo	17	"MT is more fluid in style, making for a read that flows. The arrangement of the details feels more mechanical in HT, where MT reads properly."
6	C1. Comprehension / ease of following	17	"HT is all mixed up in the narrative and it is not as logically rolled out as in MT. There is a significantly better flow in MT and it makes it more understandable and easier to follow."
7	B9. Pacing / information delivery / wordiness	17	"MT has a tighter feel that comes across as more of a conversation than HT does."
8	A2. Word choice: clarity	16	"I thought HT was referring to a literal 'band' of boys because of the instrument talk, so I prefer MT because the word 'gang' was much clearer."
9	D3. Faithfulness to original	13	"MT has a marginally better flow and leans more faithfully on the Japanese, which results in a slightly awkward English reading experience, but it is not by any means bad."
10	C2. Smoothness / reading effort	12	"MT was a better translation as it was generally easier to read with a better natural flow."

Table 47: Top Q5 POSITIVE labels assigned to MT in the chunk-level comparison task. Counts are label occurrences.

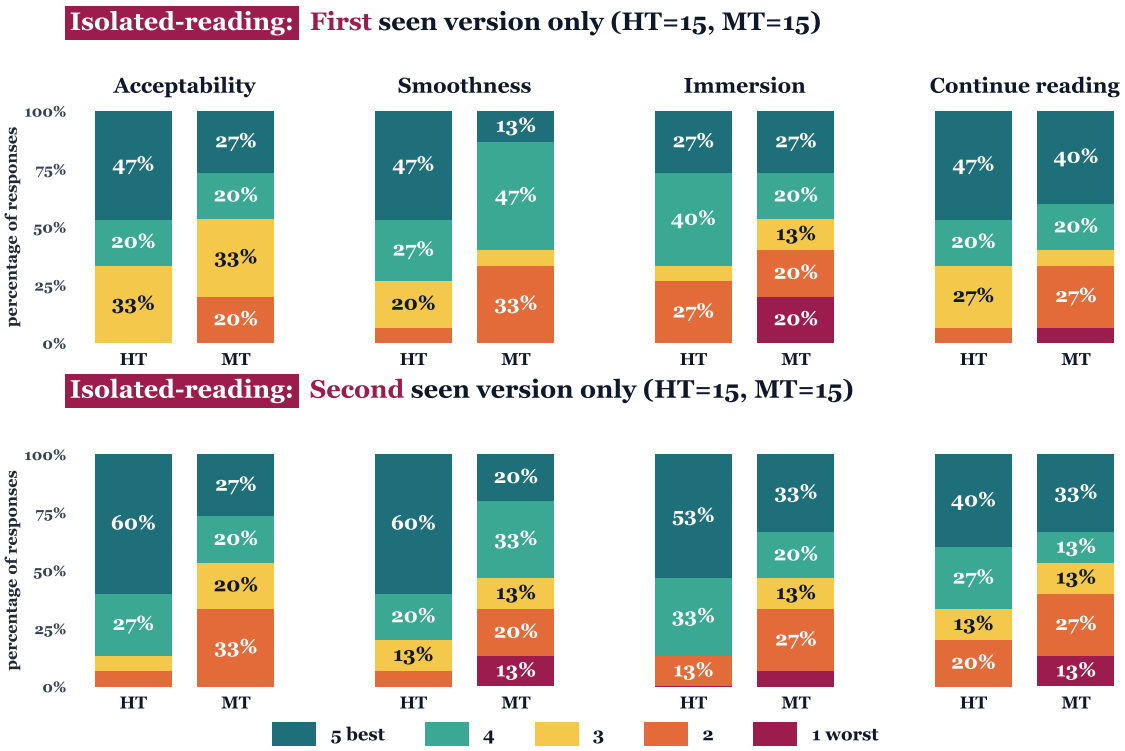
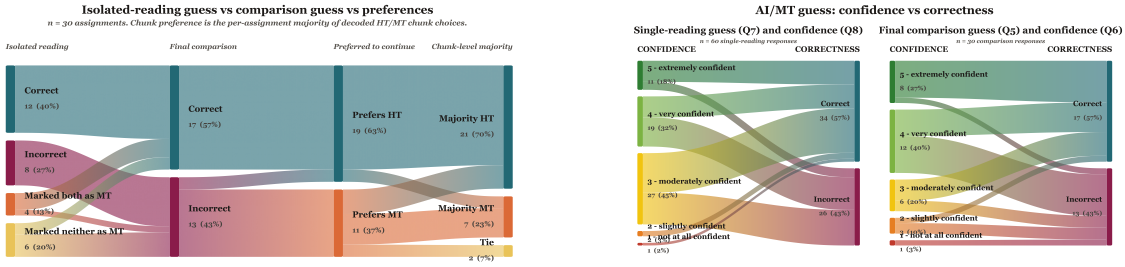


Figure 17: Immersive-reading ratings by presentation order. The figure compares ratings assigned after the first and second isolated readings, grouped by whether the human translation (HT) or machine translation (MT) was shown first.



