Creativity Bias: How Machine Evaluation Struggles with Creativity in Literary Translations

Kyo Gerrits, Rik van Noord, Ana Guerberof Arenas
Centre for Language and Cognition, University of Groningen
k.gerrits@rug.nl

Abstract

This article investigates the performance of automatic evaluation metrics (AEMs) and LLM-as-a-judge evaluation on literary translation across multiple languages, genres, and translation modalities. The aim is to assess how well these tools align with professionals when evaluating translation creativity (creative shifts & errors) and see if they can substitute laborious manual annotations. A dataset of literary translations across three modalities (human translation, machine translation, and post-editing), three genres and three language pairs was created and annotated in detail for creativity by experienced professional literary translators. The results show that both AEMs and LLM-as-a-judge evaluations correlate poorly with professional evaluations on creativity, with LLM-as-a-judge showing a systematic bias in favour of machine-translated texts and penalising creative and culturally appropriate solutions. Moreover, performance is consistently worse for more literary genres such as poetry. This highlights fundamental limitations of current automatic evaluation tools for literary translation and the need to create new tools that do not frequently consider out of routine translations as errors.

1 Introduction

As machine translation (MT) has continued to improve significantly in recent years, literary MT is used more and more and its implementation in publishing houses is now a reality (Klemin, 2024; Warner, 2025). This also creates an increased focus on evaluating literary MT. Within the MT community, automatic evaluation metrics (AEMs) are standard tools for assessing translation quality (Schmidtova et al., 2024). Recently, LLMs have also been used as judges for detailed translation annotation (Fu et al., 2024; Wang et al., 2023; Zhang et al., 2025a). Professional annotation and quality assessment is still seen as the gold standard, but it can be expensive and time-intensive (Chaganty et al., 2018). It would save costs and labour if such annotation could be performed automatically.

Although these metrics are widely adopted within the NLP community as proxies for translation quality (Lavie et al., 2025; Schmidtova et al., 2024), translation scholars have expressed critique and doubt about their usefulness and validity: they argue that these metrics are designed to capture superficial similarity to the source or a reference, and penalise all shifts and deviations, even when these shifts are exactly what makes a translation creative, culturally suited and literary accomplished (Zhang et al., 2025b; Blagec et al., 2022; Way, 2018). Our previous study analysing creativity in MT output from LLMs (Du et al., 2025) found little correlation between AEMs and human annotation, but this was not the main focus of the study. Here we would like to address this issue in more depth.

In this paper, we report on an analysis of automatic metric performances across multiple literary translations, including three language pairs, genres and translation modalities (human translation (HT), post-editing (PE) & MT). We compare these to fine-grained annotations for creativity (errors & creative shifts) from experienced professional literary translators. Specifically, we ask three questions:

RQ1: How well do automatic evaluation metrics align with professional evaluation in literary translation?

RQ2: How well does LLM-as-a-judge evaluate

creativity in literary translation (creative shifts & errors), compared to professionals?

RQ3: Do these results change across genres?

Our main findings and contributions are:

1. We introduce a **dataset** of literary translations across 2 source languages, 2 target languages, 3 genres, and 3 modalities (HT, MT, PE). HT and PE are created by professional literary translators and all texts are annotated in detail for creativity (errors and creative shifts).

2. **Automatic evaluation metrics correlate negatively to weakly with professional annotations** on creativity in literary translation.

3. **LLM-as-a-judge struggles to identify errors**, both ignoring major errors and including non-errors. It also marks significantly fewer errors in MT and significantly more in HT compared to professional annotation.

4. **Both AEMs and LLM-as-a-judge ignore or penalise creativity in translation**. LLM-as-a-judge marks creative solutions as errors, while AEMs correlate weakly to negatively with creative shifts and both underperform on highly literary and creative genres such as poems compared to thrillers.

2 Related work

We will first cover recent findings on MT-quality for literary texts in WMT, then we will look at how creativity in translation has been operationalised and quantified, and lastly we will move to research on AEMs and LLM-as-a-judge evaluation.

2.1 Literary translation and MT

Recent WMT findings indicate that MT quality for literary translation of top-performing systems is comparable to human translated texts for multiple languages (Kocmi et al., 2025). Yet the literary texts used were extracted from a fanfiction website; fanfiction adheres to different rules than literature (in terms of characterisation, plot and narrative construction) (Jacobsen et al., 2024; Zadeh et al., 2021), and texts might be written by non-native authors or GenAI (Alfassi et al., 2026). Furthermore, studies analysing literary MT in more detail continue to find issues, especially those studies focusing on certain literary features: transferring meaning and especially idiomatic phrases and culturally charged expressions continue to be difficult (Matusov, 2019; Karpinska and Iyyer, 2023; Corpas Pastor and Noriega-Santiáñez, 2024).

Post-editing (PE) can be seen as a solution for MT issues: it helps remove MT-errors, while being less time-consuming and expensive compared to HT in some cases (Castaldo et al., 2025; Tewari and Baghel, 2026), but this is not always found (Terribile, 2024). Nevertheless, PE presents multiple issues when compared with HT: PE reduces authorial literary voice (Kenny and Winters, 2020; Mohar et al., 2020), and MT can prime post-editors, retaining errors and stylistic features of MT, making PE more like MT than HT (Macken et al., 2022; Daems et al., 2024).

2.2 Creativity and creativity annotations

Creativity in translations concerns something both new and appropriate: it is a skill often highlighted when referring to literary translation, as translators need to go beyond the word equivalence to create a text that reflects the voice and intent of the author in another culture (Bassnett et al., 2022; Rojo, 2017; Kaufman and Baer, 2012). Some researchers have operationalised creativity by using a framework for manual annotation. For example, Guerberof-Arenas and Toral (2020) measure creativity in literary translations by comparing professional translations with MT outputs in literary texts (see Section 3.3). This framework is based on earlier research by Kussmaul (1991; 1995; 2000) and Bayer-Hohenwarter, who operationalises creativity to analyse the differences between students and professionals' translations (2009; 2011; 2013).

MT outputs consistently show lower creativity than professional translations (Guerberof-Arenas and Toral, 2020; Guerberof-Arenas and Toral, 2022; Corpas Pastor and Noriega-Santiáñez, 2024). Research analysing creativity in automatically generated texts (so non-translations) using human judges or creativity frameworks also finds lower creativity scores for machine-generated texts when compared to texts created by writers (Belouadi and Eger, 2023; Chakrabarty et al., 2024; Chen et al., 2024). We expect lower creativity for our MT-texts, and we are interested to see whether AEMs and LLM-as-a-judge also reflect this.

2.3 Automatic evaluation: AEM and LLM-as-a-judge

AEMs are often based on the notion that the more similar a translation is to the reference–through n-gram overlap or learned representations that cap-

ture semantic, syntactic or contextual proximity– the higher the score. AEMs can be very useful to test MT systems during training, but their validity is less pertinent when applied to literary translations where different creative solutions are possible and stylistic divergences, cultural shifts and literary features resist a single interpretation and therefore a single translation (Zhang et al., 2025b). BLEU (Papineni et al., 2002), for instance, has been criticised for its over-reliance on strict ordering of surface elements as early as 2006 (Callison-Burch et al., 2006). Other string-based methods like TER (Snover et al., 2006) and chrF (Popović, 2016) also do not correlate well with professional scores (Novikova et al., 2017; Mathur et al., 2020).

Pre-trained models like COMET (Rei et al., 2020), BERTscore (Zhang et al., 2020) and BLEURT (Sellam et al., 2020) outperform string-based metrics (Freitag et al., 2022; Thai et al., 2022). Of these, COMET is used frequently and often outperforms others (Amrhein et al., 2022; Zouhar et al., 2024; Wu et al., 2024). Still, pre-trained metrics also fail to differentiate between critical and non-critical errors (Saadany and Orasan, 2021) and stylistic or nuanced differences (Mukherjee et al., 2025; Agrawal et al., 2024; Hanna and Bojar, 2021). They also struggle more at segment-level than at document-level (Moghe et al., 2023; Freitag et al., 2023).[1] Zhang et al. (2025b) showed that even recent SOTA metrics prefer machine-generated translations above those translated by professionals.

Recently, LLMs are also used for more fine-grained evaluation of translations through prompting. LLM-as-a-judge returns error spans, categories and severity, which helps identifying problematic segments and facilitates improving or post-editing texts (Zouhar et al., 2025; Xu et al., 2023) while also giving a better understanding of how LLMs analyse errors. Most prompting strategies have similar bases and focus on error evaluation using the MQM-framework (Lommel et al., 2024), including error categories and severity weights. However, to our knowledge, these evaluations have not been tested for literary translation, as presented in this paper.

More recently, using LLMs in a multi-agent framework for evaluation has gained more attention. In this paradigm, agents are assigned specific features of evaluation which are then discussed and weighted (Feng et al., 2025; Kim et al., 2025). We did not incorporate this in our study as multi-agent frameworks introduce considerable additional complexity (system design, prompt engineering, and computational cost) and are therefore beyond the scope of our exploratory research.

To our knowledge, there have been no attempts to have LLM-as-a-judge evaluate creativity. Some studies have however attempted to analyse creativity automatically in texts (Qiu and Hu, 2025; Bielova, 2025; Atmakuru et al., 2024; Noriega-Santiáñez and Pastor, 2023; Kovalkov et al., 2021), and although these models and methods show potential in recognising creativity, they still struggle to identify creative segments and underperform significantly compared to professionals. The novelty of our approach is that we use a taxonomy of creativity when prompting the LLM-as-a-judge (as explained in Section 3.5), to seek annotation results that are closer to professional annotation on creativity.

3 Methodology

In this section, we will discuss how the texts were selected, the translations and annotations processed and how AEM and LLM-as-a-judge were used in the experiment.

3.1 Texts: language pairs and genres

This study employs six source texts from two source languages (English (EN) and Russian (RU)) into two target languages (Dutch (NL) and Catalan (CA)) from three genres (poems, literary short stories and thrillers) across three modalities (HT, PE, & MT). The texts are approximately 150 words each, which is short and could limit the results, see Appendix A. None of them have been translated and published into these target languages to our knowledge before. An overview of the texts can be found in Table 1.

Genres The three genres (poem, short story & thriller) are included to analyse whether automatic metrics perform better on certain genres, specifically whether they perform better on 'lower' and relatively less creative genres. Research suggests that the impact of MT on creativity and readers is higher for literary texts that favour style over action (Guerberof-Arenas and Toral, 2020). We also hypothesise that genre-fiction, such as thrillers, is

[1] However, we use document-level as segment-level created little output, for Limitations see Appendix A.

