

AI translation of literary texts is ‘fine’, but readers still prefer human translations

Yves Ferstler^A, Adam Podoxin^E, Ty Brassington^E, Gaëlle Laperrière^A,
Roman Grundkiewicz^E, Marie-Jean Meurs^A, Maite Taboada^E, Marzena Karpinska^E

^ESimon Fraser University ^AUniversité du Québec à Montréal ^EMicrosoft

Correspondence: ferstler.yves@courrier.uqam.ca, karpinsk@sfu.ca

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 poorly captured by automatic metrics or human evaluation targeting fluency and adequacy. We ask 15 avid readers to compare recently published human translations (HT) to machine translations (MT) generated with an agentic language model-based pipeline, for 15 recent novels in French, Polish, and Japanese translated into English. Readers evaluated ~8K-word excerpts in two conditions: *immersive reading* of the whole excerpt (30 comparisons) and *close reading* of 386 aligned HT–MT chunk pairs (772 comparisons), with two readers per book. Overall, readers find MT “fine”, but prefer HT (descriptively at the excerpt level, 19/30, and significantly at the chunk level, 522/772) for its ease, clarity, and immersive nature. Readers’ highlights show that MT’s quality varies more within one book than HT’s does. Crucially, readers cannot reliably tell the two apart (17/30 guess correctly) and tend to prefer the version they believe to be human. Automatic metrics, including LLM-as-a-judge approaches, fail to recover reader preferences and favor MT. 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.

 github.com/Yves575/lait

 <http://lait.cs.sfu.ca/>

1 Introduction

Machine-translated literature is reaching real readership.¹ For instance, Amazon introduced Kin-

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

dle Translate for self-publishing authors (Amazon Staff, 2025), while services such as GlobeScribe market AI translation to smaller publishers (GlobeScribe.AI Ltd). Similarly, Veen Bosch & Keuning, the largest Dutch trade publisher, announced AI-assisted translation (Creamer, 2024), and Japan’s Shogakukan released an app offering AI-assisted translations of light novels to US and Canadian readers (Nishioka, 2024).

This raises an interesting question. We know that readers sometimes even prefer fine-tuned AI fiction to expert human writing, at least on short excerpts (Chakrabarty et al., 2025). So how does *AI translation* shape the reading experience and the perception of the author’s work compared to a professional human translation, especially once readers stay with the text for 25–30 pages? Readers respond to rhythm, emotions, voice, and deliberate word choices (Brems and Ramos Pinto, 2013; Walker, 2021), but this experience cannot be captured by current evaluation setups which often concentrate on 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 machine 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 (~30 book pages) translated from Japanese, Polish, and French into English² by professional literary translators (HT) 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**

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

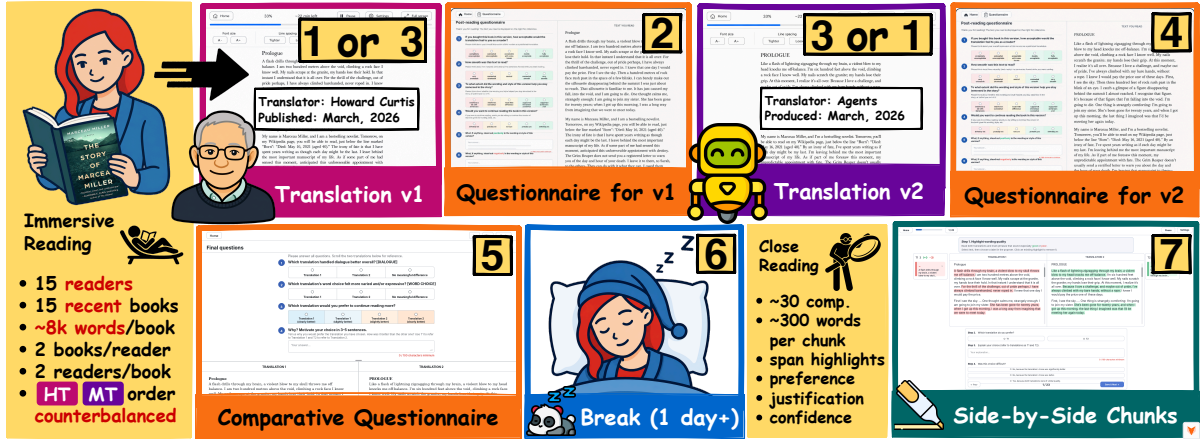


Figure 1: **Evaluation pipeline:** Avid readers of published fiction evaluate a **human translation (HT)** and an **AI-generated machine translation (MT)** of an 8,000-word book excerpt. Participants (1) read the first translation, (2) complete a questionnaire, (3–4) repeat both steps for the competing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

or **poor** before indicating which version they prefer (*close reading*). Crucially, besides ratings, preference choices, and chunk-level highlights, we also ask the readers to provide detailed comments and guess which text is MT (Figure 1).

Overall, readers show a slight preference for HT, which is significant under *close reading* (522/772) but only descriptive under *immersive reading* (19/30). However, readers do not completely reject MT as they often find it readable and sometimes preferred. Notably, chunk-level span annotations show that MT quality varies much more within a single book than HT, and its overall success depends more on the book than the source language. Reader comments suggest that HT is praised more for immersion, naturalness, and literary rendering (e.g., “*having soul*” (HT) vs. “*translation for dummies*” (MT)). MT sometimes wins due to readability or local word choices.

Readers also struggle to identify MT reliably. Their accuracy remains close to chance 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 swear words. Finally, automatic metrics, including LLM-as-a-judge approaches, fail to recover reader preferences and *all* favor MTs.

To summarize, our contributions are threefold:

■ **DATA:** We introduce 📖 **LAIT**, a reader-annotated dataset of 15 FR/PL/JA→EN novel

openings (2025–2026), each aligned across source, published HT, and agentic MT at the excerpt, chunk, and paragraph levels. Two readers per book gave 2K ratings and preferences, 1K comments, and 7.2K span-level annotations; an unannotated 16-book development set adds 1.7K paragraph-level alignments across five candidate MT systems.

🔍 **ANALYSIS:** We show that modern literary MT into English can be readable and hard to detect, but its quality varies within a book more than HT, and readers still value HT.

📖 **METHODOLOGY:** We introduce a reader-centered evaluation protocol combining *immersive* and *close reading*, and release our guidelines and annotation software to support reproducibility and future work.

2 Data & Methods

2.1 Constructing 📖 LAIT

Book selection. We selected 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; Kocigit et al., 2025). Overall, we obtained **15 books** for human evaluation,³ with five books per source language (see Table 1), and reserved **16**

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

	HT			MT			SRC
	TOKENS	WORDS	SENTS	TOKENS	WORDS	SENTS	TOKENS
■ Excerpts	(n=15)						
MEAN	9,680	7,623	547	9,582	7,676	539	13,307
ST. DEV.	535	449	74	626	529	73	2,316
MAX	10,261	7,995	668	10,500	8,595	683	19,116
MIN	8,600	6,747	440	8,474	6,928	447	9,882
■ Chunks	(n=386)						
MEAN	376	296	21	372	298	21	442
ST. DEV.	66	49	8	73	57	8	130
MAX	611	486	51	644	530	50	870
MIN	118	100	4	121	102	4	79

Table 1: Summary statistics for the evaluation dataset. Words are whitespace-delimited counts. Tokens are computed with tiktoken (o200k_base). Sentence counts for HT and MT use spaCy (en_core_web_trf).

books for selecting the MT pipeline, for a total of **31 books** (see §A for the full book list).⁴

Source and target languages. We selected French, Polish, and Japanese as source languages because they span different language families, cultures, and literary traditions. We chose 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; Zhu et al., 2024). Furthermore, the researchers’ familiarity with all the languages facilitated data collection and analysis.

Excerpt extraction. For each book, we selected an opening excerpt from the English human translation (HT) of ~8K words ($\mu = 7,623$, $SD = 449$) under whitespace tokenization while preserving paragraph boundaries.⁵ This length supported *immersive* reading while keeping the task feasible for readers’ evaluation. We then manually identified the corresponding source excerpt to obtain source-target mapping with HT. Preprocessing details are provided in §A.

MT pipeline selection. To select a strong MT pipeline for the evaluation, we compared five configurations on a 16-book development set: two prompting pipelines, P1 full-excerpt translation with post-editing and P2 chunk-level translation with excerpt-level post-editing, each using either GPT-5.4 (OpenAI, 2026) or Gemini 3.1 Pro

⁴All source books and translations 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.

⁵Excerpts contained an average of 543 sentences ($SD = 72$), equivalent to ~30 pages of a typical printed novel. This corresponds to 32,578 sentences or ~1,800 pages read by all participants combined.

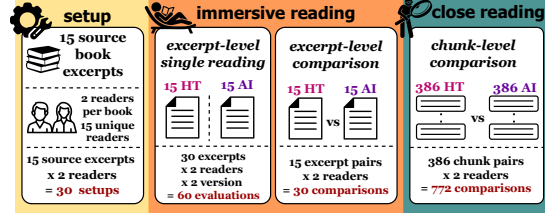


Figure 2: Study design and evaluation counts. Two readers evaluated each of the 15 book excerpts (30 book-reader evaluations).

(Google DeepMind, 2026), and P3, an AutoFiction-inspired agentic pipeline (Pham et al., 2026) using Claude Code and Codex.⁶

We used a blind five-way preference task: six raters⁷ each assessed 2–3 books. For each book, raters read five anonymized translations of the same excerpt presented in random order, selected the version they most preferred as a literary translation, and provided a short comment. The configurations were closely matched, with one exception: P3 was chosen most often, but only by a small margin, while P1 with GPT-5.4 was the clear outlier.⁸ We adopted P3, as it produces good translations, not because of evidence that it is decisively better. See §B and §C for details.

2.2 Human Evaluation Protocol

Overview. We evaluated 15 book excerpts (~8K words) translated by professional translators (published HT) and by our pipeline (agentic MT). Each excerpt was evaluated by two readers, and each reader evaluated both versions of two books⁹ (within-subjects), yielding 60 evaluations and 30 comparisons on the excerpt level (*immersive reading*; Figure 2). After a one-day break, readers re-evaluated the same excerpts chunk by chunk, yielding 772 chunk-pair comparisons (*close reading*). Figure 1 shows the full pipeline.

Participants. We recruited 15 avid readers through Upwork¹⁰ (self-reported 12–120

⁶We chose GPT and Gemini models because their model families performed strongly in WMT25 (Kocmi et al., 2025); we included coding agents because they can operate over long contexts and multiple reference files, and have been shown to produce better long-context literary fiction (Huot et al., 2025).

⁷Five of the raters were authors and one was a volunteer colleague.

⁸First-choice votes were: P3 (5/16), P1 with Gemini 3.1 Pro (4/16), P2 with Gemini 3.1 Pro (3/16), P2 with GPT-5.4 (3/16), and P1 with GPT-5.4 (1/16).

⁹Evaluations were consecutive, i.e., readers finished one book before starting the next.

¹⁰upwork.com

Question	Type
Immersive reading — single-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 §D for the full wording.

books/year) who primarily read in English and were very comfortable with long literary texts, i.e., the *target readers* of the published translations. Where possible, we matched readers to their preferred genres/story types. Participants were paid \$110 per book (~4h of work; \$27.50/h) plus a \$5–10 completion bonus; the total cost including bonuses and Upwork fees was **\$4K USD**. The protocol was approved by the Institutional Review Board. See §D for details.

Procedure. Each participant provided informed consent and completed a demographic survey on AI use and reading habits. Upon submission of a passphrase embedded in the guidelines the participants received access to the study interface designed to mimic e-book readers. They then completed the *immersive reading* task, the *close reading* task (unlocked after 24h), and a *post-study* questionnaire (see Table 2 for questions overview).

Readers' preferences: **Immersive** vs **Close** reading

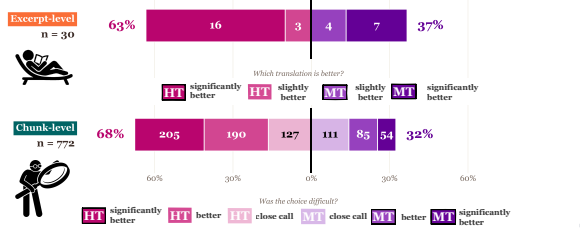


Figure 3: The distribution of readers' preferences between the **human translation (HT)** and the **machine translation (MT)** after excerpt-level immersive reading and chunk-level close reading. The readers were judging which version they preferred and how strong the preference was.

Immersive reading. Participants read one translation of the entire excerpt and answered eight *single-reading* questions, seeing only that version; they then repeated this for the second translation. Only after both readings did they compare the two versions directly (seven comparative questions). Reading order was counterbalanced for each book, i.e., one reader started with HT and the other with MT, and each reader started with HT for one of their books and MT for the other. Participants were not told which version was MT.

Close reading. Participants then assessed aligned ~300-word HT and MT chunks side by side: for each pair, they highlighted especially good or poor wording, chose the preferred translation, justified the choice, and rated its difficulty.

Quality control. We logged screen-reading time and flagged anomalies such as fast scrolling; two readers were excluded (fast reading, sparse comments) and replaced, keeping the sample at 15. All participant comments were additionally screened with the Pangram AI detector (Emi and Spero, 2024); none were flagged as AI-generated or AI-assisted.

3 Results & Analysis

Readers prefer HT, especially in close reading. After reading both excerpt-level translations, 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 (exact 95% CI for the HT share [44%, 80%]), with 16 clear preferences for HT compared to 7 clear preferences for MT (Figure 3). The preference for HT was not statistically significant (order-adjusted

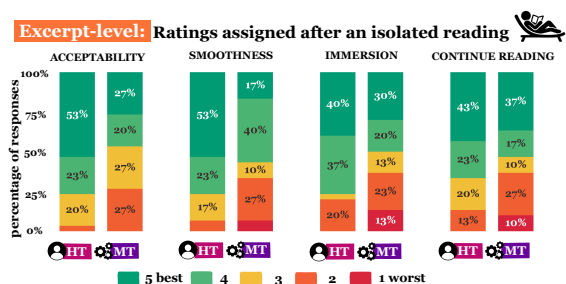


Figure 4: The distribution of readers' ratings after immersive reading of each excerpt **in isolation**. The readers were judging **acceptability** (whether it is acceptable as a published translation), **smoothness**, **immersiveness** (whether it supported immersion), and **continue reading** (whether they would continue reading).

HT probability = 63.4%; OR = 1.73, 95% CI [0.82, 3.65]; $p=.148$).¹¹

Preference for HT is clearer in *close reading* (ch-Q2) with 522 of 772 chunk-level judgments favoring HT, including 205 strong preferences, compared to 250 judgments favoring MT, including 54 strong preferences (Figure 3). The chunk-level HT preference was significant (OR = 3.28, 95% CI [1.41, 7.59], $p=.0056$), and remained significant when preference strength was modeled directly (signed mean = 0.91, 95% CI [0.24, 1.58], $p=.010$).¹²

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 reading (Figure 4). HT received a large proportion of top ratings (4 or 5) and no rating of 1. Statistical analysis shows that HT had significantly higher odds of receiving better ratings for acceptability as a published translation and for smoothness. Compared with MT, HT had 4.0 times the odds of a higher acceptability rating ($p=.0069$; at most .043 across all six random-effects specifications, Table 38) and 4.3 times the odds of a higher smoothness rating ($p=.0029$; at most .018).

¹¹Unless otherwise stated, statistical tests in this section use mixed-effects regression models with random intercepts for reader and book, and, where supported by a likelihood-ratio test, their interaction, with a model family appropriate for the response variable type. When a mixed-effects model is singular, we report the corresponding simpler model. See §H for details.

¹²Reading order (HT-first vs. MT-first) was included as a predictor in the preference models and was not significant (excerpt-level: $p=.705$, chunk-level: $p=.327$).

Span density by version and source language

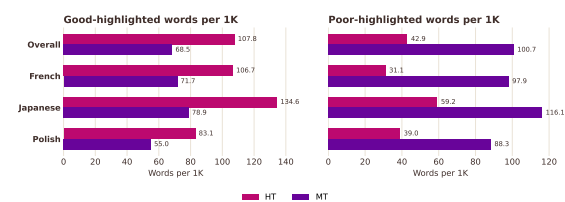


Figure 5: Distribution of **good** and **poor** span highlights by source language and translation type.

The differences for immersion and willingness to continue reading favored HT but were not significant ($p=.0815$ and $.1150$).¹³

HT is preferred on dialogue and word choice.

At the excerpt level, readers also compared the versions on dialogue and word choice (Figure 10), favoring HT on both (17 vs. 9 and 18 vs. 10, with 4 and 2 ties), though neither difference was significant (signed means = 0.27; $p=.109$ and $.131$; Table 35).

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 3). After *immersive reading*, 54% of MT responses indicated moderate to strong willingness to continue reading (a rating of 4 or 5), compared with 66% for HT (Figure 4). And in the *close reading*, roughly one-third (32%) of chunk-level choices favored MT.

MT preference varies by book but not by language.

The one-third MT share in close reading was stable across source languages (French 32%, Japanese 34%, and Polish 31%).¹⁴ By contrast, preference toward MT varied across books, from 4% to 88%, and book identity was strongly associated with chunk-level preference ($p<.001$). 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.

Readers highlight more positive evidence in HT and more negative evidence in MT. The span annotations show the same pattern as the preference judgments, i.e., readers highlighted more positive

¹³Models with random slopes for type (on reader or book) yield comparable results; see Table 38.

