Findings of the Quality Estimation Shared Task at WMT 2024 Are LLMs Closing the Gap in QE?

Chrysoula Zerva[1,2]*, Frédéric Blain[3]*, José G. C. de Souza[4], Diptesh Kanojia[5], Sourabh Deoghare[6], Nuno M. Guerreiro[1,2,4,10], Giuseppe Attanasio[1], Ricardo Rei[2,4,7], Constantin Orăsan[5], Matteo Negri[8], Marco Turchi[11], Rajen Chatterjee[9], Pushpak Bhattacharyya[6], Markus Freitag[12], André F. T. Martins[1,2,4]

[1]Instituto de Telecomunicações, [2]Instituto Superior Técnico, Universidade de Lisboa, [3]Tilburg University, [4]Unbabel, [5]University of Surrey, [6]IIT Bombay, [7]INESC-ID, [8]FBK, [9]Apple Inc. [10]MICS, CentraleSupélec, Université Paris-Saclay, [11]Zoom Video Communications, [12]Google Inc.

wmt-qe-organizers@googlegroups.com

Abstract

We report the results of the WMT 2024 shared task on Quality Estimation, in which the challenge is to predict the quality of the output of neural machine translation systems without access to reference translations. In this edition, we continue to focus both on predicting sentence-level scores and on detecting error spans. Further, we expanded our scope to assess the potential for quality estimation to help in the correction of translated outputs, hence including an automated post-editing (APE) task.

We publish new test sets with human annotations that target two directions: providing new Multidimensional Quality Metrics (MQM) annotations for three multi-domain language pairs (English to German, Spanish and Hindi) and extending the annotations on Indic languages, providing direct assessments and post edits for translation from English into Hindi, Gujarati, Tamil and Telugu. We also perform a detailed analysis of the behaviour of different models with respect to different phenomena, including gender bias, idiomatic language, and numerical and entity perturbations. We received submissions based on both traditional encoder-based approaches and large language models (LLMs) and attempted to draw some comparisons in terms of performance and robustness to different phenomena.

1 Introduction

This edition of the shared task on Quality Estimation (QE) for machine translation builds upon previous iterations and findings, to further benchmark methods for estimating the quality of neural Machine Translation (MT) output at run-time, *i.e.* without relying on reference translations. The shared task introduces (sub)tasks that assess translation quality from multiple perspectives, examining errors both at a higher level (segment scores) and with a more fine-grained view (error spans). Additionally, we expand our scope to generating corrected outputs through Automatic Post-Editing (APE).

Recently we have observed a gradual shift in the QE paradigms and methodologies, enabled by the advancement of neural metrics as well as large language models. Specifically, we have seen consistently strong performance across different language pairs and setups at sentence-level QE (Specia et al., 2021; Zerva et al., 2022; Blain et al., 2023), alongside increased efforts towards more finer-grained, explainable, and actionable evaluation of translations that focusses on error identification and explanation (Blain et al., 2023; Fernandes et al., 2023b; Guerreiro et al., 2023). The proliferation of LLM applications has led to significant performance improvements in MT, elevating the importance of advancing methodologies for quality estimation, and at the same time, it has allowed for novel perspectives and tasks related to quality estimation (Fabbri et al., 2022).

In light of the above, in this edition, we emphasise –beyond multilingual quality estimation– the analysis of the behaviour and abilities of submitted models with respect to different linguistic phenomena as well as their robustness to different error types and biases. Furthermore, we attempt to explore the degree to which quality estimation signals can be leveraged to improve translation quality via downstream automatic post-editing (Chatterjee et al., 2018b; Deoghare et al., 2023). We thus **bring APE under the QE umbrella** to make it easier for participants to develop QE systems and explore different techniques to apply it in APE shared task. These considerations collectively contribute to progress toward trustworthy and dependable QE systems that could facilitate real-time, reliable assessments of translation quality, as well as inform APE systems towards generating a corrected trans-

*Main organisers

Proceedings of the Ninth Conference on Machine Translation, pages 82–109
November 15-16, 2024

lation.

In this edition of the shared task, we further expand the provided resources for sentence-level and fine-grained QE, providing new test sets and expanding to new language pairs. Following the previous editions, we provide annotations for *direct assessments* (DA; English-Tamil, English-Hindi, English-Telugu and English-Gujarati), *post-edits* (PE; English-Tamil and English-Hindi) and *Multidimensional Quality Metrics* (MQM; English-Hindi, English-Spanish and English-German) (Lommel et al., 2014). We describe in detail the annotation process and provide statistics for the new resources in Section 3.

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 and provide new test sets emphasising low- and medium-resource languages and zero-shot approaches;
- To continue investigating the potential of fine-grained quality estimation;
- To study the robustness of QE approaches to different linguistic phenomena, error types and biases;
- To continue monitoring the computational efficiency of proposed approaches for sustainability purposes; and
- To study whether we can leverage QE signals to improve translation quality via downstream APE task.

