# Findings of the WMT 2023 Shared Task on Quality Estimation

**Frédéric Blain[(1)], Chrysoula Zerva[(2,3)], Ricardo Rei[(3,4,5)], Nuno M. Guerreiro[(2,3,4,8)], Diptesh Kanojia[(6)], José G. C. de Souza[(4)], Beatriz Silva[(4)], Tânia Vaz[(4)], Yan Jingxuan[(4)], Fatemeh Azadi[(7)], Constantin Orăsan[(6)], André F. T. Martins[(2,3,4)]**

[(1)]Tilburg University, [(2)]Instituto de Telecomunicações, [(3)]Instituto Superior Técnico, [(4)]Unbabel, [(5)]INESC-ID, [(6)]University of Surrey, [(7)]University of Tehran, [(8)]MICS, CentraleSupélec, Université Paris-Saclay

f.l.g.blain@tilburguniversity.edu, {d.kanojia,c.orasan}@surrey.ac.uk, jose.souza@unbabel.com

ft.azadi@ut.ac.ir, {chrysoula.zerva,ricardo.rei,nuno.s.guerreiro,andre.t.martins}@tecnico.ulisboa.pt

## Abstract

We report the results of the WMT 2023 shared task on Quality Estimation, in which the challenge is to predict the quality of the output of neural machine translation systems at the word and sentence levels, without access to reference translations. This edition introduces a few novel aspects and extensions that aim to enable more fine-grained, and explainable quality estimation approaches. We introduce an updated quality annotation scheme using Multidimensional Quality Metrics to obtain sentence- and word-level quality scores for three language pairs. We also extend the provided data to new language pairs: we specifically target low-resource languages and provide training, development and test data for English-Hindi, English-Tamil, English-Telegu and English-Gujarati as well as a zero-shot testset for English-Farsi. Further, we introduce a novel fine-grained error prediction task aspiring to motivate research towards more detailed quality predictions.

## 1 Introduction

This edition of the shared task on Quality Estimation (QE) aims to build on previous editions and findings to further benchmark methods for estimating the quality of neural machine translation (MT) output at run-time, without the use of reference translations. It includes (sub)tasks that consider the quality of machine translations at word- and sentence-level.

Over the past years, the QE field has been moving towards explainable, large, multilingual models that have been shown to achieve high performance, especially at sentence-level (Specia et al., 2021; Zerva et al., 2022). The recent proliferation of Large Language Model (LLM) technology and the consequential performance improvements in MT elevate the significance of advancing methodologies for quality estimation. In light of this, emphasis should be placed on multilingual quality estimation, in particular for low- and medium-resource languages, necessitating the development of more precise and interpretable quality assessment techniques. Additionally, it is important to address the challenge of robustness to hallucinations, prioritise sustainability, and optimise computational efficiency. These considerations collectively contribute to progress toward trustworthy and dependable QE systems that could facilitate real-time, reliable assessments of translation quality.

In this edition of the shared task, we further expand the provided resources, introducing new low-resource language pairs for Indian languages, namely Marathi, Tamil, Telugu, Gujarati and Hindi, as well as Farsi and Hebrew. Following the previous editions, we provide both annotations for direct assessments (DA), post-edits (PE) and Multidimensional Quality Metrics (MQM) (Lommel et al., 2014). We describe in detail the annotation process and provide statistics for the different language pairs in Section 2.

Overall, in addition to advancing the state-of-the-art at all prediction levels, our main goals are:

- to extend the languages covered in our datasets with low- and medium-resource languages;
- to investigate the potential of fine-grained quality estimation;
- to investigate new multilingual and language independent approaches esp. with regards to zero-shot approaches;
- to study the robustness of QE approaches to hallucinations; and
- to continue monitoring the computational efficiency of proposed approaches for sustainability purposes.

We thus designed two tasks this year:



*Proceedings of the Eighth Conference on Machine Translation (WMT)*, pages 629–653
December 6–7, 2023. 

**Task 1** The core QE task, which consists of separate sentence-level and word-level sub-tasks. For the sentence-level sub-tasks, the goal is to predict a quality score for each segment in a given test set, which can be a variant of DA (§2.1) or MQM (§2.2). For the word-level sub-task, participants had to predict translation errors in the form of binary quality tags (see §3.1.3).

**Task 2** The fine-grained error prediction task, where participants were asked to detect error spans alongside error severities (§3.2).

The tasks make use of large datasets annotated by professional translators with either 0-100 DA scoring, post-editing or MQM annotations. We update the training and development datasets of previous editions and provide new test sets for Tasks 1 and 2. The datasets and models released are publicly available[1].

Apart from the data made available through the QE shared task, participants were also allowed to explore any additional data and resources deemed relevant, across tasks. In addition, LLMs could also be used both to extend resources and to complement predictions.

The shared task uses CodaLab as a submission platform, where each sub-task corresponds to a separate competition instance. Participants (Section 5) could submit up to a total of 10 submissions per sub-task. Results for all tasks evaluated according to standard metrics are given in Section 6. Baseline systems were trained by the task organisers and entered into the platform to provide a basis for comparison (Section 4). A discussion on the main goals and findings from this year's task is presented in Section 7.

## 2 Datasets

### 2.1 DA & Post-edit data:

For all language pairs, the data provided is selected from publicly available resources. Specifically for training, we used the language pairs from the MLQE-PE dataset (Fomicheva et al., 2022), as well as newly annotated data for languages spoken in India (Hindi, Tamil, Telugu and Gujarati). Overall, we provided training data for 15 language pairs with DA annotations, 12 with post edits, and 3 with MQM annotations, accounting for a mix of high, medium and low-resource languages. The statistics for the provided data are detailed in Table 1.

For the English-Marathi language pair included in the last edition, we provided a novel test set this year. To expand on language resources for the QE shared task, we chose Hindi (Hi) and Gujarati (Gu) as target languages from the Indo-Aryan language family, Tamil (Ta) and Telugu (Te) were chosen from the Dravidian language family. For En-Hi, En-Ta, En-Te, and En-Gu, dataset curation and annotation were performed with the help of professional translators who were native speakers of the target language. The annotators were provided with guidelines which discussed DA score ranges with various error types. Additionally, parallel segments were curated from the following parallel corpora: *i)* *Anuvaad* parallel corpus[2] (General, Healthcare and Legal domain; *ii)* IITB English-Hindi parallel corpus[3] (Kunchukuttan et al., 2018) (Culture/-Tourism domain), and parallel segments scraped from NPTEL[4]; and *iii)* SpokenTutorials[5] (Education domain). The curated segments were selected from the above-mentioned domains to ensure cross-domain impact and performance.

From the *Anuvaad* parallel corpus, we filtered source and parallel segments based on LaBSE (Feng et al., 2022) at high threshold values in the range [0.85, 0.99]. This helps us ensure the presence of good-quality reference translations from a noisy parallel corpus. We then selected source sentences for the dataset by varying token length in buckets of $0-10$, $10-20$, and $20-30$ tokens. This allows us to get annotations on various sentence lengths and helps manage the annotation cost to a certain extent. Moreover, translation models tend to generate erroneously over longer sequences (Varis and Bojar, 2021), and ensuring short and longer source sentences are a part of the data helps us presume a balanced DA distribution in the human annotation. We obtained the translation with the 1.3B parameter NLLB model (Costa-jussà et al., 2022) from HuggingFace[6]. The inference was performed with 5 beams, limiting the n-gram repetition to 2 and maximum length to 80 tokens, with early stopping enabled. The curation of source

[1] https://github.com/WMT-QE-Task/wmt-qe-2023-data

[2] https://github.com/project-anuvaad/anuvaad-parallel-corpus

[3] Unreleased parallel segments, to be released here in v3.2: https://www.cfilt.iitb.ac.in/iitb_parallel/

[4] https://nptel.ac.in/

[5] https://spoken-tutorial.org/

[6] https://huggingface.co/facebook/nllb-200-distilled-1.3B

segments from parallel corpora allowed us to compare the performance with IndicTrans (Ramesh et al., 2022) and 600M parameter NLLB model, in terms of TER and BLEU, helping us select the model and parameters above.

During the annotation, weekly validation of randomly selected instances was performed by an unbiased native speaker who provided feedback to further improve annotations during the data curation. After all three annotators performed the DA annotations, we separated the data into training, development, and test sets while filtering for a balanced distribution of DA scores across all sets.

For the En-Fa dataset, we used the post-edited data provided in Azadi et al. (2022) to get the word-level quality annotations. It contains 1K sentences derived from some English scientific articles in the domains of technology, computer science, and humanities. These sentences were firstly translated to Farsi, using an RNN-based commercial MT system named Faraazin[7]. Then, each sentence was given to a professional human translator to be post-edited and provide the correct translation with minimum edits. These post-edits were finally validated by another annotator to ensure their quality.

### 2.2 MQM Data

As **training data**, we used the annotations released for the Metrics and QE shared tasks in the previous years (Freitag et al., 2021a,b). Together, these annotations, cover 3 high-resource language pairs, namely: Chinese-English (Zh-En), English-German (En-De) and English-Russian (En-Ru), and span across two domains (News and Ted Talks).

As **test data**, we annotated new evaluation sets for three language directions. A low-resource language pair, Hebrew-English (He-En), and two high-resource language pairs, English-German and Chinese-English. The evaluation sets were annotated by professional translators following a MQM typology (Burchardt, 2013) and specific guidelines[8].

The documents used for the evaluation sets are shared with the General MT task in WMT and follow the same distribution of domains in that data. These documents were translated using the NLLB (Team et al., 2022) model of 1.3B parameters[9], the same model used in Section 2.1. We note that the En-De sources were originally organised in document-level, and we opted for converting them to smaller segments, so that we do not divert from the processing applied for the other LPs. Hence we first applied sentence splitting and then followed the same translation and annotation process described in this section.

