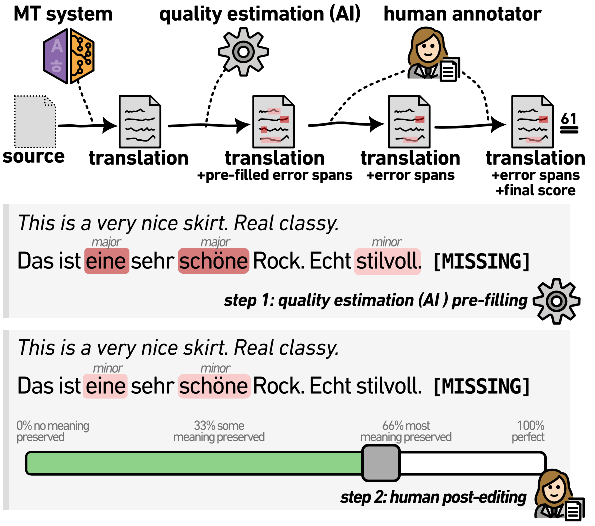
# AI-Assisted Human Evaluation of Machine Translation

**Anonymous ACL submission**

## Abstract

Annually, research teams spend large amounts of money to evaluate the quality of machine translation systems (WMT, Kocmi et al., 2023, inter alia). This is expensive because it requires detailed human labor. The recently proposed annotation protocol, Error Span Annotation (ESA), has annotators marking erroneous parts of the translation. In our work, we help the annotators by pre-filling the span annotations with automatic quality estimation. With AI assistance, we obtain more detailed annotations while cutting down the time per span annotation by half (71s/error span → 31s/error span). The biggest advantage of ESA$^{\text{AI}}$ protocol is an accurate priming of annotators (pre-filled error spans) before they assign the final score as opposed to starting from scratch. In addition, the annotation budget can be reduced by up to 24% with filtering of examples that the AI deems to be very likely to be correct.



*This is a very nice skirt. Real classy.*

Das ist eine sehr schöne Rock. Echt stilvoll. [MISSING]

*This is a very nice skirt. Real classy.*

Das ist eine sehr schöne Rock. Echt stilvoll. [MISSING]

Figure 1: The pipeline (top) and annotation user interface (bottom) with Error Span Annotation pre-filled with AI. In the example, the user: (1) lowered the severity of the gender agreement error, (2) removed incorrectly marked error span, and (3) assigned the final score.

## 1 Introduction

The quality of machine translation (MT) systems is periodically evaluated by academic and industry teams to measure progress and inform deployment decisions. This undertaking at scale, such as the WMT campaigns (Kocmi et al., 2022, 2023, inter alia), is extremely expensive, when requiring high annotation quality. Despite recent advancements in automated metric design (Freitag et al., 2023), these metrics remain misaligned with the ideal measure of text quality and human evaluation remains the most accurate and reliable standard.

Human evaluation protocols range from ranking different system outputs against each other (Novikova et al., 2018), to predicting scores (direct assessment, DA, Graham et al., 2015), or predicting specific error spans, types, and their severities (Multidimensional Quality Metrics, MQM, Lommel et al., 2014; Freitag et al., 2021). Kocmi et al. (2024) simplified this protocol into Error Span Annotation (ESA), which focuses only on the error span severities but not the actual error types, and is thus faster. One of the problems of the existing annotation protocols is their either very high cost, or low quality. In this work we aim to make the MT evaluation process with ESA less expensive.

We pose that human evaluation of MT can benefit from AI assistance. Despite the risk of automatic bias, human-AI collaboration can be faster and more accurate than human or AI alone (Bondi et al., 2022). Thus, instead of showing annotators just the source and the system translation, we pre-fill the translation with error annotations from an AI system (Figure 1 bottom). This setup, which we call **ESA$^{\text{AI}}$**, is enabled by the advancements in quality estimation systems (Guerreiro et al., 2023; Fernandes et al., 2023; Kocmi and Federmann, 2023), which provide accurate initial error spans. The main advantage of ESA$^{\text{AI}}$ comes from priming the user with possible translation errors.

[0] Code & collected data will be publicly available.

ESA$^{\text{AI}}$ yields 1.6 error spans per translation segment in contrast to 0.5 for human-only ESA. While the overall ESA$^{\text{AI}}$ annotation time is slightly lower to that of ESA (58s→52s/segment), ESA$^{\text{AI}}$ halves the time per span annotation (71s→31s/error span). In majority of cases where the AI did not predict any errors, the annotators did not add any new error span. We find that we can pre-filter such examples from the evaluation, save up to 24% of the budget, and the evaluation results will be almost identical. In addition, because of the unified priming, the annotators also become more self-consistent and have higher inter-annotator agreement, suggesting higher annotation quality.

## 2 Related Work

**Human evaluation.** One of the goals of MT evaluation is to compare systems to inform decisions (e.g. which system to deploy). Reference-based metrics compare the system translation to gold human translation, which can introduce evaluation bias (Freitag et al., 2020, 2023; Zouhar and Bojar, 2024). Reference-less approaches, known as **quality estimation**, do not have this problem, but they do not always correspond to human perception of quality (Freitag et al., 2023; Zouhar et al., 2024; Falcão et al., 2024) because the task is more complex. In higher-stakes settings, human annotators are employed to judge the translation quality.

The simplest option for human evaluation is to show the source and the translation and ask the annotators to give a number from 0 to 100, DA (Graham et al., 2015; Kocmi et al., 2022). and has the issue of low reliability and agreement. To make the annotations more objective, we can ask the annotators to mark specific errors in the translation (Multidimensional Quality Metrics, **MQM**, Lommel et al., 2014; Freitag et al., 2021). The marking is done based on their **severity** (e.g. minor or major) but also type (e.g. "inconsistent terminology"). This requires well-trained annotators and is thus expensive. In addition, this protocol does not yield scores, but only error spans, which are turned into the final score with a handcrafted formula.[1]

To simplify this process and align it with the goal of objective translation quality estimation, Kocmi et al. (2024) proposed **ESA**, which uses non-experts and asks them to provide only the error severity (not its type) and a final translation score. This combines both approaches in that the annotators are primed with their marked errors to provide high quality final scoring. The modalities are depicted in App. Figure 6.