¹⁴There was no evidence that language predicted MT preference after accounting for reading order and repeated judgments by reader and book ($p=.928$).

evidence in HT than in MT (107.8 vs. 68.5 highlighted words per 1K words) and more negative evidence in MT than in HT (100.7 vs. 42.9), a gap that holds across all three source languages (Figure 5). Overall, HT received significantly more positive evidence spans (rate ratio = 1.69) and fewer negative spans (rate ratio = 0.43; both $p < .001$; Figure 6). Negative evidence is also more concentrated in MT chunks with 41.7% having at least 100 negative-highlighted words per 1K, versus 11.9% of HT chunks (Figure 25). Figure 15 (upper row) shows an example where both readers preferred HT, with their comments and highlights.

Span annotations closely track explicit preferences. For each chunk judgment, we computed a net span score per version (good minus poor highlighted words); in cases with no ties, the version with the higher score matched the reader’s stated preference in 691 of 755 judgments (91.5%), rising to 692 of 722 (95.8%) with span counts instead of word counts (both exact binomial $p < .001$). Thus, span annotations directly capture the local textual evidence driving preferences.

Both MT and HT vary in quality, but MT has more variable weak spots. Aggregating readers’ highlights by chunk within each book, poor-wording evidence was more variable for MT than HT in 13 of 15 books ($p = .0074$, paired sign test), and net quality (good minus poor per 1K words) in 12 of 15 ($p = .0352$).¹⁵

The excerpt-level show similar pattern. MT ratings span the full 5-point scale, whereas HT is never rated 1, and within-participant variability was directionally higher for MT though not significantly ($p = .285$, §E). While readers sometimes find MT smooth or readable, they also encounter chunks (or entire excerpts) with markedly more visible issues.

Literary preference has shared and subjective components. Since every excerpt was evaluated by two readers, we further look at inter-reader agreement. Specifically, we look at whether paired readers preferred the same translation and whether they assigned similar ratings. Excerpt preference was the most subjective case, where readers chose the same version in only 6 of 15 books (40.0%; $AC1 = -0.120$). This was higher for chunk-level

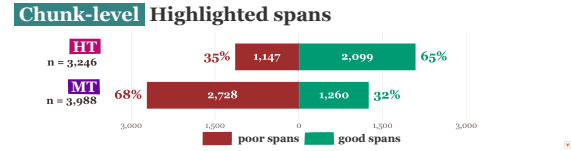


Figure 6: Distribution of span-level highlights in the close-reading task.

choices (63.7%; $AC1 = 0.355$), and even more so once the chunks marked as *similar quality* were excluded (72.2%; $AC1 = 0.546$), with the highest agreement when both readers marked one version as *significantly better* (79.5%; $AC1 = 0.667$). Perhaps unsurprisingly, this suggests that disagreement is concentrated in global preferences and when the choice between the two translations is less apparent, while clearer local preferences are shared more often between the readers.

For single-reading quality ratings, overall agreement was moderate ($AC2 = 0.426$) but notably stronger for HT than for MT (0.616 vs. 0.216). It was clearest for concrete craft ratings of HT (acceptability 0.765, smoothness 0.725) and weakest for experiential ones on MT (immersion 0.057, continue_reading 0.089; §E). Together, these results suggest readers agree on textual quality, especially locally, but diverge on how much an uneven text still functions for them as literature. Because each reader compared HT and MT directly (within-subjects), this disagreement is unlikely to reflect differences in scale use; we therefore view it as an integral part of literary judgment, combining a shared sensitivity to textual quality with reader-specific priorities (Plank, 2022; Guerberof-Arenas and Toral, 2024; Marco et al., 2025).

Coding comments. We collected 180 comments during *immersive reading* and 772 during *close reading*. Following an inductive coding approach (Thomas, 2006), we developed two schemas: one for translation quality (29 categories in four families) and one for cues that a translation was judged AI-generated (10 categories). Two researchers manually coded all *immersive reading* comments, resolving disagreements. For *close reading*, following prior work (Ziems et al., 2024), we validated GPT-5.4 against a manually double-coded 10% sample, then coded the rest automatically.¹⁶ See §E for labels and details.

¹⁵Chunks containing both good and poor spans were uncommon, but slightly more common for MT (6.7%) than HT (3.9%).

¹⁶We measure agreement as the average Jaccard overlap between the human- and model-annotated label sets (multiple labels), which was 0.727 (high overlap). See §F for details.

	HT top comment labels	MT top comment labels
Single-reading What stood out <u>positively</u> ? 	C2 Smoothness / reading effort 19/30 C1 Comprehension / ease of following 12/30 C3 Engagement / immersion 12/30 B3 Description / vividness 9/30	C2 Smoothness / reading effort 11/30 C1 Comprehension / ease of following 9/30 B1 Dialogue: naturalness 7/30 C3 Engagement / immersion 5/30
Single-reading What stood out <u>negatively</u> ? 	A3 Word choice: naturalness 5/30 A4 Word choice: idiomaticity 5/30 A6 Word choice: richness vs. blandness 5/30 A1 Grammar & spelling 4/30	A8 Sentence structure 15/30 C1 Comprehension / ease of following 11/30 C2 Smoothness / reading effort 11/30 A1 Grammar & spelling 8/30
Comparison Why this <u>excerpt</u> ? 	C2 Smoothness / reading effort 15/19 C1 Comprehension / ease of following 8/19 A3 Word choice: naturalness 6/19 A8 Sentence structure 6/19	C5 Enjoyment / overall positive affect 5/11 C2 Smoothness / reading effort 5/11 C3 Engagement / immersion 5/11 A6 Word choice: richness vs. blandness 4/11
Per chunk pair Why this <u>chunks</u> ? 	C2 Smoothness / reading effort 92/522 A2 Word choice: clarity 87/522 A3 Word choice: naturalness 80/522 C1 Comprehension / ease of following 64/522	A3 Word choice: naturalness 35/250 A5 Word choice: register & contextual fit 29/250 C2 Smoothness / reading effort 26/250 A6 Word choice: richness vs. blandness 23/250

Figure 7: The most frequent labels in readers' comments about the **human translation (HT)** and the **machine translation (MT)**. Rows separate single-reading praise, single-reading criticism, excerpt-level preference reasons, and chunk-level preference reasons. Comments could receive multiple labels.

Both HT and MT are praised for readability, but HT receives more literary praise. In the *single-reading* comments, readers **praised** both versions for basic readability (Figure 7, top row). Smoothness was the most frequent positive mention for both, though more often for HT (19/30 vs. 11/30), and comprehension showed a smaller gap (12/30 vs. 9/30), indicating that MT often cleared a basic readability bar. The gap widened for more literary and experiential qualities: HT received more praise for engagement/immersion (12/30 vs. 5/30), and description/vividness was among its top labels (9/30). MT was praised for dialogue naturalness (7/30), showing that readers did sometimes credit MT with literary strengths. Thus the comments do not suggest MT was unreadable; rather, HT was more often credited with qualities that made a translation feel engaging, vivid, and literary. After readers compared both excerpts directly, the reasons for choosing HT or MT became more distinct.

Readers choose HT because its fluency keeps them inside the story. In *immersive reading*, readers who preferred HT for smoothness/reading effort (15/19), comprehension (8/19), natural wording (6/19), and sentence structure (6/19; Figure 7, third row). They described HT as “*far easier to follow in terms of flow and word choice*,” especially when MT made it “*difficult to determine who was speaking*.”

In *close reading*, similarly, HT is chosen when

its wording is smooth, clear, and natural enough to keep the passage moving as literary prose (92/522 smoothness/reading effort, 87/522 word-choice clarity, 80/522 natural wording; Figure 7, fourth row). For chunks where HT won, readers often also mentioned issues with MT (84%, 439/522), with criticisms concentrated on word choice, sentence structure, comprehension, and smoothness.

Readers choose MT for engagement and effective local wording. In *immersive reading*, readers who preferred MT described it as enjoyable, smooth, or even more expressive than the HT alternative (5/11 enjoyment, 5/11 smoothness/reading effort, 5/11 engagement/immersion, 4/11 word-choice richness; Figure 7, third row). One reader praised MT’s “*more intricate and varied*” wording and its “*clearer picture of personalities and relationships*”.

In *close reading*, readers reward particular wording that better matches the scene, register, or character voice (35/250 natural wording, 29/250 register or contextual fit, 26/250 smoothness, 23/250 word-choice richness; Figure 7, bottom row). For instance, readers preferred “*bento containers*” over “*plastic containers*,”. Together, the excerpt-level and chunk-level comment evidence shows the similar patterns. MT can win when readers experience it as the more engaging excerpt overall, or when a local wording choice creates a better literary effect than the corresponding HT. More details in §E.

Negative comments highlight different issues in HT and MT. In *single-reading* criticism, complaints about HT were diffuse and centered on local word choice (naturalness, idiomaticity, and blandness each appeared in 5/30 comments) followed by grammar, register, smoothness, and AI tells (4/30 each; Figure 7, second row). Complaints about MT were more concentrated around reading friction, specifically sentence structure (15/30), smoothness/reading effort (11/30), and comprehension/ease of following (11/30), with grammar and word naturalness also frequent (8/30 each). Readers described MT as “*abrupt and choppy*,” with “*longwinded run-on sentences*,” or as forcing them to “*stop and reread parts*.” In six HT readings, readers found nothing negative to report, compared with only two MT readings.

Readers were surprised by the MT quality. In the post-study questionnaire, 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; e.g., “*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*”). See §E for details.

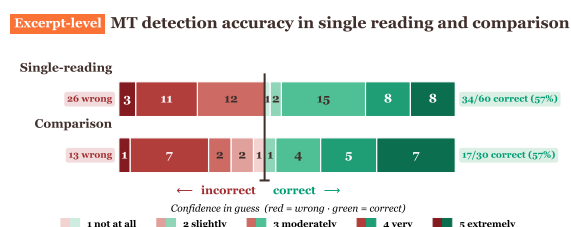


Figure 8: Machine-translation (MT) identification accuracy and confidence in the single-reading questionnaire (n=60) and comparative questionnaire (n=30). Colors represent confidence levels, from *not at all* to *extremely* confident.

Readers cannot reliably tell MT from HT. In the single-reading questionnaire, readers correctly labeled the MT excerpt as machine-translated in 16 of 30 cases and the HT excerpt as human-translated in 18 of 30 cases. Overall, we do not find any evidence that the readers were able to distinguish MT from HT in single-reading above the chance level (34/60, $p=.413$) with no difference between the MT and HT accuracy ($p=.557$).

Moreover, even after seeing both translations, readers selected the actual MT version as more likely to be AI-translated only in 17 out of 30 pairs ($p=.585$; Figure 8). While correct guesses were more confident after seeing both translations, that increase was not significant ($p=.129$). Moreover, the high confidence did not guarantee correctness. After reading both excerpts, 8 of the 20 “very” or “extremely” confident guesses were wrong (40%), including 1 of 8 “extremely” confident guesses (12.5%). See §E for more details.

Readers prefer the translation they perceive as human. Readers’ final MT guess was connected to their preferences; in 28 of 30 comparisons they judged their preferred translation to be the human one (all 11 readers who preferred MT believed it was human, as did 17 of 19 who preferred HT). Only in two cases did the readers prefer HT while still judging it to be MT. In one case the reader was misled by *em-dashes* (i.e., a strong belief that AI uses more *em-dashes*). In another one the reader noticed French structure in the HT for a book that was originally written in French, which convinced them that this translation, despite being better, is MT.

What cues do readers use to identify MT? Word choice was the most frequently cited cue, for both correct (8/17) and incorrect (9/13) guesses, making it non-discriminative. Correct guesses described wording as “awkward,” incorrect ones as “generic” or “weird” (“*people just don’t speak like that*”).

More telling was whether the readers relied on concrete textual evidence or on “folk theories” about AI (i.e., the common idea of what AI writing is). The strongest separator was formatting (M4) such as scene dividers or paragraph breaks, which made it hard to follow. Another cue with stronger discriminative power was literal rendering (M1; 6/17 vs. 3/13, e.g., MT described as “*like a literal dictionary translation*”) along with poor sentence flow or grammar (M3; 7/17 vs. 4/13, e.g., run-on and cut-off sentences).

The most deceptive were “folk theories,” which resulted in accuracy no better than chance (M10; 5/17 vs. 5/13). These include *em-dashes* perceived as “*a classic sign of AI*,” or a belief that AI “prefers to avoid” swearing (Russell et al., 2025). Least reliable were references to subjective literary quality, i.e., voice and emotional tone (M7) was cited mostly by readers who guessed wrong (2/17 vs.

5/13), who called the human translation “flat” and “mechanical” and so took it for the machine. See §E and Figure 23.

Can automatic metrics capture readers’ preferences? We compare readers’ chunk-level preferences with three automatic metrics: METRICX-QE (Juraska et al., 2024), COMETKIWI (Rei et al., 2022), and the LLM-as-a-judge framework LiTRANSProQA (Zhang et al., 2025b). All metric variants favored MT over HT, contrary to readers’ significant preference for HT in close reading. Their correlations with human judgments were also weak or negative (Kendall’s $\tau = -0.075$ to -0.318). This result is consistent with results reported in Thai et al. (2022); Zhang et al. (2025b,a). See §I for details.

How about translations into languages other than English? To examine whether our findings extend to other target languages, we conducted a small exploratory study with two books translated out of Japanese and Polish into Spanish, French, Polish, and Japanese, evaluated by five native-speaker readers following the main-study protocol. We observe strong preference towards HT with four of five readers preferring it at excerpt level all preferring it at chunk level (103/111 judgments, 92.8%). Similarly, four of five readers identified MT correctly with high confidence.

The gap was most visible for the Japanese and Polish targets with HT rated 1.75–2 points higher than MT on every single-reading scale; for the Spanish and French targets the rating gap was small or absent, with the French target showing the lowest HT chunk share (17/22). These preliminary results suggest MT’s strong English-target performance may not extend to other target languages (see §G).

4 Related Work

Automatic evaluation. Automatic evaluation is practical but surface-level (Papineni et al., 2002; Popović, 2015) and captures only partial signals in literary MT (Griebel and Underwood, 2026), while trained neural metrics (Rei et al., 2020, 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 or creative tasks were developed but tested only on

short passages (Zhang et al., 2025b; Shafayat et al., 2024; Zhang et al., 2026) and it is unclear how they handle longer dependencies. Recent document-level and long-document MT work often combines automatic metrics with LLM-as-a-judge evaluation (Läubli et al., 2018, 2020; Wang et al., 2026), but such evaluations still do not directly measure reader experience.

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. Prior work on literary MT and reception remains narrower than our reader-facing comparison of professional HT and recent MT over long fiction excerpts (Guerberof-Arenas and Toral, 2020, 2022, 2024; Gerrits and Guerberof-Arenas, 2025; Moorkens et al., 2018; Castaldo et al., 2025; Matusov, 2019; Taivalkoski-Shilov, 2019b; Webster et al., 2020; Van Egdom et al., 2023). This reader-facing framing also relates to narrative transportation, AI-perception, literary translator voice, and ethics in machine-assisted literary translation (Green and Brock, 2000; Porter and Machery, 2024; Zhu et al., 2025; Kenny and Winters, 2020; Taivalkoski-Shilov, 2019a). 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 HT was favored descriptively at the excerpt level and significantly 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 copyrights restrictions and annotation costs we 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 does 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 with a total of 15 readers. This scale was necessary for immersive, long-form evaluation to be financially feasible, but could limit statistical power. To partially mitigate this, 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 for statistical analysis (Allen, 2017).

Ethics Statement

Human participants. This study was approved by the Simon Fraser University Research Ethics Board (#30003790). Participants gave informed consent before the evaluation (§D), 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.

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 overrepresents 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 often prefer human-authored translations. As LLM-translated fiction enters trade and self-publishing channels, **we hope that better reader-centered evaluation and full disclosure 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.

AI use disclosure. We utilize coding agents to produce plots and edit table layout. We further use language models as writing assistants but no part of this paper was generated from scratch.

Acknowledgements

We would like to thank the Upwork participants for their help in evaluating the different book excerpts, and for taking the time to answer our questions with detailed and thoughtful responses that were valuable to our study. We are grateful to the volunteer and researcher readers who contributed to the multilingual case study. This work was supported by the following grants and awards: Simon Fraser University’s Rajan Recruitment Award for Women in Computing Science [M Karpinska], Natural Sciences and Engineering Research Council of Canada (NSERC) [MJ Meurs, Discovery Grant RGPIN-2025-07163], [M Taboada, Discovery Grant RGPIN-2022-04481], [M Karpinska, Discovery Grant RGPIN 2026-04987], and the Fonds de recherche du Québec

(FRQ) Chaire de recherche du Québec sur la découvrabilité des contenus scientifiques en français [MJ Meurs, Chaire de recherche du Québec sur la découvrabilité des contenus scientifiques en français, DOI 10.69777/358425].

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.
- Anthropic. 2026. Claude Opus 4.6 System Card. <https://www.anthropic.com/system-cards>. Accessed 2026-07-02.
- Dale J. Barr. 2013. [Random effects structure for testing interactions in linear mixed-effects models](#). *Frontiers in Psychology*, Volume 4 - 2013.
- Elke Brems and Sara Ramos Pinto. 2013. [Reception and translation](#). In Yves Gambier and Luc van Doorslaer, editors, *Handbook of Translation Studies*, volume 4, pages 142–147. 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.
- Tuhin Chakrabarty, Jane C Ginsburg, and Paramveer Dhillion. 2025. Readers prefer outputs of ai trained on copyrighted books over expert human writers. *arXiv preprint arXiv:2510.13939*.
- 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.
- 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 and Ana Guerberof-Arenas. 2025. To mt or not to mt: An eye-tracking study on the reception by dutch readers of different translation and creativity levels. In *Proceedings of Machine Translation Summit XX: Volume 1*, pages 516–537.
- 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://globscribe.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.
- Melanie C Green and Timothy C Brock. 2000. The role of transportation in the persuasiveness of public narratives. *Journal of personality and social psychology*, 79(5):701.
- Sarah Griebel and Ted Underwood. 2026. Fluency and faithfulness in human and machine literary translation. *arXiv preprint arXiv:2605.15282*.
- Ana Guerberof-Arenas and Antonio Toral. 2020. The impact of post-editing and machine translation on creativity and reading experience. *Translation Spaces*, 9(2):255–282.
- Ana Guerberof-Arenas and Antonio Toral. 2022. Creativity in translation: Machine translation as a constraint for literary texts. *Translation spaces*, 11(2):184–212.