All evaluation sets were annotated by professional translators and, for En-De and Zh-En the annotations were reviewed by a separate group of professional translators that amended any incoherences or disagreements from the first round of annotation. Regarding the domains of the data, for He-En, two different evaluation sets were annotated, one with newswire articles and another from product user reviews. For En-De, documents from four domains were annotated: transcriptions of meetings, newswire articles, social media posts, and product user reviews. For Zh-En, documents from three domains were used: manuals from information technology software or devices, newswire articles and product user reviews.

## 3 Quality Estimation tasks

In what follows, we briefly describe each sub-task, including the datasets provided for them.

### 3.1 Task 1: Predicting translation quality

The ability to accurately estimate the quality of translations on sentence- or word-level on-the-fly, i.e., without access to human-references is at the core of the QE shared task. Sentence and word-level estimates can provide complementary views of the quality of a sentence capturing different aspects (e.g. overall fluency versus specific mistranslations).

Following last edition, the data was produced in the following ways:

1. DA sentence level scores: The quality of each source-translation pair is annotated by at least 3 independent expert annotators, using DA on a scale 0-100.

2. MQM approach: Each source-translation pair is evaluated by at least 1 expert annotator, and errors identified in texts are highlighted and classified in terms of severity (minor, major, critical) and type (grammar correctness, omission, style, mistranslation, among others). We use this information for both word and sentence level quality scores.

[7] https://www.faraazin.ir/

[8] http://bit.ly/mqm-guidelines

[9] Model identifier FACEBOOK/NLLB-200-1.3B

3. Post-editing approach: The translation is post-edited to obtain the closest possible, fully correct translation of the source. By considering the alignment between the source, translation and post-edited sentence, we can propagate the errors from the source to the translated sentence and annotate the segments that were potentially mistranslated and/or not translated at all. We use this information to infer word-level quality scores.

The DA and MQM sentence level annotations were further processed to obtain normalised quality scores that have the same direction between high and low quality. We provide more details on the required pre-processing in §3.1.1 and §3.1.3.

#### 3.1.1 Sentence-level quality prediction

This year we used a single competition instance both for DA and MQM-derived annotations aiming to motivate the submission of models that are robust to both annotation formats. To that end, we also aligned the scores by processing and normalising them as follows:

- For the **DA** scores we standardize the scores with respect to each annotator and then compute the mean average of standardized scores for each sentence.
- For the **MQM** scores we need to first compute the overall score from the individual errors. Hence for each annotator, we first compute the sentence-level score as

$$MQM^{sent}(hyp) = \frac{100 - \sum_{e \in hyp} severity(e)}{|hyp|}, \quad (1)$$

where $hyp$ is a hypothesis sentence represented as a sequence of tokens, $e$ is an error annotated in that sentence and the $severity$ is computed but adding:

  + 1 point for minor errors
  + 5 points for major errors
  + 10 points for critical errors

To align with DA annotations we subtract the summed penalties from 100 (perfect score) and we then divide by the sentence length (computed as number of words). We then normalise per annotator as in the DA case and compute the mean average in the case of multiple annotators.

Regarding evaluation, systems in this task (both for DA and MQM) are **evaluated against the true z-normalised sentence scores using Spearman's rank correlation coefficient $\rho$ as the primary metric**. This is what was used for ranking system submissions. Pearson's correlation coefficient, $r$, and Kendall $\tau$ were also computed as secondary metrics but not used for the final ranking of systems.

#### 3.1.2 Hallucinations

Hallucinations are highly pathological translations that contain content that is detached from the source (Raunak et al., 2021). As such, they can have devastating impact when models are deployed *in the wild* for real-world applications. Quality estimation systems are an appealing and attractive strategy to identify and flag these translations before they reach end-users. However, recent research has found that QE models may not appropriately penalize hallucinations and other critical errors (Raunak et al., 2022; Guerreiro et al., 2023c). This concern is further amplified for low-resource languages, where this undesirable behavior may arise even more frequently (Dale et al., 2023b). As such, in this edition of the shared task, we created data to assess the capability of submitted QE models in detecting hallucinations.

The data was created through a three-step process: (i) we started by generating translations for all language pairs of this year's shared task with NMT models[10], using the FLORES `devtest` and `test` splits (Goyal et al., 2022), as well as WikiMatrix data available through OPUS (Schwenk et al., 2019); then (ii) we automatically detected hallucinations generated by the models; and finally (iii) manually verified the flagged translations in order to guarantee that they are hallucinations. To automatically detect the hallucinations, we followed the procedure from Guerreiro et al. (2023a), which directly draws from several relevant contributions from research works in the literature of hallucination detection (Ferrando et al., 2022; Dale et al., 2023a; Guerreiro et al., 2023c).

To evaluate the performance of the submissions, we created, for each language pair, an *evaluation* set that consists of: all the hallucinations for the language pair, and the samples whose gold score is above the 25$^{th}$ percentile. This is to ensure that the non-hallucinations in the evaluation set are not

[10]We used the massively multilingual models (175M and 615M parameters) released in Goyal et al. (2022).

| | Language Pairs | Sentences Train / Dev / Test23 | Tokens Train / Dev / Test23 | DA | PE | MQM | Data Source | Release |
|---|---|---|---|---|---|---|---|---|
| DA & post edits | En-De | 10,000 / – / – | 148,044 / – / – | ✓ | ✓ | | Wikipedia | 2021/22 |
| | En-Zh | 10,000 / – / – | 148,529 / – / – | ✓ | ✓ | | Wikipedia | 2021/22 |
| | Ru-En | 10,000 / – / – | 105,871 / – / – | ✓ | ✓ | | Reddit | 2021/22 |
| | Ro-En | 10,000 / – / – | 154,825 / – / – | ✓ | ✓ | | Wikipedia | 2021/22 |
| | Et-En | 10,000 / – / – | 126,547 / – / – | ✓ | ✓ | | Wikipedia | 2021/22 |
| | Ne-En | 10,000 / – / – | 135,095 / – / – | ✓ | ✓ | | Wikipedia | 2021/22 |
| | Si-En | 10,000 / – / – | 140,932 / – / – | ✓ | ✓ | | Wikipedia | 2021/22 |
| | Ps-En | 2,000 / – / – | 54,459 / – / – | ✓ | ✓ | | Wikipedia | 2021/22 |
| | Km-En | 2,000 / – / – | 44,029 / – / – | ✓ | ✓ | | Wikipedia | 2021/22 |
| | En-Ja | 2,000 / – / – | 41,272 / – / – | ✓ | ✓ | | Wikipedia | 2021/22 |
| | En-Cs | 2,000 / – / – | 40,638 / – / – | ✓ | ✓ | | Wikipedia | 2021/22 |
| | En-Yo | 1,010 / – / – | 21,238 / – / – | ✓ | ✓ | | | 2021/22 |
| | En-Mr | 27,000 / 1,000 / 1,086 | 717,581 / 26,253 / 27,951 | ✓ | ✓ | | multi-domain/multi-corpus | 2022/23 |
| | En-Hi | 7,000 / 1,000 / 1,074 | 181,336 / 25,943 / 28,032 | ✓ | | | multi-domain/multi-corpus | 2023 |
| | En-Gu | 7,000 / 1,000 / 1,075 | 153,685/ 21,238 / 23,084 | ✓ | | | multi-domain/multi-corpus | 2023 |
| | En-Ta | 7,000 / 1,000 / 1,067 | 150,670 / 21,655/ 20,342 | ✓ | | | multi-domain/multi-corpus | 2023 |
| | En-Te | 7,000 / 1,028 / 1,000 | 147,492 / 20,686 / 22,640 | ✓ | | | multi-domain/multi-corpus | 2023 |
| | En-Fa | – / – / 1,000 | –/ – / 26,807 | | ✓ | | news (multi-domain) | 2023 |
| MQM | En-De | 30,425 / – / 1,897 | 877,066 / – / 37,996 | | | ✓ | multi-domain | 2021/23 |
| | En-Ru | 17,144 / – / – | 395,045 / – / – | | | ✓ | multi-domain | 2021/22 |
| | Zh-En | 36,851 / – / 1,675 | 1,654,454 / – / 39,770 | | | ✓ | multi-domain | 2021/23 |
| | He-En | – / – / 1,182 | – / – / 35,592 | | | ✓ | multi-domain | 2023 |

Table 1: Statistics of the data used for Task 1 and Task 2. The number of tokens is computed based on the source sentences. Hallucinated data included in the calculations for the 2023 testsets.

highly pathological translations (they may however be incorrect translations). We report the Area Under the Receiver Operating Characteristic curve (AUROC) and Recall at $k$ (R@$k$), where $k$ is defined as the number of hallucinations in the evaluation set. A perfect QE detector would have 100 AUROC and 100% Recall at $k$. We report the statistics of the evaluation sets in Table 8.

#### 3.1.3 Word-level quality prediction

This sub-task focuses on detecting word-level errors in the MT output. The goal is to automatically predict the quality of each token using a binary decision, i.e., using OK as a label for tokens translated correctly and BAD otherwise.

We follow the annotation conventions of the previous edition, i.e., we do not consider source-side annotations, and incorporate omission errors to the target token annotations. Specifically, to account for omission errors, we consider the following convention: the token on the right side of the omitted text in the translation is annotated as "BAD". An additional `<EOS>` token is appended at the end of every translation segment to account for omissions at the end of each sentence. This allows the provision of a unified framework for both the post-edit originated annotations and the MQM annotations.

We thus use the same source-translation pairs used for the sentence-level tasks and obtain the binary tags as follows:

- For post-edited data, we use the methodology to obtain *translation error distance* (TER) scores (Snover et al., 2006) to obtain alignments between translation and post-edit and annotate the misaligned tokens as BAD.
- For MQM data, the tokens that fall within the text-spans annotated as errors (or any severity or category) are annotated as BAD. If the whitespace between two words is annotated as an error, then this is considered an omission, and the next token is annotated as BAD.

For the word-level task, **submissions are ranked using the Matthews Correlation Coefficient (MCC, Matthews, 1975) as the primary metric**, while F1-scores are provided as complementary information.