(a) Guess correctness

(b) Guess confidence

Figure 18: Translation-origin guess flows after the comparison task. The first panel summarizes whether guesses were correct for HT and MT; the second panel shows reported confidence across those identification decisions.

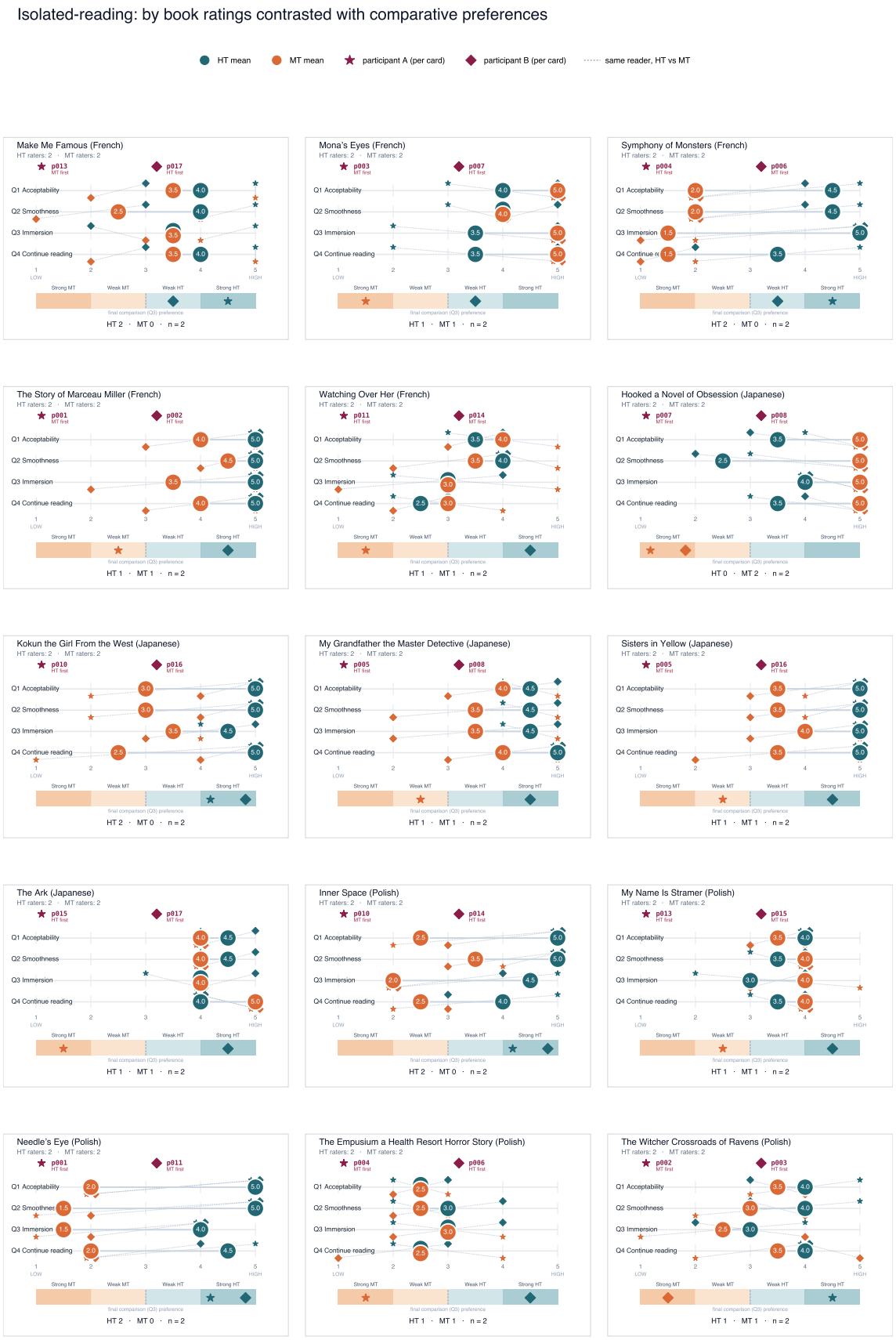


Figure 19: Per-book immersive-reading ratings for HT and MT. The figure breaks down participant ratings by book, making cross-book variation visible alongside the aggregate trends reported in the main paper.

