

Emotional Arc Preservation in LLM Literary Translation

Stanford CS224N Custom Project

Chloe Murdoch

Department of Computer Science
Stanford University
murdochc@stanford.edu

Esi Eneinyang

Department of Computer Science
Stanford University
esi@stanford.edu

Julia Rhee

Department of Computer Science
Stanford University
jrhee25@stanford.edu

Abstract

We investigate whether LLM-generated literary translations preserve the emotional arc of the original text, and whether they do so more faithfully than published human translations. We build a cross-lingual sentiment pipeline using XLM-RoBERTa to extract 100-point emotional trajectories from four English novels and their French translations (both human and LLM-generated). We compare three LLM translation strategies: a baseline parallel translation, a calibrated temperature sweep that targets sentiment magnitude, and a best-of- N shape-guided reranking that targets trajectory correlation. Across three of four novels, all LLM methods outperform the human translation in arc shape fidelity (Pearson r), with the best-of-5 method achieving $r = 0.82$ vs. human $r = 0.24$ on *Jekyll & Hyde*. Prompt-level interventions yield only marginal gains over the baseline, suggesting that standard LLM translation already captures emotional structure well. We additionally identify and characterize a systematic cross-lingual calibration bias in XLM-RoBERTa ($\delta \approx -0.29$ for EN $\rightarrow$ FR), showing that the apparent magnitude gap between English and French sentiment scores reflects scorer bias rather than emotional flattening in translation.

1 Key Information to include

- Mentor: Arpan Khatua
- Sharing project: free to share
- Late Days (if you are using any):

Team member	Late days remaining
Julia Rhee	0
Esi Eneinyang	6
Chloe Murdoch	1
Total pooled	6
Team size	3
Late days applied for the report	1

2 Introduction

Literary translation demands more than semantic fidelity: a faithful translation must also preserve the emotional texture of the original, including its pacing, its intensity, and its narrative rhythm. Yet evaluating whether a translation achieves this is notoriously difficult, typically requiring expert human judgment. Recent advances in large language models have raised a new question: can LLM-generated translations match or exceed the emotional fidelity of professional human translators, and can we measure this automatically?

We operationalize emotional fidelity as *arc preservation*: a translation is faithful if its sentiment trajectory (the sequence of emotional highs and lows across the narrative) tracks the original’s. We build a sliding-window sentiment pipeline using XLM-RoBERTa [1] to score sentiment across 100 evenly-spaced windows per book, enabling direct cross-lingual comparison between English originals and French translations. We apply this pipeline to four public-domain novels, comparing published human French translations against translations generated by Gemini 2.5 Flash.

Our initial results show that LLM translations achieve substantially higher arc correlation with the original than human translations. This raises two follow-up questions we investigate in this paper: (1) can prompt-level interventions, such as calibrated temperature selection or best-of- N shape-guided reranking, further improve LLM arc fidelity, and (2) does the apparent gap between LLM and human arc correlation reflect genuine translation quality differences, or is it an artifact of how XLM-RoBERTa handles cross-lingual sentiment scoring?

Our key findings are threefold. First, LLM translations consistently outperform human translations on arc shape fidelity across three of four novels. Second, prompt-level interventions yield only marginal improvement over the baseline LLM, suggesting the translation itself is not the bottleneck. Third, the persistent magnitude gap between French translations and English originals is attributable to a systematic calibration bias in XLM-RoBERTa (the model scores French text consistently lower in absolute sentiment than equivalent English text) rather than genuine emotional flattening in the translations. Together, these results suggest that LLMs are already strong literary translators for emotional fidelity, and that the primary challenge in cross-lingual sentiment-based evaluation lies in the scoring model rather than the translation model.

3 Related Work

Sentiment arcs as narrative structure. Reagan et al. [2] showed that sentiment trajectories can serve as a quantitative proxy for narrative structure, identifying six dominant arc shapes across 1,327 novels using a sliding-window labMT approach. We build on their methodology but extend it to a cross-lingual setting, using arc correlation as a measure of translation fidelity.