Title	# Words	Author	SL	Genre
A Bath	172	Amy Lowell	EN	Poem
Afternoon in Summer*	182	Sylvia Townsend Warner	EN	Short Story
Bright Young Women*	154	Jessica Knoll	EN	Thriller
Your house*	97	Bella Akhamadulina	RU	Poem
Yasha's Dream*	169	Anna Starobinets	RU	Short Story
Night Watch*	151	Sergei Lukyanenko	RU	Thriller

Table 1: An overview of the texts used, including word counts. Russian titles are in English for readability. * indicates that only a fragment was used.

easier to translate for LLMs due to its conventionality (AbdulGhaffar, 2024); publishing houses also make this distinction when deciding to use MT to translate their books (Creamer, 2024). High(er) literature, on the other hand, might pose more of an issue for these models as the focus is not only on the narrative but also on style. Poetry, with the tension between form and content, especially represents a challenge (Sharma and Yadav, 2026; Chakrabarty et al., 2021). On the other hand, the poetic license afforded to poetry could actually give some allowance to LLMs in coming up with unexpected text and language use, or surprising collocations. We analyse whether genre impacts human evaluation and specifically whether the automatic metrics also perform differently across these genres.

Languages We include multiple language pairs to investigate if there are indications of different results for more distant language pairs. We include two distinct source languages (EN & RU) and two target languages (NL & CA), one of which is considered a low-resource language (CA).

3.2 Translations: HT, PE and MT

HT & PE The texts were translated by experienced professional literary translators. We contracted two EN-NL and RU-NL translators, who have each translated more than 15 novels in these language combinations.[2] They each translated half of the texts on their own (HT) and the other half post-edited (PE). This was alternated between languages, so, for example, if one translated the English thriller on their own, then this translator post-edited the Russian thriller (see Appendix C for details). Balancing languages and modalities across translators attempts to mitigate interpreting individual translator's effects as effects of language or modality. For CA, two EN-CA translators were hired to carry out the task. They have translated more than 12 literary novels from English.[3] We followed the same process as before, one translator did half of the texts on their own, and post-edited the second half, while the other one followed the opposite order. We did not apply any discounted rate for post-editing. The RU-CA was carried out by two other translators and is still under evaluation. It is therefore not included in the results.

MT To create our MT output we tried out different systems, including different prompts for the LLMs. This was done to make sure the MT output generated showed high levels of accuracy, fluency and style. We tried out multiple different strategies for the prompt, including a persona (He, 2024), zero-shot or few-shot (Castaldo and Monti, 2024; Zhang et al., 2023b; Zhang et al., 2023a), information about the author, author's style, context or content of the fragment (Opaluwah, 2025; Gao et al., 2024; Cui et al., 2024; Egdom et al., 2024; Yamada, 2023), or specifying to "translate creatively" (Du et al., 2025). Outcomes were evaluated per sentence based on CSs and error, and on translation preference. A zero-shot English-language prompt to translate creatively, including information on author and the context of the fragment in GPT-4o (the then most recent model, March 2025) worked best across all conditions.[4] The MT also formed the basis for the PE task.

3.3 Annotation framework

The professional annotations serve as a gold reference to compare with automatic evaluations. The annotations were based on the framework to measure creativity in literary translations, already mentioned in Section 2.2. This framework focuses on units of creative potential (UCPs): problematic units in an ST that translators cannot translate routinely and for which they have to use problem-solving abilities–their creative skills (Bayer-Hohenwarter, 2011). These can be metaphors, colloquial language, idioms, compar-

[2]Found via the *Expertisecentrum Literair Vertalen*, a Dutch literary translation database.

[3]Found via *AELC*, a Catalan literary translation database.

[4]See Appendix B for the prompt template. All prompts and evaluation protocols are available via `https://github.com/INCREC/Creativity_bias`.

ST	UCP	Modality	TT	Classification
Sally had emerged from her book	had emerged	HT	havia aixecat el cap (had lifted the head)	CS
		PE	havia emergit (had emerged)	R
		MT	havia sortit (had come out)	R

Table 2: A sample of creativity annotations, with the ST sentence, the UCP in question, the three modalities, the translation in the modalities and their classification. PE & MT are reproductions (R) as they reproduce the structure, form and meaning of the ST exactly, whereas HT changes the form and makes the movement more explicit, making it a creative shift (CS).

isons, etc.[5] In the TT, these UCPs can be resolved in a number of ways, the most common ones are:

1. Creative Shift (CS)
2. Reproduction (R)

CSs are translated UCPs in which the translation deviates from the original. Reproductions, on the other hand, are translations that do not deviate from the original or already-coined translations. A UCP with a reproduction and creative shift is shown as an example in Figure 1 for EN-NL: the adverb-gerund subclause in the original is reproduced in PE, whereas HT changes the structure to a main clause, slightly changes the meaning of words and makes it more idiomatic Dutch–in other words, a creative shift.

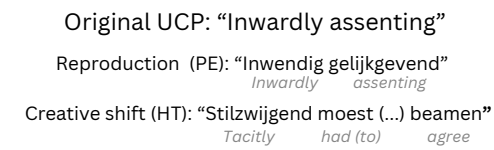

Figure 1: Example of a UCP in the ST with a Reproduction in the PE and a CS in the HT with English glosses underneath.

The creative shift annotations were done by 3 doctoral students who are proficient with the framework that were also translators and were native or proficient speakers for the language pair. They received the texts and the UCP annotations (created by 2 of the doctoral students) and could mark a solution as Creative Shift, Reproduction or Omission. An example of these annotations is shown in Table 2.

Creative solutions should not only be new but also correct. To verify this, professionals also annotated the translations for errors based on the MQM-framework (Lommel et al., 2024), including error type and severity. The same professional literary translators who translated the texts also annotated the translations for errors (see Section 3.2). Annotators could also mark *Kudos* for specifically well-found translation solutions. Annotations were made through the INCEpTION-platform (Klie et al., 2018).

The annotations for errors and creative shifts are combined to calculate a creativity score, also known as the creativity index (Guerberof-Arenas and Toral, 2020):

$$\text{CI} = (\frac{\#CS}{\#UCPs} - \frac{\#ErrorPoints - \#Kudos}{\#Words\_in\_ST}) * 100$$

This score rewards creative solutions, while penalising error points (errors weighted for severity) based on the number of words and UCPs in the ST. This model has been used for creativity annotations in literary translation before (Castaldo et al., 2025; Gerrits and Guerberof-Arenas, 2025; Du et al., 2025).

The professional annotations for the translations used in this paper show that HT was rated significantly higher in creativity than both MT and PE.[6] This is in line with previous research, with HT showing higher creativity than PE, and an even more pronounced difference with MT (Corpas Pastor and Noriega-Santiáñez, 2024; Guerberof-Arenas and Toral, 2022; Guerberof-Arenas and Toral, 2020). Looking at genres (for RQ3), we see that the 'high' literary genres (poems & short stories) have more CSs and a higher CI than thrillers, which indicates that these 'higher' genres are also more creative.

3.4 Automatic evaluation metrics (AEMs)

Our study explores two main types of AEMs, string-based and pre-trained models, besides reference-free and question-based models. An overview of the models used is shown in Table 3. The first six metrics were run through MATEO (Vanroy et al., 2023), while the others were run through their GitHub implementations. The HT created for this study were used as reference.

[5]For more, see Guerberof-Arenas and Toral (2022), p.191.

[6]For detailed results, see Appendix D.

AEM	Type	Ref	Source
BLEU	string-based	Yes	Papineni et al. (2002)
TER	string-based	Yes	Snover et al. (2006)
ChrF2	string-based	Yes	Popović (2016)
COMET	pre-trained	Yes	Rei et al. (2020)
BERTscore	pre-trained	Yes	Zhang et al. (2020)
BLEURT	pre-trained	Yes	Sellam et al. (2020)
COMETKiwi	pre-trained	No	Rei et al. (2022)
MetricX24	pre-trained	Yes	Juraska et al. (2024)
LiTransProQA	question-answering	No	Zhang et al. (2025c)

Table 3: The AEMs analysed in this paper.

3.5 LLM-as-a-judge evaluation prompting

We also wanted to investigate how well LLM-as-a-judge recognises and identifies errors and creative shifts. We let LLM-as-a-judge annotate errors and creative shifts separately, which was then combined in an LLM-as-a-judge CI score.

For the error annotation with LLM-as-a-judge, we opted for Tagged Span Annotation by Yeom et al. (2025) using GPT-5.2 with high reasoning effort, after trying out different prompting strategies across different models.[7]

For our automatic creative shifts analysis, we employ a similar prompt setup as for the error prompting, with a short persona ("You are a careful and balanced annotator for creativity in translation"), evaluation guidelines and creativity categories. The models were presented with the list of already identified UCPs and classified these, based on the protocol given to the researchers who identified the UCPs (see Section 3.3).[8] Using the already established list of UCPs we compare the annotations of the model to the annotations of the language experts, both in terms of CS classification and calculated CI score.

3.6 Comparing professional and computer evaluations

RQ1 To answer how well AEMs align with professional evaluation in literary translation, we correlate the AEM scores with the professional annotations for error points, creative shifts and the combined creativity index (CI) score.

RQ2 To determine how well LLM-as-a-judge for errors and creative shifts align with professional evaluation in literary translation, we first analyse the number of errors from the model and the professionals and see how often they match, including how well they match in categorisation of errors. By looking at the errors the model marked we try to understand its workings. We then look at how well LLM-as-a-judge classifies creative shifts in translation compared to professionals, and, finally, we analyse whether professional and automatic CI scores correlate.

RQ3 To investigate whether these aspects change across genres, we analyse the results for RQ1 and RQ2 across genres in more detail. In professional annotation, thrillers contained less creative solutions and received less Kudos than the other two genres (see Section 3.3), and we would like to know if this pattern is also present in automatic evaluation and whether automatic evaluations perform better on relatively less creative genres.

Limitations This study is not exempt from limitations as mentioned throughout. For more detailed information on these, see Appendix A.

4 Results

The results will be presented in three subsections, each one representing the results for each RQ.