Comment source	HT top labels	MT top labels
(a) Isolated reading (+)	Smoothness & reading effort (C2): 19/30 Comprehension & ease of following (C1): 12/30 Engagement & immersion (C3): 12/30 Description & vividness (B3): 9/30 Dialogue naturalness (B1): 7/30	Smoothness & reading effort (C2): 11/30 Comprehension & ease of following (C1): 9/30 Dialogue naturalness (B1): 7/30 Engagement & immersion (C3): 5/30 Word choice: naturalness (A3): 4/30
(b) Isolated reading (-)	Word choice: naturalness (A3): 5/30 Word choice: idiomaticity (A4): 5/30 Word choice: richness vs. blandness (A6): 5/30 Grammar & spelling (A1): 4/30 Word choice: register & contextual fit (A5): 4/30	Sentence structure (A8): 15/30 Comprehension & ease of following (C1): 11/30 Smoothness & reading effort (C2): 11/30 Grammar & spelling (A1): 8/30 Word choice: naturalness (A3): 8/30
(c) Excerpt preference (+)	Smoothness & reading effort (C2): 15/19 Comprehension & ease of following (C1): 8/19 Word choice: naturalness (A3): 6/19 Sentence structure (A8): 6/19 Word choice: clarity (A2): 5/19	Enjoyment & positive affect (C5): 5/11 Smoothness & reading effort (C2): 5/11 Engagement & immersion (C3): 5/11 Word choice: richness vs. blandness (A6): 4/11 Comprehension & ease of following (C1): 4/11
(d) Chunk preference (+)	Smoothness & reading effort (C2): 92/522 Word choice: clarity (A2): 87/522 Word choice: naturalness (A3): 80/522 Comprehension & ease of following (C1): 64/522 Sentence structure (A8): 39/522	Word choice: naturalness (A3): 35/250 Word choice: register & contextual fit (A5): 29/250 Smoothness & reading effort (C2): 26/250 Word choice: richness vs. blandness (A6): 23/250 Description & vividness (B3): 22/250

Table 48: The most frequent labels in readers' comments about the **human translation (HT)** and the **machine translation (MT)**. Rows separate isolated-reading praise, isolated-reading criticism, excerpt-level preference reasons, and chunk-level preference reasons. Comments could receive multiple labels; the full annotation scheme is in Table 42.

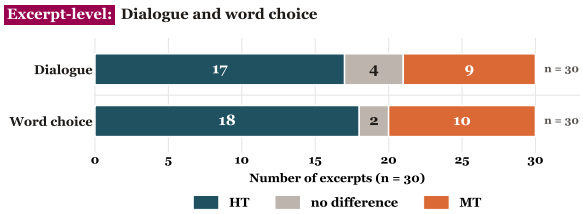


Figure 20: The distribution of readers' excerpt-level preferences for **dialogue handling** and **word choice** after reading both the **human translation** (**HT**) and the **machine translation** (**MT**). The readers were judging which version handled dialogue better overall and which version had more varied or expressive word choice.

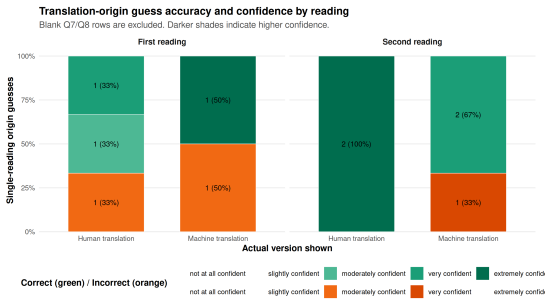


Figure 21: Translation-origin identification accuracy and reported confidence across the first and second readings. Bars are grouped by the actual version shown, human translation (HT) or machine translation (MT). Green segments indicate correct origin guesses, whereas orange segments indicate incorrect guesses; darker shades represent higher confidence. Blank responses to the relevant identification and confidence questions are excluded.