### 3.2 Task 2: Fine-grained error detection

For this task we attempt to focus on finer-grained quality predictions, taking advantage of the detailed information provided in the MQM annotation schema. Specifically, the MQM schema allows the annotation of additional information for each identified error. Specifically, each error span is annotated with error severity (*minor*, *major*, *critical*)

as well as error type (see also Figure 1). Such information allows for a more detailed analysis of the errors of MT systems, an understanding of their failure points and can provide the basis towards more explainable quality estimation.

Ideally, a fine-grained QE system is expected to be able to predict both the error type and its corresponding severity. However, the cardinality of error categories, the complexity of disentangling between them and the scarcity of MQM annotations render such a classification task particularly challenging. Hence, in this first attempt, we chose to focus only on **error severity** and merged together the major and the critical labelled errors due to the scarcity of the latter (see Table 2). As a result, we aimed to classify error spans as either *minor* or *major*.

| LP | minor | major | critical |
|---|---|---|---|
| En-De | 652 | 595 | 81 |
| Zh-En | 633 | 1063 | 242 |
| He-En | 792 | 1837 | 3 |

Table 2: Error severities (original: before merging critical and major severities) for the 2023 MQM test set.

As such, the information used for this task consists of: *i*) start and end index positions for each error span; and *ii*) the simplified error severity. The error spans are identified as sequences of continuous characters within a target hypothesis, allowing for annotations of single white spaces and punctuation marks in order to account for omission and punctuation errors respectively. Aiming to mimic the human annotations and simplify the task, overlapping error spans were allowed. Figure 1 shows an example of annotations.

For the evaluation, the primary metric was **F1-score**, computed on the character level and weighted to allow for half points for correctly identified span but misclassified severity. Precision and recall were also provided as complementary metrics. The evaluation approach is inspired by (Fonseca et al., 2019) but does not consider document-level annotations. With respect to overlapping annotations, we allow for multiple character level annotations [11] and will consider the best matching annotation per character position. As such for each segment we compute recall for the characters in

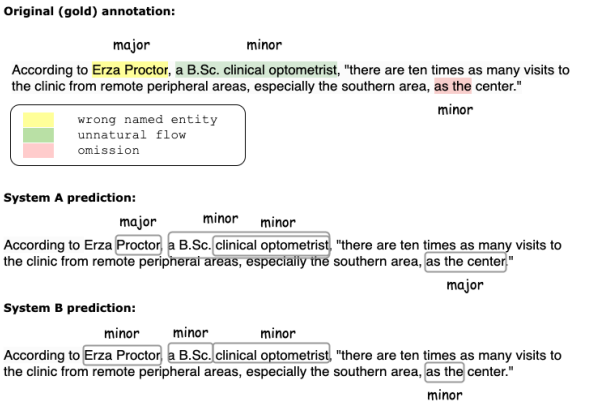


Figure 1: Example of gold annotations (MQM) for Task 2 (top) and respective prediction examples (bottom). Example taken from He-En test set.

gold annotation text spans, by computing the ratio between the overlap with system error spans and the gold error span length and weighting severity mismatches by 0.5. Respectively, we compute precision with respect to the system error span length and apply the same weighting convention (down-weighting by 0.5 for mismatched error severities). Figure 1 and Table 3 shows an example of the aforementioned process [12].

## 4 Baseline systems

### 4.1 Task 1: quality estimation

For the **sentence-level** sub-task, we opted for COMET-QE models (Rei et al., 2021) respectively pretrained on the DA and MQM QE data from WMT'21. Models are publicly available to download[13].

For the **word-level** sub-task, we trained a simplified architecture inspired by COMETKIWI (Rei et al., 2022a). More specifically, we used the multi-task architecture combining the sentence-level target and the binary word-level targets. However, we did not pretrain on HTER scores or Metrics data, and we skipped the few-shot language adaptation and language-specific tuning of task weights. The architecture of the baseline model is shown in Figure 2. The list of hyperparameters and their corresponding values can be found in appendix A.

[11] The gold data was processed to remove identical segments that correspond to the same span but have different error categories, but it preserved any partially overlapping segments that correspond to different error categories and/or severities.

[12] The link to evaluation scripts can be found at: https://github.com/WMT-QE-Task/qe-eval-scripts/blob/main/wmt23/

[13] https://wmt-qe-task.github.io/subtasks/task1/

| Systems | Precision | Recall | F1-score |
|---|---|---|---|
| System A | $\frac{1*7+1*28+0.5*6}{7+28+13} = 0.79$ | $\frac{1*7+1*28+0.5*6}{12+28+6} = 0.83$ | 0.81 |
| System B | $\frac{0.5*12+1*28+0.5*6}{12+28+6} = 0.80$ | $\frac{1*12+1*28+0.5*6}{12+28+6} = 0.80$ | 0.80 |

Table 3: Example of Precision and Recall computations for each annotation in the example of Figure 1.

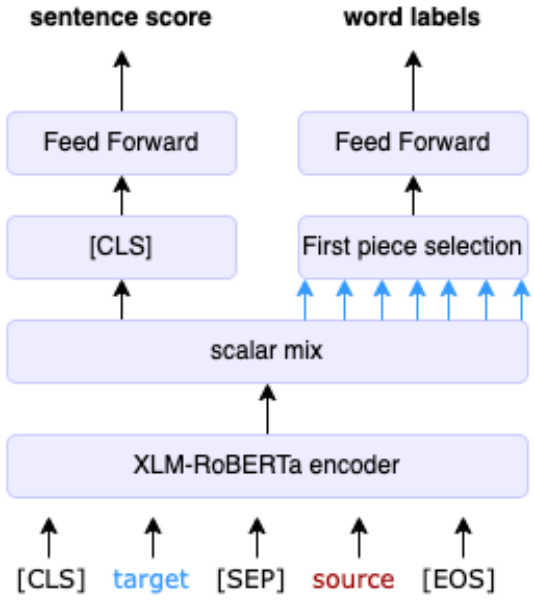


Figure 2: Baseline model for Task1 word level. Figure adapted from (Rei et al., 2022a)

### 4.2 Task 2: fine-grained error detection

For **Task 2** we constructed a simple baseline by using the same model used for Task 1 word-level, and post-process the predictions as follows:

- Detokenize the sentence
- Annotate continuous BAD tokens as a single text span
- Assume all errors are major

For all models, a large XLM-RoBERTa pre-trained encoder was used, without additional language tuning. The specific hyperparameters used are presented in Table 11.

## 5 Participants

Table 4 lists teams who officially took part to the QE shared task this year. In the remaining of this section, we report a brief system description gathered from each participant. For each team, we indicate the task(s) and sub-task(s) (*i.e.* language-pair(s)) they participated into.

**Bering Lab (T1-SL; En-De, Zh-En):** For each language pair, the team used an ensemble of the best three models from a pool of 10 models jointly trained for the word and sentence level tasks using a novel relative ranking loss function and Adversarial Weight Perturbation (AWP) to improve the robustness of the model. Using no additional pseudo-generated data, they pre-train the models using publicly available data from the previous WMT conferences that were augmented using the TER tool to generate binary word tags. They then fine-tuned 10 separate models on the labelled data from the WMT 2022 QE task randomly split into 10 folds. The models are fine-tuned in two steps. In the first step, the models are fine-tuned without AWP using the same objective as the pre-training step. Then, the best checkpoint from the first step is selected and tuned on the same objective, but with AWP. For the final submission they ensemble (average) the z-normalized scores from the top three models to get the final predictions.

**(T1-WL; En-De, Zh-En):** For each language pair, they use an ensemble of five models jointly trained for the word and sentence level tasks using a novel relative ranking loss function. Using no additional pseudo-generated data, they pre-train the models using publicly available data from the previous WMT conferences that were augmented using the TER tool to include word tags. They then split the labelled data from the WMT 2022 QE task into 20 folds and chose the best combination of five folds based on the minimum mean of the Kolmogorov-Smirnov goodness-of-fit scores between each validation set. Using these 5 folds, they fine-tune 5 final models using the same objective as the pre-training step. For the final prediction, the team chose the max score from each model for each token to get the final predictions.

**NJUNLP (T1 & T2; En-De):** Inspired by Direc-

tQE (Cui et al., 2021) and CLQE (Geng et al., 2023b), NJUNLP submission continues exploring pseudo data methods for QE. They generate pseudo MQM data using parallel data from the WMT translation task. Specifically, they replace the reference tokens with these tokens sampled from translation models. To simulate translation errors with different severities, they sample tokens with lower generation probabilities for worse errors. They pre-train the XLMR large model on pseudo MQM data, then fine-tune it on real QE data (including PE data). At both stages, they jointly learn sentence-level scores (MSE loss and margin ranking loss) and word-level tags (cross-entropy loss). For Task 1, the QE model outputs the sentence scores and the 'OK' probability of each token.

For Task 2, they set different thresholds for the 'OK' probability to predict fine-grained severity. They regard consecutive 'BAD' tokens as a whole span and take the worse severity of each token as the result. They train different models with different parallel data and ensemble their results as the final submission.

**HW-TSC (T1-SL; En-De, Zh-En, En-Mr, En-Hi, En-Ta, En-Te, En-Gu):** HW-TSC uses CrossQE, the same model as the one reported in (Tao et al., 2022), which consists of a multilingual base model and a task-specific downstream layer. The input is the concatenation of the source and the translated sentences. To enhance the performance, they finetuned and ensembled multiple base models using multilingual encoders such as XLM-RoBERTa, InfoXLM and RemBERT as well as a CometKiwi model. Moreover, they introduce a new corruption-based data augmentation method, which generates deletion, substitution and insertion errors in the original translation and uses a reference-based QE model to obtain pseudo scores.