**AI Assistance.** Prior work shows that annotators can benefit from AI assistance (Devarajan et al., 2023; Pavoni et al., 2022). However, the use of AI in evaluation is not straightforward because the AI might bias the user or induce over-reliance (Buçinca et al., 2021). Human annotators usually have a financial incentive to optimize their work. Veselovsky et al. (2023) showed, that up to 46% of annotators did use LLMs for abstractive summarization. Including AI assistance in the annotation directly could decrease the use of undisclosed tools. We do so with quality estimation (QE/**AI**) systems that mark error spans in the output. The most popular QE systems are xCOMET (Guerreiro et al., 2023), AutoMQM (Fernandes et al., 2023) and GEMBA (Kocmi and Federmann, 2023). The QE system is not always correct, but the output is vetted by a human annotator. The QE thus still offloads some of the work that a human would do and better primes the annotators for evaluation.

## 3 Setup

**Pipeline.** We implement our study in Appraise (Federmann, 2018) and use GEMBA, a GPT-based quality estimation system. We adapt the ESA protocol, where errors are marked and annotated as either minor or major.[2] The initial error markings are done by the AI and then post-edited by annotators. Afterwards, they manually assign a final score on the scale from 0 to 100% (not with AI). The error annotation part thus works as priming of the annotators in giving more accurate scores. The complete pipeline is shown in Figure 1 top. We also run the ESA$^{\text{AI}}$ setup twice with a different set of annotators to be able to determine the inter-annotator agreement and annotation stability.

**Dataset and collected data.** We use the data of WMT23 Metrics Shared Task (Freitag et al., 2023) which has been annotated with MQM and ESA. For maximum compatibility, we use the setup identical to that of Kocmi et al. (2024). We focus on English→German where 13 systems were submitted, one of which is the human reference translation. For each system, we have 207 segments (average 18 words per segment) from 74 source documents.

[1] With some exceptions, the score computation from spans is a sum across all errors with -1 for minor and -5 for major.

[2] Minor: style/grammar/lexical choice could be better; Major: changes meaning, lowers usability. See Appendix C.

We first examine the high-level distribution of the collected data in Table 1. For ESA$^{AI}$, the total number of reported error spans is three times higher than for ESA, which is due to the high number of suggested annotations by GEMBA. The split between minor and major errors is similar, though ESA$^{AI}$ annotators prefer major errors as opposed to ESA, even slightly more than those produced by GEMBA. Finally, the overall translation score is lower for ESA$^{AI}$ as opposed to just ESA. This is potentially caused by the priming effect of initially annotated error spans by GEMBA which highlight the negative aspects of the translation.

| Protocol/method | #errors | %minor/%major | Score |
|---|---|---|---|
| ESA | 0.45 | 63% / 37% | 81.8 |
| ESA$^{AI}$ | 1.63 | 54% / 46% | 76.7 |
| GEMBA (AI only) | 1.51 | 55% / 45% | × |

Table 1: Average number of error spans and scores across ESA, ESA$^{AI}$, and GEMBA.

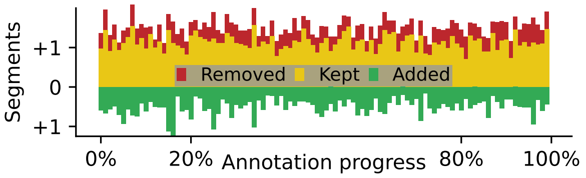


Figure 2: Number of removed/kept/added error spans from GEMBA with respect to annotator progress.

## 4 Analysis

To evaluate the new ESA$^{AI}$ annotation pipeline, we consider two main aspects: (1) the annotation process, including its reliability and human effort, and (2) its usefulness for MT system comparisons.

### 4.1 ESA$^{AI}$ Evaluation Process

Not all post-editing operations are of equal value. For example, moving the error span by a few characters to the left is less important than adding a new error span for a missing translation. We point out two post-editing types: (1) changing the error span severity, and (2) editing the error span boundaries (App. Table 5). In 11% of cases, the users only changed the severity. This is important from the workflow perspective, because it only requires clicking the error span. In many cases, the error span was only moved. Time-wise this is more expensive because it requires the original error span to be removed and a new one created in its place. This operation can be skipped because it does not contribute to the ESA score. Therefore, the annotators could be instructed more specifically to not try to post-edit errors as long as they are approximately correct. See App. Example 2 for post-editing types.

**Do annotators blindly accept AI hints?** Gradual overreliance (Holford, 2022) is a type of automation bias which arises through repetition of non-problematic examples. Especially when there are no repercussions, the annotator might be tempted to only confirm the AI suggestion without actually doing any post-editing work. We first examine this through the perspective of changing in annotator's behaviour thorough the annotation. In Figure 2 we show that the annotators make the same number of edits at the beginning as at the end, therefore excluding the possibility of automation bias.

**Do annotators pay attention?** We use attention checks, where the translation is malformed but GEMBA does not show an error (App. Example 1). The range for passing the attention checks for both protocols is the same—around 75% (App. Table 6). ESA$^{AI}$ is at a disadvantage because GEMBA spans were showing that the perturbed span was correct (i.e. no error spans). Therefore, the pertrubed examples were even more out-of-distribution and attention in-distribution is likely higher.

**Do AI mistakes affect annotators?** Showing incongruent examples, where AI predictions are clearly wrong, has the potential of reducing the user's trust in the AI and subsequent collaborating (Dhuliawala et al., 2023). Such examples are our attention checks. 84% and 73% of GEMBA-suggested spans are accepted for the document directly before and after the perturbed one. This is a slight decrease in trust, but it does not render the collaboration ineffective.

**How long do annotations take?** One of the motivations of the AI-assisted setup is speeding up the annotations and leading to lower costs. The variance in individual annotator time can be explained by how much they post-edited the GEMBA error span annotation (see Figure 3). Per segment, ESA$^{AI}$ annotators required 52s while ESA required 58s, which is comparable. The time is 71s per single error span for ESA but 31s per single error span for ESA$^{AI}$, making the latter more efficient in detailed annotation. In addition, the annotators get faster as the annotation progresses (Appendix B).