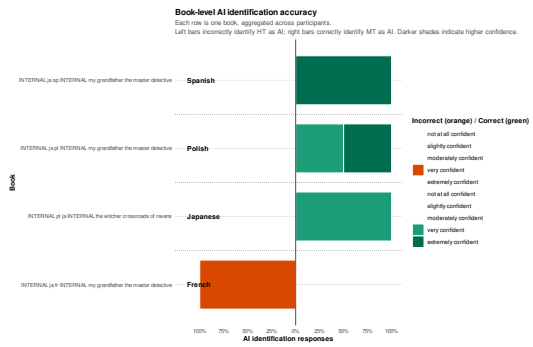


Figure 22: Book-level AI-identification results, aggregated across participants. For each book, the left-hand bars show cases in which the human translation was incorrectly identified as AI-generated, while the right-hand bars show cases in which the machine translation was correctly identified as AI-generated. Green denotes correct identification and orange denotes incorrect identification, with darker shades corresponding to greater confidence.

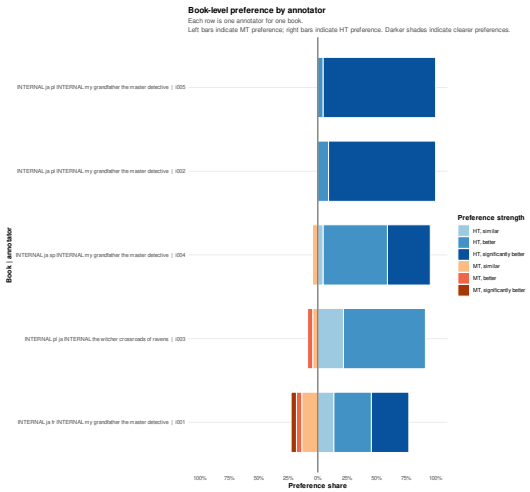


Figure 23: Overall translation preference by book and annotator. Each row represents one annotator evaluating one book. Bars extending to the left indicate a preference for the machine translation (MT), while bars extending to the right indicate a preference for the human translation (HT). The shading distinguishes similar quality, better quality, and significantly better quality judgments, with darker colors representing stronger preferences.

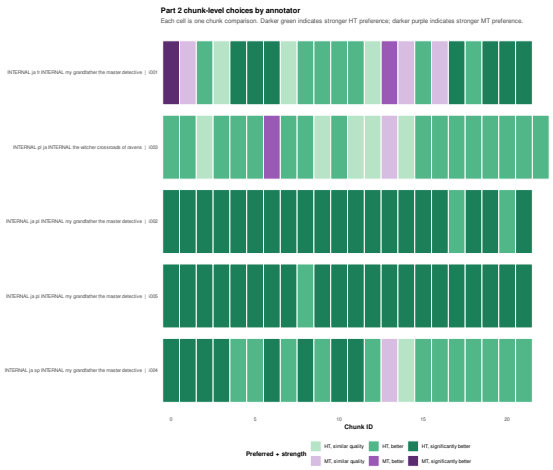


Figure 24: Chunk-level translation preferences in Part 2, shown separately for each annotator and book. Each cell corresponds to one chunk comparison. Green cells indicate a preference for the human translation (HT), whereas purple cells indicate a preference for the machine translation (MT). Darker shades represent stronger preference judgments, allowing both consistent and locally varying preferences across chunks to be identified.

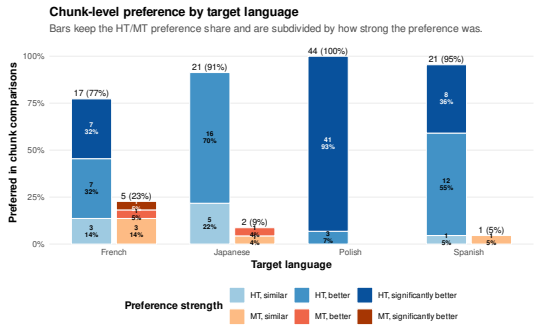


Figure 25: Chunk-level translation preference grouped by source language. For each source language, the bars summarize the proportion of chunk comparisons favoring the human translation (HT) or machine translation (MT). Each bar is subdivided by preference strength, distinguishing judgments of similar quality, better quality, and significantly better quality. The figure therefore shows both the direction and intensity of translation preferences across languages.

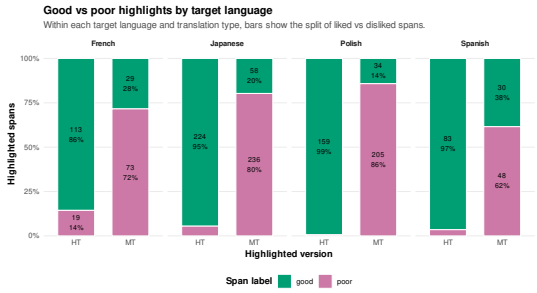


Figure 26: Distribution of positively and negatively evaluated highlighted spans by source language and translation type. Within each source language, separate bars are shown for highlighted segments from the human translation (HT) and the machine translation (MT). Green segments correspond to spans judged as good, while purple segments correspond to spans judged as poor. The annotations report the number and percentage of spans in each category.