- Ana Guerberof-Arenas and Antonio Toral. 2024. To be or not to be: A translation reception study of a literary text translated into dutch and catalan using machine translation. *Target*, 36(2):215–244.
- 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.
- Fantine Huot, Reinald Kim Amplayo, Jennimaria Palomaki, Alice Shoshana Jakobovits, Elizabeth Clark, and Mirella Lapata. 2025. [Agents’ room: Narrative generation through multi-step collaboration](#). In *The Thirteenth International Conference on Learning Representations*.
- 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.
- Dorothy Kenny and Marion Winters. 2020. Machine translation, ethics and the literary translator’s voice. *Translation Spaces*, 9(1):123–149.
- 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 Lakouagna, 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.
- Muhammed Yusuf Kocyigit, Eleftheria Briakou, Daniel Deutsch, Jiaming Luo, Colin Cherry, and Markus Freitag. 2025. [Overestimation in LLM evaluation: A controlled large-scale study on data contamination’s impact on machine translation](#). In *Forty-second International Conference on Machine Learning*.
- Samuel Läubli, Sheila Castilho, Graham Neubig, Rico Sennrich, Qinlan Shen, and Antonio Toral. 2020. A set of recommendations for assessing human–machine parity in language translation. *Journal of artificial intelligence research*, 67:653–672.
- Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 4791–4796.
- 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.
- Guillermo Marco, Julio Gonzalo, and Víctor Fresno. 2025. [The Reader is the Metric: How Textual Features and Reader Profiles Explain Conflicting Evaluations of AI Creative Writing](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25432–25449, Vienna, Austria. Association for Computational Linguistics.
- Evgeny Matusov. 2019. The challenges of using neural machine translation for literature. In *Proceedings of the qualities of literary machine translation*, pages 10–19.
- Joss Moorkens, Antonio Toral, Sheila Castilho, and Andy Way. 2018. Translators’ perceptions of literary post-editing using statistical and neural machine translation. *Translation Spaces*, 7(2):240–262.
- Annu Nishioka. 2024. [Japanese publisher to launch ‘light novel’ app with AI-assisted translations](#). *Nikkei Asia*. Accessed: 2026-05-24.
- OpenAI. 2026. Introducing gpt-5.4. <https://openai.com/index/introducing-gpt-5-4/>. Accessed 2026-07-02.

- 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, Atrey Desai, and Mohit Iyyer. 2026. [Autofiction](#). Web platform.
- Barbara Plank. 2022. The “problem” of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 10671–10682.
- 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.
- Brian Porter and Edouard Machery. 2024. Ai-generated poetry is indistinguishable from human-written poetry and is rated more favorably. *Scientific Reports*, 14(1):26133.
- Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. In *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)*, pages 2685–2702.
- 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*.
- Kristiina Taivalkoski-Shilov. 2019a. Ethical issues regarding machine (-assisted) translation of literary texts. *Perspectives*, 27(5):689–703.
- Kristiina Taivalkoski-Shilov. 2019b. Free indirect discourse: an insurmountable challenge for literary mt systems? In *Proceedings of the qualities of literary machine translation*, pages 35–39.
- 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.
- Gys-Walt Van Egdom, Onno Kusters, and Christophe Declercq. 2023. [The riddle of \(literary\) machine translation quality](#). *Tradumàtica tecnologies de la traducció*, (21):129–159.
- 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.
- Yutong Wang, Xuebo Liu, Derek F Wong, Zhilin Li, Rongqing Jiang, Min Zhang, Shimin Tao, and Daimeng Wei. 2026. Loong: A human-like long document translation agent with observe-and-act adaptive context selection. *arXiv preprint arXiv:2605.30274*.
- Rebecca Webster, Margot Fonteyne, Arda Tezcan, Lieve Macken, and Joke Daems. 2020. Gutenberg goes neural: Comparing features of Dutch human translations with raw neural machine translation outputs in a corpus of English literary classics. In *Informatics*, volume 7, page 32. MDPI.
- Minghao Wu, Jiahao Xu, Yulin Yuan, Gholamreza Haffari, Longyue Wang, 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, Steffen Eger, Arda Tezcan, Wei Zhao, Simone Paolo Ponzetto, and Lieve Macken. 2026. Beyond reproduction: A paired-task framework for assessing llm comprehension and creativity in literary translation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 40862–40884.

- 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.
- Tiffany Zhu, Iain Weissburg, Kexun Zhang, and William Yang Wang. 2025. Human bias in the face of ai: Examining human judgment against text labeled as ai generated. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25907–25914.
- Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. [Multilingual machine translation with large language models: Empirical results and analysis](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.
- Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. [Can large language models transform computational social science?](#) *Computational Linguistics*, 50(1):237–291.

A Dataset

In this section of the appendix, we provide additional information regarding our dataset.

Corpus construction. A total of 35 long excerpts have been used for this study. 15 for the evaluation, 16 for the development and 4 for the case-study. Details about the selected books, such as the title, authors, but also translators and release dates are presented in [Table 3](#).

Dataset Excerpt 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 §D for details).

Dataset statistics. Additional details about the corpus size for the excerpt and chunks are presented in [Table 4](#). It also includes statistical information about the participants’ comments written during the two reading phases.

B Machine Translation

In this section of the appendix, we provide more details on our machine translation pipeline, including the selection of a reasonable translation approach.

Machine translation pipelines. Overall, we experiment with three pipelines in five setups:

- 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.

Agentic pipeline. We adapted the AutoFiction long-form fiction generation framework ([Pham et al., 2026](#)) to literary translation. Specifically,

we deployed two coding agents, Claude Code with Claude Opus 4.6 (Anthropic, 2026) and Codex with GPT-5.4,¹⁷ to produce and validate the translations. Our pipeline consisted of three main stages (see Figure 9):

1. **⚙️ Preprocessing:** Claude Code analyzed the source excerpt to produce translation guidelines (e.g., register, terminology, translation memories). The source was then divided into 1K-token chunks while preserving paragraphs.
2. **📖 Chunk-level:** Codex translated each chunk, which was then reviewed along two dimensions: (1) translation quality (e.g., faithfulness) using a modified version of LiTransProQA (Zhang et al., 2025b) (Codex), and (2) literary quality review (e.g., voice, style) adapted from AutoFiction (Claude Code). Based on the output of these reviews, an acceptance gate¹⁸ determined whether the chunk proceeded to the next stage or was sent for revision. We allowed up to three local cycles total due to the high associated cost and small gains.¹⁹
3. **📖 Excerpt-level:** Revised chunks were merged into a full draft and reviewed for (1) global literary quality and (2) cross-chunk consistency (Codex). The results passed through a final acceptance gate, which triggered revision by Claude Code if minimum quality thresholds were not met.²⁰ We allowed up to two global cycles.

Chunking. To prepare the data for the chunk-level *close reading*, we split each HT excerpt into

¹⁷We set the reasoning effort to the maximum value, which is max for Claude Code and xhigh for Codex. Both agents were accessed via a monthly subscription at the cost of \$400 USD for the entire experiment.

¹⁸The gate uses the review rubrics described in §C; chunks pass only when the automatic reviewers find no major faithfulness or literary-quality problems.

¹⁹All 16 development-book runs used up to three local cycles. The third local cycle yielded limited marginal gains in the automatic gate outcomes, clearing 8 of 130 chunks still failing after cycle 2 (6.2%); 12 of the 17 three-cycle runs had an unchanged failing-chunk set. We therefore capped subsequent main evaluation and multilingual target-language case-study runs at two local cycles. The exception is *Make Me Famous*, which was generated during the earlier development batch and was later used in the evaluation set after replacing an evaluation book with sensitive content.

²⁰The final gate uses the excerpt-level rubrics described in §C; drafts pass only when the automatic reviewers find no major global-quality or cross-chunk consistency problems.

~300-word chunks ($\mu = 296$, $SD = 49$) preserving the paragraph boundaries, yielding **386** chunks across the 15 evaluation books. We then aligned these chunks using Gemini 3.1 Pro to the corresponding chunks from MT and source text. All alignments were verified manually.

Prompts. We deploy different prompts for each pipeline. As the pipelines include multiple long prompts we refrain from including the entire prompt texts in the appendix and make them accessible through the repository instead (github.com/Yves575/lait).

Details on prompting pipelines. P1 and P2 used the same two-stage prompting procedure. In the first stage, the model was prompted as an expert literary translator and asked to translate the source excerpt into English while preserving meaning, tone, voice, register, rhythm, stylistic effect, ambiguity, characterization, and differences between narrator and character voices. The prompt also instructed the model to avoid unjustified omissions or additions and to output only the translation. In the second stage, the model received the original source excerpt together with its own first-pass English translation and was tasked with revision.

[P1] Full-excerpt translation. For P1, each complete source excerpt was translated in a single first-pass call. This condition therefore gave the model the entire opening excerpt at once, including all available discourse context, recurring names, register shifts, and dialogue patterns. The resulting complete English draft was then passed to the post-editing stage together with the complete source excerpt. The post-editing pass was also performed at the excerpt level, so both stages in P1 had access to the full source span.

[P2] Chunk-level translation with excerpt-level post-editing. For P2, the first-pass translation was performed over smaller source chunks. We counted tokens with the tiktoken o200k_base encoding and targeted a maximum of 1K source tokens per chunk. Chunking preserved paragraph boundaries: we split the source excerpt on blank-line paragraph breaks, added whole paragraphs to the current chunk until adding the next paragraph would exceed the token limit, and then started a new chunk. If a single paragraph exceeded the limit, it was kept intact as an oversized chunk rather than split mid-paragraph. Each chunk was

Agentic Pipeline for Literary Machine Translation

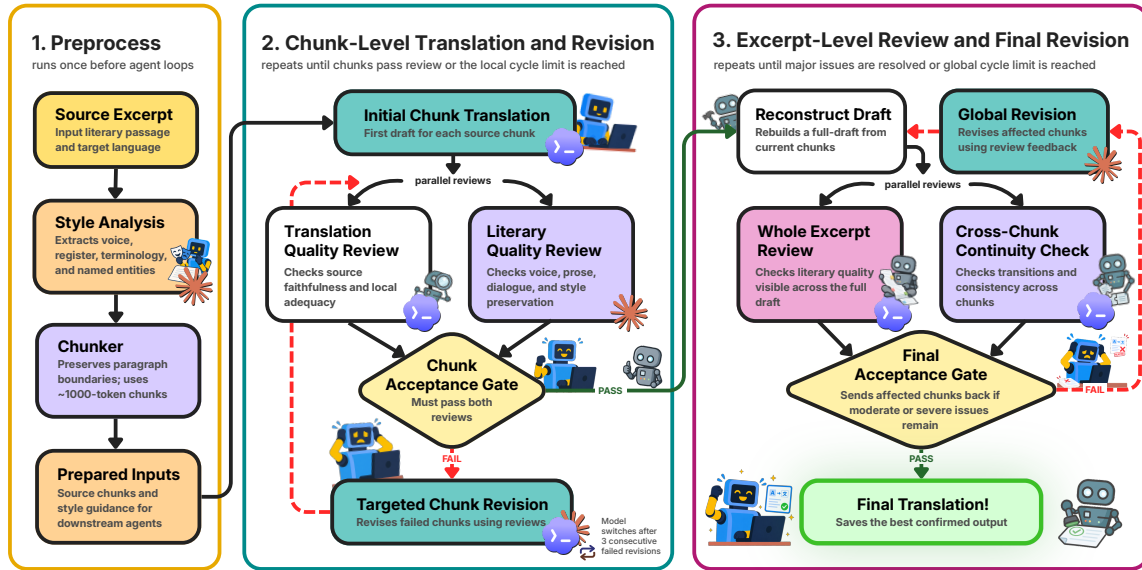
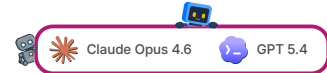


Figure 9: Agentic literary MT pipeline used in this study. Source excerpts are chunked and paired with style guidance, then translated and revised through chunk-level and full-draft review loops.

translated independently with the same literary-translation prompt used in P1. The chunk translations were then concatenated in source order with paragraph breaks preserved, producing a complete first-pass English draft. Finally, as in P1, the post-editing stage received the complete source excerpt and the complete concatenated English draft, allowing the model to revise across chunk boundaries for consistency of voice, terminology, register, narrative perspective, and local fluency.

[P3] Agentic pipeline. The entire pipeline is presented in Figure 9. We report the agent responsible for each task and the associated checkpoint in Table 5. The core parts of the pipelines are: (1) data preparation including style analysis, (2) chunk-level translation, and (3) excerpt-level review. For (2) and (3) we implement metrics and guard gates that send translations for edits if any issues are spotted. Below we explain the rubrics and acceptance gate mechanism.

[P3] Agent review rubrics and acceptance gates. The chunk-level acceptance gate combined two automatic reviews. First, the modified LiTransProQA review asked 25 yes/no/maybe questions about translation quality, such as faithfulness, cultural transfer, voice and more. A chunk passed this quality-estimation component only if it received no NO judgments and at most five MAYBE judgments.

Second, a literary review checked whether the chunk preserved the source’s accuracy, voice, dialogue, and prose, and recorded findings with LOW, MEDIUM, HIGH, or CRITICAL severity. A chunk passed the literary component only if every lens passed and there were no MEDIUM+ findings.²¹ The local acceptance gate failed any chunk that failed either component and sent only those chunks to the next revision cycle.

The excerpt-level final gate combined a full-draft book review with a cross-chunk audit. The book review targeted global translation defects such as repeated translationese, flattening, loss of ambiguity, etc., visible only when reading the excerpt as a whole. The cross-chunk audit targeted boundary and consistency problems, such as inconsistent terms, narrator or register drift, seam issues, etc. Both reviews used the same four-level severity scale, and the final gate passed only when neither review found any MEDIUM+ issue; otherwise, affected chunks were sent to final revision.

Pipeline selection. We translate 8K-words²² excerpts from 16 development books under all five

²¹Following Pham et al. (2026) we denote MEDIUM+ as any issue of severity medium or higher that would usually require revision. This excludes small issues marked as LOW.

²²The book length limit should not exceed 10% of the original word count, with a maximum limit of 8k-words, and it can be lower.

setups (Table 6).²³ In a blind five-way preference task, six raters (the authors and volunteer colleagues), each assessing 2–3 books, together ranked the outputs across all 16 development excerpts. The raters were asked to read at least 2K words to assure everyone reads past the chunk boundary. The raters can choose to read more. Besides choosing the best performing translation in their set of five candidates, each rater shares their observations on other translations. Overall, most outputs were comparable, except GPT-5.4 (P1) which noticeably underperformed²⁴. 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.

C Agentic pipeline runs

In this section of the appendix we provide more details on the agentic pipeline runs and artifacts it produced. All artifacts from agentic runs are shared at github.com/Yves575/lait.

Runs summary. We report the details of all runs in Table 7, including the number of chunks and excerpts which pass acceptance gate within the enforced limits. We note that even when the final translation failed to pass within the enforced limits we observe that increasing the limits did not show justifiable improvements, while the “failed” translation was often acceptable.

Artifact footprint. The selected agentic translation pipeline outputs comprise 35 completed runs: 15 evaluation-book runs, 16 development-book runs, and 4 multilingual target-language runs. Across these runs, the pipeline processed 441 source-language chunks and produced 1,059 chunk-cycle translation or revision artifacts, counting each completed translation or revision cycle for a chunk once. Job-level logs were available for 31 runs and record 2,930 agent jobs, including 19 failed jobs; four development-book runs were completed but untracked in the job-metric logs. At the automatic final acceptance gate, 13 runs passed and 22 retained flagged issues, with 74 chunks flagged in total (Table 8).

²³Because agents make many model calls, running them via API is prohibitively expensive; instead we used subscription services, which cap our translation count.

²⁴P3 (5/16), P1 with Gemini 3.1 Pro (4/16), P2 with Gemini 3.1 Pro (3/16), P2 with GPT-5.4 (3/16), and P1 with GPT-5.4 (1/16).

D Human evaluation

In this section of the appendix, we add more details about our human evaluation protocol, including participant recruitment, participant profiles, guidelines, and interface screenshots.

Participant recruitment. The participants for the experiments were recruited using a freelancing platform Upwork. We limit our search to avid readers, who read at least one book in English per month and have native-level command of the language. We asked them to complete an entry questionnaire (Table 13). We report their demographics in Table 9 and AI use frequency in Table 10.

Procedure. Before the experiment all participants were asked to sign the consent form (see Table 11 and Table 12). They were also asked to fill in a demographic survey (Table 13). All surveys were administered through Microsoft Forms. During the recruitment stage, we additionally ask them to report their reading habits.

Questions. During the human evaluation, the readers fill in three types of questionnaires embedded in the interface. In the *immersive reading* they are asked to fill in a questionnaire about each text separately (single-reading; Table 14), and then about both texts (comparison; Table 15). In the *close reading* each chunk pair is accompanied by a short questionnaire asking them to compare both translations (Table 16).