**Do annotators agree?** For a robust and objective annotation protocol, the scores by two independent annotators should be similar. To test this,

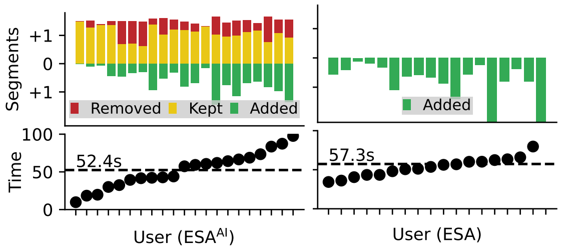


Figure 3: Annotation actions (remove/keep/add an error span) and time per segment. Each dot is an annotator.

we ran the annotations again with different annotators. App. Table 7 shows that ESA$^{AI}$ has a much larger inter-annotator agreement. For the MQM-like score computation, this is due to the bias by the pre-filled error spans. Still, the agreement is much higher also for the direct scoring, likely due to the priming. This is consistent with much higher ESA$^{AI}$ *intra*-annotator agreement (App. Figure 9), i.e. how much annotators agree with themselves.

### 4.2 ESA$^{AI}$ for Evaluation of WMT Systems

Our goal is for ESA$^{AI}$ to be as reliable or more than ESA in ranking MT systems. We consider MQM collected by Freitag et al. (2023) as the human gold standard and show the system-level correlations with our protocol in App. Figure 8. Both ESA and ESA$^{AI}$ have similar correlations with MQM$^{WMT}$, justifying our setup. In Table 2 we show that this protocol does not stray far away from existing ones in terms of segment-level rating. Many of these cross-protocol correlations are on part with inter-annotator agreement, which is naturally the upper bound. Notably, ESA$^{AI}$ has higher correlation than ESA or MQM by Kocmi et al. (2024) alone.

| | ESA | ESA$^{AI}$ | MQM | GEMBA |
|---|---|---|---|---|
| MQM$^{WMT}$ | 0.240 | 0.292 | 0.239 | 0.416 |

Table 2: Kendall $\tau$ segment-level correlations between evaluation protocols, ESA and ESA$^{AI}$ use direct scores.

**Can cost be further lowered?** GEMBA is recall-focused and therefore the occurrence of "false positive" segments is low. In 89% of cases, spans that were marked by GEMBA to have 0 errors retained 0 errors after annotation (App. Table 4), and such segments had an average score of 95. This makes it possible to also use GEMBA as a pre-filtering step. If we replace all such segments with 100 (to not overfit), all but one system comparisons remain the same. Alternatively, one can also filter segments for which GEMBA marks 0 errors for most systems, which has the advantage that we do not alter the data. For this method, again all but one system comparison would be the same (Figure 4). Pre-filtering can thus result in almost 25% budget saving (~52 segments per system).

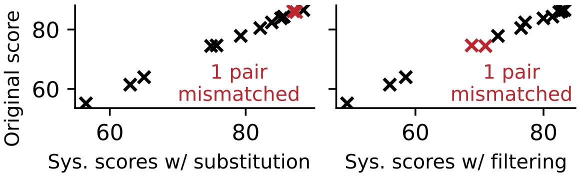


Figure 4: Average system scores with either substitution or filtering of segments with no GEMBA errors.

**How many annotations are needed?** With large enough evaluation, even noisy annotation schemes yield the true system ordering. Conversely, only robust annotation schemes are good on a small scale. We formalize this in Appendix A to show that ESA$^{AI}$ leads to better annotations than ESA or MQM. For each subset size, e.g. 30 source sentences, we select 1000 random subsets and compute the system ranking accuracy against the whole dataset. Results in Figure 5 show that GEMBA alone is the most consistent because of the lack of inter-annotator confusion. However, it also increases the stability and quality of scores that annotators assign manually in ESA$^{AI}$. In addition in practice, one can annotate fewer examples (e.g. 2000 for ESA$^{AI}$) to obtain the same system-level accuracy as a slower protocol (e.g. 2500 for ESA), lowering the costs.

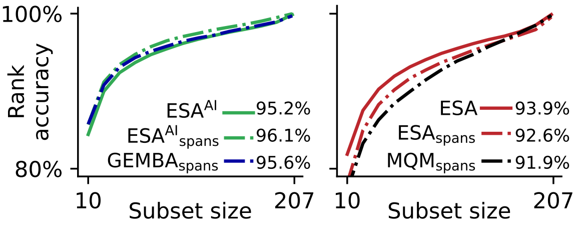


Figure 5: Accuracy of a system ranking only on a subset against ranking on full data. Percentages are averages. See numbers in App. Table 8.

## 5 Conclusion

Our AI-assisted protocol of human evaluation of MT is faster and cheaper. This protocol is more robust and self-consistent and increases inter-annotator agreement by priming the annotators with pre-annotated error spans. Our analysis also shows that the annotators did not over-rely on the AI and were able to maintain evaluation quality. The inclusion of AI in evaluation also opens many options for further evaluation economy.

## Limitations

Despite the advantages in lower costs per error span of the presented setup, we urge practitioners to not use this approach when metrics evaluation is one of the expected tasks due to the particular bias to the used metric in the setup. The intended application of this pipeline is purely a more efficient evaluation of machine translation system quality.

Both ESA$^{\text{AI}}$ and GEMBA rank GPT-4-5shot as the best system, a system that uses the same LLM to translate sentences as we use to generate for GEMBA. This indicated a weakness that our approach is biased towards systems build on top of the same underlying LLM. Liu et al. (2023) described this phenomena when the same system used for generating output should not be used to also evaluate them. This issue could be mitigated by using two different LLMs to generate error spans.

Lastly, we use GEMBA, a GPT4-based system, for the quality estimation and work with WMT 2023 data. Unfortunately, we can not exclude the possibility of the QE system being trained on this data, though the texts and scores are kept in two separate large files with non-linear mappings.

## Ethics Statement

The annotators were paid a standard commercial translator wage in the respective country. No personal data was collected and the showed data was screened for potentially disturbing content.

## References

Elizabeth Bondi, Raphael Koster, Hannah Sheahan, Martin Chadwick, Yoram Bachrach, Taylan Cemgil, Ulrich Paquet, and Krishnamurthy Dvijotham. 2022. Role of human-ai interaction in selective prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36/5, 5286–5294.

Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–21.

Ganesh Gopal Devarajan, Senthil Murugan Nagarajan, Sardar Irfanullah Amanullah, SA Sahaaya Arul Mary, and Ali Kashif Bashir. 2023. Ai-assisted deep nlp-based approach for prediction of fake news from social media users. *IEEE Transactions on Computational Social Systems*.