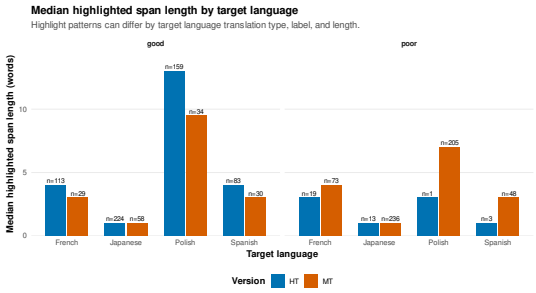


Figure 27: Median length of highlighted spans, measured in words, by source language, translation type, and evaluation label. The figure compares spans highlighted as good or poor in the human translation (HT) and machine translation (MT) outputs. Sample sizes are displayed above the corresponding values, making it possible to interpret differences in highlighted-span length alongside the amount of available annotation evidence.

D.4 Post Questionnaire Answers

The tables below give the post-study question key and complete participant responses.

Question	Wording
Q1	What were your expectations for the quality of AI translation before starting this experiment?
Q2	You <correctly/incorrectly> identified the AI-translated version of Book 1, and <correctly/incorrectly> identified the AI-translated version of Book 2. Are you surprised by your accuracy?
Q3	What do you think about the quality of AI translation now, after reading the translation?
Q4	Any other comments?

Table 49: Question key for the post-study responses.

Q	Response
P1	**Correctly identified Book 1?**: yes **Correctly identified Book 2?**: no **Expected worse**: yes
Q1	I expected the AI translations to be more obviously written by AI. I was impressed even with the translations I didn't prefer.
Q2	I'm not that surprised actually, book 2 was by far harder to pick a favorite of. I had a really hard time even finding flaws, and when I did I'd find them more on a preference level and then both translations had some.
Q3	I think it's so good, hence my mistaking it as human. I think it read really smoothly and if it was one of my books, I would have no inkling of it being translated by AI
Q4	no other comments come to mind! Thanks for having it, it was so fun!
P2	**Correctly identified Book 1?**: yes **Correctly identified Book 2?**: yes **Expected worse**: yes
Q1	My expectations for the quality of AI translation before starting this experiment were that they would be a little confusing to read because they would be directly translated, which sometimes does not mean translated correctly. It was my impression that usually with different languages, the English translations were not exactly correct. So I figured there would be some awkward or incorrect wording.
Q2	I am not surprised by my accuracy in correctly identifying the AI translations of both Book 1 and Book 2.
Q3	My opinion of AI and its quality is the same as it was before, which is that it can be helpful but at the end of the day only a real human can get something to the way it needs to be. I don't think AI has the ability to understand when something doesn't sound right or use context to correct something. A real person, on the other hand, can take something translated poorly and rework it in order for it to make sense.
Q4	N/A
P3	**Correctly identified Book 1?**: no **Correctly identified Book 2?**: no **Expected worse**: yes
Q1	I expected the writing quality to be a lot worse. I thought that the resulting text would be more in line with that of AI-written text (not translated, entirely AI-generated from prompts).
Q2	Not particularly. I leaned towards one being AI over the other mainly because of quality differences. I do still believe that with a few copies of each, humans could learn to pick up the patterns and correctly detect which is AI-generated.
Q3	It's substantially better than I expected. I definitely feel that it's one of the best applications of LLMs that I've seen.
Q4	The quality of the human translations were quite a bit worse than I expected. It's almost like they didn't use an editor to review it afterwards. There were pretty clear issues that shouldn't have slipped through. I think that a version that's objectively higher quality than both could be made by combining the better parts of each translation.
P4	**Correctly identified Book 1?**: no **Correctly identified Book 2?**: no **Expected worse**: yes
Q1	My expectations of the AI translations were low, based on the poor quality of many of the AI-generated books I edit and humanize.
Q2	I am somewhat surprised because I was fairly sure. However, the quality was much better than what I'm used to, and the books didn't have many of the usual "tells" I often see.
Q3	I think it has improved remarkably in a very short time.
Q4	N/A