We thus designed three tasks this year:

Task 1 The core QE task, which consists of separate sentence-level sub-tasks for different language pairs (§**??**). The goal is to predict a quality score for each segment in a given test set, which can be a variant of DA (§3.2) or MQM (§3.3).

Task 2 The fine-grained error prediction task, where participants were asked to detect error spans alongside error severities (*Major* versus *Minor*) (§2.2).

Task 3 A newly introduced task, which requires participants to combine quality estimation and automatic post-editing in order to correct the output of machine translation. (§2.3).

The tasks make use of large datasets annotated by professional translators with either $0 - 100$ DA scoring, post-editing or MQM annotations. We provide new training, development and test data for Task 3 as well as fresh new test sets for Tasks 1 and 2. The datasets and models released are publicly available[1].

Besides 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 *CodaBench* 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). We provide an additional evaluation focussed on robustness against different phenomena and biases in Section 7. A discussion on the main findings from this year's task is presented in Section 8.

2 Quality Estimation tasks

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

2.1 Task 1: Predicting translation quality

The ability to accurately estimate the quality of translations on-the-fly, i.e., without access to human references, is at the core of the QE shared task. This year, we focus on sentence-level quality, attempting to disentangle finer-grained analysis or post-edits that are tackled in Tasks 2 and 3.

Similar to the last edition, the data was produced as follows:

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 annotation: Each source-translation pair is evaluated by at least 1 expert annotator, and errors identified in texts are highlighted

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

and classified in terms of severity (minor, major, critical) and type (grammar correctness, omission, style, mistranslation, among others).

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 §2.1.1.

2.1.1 Sentence-level quality prediction

Similarly to the previous 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. Hence, 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 ρ as the primary metric**. This is what was used for ranking system submissions. Pearson's correlation coefficient, r, and Kendall τ were also computed as secondary metrics but not used for the final ranking of systems.

2.1.2 Finer-grained Evaluation and Challenge Sets

To assess the robustness and capabilities of automatic machine translation evaluation systems, we created a challenge set focusing on five different phenomena for the En-De and En-Es language pairs. Each category tests a particular aspect of translation quality that may have impact in real-world applications. The challenge set aims to determine whether evaluation systems can distinguish between correct translations—which we designate as hyp—and those containing subtle but relevant variations—which we designate as con.

Currency and date formatting This set tests the detection of format changes in currency symbols and date expressions. The hyp preserves the original source format (e.g., keeping "$100" or "MM/DD/YYYY"), while the con presents localized versions (e.g., "100 USD" or "DD/MM/YYYY"). Note that here it is the case that con is also a good-quality translation.

Word order This category examines the handling of word order variations. The hyp consists of monotonic translations that closely follow the source sentence order, while the con presents non-monotonic translations that rearrange words while preserving meaning. Evaluation models might have a preference towards one or the other, even though both preserve the meaning of the source.

Detached translations and omissions This set focuses on critical divergences from the source text. The hyp provides accurate and complete translations of the source. In contrast, the con includes examples where translations start correctly but then veer into unrelated topics or omit substantial portions of the source text. Evaluation systems are expected to detect these critical errors.

Idiomatic translations This category tests the handling of idiomatic expressions. The hyp presents idiomatic renderings that accurately convey the meaning in the target language, while the con offers literal word-for-word translations that may render the target text non-sensical. Evaluation systems should appropriately score translations that

prioritize conveying the correct meaning over strict word-for-word translation.

We generated data for all the phenomena listed above using GPT-3.5 (`gpt-3.5-turbo-0125`) and GPT-4 (`gpt-4-1106-preview`). Then, we conducted a human annotation study to discard erroneous triples.

Gender Subset The gender subset of the challenge set aims to study QE metrics and gender inflection in grammatical gender languages.

Following Zaranis et al. (2024), we collected unmodified instances from the counterfactual subsets (Es and De) of MT-GenEval (Currey et al., 2022), an evaluation set for sentence-level gender bias in machine translation. In these examples, sources from English Wikipedia mention exactly one human entity and contain intra-sentence lexical clues that help disambiguate the entity's gender identity.[2] Each source is provided with a masculine (M) and a feminine (F) variant (e.g., "She/He is a graduate of Harvard, but rarely applies such skills."). Human references are included as well.

We compiled the gender subset by constructing contrastive pairs as follows. First, we sampled 150 instances from the original MT-GenEval's subset. Fifty unique sources have a female referent and fifty a male referent. From each instance, we created a triplet with the source, the reference with correct gender inflection used as hypothesis, and the reference with wrong gender inflection used as contrast. Then, to isolate the impact of the source content, we created two triplets for each of the remaining fifty instances. The source in the triples is identical except for the gender identity of the entity. This step yields 100 more examples. The gender subset hence counts 200 contrastive triplets in total.