**(T2; all):** For Task 2 they they convert the original word-level binary classification to a 3-way classification to adapt to Task 2 severities (no-error, minor, major). They then use a multitasking Comet model based on CometKiwi (Rei et al., 2023) which combines sentence scores and word-level tags using a weighted loss function. They set the weight of the sentence score sub-module to 0. They use InfoXLM-large and XLM-RoBERTa-large as the pre-trained encoders used during training and train on different data subsets for each LP. They finally use CometKiwi-da and continue to train a model based on CometKiwi-da. They finally combine the results over five checkpoints using the union of the predicted spans, which out-performed token-level majority voting.

**KUNMT (T2; En-De, Zh-En):** KUNMT proposes the use of different models to decompose tasks and post-editing with a large language model. In the process of error determination, span extraction, and severity assessment for each error span, distinct models were employed sequentially. The error determination model determines if an error exists in the sentence, and then the span assessment model explores the parts of the sentence where the error exists. For the spans where the error exists, the severity evaluation model evaluates whether the severity of the error is minor or major. All models were built upon XLM-RoBERTa-large, with some incorporating prompt-based learning. Results were subsequently calibrated using a large language model and tailored prompt engineering for the specific task.

**Unbabel-IST (T1 & T2; all):** the submission for Task 1 (word-level and sentence-level) follows their work from last year (Rei et al., 2022b). The major difference is the inclusion of the data from this year (e.g. sentence-level DA's for En-Te, En-Hi, En-Gu, En-Ta) and scaling the size of the pretrained encoder from InfoXLM to XLM-R XL and XXL (XXL was only used for sentence-level). They ensemble multiple checkpoints for the sentence-level subtask, using a weighted averaging of the predicted scores, optimised by LP.

For Task 2 they experimented with word-level models from Task 1 with GPT-4 prompts and with XCOMET (Guerreiro et al., 2023b). Their primary submission uses XCOMET which stands for eXplainable COMET. This model is trained with references to perform regression and error span identification. During inference the model can be used without refer-

ences, yet, for this task they found that using pseudo-references yields better performance if used with a simple heuristic where they first use a sentence-level QE system trained for Task 1 to evaluate the pseudo-reference. If the pseudo-reference is of high-quality, they give more weight to it otherwise, they give more weight to the source.

**IOL Research (T1-SL; all):** The IOL team experimented with several pretrained language models with extra modules to predict sentence level score and word tags including mBERT, XLM-RoBERTa-large, mDeberta, RemBert and InfoXLM. They first finetuned these models on DA and MQM scores data of QE and Metrics tasks in the previous years. Then, source text and its translation are fed into finetuned models added with extra modules for both sentence and word-level tasks. For sentence level, they separate embeddings of source text and translation of each layer in transformer models, and make a weighted sum among different layers for source and translation. Then the weighted embeddings of source and translation are concatenated and fed into a two-layer deep neural network to get score prediction with mean squared error (MSE) loss.

**(T1-WL; all):** For the word-level subtask, they use BiLSTM layer or one-layer DNN to do tag prediction on each token of translation with cross-entropy loss. The best checkpoint of each model is chosen by determining which checkpoint is best with respect to either Spearman correlation coefficient or MCC score, after training for three (3) epochs. The model of each language pair is tuned individually. The final result for each language pair is predicted by a weighted ensemble of different model checkpoints with LP-specific weights computed through weight searching using Optuna.

**MMT (T1-SL; En-De, En-Mr, En-Hi, En-Ta, En-Te, En-Gu):** For the studied language pairs, the MMT team enriched the training dataset through the application of eleven distinct data augmentation techniques, such as synonym substitution and back-translation, individually on the source sentence of each training instance. The results were generated using the best-performing model chosen from those trained on the corresponding augmented training datasets (in the case of English-German, the chosen model was trained on the augmented dataset created by applying the top four effective data augmentation techniques to each source sentence). The training methodology adheres to the COMET framework, with the foundational pre-trained model being XLM-RoBERTa-large.

**SurreyAI (T1-SL; En-Mr, En-Hi, En-Ta, En-Te, En-Gu):** The team proposes ensembleTQ as the main model, for which they train multiple multilingual QE models by fine-tuning pretrained language models (PTLMs) with autoencoder architecture. To that end they use the MonoTransQuest (Ranasinghe et al., 2020) architecture and report mean z-scores. The PTLMs that they combine in the ensemble are InfoXLM-large, XLMV-base and XLM-RoBERTa-large.

## 6 Results

In this section, we present and discuss the results of our shared task. Please note that for all the three sub-tasks we used statistical significance testing with $p = 0.05$.

### 6.1 Task 1

As we have seen in Task 1 description sentence-level submissions are evaluated against the true z-normalised sentence scores using Spearman's rank correlation coefficient $\rho$ along with the following secondary metrics: Pearson's correlation coefficient, $r$, and Kendall's $\tau$. Nonetheless, the final ranking between systems is calculated using the primary metric only (Spearman's $\rho$). Statistical significance was computed using William's test.

For the word-level task, the submissions are ranked using the Matthews correlation coefficient (MCC). F1-scores are provided as complementary information only and statistical significance was computed using randomisation tests (Yeh, 2000) with Bonferroni correction (Abdi, 2007) for each language pair.

The results for Task 1 are described in Tables 5 and 6. Looking at the obtained scores, we can observe an overall improved performance for the sentence-level scores, compared to previous years. While it is hard to make direct comparisons since

| ID | Affiliations | |
|---|---|---|
| BeringLab | BeringLab | – |
| HW-TSC | Huawei Translation Services Center, China | (Li et al., 2023) |
| IOL Research | Transn IOL Research, China | (Yan, 2023) |
| KUNMT | Korea University, South Korea | – |
| MMT | University of Manchester, UK & ASUS Intelligent Could Services, Singapore & University of Melbourne, Australia | (Wu et al., 2023) |
| NJUNLP | Nanjing University & Huawei Translation Services Center, China | (Geng et al., 2023a) |
| Surrey AI | University of Surrey & Aston University, UK | (Sindhujan et al., 2023) |
| Unbabel-IST | Unbabel & INESC-ID & Instituto de Telecomunicações & Instituto Superior Técnico, Portugal | (Rei et al., 2023) |

Table 4: Participants to the WMT23 Quality Estimation shared task.

the test-sets are new (and many language pairs are introduced for the first time), we can see that top performers obtain higher scores for En-Mr and Zh-En compared to the previous edition, and only for En-De we observe a relative drop (potentially justifiable by the introduction of more domains in the test set this year). Interestingly, for the word-level scores we observe higher correlations for the zero-shot tasks, as opposed to the ones where more training and development resources were made available.

We observe that especially for the sentence-level task all participants this year submitted ensembled predictions, with the ensemble size ranging from 3 to 12 models. We note that several teams combined models using different pre-trained encoders (HW-TSC, IOL Research, Surrey) and some participants focused their efforts on optimising the ensembled scores. Most notably, Bering Lab use a multi-step training where they select the best models on the first step, retrain with their proposed Adversarial Weight Perturbation method and then ensemble the top-3 models for the final submission. IOL research and Unbabel-IST also optimise the ensemble weights using optuna search.

We finally observe that following the trend of previous editions several participants experiment with training data augmentation techniques. Notably, most approaches this year focus on data augmentation that relates to the word-level or fine-grained annotations, either by computing TER-based word tags (Bering Lab), or by corrupting the target translations to generate pseudo-data with artificially generated error spans (NJUNLP, HW-TSC). For the latter, NJUNLP replace tokens and make use of the token distribution to approximate major versus minor errors (i.e, lower versus higher generation probabilities) and generate MQM-style annotations. Instead, HW-TSC propose to randomly corrupt the target (where corruption corresponds to insertion, deletion or replacement of a token) and use a heuristic score of the corrupted target to approximate a DA annotation style.

**Best performers** A total of seven teams participated in the **sentence-level sub-task**, yet only Unbabel-IST and IOL Research participated for all language pairs (including the zero-shot language pair, *He-En*), with Unbabel winning in the multilingual setting. However, for the individual language pairs, we observe different teams ranking at the top for different language pairs. Specifically, HW-TSC ranks at the top for all Indic language pairs, sharing the win with Unbabel-IST for En-Mr, En-Hi and En-Gu. On the MQM annotations, Unbabel-IST won the Zh-En and He-En language pairs, while IOL-Research and NJUNLP ranked top for En-De.

A total of four teams participated in the **word-level sub-task**, and similarly to the sentence-level only Unbabel-IST and IOL Research participated for all language pairs (including both zero-shot language pairs: *He-En* and *En-Fa*). NJUNLP won the task for the En-De language pair while Unbabel-IST ranked at the top for Zh-En, He-En, En-Mr and the multilingual task. IOL Research tied at the top for En-Fa.