Continued on next page

Continued from previous page

Q	Response
P5	**Correctly identified Book 1?**: no **Correctly identified Book 2?**: no **Expected worse**: no
Q1	I expected them to be good. I use ChatGPT multiple times a day, for many reasons. A lot of times it's to get answers for random things that pop in my head that I'm curious about, (that happens quite a bit, maybe more than most people, my brain is funny like that). But I also have been using it a ton the past few months, daily, with my resume as I'm job hunting. I've seen AI give amazing responses, but I've also seen it be incorrect and inconsistent. So while my expectations were high, I had a constant hesitation wondering how much faith I wanted to put into the machine. If that makes sense.
Q2	I'm not surprised now that I've completed them. For both books, when it came to Part 2 and I was annotating them and comparing them, I knew I'd made the wrong choices. Reading them one at a time and a day apart made a huge, huge difference and that surprised me.
Q3	I think it's inconsistent, for starters. I don't think it's safe to hand projects over to and be confident that you'll get a flawless result. I've often wondered about how a machine can replace a human in instances like this, when tone and even the smallest nuances, make the biggest differences. They're also so literal with the words, they can't grasp the context. I think AI translations make for a great rough draft. Maybe that's the best way to put it.
Q4	I'd be curious to know if my annotations, the translations that I critiqued and disliked the most, were AI. When I got to that point is where I feel I realized what was what. Or maybe I just didn't care for how a person translated them, and was critiquing them! Oh gosh.. haha!
P6	**Correctly identified Book 1?**: yes **Correctly identified Book 2?**: yes **Expected worse**: no
Q1	I expected AI to produce a competent literal translation but to slip up with idiomatic language and polysemous words, and its translation of scenes that are supposed to evoke emotion.
Q2	Not really surprised (but certainly gratified) because translating fiction is tricky even for humans. I think it was with Henning Mankell that I first started paying attention to the names of translators when one of the books in his Wallander series sounded "off". And he used top-notch human translators!
Q3	On the whole, I assume that AI accurately reproduced the text from the original to English but it tends to flatten the author's voice. That is not a problem when all the reader wants is a functional translation of information. But that is not why we read fiction.
Q4	An AI translation is still a lot like CliffsNotes: function without form. I doubt "War and Peace" will get an AI makeover any time soon! But I shouldn't be too cock-a-hoop. AI learns 24/7 and I will probably have to buy another hat to eat in the not too distant future.
P7	**Correctly identified Book 1?**: no **Correctly identified Book 2?**: no **Expected worse**: yes
Q1	My expectations were pretty low. I was paid to "clean up" a machine translation of a non-fiction book about the outbreak of war in Ukraine in 2022 for Ibidem press, and it was pretty clunky. The idioms were very off. This was much better.
Q2	I am only a little surprised by my accuracy, now that both tasks are finished.
Q3	I think that rather than demonstrating the quality of AI translation, this project demonstrated to me how challenging it is to judge the quality of any translation of a piece of literature from the opening pages. The Japanese novel really seemed "off" to me in how it was depicting the protagonist – she seemed much too squirrely to be a successful businesswoman. I attributed this to an AI-translator, but having looked up the book, I now know she was supposed to seem squirrelly because she is a stalker on the verge of giving in to obsession. After reading those texts, I can definitely see the appeal of working with an AI-translator rather than a human-translator for an author looking to truly preserve their voice in a piece of fiction. The two tasks have not impacted how i feel about AI writing at all.
Q4	This project inspired me to dust off the now 15-year-old translation of a Renaissance Latin text in the style of Sallust that was a core component of my doctoral project and which, because the text seems just so unappealing to English readers, I had not ever planned on returning to... and the LLMs' analysis and discussion with me of the text confirmed that the text just isn't translateable today ... when no one wants to read Gibbons-esque prose. Thank you for this opportunity. It was a fun task, and I am enjoying reading Hooked so much that I have sent copies to friends.
P8	**Correctly identified Book 1?**: yes **Correctly identified Book 2?**: yes **Expected worse**: yes
Q1	I had a realistic expectation of accurate translations in terms of word choice, but I wasn't sure about the sentence construction on a grammar level. I wasn't expecting a 1v1 direct translation, because that doesn't work from Japanese to English, but I thought there would be more… mechanical issues.