Questionnaires. We used four questionnaires. The *immersive reading* task included two: (1) a single-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 on translations. See Table 2 for questions.

Quality control. To ensure data quality, we required participants to submit a passphrase embedded in the guidelines before they received their credentials, confirming they had read the instructions. We also recorded screen-reading time and detected anomalies such as fast scrolling.²⁵ Finally,

²⁵Two readers were excluded for insufficient comment detail and fast reading, and replaced by two others to keep the

we ran the Pangram AI detector (Emi and Spero, 2024) over participant comments to flag potentially AI-generated text.²⁶

Interface. Participant-facing guidelines and interface screenshots are provided in the supplementary materials section of the GitHub repository (github.com/Yves575/lait).

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.

E Human evaluation results

In this section of the appendix, we report additional results and descriptive statistics of the human evaluation data.

Overall results. At the excerpt level, readers preferred HT more often for both dialogue handling and word choice (Figure 10). Close-reading judgments also favored HT overall: MT accounted for about one-third of preferences in each source language, although choices varied substantially across books and individual chunks (Figure 29 and Figure 19). In the smaller multilingual case study, HT was preferred in 103 of 111 chunk judgments. The book-level results show the only excerpt-level MT preference, while the span annotations show mostly positive evidence for HT and negative evidence for MT across target languages (Figure 13, Figure 12, and Figure 11).

Agreement. We report inter-reader agreement at three levels: (1) single-reading ratings, (2) binary preference and origin judgments (MT or HT), and (3) span highlights. For the **single-reading** 5-point ratings, we report Gwet’s AC2 (Gwet, 2021) for acceptability, smoothness, immersiveness, would continue (Table 18) as it is more fit for skewed distribution data than Krippendorff’s alpha and handles ordinal variables.

sample size at 15.

²⁶No comments were flagged as AI-generated or AI-assisted.

Excerpt-level: Dialogue and word choice

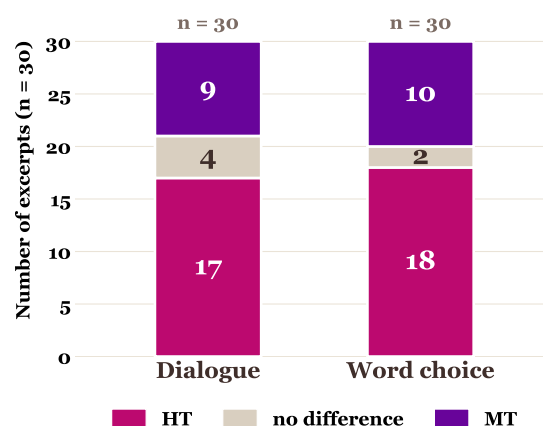


Figure 10: The distribution of readers’ excerpt-level preferences for **dialogue handling** and **word choice** after reading both the **human translation (HT)** and the **machine translation (MT)**. The readers were judging which version handled dialogue better overall and which version had more varied or expressive word choice.

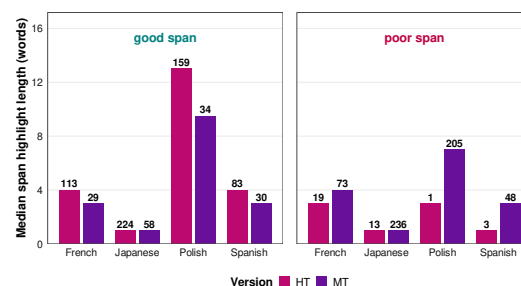


Figure 11: Median span highlight length in the multilingual case study. Bars show the median length of good and poor spans marked for HT and MT by target language.

For **binary judgments**, we report the Gwet’s AC1 (nominal data; Table 19) for excerpt-level MT identification, excerpt-level preference, and chunk-level preference. Chunk-level agreement is reported for all chunks, for chunks excluding *similar quality* judgments, and for the subset where both readers judged one translation as *significantly better*. We also report chunk-level agreement measured separately by book (Table 20), showing how it varies across the items and after excluding *similar quality* judgments.

²⁷HT = human-translated; MT = machine-translated; n = paired ratings. AC2 colors follow Landis and Koch heuristic bands: red = slight, yellow = fair, green = moderate, blue = substantial, purple = almost perfect.

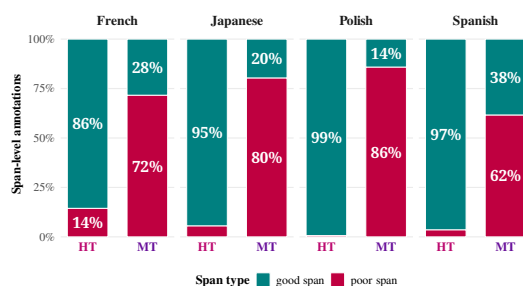


Figure 12: Span-level annotations by target language in the multilingual target-language case study. Bars show the proportion of good and poor spans marked in HT and MT for each target language.

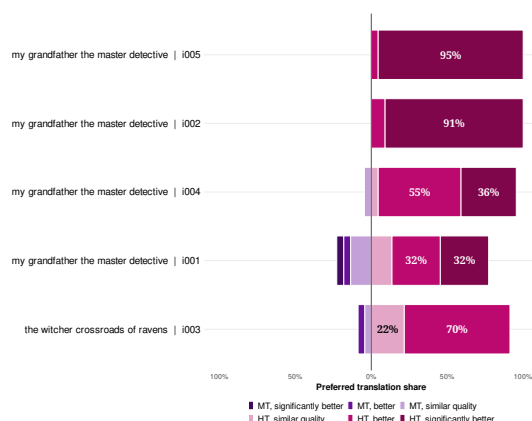


Figure 13: Book-level preferences in the multilingual case study. Each row is one book–participant pair; left/right bars indicate MT/HT preference, with darker shades marking stronger choices.

For **span highlights**, we report overlap diagnostics rather than a chance-corrected coefficient. A span is counted as overlapping when it intersects a span from the other reader on the same translation version; “label-aware” overlap additionally requires the same *good/poor* label. Overall span overlap was 0.307, and label-aware overlap was 0.217. Readers who agreed on the preferred translation overlapped more in their highlighted evidence (0.322 span overlap; 0.263 label-aware overlap) than readers who disagreed (0.280 span overlap; 0.137 label-aware overlap). Table 23 shows examples and summarizes the top preference labels and representative comments.

Comment coding. We collected 180 comments during *immersive reading* and 772 during *close*

²⁸ n is paired judgments. Both HT, Both MT, and Diff. count paired annotator outcomes. For excerpt AI/MT identification, HT and MT refer to which version readers guessed was MT. AC1 colors follow the same heuristic bands as Table 18.

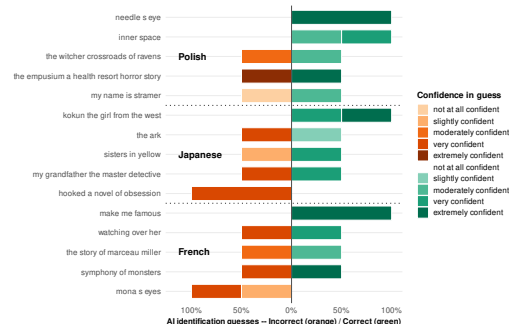


Figure 14: MT identification by book. Orange bars show HT identified as MT, green bars show MT identified as MT, and darker shades indicate higher confidence.

reading.²⁹ Following an inductive coding approach (Thomas, 2006), we developed two coding schemas: (1) one for translation quality (29 categories) and (2) one for why a translation was judged AI-generated (10 categories).³⁰ The coding schema was developed by first running all comments through GPT-5.4 and Claude Opus 4.7, which helped us to obtain a list of candidate categories. These were used as support for the *inductive coding* conducted by the researchers who further refined the schema.

All *immersive reading* comments were manually coded by two researchers, who resolved disagreements. For *close reading*, we validated GPT-5.4³¹ coding against a 10% sample coded manually by two researchers, then coded the rest automatically.³² The coded comments are released in the [repository](#).

The category labels for translation quality and the reasons for judging a translation as AI-generated are described in Table 21 and Table 22 respectively. We present comment examples in Table 33.

Comment analysis. Table 27 shows the top positive labels assigned to comments by readers who preferred HT in the immersive reading comparative questionnaire to indicate why they preferred

²⁹For *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.

³⁰The quality schema groups its 29 categories into four families: (A) language-level, (B) narrative-level, (C) reader experience, and (D) meta-translation.

³¹We deploy the model through Codex utilizing the monthly subscription.

³²As comments can receive multiple labels, we measure agreement as the average Jaccard overlap between the human- and model-annotated label sets, which was 0.727 (high overlap).

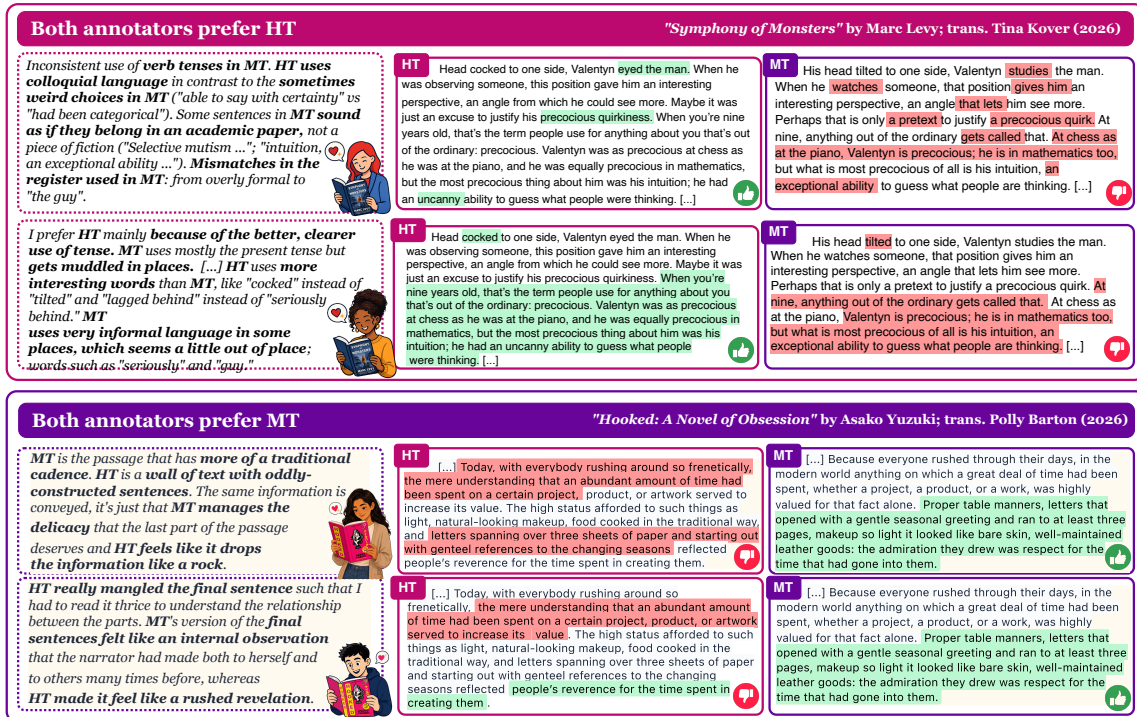


Figure 15: Examples of span-level preference evidence from the side-by-side chunks evaluation. Participants compared aligned HT and MT chunks, labeled spans as **good** or **poor**, and chose the preferred translation. Green highlights indicate **good** spans and red highlights indicate **poor** spans. The examples show cases where both readers chose the same translation, one favoring HT and one favoring MT, while marking different local evidence for their choices.

it. Table 26 shows the top positive labels for comments from readers who preferred MT in that questionnaire. Table 25 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 24 shows the top positive labels for comments from readers who preferred MT in that questionnaire. Table 23 summarizes the top preference labels and representative comments.

We further report tendencies mentioned in comments in Figure 16, which summarizes positive and negative aspects readers mentioned for MT and HT in the comparison comments. Similarly, Figure 17 shows the positive reasons readers gave for preferring HT or MT, separated by excerpt-level and chunk-level comparisons.

As readers often mention not only positive things about their chosen translation but also negative things about the other candidate, we aim at visualizing these in Figure 18.

Finally, we present a book-level view of AI guesses per language (Figure 14), while Figure 20 shows book-level HT/MT preference direction and

strength for each participant.

Single-reading ratings. Figure 22 shows single-reading ratings for acceptability, smoothness, immersiveness, and would continue (reading), divided by the presentation order (i.e., ratings for what the readers saw first vs. second). We further report these ratings by individual preferences, indicating both the translation order and the readers' final preference for MT vs. HT in the excerpt-level evaluation (Figure 21).

AI-detection guesses. For AI-detection questions, we compare readers' single-reading guesses with their final comparison guesses. Figure 23 shows the most frequent cues readers mentioned when identifying a translation as machine-translated, separated by correct and incorrect guesses. Additional descriptive flows for guess correctness and confidence are shown in Figure 24a and Figure 24b.

Readers who often use AI were NOT better at detecting AI translation. Since we collect readers AI experience we check whether the more ex-

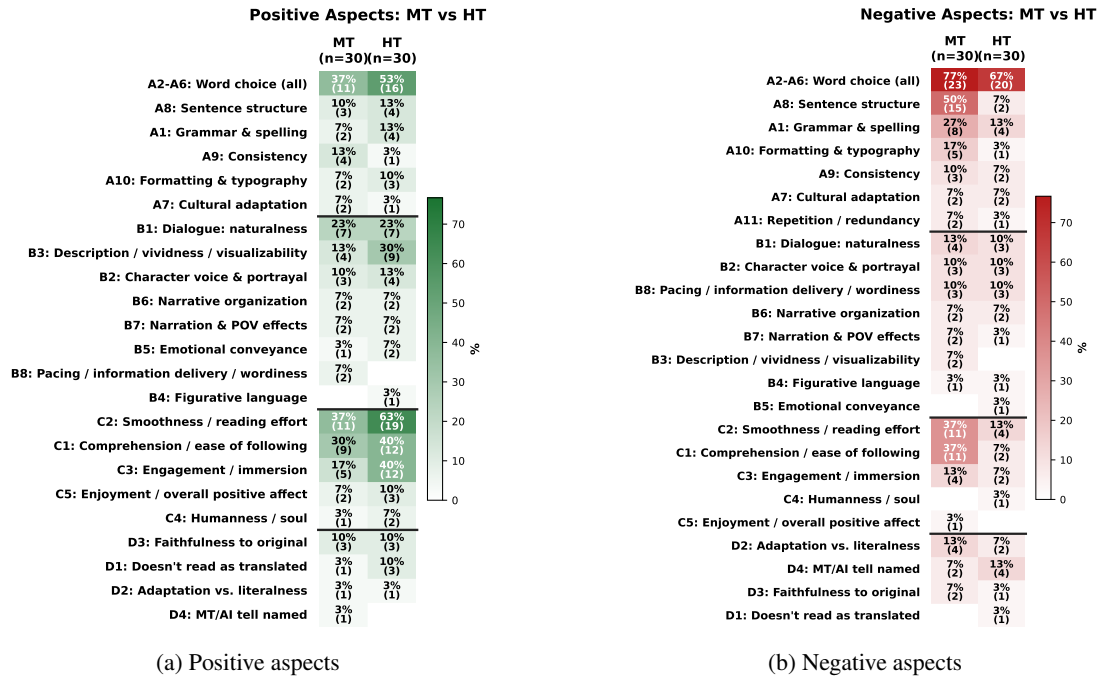


Figure 16: Positive and negative aspects mentioned by participants when comparing machine translation (MT) and human translation (HT).