4.1 Alignment of AEMs and Professional Annotations

Table 4 shows the AEM scores for each translation modality per language pair with the highest score per metric bolded for each language pair.[9] EN-NL PE has consistently higher scores than MT throughout (except for LiTransProQA), which agrees with professional annotations who also rate PE above MT. Less agreement is seen in the other language pairs. For example, for EN-CA, BERTscore, BLEURT, COMET, COMETKiwi have lower scores for PE than MT, contrary to the other metrics and professional annotations. For RU-NL, BLEURT, COMET, COMETKiwi, and LiTransProQA also score PE lower than MT. Furthermore, whereas the difference in CI for PE and MT is pronounced in the professional annotation, this is less evident in the AEMs. This suggests that

[7] This occurred after creating the MT, which is why we use a newer model. For an overview of our prompting trial, see Appendix E, and Appendix F for the prompt.

[8] We tried having the model detect UCPs, but it marked almost all sentences as UCP, even with strict guidelines. Moreover, comparing the model's performance annotating UCPs when different sets of UCP are used would be impossible.

[9] Details and individual AEM scores are in Appendix G.1.

		CI	BERTscore↑	BLEU↑	BLEURT↑	chrF2↑	COMET↑	COMETKiwi↑	LiTransProQA↑	MetricX24↓	TER↓
EN-NL	HT	**62.2**						76.7	**22.3**		
	PE	28.9	**87.2**	**37.8**	**70.5**	**57.7**	**82.4**	**80.0**	19.2	**2.66**	**50.4**
	MT	-15.9	84.2	31.4	68.0	54.2	80.0	77.5	21.4	2.91	52.5
EN-CA	HT	**30.0**						63.6	14.6		
	PE	15.2	79.3	**16.0**	48.0	**36.9**	67.6	65.5	**22.1**	5.48	**81.3**
	MT	-31.6	**79.9**	14.6	**48.3**	35	**68.8**	**71.4**	21.2	**5.06**	83.6
RU-NL	HT	**60.1**						**65.1**	20.0		
	PE	32.2	**68.3**	**20.3**	49.5	**42.4**	72.7	63.1	20.0	4.99	**71.8**
	MT	-50.8	68.1	14.5	**50.3**	40.0	**73.3**	63.8	**20.6**	**4.18**	76.2

Table 4: Overview of all AEM scores per modality for each language pair, with CI for comparison with human evaluations.

AEMs perform worse on low-resource languages and in pairs that do not involve English.

Correlations We want to see how well AEMs align with professional evaluations on errors, creative shifts (CS) and the combined creativity index (CI).[10] To do so, we created Spearman correlations between these professional evaluations and the AEM scores across all texts collapsing Modality, Genre and Languages. We used Spearman instead of Pearson as AEM-scores are not linearly spread and have different scales. Spearman is also more robust to outliers.

The results in Figure 2 show weak correlations. Errors have the strongest correlations (with 0.39 for BLEURT), but even this is moderate to weak.[11] For CS, we see weak correlations too, with 4 negative correlations and 0.2 as highest correlation (BLEU & chrF2). String-based metrics perform similar to pre-trained models, even though the latter are considered much more reliable (Freitag et al., 2022). COMET especially is surprising, as it is considered one of the best AEMs and is frequently used to compare model performance (Zouhar et al., 2024; Wu et al., 2024). The short length of our texts could influence the AEM's scores and our results. Still, AEMs here correlate weakly with errors and disregard or even penalise creative shifts. This lack of correlation is also shown for CI, which is understandable as it rewards creative shifts.

4.2 LLM-as-a-judge

Our second RQ focuses on how well LLM-as-a-judge correlates with professional annotations for creative shifts and errors. Table 5 shows the number of errors as annotated by professionals and GPT-5.2, and the number of matches between the

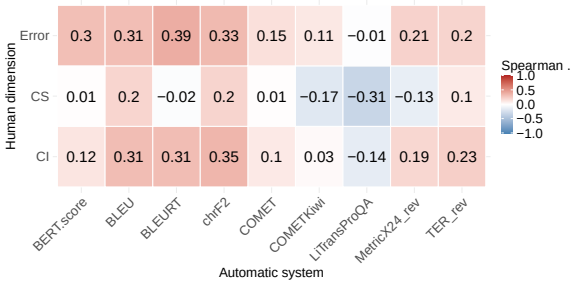

Figure 2: Heat map of the Spearman correlations between AEMs and professional annotations.

			Poem			Short story			Thriller		
ST	**TT**	**From**	**HT**	**PE**	**MT**	**HT**	**PE**	**MT**	**HT**	**PE**	**MT**
EN	NL	Human	3	8	16	4	5	17	6	11	9
		LLM	8	12	12	14	8	12	8	9	9
		Match	0	3	6	2	0	2	1	4	1
EN	CA	Human	9	7	16	3	7	25	4	7	18
		LLM	13	16	9	14	6	8	9	10	5
		Match	4	4	7	1	1	4	0	1	4
RU	NL	Human	7	6	14	4	14	14	9	15	28
		LLM	19	23	24	10	13	10	9	13	20
		Match	4	5	8	2	6	4	0	5	12

Table 5: Number of errors from professional and machine annotations, including matches.

two. A match means that the model and professionals identify the same error, but not necessarily that they classify the error under the same category.

The low number of matches for all translations shows that LLM-as-a-judge usually does not identify the same errors as the professionals. Looking at translation modalities, we see that HT is assigned more errors by the LLM-as-a-judge than by professionals, except for the RU-NL thriller (where it receives the same number of errors, but no matches). Similarly, MT is assigned more errors by professionals than by the LLM-as-a-judge, except for RU-NL Poem and all thrillers.[12] To examine whether these differences across transla-

[10] We focus on correlations between the creativity and the AEMs, so potential differences across other variables (except for Genre, see Section 4.3) are beyond our scope.

[11] Errors are inverted for positive correlations.

[12] For more on Genre, see Section 4.3.

tion modalities are significant, we fitted a binomial generalised linear model (GLM) predicting the source of the error annotation (professional vs LLM). Modality was included as main predictor, with Genre, ST and TT as additional control variables. We find a significant effect of Modality on error markings. Specifically, LLM-as-a-judge is more likely to mark errors in HT (z = 2.07, p = .038), whereas it is less likely to mark errors in MT (z = 6.16, $p < .000$) compared to professionals. This suggests that LLM-as-a-judge overestimates errors in HT and underestimates them in MT, the opposite pattern from professionals. This also shows that LLM-as-a-judge cannot reliably differentiate between modalities based on error counts, unlike professionals where most errors are marked in MT, followed by PE and then HT.

To explore in depth how the LLM-as-a-judge analyses errors, we look at the specific errors it marked. What stood out was that 'cultural' creative shifts and segments marked as Kudos by the professional translators were often marked as error (some examples shown in Table 6). Kudos were marked as error 15 out of 38 times they appeared. This was especially visible in EN-NL (9 out of 15 times, 60%) and EN-CA (5 out of 10 times, 50%), with 7 even marked as major error. A similar phenomenon occurs for cultural references: the model penalises creative shifts that are correct, and fails to mark incorrectly translated cultural references. This shows that the model fails to distinguish between errors and creative shifts and incorrectly marks creative segments–and thus more creative texts–down when such solutions should rather be promoted. In other words, the model biases against creative solutions.[13]

Categorisation The last aspect to explore for the error evaluation was their categorisation according to the MQM framework. Table 7 shows a confusion matrix with the distribution of errors per category (Style, Accuracy, Linguistic Convention, and No error) across professional and LLM-as-a-judge annotations. Colours indicate matches: green shows that the model and professionals agreed on both error span and category (there was for instance only 1 error that both mark as Style error), yellow indicates that the model and professionals both marked the error but categorised it differently (for example, both marked *waterbassin* (water basin) as error in the Dutch PE poem, but the model categorised it as an Accuracy error, whereas the professionals marked it as a Style error), red lastly indicates that the error marked by the model or professionals was not marked at all by the other (so professionals mark 88 style errors that the model does not mark at all). In other words: green indicates span and category matches, yellow only span matches and red no match.

Of all errors marked by the model, only 92 (the sum of the green and yellow cells) match in terms of span with professionals (28.3%). Of those span matches, 66.3% also match the category of the professionals. This seems due mainly to accuracy, where 52 error categories match, but the model also overestimates accuracy errors (only 73 (52 + 18 + 3) of the 255 accuracy errors from the model were also marked by professionals). The success rate is much lower for style (only 1 category match out of 4 span matches) and linguistic convention (8 out of 15). So although the model technically performs above chance when categorising errors, this is mainly due to Accuracy which the model assigns incorrectly to many errors. Furthermore, this performance only applies to the small subset of matching error spans which was low to begin with, so its actual performance in error annotation is still poor overall compared to professionals.

CS classification We are also interested to see if LLM-as-a-judge performs better on categorising creative shifts. As described in Section 3.5, the model classified an already-existing list of UCPs into CS, reproduction or omission. Table 8 shows a confusion matrix with the UCP-classification for professionals and LLM-as-a-judge over all texts. The table shows how often professionals and LLMs assigned CS, reproduction (R) and Omission (O) and how well they match. We can see for instance that both marked 90 UCPs as CS, but 110 UCPs were marked by professionals as CS but as reproduction (R) by the model.

Looking at the table, the model classified 63.6% (359 out of 564) of the UCPs correctly, shown in the green cells. We also see that professionals assign more CSs than LLM-as-a-judge (204 vs 165). We want to know if this difference is significant; as the data is paired (both classified the same set of UCPs) and nominal, we run a McNemar–Bowker test on the entire distribution table. This shows that the classification of UCPs significantly differs between professionals and LLM-as-a-judge ($\chi^2(1)$ = 17.15, p = .001), post-hoc pairwise symmetry tests

[13] More on bias, Stureborg et al. (2024) & Gao et al. (2025).

ST	Marked as error	Back translation	Explanation
public library	bibliotheek (HT, PE), biblioteca (HT, PE)	library	both commonly used without specifying 'public', which MT includes (error for professionals, not for LLM-as-a-judge).
second floor	eerste verdieping (HT)	first floor	US second floor is European first floor (CSI), same phrasing not marked in PE, MT had 'tweede verdieping' (second floor).
thirteen feet and two inches	quatre metres i deu centímetres (HT, PE)	4 meter and 2 centimetres	Metrical shift: EU uses the metric system, not the imperial one, MT retains imperial system and is not marked down by the model.
fifteen feet	vijf meter (HT, PE, MT)	5 meter	Metrical shift: EU uses the metric system, not the imperial one.
pours	klatert	splashes	Marked as Kudos by professionals, but as error by model.
inwardly assenting	stilzwijgend moest (...) beamen	tacitly (...) had to agree	Marked as Kudos by professionals, but as error by model (Figure 1).

Table 6: Examples of Kudos and culture-specific items (CSI) that are marked as errors by the model.