Continued on next page

Continued from previous page

Q	Response
Q2	I am and I'm not. I think it comes down to the nature of the translation. Book 2 was much easier to detect AI on because the human-led translation was so tightly done on a narrative level. It flowed very well, with some variance in terms of paragraph structures that I attributed to the nature of the original language. Book One was hard to figure out, but I chose the version I did because of the work I currently do. Most of my authors use AI to wholly or in part generate a manuscript which they submit to me for humanizing. These manuscripts have certain linguistic conventions (truncated sentences, over-reliance on mechanical or stage direction language, certain buzzwords to describe human emotions: the architecture of a face, etc.) that made me view the translations through that lens. I'm trained to spot short, choppy sentences and odd word choices as being AI-generated.
Q3	I'm pleasantly surprised at the quality. It's not perfect, but it's definitely not going to stop me from reading.
Q4	I think the style of the human translation played a large part in how I viewed the AI version. The translator who chose a more Western style felt more natural to me, and that made it easier to pick out an AI version. Book One's translator felt more faithful to the tone and style of the Japanese original, which read oddly to my English eyes. This was an honestly fascinating look at a side of my industry that I didn't have before. I greatly appreciated this chance to be part of the project, and it's solidified my desire to write down my observations on how AI handles fiction.
P9	**Correctly identified Book 1?**: yes **Correctly identified Book 2?**: yes **Expected worse**: yes
Q1	I expected the translation to be a lot worse than it actually was (almost like how Google Translate translates a sentence - usually in a way no English speaker would say it).
Q2	I'm not surprised by the accuracy - I wasn't completely confident, of course, but I didn't expect a human to miss the obvious issues in the AI translations (dialogue issues and scene break/double hyphen issues).
Q3	The issues in the AI translations were more to do with style than actual major wording problems, which I think is amazing.
Q4	I think the books were translated very well in all versions, a few easily-fixable issues aside. If AI could do the brunt work in translations, then a human can just review it and fix those small issues (or train the model to fix it itself in the future!)
P10	**Correctly identified Book 1?**: no **Correctly identified Book 2?**: yes **Expected worse**: yes
Q1	My expectations for AI translation were pretty low coming into this experiment. I have seen some pretty bad translations in the past.
Q2	I'm surprised! But I do think Book 1's translation was a lot better than Book 2's. It was easier in Book 2 to differentiate between the two because of the amount of dialogue, in my opinion, and the way things were phrased.
Q3	I think there is some merit to exploring AI-assisted translations, but I think they need human revisions to streamline the overall translation.
Q4	Thanks for giving me the opportunity to participate in the study!
P11	**Correctly identified Book 1?**: yes **Correctly identified Book 2?**: no **Expected worse**: yes
Q1	I expected the AI translation to be obvious. I read a lot of AI generated work and it has very specific tells and often devolves into nonsense.
Q2	I am not surprised. Book 1 felt very obvious when comparing the two translations. Book 2 was much harder. I'm not even sure which one I chose because the translations were of similar quality in my opinion (both felt kind of off).
Q3	I was surprised by the quality of the translation in Book 1, it was much better than expected, but the human-generated translation was even better and provided context I don't think an LLM is capable of. I can see a good use case for it. I struggled a little with both translations of Book2 so it's hard to draw any conclusions there.
Q4	As a lover of books and languages I really enjoyed doing this! Thank you for the experience. Oh! And I'll also add that if I had only read T1 from book 1, I would not have guessed it was AI translated. It was only obvious by comparison. I was very suspicious of there being a double blind on the second book *[face with tears of joy emoji]*
P12	**Correctly identified Book 1?**: yes **Correctly identified Book 2?**: yes **Expected worse**: yes
Q1	I think I was expecting less than what I read. As in, I thought it would be easier to detect the AI.
Q2	No, I'm not surprised. I do a LOT of work with AI and there are some tells that give it away. There is also the conversational tone that AI hasn't quite nailed down.