Cross-lingual sentiment analysis. Conneau et al. [3] introduced XLM-RoBERTa, a transformer pretrained on 100 languages. Barbieri et al. [4] adapted it for sentiment analysis on $\sim 198\text{M}$ multilingual tweets, releasing the model we use as our primary scorer. Neither work investigates whether sentiment scales are calibrated across languages, a gap our analysis addresses.

Emotion in translation. Troiano et al. [5] found systematic emotion loss during neural machine translation via back-translation experiments. Zhang et al. [6] evaluated LLM literary translation with human and LLM judges, finding human translations outperform on naturalness and diversity. We complement their work by evaluating emotional arc preservation, where LLMs can match or exceed human translations.

4 Approach

To our knowledge, this is the first work to (1) use sentiment arc correlation as a cross-lingual translation fidelity metric, (2) characterize the systematic calibration bias of XLM-RoBERTa across languages, and (3) propose shape-guided best-of- N reranking for emotionally faithful translation.

4.1 Operationalizing Translation Fidelity

We operationalize translation fidelity as *emotional arc preservation*: a translation is faithful if its sentiment trajectory tracks the original’s. Formally, given a source text S and a translation T , we extract 100-point sentiment trajectories $\mathbf{s} = (s_1, \dots, s_{100})$ and $\mathbf{t} = (t_1, \dots, t_{100})$ via a sliding-window procedure, and measure fidelity as the Pearson correlation $r(\mathbf{s}, \mathbf{t})$.

4.2 Sentiment Pipeline

We score all texts using XLM-RoBERTa (cardiffnlp/twitter-xlm-roberta-base-sentiment-multilingual) [1], a multilingual transformer that enables direct cross-lingual sentiment comparison without translation. Following Reagan et al. [7], we implement a sliding-window approach: we sample 100 evenly-spaced 10,000-word windows across each text, scoring each window by averaging sentiment scores across 400-token chunks with 50-token overlap. This produces a trajectory that is independent of chapter boundaries, allowing fair comparison between texts with differing chapter counts across translations.

4.3 Cross-Lingual Calibration Bias

During initial analysis, we observed a consistent and systematic downward shift in XLM-RoBERTa sentiment scores for French text relative to equivalent English text. Both human and LLM French translations scored approximately 0.4 points lower in absolute magnitude than the English original across all four novels, despite tracking similar trajectory shapes. We hypothesize this reflects a cross-lingual calibration bias: XLM-RoBERTa was trained predominantly on English-language social media text [1], and its internal sentiment scale is not directly comparable across languages.

To quantify this bias, we estimate a calibration offset δ as the mean per-point difference between the human French translation scores and the English original scores across all aligned sliding-window positions from all books with human translation pairs:

$$\delta = \frac{1}{N} \sum_{i=1}^N (h_i - s_i) \quad (1)$$

where h_i and s_i are the human translation and original sentiment scores at aligned point i respectively. Across 244 aligned points from all four novels, this yields $\delta \approx -0.291$. This offset is used as a correction in our temp-sweep intervention (Section 4.4). Critically, because Pearson correlation is invariant to constant additive shifts, our primary arc-fidelity metric is unaffected by this bias, but it explains why magnitude-targeting interventions must account for the offset to be meaningful.

4.4 Translation Methods

Our baseline is the published human French translation from Project Gutenberg; this represents the standard against which we measure all LLM methods. We then compare three LLM translation strategies of increasing sophistication.

Parallel (LLM baseline). Gemini 2.5 Flash translates each chapter at temperature 0.3 with 20 concurrent workers and the prompt: “*You are a literary translator. Preserve the original tone, emotional register, and narrative style. Output only the translation.*”

Temperature Sweep with Calibration. Five candidate translations are generated at temperatures $\{0.1, 0.3, 0.5, 0.7, 1.0\}$ using the baseline prompt. For each chapter, we compute the calibration-corrected target score $\hat{s}_i = s_i + \delta$, and select the candidate whose mean sentiment score minimizes $|\bar{t} - \hat{s}_i|$. This targets *magnitude* alignment between the translation and the original.