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: idiomatity (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%)
C. Reader experience		
Smoothness & reading effort (C2)	5 (45.5%)	15 (76.9%)
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%)	36 (15.2%)
Word choice: clarity (A2)	16 (6.4%)	42 (16.2%)
Word choice: register & contextual fit (A5)	29 (11.6%)	26 (10.4%)
Sentence structure (A8)	11 (4.4%)	39 (15.6%)
Word choice: richness vs. blandness (A6)	23 (9.2%)	24 (9.6%)
Grammar & spelling (A1)	8 (3.2%)	22 (8.8%)
Word choice: idiomatity (A4)	3 (1.2%)	15 (5.9%)
Cultural adaptation (A7)	9 (3.6%)	4 (1.6%)
Formatting & typography (A10)	3 (1.2%)	7 (2.8%)
Consistency (A9)	2 (0.8%)	4 (1.6%)
Repetition or redundancy (A11)	0 (0.0%)	1 (0.4%)
B. Narrative-level features		
Description & vividness (B3)	22 (8.8%)	39 (15.6%)
Pacing & information delivery (B8)	17 (6.8%)	31 (12.4%)
Narrative organization (B6)	6 (2.4%)	17 (6.8%)
Character voice & portrayal (B2)	16 (6.4%)	4 (1.6%)
Dialogue: naturalness (B1)	6 (2.4%)	15 (5.9%)
Emotional conveyance (B5)	7 (2.8%)	8 (3.2%)
Figurative language (B4)	7 (2.8%)	5 (1.9%)
Narration & POV effects (B7)	8 (3.2%)	4 (1.6%)
C. Reader experience		
Smoothness & reading effort (C2)	26 (10.4%)	52 (20.8%)
Comprehension & ease of following (C1)	17 (6.8%)	54 (21.6%)
Enjoyment & overall positive affect (C5)	4 (1.6%)	14 (5.6%)
Engagement & immersion (C3)	3 (1.2%)	9 (3.6%)
Humanness & soul (C4)	2 (0.8%)	5 (1.9%)
D. Meta-translation		
Faithfulness to original (D3)	13 (5.2%)	19 (7.6%)
Adaptation vs. literalness (D2)	1 (0.4%)	8 (3.2%)
AI tell named (D4)	0 (0.0%)	4 (1.6%)
Doesn't read as translated (D1)	0 (0.0%)	3 (1.2%)

(b) Chunk-level preferences

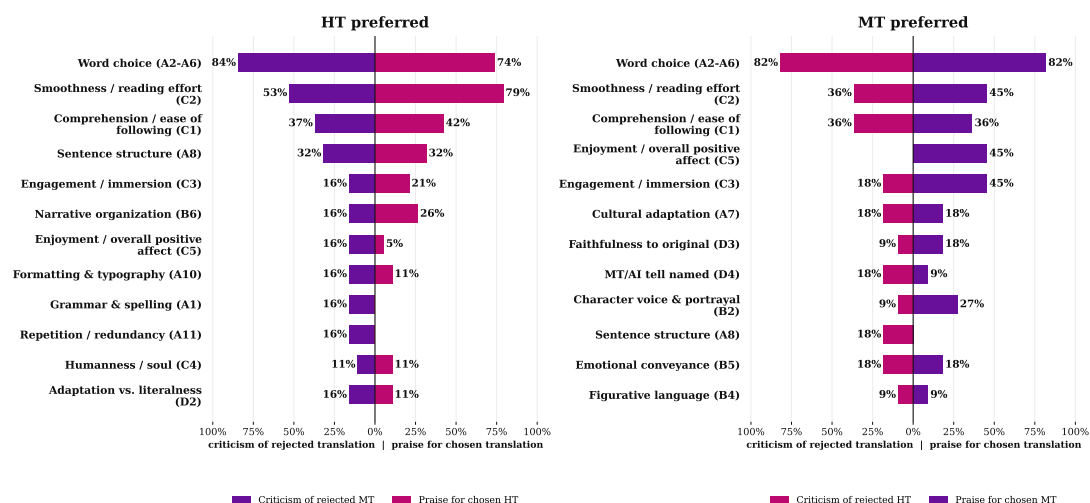
Figure 17: Positive reasons readers reported for preferring one translation over the other, separated by preferred system. Counts and percentages are calculated within each preferred-system group.

perienced readers were better at detecting AI translation. Overall, we do not see any apparent trends or correlation between the AI experience and detection accuracy (Table 17).

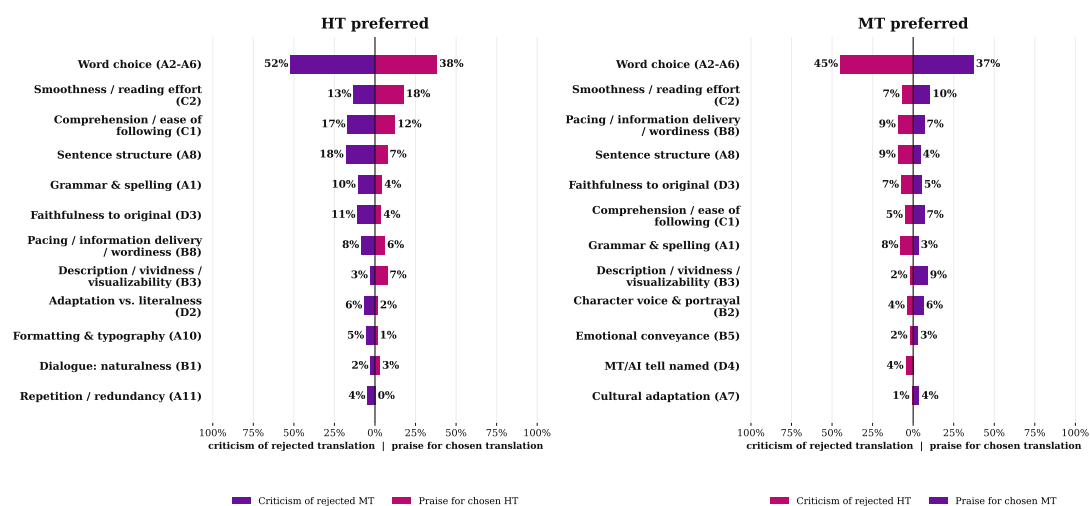
Good and poor span-level highlights. In the *close reading*, participants highlight spans they perceive as well and poorly worded. Poor evidence was more concentrated in MT chunks: 41.7% had at least 100 poor-highlighted words per 1K words, compared with 11.9% of HT chunks (Figure 25).

Post-study interviews. Participants were asked to answer several questions about their opinions on LLM translations before and after the study. The questions can be found in Table 28. We report readers' post-study responses across four tables: Table 29, Table 30, Table 31 and Table 32. The tables are split by participant group for readability.

Participants were initially skeptical about the quality of LLM translation: some, who have previously edited AI translations, recall them being of poor quality, and thus adjust their expectations accordingly, while other non-experts did not trust



(a) Excerpt-level preferences



(b) Chunk-level preferences

Figure 18: 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.

LLMs to produce a good narrative from anecdotal experience. However, most participants, even those who correctly discriminated HT from MT, were pleasantly surprised by MT's quality, despite their general preference towards HT, one participant even preferred MT over HT despite correctly identifying them! Furthermore, several participants noted that the LLM translations did not contain the usual cues that they typically associate with LLM translations and writing: one participant states that they would “*have no inkling of it being translated by AI*” if it was surrounded by human-written books in their personal book collection. This view is not universal among participants, however. Some state that despite MT texts being semantically accurate, which is the goal of

a general-purpose translation system, they write too literally, frequently miss subtle context cues in the story, mistranslate idioms or other culturally-sensitive phrases, and lack the stylistic and narrative requirements of fiction, “[*MT*] tends to flatten the author's voice.” Conversely, many participants were disappointed in the quality of HT excerpts of some of the books, either due to mistakes that should have been caught during editing or overall poor quality. The participants' general consensus was that agentic LLM translation systems are not ready for standalone publication, but may be integrated to generate first drafts to be further edited by expert human translators.



Figure 19: Close-reading chunk-level preferred translation. Each cell shows one chunk comparison for an excerpt and reader. Green indicates HT preference, purple indicates MT preference, and darker shades indicate stronger preference.

#	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 28: Post-study questionnaire. All responses are free text.

F Topic modeling support

This section describes the label-blind topic-model analysis used to validate the human inductive coding.

Model-assisted inductive coding. We use GPT-5.4 as a topic model to support the inductive coding process. We pool 922 comments from the open-ended questions about translation quality (i-Q5, i-Q6, co-Q4, and ch-Q3). We exclude 7 comments because they do not express an opinion about translation quality, leaving 915 comments. We also analyze 30 AI-detection explanations from co-Q7 separately. The model receives the participant com-

ments and the question, but no manual annotations or codebooks. It assigns each comment to one, none or more topics, and a short description. The exact prompts and run settings are available in our GitHub repository ([analysis/human_eval](#)). The resulting topic inventories are reported in Table 34.

Topic inventory selection. We repeat label discovery ten times for each inventory. Without consulting the manual labels, we match topics across runs using best-match Jaccard similarity. We then select the run whose topics are, on average, most similar to those from the other nine runs. The selected inventories contain 24 translation-quality topics and 15 AI-detection-rationale topics (mean Jaccard: 0.568 and 0.687, respectively). Using the selected topic list, we ask GPT-5.4 to independently assign topics to every comment three times. Across these runs, topics contain similar sets of comments (mean Jaccard: 0.653 for translation quality and 0.827 for AI detection), and individual comments receive similar sets of topics (0.746 and 0.814, respectively). Final assignments retain topics identified in at least two runs.

We compare these label-blind topics with the final human schemas (29 translation-quality labels

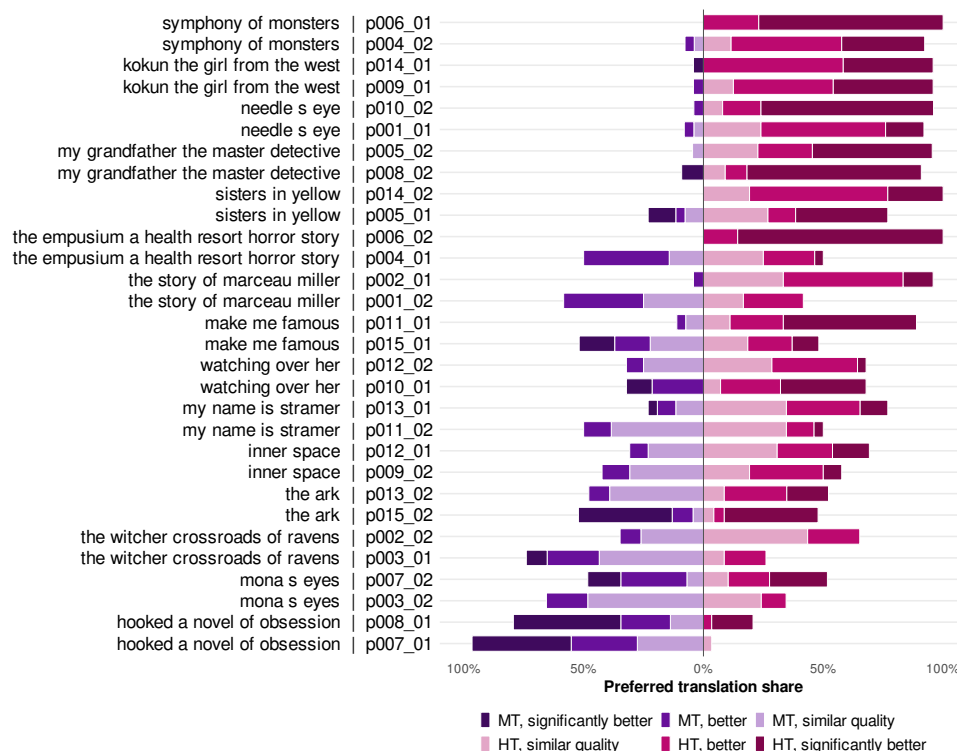


Figure 20: Preferred translation by excerpt and reader. Bars show the share of chunk-level choices favoring MT on the left and HT on the right; darker shades indicate greater preference strength.

and 10 AI-detection labels) to assess whether the themes identified through human inductive coding can be independently recovered. This provides schema-level validation, not annotator agreement.

G Exploratory multilingual case study

In this section of the appendix, we detail our case study targeting translation into languages other than English, specifically Spanish, French, Polish, and Japanese.

Setup. In order to explore whether the strong performance of MT holds beyond English, we run a small multilingual case study in which readers compared HT and MT into Spanish, French, Polish, and Japanese.³³ We translate *My Grandfather, the Master Detective* (Japanese source) into Spanish, French, and Polish, and *The Witcher: Crossroads of Ravens* (Polish source) into Japanese, choosing these two books because both have recently been professionally translated and published in the respective target languages.

³³For Spanish, French, and Polish the readers were researchers and volunteers who are native speakers of the given language; we used two readers for the Polish book. For Japanese we followed the English procedure and hired one reader via Upwork, paid \$110 USD.

MT identification. Readers reliably identified MT for Polish (two raters), Spanish, and Japanese in the final excerpt-level comparison, while the French reader was incorrect (Figure 28). We also note that the Spanish reader was initially incorrect, identifying MT as HT when they first encountered it in single-reading (Figure 26), but they were able to reliably distinguish both during comparison, suggesting that perhaps the quality of translation into French and Spanish was higher than into Polish and Japanese.

Readers identified MT through concrete translationese. Unlike in the English-facing study, where MT origin was difficult to detect, for translations into other languages the detection accuracy was better with 4 out of 5 readers identifying MT correctly and with strong average confidence (mean 4.4/5). Furthermore, the readers' explanations point to concrete issues such as translationese. One Polish reader states they were 100% sure that the MT “cannot possibly be human-translated (...) [at] around 10% of reading.” Readers flagged grammatical and structural problems, such as sentences “hanging there as if directly translated from the source, with no predicate,” as well as source language calques, including “the third person for

*Kaede when she speaks about herself.*³⁴ One recurring cue was also vocabulary, especially in cases where MT failed to use a standard term and instead supplied description “*like a student who has not mastered the vocabulary fully yet.*” For example, where the human translator in Polish wrote *nekrolog* (“obituary”), the MT wrote *prasowe artykuły pośmiertne* (“posthumous press articles”). Other times MT used a wrong or non-idiomatic term, for instance, in Spanish, *librerías de viejo* where the human wrote *librerías de segunda mano* (“second-hand bookshops”), a phrasing the reader judged “*not idiomatic in any variety of Spanish.*”

HT is strongly preferred for translations into languages other than English. The HT vs. MT contrast was much sharper than in translation into English. At the chunk level, every reader preferred HT overall, with HT favored in 103 of 111 chunk judgments (92.8%). At the excerpt level, only the French reader chose MT; however, they still rated HT highly in single-reading, with just a one-point difference between MT and HT (5 vs. 4 for acceptability, smoothness, etc.). The remaining four readers preferred HT, with single-reading ratings favoring HT most for translation into Japanese and Polish. Even the French reader changed their preference toward HT in the close-reading task (Figure 13, Figure 30, and Figure 27).

Span annotations. Across target languages, readers marked mostly positive evidence in HT and mostly negative evidence in MT (Figure 12), with median span lengths reported to indicate how localized these judgments were (Figure 11).

H Statistical analysis

In this section of the appendix, we report details on the statistical models used for the quantitative results. The statistical models were chosen based on the type of response variable and data distribution. All analyses were conducted in R using lme4 for mixed-effects models, lmerTest for inference in linear mixed-effects models, emmeans for order-adjusted marginal means and contrasts, and ordinal for cumulative-link models. The exact formulas and outputs are provided in Table 35 for preference judgments, Table 36 for span high-

light, Table 37 for Likert-type ratings, and Table 39 for AI detection.

Binary preferences. For binary choices (HT vs MT), we use logistic generalized linear mixed-effects models (GLMMs), with HT coded as 1 and MT as 0. These models estimate whether readers are more likely to choose HT than MT. When estimable, we include random intercepts for *reader* and *book* to account for repeated judgments by the same readers and repeated judgments of the same books.

For dialogue and word-choice comparisons, responses had three levels: HT, MT, and no meaningful difference. We coded these as signed attribute preferences, with HT = +1, no difference = 0, and MT = -1. Crossed mixed-effects models with *reader* and *book* random intercepts were singular, hence we report order-adjusted fixed-effect linear models.

Preference strength. For preference strength, we convert choices into signed numeric scores. HT preferences are coded as positive and MT preferences as negative, with larger absolute values indicating stronger preferences. We analyze these signed scores with linear mixed-effects models when the random-effects structure is estimable (random intercepts for *reader* and *book*).

Single-reading ratings. For 5-point Likert ratings, such as acceptability, smoothness, immersion, and willingness to continue reading, we use cumulative-link mixed models with random intercepts for *reader* and *book* when estimable. These models are designed for ordered categorical responses, where the difference between adjacent scale points should not be treated as necessarily equal. We additionally tested by-book and by-reader random slopes for translation type. Across the six tested random-effects specifications, the worst-case *p*-values were .043 for acceptability, .018 for smoothness, .199 for immersion, and .307 for willingness to continue reading. Acceptability and smoothness therefore remained significant, whereas immersion and willingness to continue remained non-significant.

AI detection. For AI-detection questions, responses were coded as binary correctness outcomes with 1 if the reader correctly identified the translation origin and 0 otherwise. In the single-reading task, MT excerpts were counted as correct when labeled machine-translated, and HT ex-

³⁴It is common in Japanese to refer to yourself in the third person but in other languages it often seems either childish or weird.

cerpts when labeled human-translated. We analyzed these responses with logistic generalized linear mixed-effects models, including random intercepts for *reader* and *book* when estimable. In the direct-comparison task, correctness meant selecting the actual MT version as more likely AI-translated; the crossed mixed-effects model was singular, so we report the corresponding intercept-only logistic model and an exact binomial test against chance. Confidence ratings were analyzed with linear mixed-effects models, restricted to correct guesses when testing whether correct responses were more confident after direct comparison.

Highlight counts. For highlighted evidence spans, we use negative-binomial mixed-effects count models. These models are appropriate for count data when the variance is larger than expected under a simple Poisson model.

Random effects. Across mixed-effects analyses, we include random intercepts for *reader* and *book* whenever estimable (i.e., does not result in a *singular fit*). If a mixed-effects model is singular, meaning that one or more random-effect variances are estimated as zero or near-zero, we report the corresponding simpler model.

I Automatic metrics preferences

In this section, we explore how well automatic metrics can capture readers’ preferences for the chunk-level judgments.

Setup. We compare chunk-level human judgments with automatic metrics. Since several metrics are constrained by context-window size, we use the *par3* aligner (Thai et al., 2022) to obtain paragraph-level source–HT–MT alignments for the evaluation data. We then score the paragraphs and map the scores to chunks containing these paragraphs. A chunk-level score was computed by taking the average of paragraph-level scores within each chunk. See §I for details.