		Professionals				
		Style	Acc	Ling Con	No error	Total LLM
LLM	Style	1	3	0	15	19
	Acc	18	52	3	182	255
	Ling Con	4	3	8	36	51
	No error	88	59	45		192
Total human		111	117	56	233	

Table 7: Error categorisation. Green shows matches in span & categories, yellow only span and red no match.

indeed show that the model assigns significantly fewer CSs than the professionals do (p_{adj} = .012). Looking at translation modalities, LLM-as-a-judge assigns more CSs to MT than professionals do (54 vs 35).[14] This could be another indication that LLM-as-a-judge favours MT, but we also see that the model assigns very similar numbers of CSs to each modality (HT: 55, PE: 56, MT: 54), which is in contrast with professionals who assign more CSs to HT and fewer to MT. More research into prompting strategies for creative shifts is needed to see if the model indeed favours MT or if it simply assigns similar numbers of CSs regardless of the modality.

		Professionals			
		# CS	# R	# O	Total LLM
LLM	# CS	90	73	2	165
	# R	110	265	2	377
	# O	4	14	4	22
Total Professional		204	352	8	564

Table 8: Distribution of categorisation of CSs, reproductions (R) and omissions (O).

CI scores Lastly, we want to correlate the CI scores from professionals and LLM-as-a-judge.[15] To analyse this correlation, we created a scatter plot, shown in Figure 3. The figure shows the CI score from the model (X-axis) and the professionals (Y-axis) for each text. A dashed trend line shows overall correlation, with coloured lines indicating the correlation for each translation modality.[16] Looking at overall correlation between the model's and professionals' CI, we see a weak and insignificant correlation (ρ = .161, p =.422). Looking at the translation modalities, an interesting difference arises: the correlations almost flatten out for PE and HT, but the correlation increases for MT (ρ = .43) although this is not significant (p = .25). Still, it suggests that the model aligns better with professionals on MT. Besides correlation, we can also look at distinguishing texts with CI: for professionals we know see) that CI differentiates modality rather cleanly.[17] This pattern is absent from the LLM-as-a-judge CI on the X-axis which shows that CI of the model does not differentiate between the modalities. We see overall that CI scores from LLM-as-a-judge do not align with professional CI scores and are uninformative for creativity estimation.

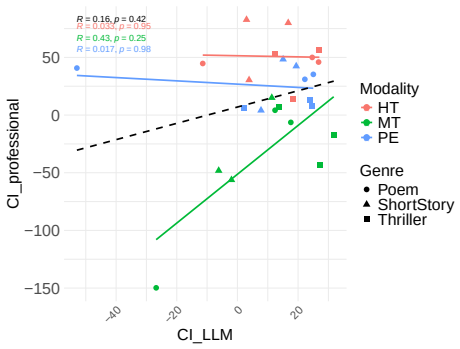

Figure 3: Scatterplot of CI from professional and LLM annotations. Colouring indicates the modality and shapes genre.

4.3 Genre

Our last RQ investigates whether the correlations between AEMs and LLM-as-a-judge for the first

[14] More details on CSs across modalities are in Appendix G.2.
[15] Detailed CI per text are in Appendix G.3.
[16] The shapes reflect genre, see Section 4.3.
[17] MT scores lowest, followed by PE and topped by HT (as shown in the figure by the green dots towards the lower end of the Y-axis, the blue in between and the red at the top).

two RQs differ significantly between genres.

First, we wanted to see if correlations between AEMs and professional annotations were higher for certain genres than for others. To analyse this, we created the same heat map as in Figure 2, but divided by genre, shown in Figure 4. Here, we see something striking: correlations between the AEMs and professionals for CSs and CI are much higher for thrillers, especially compared with poems. Our AEM analysis (see Section 4.1) showed that AEMs struggle with CSs and ignore or even penalise it. Section 3.3 further discussed that thrillers are a relatively low and less creative genre compared to high literary genres like poems, which the professional annotations showed as thrillers had significantly fewer CSs and a lower CI than the poems. Perhaps AEMs perform better on thrillers because thrillers are less creative and AEMs perform better on less creative genres. This might also explain why some studies do find higher correlations between AEMs and texts that are less literary, such as fan fiction, as used in WMT25 (Kocmi et al., 2025). It will be interesting to test this on a larger corpus to see if the trend remains.

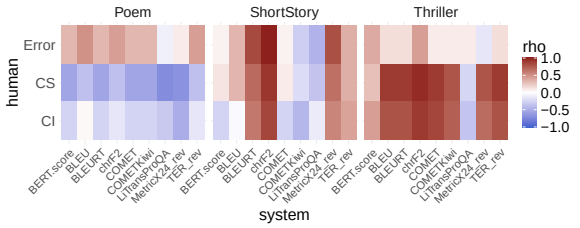

Figure 4: Heat map of Spearman correlations between AEMs and professional evaluations across genres.

# Errors	Poem	Short Story	Thriller
Human	84	93	107
LLM	137	95	93
Match	42	22	28

Table 9: Number of errors from professionals and LLM-as-a-judge (including matches) across genre.

We also analysed LLM-as-a-judge's performance on errors across genre, shown in Table 9. We see that while professionals mark most errors in thrillers, followed by short stories and poems, this is the opposite for LLM-as-a-judge. This suggests that LLM-as-a-judge struggles with high literary and more creative texts, ranking those lower than professionals would.

For creativity evaluation by LLM-as-a-judge, genres were included as shapes in Figure 3. No pattern arises directly from the genre. The two severe outliers are poems (the RU-NL PE Poem, scoring a CI of 41 from professionals but -53 from the LLM, and the RU-NL MT Poem, scoring a CI of -150 from the professionals and -27 from the LLM). This suggests LLM-as-a-judge struggles more with evaluating poetry compared to the other genres. Looking at the distribution along the X-axis (CI_LLM), we also see that thrillers have received relatively high scores from the model, which is not always matched by the professionals (as seen in the spread along the Y-axis).[18]

5 Conclusion

This paper set out to explore automatic evaluation performance on literary translation, specifically focusing on how well it correlates with human evaluation on creativity. Our analyses show that both AEMs and LLM-as-a-judge only partially and weakly correlate with professional annotations. They even penalise some creative shifts, such as correctly adapted cultural references. A closer look at their performances across genres reveals that they have lower correlations with professional annotations for high literary genres like poems compared to a relatively less creative genre like thrillers.

Although our dataset is small, these results suggest a key problem in the current metrics' ability to assess literary translation, as creative features are exactly what makes a literary translation stand out. This study begins to help explain the discrepancy between what professional annotators see in a literary text as creative and what AEMs or LLMs consider a quality translation, as in WMT2025 (Kocmi et al., 2025). Future work across a larger dataset using more annotators will show further details and differences between professional and automatic evaluation. Publishers who use these metrics to assess MT-quality and PE-rates should consider whether these metrics can accurately give an indication of the creative work needed. We think future work combining both creative shift and error annotations through multi-agent frameworks could reveal more to us about the workings, problems and potential solutions for machine evaluation. We need to find a new way of measuring that reflects creativity in literary translation better, which we believe real interdisciplinary research can achieve.

[18] See Appendix G.3 for detailed CI scores.

6 Funding & acknowledgments

This project has received funding from the EU ERC Consolidator Grant 101086819.

We would like to thank all translators and annotators who contributed to this project.

7 Sustainability statement

We did not run any of our experiments on dedicated local hardware, but queried LLMs such as ChatGPT and Claude both through their web interface and the ChatGPT API. To the best of our knowledge, the average CO_2 emissions of ChatGPT and Claude models are not publicly disclosed, so we will estimate our emissions for both the MT creation and the LLM-as-a-judge evaluations.

For the MT creation, we used LLMs through their web interface. In total, we submitted 63 requests to ChatGPT, 43 requests to Claude and 12 to both Google Translate and DeepL. A calculation by a third party estimates that each message sent to ChatGPT produces approximately 4.32g CO_2 (Wong, 2024), while each message sent to Claude produces approximately 3.5g (Hall, 2025). This would mean creating our MT emitted 0.4kg CO_2.

For the LLM-as-a-judge evaluations we submitted 1783 API requests to ChatGPT, leading to the processing of 1,563,995 tokens (combining inputs and outputs). Using the 4.32g CO_2 figure above, our LLM-as-a-judge experiments would have emitted 7.7kg CO_2.

In total, we estimate that our experiments emitted about 8.1kg CO_2, which is equivalent of driving about 55 km.

References

AbdulGhaffar, Nehal Ali. 2024. Beyond literal meaning: Neural machine translation constraints in translating the poetic depth of Al-Mutanabbi's "Tell My Beloved" . *Evolutionary Studies in Imaginative Culture*, pages 364–374.

Agrawal, Sweta, António Farinhas, Ricardo Rei, and Andre Martins. 2024. Can automatic metrics assess high-quality translations? In Al-Onaizan, Yaser, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14491–14502, Miami, Florida, USA, November. Association for Computational Linguistics.

Alfassi, Roi, Angelora Cooper, Zoe Mitchell, Mary Calabro, Orit Shaer, and Osnat Mokryn. 2026. Fanfiction in the age of ai: Community perspectives on creativity, authenticity and adoption. *International Journal of Human–Computer Interaction*, 42(5):3062–3094.

Amrhein, Chantal, Nikita Moghe, and Liane Guillou. 2022. ACES: Translation accuracy challenge sets for evaluating machine translation metrics. In Koehn, Philipp, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 479–513, Abu Dhabi, United Arab Emirates (Hybrid), December. Association for Computational Linguistics.

Atmakuru, Anirudh, Jatin Nainani, Rohith Siddhartha Reddy Bheemreddy, Anirudh Lakkaraju, Zonghai Yao, Hamed Zamani, and Haw-Shiuan Chang. 2024. Cs4: Measuring the creativity of large language models automatically by controlling the number of story-writing constraints.

Bassnett, Susan, Lawrence Venuti, Jan Pedersen, and Ivana Hostová. 2022. Translation and creativity in the 21st century. *World Literature Studies*, 14(1):3–17.

Bayer-Hohenwarter, Gerrit. 2009. Translation creativity: How to measure the unmeasurable. In Göpferich, Susanne, Arnt Lykke Jakobsen, and Inger M. Mees, editors, *Behind the Mind: Methods, Models and Results in Translation Process Research*, pages 39–59. Samfundslitteratur, Copenhagen, DK.

Bayer-Hohenwarter, Gerrit. 2011. Creative shifts as a means of measuring and promoting translation creativity. *Meta*, 56(3):663–692.