Shehzaad Dhuliawala, Vilém Zouhar, Mennatallah El-Assady, and Mrinmaya Sachan. 2023. A diachronic perspective on user trust in AI under uncertainty. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 5567–5580. Association for Computational Linguistics.

Júlia Falcão, Claudia Borg, Nora Aranberri, and Kurt Abela. 2024. COMET for low-resource machine translation evaluation: A case study of English-Maltese and Spanish-Basque. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 3553–3565, Torino, Italia. ELRA and ICCL.

Christian Federmann. 2018. Appraise evaluation framework for machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, 86–88. Association for Computational Linguistics.

Patrick Fernandes, 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 *Proceedings of the Eighth Conference on Machine Translation*, 1066–1083. Association for Computational Linguistics.

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

Markus Freitag, David Grangier, and Isaac Caswell. 2020. BLEU might be guilty but references are not innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 61–71. Association for Computational Linguistics.

Markus Freitag, 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 *Proceedings of the Eighth Conference on Machine Translation*, 578–628. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, and Nitika Mathur. 2015. Accurate evaluation of segment-level machine translation metrics. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1183–1191. Association for Computational Linguistics.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2023. xCOMET: Transparent machine translation evaluation through fine-grained error detection.

W David Holford. 2022. Design-for-responsible algorithmic decision-making systems: a question of ethical judgement and human meaningful control. *AI and Ethics*, 2(4):827–836.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, 1–42. Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, 1–45. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, 768–775. Association for Computational Linguistics.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024. Error Span Annotation: A balanced approach for human evaluation of machine translation.

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

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica*, 12:0455–463.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2018. RankME: Reliable human ratings for natural language generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 72–78. Association for Computational Linguistics.

Gaia Pavoni, Massimiliano Corsini, Federico Ponchio, Alessandro Muntoni, Clinton Edwards, Nicole Pedersen, Stuart Sandin, and Paolo Cignoni. 2022. Taglab: Ai-assisted annotation for the fast and accurate semantic segmentation of coral reef orthoimages. *Journal of Field Robotics*, 39(3):246–262.

Parker Riley, Daniel Deutsch, George Foster, Viresh Ratnakar, Ali Dabirmoghaddam, and Markus Freitag. 2024. Finding replicable human evaluations via stable ranking probability.

Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. 2023. Artificial artificial artificial intelligence: Crowd workers widely use large language models for text production tasks. *arXiv preprint arXiv:2306.07899*.

Vilém Zouhar and Ondřej Bojar. 2024. Quality and quantity of machine translation references for automatic metrics. In *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval) @ LREC-COLING 2024*, 1–11, Torino, Italia. ELRA and ICCL.

Vilém Zouhar, Shuoyang Ding, Anna Currey, Tatyana Badeka, Jenyuan Wang, and Brian Thompson. 2024. Fine-tuned machine translation metrics struggle in unseen domains.

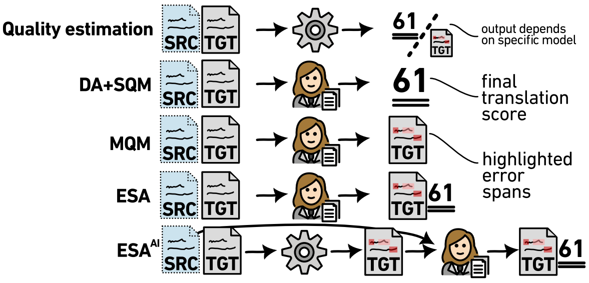


Figure 6: Overview of inputs and outputs of various machine translation evaluation approaches.

## A Subset Consistency Formalization

This section justifies the setup in Section 4.2 and is reminiscent of the work of Riley et al. (2024). A key distinction is that we are considering ranking stability with respect to the protocol itself. We do so by bootstrapping subsets of the data.

Our goal is tho show that a protocol with lower annotation error has higher system-level ranking accuracy. We assume that the annotation schemes are not biased towards a particular system but are noisy. We also assume a simplified model of system performance, where the annotation output $y_{m,i}$ of system $m$ on segment $i$ can be approximated by the system ability $a_m$ (e.g. average across a real life distribution) from which segment-specific variance $d_i$ is subtracted and error term $\epsilon$ is added. The annotation output $y_{m,i}$ is dependent on the specific annotation scheme, which is not indicated for brevity. We would like to find the system abilities $a_m$ but we only have access to $y_{m,i}$. This notation can also be extended to a collection of segments $I$:

$$y_{m,i} = a_m - d_i + \epsilon_{m,i} \tag{1}$$

$$Y_{m,I} = \frac{\sum_{i\in I} y_{m,i}}{|I|} \tag{2}$$

$$= a_m - \frac{\sum_{i\in I} d_i}{|I|} + \frac{\sum_{i\in I} \epsilon_{m,i}}{|I|} \tag{3}$$

On a large enough set of segments with the law of large numbers, we can assume $\frac{\sum_{i\in I}\epsilon_{m,i}}{|I|} \approx 0$ as $\epsilon$ is unbiased. If we want to estimate $\epsilon_{m,i}$, we could subtract from sample $i$ the average from all dataset, $Y_{m,D}$. Unfortunately, this would still leave the segment-specific difference $d_i$:

$$y_{m,i} - Y_{m,D} = -d_i + \epsilon_{m,i} \tag{4}$$

To separate $\epsilon_{m,i}$, we could consider subsets $I \subsetneq D$ for which $\frac{\sum_{i\in I} d_{m,i}}{|I|} \approx 0$ but $\frac{\sum_{i\in I}\epsilon_{m,i}}{|I|} \not\approx 0$. Apart from the difficulty of finding such subsets, our goal is to have a good estimation of the ranking of the systems. For this, we define system ordering $>_I$ given by the observed subset $I$:

$$m_1 >_I m_2$$

$$\overset{\text{def}}{\Leftrightarrow} \frac{\sum_{i\in I} y_{i,m_1}}{|I|} > \frac{\sum_{i\in I} y_{i,m_2}}{|I|} \tag{5}$$

$$\Leftrightarrow \sum_{i\in I} y_{i,m_1} > \sum_{i\in I} y_{i,m_2} \tag{6}$$

$$\Leftrightarrow a_{m_1} - \sum_{i\in I} d_i + \sum_{i\in I} \epsilon_{i,m_1} > a_{m_2} - \sum_{i\in I} d_i + \sum_{i\in I} \epsilon_{i,m_2} \tag{7}$$

$$\Leftrightarrow a_{m_1} + \sum_{i\in I} \epsilon_{i,m_1} > a_{m_2} + \sum_{i\in I} \epsilon_{i,m_2} \tag{8}$$