We observe that while submissions consist of a mix of monolingual and multilingual submissions, and participants adopted a set of different strategies to design their architectures and tuning process, the top-ranking participants do share some common methodological choices. Specifically, all aforementioned participants tune their models in a multitasking setup, taking advantage of not only the sentence-level scores but also the word-level

| Model | Multi | Multidimensional Quality Metric (MQM) | | | Direct Assessment (DA) | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | En-De | En-Zh | He-En | En-Mr | En-Hi | En-Ta | En-Te | En-Gu |
| Unbabel-IST | **0.594** | 0.456 | **0.493** | **0.668** | **0.704** | 0.598 | 0.739 | **0.388** | **0.714** |
| IOL Research | 0.556 | **0.483** | **0.482** | 0.575 | 0.505 | 0.600 | 0.740 | 0.376 | 0.695 |
| BASELINE | 0.372 | 0.340 | 0.447 | 0.475 | 0.392 | 0.281 | 0.507 | 0.193 | 0.337 |
| HW-TSC | – | 0.437 | 0.460 | – | **0.692** | **0.644** | **0.775** | **0.394** | 0.691 |
| MMT | – | 0.316 | – | – | 0.650 | 0.494 | 0.547 | 0.337 | 0.540 |
| SurreyAI | – | – | – | – | 0.596 | 0.551 | 0.674 | 0.349 | 0.649 |
| BeringLab | – | 0.380 | 0.384 | – | – | – | – | – | – |
| NJUNLP | – | **0.479** | – | – | – | – | – | – | – |

Table 5: Spearman correlation for the official submissions to WMT23 Quality Estimation **Task 1 Sentence-level**. For each language pair, results marked in bold correspond to the winning submissions, as they are not significantly outperformed by any other system according to the Williams Significance Test (Williams, 1959). Baseline systems are highlighted in grey

| Model | Multi | Multidimensional Quality Metric (MQM) | | | Post-Editing (PE) | |
|---|---|---|---|---|---|---|
| | | En-De | Zh-En | He-En | En-Mr | En-Fa |
| Unbabel-IST | **0.329** | 0.246 | **0.302** | **0.402** | **0.347** | 0.345 |
| IOL Research | 0.298 | 0.256 | 0.250 | 0.359 | 0.334 | **0.351** |
| BASELINE | 0.252 | 0.179 | 0.225 | 0.275 | 0.287 | 0.293 |
| BeringLab | – | 0.233 | 0.241 | – | – | – |
| NJUNLP | – | **0.297** | – | – | – | – |

Table 6: Matthews Correlation Coefficient (MCC) for the official submissions to WMT23 Quality Estimation **Task 1 Word-level**. For each language pair, results marked in bold correspond to the winning submissions, as they are not significantly outperformed by any other system according to the Williams Significance Test (Williams, 1959). Baseline systems are highlighted in grey

| Model | Multi | Multidimensional Quality Metric (MQM) | | |
|---|---|---|---|---|
| | | En-De | En-Zh | He-En |
| Unbabel-IST | **0.220** | 0.273 | **0.288** | **0.279** |
| HW-TSC | 0.165 | 0.166 | 0.235 | 0.266 |
| BASELINE | 0.156 | 0.167 | 0.219 | 0.227 |
| KUNMT | – | 0.214 | 0.210 | – |
| NJUNLP | – | **0.284** | – | – |

Table 7: F1-score for the official submissions to WMT23 Quality Estimation **Task 2 Error Span Detection**. For each language pair, results marked in bold correspond to the winning submissions, as they are not significantly outperformed by any other system according to the Williams Significance Test (Williams, 1959). Baseline systems are highlighted in grey

quality tags. This inferred alignment between fine-grained quality annotations and overall quality at the segment level seems to be a promising direction for further improvements in quality estimation.

### 6.2 Task 2

For Task 2, the submissions are ranked using the F1-score, computed at character level for the annotated error spans, as described in Section 3.2. Precision and Recall scores are also provided as complementary information to help contextualise the performance observed. Statistical significance was computed using randomisation tests (Yeh, 2000) with Bonferroni correction (Abdi, 2007) for each language pair. The results for Task 2 are described in Table 7.

For this subtask we also had participants using pretrained large language models to enhance their submissions. Both KUNMT and Unbabel-IST (for the complementary submission of the latter) used GPT-4 with prompts tailored to fine-grained error span detection. KUNMT use an approach where they combine two prompts in a chain-of-thought manner, asking the model to act as an expert that either:

- Acts as an expert annotator that evaluates the translation and annotates error spans and severities (following the task instructions); or
- Acts as annotation validator and edits previous annotations or marks them as good.

We provide the full prompts in the Appendix E. In turn, Unbabel considers GPT4 both for the word level part of Task 1 and for Task 2, using prompts inspired by (Fernandes et al., 2023).

Aside from the use of LLMs, there are two main approaches in participating submissions: *i*) Participants who extended the word-level approach to obtain fine-grained error spans (HW-TSC, NJUNLP);

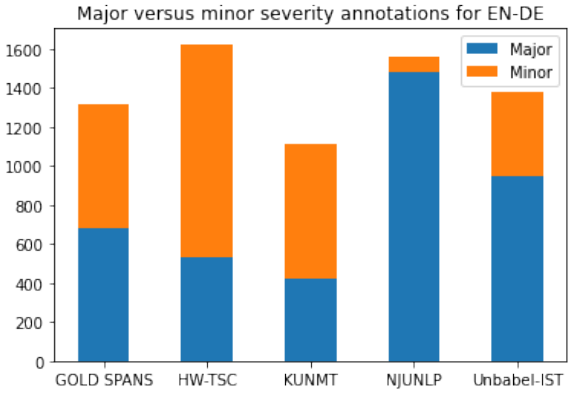


Figure 3: Balance between major and minor errors for each team and the original annotation (first bar).

and *ii*) participants who designed a methodology directly targeting error-spans (KUNMT, Unbabel-IST). To expand the word-level approach, NJUNLP maintain the binary OK/BAD labels and tune different thresholds to decide the severity of the BAD tags. They resolve severity inconsistencies over spans by taking the worst severity. Instead, HW-TSC convert the binary task to a 3-way classification ([OK, MINOR, MAJOR]) and tune on the Task 2 data. On the other hand, in their main approach, Unbabel-IST depart from the word-level approach and tune XCOMET directly on error span annotations, in a multitask setting (predicting both overall sentence score, and severities over spans).

Analysing the results, we can observe that with the exception of KUNMT, most submissions obtain higher *Recall* scores compared to *Precision*. Moreover, if we look at the distribution of identified error severities for e.g. En-De (see Figure 3) we can also observe a difference in the severity proportions as NJUNLP and Unbabel-IST identify mostly "major" errors and obtain better performance compared to KUNMT and HW-TSC that predict less skewed severities, with proportions closer to the gold data.

**Best performers.** Four teams participated in Task 2, IST-Unbabel, HW-TSC, KUNMT and NJUNLP, with only Unbabel-IST and HW-TSC participating in all tasks. NJUNLP ranked first for the En-De language pair while Unbabel-IST ranked first for the rest, including the multilingual track. We note that both top-ranking participants are using ensemble approaches as well as enhancing their approaches with pseudo-data (pseudo-references in the case of Unbabel-IST and pseudo-MQM scores in the case of NJUNLP).

## 7 Discussion

In what follows, we discuss the main findings of this year's shared task based on the goals we had previously identified for it.

**Granularity of quality annotations** We note that while performance for the sentence-level quality scores is better, the finer-grained annotations are also contributing towards these improvements. Indeed, while in comparison participants achieve lower performance for the fine-grained and word-level tasks, **most of the top-ranking submissions across tasks constitute multi-task approaches**, that combined information from the finer-grained annotations and the sentence level scores.

Moreover, comparing the F1-scores between the word-level Task 1 and Task 2, while F1-scores for Task 2 are somewhat lower, considering the additional complexity of considering error severities and multiple spans, the performance seems promising. To further encourage future participation and improved performance we aim to focus on extending MQM annotations for the next iterations, but also revising the error severity definition to potentially include '*critical*' errors.

**Zero shot predictions** We observe that performance for the zero-shot language pairs, He-En and En-Fa was not hampered by the lack of training and development resources. While fewer participants submitted predictions for these languages, their performance was on par with other language pairs. For for the word-level task, scores were actually higher than those observed for other language pairs. Looking closer at the approach adopted by the participants in these tasks, we can see that besides relying on multi-lingual encoders no additional data was used to train for these languages, across tasks. These findings are encouraging towards annotating a wider range of language pairs (maintaining the emphasis on low and medium source languages) to test on for the upcoming editions, even when training resources are scarce.

**Hallucinations** We report the hallucination detection results in Table 9. Overall, the results indicate that good-quality QE models are capable of detecting hallucinations very satisfactorily. Some submissions, in fact, obtain perfect or near-perfect results for some language pairs, which indicates that they are able to appropriately penalise the severity of hallucination errors. Not only that, but they can

| En-De | | Zh-En | | He-En | | En-Mr | | En-Hi | | En-Ta | | En-Te | | En-Gu | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| No. Halls | Total | No. Halls | Total | No. Halls | Total | No. Halls | Total | No. Halls | Total | No. Halls | Total | No. Halls | Total | No. Halls | Total |
| 9 | 1422 | 13 | 1260 | 48 | 896 | 86 | 836 | 74 | 824 | 75 | 824 | 75 | 811 | 75 | 825 |

Table 8: Statistics (number of hallucinations and total number of samples) of the hallucination detection evaluation sets for each language pair.

| | En-De | | Zh-En | | He-En | | En-Mr | | En-Hi | | En-Ta | | En-Te | | En-Gu | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | AUC | R@$k$ | AUC | R@$k$ | AUC | R@$k$ | AUC | R@$k$ | AUC | R@$k$ | AUC | R@$k$ | AUC | R@$k$ | AUC | R@$k$ |
| NJUNLP | 96.25 | 44.44 | – | – | 41.62 | 25.00 | – | – | – | – | – | – | – | – | – | – |
| Unbabel-IST | 99.99 | 88.89 | 98.97 | 84.62 | 97.43 | 83.33 | 99.65 | 91.86 | 99.92 | 97.30 | 99.82 | 92.00 | 99.83 | 93.33 | 99.82 | 92.00 |
| IOL Research | 100.00 | 100.00 | 99.60 | 84.62 | 97.04 | 87.50 | 98.59 | 87.21 | 99.83 | 97.30 | 99.48 | 89.33 | 99.51 | 88.00 | 99.95 | 98.67 |
| BeringLab | 80.11 | 0.00 | 90.40 | 7.69 | – | – | – | – | – | – | – | – | – | – | – | – |
| MMT | 90.71 | 0.00 | 96.97 | 30.77 | – | – | 92.53 | 62.79 | 95.85 | 78.38 | 91.29 | 65.33 | 96.57 | 68.00 | 52.58 | 8.00 |
| HW-TSC | 99.97 | 88.89 | 98.77 | 76.92 | – | – | 99.51 | 89.53 | 99.42 | 95.95 | 99.49 | 93.33 | 99.99 | 97.33 | 99.94 | 98.67 |
| Surrey AI | – | – | – | – | – | – | 85.79 | 48.84 | 98.86 | 85.14 | 98.24 | 82.67 | 98.82 | 82.67 | 99.28 | 88.00 |

Table 9: Hallucination detection performance for the official submissions to **Task 1 Sentence-level**. We report AUROC (↑) and R@$k$, in percentage (↑).