Bayer-Hohenwarter, Gerrit. 2013. Triangulating translation creativity scores. In *Tracks and Treks in Translation Studies: Selected Papers from the EST Conference, Leuven 2010*, pages 63–85.

Belouadi, Jonas and Steffen Eger. 2023. ByGPT5: End-to-end style-conditioned poetry generation with token-free language models. In Rogers, Anna, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7364–7381, Toronto, Canada, July. Association for Computational Linguistics.

Bielova, Maryna. 2025. Machine vs. human translation of stylistic neologisms in english language chick lit into ukrainian. *Respectus Philologicus*, (48 (53)):109–123, Oct.

Blagec, Kathrin, Georg Dorffner, Milad Moradi, Simon Ott, and Matthias Samwald. 2022. A global analysis of metrics used for measuring performance in natural language processing. In Shavrina, Tatiana, Vladislav Mikhailov, Valentin Malykh, Ekaterina Artemova, Oleg Serikov, and Vitaly Protasov, editors, *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*, pages 52–63, Dublin, Ireland, May. Association for Computational Linguistics.

Callison-Burch, Chris, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of Bleu in machine translation research. In McCarthy, Diana and Shuly Wintner, editors, *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento, Italy, April. Association for Computational Linguistics.

Castaldo, Antonio and Johanna Monti. 2024. Prompting large language models for idiomatic translation. In Vanroy, Bram, Marie-Aude Lefer, Lieve Macken, and Paola Ruffo, editors, *Proceedings of the 1st Workshop on Creative-text Translation and Technology*, pages 32–39, Sheffield, United Kingdom, June. European Association for Machine Translation.

Castaldo, Antonio, Sheila Castilho, Joss Moorkens, and Johanna Monti. 2025. Extending CREAMT: Leveraging large language models for literary translation post-editing. In Bouillon, Pierrette, Johanna Gerlach, Sabrina Girletti, Lise Volkart, Raphael Rubino, Rico Sennrich, Ana C. Farinha, Marco Gaido, Joke Daems, Dorothy Kenny, Helena Moniz, and Sara Szoc, editors, *Proceedings of Machine Translation Summit XX: Volume 1*, pages 506–515, Geneva, Switzerland, June. European Association for Machine Translation.

Chaganty, Arun, Stephen Mussmann, and Percy Liang. 2018. The price of debiasing automatic metrics in natural language evaluation. In Gurevych, Iryna and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653, Melbourne, Australia, July. Association for Computational Linguistics.

Chakrabarty, Tuhin, Arkadiy Saakyan, and Smaranda Muresan. 2021. Don't go far off: An empirical study on neural poetry translation. In Moens, Marie-Francine, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7253–7265, Online and Punta Cana, Dominican Republic, nov. Association for Computational Linguistics.

Chakrabarty, Tuhin, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2024. Art or artifice? large language models and the false promise of creativity. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Chen, Yanran, Hannes Gröner, Sina Zarrieß, and Steffen Eger. 2024. Evaluating diversity in automatic poetry generation. In Al-Onaizan, Yaser, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19671–19692, Miami, Florida, USA, November. Association for Computational Linguistics.

Corpas Pastor, Gloria and Laura Noriega-Santiáñez. 2024. Human versus neural machine translation creativity: A study on manipulated mwes in literature. *Information*, 15(9).

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

Cui, Menglong, Jiangcun Du, Shaolin Zhu, and Deyi Xiong. 2024. Efficiently exploring large language models for document-level machine translation with in-context learning. In Ku, Lun-Wei, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10885–10897, Bangkok, Thailand, August. Association for Computational Linguistics.

Daems, Joke, Paola Ruffo, and Lieve Macken. 2024. Impact of translation workflows with and without MT on textual characteristics in literary translation. In Vanroy, Bram, Marie-Aude Lefer, Lieve Macken, and Paola Ruffo, editors, *Proceedings of the 1st Workshop on Creative-text Translation and Technology*, pages 57–64, Sheffield, United Kingdom, June. European Association for Machine Translation.

Du, Shuxiang, Ana Guerberof Arenas, Antonio Toral, Kyo Gerrits, and Josep Marco Borillo. 2025. Optimising ChatGPT for creativity in literary translation: A case study from English into Dutch, Chinese, Catalan and Spanish. In Bouillon, Pierrette, Johanna Gerlach, Sabrina Girletti, Lise Volkart, Raphael Rubino, Rico Sennrich, Ana C. Farinha, Marco Gaido, Joke Daems, Dorothy Kenny, Helena Moniz, and Sara Szoc, editors, *Proceedings of Machine Translation Summit XX: Volume 1*, pages 578–591, Geneva, Switzerland, June. European Association for Machine Translation.

Egdom, Gys-Walt, Christophe Declercq, and Onno Kosters. 2024. 'can make mistakes'. prompting ChatGPT to enhance literary MT output. In Vanroy, Bram, Marie-Aude Lefer, Lieve Macken, and Paola Ruffo, editors, *Proceedings of the 1st Workshop on Creative-text Translation and Technology*, pages 10–20, Sheffield, United Kingdom, June. European Association for Machine Translation.

Feng, Zhaopeng, Jiayuan Su, Jiamei Zheng, Jiahan Ren, Yan Zhang, Jian Wu, Hongwei Wang, and Zuozhu Liu. 2025. M-MAD: Multidimensional multi-agent debate for advanced machine translation evaluation. In Che, Wanxiang, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the*

Association for Computational Linguistics (Volume 1: Long Papers), pages 7084–7107, Vienna, Austria, July. Association for Computational Linguistics.

Fernandes, Patrick, Daniel Deutsch, Mara Finkelstein, Parker Riley, André Martins, Graham Neubig, Ankush Garg, Jonathan Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In Koehn, Philipp, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 1066–1083, Singapore, December. Association for Computational Linguistics.

Freitag, Markus, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In Koehn, Philipp, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid), December. Association for Computational Linguistics.

Freitag, Markus, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In Koehn, Philipp, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore, December. Association for Computational Linguistics.

Fu, Jinlan, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. GPTScore: Evaluate as you desire. In Duh, Kevin, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576, Mexico City, Mexico, June. Association for Computational Linguistics.

Gao, Yuan, Ruili Wang, and Feng Hou. 2024. How to design translation prompts for chatgpt: An empirical study. In *Proceedings of the 6th ACM International Conference on Multimedia in Asia Workshops*, MMAsia '24 Workshops, New York, NY, USA. Association for Computing Machinery.

Gao, Zu, Lingbo Tong, and Zhiyong Zhana. 2025. Detecting and evaluating bias in large language models: Concepts, methods, and challenges.

Gerrits, Kyo 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 Bouillon, Pierrette, Johanna Gerlach, Sabrina Girletti, Lise Volkart, Raphael Rubino, Rico Sennrich, Ana C. Farinha, Marco Gaido, Joke Daems, Dorothy Kenny, Helena Moniz, and Sara Szoc, editors, *Proceedings of Machine Translation Summit XX: Volume 1*, pages 516–537, Geneva, Switzerland, June. European Association for Machine Translation.

Guerberof-Arenas, Ana and Antonio Toral. 2020. The impact of post-editing and machine translation on creativity and reading experience. *Translation Journal*, 9(2):255–282.

Guerberof-Arenas, Ana and Antonio Toral. 2022. Creativity in translation: Machine translation as a constraint for literary texts. *Translation Space*, 11(2):184–212.

Hall, Christine. 2025. What's Your Chatbot's Carbon Footprint? `https://fossforce.com/2025/04/whats-your-chatbots-carbon-footprint/`. Accessed: 2026/23/03.

Hanna, Michael and Ondřej Bojar. 2021. A fine-grained analysis of BERTScore. In Barrault, Loic, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Tom Kocmi, Andre Martins, Makoto Morishita, and Christof Monz, editors, *Proceedings of the Sixth Conference on Machine Translation*, pages 507–517, Online, November. Association for Computational Linguistics.

He, Sui. 2024. Prompting ChatGPT for translation: A comparative analysis of translation brief and persona prompts. In Scarton, Carolina, Charlotte Prescott, Chris Bayliss, Chris Oakley, Joanna Wright, Stuart Wrigley, Xingyi Song, Edward Gow-Smith, Rachel Bawden, Víctor M Sánchez-Cartagena, Patrick Cadwell, Ekaterina Lapshinova-Koltunski, Vera Cabarrão, Konstantinos Chatzitheodorou, Mary Nurminen, Diptesh Kanojia, and Helena Moniz, editors, *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 316–326, Sheffield, UK, June. European Association for Machine Translation (EAMT).

Jacobsen, Mia, Yuri Bizzoni, Pascale Feldkamp Moreira, and Kristoffer L. Nielbo. 2024. Patterns of quality: Comparing reader reception across fanfiction and commercially published literature. In *Proceedings of the Computational Humanities Research Conference 2024*, pages 718–739.

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

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

Kaufman, James C. and John Baer. 2012. Beyond new and appropriate: Who decides what is creative? *Creativity Research Journal*, 24(1):83–91.

Kenny, Dorothy and Marion Winters. 2020. Machine translation, ethics and the literary translator's voice. *Translation Spaces*, 9(1):123–149.

Kim, Junghwan, Kieun Park, Sohee Park, Hyunggug Kim, and Bongwon Suh. 2025. Mas-liteval : Multi-agent system for literary translation quality assessment.

Klemin, Jeremy. 2024. The last frontier of machine translation. *The Atlantic*, January.

Klie, Jan-Christoph, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico.

Kocmi, Tom and Christian Federmann. 2023. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In Koehn, Philipp, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore, December. Association for Computational Linguistics.

Kocmi, Tom, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Christof Monz, Kenton Murray, Masaaki Nagata, Stefano Perrella, Lorenzo Proietti, Martin Popel, Maja Popović, Parker Riley, Mariya Shmatova, Steinthór Steingrímsson, Lisa Yankovskaya, and Vilém Zouhar. 2025. Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets. In Haddow, Barry, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Tenth Conference on Machine Translation*, pages 355–413, Suzhou, China, November. Association for Computational Linguistics.

Kovalkov, Anastasia, Benjamin Paaßen, Avi Segal, Niels Pinkwart, and Kobi Gal. 2021. Automatic creativity measurement in scratch programs across modalities. *IEEE Transactions on Learning Technologies*, 14(6):740–753.

Kussmaul, Paul. 1991. Creativity in the translation process: Empirical approaches. In Leuven-Zwart, Kitty M. and Ton Naaijkens, editors, *Translation Studies: The State of the Art. Proceedings of the First James S. Holmes Symposium on Translation Studies*, pages 91–101. Rodopi, Amsterdam, NL.