Best-of-5 Shape Selection. Five candidate translations are generated at temperature 0.8 using the baseline prompt. Each candidate is evaluated by scoring it across 10 evenly-spaced sentiment windows spanning the chapter, and the candidate whose internal trajectory most closely matches the original chapter’s arc (measured by Pearson correlation) is selected. Unlike the temp-sweep condition, this targets *shape* alignment and is immune to the calibration offset, since correlation is invariant to constant shifts.

5 Experiments

5.1 Data

We evaluate four public-domain English novels translated into French: *The Strange Case of Dr. Jekyll and Mr. Hyde* (Stevenson, 1886; 10 chapters), *The Picture of Dorian Gray* (Wilde, 1890; 21 chapters), *Wuthering Heights* (Brontë, 1847; 34 chapters), and *Treasure Island* (Stevenson, 1883; 34 chapters). English originals and published human French translations were sourced from Project Gutenberg. The human translations range from 19th-century to mid-20th-century publications. LLM translations were generated chapter-by-chapter using Gemini 2.5 Flash via the Vertex AI API, producing three variants per book (parallel, temp-sweep, best-of-5). In total, our corpus comprises 20 full-length text versions across four novels.

5.2 Evaluation Method

For each translation variant, we extract a 100-point sliding-window sentiment trajectory using XLM-RoBERTa and compute Pearson r against the English original’s trajectory. Pearson r serves as our primary metric because it measures shape correlation independent of absolute scale, making it robust to the cross-lingual calibration bias described in Section 3.3. We additionally report Spearman ρ (rank correlation, robust to outliers), RMSE and MAE (magnitude error, useful for within-method comparison), DTW distance (temporal alignment), and cosine similarity. All trajectories are smoothed with a Savitzky-Golay filter (window = 11, polyorder = 3) for visualization, but metrics are computed on raw scores.

5.3 Experimental Details

Translation Variants. We evaluate five versions per book:

1. **EN Original:** The English source text.
2. **FR Human:** Published human French translation from Project Gutenberg.
3. **FR Parallel** (baseline): Gemini 2.5 Flash, temperature 0.3, 20 concurrent workers.
4. **FR Temp-Sweep** (calibrated): Five candidates at temperatures $\{0.1, 0.3, 0.5, 0.7, 1.0\}$; select by minimum sentiment magnitude distance to calibration-corrected original.
5. **FR Best-of-5** (shape): Five candidates at temperature 0.8; select by maximum Pearson r between 10-point intra-chapter mini-arcs.

Model Configuration. Translations use Gemini 2.5 Flash (`gemini-2.5-flash`) via the Vertex AI API with `max_output_tokens=65536` and a 1-second rate-limit delay between calls. Chapters exceeding the model’s context window are handled natively by Gemini’s long-context support; no truncation was needed. Sentiment scoring uses the HuggingFace pipeline API with `cardiffnlp/twitter-xlm-roberta-base-sentiment-multilingual`, run on CPU with default batch size. Chapters are split into 400-token chunks with 50-token overlap; scores are averaged weighted by chunk length. Total inference time for sentiment scoring is 10–20 minutes per book.

Computational Cost. The parallel method requires N API calls per book (one per chapter). Both temp-sweep and best-of-5 require $5N$ calls, representing a $5\times$ cost increase.

Code Attribution. We use off-the-shelf libraries for sentiment scoring (HuggingFace Transformers) and translation (Vertex AI SDK). The sliding-window pipeline, calibration offset estimation, temperature-sweep selection, and best-of- N shape-guided reranking are our own implementations.