**Source:**
大学于1896至1991年，及1912至1968年间分别接管了王家音乐学院和王家安大略博物馆；后两者如今作为独立的机构仍与大学保留着密切的联系。

**Hallucination in zh-en not detected by the Unbabel-IST and IOL Research submissions:**
The University took over the **Oscillation:Royal Academy of Music and the Royal Academy of Music** between 1896 and 1991, and between 1912 and 1968, and the Royal Academy of Music and the Royal Academy of **Oscillation:Art and Art** Museum, which are now in close connection with the University as an independent institution.

**Source:**
Intelligent nursing care bed / smart bed Nursing care beds with technical equipment including sensors and notification functions are known as “intelligent” or “smart” beds.

**Hallucination in en-de not detected by the Unbabel-IST submission:**
Intelligente Pflege-Bett / Smart-**Oscillation: Bett-Bett-Bett-Bett-Bett-Bett-Bett** mit technischer Ausstattung einschließlich Sensoren und Notifizierungsfunktionen sind als "intelligente" oder "intelligente" Bett-Bett bekannt.

Table 10: Examples of hallucinations not flagged in the bottom $k$ of the respective language pair by the best systems in Task 1 Sentence-level.

also distinguish the hallucinations from other translations whose quality may not necessarily be high. This is a property that was not observed in previous iterations of QE models (Amrhein and Sennrich, 2022; Raunak et al., 2022), in particular those that were based on dual encoding of the source and the translation (e.g., COMET-QE (Rei et al., 2020)).

Nevertheless, even the top-performing QE systems, including the winning submissions, may struggle with localised critical errors such as oscillations. We show two such examples in Table 10. In fact, although most pathological hallucinations are detected, some egregious examples have not been detected by both the IOL Research and Unbabel-IST systems (e.g., a he-en translation that contains 70 hallucinated <unk> tokens, and a zh-en translation that contains the oscillation "Tropical and Sub-Tropical Plains and Plains, Tropical and Sub-Tropical Plains and Plains, Tropical and Sub-Tropical High Plains, Sub-Tropical Plains and Sub-Tropical Plains, Sub-Tropical Plains and Sub-Tropical Plains").

One hypothesis for this undesirable behavior is that such samples are out-of-distribution for the QE systems. As such, augmenting the training sets with examples of such hallucinations (e.g., as done in XCOMET (Guerreiro et al., 2023b)) may be a straightforward yet effective approach for correcting this behavior.

## 8 Conclusions

This year's edition of the QE Shared Task introduced a number of new elements: new low-resource language pairs (including two zero-shot ones), new test sets, and new fine-grained error detection task that we aspire to continue in future editions. It also introduced a mix of hallucinated data together with the original translations, allowing us to assess the robustness of submissions and detect failure patterns that will hopefully help develop more robust QE systems in the future.

The tasks attracted a steady number of participating teams and we believe the overall results are a great reflection of the evolution of the QE field. We note that the gold labels and best submissions to all tasks are made available for those interested in further analysing the results. We aspire for the future editions to continue the efforts set in this and previous years and expand the resources and coverage of QE, while further exploring recent and more challenging subtasks such as fine-grained QE and explainable QE.

## 9 Ethical Considerations

MQM and DA annotations in this paper are done by professional translators. They are all paid at professional rates.

Organisers from Unbabel and University of Surrey have submitted to this task without using prior access to test sets nor using any insider information.

## Acknowledgements

Part of this work was supported by EU's Horizon Europe Research and Innovation Actions (UTTER, contract 101070631), by the project DECOLLAGE (ERC-2022-CoG 101088763), by Fundação para a Ciência e Tecnologia through contract UIDB/50008/2020, and by the Portuguese Recovery and Resilience Plan through project C645008882- 00000055 (Center for Responsible AI).

We thank the annotation agencies Zibanka Media Services Pvt. Ltd. and Techliebe for working with us towards annotating DA data for Indic language pairs. We also thank the European Association for Machine Translation (EAMT) for sponsoring our Indic language pair annotation project at the University of Surrey.

## References

Hervé Abdi. 2007. The bonferroni and šidák corrections for multiple comparisons. *Encyclopedia of measurement and statistics*, 3:103–107.

Chantal Amrhein and Rico Sennrich. 2022. Identifying weaknesses in machine translation metrics through minimum Bayes risk decoding: A case study for COMET. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1125–1141, Online only. Association for Computational Linguistics.

Fatemeh Azadi, Heshaam Faili, and Mohammad Javad Dousti. 2022. Mismatching-Aware Unsupervised Translation Quality Estimation for Low-Resource Languages. *arXiv preprint arXiv:2208.00463*.

Aljoscha Burchardt. 2013. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Qu Cui, Shujian Huang, Jiahuan Li, Xiang Geng, Zaixiang Zheng, Guoping Huang, and Jiajun Chen. 2021. Directqe: Direct pretraining for machine translation quality estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12719–12727.

David Dale, Elena Voita, Loic Barrault, and Marta R. Costa-jussà. 2023a. Detecting and mitigating hallucinations in machine translation: Model internal workings alone do well, sentence similarity Even better. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36–50, Toronto, Canada. Association for Computational Linguistics.

David Dale, Elena Voita, Janice Lam, Prangthip Hansanti, Christophe Ropers, Elahe Kalbassi, Cynthia Gao, Loïc Barrault, and Marta R. Costa-jussà. 2023b. Halomi: A manually annotated benchmark for multilingual hallucination and omission detection in machine translation.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André FT Martins, Graham Neubig, Ankush Garg, Jonathan H Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. *arXiv preprint arXiv:2308.07286*.

Javier Ferrando, Gerard I. Gállego, and Marta R. Costa-jussà. 2022. Measuring the mixing of contextual information in the transformer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8698–8714, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Marina Fomicheva, Shuo Sun, Erick Fonseca, Chrysoula Zerva, Frédéric Blain, Vishrav Chaudhary, Francisco Guzmán, Nina Lopatina, Lucia Specia, and André F. T. Martins. 2022. MLQE-PE: A multilingual quality estimation and post-editing dataset. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4963–4974, Marseille, France. European Language Resources Association.

Erick Fonseca, Lisa Yankovskaya, André FT Martins, Mark Fishel, and Christian Federmann. 2019. Findings of the wmt 2019 shared tasks on quality estimation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 1–10.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021a. 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, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021b. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

Xiang Geng, Zhejian Lai, Yu Zhang, Shimin Tao, Hao Yang, Jiajun Chen, and Shujian Huang. 2023a. Njunlp's participation for the wmt2023 quality estimation shared task. In *Proceedings of the Eigth Conference on Machine Translation*, Singapore. Association for Computational Linguistics.

Xiang Geng, Yu Zhang, Jiahuan Li, Shujian Huang, Hao Yang, Shimin Tao, Yimeng Chen, Ning Xie, and Jiajun Chen. 2023b. Denoising pre-training for machine translation quality estimation with curriculum learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12827–12835.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Nuno M. Guerreiro, Duarte Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023a. Hallucinations in large multilingual translation models.

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

Nuno M. Guerreiro, Elena Voita, and André Martins. 2023c. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1059–1075, Dubrovnik, Croatia. Association for Computational Linguistics.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2018. The IIT Bombay English-Hindi parallel corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Yuang Li, Chang Su, Ming Zhu, Mengyao Piao, Xinglin Lyu, Min Zhang, and Hao Yang. 2023. Hw-tsc 2023 submission for the quality estimation shared task. In *Proceedings of the Eigth Conference on Machine Translation*, Singapore. Association for Computational Linguistics.

Arle Lommel, Aljoscha Burchardt, Maja Popović, Kim Harris, Eleftherios Avramidis, and Hans Uszkoreit. 2014. Using a new analytic measure for the annotation and analysis of mt errors on real data. In *Proceedings of the 17th Annual conference of the European Association for Machine Translation*, pages 165–172.

Brian W Matthews. 1975. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Tharindu Ranasinghe, Constantin Orašan, and Ruslan Mitkov. 2020. Transquest: Translation quality estimation with cross-lingual transformers. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5070–5081.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183, Online. Association for Computational Linguistics.

Vikas Raunak, Matt Post, and Arul Menezes. 2022. SALTED: A framework for SAlient long-tail translation error detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5163–5179, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ricardo Rei, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. Martins, and Alon Lavie. 2021. Are references really needed? unbabel-IST 2021 submission for the metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040, Online. Association for Computational Linguistics.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. 2023. Scaling up cometkiwi: Unbabel-ist 2023 submission for the quality estimation shared task. In *Proceedings of the Eigth Conference on Machine Translation*, Singapore. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Unbabel's participation in the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 911–920, Online. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José GC de Souza, Taisiya Glushkova, Duarte M Alves, Alon Lavie, et al. 2022a. Cometkiwi: Ist-unbabel 2022 submission for the quality estimation shared task. *WMT 2022*, page 634.

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

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia.

Archchana Sindhujan, Diptesh Kanojia, Constantin Orăsan, and Tharindu Ranasinghe. 2023. Surreyai 2023 submission for the quality estimation shared task. In *Proceedings of the Eigth Conference on Machine Translation*, Singapore. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard 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.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Chrysoula Zerva, Zhenhao Li, Vishrav Chaudhary, and André F. T. Martins. 2021. Findings of the WMT 2021 shared task on quality estimation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 684–725, Online. Association for Computational Linguistics.

Shimin Tao, Su Chang, Ma Miaomiao, Hao Yang, Xiang Geng, Shujian Huang, Min Zhang, Jiaxin Guo,

Minghan Wang, and Yinglu Li. 2022. Crossqe: Hwtsc 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 646–652.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Dusan Varis and Ondřej Bojar. 2021. Sequence length is a domain: Length-based overfitting in transformer models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8246–8257, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Evan J. Williams. 1959. *Regression Analysis*, volume 14. Wiley, New York, USA.