Notice that $>_I$ is independent of the segment-specific term $d_i$ because both systems are evaluated on the same segments. We compare this empirical ordering with that of the true system ranking. This is done across a set of systems $\mathcal{M}$ using pairwise accuracy, i.e. how many system pairs are ranked in the same way as by the true system ranking:

$$\textsc{Acc}(I) \overset{\text{def}}{=} \sum_{m_1,m_2\in\mathcal{M}} \frac{\mathbb{1}[(m_1 >_I m_2) \Leftrightarrow (a_{m_1} > a_{m_2})]}{|\mathcal{M}|^2} \tag{9}$$

With higher accuracy we can assume that the relative $\epsilon$ is lower, at least for the purposes of ordering. This is because **if** the accumulated error terms are low (10), the indicator in Equation (9) is true (11), which is **equivalent** to high accuracy (12):

$$\sum_{i\in I} \epsilon_{i,m_1} \to 0 \land \sum_{i\in I} \epsilon_{i,m_2} \to 0 \quad \Rightarrow \tag{10}$$

$$\Big(a_{m_1} + \sum_{i\in I} \epsilon_{i,m_1} > a_{m_2} + \sum_{i\in I} \epsilon_{i,m_2} \Leftrightarrow a_{m_1} > a_{m_2}\Big) \tag{11}$$

$$\Leftrightarrow \quad \textsc{Acc}(I) \to 1 \tag{12}$$

To obtain Acc, we would need to know if $a_{m_1} > a_{m_2}$. In our setup, we do not know this true ranking and obtaining it would require large-scale super-human annotations. However, for large-enough $I$, we can assume that $\frac{\sum_{i\in I}\epsilon_{m,i}}{|I|} \approx 0$. Therefore, for the true ordering, we use the ordering by that particular annotation scheme on all data. Now we established a link between accumulated annotation noise, $\sum_{i\in I} \epsilon_{i,m}$, and accuracy, which we can measure.

The accuracy will be high if the error terms are low and therefore the annotations are of high quality. This can be used to measure the annotation protocol usefulness. In addition, this has practical implications as we could solicit fewer annotations to obtain the same results as if we had more.

## B Learning to Annotate

Kocmi et al. (2024) showed that despite ESA being faster than MQM, the users *learn* to perform the MQM annotations slightly faster. We show similar results in Figure 7, though in our case the workers learn to perform the post-editing of GEMBA error span annotation slightly faster (0.18s faster with every segment). This effect is present despite the ESA annotators being at an advantage because there was fewer of them and they thus each individually processed more segments. Even though the speedup happens thorough the whole annotation, it is mostly present in the first 15% of segments (green box in Figure 7). For ESA, this is -2.1s/segment and for ESA$^{\text{AI}}$ this is -1.9s/segment. In addition, users in the post-editing task seem to be more consistent. For ESA, the user's deviation from their personal average is 43.3s, while for post-editing GEMBA this is only 32.1s. Overall, this makes the human effort more consistent and predictable but also showcases that the nature of the annotation task changes.

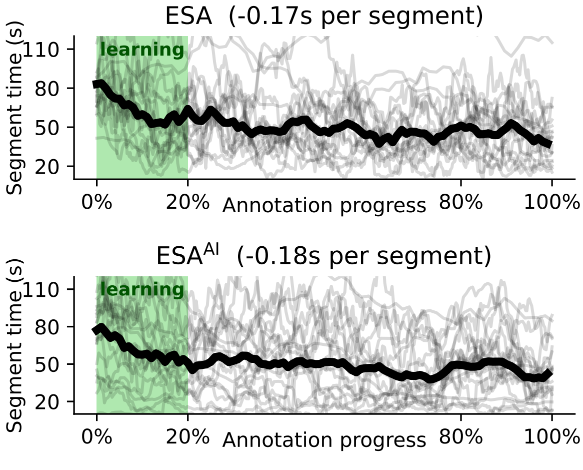


Figure 7: Time per segment with respect to progression in the annontation. Each annotator is the gray faint line and their average is in black. The lines are smoothed with a window of size 15 segments. We also compute the average speed at the beginning and at the end, which yields the *learned speedup*. This is how much the annotator speeds up per working on one segment.

We now examine what specifically makes some segments take longer than others. We do so using feature-level correlations as shown in Table 3. Many of these features are co-dependent. For example, the longer the translation, the more likely GEMBA finds more error spans and the lower the final score. Nevertheless it gives us insights that ESA$^{\text{AI}}$ users learn to become faster. The number of words in the translation, together with the number of error spans is a strong predictor of annotation time. For MQM this is the highest, which can be explained by each error span requiring the most work in the MQM annotation scheme because the annotators have to also assign the error type. The longer the document (number of translation paragraphs), the lower the annotation time, which is likely due to shared context. With longer documents, the annotator does not have to switch between domains and contexts. Contrastively, the ESA$^{\text{AI}}$ annotators are slightly less affected by the translation length in contrast to ESA.

| | **MQM** | **ESA** | **ESA$^{\text{AI}}$** |
|---|---|---|---|
| Progress | -0.12 | -0.13 | -0.13 |
| Translation word count | 0.30 | 0.19 | 0.16 |
| GEMBA error spans | 0.12 | 0.07 | 0.12 |
| Error spans | 0.06 | 0.04 | 0.12 |
| Score | -0.07 | -0.03 | -0.08 |
| Document size | -0.14 | -0.17 | -0.17 |

Table 3: Individual Pearson correlation between features and annotation times. The higher the absolute value, the more it affects the annotation time.

## C User Guidelines

The following are are annotation guidelines for our two local ESA$^{\text{AI}}$ campaigns, which is closely based on the setup of Kocmi et al. (2024).

**Highlighting errors:** Highlight the text fragment where you have identified a translation error (drag or click start & end). Click repeatedly on the highlighted fragment to increase its severity level or to remove the selection.