Continued on next page

Continued from previous page

Q	Response
Q3	I think the first book, the one about the spaceship, was much better. I don't know if it's the original author writing the book in a more conversational tone, or if the AI chose better words. The Michealangelo book was extremely difficult to read. Both version were challenging, which tells me that's the way the author wrote it.
Q4	N/A
P13	**Correctly identified Book 1?**: yes **Correctly identified Book 2?**: no **Expected worse**: yes
Q1	I expected the quality of AI translation to be stilted and mechanical with the words not always in the right order.
Q2	Yes, I'm surprised by incorrectly identifying the AI-translation version of Book 2.
Q3	I still feel that the AI translation lacks a human aspect, with less vivid descriptions.
Q4	The human translation can vary depending on the experience/quality of the translator. A human translator who is familiar with a particular place or culture can bring their own human experience to the translation.
P14	**Correctly identified Book 1?**: **Correctly identified Book 2?**: **Expected worse**: yes
Q1	I was expecting the quality to be somewhat worse than what I read in the provided texts. More missing context due to language/cultural differences, more literal translations without replacing idioms, etc.
Q2	I thought it would be easier to identify the AI translation than it was in reality. When I had two texts side by side it was quite easy, but when I read them separately, it was quite similar to poorly done translation by a human amateur.
Q3	Both AI translations in the text were more or less acceptable from the point of view of literal translation. However, they were missing emotional import, poetic sense and failed to take into consideration that some things should be changed in translation based on the cultural and language differences in the output language.
Q4	I think AI can be a useful tool if we are stuck and are trying to see how something can be said or written, but it should only be treated as a tool or a guide, there is no way that AI can properly convey everything - emotion, language style, etc, and take into consideration many, many things around the language that we do not think of but notice the difference when reading the text.
P15	**Correctly identified Book 1?**: yes **Correctly identified Book 2?**: yes **Expected worse**: no
Q1	I expected a grammatically correct translation of AI. It has become much better in the past few months and I anticipated a robust and accurate interpretation of the narrative. In the first book, one translation was significantly better than the other. In the second book, I was disappointed by the grammar and use of words. There was lack of depth and engagement.
Q2	I am not surprised by my accuracy. I read a lot of AI and work with it every day. So I can tell by grammar, use of colons and em dashes, and choice of words pretty clearly.
Q3	I think the translations need work. The first book was superior to the second book. The translations were poorly organized and lacked compelling narrative.
Q4	It would be fairly easy to fix the translations so they don't sound so AI driven. HUmanizing does not take that much time. Please feel free to reach out if you need help with anything else.

Table 50: Post-study responses grouped by participant.

E Comparison of Machine Translation Systems

MT systems converge on similar renderings. Although the agentic pipeline performed best in our internal evaluation, we also compare how much other MT systems varies for the same excerpts. For this analysis, we compare all available MT outputs: the agentic pipeline and two LLM-based pipelines with nine models, giving 19 MT systems in total. We compare these systems with the professional HT across the full corpus, including both the evaluation and development sets (40 excerpts). All outputs are aligned at the paragraph level using Par3.

Although all outputs are translated from the same source excerpts, the MT systems are much more similar to one another than to HT. In embedding space, MT outputs form a cluster regardless of model family or pipeline type, while HT lies outside this cluster (Figure 7). Average MT-MT cosine similarity is 0.8854, compared with 0.8021 for MT-HT. We observe the same convergence when we remove function words, lemmatize the remaining content words, and compare ordered n-grams. For combined 3-5-grams, MT-MT Jaccard overlap is 0.168, compared with 0.068 for MT-HT (Figure 8). Even for literary source texts that allow multiple valid translations, MT systems tend to converge on similar English renderings, whereas professional HT appears to make more varied translation choices.

Can the difference between MT and HT be captured with automatic metrics? We compare our chunk-level human judgments with automatic metrics. Since some metrics cannot handle long texts we use the `par3` aligner (Thai et al., 2022) to align paragraphs from source text to the HT and MT, and then map them to the chunks before computing chunk-level scores.[23] We run METRICX-QE (Juraska et al., 2024), COMETKIWI (Rei et al., 2022), and LITRANSPROQA (Zhang et al., 2025b) with Gemini 3.1 Pro (Google DeepMind, 2026). Overall, we observe all metrics preferring MT over HT on the system level (LITRANSPROQA↑ MT: 0.996, HT: 0.920; COMETKIWI↑ MT: 0.760, HT: 0.722; METRICX-QE mean↓ MT: 5.184, HT: 5.671). We further compute Kendall τ (Kendall, 1938) correlation between each metric and human judgments on chunk-level. We compare for : (a) the two human judgments (772), and (b) whenever they agree on the same judgments (386), where ties are excluded. The results show week to negative correlation for all metrics : LITRANSPROQA ($\tau = -0.124$), COMETKIWI ($\tau = -0.318$), and METRICX-QE ($\tau = -0.216$). Both (a) and (b) give the same τ because the preference pattern is unchanged: clear human agreements tend to favor HT, whereas the metrics often favor MT or fail to match the human preference. Thus, automatic metrics fail to recover reader preferences under either aggregation, consistent with previous studies (Thai et al., 2022; Zhang et al., 2025b,a).

[23] As one chunk would consist of multiple paragraphs we compute the mean scores for all paragraphs within a chunk to obtain the chunk-level score.