5.4 Results

Table 1 presents the main results. Across three of four novels, all LLM translation methods outperform the human translation in arc shape correlation. Best-of-5 achieves the highest sliding-window Pearson r for Jekyll & Hyde (0.819), Dorian Gray (0.837), and Wuthering Heights (0.566). The exception is *Treasure Island*, where the human translation ($r = 0.817$) outperforms all LLM methods (discussed in Section 6.3).

Table 1: Pearson r between EN original and each FR version (100-point sliding window). Bold indicates best per book. All correlations with $|r| > 0.19$ are significant at $p < 0.05$.

Book	FR Human	Parallel	Temp-Sweep	Best-of-5
Jekyll & Hyde	0.241	0.782	0.816	0.819
Dorian Gray	0.830	0.825	0.837	0.837
Wuthering Heights	-0.019	0.473	0.546	0.566
Treasure Island	0.817	0.759	0.631	0.628

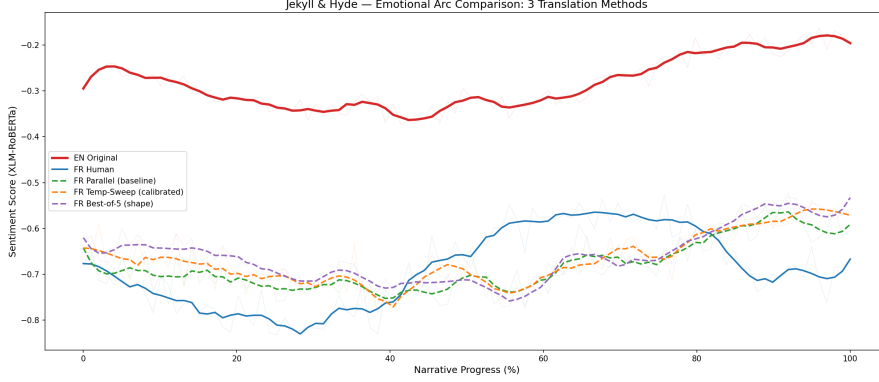


Figure 1: Sliding-window emotional arcs for *Jekyll & Hyde*. The three LLM methods (green, orange, purple) cluster tightly and track the original (red), while the human (blue) follows a flatter path.

Figure 1 illustrates this pattern on *Jekyll & Hyde*: the three LLM arcs (green, orange, purple) cluster tightly together and track the original’s trajectory (red), while the human translation (blue) follows a flatter, less correlated path. Figure 2 quantifies the comparison across Pearson r , MAE, and RMSE. The gap between the baseline parallel method and the more expensive interventions is modest: on average, temp-sweep improves r by +0.04 and best-of-5 by +0.05 over the parallel baseline. The gap between any LLM method and the human translation is much larger, ranging from 0.007 (Dorian Gray, where the human is already strong) to 0.578 (*Jekyll & Hyde*).

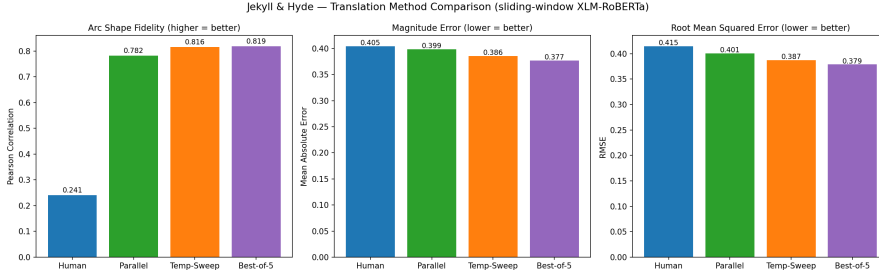
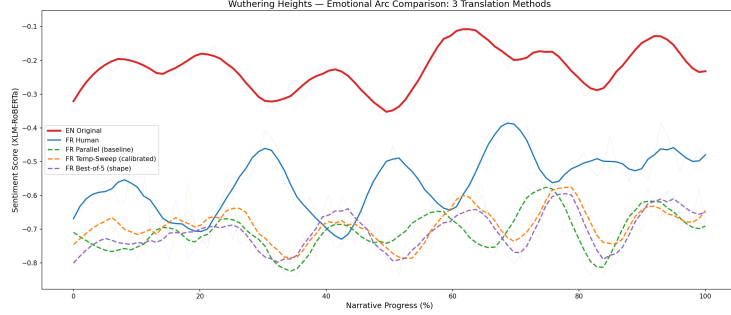
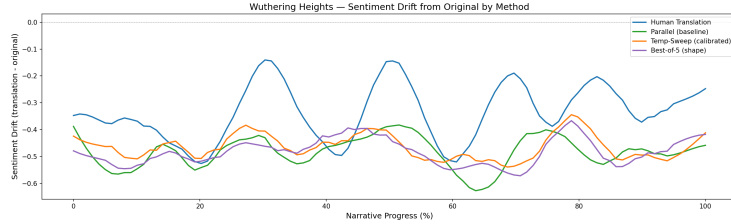