Evaluation methods. We test the three most popular evaluation methods: METRICX-QE, COMETKIWI, and LiTRANSProQA with Gemini 3.1 Pro (Google DeepMind, 2026) as a judge. We score the paragraphs directly with METRICX-QE and LiTRANSProQA as they both can handle chunk length. We additionally score the paragraphs and then compute the average chunk score with

COMETKIWI and METRICX-QE.³⁵

Can the difference between MT and HT be captured with automatic metrics? Across the close-reading chunks, all automatic metric variants prefer MT over HT (LiTRANSProQA $\uparrow$ MT: 0.996, HT: 0.920; METRICX-QE $\downarrow$ MT: 8.347, HT: 8.657; COMETKIWI paragraph mean $\uparrow$ MT: 0.760, HT: 0.722; METRICX-QE paragraph mean $\downarrow$ MT: 5.184, HT: 5.671). Kendall τ (Kendall, 1938) correlation with human judgments at chunk-level is weak to negative for all metric variants (LiTRANSProQA $\tau = -0.124$, METRICX-QE $\tau = -0.075$, COMETKIWI PARAGRAPH MEAN $\tau = -0.318$, METRICX-QE PARAGRAPH MEAN $\tau = -0.216$), and is identical whether computed over all judgments or only those where readers agree, since clear human agreement usually favors HT while metrics favor MT.

Paragraph-level alignment. Because several automatic metrics are constrained by context-window size, we evaluate the translations on shorter aligned paragraph units rather than complete 8K-word excerpts or even 300-word chunks. We follow the *par3* alignment procedure (Thai et al., 2022). For each excerpt, source paragraphs are paired with paragraph-level Microsoft Translate output, which serves as an English pivot. Each candidate English version, including professional HT and each MT system output, is sentence-aligned to this pivot. The candidate sentences aligned to each pivot paragraph are then grouped under the original source paragraph boundaries. After alignment, we compute diagnostics for each recovered paragraph pair and manually inspect flagged cases. We flag a length issue when the larger paragraph is more than three times longer than the smaller paragraph, ignoring pairs where both sides are shorter than 10 characters. We also flag lexical-overlap issues when a candidate paragraph has a pairwise BLEU score below 1.5, against a majority of the other MT outputs aligned to the same source paragraph. We employ these heuristics to filter misaligned paragraphs.

Paragraph-level metrics. On the aligned paragraph pairs, we run two metric families: COMET and METRICX. For each family, we use one reference-based metric and one reference-free

³⁵We compute METRICX-QE over both chunk and paragraph as it is likely better adapted to shorter lengths. For LiTRANSProQA we report only chunk-level evaluation due to the cost and since the model can handle 300-word texts.

quality-estimation metric: COMET-22 and METRICX use the professional HT as reference, while COMETKIWI (Rei et al., 2022) and METRICX-QE (Juraska et al., 2024) use only the source and candidate translation. COMET scores are higher-is-better, whereas METRICX scores are lower-is-better. The metric input limits are 512 tokens for the COMET family and 1,536 tokens for the METRICX family. For reference-based metrics, this budget includes the source, candidate translation, and reference. For reference-free metrics, it only includes the source and candidate translation. Paragraph pairs that exceed the corresponding model limit are skipped for that metric. Reference-based metrics are not reported for the HT row, since this would score the reference against itself. For each book and system, we report the mean of all remaining paragraph scores for each metric.

Paragraph-level par3 automatic-evaluation results for all systems are summarized in Table 40.

Chunk-level comparison with human judgments. We evaluate the same approximately 300-word aligned chunks used in the chunk-level *close reading* task. We score the HT and agentic/P3 MT chunks directly with METRICX-QE and with LITRANSROQA (Zhang et al., 2025b), using Gemini 3.1 Pro as the question-answering judge (Google DeepMind, 2026). In addition, we map the paragraph-level COMETKIWI and METRICX-QE scores onto the close-reading chunks. This mapping matches paragraph-level source text to the chunk source text in order and averages all paragraph scores assigned to a chunk. When both HT and MT have paragraph scores for the same metric, we average over the shared paragraph indices so that the two systems are compared on the same source material. For every metric, we compute an oriented HT–MT delta: positive values favor HT, and negative values favor the agentic/P3 MT. Chunk-level HT–MT metric deltas used in the human-judgment comparison are summarized in Table 41.

Pair	Original Title	Author	Translation Title	Translator	Original Date	Translation Date
Main human evaluation: source to English						
FR→EN	<i>Le Roman de Marceau Miller</i>	Marceau Miller	<i>The Story of Marceau Miller</i>	Howard Curtis	Jan. 2025	Mar. 2026
FR→EN	<i>La Symphonie des monstres</i>	Marc Levy	<i>Symphony of Monsters</i>	Tina Kover	Oct. 2023	Jan. 2026
FR→EN	<i>Veiller sur elle</i>	Jean-Baptiste Andrea	<i>Watching Over Her</i>	Frank Wynne	Aug. 2023	Aug. 2025
FR→EN	<i>Les Yeux de Mona</i>	Thomas Schlessler	<i>Mona's Eyes</i>	Hildegard Serle	Jan. 2024	Sept. 2025
FR→EN	<i>Célèbre</i>	Maud Ventura	<i>Make Me Famous</i>	Gretchen Schmid	Aug. 2024	May 2025
JA→EN	<i>Hakobune</i>	Haruo Yuki	<i>The Ark</i>	Jim Rion	Sept. 2022	Feb. 2026
JA→EN	<i>Meitantei no Mama de Ite</i>	Masateru Konishi	<i>My Grandfather, the Master Detective</i>	Louise Heal Kawai	Jan. 2023	Mar. 2026
JA→EN	<i>Kokun, Vol. 1: Nishi kara Kita Shoji</i>	Nahoko Uehashi	<i>Kokun: The Girl from the West</i>	Cathy Hirano	Mar. 2022	Feb. 2026
JA→EN	<i>Nairu Pachi no Joshikai</i>	Asako Yuzuki	<i>Hooked: A Novel of Obsession</i>	Polly Barton	Mar. 2015	Mar. 2026
JA→EN	<i>Kiirōi Ie</i>	Mieko Kawakami	<i>Sisters in Yellow</i>	Laurel Taylor and Hitomi Yoshio	Feb. 2023	Mar. 2026
PL→EN	<i>Rozdroże kruków</i>	Andrzej Sapkowski	<i>The Witcher: Crossroads of Ravens</i>	David French	Oct. 2024	Sept. 2025
PL→EN	<i>Stacja</i>	Jakub Szamałek	<i>Inner Space</i>	Kasia Beresford	May 2023	July 2025
PL→EN	<i>Stramer</i>	Mikołaj Łoziński	<i>My Name is Stramer</i>	Antonia Lloyd-Jones	Oct. 2019	Aug. 2025
PL→EN	<i>Ucho Igielne</i>	Wiesław Myśliwski	<i>Needle's Eye</i>	Bill Johnston	Oct. 2018	Oct. 2025
PL→EN	<i>Empuzjon. Horror przyrodolecznicy</i>	Olga Tokarczuk	<i>The Empusium: A Health Resort Horror Story</i>	Antonia Lloyd-Jones	June 2022	Sept. 2024
Multilingual target-language case study						
JA→FR	<i>Meitantei no Mama de Ite</i>	Masateru Konishi	<i>Les Histoires de Kaede</i>	Mathilde Tamae-Bouhon	Jan. 2023	May 2024
JA→PL	<i>Meitantei no Mama de Ite</i>	Masateru Konishi	<i>Mój dziadek – genialny detektyw</i>	Dariusz Latoś	Jan. 2023	Sept. 2025
JA→ES	<i>Meitantei no Mama de Ite</i>	Masateru Konishi	<i>La pequeña habitación de los misterios</i>	Juan Francisco González Sánchez	Jan. 2023	Sept. 2025
PL→JA	<i>Rozdroże kruków</i>	Andrzej Sapkowski	<i>Witcher: Karasu no Jujiro</i>	Yasuko Kawano and Aya Sugiura	Oct. 2024	Sept. 2025
Development books for MT pipeline selection and ablations						
FR→EN	<i>Les guerres précieuses</i>	Perrine Tripier	<i>Our Precious Wars</i>	Alison Anderson	Jan. 2023	Nov. 2025
FR→EN	<i>Naufrage</i>	Vincent Delecroix	<i>Small Boat</i>	Helen Stevenson	Aug. 2023	Mar. 2025
FR→EN	<i>Déserteur</i>	Mathias Énard	<i>The Deserters</i>	Charlotte Mandell	Aug. 2023	May 2025
FR→EN	<i>Ilaria ou la conquête de la désobéissance</i>	Gabriella Zalapi	<i>Ilaria, or The Conquest of Disobedience</i>	Adriana Hunter	Aug. 2024	Nov. 2025
FR→EN	<i>Le vieil incendie</i>	Elisa Shua Dusapin	<i>The Old Fire</i>	Aneesa Abbas Higgins	Aug. 2023	Jan. 2026
JA→EN	<i>Jakuson Hitori</i>	Jose Ando	<i>Jackson Alone</i>	Kalau Almony	Nov. 2022	Jan. 2026
JA→EN	<i>Machi to Sono Futashika na Kabe</i>	Haruki Murakami	<i>The City and Its Uncertain Walls</i>	Philip Gabriel	Apr. 2023	Nov. 2024
JA→EN	<i>Baba Yaga no Yoru</i>	Akira Otani	<i>The Night of Baba Yaga</i>	Sam Bett	Oct. 2020	July 2024
JA→EN	<i>Kai ni Tsuzuku Bashonite</i>	Mai Ishizawa	<i>The Place of Shells</i>	Polly Barton	July 2021	Mar. 2025
JA→EN	<i>Kyūkanbi no kanojotachi</i>	Emi Yagi	<i>When the Museum Is Closed</i>	Yuki Tejima	Mar. 2023	Jan. 2026
PL→EN	<i>Powrót z gwiazd</i>	Stanisław Lem	<i>Return from the Stars</i>	Barbara Marszał	Jan. 1961	June 1989
PL→EN	<i>Księgi Jakubowe</i>	Olga Tokarczuk	<i>The Books of Jacob</i>	Jennifer Croft	Oct. 2014	Nov. 2021
PL→EN	<i>Lód</i>	Jacek Dukaj	<i>Ice</i>	Ursula Phillips	Jan. 2007	Nov. 2025
PL→EN	<i>Mgła</i>	Kaja Malanowska	<i>Fog</i>	Bill Johnston	Jan. 2015	Mar. 2021
PL→EN	<i>Lakome</i>	Małgorzata Lebda	<i>Voracious</i>	Antonia Lloyd-Jones	Aug. 2023	Jan. 2025
PL→EN	<i>Poufne</i>	Mikołaj Grynberg	<i>Confidential</i>	Sean Gasper Bye	June 2020	Jan. 2025

Table 3: Books used in the human evaluation, multilingual target-language case study, and MT-pipeline development experiments. 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.

			Excerpts 📖	Chunks 📦	Participant Comments 🗣️					
			(n=15)	(n=386)		(n=60)	(n=60)	(n=30)	(n=30)	(n=772)
					#/Ex	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 4: Summary statistics for evaluation excerpts, chunks, and participant free-text comments, split into a **WORDS** sub-table (whitespace-delimited word counts; we do not report **SRC** word counts because Japanese cannot be split on whitespaces) and a **TOKENS** sub-table (tiktoken o200k_base token counts). Excerpts and chunks are reported separately for human-translated (**HT**) and machine-translated (**MT**) versions in the **WORDS** sub-table; the **TOKENS** sub-table also includes source-language (**SRC**) versions. The #/EX (chunks per excerpt) 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.

Isolated-reading: by book ratings contrasted with comparative preferences

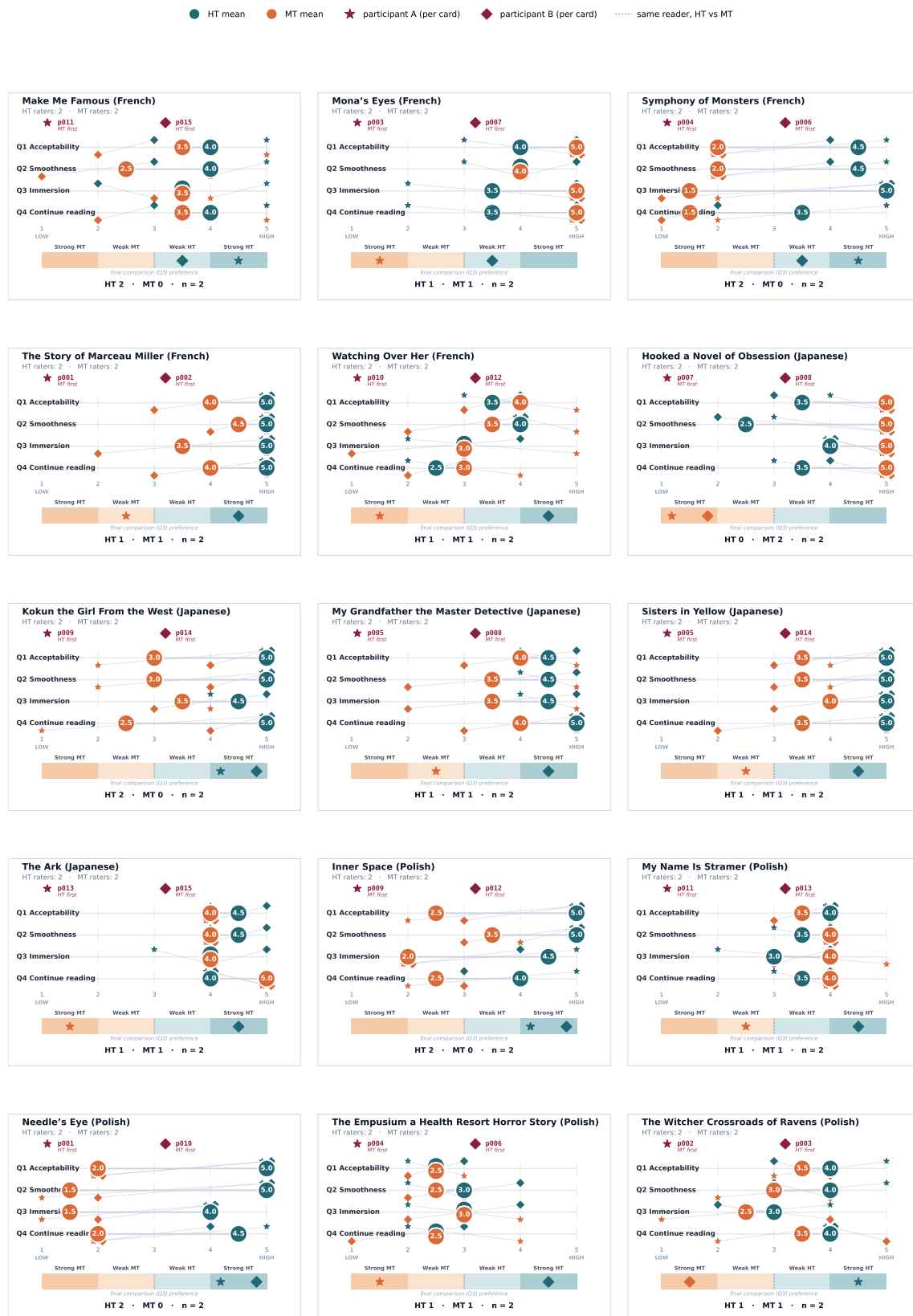


Figure 21: Per-book immersive-reading ratings for HT and MT. The figure breaks down participant ratings by book, making cross-book variation visible alongside the aggregate trends reported in the main paper.

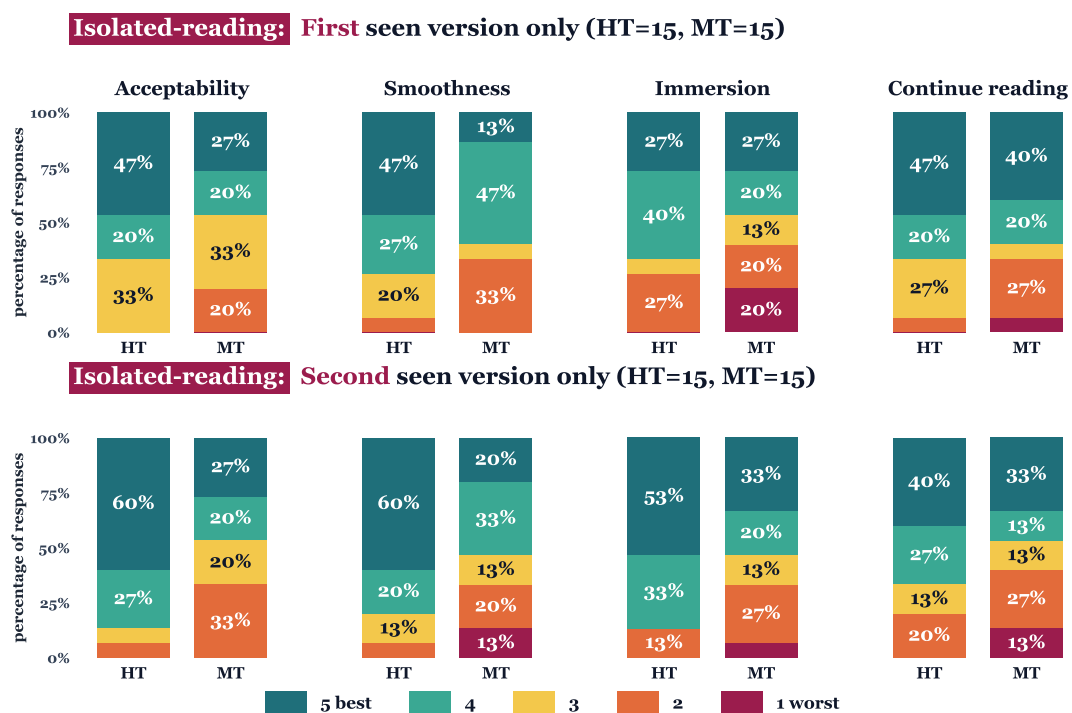


Figure 22: Immersive-reading ratings by presentation order. The figure compares ratings assigned after the first and second single readings, grouped by whether the human translation (HT) or machine translation (MT) was shown first.