Kussmaul, Paul. 1995. *Training the Translator*. John Benjamins Publishing, Amsterdam, NL.

Kussmaul, Paul. 2000. A cognitive framework for looking at creative mental processes. In Olohan, Maeve, editor, *Intercultural Faultlines Research Models in Translation Studies: v. 1: Textual and Cognitive Aspects*, pages 59–71. Routledge, London, UK.

Lavie, Alon, Greg Hanneman, Sweta Agrawal, Diptesh Kanojia, Chi-Kiu Lo, Vilém Zouhar, Frederic Blain, Chrysoula Zerva, Eleftherios Avramidis, Sourabh Deoghare, Archchana Sindhujan, Jiayi Wang, David Ifeoluwa Adelani, Brian Thompson, Tom Kocmi, Markus Freitag, and Daniel Deutsch. 2025. Findings of the WMT25 shared task on automated translation evaluation systems: Linguistic diversity is challenging and references still help. In Haddow, Barry, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Tenth Conference on Machine Translation*, pages 436–483, Suzhou, China, November. Association for Computational Linguistics.

Liu, Yang, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In Bouamor, Houda, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore, December. Association for Computational Linguistics.

Lommel, Arle, Serge Gladkoff, Alan Melby, Sue Ellen Wright, Ingemar Strandvik, Katerina Gasova, Angelika Vaasa, Andy Benzo, Romina Marazzato Sparano, Monica Foresi, Johani Innis, Lifeng Han, and Goran Nenadic. 2024. The multi-range theory of translation quality measurement: Mqm scoring models and statistical quality control.

Lu, Qingyu, Baopu Qiu, Liang Ding, Kanjian Zhang, Tom Kocmi, and Dacheng Tao. 2024. Error analysis prompting enables human-like translation evaluation in large language models. In Ku, Lun-Wei, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL*

2024, pages 8801–8816, Bangkok, Thailand, August. Association for Computational Linguistics.

Macken, Lieve, Bram Vanroy, Luca Desmet, and Arda Tezcan. 2022. Literary translation as a three-stage process: machine translation, post-editing and revision. In Moniz, Helena, Lieve Macken, Andrew Rufener, Loïc Barrault, Marta R. Costa-jussà, Christophe Declercq, Maarit Koponen, Ellie Kemp, Spyridon Pilos, Mikel L. Forcada, Carolina Scarton, Joachim Van den Bogaert, Joke Daems, Arda Tezcan, Bram Vanroy, and Margot Fonteyne, editors, *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 101–110, Ghent, Belgium, June. European Association for Machine Translation.

Mathur, Nitika, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In Jurafsky, Dan, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online, July. Association for Computational Linguistics.

Matusov, Evgeny. 2019. The challenges of using neural machine translation for literature. In Hadley, James, Maja Popović, Haithem Afli, and Andy Way, editors, *Proceedings of the Qualities of Literary Machine Translation*, pages 10–19, Dublin, Ireland, August. European Association for Machine Translation.

Moghe, Nikita, Tom Sherborne, Mark Steedman, and Alexandra Birch. 2023. Extrinsic evaluation of machine translation metrics. In Rogers, Anna, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13060–13078, Toronto, Canada, July. Association for Computational Linguistics.

Mohar, Tjaša, Sara Orthaber, and Tomaž Onič. 2020. Machine translated Atwood: Utopia or dystopia? *ELOPE: English Language Overseas Perspectives and Enquiries*, 17(1):125–141.

Mukherjee, Aniruddha, Vikas Hassija, Vinay Chamola, and Karunesh Kumar Gupta. 2025. A detailed comparative analysis of automatic neural metrics for machine translation: BLEURT & BERTScore. *IEEE Open Journal of the Computer Society*, 6:658–668.

Noriega-Santiáñez, Laura and Gloria Corpas Pastor. 2023. Machine vs human translation of formal neologisms in literature: Exploring E-tools and creativity in students. *Revista Tradumàtica. Tecnologies de la traducció*, 21:233–264.

Novikova, Jekaterina, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In Palmer, Martha, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark, September. Association for Computational Linguistics.

Opaluwah, Adeyola. 2025. Prompt-oriented output of culture-specific items in translated african poetry by large language models: An initial multi-layered tabular review. *American Journal of Computer Science and Technology*, 8(2):85–101, June.

Papineni, Kishore, 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.

Popović, Maja. 2016. chrF deconstructed: beta parameters and n-gram weights. In Bojar, Ondřej, Christian Buck, Rajen Chatterjee, Christian Federmann, Liane Guillou, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Aurélie Névéol, Mariana Neves, Pavel Pecina, Martin Popel, Philipp Koehn, Christof Monz, Matteo Negri, Matt Post, Lucia Specia, Karin Verspoor, Jörg Tiedemann, and Marco Turchi, editors, *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 499–504, Berlin, Germany, August. Association for Computational Linguistics.

Qiu, Ziliang and Renfen Hu. 2025. Deep associations, high creativity: A simple yet effective metric for evaluating large language models. In Christodoulopoulos, Christos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10859–10872, Suzhou, China, November. Association for Computational Linguistics.

Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In Webber, Bonnie, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online, November. Association for Computational Linguistics.

Rei, Ricardo, 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 Koehn, Philipp, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz,

Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid), December. Association for Computational Linguistics.

Rojo, Ana, 2017. *The Role of Creativity*, chapter 19, pages 350–368. John Wiley & Sons, Ltd.

Saadany, Hadeel and Constantin Orasan. 2021. BLEU, METEOR, BERTScore: Evaluation of metrics performance in assessing critical translation errors in sentiment-oriented text. In Mitkov, Ruslan, Vilelmini Sosoni, Julie Christine Giguère, Elena Murgolo, and Elizabeth Deysel, editors, *Proceedings of the Translation and Interpreting Technology Online Conference*, pages 48–56, Held Online, July. INCOMA Ltd.

Schmidtova, Patricia, Saad Mahamood, Simone Balloccu, Ondrej Dusek, Albert Gatt, Dimitra Gkatzia, David M. Howcroft, Ondrej Platek, and Adarsa Sivaprasad. 2024. Automatic metrics in natural language generation: A survey of current evaluation practices. In Mahamood, Saad, Nguyen Le Minh, and Daphne Ippolito, editors, *Proceedings of the 17th International Natural Language Generation Conference*, pages 557–583, Tokyo, Japan, September. Association for Computational Linguistics.

Sellam, Thibault, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In Jurafsky, Dan, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online, July. Association for Computational Linguistics.

Sharma, Priya and Tanuja Yadav. 2026. Poetry and emotions: Investigating the limitations of AI translation. In Uddin, Mohammad Shorif, Dharm Singh, Thittaporn Ganokratanaa, and Gaurav Kumawat, editors, *Proceedings of International Conference on Innovations in Data Science*, pages 401–409, Singapore. Springer Nature Singapore.

Snover, Matthew, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA, August 8-12. Association for Machine Translation in the Americas.

Stureborg, Rickard, Dimitris Alikaniotis, and Yoshi Suhara. 2024. Large language models are inconsistent and biased evaluators.

Terribile, Silvia. 2024. Is post-editing really faster than human translation? *Translation Spaces*, 13(2):171–199.

Tewari, Pragya and Anurag Singh Baghel. 2026. Stylistically-aware Hindi-English poetic translation with mbart and LLM-based post-editing. *Journal of Theoretical and Applied Information Technology*, 104(3):560–569.

Thai, Katherine, 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 Goldberg, Yoav, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9882–9902, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Vanroy, Bram, Arda Tezcan, and Lieve Macken. 2023. MATEO: MAchine Translation Evaluation Online. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 499–500, Tampere, Finland, June. European Association for Machine Translation.

Wang, Jiaan, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is ChatGPT a good NLG evaluator? a preliminary study. In Dong, Yue, Wen Xiao, Lu Wang, Fei Liu, and Giuseppe Carenini, editors, *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11, Singapore, December. Association for Computational Linguistics.

Warner, Emily. 2025. 'dumbing down' or 'momentous' opportunity? industry divided over ai literary translation. *The Bookseller*, July.

Way, Andy, 2018. *Quality Expectations of Machine Translation*, pages 159–178. Springer International Publishing, Cham.

Wong, Vinnie. 2024. Gen AI's Environmental Ledger: A Closer Look at the Carbon Footprint of ChatGPT. `https://piktochart.com/blog/carbon-footprint-of-chatgpt/`. Accessed: 2026/23/03.

Wu, Guojun, Shay B Cohen, and Rico Sennrich. 2024. Evaluating automatic metrics with incremental machine translation systems. In Al-Onaizan, Yaser, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2994–3005, Miami, Florida, USA, November. Association for Computational Linguistics.

Xu, Wenda, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Wang, and Lei Li. 2023. INSTRUCTSCORE: Towards explainable text generation evaluation with automatic feedback. In Bouamor, Houda, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5967–5994, Singapore, December. Association for Computational Linguistics.

Yamada, Masaru. 2023. Optimizing machine translation through prompt engineering: An investigation into ChatGPT's customizability. In Yamada, Masaru and Felix do Carmo, editors, *Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track*, pages 195–204, Macau SAR, China, September. Asia-Pacific Association for Machine Translation.

Yeom, Taemin, Yonghyun Ryu, Yoonjung Choi, and Jinyeong Bak. 2025. Tagged span annotation for detecting translation errors in reasoning LLMs. In Haddow, Barry, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Tenth Conference on Machine Translation*, pages 878–886, Suzhou, China, November. Association for Computational Linguistics.

Zadeh, Zhivar Sourati Hassan, Nazanin Sabri, Houmaan Chamani, and Behnam Bahrak. 2021. Quantitative analysis of fanfictions' popularity. *Social Network Analysis and Mining*, 12(42).

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

Zhang, Biao, Barry Haddow, and Alexandra Birch. 2023a. Prompting large language model for machine translation: A case study. In Krause, Andreas, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 41092–41110. PMLR, 23–29 Jul.

Zhang, Xuan, Navid Rajabi, Kevin Duh, and Philipp Koehn. 2023b. Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with QLoRA. In Koehn, Philipp, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 468–481, Singapore, December. Association for Computational Linguistics.

Zhang, Qiyuan, Yufei Wang, Yuxin Jiang, Liangyou Li, Chuhan Wu, Yasheng Wang, Xin Jiang, Lifeng Shang, Ruiming Tang, Fuyuan Lyu, and Chen Ma. 2025a. Crowd comparative reasoning: Unlocking comprehensive evaluations for LLM-as-a-judge. In Che, Wanxiang, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5059–5074, Vienna, Austria, July. Association for Computational Linguistics.