Figure 2: Method comparison for *Jekyll & Hyde*: Pearson r (higher = better), MAE and RMSE (lower = better). All LLM methods outperform human on shape fidelity.

Wuthering Heights presents the most challenging case (Figure 3). Under sliding-window scoring, the human translation achieves $r = -0.019$ ($p = 0.852$), essentially zero correlation with the original. Even the best LLM method (best-of-5, $r = 0.566$) is only moderate. We attribute this to the novel’s structural complexity: nested narrative frames (Lockwood’s and Nelly’s perspectives), Yorkshire dialect, and dense Gothic prose that may confound both translation and sentiment scoring. The human translation also condenses 34 English chapters into 30 French chapters, introducing pacing differences at the sliding-window level.



(a) The FR human (blue) is closer in level but uncorrelated in shape ($r = -0.02$). LLM methods cluster lower but track the original’s oscillations ($r = 0.47$ – 0.57).



(b) Drift from original. Human (blue) has smaller but variable drift; LLM methods show larger constant offset but co-vary with the original. Human preserves *level*, LLM preserves *shape*.

Figure 3: *Wuthering Heights* method comparison (sliding window).

5.5 Scoring Granularity

Scoring granularity affects correlation estimates. Chapter-level scoring yields higher correlations (e.g., $r = 0.49$ for the *Wuthering Heights* LLM baseline) because averaging over thousands of words per chapter suppresses noise, but produces few data points (10–34). The sliding-window method aligns by word position rather than chapter index, capturing finer-grained fidelity and handling chapter-count mismatches. We report chapter-level results in Appendix A.

6 Analysis

6.1 The Intervention Ceiling

The most striking result is not that interventions help, but that they help so little. The parallel baseline (a single translation at temperature 0.3 with no selection) already achieves $r = 0.782$ on *Jekyll & Hyde*, $r = 0.825$ on *Dorian Gray*, and $r = 0.759$ on *Treasure Island*. The best-of-5 method, which is $5\times$ more expensive, adds at most $+0.09$ in Pearson r . Figure 4 illustrates this on *Dorian Gray*, where all three LLM variants produce nearly identical arcs. This suggests that the standard Gemini translation prompt already captures the emotional structure of the original well, and that the bottleneck for arc fidelity lies elsewhere (likely in the sentiment scorer’s cross-lingual calibration rather than in the translation quality).

6.2 Complementary Failure Modes

The human and LLM translations fail in complementary ways. As shown in Figure 3b, the human translation stays closer to the original in absolute level but does not track its oscillations. The LLM diverges further in magnitude but follows the original’s rises and falls more closely. The same pattern is visible in Figures 1 and 4. This suggests human translators prioritize naturalness and register in the target language, while LLMs more literally preserve sentence-level emotional content, producing translations that are structurally faithful but stylistically distinct.