- **Minor Severity:** Style/grammar/lexical choice could be better/more natural.
- **Major Severity:** Seriously changed meaning, difficult to read, decreases usability.

If something is missing from the text, mark it as an error on the **[MISSING]** word. The highlights do not have to have character-level precision. It's sufficient if you highlight the word or rough area where the error appears. Each error should have a separate highlight.

**Score:** After highlighting all errors, please set the overall segment translation scores. The quality levels associated with numerical scores on the slider:

- 0%: No meaning preserved: Nearly all information is lost in the translation.

- 33%: Some meaning preserved: Some of the meaning is preserved but significant parts are missing. The narrative is hard to follow due to errors. Grammar may be poor.
- 66%: Most meaning preserved and few grammar mistakes: The translation retains most of the meaning. It may have some grammar mistakes or minor inconsistencies.
- 100%: Perfect meaning and grammar: The meaning and grammar of the translation is completely consistent with the source.

| GEMBA | ——Removed—— | | | No edit | ——Added—— | | | |
|---|---|---|---|---|---|---|---|---|
| #err. (freq.) | =2 | =1 | =0 | | =0 | =1 | =2 | $\geq$3 |
| 0 (23.8%) | 0% | 0% | 100% | 88% | 88% | 8% | 2% | 2% |
| 1 (38.0%) | 0% | 28% | 72% | 62% | 81% | 14% | 3% | 3% |
| 2 (18.8%) | 15% | 16% | 69% | 54% | 71% | 13% | 9% | 7% |
| 3 (10.4%) | 11% | 20% | 62% | 51% | 68% | 16% | 7% | 10% |
| 4 (8.9%) | 11% | 13% | 69% | 54% | 65% | 13% | 10% | 12% |

Table 4: Distribution of error span post-editing based on original GEMBA-reported error spans (2nd column). Percentages in the table are proportions within the number of GEMBA error spans. For example, second row shows that 62% of segments with exactly one GEMBA error span received no post-editing from annotators and in 28% the annotators removed the single error. ESA is comparable to ESA$^{\text{AI}}$.

| Operation | | Frequency |
|---|---|---|
| Severity change | | 12.0% |
| | Increase severity | 60.0% |
| | Decrease severity | 40.0% |
| Move span $\leq$5 | | 13.1% |
| Move span $\leq$10 | | 17.2% |
| Move span $\leq$20 | | 23.3% |
| Resize | Increase error span size | 21.5% |
| | Decrease error span size | 78.5% |

Table 5: Distribution of two ESA$^{\text{AI}}$ post-editing types: changing the severity, and moving the error span. A span is considered to be *moved* if the distance between old and new endpoints is at most 5, 10, or 20 characters. Many GEMBA errors are only misplaced or have the wrong severity. See specific cases in Example 2.

**SRC**: *Sie haben gestern das Treffen wieder verschoben.*
**TGT**: *He postponed the meeting again yesterday.*
**TGT$^{\mathbf{P}}$**: *He postponed the meeting squirrels are never.*

Example 1: An example of a perturbed translation **TGT$^{\mathbf{P}}$** based on the original system translation **TGT**. GEMBA correctly annotated the error span he (correctly the pronoun is *they*) but the perturbed part is left intentionally unannotated as an attention check.

| | | Original | Perturbed | OK |
|---|---|---|---|---|
| **ESA** | Score | 79.5 | 52.6 | 86% |
| | Span count | 0.85 | 1.86 | 54% |
| | Perturbation marked | | | 56% |
| **ESA$^{\text{AI}}$** | Score | 75.8 | 52.6 | 76% |
| | Span count | 2.19 | 4.48 | 61% |
| | Perturbation marked | | | 71% |

Table 6: Annotations assigned to perturbed attention check items (either scores or number of spans). **OK** is percentage in how many cases the non-perturbed item received a higher score or had fewer error spans, and how often the pertrubed span was marked by the annotator.

| Scoring | ESA | ESA$^{\text{AI}}$ |
|---|---|---|
| direct score | 0.376 | 0.533 |
| from spans | 0.327 | 0.671 |

Table 7: Inter-annotator agreement with direct scores and scores computed from error spans with MQM formula, as measured with Spearman correlation. ESA$^{\text{AI}}$ $_{\text{spans}}$ have the highest inter-annotator agreement, which is however caused by the GEMBA pre-filling. Still, the scores from ESA$^{\text{AI}}$, solely by humans, have the highest inter-annotator agreement.

| Protocol/ method | Subset size 10 | 40 | 115 | 190 |
|---|---|---|---|---|
| ESA$^{\text{AI}}$ | 84.41% | 92.38% | 96.69% | 98.88% |
| ESA$^{\text{AI}}_{\text{spans}}$ | 85.69% | 93.43% | 97.46% | 99.49% |
| GEMBA$_{\text{spans}}$ | 85.73% | 93.10% | 96.86% | 98.94% |
| ESA | 81.86% | 90.26% | 95.52% | 98.52% |
| ESA$_{\text{spans}}$ | 78.11% | 88.28% | 94.48% | 97.94% |
| MQM$_{\text{spans}}$ | 77.19% | 86.30% | 93.89% | 98.50% |

Table 8: Specific values of Figure 5. Subset accuracy across annotation schemes. ESA$^{\text{AI}}$ $_{\text{spans}}$ has the highest subset consistency, though this is likely biased by the spans from GEMBA, which as 100% inter-annotator agreement. However, ESA$^{\text{AI}}$ (direct scores) is based solely on human scorings, which has the second-highest subset consistency of any protocol.