Zhang, Ran, Wei Zhao, and Steffen Eger. 2025b. How good are LLMs for literary translation, really? literary translation evaluation with humans and LLMs. In Chiruzzo, Luis, Alan Ritter, and Lu Wang, editors, *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, April. Association for Computational Linguistics.

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

Zouhar, Vilém, Shuoyang Ding, Anna Currey, Tatyana Badeka, Jenyuan Wang, and Brian Thompson. 2024. Fine-tuned machine translation metrics struggle in unseen domains. In Ku, Lun-Wei, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 488–500, Bangkok, Thailand, August. Association for Computational Linguistics.

Zouhar, Vilém, Tom Kocmi, and Mrinmaya Sachan. 2025. AI-assisted human evaluation of machine translation. In Chiruzzo, Luis, Alan Ritter, and Lu Wang, editors, *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 4936–4950, Albuquerque, New Mexico, April. Association for Computational Linguistics.

A Limitations

These are the limitations of this exploratory study.

- Our dataset is small, as it consists of six source texts from 2 languages translated into three modalities for two languages (HT, PE, MT). This means we have 27 translations (3 modalities * 3 genres * 3 language pairs) in total.

- The texts are also small, as they are about 150 words each. Some of the AEMs and perhaps even LLM-as-a-judge might perform better overall on longer texts, especially if they do not score or judge on sentence-level but on document-level (such as LiTransProQA).

- We only had a few human annotators available, which means the texts have been annotated by two to four people depending on the language pair. This makes it more difficult to see which annotations are more idiosyncratic than others and to see how the automatic metrics deal with those differences.

- We evaluated LLM-as-a-judge on segment-level (per sentence), as document-level created very little output (see Appendix E for details). However previous research has indicated this paradigm struggles more on segment-level compared to system-level (which we were limited to because of the dataset size)

- We only used one LLM and one method for analysis. The model was SOTA (ChatGPT's GPT-5.2) and we did try out multiple other LLMs and methods to make sure our LLM-as-a-judge was of high quality, but it would be interesting to see how other models and prompting strategies compare to each other.

B Prompt template for the creation of the machine translation

As mentioned in Section 3.2 we tried out different MT systems and prompts. Eventually we decided on a zero-shot English-language prompt to translate creatively, including information on author and the context of the fragment. We include the template of the prompts below using square brackets to indicate variable content such as authorial information and context. The individual prompts are included in our GitHub repository.[19]

```
You are a professional
translator.
Translate the following text from
[source text] into [target text]
creatively. It is a fictional
text, a segment from a [genre]
[title] by [author]. She/he is
a(n) [country of origin] writer,
[information on author and their
style].
This is a story about [content,
including point of view, plot,
and where the segment occurs
in the narrative]. In the
translation, please retain
the style and the feel of
the original. Please also
explain your steps and reasoning
throughout the process before
giving me the final translation.
```

[19] https://github.com/INCREC/Creativity_bias

C Counterbalancing of translations

The translators were counterbalanced across modalities, languages and genres for both translations and the annotations, shown in Table 10.

Language	Genre	Modality	Transl.	Ann.
EN-NL	Poem	HT	T1	T2
		PE	T2	T1
		MT		T1
EN-NL	Short story	HT	T2	T1
		PE	T1	T2
		MT		T2
EN-NL	Thriller	HT	T2	T1
		PE	T1	T2
		MT		T2
RU-NL	Poem	HT	T2	T1
		PE	T1	T2
		MT		T2
RU-NL	Short story	HT	T1	T2
		PE	T2	T1
		MT		T1
RU-NL	Thriller	HT	T1	T2
		PE	T2	T1
		MT		T1
EN-CA	Poem	HT	T3	T4
		PE	T4	T3
		MT		T3
EN-CA	Short story	HT	T3	T4
		PE	T4	T3
		MT		T3
EN-CA	Thriller	HT	T4	T3
		PE	T3	T4
		MT		T3

Table 10: Overview of which translators translated (transl.) and annotated (ann.) which texts.

D Analysis of professional annotations

Section 3.3 briefly discussed the results from the professional annotations by discussing CI scores for the translated texts across modalities. Table 11 shows the more detailed professional annotations for each text, including number of creative shifts, error points, Kudos and CI score. We created box plots for modality and genre across our human judgments variables, except for Kudos as this ranged only from 0 to 4, which did not result in informative box plots. These are shown in Figures 5 and 6.

To test whether any of the differences were significant, we ran statistical tests. As there were relatively few texts we used non-parametric tests, specifically we used Kruskal-Wallis rank sum tests

Mod.	Lang	Genre	# CS	# Errors	# EP	# Kudos	CI
HT	EN-NL	Poem	12	3	3	3	50.0
PE	EN-NL	Poem	8	8	8	4	31.0
MT	EN-NL	Poem	5	16	48	1	-6.3
HT	EN-NL	Short Story	19	4	8	3	80.0
PE	EN-NL	Short Story	10	5	5	2	42.4
MT	EN-NL	Short Story	2	17	109	1	-48.2
HT	EN-NL	Thriller	12	7	6	1	56.8
PE	EN-NL	Thriller	5	11	18	0	13.3
MT	EN-NL	Thriller	5	9	29	1	6.8
HT	EN-CA	Poem	12	9	9	2	46.0
PE	EN-CA	Poem	10	7	11	0	35.3
MT	EN-CA	Poem	6	16	36	0	4.2
HT	EN-CA	Short Story	7	3	3	3	30.4
PE	EN-CA	Short Story	3	7	19	2	4.1
MT	EN-CA	Short Story	1	25	115	0	-56.2
HT	EN-CA	Thriller	3	4	4	2	13.7
PE	EN-CA	Thriller	2	7	7	1	6.1
MT	EN-CA	Thriller	0	18	66	0	-42.9
HT	RU-NL	Poem	18	7	29	2	44.7
PE	RU-NL	Poem	16	5	23	0	40.8
MT	RU-NL	Poem	7	14	176	0	-149.8
HT	RU-NL	Short Story	10	4	4	3	82.7
PE	RU-NL	Short Story	7	14	18	2	48.4
MT	RU-NL	Short Story	3	14	18	2	15.1
HT	RU-NL	Thriller	10	9	9	0	52.9
PE	RU-NL	Thriller	7	14	18	2	48.4
MT	RU-NL	Thriller	4	29	63	1	-17.5

Table 11: Overview of human evaluation features (creative shifts (CS), errors, error points, Kudos and CI score) for each text. EP is short for Error Points.

as the variables each have three levels. If this reaches significance, we run post-hoc comparisons using Wilcoxon Rank Sum tests with Holm-Bonferroni correction. The results of these tests are shown in Table 12.

	Errors		CS		CI	
	χ^2	p	χ^2	p	χ^2	p
Modality	**16.89**	**<.000*** **	**11.28**	**.004****	**17.60**	**<.000*** **
HT-MT		<.000***		.008**		<.000***
HT-PE		.011*		.11		.011*
PE-MT		.004**		.11		.004**
Genre	**0.53**	**.77**	**6.05**	**.049**	0.45	**.79**
Poem-Thriller				.04*		
Thriller-ShortStory				.26		
ShortStory-Thriller				.69		

Table 12: Number of errors from professionals and LLM-as-a-judge (including matches) across genre. ***p <.001, **p <.01, *p <.05

E Overview of LLM-as-a-judge prompting strategies

As mentioned in the main body, we experimented with different prompting strategies across different models (see Section 2.3). We did not run systematic analyses of all texts across all strategies and models as this was not the main aim of the paper and it would have led to an (unnecessary)

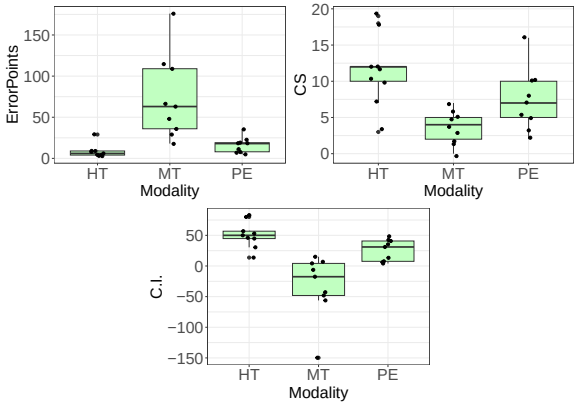

Figure 5: Box plots for errors, creative shift (CS) and creativity index (CI) per each modality level.

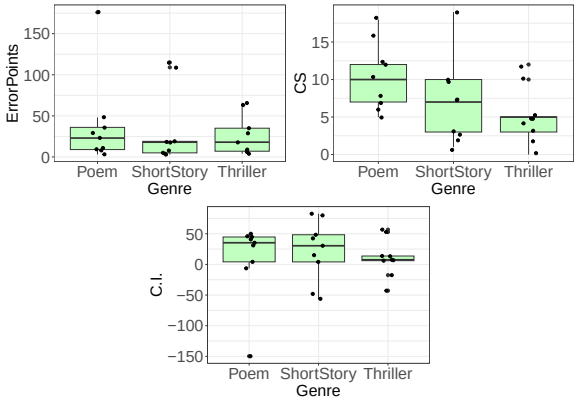

Figure 6: Box plots for errors, creative shift (CS) and creativity index (CI) per each genre level.

increase in costs and environmental impact, but we did try out multiple models and prompts using selected texts. Specifically, we tried out TSA (Yeom et al., 2025), AutoMQM (Fernandes et al., 2023), GEMBA-MQM (Kocmi and Federmann, 2023) and EAPrompt (Lu et al., 2024).

First, we tried out different strategies for the texts at document-level, as studies have shown these strategies tend to perform better at document-level than segment-level (Moghe et al., 2023; Freitag et al., 2023). However, this created outputs of only about 1 to 6 errors on MT-texts, whereas the professional annotation found on average 18 errors per MT-text. These errors also did not overlap, meaning it missed all professionally-marked errors and included non-errors.

We thus decided to annotate on **segment-level (per sentence)**. This increased the number of errors included in each annotation. However, it also caused multiple instances of omission/addition when segments where only moved across sentences. To mitigate this we **included preced-**

ing and subsequent sentence as context to the prompt as well as extra instructions in the prompt about omissions.Although not all instances were solved, this did decrease the (superfluous) instances of omissions and deletions overall and was thus kept for the final prompt. After careful consideration, **Tagged Span Annotation (TSA)** by Yeom et al. (2025) was chosen as the prompt basis as this returned the most accurate results for our dataset among the existing prompts.