Pipeline step	Agent backend	Checkpoint	Runs
Style Analysis	Claude	claude-opus-4-6	35
Initial Chunk Translation	GPT	gpt-5.4	31
Translation Quality Review	GPT	gpt-5.4	35
Literary Quality Review	Claude	claude-opus-4-6	35
Targeted Chunk Revision	GPT	gpt-5.4	31
Targeted Chunk Revision (after 3 failed revisions)	Claude	claude-opus-4-6	4
Whole Excerpt Review	GPT	gpt-5.4	35
Cross-Chunk Continuity Check	GPT	gpt-5.4	35
Global Revision	Claude	claude-opus-4-6	35

Table 5: 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.

	Pipeline	Model	TOTAL	MEAN	MEDIAN	ST. DEV.	MIN	MAX
WORDS	Pipeline 1	Gemini	79,232	4,952.0	4,019	2,397.4	2,508	9,272
		GPT-5.4	78,920	4,932.5	3,970	2,383.9	2,498	9,218
	Pipeline 2	Gemini	80,614	5,038.4	4,142	2,442.1	2,534	9,379
		GPT-5.4	79,872	4,992.0	4,058	2,448.5	2,518	9,378
	Pipeline 3	Agents	79,851	4,990.7	4,057	2,467.6	2,476	9,012
TOKENS	Pipeline 1	Gemini	100,622	6,288.9	5,042	3,109.0	3,150	11,974
		GPT-5.4	99,219	6,201.2	4,885	3,034.6	3,223	11,830
	Pipeline 2	Gemini	102,331	6,395.7	5,185	3,180.4	3,204	12,171
		GPT-5.4	100,270	6,266.9	5,012	3,102.4	3,206	11,938
	Pipeline 3	Agents	99,640	6,227.5	5,012	3,124.3	3,145	11,670

Table 6: 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 summarize 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	16
Run group	Case study evaluation books	4
Source language	French	10
Source language	Japanese	13
Source language	Polish	12
Target language	English	31
Target language	French	1
Target language	Japanese	1
Target language	Polish	1
Target language	Spanish	1
Final Acceptance Gate	FAIL	22
Final Acceptance Gate	PASS	13
Max chunk cycles	2	18
Max chunk cycles	3	17

Table 7: Corpus and run overview for the completed agentic pipeline runs.

Artifact / log unit	Count	Scope
Completed pipeline runs	35	Evaluation, development, and multilingual target-language runs used in this paper.
Source-language chunks processed	441	All selected chunks entering the agentic pipeline.
Chunk-cycle translation/revision artifacts	1,059	Source chunks $\times$ completed chunk cycles.
Runs with job-metric logs	31	Excludes four development-book runs with unavailable metrics.
Logged agent jobs	2,930	All recorded style, translation, review, revision, and continuity jobs.
Failed agent jobs	19	Logged failed jobs across runs with metrics.
Final-gate flagged chunks	74	Chunks flagged by the final acceptance gate.

Table 8: Artifact counts generated by the selected agentic pipeline runs.

CORRECT	INCORRECT
Word choice (M2) 8/17	Word choice (M2) 9/13
Sentence flow, structure & grammar (M3) 7/17	Folk theory invoked (M10) 5/13
Formatting & surface artifacts (M4) 6/17	Voice & emotional life (M7) 5/13
Literalism / source-bleed (M1) 6/17	Sentence flow, structure & grammar (M3) 4/13
Folk theory invoked (M10) 5/17	Meaning errors (M8) 4/13
Meaning errors (M8) 3/17	Wrong tone or mixed register (M5) 3/13
Wrong tone or mixed register (M5) 3/17	Literalism / source-bleed (M1) 3/13
Wordiness / over-explaining (M6) 3/17	Comprehension difficulty (M9) 2/13
Voice & emotional life (M7) 2/17	Formatting & surface artifacts (M4) 1/13

Figure 23: 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.

ID	Age	Gender	Education	Nationality	Native language	Additional languages	English literary fiction	Translated fiction in English	Books per year
1	25–34	Woman	Undergraduate	American	English	None	Weekly or more	Almost never	100
2	25–34	Woman	Undergraduate	American	English	None	Weekly or more	A few times a year	40
3	18–24	Man	Undergraduate	South African	English	None	A few times a month	Not sure	12
4	45+	Woman	Undergraduate	South African	English	None	About once a month	A few times a year	12
5	45+	Woman	High school	American	English	None	Weekly or more	Almost never	12
6	45+	Woman	Master's	South African	English	None	Weekly or more	A few times a year	50
7	35–44	Woman	Doctoral	Canadian	English	French	About once a month	Almost never	24
8	45+	Woman	Undergraduate	American	English	None	Weekly or more	A few times a month	75
9	25–34	Woman	Undergraduate	South African	English	None	Weekly or more	A few times a year	120
10	25–34	Woman	Undergraduate	American	Spanish	None	Weekly or more	A few times a year	35
11	25–34	Woman	Undergraduate	American	English	None	Weekly or more	A few times a year	100
12	45+	Woman	High school	American	English	None	Weekly or more	Almost never	103
13	45+	Woman	Undergraduate	British	English	French	A few times a month	About once a month	20
14	35–44	Woman	Master's	Polish	Polish	Polish; French	Weekly or more	Weekly or more	80
15	45+	Woman	Doctoral	American	English	None	Weekly or more	About once a month	100

Table 9: Participant demographics and reading profile. All participants reported that they primarily read in English and were very comfortable reading long literary texts in English. Nationality entries are shortened for display.

ID	AI-use frequency	Models used most often	Reported uses
1	Sometimes	ChatGPT	Deep research and summarization
2	Rarely	Google	Does not use AI
3	Often	ChatGPT; Grok Imagine; Nano Banana	Deep research and summarization; job search and career support; image generation
4	Often	ChatGPT	Writing and editing
5	Often	ChatGPT	Web search; personal advice; job search and career support
6	Often	ChatGPT; Grok	Web search; image generation
7	Often	Claude Opus; Claude Sonnet	Deep research and summarization; writing and editing
8	Often	Claude; ChatGPT; Grok	Deep research and summarization; writing and editing; translation
9	Always	ChatGPT; Claude; Kiro	Deep research and summarization; writing and editing
10	Sometimes	Gemini	Deep research and summarization; writing and editing
11	Never	None	Does not use AI
12	Always	Claude Sonnet 4.6; ChatGPT 5.3	Deep research and summarization; writing and editing
13	Never	None	Does not use AI
14	Rarely	ChatGPT	Web search
15	Always	ChatGPT; Claude; DeepSeek	Deep research and summarization; writing and editing

Table 10: Self-reported AI use.

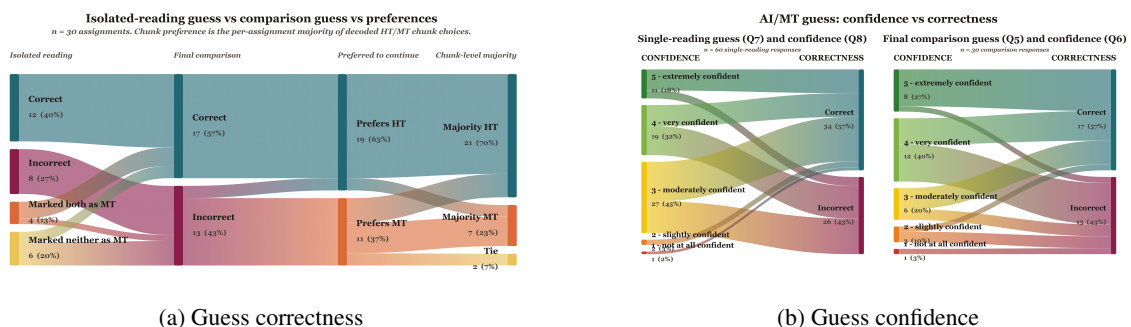


Figure 24: Translation-origin guess flows after the comparison task. The first panel summarizes whether guesses were correct for HT and MT; the second panel shows reported confidence across those identification decisions.

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].

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.**

Table 11: Consent form presented to the participants prior to the evaluation.

Consent Form Questions	
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 (<i>Required</i>)	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 (<i>Optional</i>)	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 (<i>Required</i>)	Open-ended response
Date (<i>Required</i>)	Date input field

Table 12: Consent-form response fields shown to participants.

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?	18-24; 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 13: Entry questionnaire presented to the participants prior to the evaluation.

#	Question	Response Options
1	How acceptable did this translation feel overall as a published version of the book? <i>Please think about your overall impression of this version as a published translation.</i>	completely unacceptable somewhat unacceptable acceptable with some issues mostly acceptable fully acceptable
2	How smooth was this text to read? <i>Please think about how naturally and clearly the sentences flowed while you were reading.</i>	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? <i>Please think about whether the wording and style helped you stay absorbed in the story, or pulled you out of it.</i>	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? <i>If you were to continue reading, would you be willing to continue this version given its wording and style?</i>	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? <i>(Regarding Question 7)</i>	not at all confident slightly confident moderately confident very confident extremely confident

Table 14: Single-reading questionnaire shown after reading each excerpt in isolation. 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? <i>[Dialogue]</i>	Translation 1 Translation 2 No meaningful difference
2	Which translation's word choice felt more varied and/or expressive? <i>[Word choice]</i>	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. <i>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.</i>	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? <i>(Regarding Question 5)</i>	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? <i>What are the indicators that make you think so?</i>	Open-ended response

Table 15: 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 <i>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.</i>	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 16: 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.

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 17: MT-identification accuracy by AI-use frequency. Participant-level association between AI-use frequency and MT-identification accuracy was close to zero: Pearson $r = 0.013$; Spearman $\rho = 0.121$.

SUBSET	<i>n</i>	EXACT	≤ 1	$M d $	AC2
Q1-Q4 overall	120	35.0	64.2	1.16	0.426
	60	38.3	70.0	0.98	0.616
HT only					
MT only	60	31.7	58.3	1.33	0.216
Q1 acceptability	30	36.7	73.3	0.93	0.632
	15	40.0	80.0	0.80	0.765
HT only					
MT only	15	33.3	66.7	1.07	0.468
Q2 smoothness	30	36.7	70.0	1.03	0.529
	15	40.0	73.3	0.87	0.725
HT only					
MT only	15	33.3	66.7	1.20	0.290
Q3 immersion	30	30.0	53.3	1.40	0.229
	15	33.3	53.3	1.27	0.405
HT only					
MT only	15	26.7	53.3	1.53	0.057
Q4 continue	30	36.7	60.0	1.27	0.317
	15	40.0	73.3	1.00	0.545
HT only					
MT only	15	33.3	46.7	1.53	0.089

Table 18: Inter-annotator agreement for single-reading ratings. Exact and ≤ 1 are agreement percentages; $M|d|$ is mean absolute score difference; AC2 is Gwet’s ordinal agreement coefficient.²⁷

MEASURE	<i>n</i>	EXACT	BOTH HT	BOTH MT	DIFF.	AC1
Excerpt: guessed MT	15	40.0	2	4	9	-0.179
Excerpt: preferred	15	40.0	5	1	9	-0.120
Chunk: preferred	386	63.7	191	55	140	0.355
excluding similar	194	72.2	116	24	54	0.546
chunks						
both significantly better	44	79.5	28	7	9	0.667

Table 19: Inter-annotator agreement for excerpt- and chunk-level binary judgments.²⁸

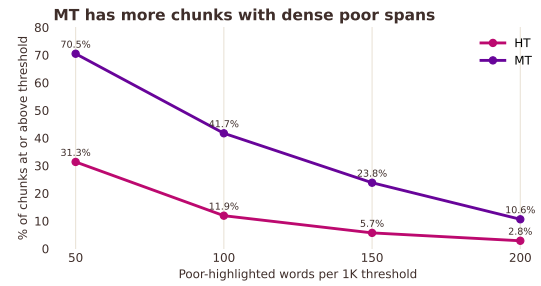


Figure 25: Share of close-reading chunks with dense poor-span evidence. For each threshold, the curve shows the percentage of HT and MT chunks with at least that many poor-highlighted words per 1K words. MT has substantially more high-density poor-span chunks.

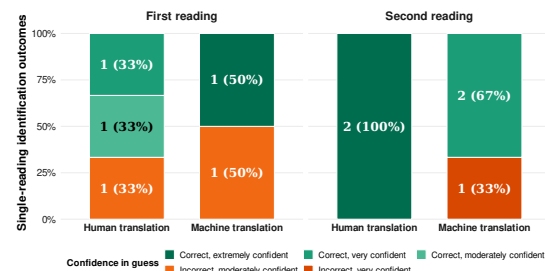


Figure 26: Multilingual case-study identification outcomes. Single-reading results for HT/MT identification in the multilingual case study; darker shades indicate higher confidence.

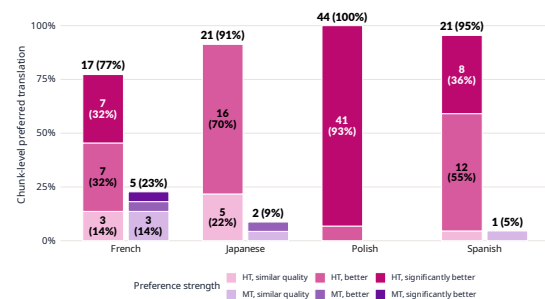


Figure 27: Close-reading preferences in the multilingual case study. Stacked bars show chunk-level preferred translation choices for HT and MT across the four evaluated translation directions, with color intensity indicating preference strength.

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	JA	29	75.9%	0.694	19 _(9/4)	73.7%	0.659
Kokun: The Girl From The West	JA	24	91.7%	0.909	21 _(3/0)	90.5%	0.895
My Grandfather, The Master Detective	JA	22	95.5%	0.948	15 _(6/2)	93.3%	0.929
Sisters In Yellow	JA	26	76.9%	0.710	14 _(9/5)	71.4%	0.622
The Ark	JA	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 20: Chunk-level preference agreement by book.

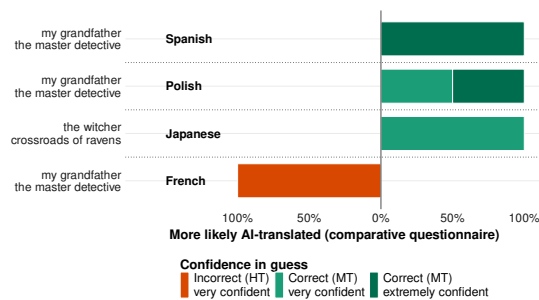


Figure 28: More likely AI-translated choices in the comparative questionnaire. After both readings, readers chose which version seemed more likely AI-translated. Orange bars show incorrect HT choices; green bars show correct MT choices, with darker shades indicating higher confidence.

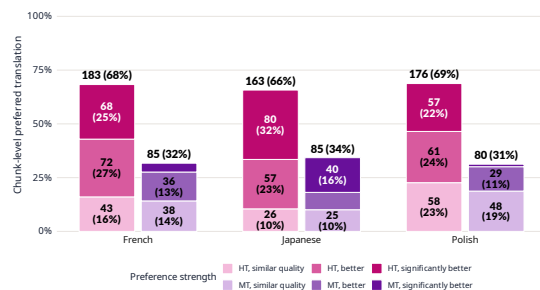


Figure 29: Close-reading chunk-level preferred translation by source language. Each pair of bars shows HT and MT preference shares within a source language; stacked segments show preference strength, and labels report counts and percentages.

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 intended.	+/-
	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 understandable, 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, choppy, 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 plausibility in speech (of this character), believable phrasing etc.	+/-
	B2	Character voice & portrayal	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.	+/-
	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	Reader explicitly articulates a specific 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 perceived as tells for AI-generated text.	+/-

Table 21: 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.

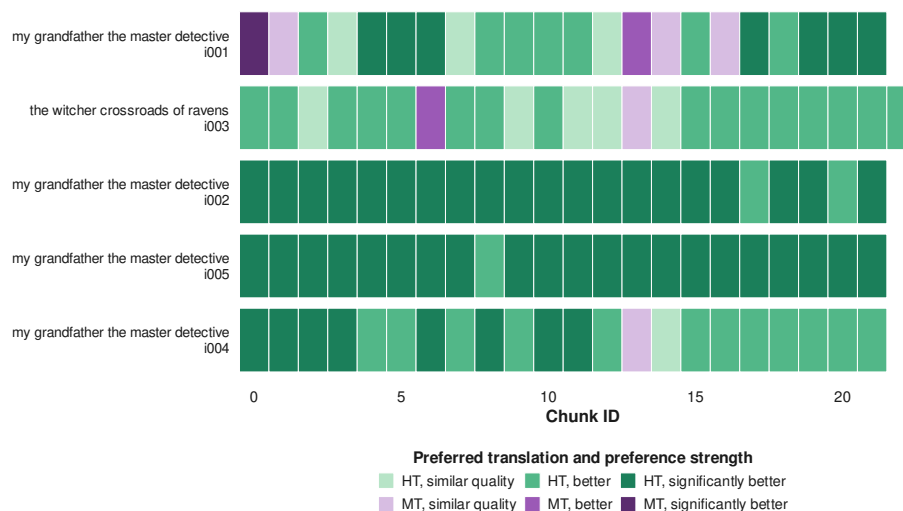


Figure 30: Chunk-level preferred translation in the multilingual target-language case study. Each row shows one excerpt-reader pair; green indicates HT preference, purple indicates MT preference, and darker shades indicate stronger preference.