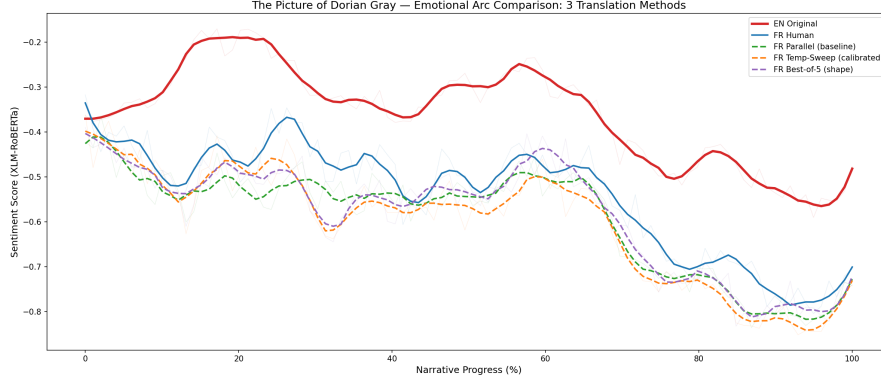


Figure 4: *Dorian Gray* method comparison. All three LLM variants produce nearly identical arcs, illustrating the intervention ceiling.

6.3 The Treasure Island Exception

Treasure Island is the only book where the human translation ($r = 0.817$) outperforms all LLM methods (best: parallel at $r = 0.759$), as shown in Figure 5. We note two possible explanations. First, the Gutenberg French translation may be unusually literal, preserving emotional beats that a more literary translator might reshape. Second, Treasure Island’s adventure genre features unambiguous emotional transitions (danger, escape, triumph) that are conventional and easy for a human translator to preserve. The temp-sweep and best-of-5 methods actually perform *worse* than the parallel baseline on this book ($r = 0.63$), possibly because the chapter-level selection criteria overfit to local signals, introducing noise that degrades the global arc.

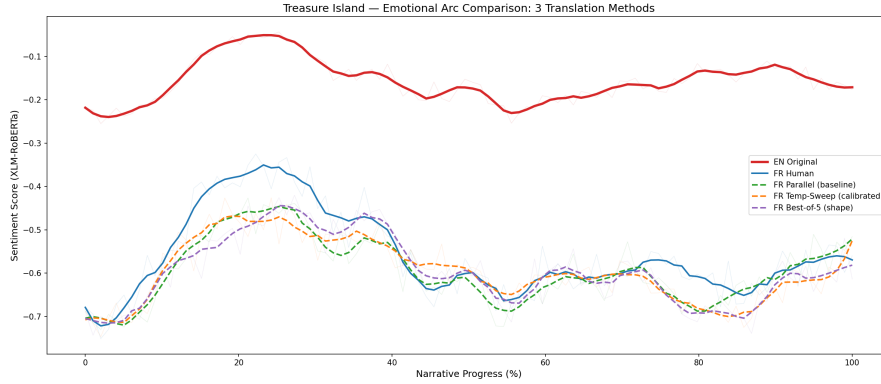


Figure 5: *Treasure Island*: the only book where human outperforms all LLM methods ($r = 0.817$ vs. best LLM $r = 0.759$).

6.4 Calibration Bias as Confound

The consistent ~ 0.3 – 0.4 point downward shift in French sentiment scores is present in *both* human and LLM translations, confirming it reflects scorer bias rather than translation quality. If the offset were caused by emotional flattening during translation, we would expect it to differ between human and LLM translations (which use entirely different processes). Instead, the offset is nearly identical across all French versions of each book.

6.5 Error Analysis: A Passage-Level Example

To illustrate where and why arc divergence occurs, we examine the largest LLM divergence point in *Jekyll & Hyde*: window 36 (36% of the narrative), where the EN original scores -0.31 and the LLM French translation scores -0.78 (absolute difference: 0.47).