For the model, we experimented with multiple models, including both reasoning models and non-reasoning ones. We focused on ones that are consistently used in literary translation according to the literature (Liu et al., 2023; Yeom et al., 2025; Kocmi and Federmann, 2023). We found that **reasoning models** worked better than non-reasoning ones,[20] and from the reasoning models **GPT-5.2** was the best performing on errors as Gemini 3 and Claude 3.7 Sonnet show erratic behaviour when marking errors.

Lastly, we also tried out different levels of reasoning effort: increasing the effort generally improved performance, but this flattened out towards higher efforts: xhigh performed similarly to high but had much higher costs (both pecuniary and time-wise). We therefore decided to set the **reasoning effort to high**, which was used for all evaluations.

F Template prompt for LLM-as-a-judge evaluation

F.1 Error evaluation prompt

For the error evaluation we eventually settled on an adapted version of the TSA (Yeom et al., 2025) as described above. The entire code can be found on our GitHub repository. The prompt looked at follows:

```
You are a careful and balanced
annotator for machine translation
quality. Your task is to
identify translation errors with
appropriate confidence.

## EVALUATION GUIDELINES:
- Be thorough but precise
- Only mark errors when confident
- Focus on objective errors, not
preferences
- Mark minimal error spans with
<v0>, <v1>, ...
- Do NOT treat cross-sentence
shifts as errors.
- If a phrase is moved to the
previous/next sentence but
translated correctly, do NOT
label it as omission or addition.
- Only mark an omission/addition
if the content is missing
entirely or appears incorrectly
in the translation as a whole,
not just in this segment.
- Use the provided previous/next
sentences to verify whether the
content is simply moved rather
than missing.

## ERROR CATEGORIES:
- Accuracy (addition, omission,
mistranslation, overtranslation,
undertranslation untranslated)
- Linguistic Convention (grammar,
punctuation, spelling)
- Style (awkward, inconsistent,
register, unidiomatic, audience
appropriateness)
- Other

## Severity:
- Major
- Minor

## CRITICAL RULES:
- Tags must be sequential:
<v0>, <v1>, <v2>...
- No explanations, comments, or
extra text
- If omission: insert empty
tags<vN></vN>

### CONTEXT (NOT FOR EVALUATION):
Previous source: {prev_source}
Previous translation:
{prev_translation}

Next source: {next_source}
Next translation:
{next_translation}
```

[20]Non-reasoning models such as Claude's Sonnet 4.6 and GPT-5.2 did not follow instructions consistently and also included error segments that were not even present in the text.

```
### SOURCE TEXT (CURRENT
SENTENCE):
{source}

### TRANSLATION (CURRENT
SENTENCE):
{translation}
```

F.2 Creative shift evaluation prompt

Based on the TSA strategy for errors, we created a new prompt for creative shifts shown below.

```
You are a careful and balanced
annotator for creativity in
translation.  Your task is to
categorise potential creative
segments (UCP) as a creative
shift (CS), Reproduction (R),
omission (O), or error (E).

## EVALUATION GUIDELINES:
- Be thorough but precise
- Consider whether there is a
direct or coined translation
available

  ## CREATIVITY CATEGORIES:
- Creative shift (CS) (All
translations that deviate from
the source with a different
idea or image are considered
creative shifts.  These are the
creative shifts (\non-literal").)
includes:
-- CSA (Abstraction):  refers to
those cases in which translators
use more vague, general or
abstract in the translation
-- CSM (Modification):  refers to
shifts that are at the same level
of abstraction (e.g.  express
a source text metaphor with a
different metaphor without the
image becoming more abstract or
concrete).  In other words, the
translation is modified for the
target culture.
-- CSC (Concretisation):  refers
to instances when the translation
evokes a more explicit, more
detailed and more precise idea
or image than the source text -
Reproduction (R): translations
that reproduce the source text
with the same idea or image,
even if they are acceptable,
are not considered a creative
shift in the translation, but a
reproduction.
- Omission (O): If a term or
expression in the source text is
omitted in the translation this
will be marked as omission.  An
omission could correspond to a)
creative solution (for example,
that text was omitted because
it does not make sense in the
translation) or b) a shortcut
solution (for example, that text
is omitted because it is rather
cumbersome to render).
- Error (E): If the translation
is not acceptable (contains too
many errors).

## CRITICAL RULES:
- No explanations, comments, or
extra text
```

G Detailed scores for AEMs and LLM-as-a-judge

G.1 Detailed AEM scores

The results of the AEM scores were briefly discussed in Section 4.1 More detailed analysis of AEMs will be discussed here. The scores are shown in Table 15 below.

G.2 Professional vs LLM-as-a-judge for creativity

Table 13 shows the distribution of creative shifts (CS) and reproductions (R) across the modalities for the professional and the LLM-as-a-judge annotations. Looking at the proportions of matching assignments by professionals and LLM-as-a-judge, we see that LLM-as-a-judge performs best in MT (128 / 188, 68%), closely followed by PE (121 / 188, 64%) with HT trailing slightly further behind (110 / 188, 59%). This seems largely due to the reproductions (110 of the 127 matches in MT are for reproductions for instance), but perhaps this is another indication that automatic metrics such

as LLM-as-a-judge performs better on MT than on other modalities.

G.3 CI-scores for professionals and LLM-as-a-judge

Table 14 shows the CI scores for each text from professionals and LLM-as-a-judge per language, genre and modality, showing the lack of correlations between professionals and LLM-as-a-judge.

			Professional			
			CS	R	O	Total
	HT	CS	41	12	2	55
		R	59	68	0	127
		O	3	2	1	6
		Total	103	82	3	188
LLM	PE	CS	32	24	0	56
		R	34	87	1	122
		O	0	8	2	10
		Total	66	119	3	188
	MT	CS	17	37	0	54
		R	17	110	1	128
		O	1	4	1	6
		Total	35	151	2	188

Table 13: Distribution of creativity classification for professionals and LLM-as-a-judge per modality.

Lang	Genre	Mod.	Human	LLM
EN-NL	Poem	HT	50.0	24.7
		PE	31.0	22.2
		MT	-6.3	17.6
	Short Story	HT	80.0	16.8
		PE	42.4	19.4
		MT	-48.2	-6.2
	Thriller	HT	56.8	27.0
		PE	13.3	24.0
		MT	6.8	13.6
EN-CA	Poem	HT	46.0	26.8
		PE	35.3	25.0
		MT	4.2	12.4
	Short Story	HT	30.4	3.8
		PE	4.1	7.8
		MT	-56.2	-2.0
	Thriller	HT	13.7	18.3
		PE	6.1	2.0
		MT	-42.9	27.2
RU-NL	Poem	HT	44.7	-11.4
		PE	40.8	-52.9
		MT	-149.8	-26.8
	Short Story	HT	82.7	3.0
		PE	48.4	15.1
		MT	15.1	11.3
	Thriller	HT	52.9	12.3
		PE	7.6	24.6
		MT	-17.5	31.8

Table 14: CI scores from professional and LLM-as-a-judge evaluation for each text.

Mod.	Lang	Genre	BERTscore↑	BLEU↑	BLEURT↑	chrF2↑	COMET↑	COMETKiwi↑	LiTransProQA↑	MetricX24↓	TER↓
HT	EN-NL	Poem						78.0	**22.6**		
PE	EN-NL	Poem	**89.3**	**38.1**	**71.4**	**59.6**	**83.0**	**78.6**	21.7	**2.80**	**39.8**
MT	EN-NL	Poem	86.4	29.9	69.3	55.7	80.0	75.3	**22.6**	2.81	45.4
HT	EN-NL	Short Story						76.6	**22.6**		
PE	EN-NL	Short Story	**85.5**	**37.6**	**69.8**	**55.5**	**83.5**	82.1	19.3	**2.22**	**62.3**
MT	EN-NL	Short Story	81.2	27.7	66.4	50.2	82.0	**82.2**	21.7	2.46	**62.3**
HT	EN-NL	Thriller						75.5	**21.6**		
PE	EN-NL	Thriller	**86.8**	**37.6**	**70.3**	**57.9**	**80.8**	**79.4**	16.7	**2.95**	**49.1**
MT	EN-NL	Thriller	85.0	36.5	68.3	56.6	77.9	75.2	19.9	3.47	49.7
HT	EN-CA	Poem						58.5	**22.6**		
PE	EN-CA	Poem	79.8	17.8	44.3	**40.3**	68.0	60.8	22.1	7.14	**76.4**
MT	EN-CA	Poem	**81.1**	**18.6**	**49.0**	39.9	**70.6**	**69.7**	22.1	**6.21**	77.7
HT	EN-CA	Short Story						70.9	20.3		
PE	EN-CA	Short Story	**85.1**	**29.8**	**69.3**	**52.3**	**81.8**	72.1	22.1	**3.72**	**56.5**
MT	EN-CA	Short Story	83.8	24.7	61.5	47.3	81.0	**77.4**	**22.6**	4.33	63.5
HT	EN-CA	Thriller						61.3	1.0		
PE	EN-CA	Thriller	73.1	0.4	30.3	**18.1**	52.9	63.7	**22.6**	5.57	111.0
MT	EN-CA	Thriller	**74.8**	**0.5**	**34.3**	17.8	**54.8**	**67.1**	18.9	**4.53**	**109.7**
HT	RU-NL	Poem						50.2	16.8		
PE	RU-NL	Poem	58.3	**7.2**	24.6	20.0	59.9	47.4	15.8	9.22	**101.8**
MT	RU-NL	Poem	**59.5**	2.7	**28.1**	**23.0**	**63.8**	**54.5**	**17.6**	**6.78**	110.9
HT	RU-NL	Short Story						**73.8**	21.6		
PE	RU-NL	Short Story	76.9	21.8	68.6	**53.5**	77.9	71.8	**22.6**	3.14	62.1
MT	RU-NL	Short Story	**78.3**	**25.2**	**72.4**	**53.5**	**78.9**	70.3	21.7	**2.43**	**59.1**
HT	RU-NL	Thriller						**71.2**	21.7		
PE	RU-NL	Thriller	**69.7**	**31.9**	**55.3**	**53.8**	**80.3**	67.0	21.7	**2.62**	**51.4**
MT	RU-NL	Thriller	66.5	15.5	50.3	43.4	77.1	66.7	**22.6**	3.32	68.5

Table 15: Overview of AEM scores for all translations in the text.