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, structure & grammar	How sentences are shaped and how they connect: choppy, truncated, run-on, abrupt, copy-paste structure, sentences that do not link, and grammatical mechanics such as tense inconsistency, verb agreement errors, modal misuse, or overuse of grammatical forms.	some things flow more smoothly in T1; more variation in sentence lengths
M4	Formatting & surface artifacts	Visible page-level mechanics not folk-theory framed: punctuation, scene dividers, quotation marks, dialogue formatting, and encoding glitches. Em-dashes as an "AI sign" are coded as M10.	slightly better formatting with proper nouns
M5	Wrong tone or mixed register	Whole passage is in a wrong tone for a novel, such as manual, textbook, technical, or Wikipedia-mode prose, or register switches and mixes inside the passage, such as US/UK slang together, source-culture and target-culture mixing, or formal-informal swings.	older style consistent with fantasy
M6	Wordiness / over-explaining	Text uses more words than needed: padding, extra words, info-dumping, or lack of economy. Different from M5; wordiness can occur in any mode.	economical; tight prose
M7	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.	more natural and conveys more emotion; made me feel like a sassy idol
M8	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.	T2 "breathed through his mouth" more logical; T2 treats speaking underwater as hypothetical rather than given
M9	Comprehension difficulty	Reader could not follow, had to reread, needed the other version, or was confused by a specific passage. About reader experience, not text defect.	T2 much easier to understand; version of the sentence clearer in T2
M10	Folk theory invoked	Reader explicitly references what AI does, such as a named stereotype, prior experience, or capability claim. Triggers on language like "classic sign of AI," "AI tends to," "I'd expect from AI," or "beyond what AI is capable of."	Human translators manage fluidity AI does not match; would not think AI capable of that leap

Table 22: Annotation scheme for participant comments on Compar. Q7 (justifying decision for which translation is perceived as machine translation).

TOP LABELS	DEFINITIONS	REPRESENTATIVE PARTICIPANT COMMENT
Top 3 reasons HT was chosen EXCERPT COMPARISON 1. 78.9% Smoothness 2. 42.1% Comprehension 3. 31.6% 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, <i>Needle's Eye</i> (Bill Johnson, Trans.)
Top 3 reasons MT was chosen EXCERPT COMPARISON 1. 45.5% Enjoyment 2. 45.5% Smoothness 3. 45.5% Engagement	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 Yuki, <i>The Ark</i> (Jim Rion, Trans.)
Top 3 reasons HT was chosen CHUNK COMPARISON 1. 17.6% Smoothness 2. 16.7% Word choice: clarity 3. 15.3% 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, <i>Symphony Of Monsters</i> (Tina Kover, Trans.)
Top 3 reasons MT was chosen CHUNK COMPARISON 1. 14.0% Word choice: naturalness 2. 11.6% Word choice: register 3. 10.4% Smoothness	Word choice: naturalness: 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, <i>My Name Is Stramer</i> (Antonia Lloyd-Jones, Trans.)

Table 23: 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.

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	C2. Smoothness / reading effort	26	"MT was a better translation as it was generally easier to read with a better natural flow."
4	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.'"
5	B3. 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."
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	B8. Pacing / information delivery / wordiness	17	"MT has a tighter feel that comes across as more of a conversation than HT does."
8	B2. Character voice & portrayal	16	"I overall liked the wording of MT better because it sounded to me more like how the character of Sarah, who is the wife of an author, might speak."
9	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."
10	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."

Table 24: Top Q5 POSITIVE labels assigned to MT in the chunk-level comparison task. Counts are response-level label presences.

Rank	POSITIVE label assigned to HT	Count	Example
1	C2. Smoothness / reading effort	92	“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.”
2	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.’”
3	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.”
4	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.”
5	A8. Sentence structure	39	“HT had a better choice of splitting it up into smaller sentences which make it easier to read.”
6	B3. 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	B8. 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.”
8	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.”
9	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.’”
10	A1. Grammar & spelling	22	“HT has a much more natural use of grammar and punctuation. I struggled to understand how these two sentences are connected when I initially read them in MT.”

Table 25: Top Q5 POSITIVE labels assigned to HT in the chunk-level comparison task. Counts are response-level label presences.

Rank	POSITIVE label assigned to MT	Count	Example
1	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."
2	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."
3	C5. 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."
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	B2. Character voice & portrayal	3	"I felt like I was able to pick up the essence of the characters more in MT, for instance, the difference between Sarah thinking 'He melts me, the idiot' and 'I'm touched.' MT just gave a clearer picture of personalities and relationships with its words."
7	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'."
8	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."
9	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.'"
10	A7. Cultural adaptation	2	"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 as someone with very little knowledge of Jewish teachings."

Table 26: Top Q3 POSITIVE labels assigned to MT in the Part 1 comparison questionnaire. Counts are response-level label presences.

Rank	POSITIVE label assigned to HT	Count	Example
1	C2. Smoothness / reading effort	15	"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	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."
4	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."
5	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."
6	B6. 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."
7	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."
8	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."
9	C4. 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."
10	D1. Doesn't read as translated	2	"HT also felt more like reading a story in English rather than a translated story. MT definitely read like someone or something translating word by word; it did not feel natural while reading."

Table 27: Top Q3 POSITIVE labels assigned to HT in the Part 1 comparison questionnaire. Counts are response-level label presences.

P1	Identified Book 1: yes 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	Identified Book 1: yes 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 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	Identified Book 1: no 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	Identified Book 1: no 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

Table 29: Post-study responses grouped by participant (part 1 of 4; participants P1–P4). Q1 asks about expectations for AI translation quality before the study; Q2 asks whether participants were surprised by their AI-identification accuracy; Q3 asks for their post-study view of AI translation quality; Q4 collects other comments.

P5	Identified Book 1: no 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	Identified Book 1: yes 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	Identified Book 1: no 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 squirrely 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 translatable 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.

Table 30: Post-study responses grouped by participant (part 2 of 4; participants P5–P7). Questions are as in Table 29.

P8	Identified Book 1: no 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.
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	Identified Book 1: yes 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	Identified Book 1: no 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	Identified Book 1: yes 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 [<i>face with tears of joy emoji</i>]

Table 31: Post-study responses grouped by participant (part 3 of 4; participants P8–P11). Questions are as in Table 29.

P12	Identified Book 1: yes 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.
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	Identified Book 1: yes 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	Identified Book 1: yes Identified Book 2: yes 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	Identified Book 1: yes 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 32: Post-study responses grouped by participant (part 4 of 4; participants P12–P15). Questions are as in Table 29.

Level	Book	Preferred	Participant comment
Excerpt	<i>Needle's Eye</i>	HT	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, and while HT also lacked some context, it was just due to the story being short and its style. With MT, it was difficult to follow because it switched around, sometimes using quotation marks and sometimes using dashes. It did not have a flow to it.
Excerpt	<i>The Story of Marceau Miller</i>	HT	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. HT also felt more like reading a story in English rather than a translated story. MT definitely read like someone or something translating word by word; it didn't feel natural while reading.
Excerpt	<i>The Story of Marceau Miller</i>	MT	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. I felt like I was able to pick up the essence of the characters more in MT, for instance, the difference between Sarah thinking "He melts me, the idiot" and "I'm touched." MT just gave a clearer picture of personalities and relationships with its words, in my opinion.
Excerpt	<i>The Witcher: Cross-roads of Ravens</i>	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	<i>Kokun: The Girl from the West</i>	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	<i>Sisters in Yellow</i> , 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	<i>The Empusium</i> , 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	<i>The Witcher: Cross-roads of Ravens</i> , 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	<i>My Name Is Stramer</i> , chunk 10	MT	I could imagine the stories better in MT with more vivid language and I liked the phrase "windmill his arms" which made the wording come alive.
Chunk	<i>Mona's Eyes</i> , 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	<i>Kokun: The Girl from the West</i> , 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 33: Examples of contextual reader judgments in excerpt-level and chunk-level preference explanations. These examples include the full comments behind selected snippets quoted in the main text and 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.

Translation quality				AI detection			
Count	<i>N</i>	Stab.	GPT-5.4-generated topic	Count	<i>N</i>	Stab.	GPT-5.4-generated topic
1	361	0.78	Word choice and diction fit	1	10	0.65	Lexical fit
2	226	0.67	Clarity and ease of understanding	2	9	0.74	Source-text literalness
3	210	0.75	Flow and readability	3	9	0.93	Sentence and paragraph flow
4	172	0.69	Sentence structure and rhythm	4	9	0.86	Punctuation and dialogue formatting
5	170	0.66	Naturalness and idiomatic English	5	9	0.70	Semantic and contextual resolution
6	114	0.60	Meaning accuracy	6	6	1.00	Register and diction level
7	98	0.59	Dialogue authenticity	7	5	0.64	Sentence-level correctness
8	97	0.73	Verbosity and concision	8	5	0.77	Emotional and tonal conveyance
9	92	0.67	Grammar and surface mechanics	9	4	0.72	Adaptive rephrasing and localization
10	76	0.68	Imagery and descriptive vividness	10	4	1.00	Lexical richness and specificity
11	65	0.51	Literary richness and atmosphere	11	3	1.00	Information density and verbal economy
12	57	0.66	Formatting and presentation	12	3	0.78	Variety and register consistency
13	53	0.73	Cultural and contextual mediation	13	3	0.83	Formatting artifact handling
14	53	0.81	Engagement and immersion	14	3	0.78	Voice differentiation
15	51	0.79	Machine-like or translationese feel	15	2	1.00	Terminology and naming consistency
16	47	0.74	Character voice and viewpoint fit				
17	44	0.56	Coherence, transitions, and organization				
18	42	0.57	Emotional resonance and nuance				
19	40	0.72	Literalness versus adaptation				
20	37	0.73	Genre, setting, and period appropriateness				
21	35	0.47	Tense and grammatical perspective consistency				
22	29	0.61	Referent and pronoun clarity				
23	28	0.64	Consistency of recurring terms and details				
24	25	0.34	Register, tone, and stylistic fit				

Table 34: GPT-5.4-generated candidate topics for translation-quality comments (*i-Q5*, *i-Q6*, *co-Q4*, and *ch-Q3*) and AI-detection explanations (*co-Q7*). *N* is the number of comments assigned to the topic by majority vote across three runs. Stab. is the mean pairwise Jaccard similarity of the topic’s comment membership across those runs. Complete topic definitions and evidence phrases are available in the GitHub repository.

COUNTS	OUTCOME CODING	EFFECT SIZE [95% CI]	<i>p</i>	ORDER <i>p</i>
excerpt_preference	glm(prefer_HT ~ order, family = binomial, data = excerpt)			
19/11	prefer_HT: HT=1, MT=0	HT probability = 63.4%; OR = 1.73 [0.82, 3.65]	.148	.705
chunk_preference	glmer(prefer_HT ~ order + (1 reader) + (1 book), family = binomial, data = chunks)			
522/250	prefer_HT: HT=1, MT=0	HT probability = 76.2%; OR = 3.20 [1.31, 7.82]	.011	.536
chunk_preference (primary)	glmer(prefer_HT ~ order + (1 reader) + (1 book) + (1 reader:book), family = binomial, data = chunks)			
522/250	prefer_HT: HT=1, MT=0	OR = 3.28 [1.41, 7.59]	.0056	.327
excerpt_strength	lm(signed_strength ~ order, data = excerpt)			
19/11	signed score: HT positive, MT negative; max · = 2	signed mean = 0.57 [-0.10, 1.23]	.092	.612
chunk_strength	lmer(signed_strength ~ order + (1 reader) + (1 book), data = chunks)			
522/250	signed score: HT positive, MT negative; max · = 3	signed mean = 0.91 [0.24, 1.58]	.010	.176
dialogue	lm(signed_attribute ~ order, data = excerpt)			
17/9/4	signed score: HT=1, no difference=0, MT=-1	signed mean = 0.27	.109	.109
word_choice	lm(signed_attribute ~ order, data = excerpt)			
18/10/2	signed score: HT=1, no difference=0, MT=-1	signed mean = 0.27	.131	.253

Table 35: Statistical analyses for HT-vs-MT preferences and excerpt-level attribute preferences. Counts are HT/MT for binary and strength analyses, and HT/MT/no-difference for dialogue and word_choice. Binary models code HT as 1 and MT as 0. Strength and attribute models use signed scores, with HT coded positive and MT coded negative. Excerpt-level mixed-effects models were singular, so simpler order-adjusted fixed-effect models are reported for excerpt outcomes; chunk-level models include random intercepts for reader and book. Effects are order-adjusted; brackets give 95% confidence intervals.

HT COUNT	MT COUNT	EFFECT SIZE [95% CI]	<i>p</i>	ORDER <i>p</i>
positive_spans	glmer.nb(count ~ version + order + (1 reader) + (1 book), data = highlights)			
2099	1260	HT/MT rate ratio = 1.69 [1.55, 1.85]	<.001	.596
negative_spans	glmer.nb(count ~ version + order + (1 reader) + (1 book), data = highlights)			
1147	2728	HT/MT rate ratio = 0.43 [0.39, 0.48]	<.001	.052

Table 36: Chunk-level highlight count statistical analysis. Positive and negative evidence spans were modeled separately with negative-binomial mixed-effects count models with random intercepts for *reader* and *book*. Rate ratios compare HT to MT; values above 1 indicate more spans on HT, and values below 1 indicate fewer spans on HT.

RESPONSE	PREDICTOR	ESTIMATE	SE	<i>z</i>	<i>p</i>	OR [95% CI]	HT ADVANTAGE
acceptability	clmm(q1 ~ type + order + (1 reader) + (1 book), data = model_data, link = "logit")						
	typeMT	-1.38	0.51	-2.70	.0069	0.25 [0.09, 0.68]	4.0× odds
	orderMT-first	0.41	0.49	0.85	.3978	1.51 [0.58, 3.90]	–
smoothness	clm(q2 ~ type + order, data = model_data, link = "logit")						
	typeMT	-1.49	0.50	-2.98	.0029	0.23 [0.08, 0.59]	4.3× odds
	orderMT-first	0.30	0.48	0.63	.5279	1.36 [0.53, 3.51]	–
immersion	clm(q3 ~ type + order, data = model_data, link = "logit")						
	typeMT	-0.83	0.48	-1.74	.0815	0.44 [0.17, 1.10]	2.3× odds
	orderMT-first	0.27	0.47	0.56	.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	.1150	0.45 [0.17, 1.21]	2.2× odds
	orderMT-first	0.07	0.49	0.14	.8860	1.07 [0.41, 2.79]	–

Table 37: Cumulative-link model results for single-reading 5-point ratings ($n = 60$ per response, 30 HT and 30 MT). HT is the reference for type; odds ratios below 1 for typeMT indicate lower odds of a higher rating for MT than HT. HT ADVANTAGE reports the inverse odds ratio for translation effects only.

RESPONSE	<i>p</i> for typeMT by random-effects specification					
	PRIMARY	+B SLOPE	+R SLOPE	+B SLOPE ⁰	+R SLOPE ⁰	MAXIMAL [†]
Acceptability	.0069	.0382	.0139	.0154	.0121	.0428
Smoothness	.0029	.0180	.0029	.0044	.0029	.0180
Immersion	.0815	.1196	.1461	.1076	.1279	.1988
Continue reading	.1150	.1883	.2764	.1732	.2726	.3072

Table 38: Sensitivity of the translation-type effect to random-effects structure (cumulative-link models; slopes on book (b) or reader (r), with ⁰ marking zero-correlation variants). All slope-bearing fits are at the boundary (intercept-slope correlation = -1); fixed effects are reported at face value following the reduction protocol of Barr (2013). For smoothness, the reader-slope fit reduces to the fallback model. [†]Saturated maximal model (60 random effects for 60 observations).

ANALYSIS	MODEL	DATA	EFFECT	MODEL <i>p</i>	EXACT <i>p</i>
Single-reading origin accuracy	glmer(correct ~ 1 + (1 person_id) + (1 book_id), family = binomial)	34/60 correct	accuracy = 58.3%	.413	.366
Single-reading HT vs. MT accuracy	glmer(correct ~ version + (1 person_id) + (1 book_id), family = binomial)	MT: 16/30; HT: 18/30	no version difference	.557	–
Direct-comparison origin accuracy	glm(correct ~ 1, family = binomial)	17/30 correct	accuracy = 56.7%	.467	.585
Correct-guess confidence	lmer(confidence ~ stage + (1 person_id) + (1 book_id))	51 correct guesses	comparison–single = 0.40	.129	–

Table 39: Statistical analysis for AI detection. In single reading, a correct response means labeling an MT excerpt as machine-translated or an HT excerpt as human-translated. In direct comparison, a correct response means selecting the actual MT version as more likely AI-translated. The requested direct-comparison crossed mixed-effects model was singular, so the intercept-only logistic model is reported; the exact binomial test is included because the comparison is against chance.

SYSTEM			AUTOMATIC METRIC			
Model	Pipeline	Human eval.	COMET-22 $\uparrow$	COMETKiwi $\uparrow$	MetricX $\downarrow$	MetricX-QE $\downarrow$
Gemini	P2		0.805	0.773	5.182	4.758
Gemini	P1		0.804	0.772	5.243	4.813
GPT-5.4	P2		0.800	0.768	5.482	5.020
GPT-5.4	P1		0.798	0.766	5.542	5.073
Agents	P3	✓	0.793	0.766	5.656	5.109
HT	–	✓	–	0.726	–	5.726

Table 40: 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
LiTransProQA	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 41: Automatic evaluation of the aligned 300-word human-evaluation chunks: professional HT vs. agentic (P3) MT. Positive oriented deltas favor HT; negative deltas favor agentic/P3 MT. Paragraph-mean rows use mapped par3 paragraph scores.