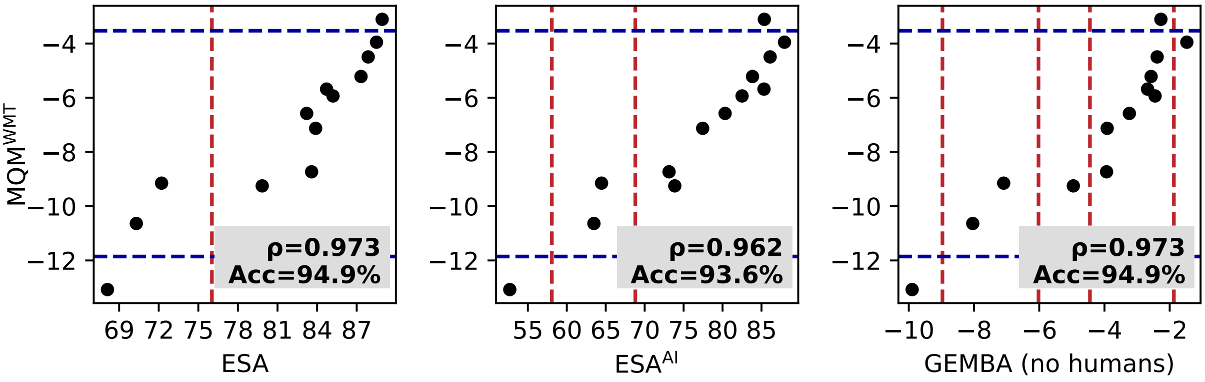


Figure 8: Each point is a system, with original MQM$^{\text{WMT}}$ scores on the $y$-axis against ESA, ESA$^{\text{AI}}$, and GEMBA before post-editing. Stripped lines indicate cluster separations with alpha threshold 0.05. Numbers show Spearman's correlations between the specific protocol and MQM$^{\text{WMT}}$. ESA and ESA$^{\text{AI}}$ have comparable system-level accuracy and correlations with MQM$^{\text{WMT}}$, making them equal in quality in this aspect.

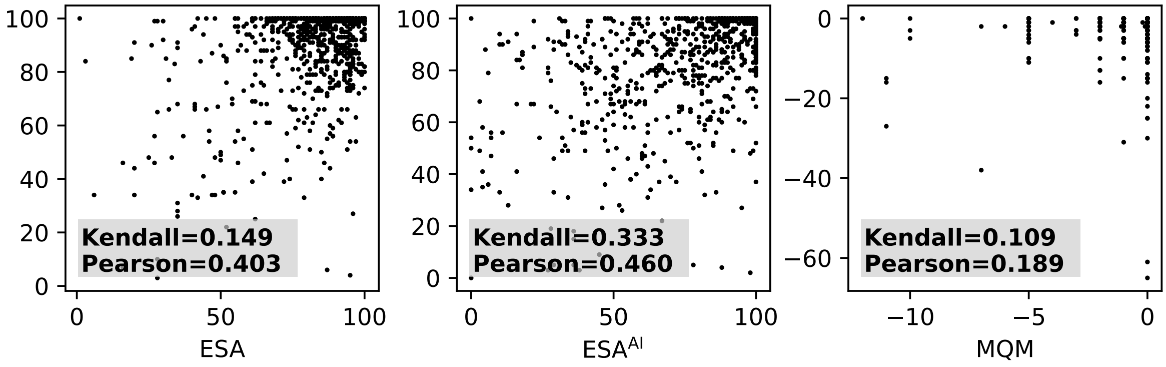
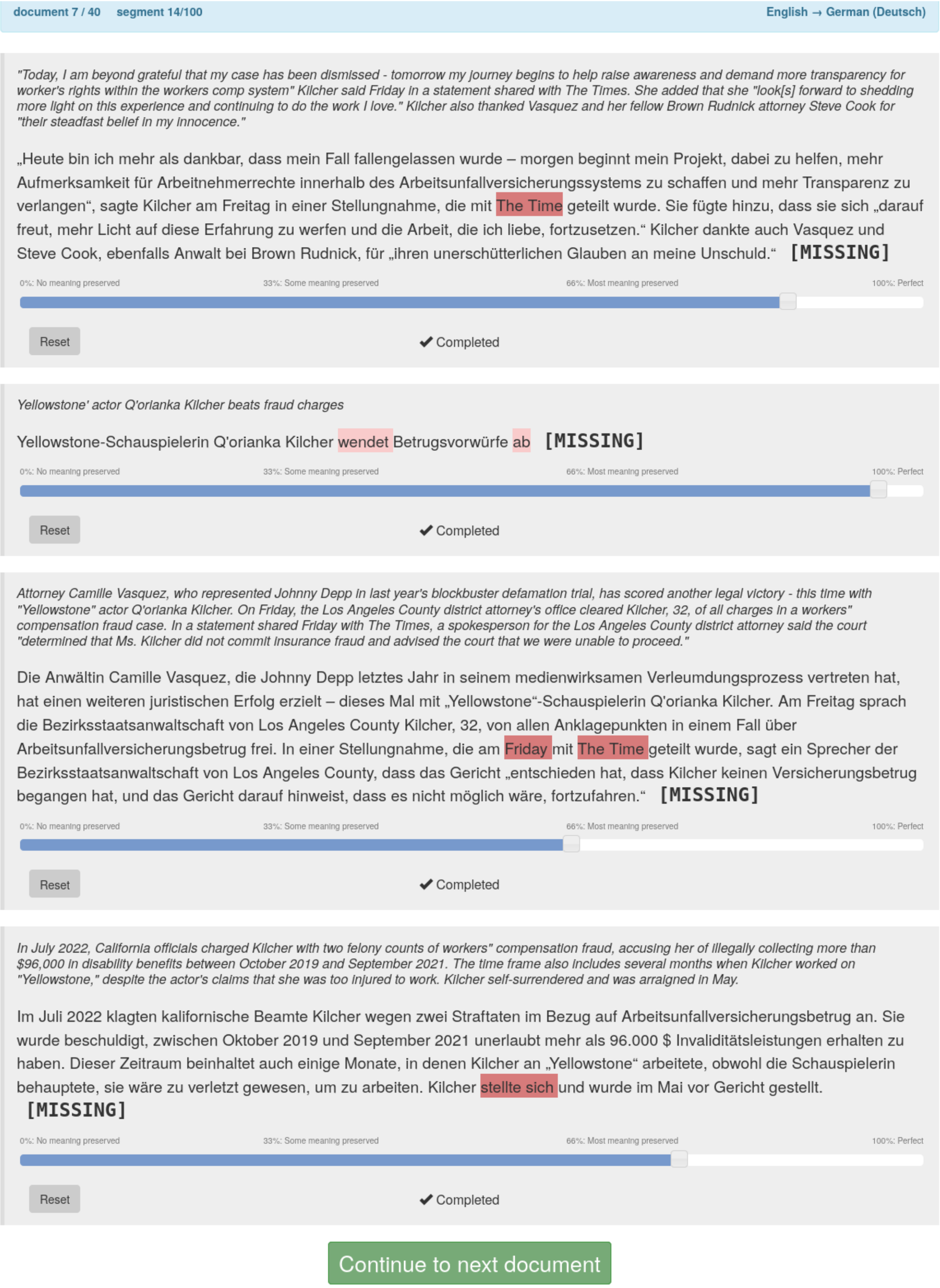