Yulong Wu, Viktor Schlegel, Daniel Beck, and Riza Batista-Navarro. 2023. Mmt's submission for the wmt 2023 quality estimation shared task. In *Proceedings of the Eigth Conference on Machine Translation*, Singapore. Association for Computational Linguistics.

Zeyu Yan. 2023. Iol research's submission for wmt 2023 quality estimation shared task. In *Proceedings of the Eigth Conference on Machine Translation*, Singapore. Association for Computational Linguistics.

Alexander Yeh. 2000. More Accurate Tests for the Statistical Significance of Result Differences. In *Coling-2000: the 18th Conference on Computational Linguistics*, pages 947–953, Saarbrücken, Germany.

Chrysoula Zerva, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José G. C. de Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orăsan, Marina Fomicheva, André F. T. Martins, and Lucia Specia. 2022. Findings of the WMT 2022 shared task on quality estimation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 69–99, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

## A Hyper-parameters of pre-trained baseline models for Task 1 Quality Estimation

| Hyper-parameter | **T1 Sentence-level** COMET-QE | **T1 Word-level** COMETKIWI |
|---|---|---|
| Encoder Model | XLM-RoBERTa (large) | XLM-RoBERTa (large) |
| Optimizer | Adam (default parameters) | Adam (default parameters) |
| n frozen epochs | 0.3 | 0.3 |
| Keep embeddings frozen | True | True |
| Learning rate | 3e-05 and 1e-05 | 1e-06 and 1e-05 |
| Batch size | 4 | 4 |
| Loss function | MSE | MSE and CE ($\epsilon = 1.0$) |
| Dropout | 0.15 | 0.1 |
| FP precision | 32 | 32 |
| Feed-Forward hidden units | [2048, 1024] | [2048, 1024] |
| Word weights | – | [0.3, 0.7] |
| Feed-Forward activation | Tanh | – |

Table 11: Hyper-parameters of both the COMET-QE and the CometKiwi models used as baselines for Task 1 Quality Estimation.

## B Official Results of the WMT23 Quality Estimation Task 1 Sentence-level

Tables 12, 13, 14, 15, 16, 17, 18, 19 and 20 show the results for all language pairs and the multilingual variants, ranking participating systems best to worst using Spearman correlation as primary key for each of these cases.

| Model | **Spearman** | Pearson | Kendall |
|---|---|---|---|
| Unbabel-IST • | 0.594 | 0.580 | 0.438 |
| IOL Research | 0.556 | 0.513 | 0.407 |
| BASELINE | 0.372 | 0.308 | 0.265 |

Table 12: Official results of the WMT23 Quality Estimation Task 1 Sentence-level **Multilingual** (average over all language pairs). Teams marked with "•" are the winners, as they are not significantly outperformed by any other system according to the Williams Significance Test (Williams, 1959). Baseline systems are highlighted in grey.

| Model | **Spearman** | Pearson | Kendall | Disk footprint (B) | # Model params | Ensemble |
|---|---|---|---|---|---|---|
| IOL Research • | 0.483 | 0.429 | 0.364 | 2,357,242,105 | 589,270,071 | 5 |
| NJUNLP • | 0.479 | 0.423 | 0.360 | 3,264,730,349 | 560,145,557 | 12 |
| Unbabel-IST | 0.456 | 0.457 | 0.346 | 42,868,104,221 | 10,716,932,147 | 6 |
| HW-TSC | 0.437 | 0.433 | 0.331 | 27,730,527,504 | 6,932,631,876 | 12 |
| BeringLab | 0.380 | 0.281 | 0.283 | 2,243,955,309 | 560,945,155 | 3 |
| BASELINE | 0.340 | 0.253 | 0.257 | 2,277,430,715 | 569,330,715 | 1 |
| MMT | 0.316 | 0.221 | 0.237 | 2,448,132,038 | 569,330,715 | 6 |

Table 13: Official results of the WMT23 Quality Estimation Task 1 Sentence-level for **Englih-German (MQM)**. Teams marked with "•" are the winners, as they are not significantly outperformed by any other system according to the Williams Significance Test (Williams, 1959). Baseline systems are highlighted in grey.

| Model | **Spearman** | Pearson | Kendall | Disk footprint (B) | # Model params | Ensemble |
|---|---|---|---|---|---|---|
| Unbabel-IST • | 0.493 | 0.423 | 0.378 | 42,868,104,221 | 10,716,932,147 | 4 |
| IOL Research • | 0.482 | 0.392 | 0.369 | 2,357,242,105 | 589,270,071 | 5 |
| HW-TSC | 0.460 | 0.369 | 0.352 | 27,730,527,504 | 6,932,631,876 | 12 |
| BASELINE | 0.447 | 0.318 | 0.342 | 2,277,430,715 | 569,330,715 | 1 |
| BeringLab | 0.384 | 0.230 | 0.288 | 2,243,955,309 | 560,945,155 | 3 |

Table 14: Official results of the WMT23 Quality Estimation Task 1 Sentence-level for **Chinese-English (MQM)**. Teams marked with "•" are the winners, as they are not significantly outperformed by any other system according to the Williams Significance Test (Williams, 1959). Baseline systems are highlighted in grey.

| Model | **Spearman** | Pearson | Kendall | Disk footprint (B) | # Model params | Ensemble |
|---|---|---|---|---|---|---|
| Unbabel-IST • | 0.668 | 0.518 | 0.499 | 42,868,104,221 | 10,716,932,147 | 4 |
| IOL Research | 0.575 | 0.424 | 0.416 | 2,357,242,105 | 589,270,071 | 5 |
| BASELINE | 0.475 | 0.363 | 0.337 | 2,277,430,715 | 569,330,715 | 1 |

Table 15: Official results of the WMT23 Quality Estimation Task 1 Sentence-level for **Hebrew-English (MQM)**. Teams marked with "•" are the winners, as they are not significantly outperformed by any other system according to the Williams Significance Test (Williams, 1959). Baseline systems are highlighted in grey.

| Model | **Spearman** | Pearson | Kendall | Disk footprint (B) | # Model params | Ensemble |
|---|---|---|---|---|---|---|
| Unbabel-IST • | 0.704 | 0.735 | 0.513 | 42,868,104,221 | 10,716,932,147 | 6 |
| HW-TSC • | 0.692 | 0.718 | 0.504 | 27,730,527,504 | 6,932,631,876 | 12 |
| MMT | 0.650 | 0.663 | 0.466 | 2,448,132,038 | 569,330,715 | 7 |
| SurreyAI | 0.596 | 0.668 | 0.423 | 2,362,232,012 | 633,305,686 | 3 |
| IOL Research | 0.505 | 0.372 | 0.353 | 2,357,242,105 | 589,270,071 | 5 |
| BASELINE | 0.392 | 0.427 | 0.274 | 2,277,430,715 | 569,330,715 | 1 |

Table 16: Official results of the WMT23 Quality Estimation Task 1 Sentence-level **English-Marathi (DA)**. Teams marked with "•" are the winners, as they are not significantly outperformed by any other system according to the Williams Significance Test (Williams, 1959). Baseline systems are highlighted in grey.

| Model | **Spearman** | Pearson | Kendall | Disk footprint (B) | # Model params | Ensemble |
|---|---|---|---|---|---|---|
| HW-TSC • | 0.644 | 0.720 | 0.477 | 27,730,527,504 | 6,932,631,876 | 12 |
| IOL Research | 0.600 | 0.667 | 0.433 | 2,357,242,105 | 589,270,071 | 5 |
| Unbabel-IST | 0.598 | 0.667 | 0.431 | 42,868,104,221 | 10,716,932,147 | 4 |
| SurreyAI | 0.551 | 0.668 | 0.395 | 2,362,232,012 | 633,305,686 | 3 |
| MMT | 0.494 | 0.570 | 0.345 | 2,448,132,038 | 569,330,715 | 7 |
| BASELINE | 0.281 | 0.245 | 0.190 | 2,277,430,715 | 569,330,715 | 1 |

Table 17: Official results of the WMT23 Quality Estimation Task 1 Sentence-level **English-Hindi (DA)**. Teams marked with "•" are the winners, as they are not significantly outperformed by any other system according to the Williams Significance Test (Williams, 1959). Baseline systems are highlighted in grey.

| Model | **Spearman** | Pearson | Kendall | Disk footprint (B) | # Model params | Ensemble |
|---|---|---|---|---|---|---|
| HW-TSC • | 0.775 | 0.778 | 0.597 | 27,730,527,504 | 6,932,631,876 | 12 |
| IOL Research | 0.740 | 0.742 | 0.557 | 2,357,242,105 | 589,270,071 | 5 |
| Unbabel-IST | 0.739 | 0.733 | 0.550 | 42,868,104,221 | 10,716,932,147 | 4 |
| SurreyAI | 0.674 | 0.710 | 0.495 | 2,362,232,012 | 633,305,686 | 3 |
| MMT | 0.547 | 0.531 | 0.384 | 2,448,132,038 | 569,330,715 | 7 |
| BASELINE | 0.507 | 0.402 | 0.354 | 2,277,430,715 | 569,330,715 | 1 |

Table 18: Official results of the WMT23 Quality Estimation Task 1 Sentence-level **English-Tamil (DA)**. Teams marked with "•" are the winners, as they are not significantly outperformed by any other system according to the Williams Significance Test (Williams, 1959). Baseline systems are highlighted in grey.

| Model | **Spearman** | Pearson | Kendall | Disk footprint (B) | # Model params | Ensemble |
|---|---|---|---|---|---|---|
| HW-TSC • | 0.394 | 0.350 | 0.269 | 27,730,527,504 | 6,932,631,876 | 12 |
| Unbabel-IST • | 0.388 | 0.362 | 0.264 | 42,868,104,221 | 10,716,932,147 | 4 |
| IOL Research | 0.376 | 0.344 | 0.257 | 2,357,242,105 | 589,270,071 | 5 |
| SurreyAI | 0.349 | 0.376 | 0.241 | 2,362,232,012 | 633,305,686 | 3 |
| MMT | 0.337 | 0.281 | 0.228 | 2,448,132,038 | 569,330,715 | 7 |
| BASELINE | 0.193 | 0.153 | 0.134 | 2,277,430,715 | 569,330,715 | 1 |