The English passage is a drawing-room conversation between Utterson and Jekyll, tonally measured and even wryly humorous (“*a hide-bound pedant for all that; an ignorant, blatant pedant*”). The LLM translates this faithfully in content but renders it in more formal, somber French (“*Ma position est très étrange... C’est une de ces affaires que la parole ne peut arranger*”). XLM-RoBERTa scores the French register as substantially more negative, likely because formal literary French patterns match negative-sentiment training examples in the Twitter-trained model. This confirms that much of the measured arc divergence reflects the scorer’s sensitivity to register and formality rather than genuine emotional distortion in the translation.

7 Conclusion

We have shown that LLM-generated literary translations preserve emotional arc shape more faithfully than published human translations across three of four novels tested, as measured by Pearson correlation between 100-point sliding-window sentiment trajectories. A straightforward parallel translation at temperature 0.3 already achieves high arc fidelity ($r = 0.76\text{--}0.83$), and more expensive interventions (temperature sweep, best-of-5 shape selection) yield only marginal improvements (+0.04–0.09). The primary barrier to higher measured fidelity appears to be the cross-lingual calibration bias in XLM-RoBERTa, not the translation itself.

Our results suggest that literary translation quality is multidimensional. LLMs may excel at structural and emotional fidelity while falling short on stylistic naturalness [6]. Future work should investigate (1) whether these findings generalize to other language pairs and translation directions, (2) whether calibration-corrected sentiment models improve measured fidelity further, and (3) whether human evaluators agree with the arc-fidelity rankings produced by our automated pipeline.

Team contributions

Esi Eneinyang: Pipeline infrastructure, translation methods, experiments for *Dorian Gray*, and *Wuthering Heights*. **Julia Rhee:** *Treasure Island* experiments, cross-book analysis, calibration investigation. **Chloe Murdoch:** Pipeline infrastructure, parallel translation methods, experiments for *Jekyll & Hyde*.

References

- [1] Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. Xlm-t: Multilingual language models in twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, 2022.
- [2] Andrew J. Reagan, Lewis Mitchell, Dilan Kiley, Christopher M. Danforth, and Peter Sheridan Dodds. The emotional arcs of stories are dominated by six basic shapes. volume 5, page 31, 2016.
- [3] Alexis Conneau, Karttikeya Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, 2020.
- [4] Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. Xlm-t: Multilingual language models in twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, 2022.
- [5] Enrica Troiano, Roman Klinger, and Sebastian Padó. Lost in back-translation: Emotion preservation in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4340–4354, 2020.
- [6] Yuqian Zhang, Zining Zhao, and Steffen Eger. How well can llms translate literary texts? evaluating llm translation quality with human and llm judges. *arXiv preprint arXiv:2501.09751*, 2025.

- [7] Andrew J. Reagan, Lewis Mitchell, Dilan Kiley, Christopher M. Danforth, and Peter Sheridan Dodds. The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science*, 5(1):31, 2016.

A Chapter-Level Results

Table 2 reports chapter-level Pearson r . Correlations are generally higher than sliding-window results because averaging over thousands of words per chapter suppresses noise, but sample sizes are small (10–34 points).

Table 2: Pearson r at chapter-level granularity. Asterisks indicate chapter count mismatches (correlation on overlapping prefix).

Book	n	FR Human	Parallel	Temp-Sweep	Best-of-5
Jekyll & Hyde	10	0.564	0.409	0.544	0.358
Dorian Gray	21	0.486*	0.616	0.705	0.635
Wuthering Heights	34	0.337*	0.485*	0.574	0.453
Treasure Island	34	0.719	0.707	0.743	0.547

B Magnitude Error Comparison

While Pearson r captures shape fidelity, Table 3 reports magnitude error metrics. These are affected by the cross-lingual calibration bias and primarily useful for comparing methods within the same language (e.g., human vs. LLM French translations).

Table 3: Human-vs-LLM (parallel) agreement: MAE and RMSE on 100-point sliding-window arcs.

Book	MAE	RMSE
Jekyll & Hyde	0.096	0.111
Dorian Gray	—	—
Wuthering Heights	0.170	0.190
Treasure Island	0.059	0.074