Figure 9: Changes in scoring by the same annotator when evaluated again. Each point represents single annotated segment with x-axis being annotator's score assigned in March and y-axis their score assigned in May. ESA$^{\text{AI}}$ has the highest *intra*-annotator agreement, showing another positive aspect of being primed by GEMBA.

| | | |
|---|---|---|
| **Increase severity** | **Source** | The physics are terrible and the people that created the game won't do anything about it |
| | **GEMBA** | Die Physik ist schrecklich und die Leute, die das Spiel entwickelt haben, werden nichts dagegen tun |
| | **ESA$^{\text{AI}}$** | Die Physik ist schrecklich und die Leute, die das Spiel entwickelt haben, werden nichts dagegen tun |
| **Decrease severity** | **Source** | Will not buy Mr. Coffee again |
| | **GEMBA** | Ich kaufe Mr. Kaffee nicht mehr. |
| | **ESA$^{\text{AI}}$** | Ich kaufe Mr. Kaffee nicht mehr. |
| **Move** | **Source** | However, I hate classes on fine arts and literature, and my school history bears it out. |
| | **GEMBA** | Aber ich hasse Kunst und Literatur, und meine Schulgeschichte bestätigt es. |
| | **ESA$^{\text{AI}}$** | Aber ich hasse Kunst und Literatur, und meine Schulgeschichte bestätigt es. [missing] |
| **Resize** | **Source** | [. . .] I'm not sure if that would work for this. |
| | **GEMBA** | [. . .] ich bin mir nicht sicher, ob das für diesen Zweck funktionieren würde. |
| | **ESA$^{\text{AI}}$** | [. . .] ich bin mir nicht sicher, ob das für diesen Zweck funktionieren würde. |

Example 2: Several post-editing operations from the collected data. Changing the severity (minor and major) is a very fast operation (only clicking the span), while moving and resizing are slow (removing the error span and creating a new one in its place takes up more of the annotator's time).



document 7 / 40 segment 14/100 English → German (Deutsch)

*"Today, I am beyond grateful that my case has been dismissed - tomorrow my journey begins to help raise awareness and demand more transparency for worker's rights within the workers comp system" Kilcher said Friday in a statement shared with The Times. She added that she "look[s] forward to shedding more light on this experience and continuing to do the work I love." Kilcher also thanked Vasquez and her fellow Brown Rudnick attorney Steve Cook for "their steadfast belief in my innocence."*

„Heute bin ich mehr als dankbar, dass mein Fall fallengelassen wurde – morgen beginnt mein Projekt, dabei zu helfen, mehr Aufmerksamkeit für Arbeitnehmerrechte innerhalb des Arbeitsunfallversicherungssystems zu schaffen und mehr Transparenz zu verlangen“, sagte Kilcher am Freitag in einer Stellungnahme, die mit The Time geteilt wurde. Sie fügte hinzu, dass sie sich „darauf freut, mehr Licht auf diese Erfahrung zu werfen und die Arbeit, die ich liebe, fortzusetzen.“ Kilcher dankte auch Vasquez und Steve Cook, ebenfalls Anwalt bei Brown Rudnick, für „ihren unerschütterlichen Glauben an meine Unschuld.“ **[MISSING]**

*Yellowstone' actor Q'orianka Kilcher beats fraud charges*

Yellowstone-Schauspielerin Q'orianka Kilcher wendet Betrugsvorwürfe ab **[MISSING]**

*Attorney Camille Vasquez, who represented Johnny Depp in last year's blockbuster defamation trial, has scored another legal victory - this time with "Yellowstone" actor Q'orianka Kilcher. On Friday, the Los Angeles County district attorney's office cleared Kilcher, 32, of all charges in a workers" compensation fraud case. In a statement shared Friday with The Times, a spokesperson for the Los Angeles County district attorney said the court "determined that Ms. Kilcher did not commit insurance fraud and advised the court that we were unable to proceed."*

Die Anwältin Camille Vasquez, die Johnny Depp letztes Jahr in seinem medienwirksamen Verleumdungsprozess vertreten hat, hat einen weiteren juristischen Erfolg erzielt – dieses Mal mit „Yellowstone“-Schauspielerin Q'orianka Kilcher. Am Freitag sprach die Bezirksstaatsanwaltschaft von Los Angeles County Kilcher, 32, von allen Anklagepunkten in einem Fall über Arbeitsunfallversicherungsbetrug frei. In einer Stellungnahme, die am Friday mit The Time geteilt wurde, sagt ein Sprecher der Bezirksstaatsanwaltschaft von Los Angeles County, dass das Gericht „entschieden hat, dass Kilcher keinen Versicherungsbetrug begangen hat, und das Gericht darauf hinweist, dass es nicht möglich wäre, fortzufahren.“ **[MISSING]**

*In July 2022, California officials charged Kilcher with two felony counts of workers" compensation fraud, accusing her of illegally collecting more than $96,000 in disability benefits between October 2019 and September 2021. The time frame also includes several months when Kilcher worked on "Yellowstone," despite the actor's claims that she was too injured to work. Kilcher self-surrendered and was arraigned in May.*

Im Juli 2022 klagten kalifornische Beamte Kilcher wegen zwei Straftaten im Bezug auf Arbeitsunfallversicherungsbetrug an. Sie wurde beschuldigt, zwischen Oktober 2019 und September 2021 unerlaubt mehr als 96.000 $ Invaliditätsleistungen erhalten zu haben. Dieser Zeitraum beinhaltet auch einige Monate, in denen Kilcher an „Yellowstone“ arbeitete, obwohl die Schauspielerin behauptete, sie wäre zu verletzt gewesen, um zu arbeiten. Kilcher stellte sich und wurde im Mai vor Gericht gestellt. **[MISSING]**

0%: No meaning preserved
33%: Some meaning preserved
66%: Most meaning preserved
100%: Perfect
Reset
✔ Completed
Continue to next document

Figure 10: Screenshot of the study interface implemented for Appraise. Multiple segments from a document are shown together for context. The AI suggests the initial error spans which the annotator post-edits and finally adds final score judgment.