Table 19: Official results of the WMT23 Quality Estimation Task 1 Sentence-level **English-Telegu (DA)**. Teams marked with "•" are the winners, as they are not significantly outperformed by any other system according to the Williams Significance Test (Williams, 1959). Baseline systems are highlighted in grey.

| Model | **Spearman** | Pearson | Kendall | Disk footprint (B) | # Model params | Ensemble |
|---|---|---|---|---|---|---|
| Unbabel-IST • | 0.714 | 0.745 | 0.529 | 42,868,104,221 | 10,716,932,147 | 4 |
| IOL Research | 0.695 | 0.742 | 0.513 | 2,357,242,105 | 589,270,071 | 5 |
| HW-TSC | 0.691 | 0.714 | 0.511 | 27,730,527,504 | 6,932,631,876 | 12 |
| SurreyAI | 0.649 | 0.700 | 0.474 | 2,362,232,012 | 633,305,686 | 3 |
| MMT | 0.540 | 0.581 | 0.386 | 2,448,132,038 | 569,330,715 | 7 |
| BASELINE | 0.337 | 0.307 | 0.230 | 2,277,430,715 | 569,330,715 | 1 |

Table 20: Official results of the WMT23 Quality Estimation Task 1 Sentence-level **English-Gujarati (DA)**. Teams marked with "•" are the winners, as they are not significantly outperformed by any other system according to the Williams Significance Test (Williams, 1959). Baseline systems are highlighted in grey.

## C Official Results of the WMT23 Quality Estimation Task 1 Word-level

Tables 21, 22, 23, 24, 25 and 26 show the results for all language pairs and the multilingual variant, ranking participating systems best to worst using Matthews Correlation Coefficient (MCC) as primary key for each of these cases.

| Model | **MCC** | F1-score |
|---|---|---|
| Unbabel-IST • | 0.329 | 0.355 |
| IOL Research | 0.298 | 0.322 |
| BASELINE | 0.252 | 0.243 |

Table 21: Official results of the WMT23 Quality Estimation Task 1 Word-level **Multilingual** (average over all language pairs). The winning submission is indicated by a •. Baseline systems are highlighted in grey.

| Model | **MCC** | F1-score | Disk footprint (B) | # Model params | Ensemble |
|---|---|---|---|---|---|
| NJUNLP • | 0.297 | 0.329 | 3,264,730,349 | 560,145,557 | 12 |
| IOL Research | 0.256 | 0.281 | 2,357,242,105 | 589,270,071 | 5 |
| Unbabel-IST | 0.246 | 0.279 | 2,252,351,787 | 563,041,309 | 1 |
| BeringLab | 0.233 | 0.269 | 2,243,955,309 | 560,945,155 | 5 |
| BASELINE | 0.179 | 0.207 | 2,252,351,659 | 563,041,309 | 1 |

Table 22: Official results of the WMT23 Quality Estimation Task 1 Word-level **English-German (MQM)**. The winning submission is indicated by a •. Baseline systems are highlighted in grey.

| Model | **MCC** | F1-score | Disk footprint (B) | # Model params | Ensemble |
|---|---|---|---|---|---|
| Unbabel-IST • | 0.302 | 0.331 | 2,252,351,787 | 563,041,309 | 1 |
| IOL Research | 0.250 | 0.268 | 2,357,242,105 | 589,270,071 | 5 |
| BeringLab | 0.241 | 0.262 | 2,243,955,309 | 560,945,155 | 5 |
| BASELINE | 0.225 | 0.255 | 2,252,351,659 | 563,041,309 | 1 |

Table 23: Official results of the WMT23 Quality Estimation Task 1 Word-level **Chinese-English (MQM)**. The winning submission is indicated by a •. Baseline systems are highlighted in grey.

| Model | **MCC** | F1-score | Disk footprint (B) | # Model params | Ensemble |
|---|---|---|---|---|---|
| Unbabel-IST • | 0.402 | 0.439 | 2,252,351,787 | 563,041,309 | 1 |
| IOL Research | 0.359 | 0.361 | 2,357,242,105 | 589,270,071 | 5 |
| BASELINE | 0.275 | 0.275 | 2,252,351,659 | 563,041,309 | 1 |

Table 24: Official results of the WMT23 Quality Estimation Task 1 Word-level **Hebrew-English (MQM)**. The winning submission is indicated by a •. Baseline systems are highlighted in grey.

| Model | **MCC** | F1-score | Disk footprint (B) | # Model params | Ensemble |
|---|---|---|---|---|---|
| Unbabel-IST • | 0.347 | 0.359 | 2,252,351,787 | 563,041,309 | 1 |
| IOL Research | 0.334 | 0.373 | 2,357,242,105 | 589,270,071 | 5 |
| BASELINE | 0.287 | 0.224 | 2,252,351,659 | 563,041,309 | 1 |

Table 25: Official results of the WMT23 Quality Estimation Task 1 Word-level **English-Marathi (Post-Editing)**. The winning submission is indicated by a •. Baseline systems are highlighted in grey.

| Model | **MCC** | F1-score | Disk footprint (B) | # Model params | Ensemble |
|---|---|---|---|---|---|
| IOL Research • | 0.351 | 0.389 | 2,357,242,105 | 589,270,071 | 5 |
| Unbabel-IST | 0.345 | 0.365 | 2,252,351,787 | 563,041,309 | 1 |
| BASELINE | 0.293 | 0.254 | 2,252,351,659 | 563,041,309 | 1 |

Table 26: Official results of the WMT23 Quality Estimation Task 1 Word-level **English-Farsi (Post-Editing)**. The winning submission is indicated by a •. Baseline systems are highlighted in grey.

## D Official Results of the WMT23 Quality Estimation Task 2 Error Span Detection

Tables 27, 28, 29 and 30 show the results for all language pairs and the multilingual variant, ranking participating systems best to worst using F1-score as primary key for each of these cases.

| Model | **F1-score** | Precision | Recall |
|---|---|---|---|
| Unbabel-IST • | 0.220 | 0.164 | 0.360 |
| HW-TSC | 0.165 | 0.177 | 0.161 |
| BASELINE | 0.156 | 0.203 | 0.128 |

Table 27: Official results of the WMT23 Quality Estimation Task 1 Word-level **Multilingual** (average over all language pairs). The winning submission is indicated by a •. Baseline systems are highlighted in grey.

| Model | **F1-score** | Precision | Recall | Disk footprint (B) | # Model params | Ensemble |
|---|---|---|---|---|---|---|
| NJUNLP • | 0.284 | 0.238 | 0.352 | 560,145,557 | 3,264,730,349 | 12 |
| Unbabel-IST | 0.273 | 0.209 | 0.394 | -1 | -1 | -1 |
| KUNMT | 0.214 | 0.224 | 0.206 | 818,245,780 | 2,235,540,305 | 3 |
| BASELINE | 0.167 | 0.229 | 0.131 | 563,041,309 | 2,252,351,659 | 1 |
| HW-TSC | 0.166 | 0.220 | 0.133 | 285,019,112 | 1,148,646,407 | 5 |

Table 28: Official results of the WMT23 Quality Estimation Task 1 Word-level **English-German (MQM)**. The winning submission is indicated by a •. Baseline systems are highlighted in grey.

| Model | **F1-score** | Precision | Recall | Disk footprint (B) | # Model params | Ensemble |
|---|---|---|---|---|---|---|
| Unbabel-IST • | 0.288 | 0.246 | 0.349 | 3,485,770,281 | 13,943,358,015 | 5 |
| HW-TSC | 0.235 | 0.221 | 0.250 | 285,019,112 | 1,148,646,407 | 4 |
| BASELINE | 0.219 | 0.259 | 0.190 | 563,041,309 | 2,252,351,659 | 1 |
| KUNMT | 0.210 | 0.216 | 0.204 | 818,245,780 | 2,235,540,305 | 3 |

Table 29: Official results of the WMT23 Quality Estimation Task 1 Word-level **Chinese-English (MQM)**. The winning submission is indicated by a •. Baseline systems are highlighted in grey.

| Model | **F1-score** | Precision | Recall | Disk footprint (B) | # Model params | Ensemble |
|---|---|---|---|---|---|---|
| Unbabel-IST • | 0.279 | 0.241 | 0.332 | 3,485,770,281 | 13,943,358,015 | 5 |
| HW-TSC | 0.266 | 0.254 | 0.279 | 285,019,112 | 1,148,646,407 | 10 |
| BASELINE | 0.227 | 0.474 | 0.150 | 563,041,309 | 2,252,351,659 | 1 |

Table 30: Official results of the WMT23 Quality Estimation Task 1 Word-level **Hebrew-English (MQM)**. The winning submission is indicated by a •. Baseline systems are highlighted in grey.

## E GPT-4 prompts for Task 2

We add below the prompts used by KUNMT team with GPT4 for Task 2.

**Expert annotator prompt:**

You are an expert in the Fine-grained error span detection task. The goal of this task is to predict the word-level translation error spans. you will be asked to predict both the error span (start and end indices) as well as the error severity (major or minor) for each segment. There can be multiple error spans, and you must indicate the severity of the error for the spans that exist. If no errors exist in the translation, the error span is (-1,-1) and the error severity is no-error.

**Expert validator prompt:**

Review this result by checking the work done by the other workers. If the work was done correctly, mark it as "GOOD"; if there were any errors, re-annotate the Error Span and Error Severity. To avoid

inconsistencies, we expect the indices of the errors spans to correspond to characters in the target string before tokenisation, i.e., the target string that will be provided as test data.'