2.2 Task 2: Fine-grained error detection

For this task, we focus on finer-grained quality predictions, taking advantage of the detailed information provided in the MQM annotation schema. Specifically, each error span is annotated with error severity (*minor*, *major*, *critical*) as well as error type (see also Figure 1). Following the findings of the previous edition, we focus on the severity annotations and do not use the other error categories annotated in the MQM schema. As a result, we aimed to (1) identify error spans and (2) classify said error spans as either *minor* or *major*. We note that we merge the critical and major categories, since in this edition we noticed particularly sparse occurrences of critical errors (even less than the previous year). Additionally, in this edition, the annotations included a *neutral* category, which was ignored as it was (1) not occurring for all language pairs and (2) they correspond to subjective opinions/preferences about translation. [3] We point readers to Figure 3 for some statistics on error severity distribution per language pair and domain.

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 continuous sequences of 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 are allowed and count towards *recall* of different errors, but overlapping annotations are flattened for both gold and system annotations (see below). Figure 1 shows an example of annotations.

For the evaluation, the primary metric is the **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. With respect to overlapping annotations, we allow for multiple character level annotations[4] and consider the best matching annotation per character position. As such, for each segment, we compute recall for the characters in 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 1 show an example of the aforementioned process [5].

[2] We acknowledge a notion of gender identity beyond the binary. However, we include only masculine and feminine examples as they are provided in the original dataset.

[3] Note that the neural errors are also not considered when computing an MQM score.

[4] 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.

[5] The link to evaluation scripts can be found at: https://github.com/WMT-QE-Task/qe-eval-scripts/blob/main/wmt24/

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

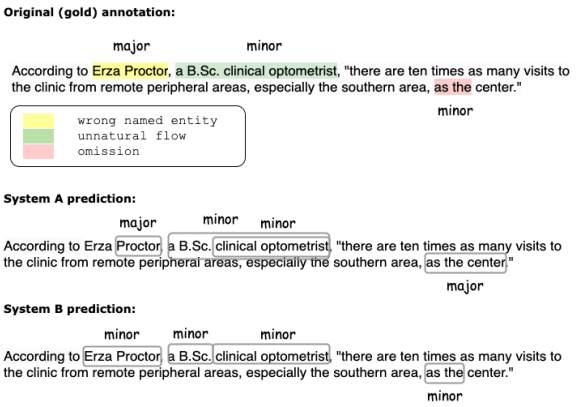


Figure 1: Example of gold annotations (MQM) for Task 2 (top) and respective prediction examples (bottom). Example taken from He-En test set.

2.3 Task 3: QE-informed APE

MT Automatic Post-Editing (APE) is the task of automatically correcting errors in a machine-translated text. As pointed out by Chatterjee et al. (2015), from the application point of view, the task is motivated by its possible uses to:

- Enhance MT output by harnessing information that is not available to the decoder or by conducting deeper text analysis, which may be prohibitively expensive during the decoding phase.
- Address systematic errors stemming from an MT system whose decoding process is inaccessible for focused modifications.
- Provide professional translators with improved MT output quality, thereby reducing the need for subsequent human post-editing.
- Tailor the output of a general-purpose MT system to align with the lexicon and style requirements of a specific application domain.

Building on the work of Chatterjee et al. (2018b); Deoghare et al. (2023), which demonstrated the potential of QE to enhance APE systems, this edition of the WMT QE shared task introduced the new QE-informed APE subtask. In this subtask, we focus on a unified *evaluation and correction* paradigm, taking advantage of the additional information provided by the human post-edits. Participants were encouraged to incorporate signals from QE systems to improve APE performance. The evaluation setup remained consistent with the previous rounds WMT APE shared tasks, requiring participants to automatically correct translations generated by a generic, domain-unadapted "black-box" NMT system. The training data consisted of human post-edits of translations produced by this system. While TER (Snover et al., 2006) and BLEU (Papineni et al., 2002) continued as the primary and secondary evaluation metrics, this year also introduced chrF (Popović, 2015) and COMET[6] for a more comprehensive automatic evaluation of the submitted APE systems.

For this year, English-Hindi and English-Tamil were the selected language pairs, with Hindi and Tamil as the target languages for post-editing. The training, development, and test data encompassed a wide range of domains, including education, legal, healthcare, culture, tourism, reviews, subtitles, and general/news.

3 Datasets

Below, we describe the datasets provided to participants for development and testing. Specifically, this year, we provided training data only for Task 3, which was newly introduced (see §3.4).

3.1 Training Resources

Overall, participants were encouraged to employ training data from a wide range of sources, including datasets from previous competitions, as well as synthetic or proprietary data.

Proposed training data for DA annotations, following the previous editions, includes the language pairs from the MLQE-PE dataset (Fomicheva et al., 2022), as well as the data from the previous QE editions (Zerva et al., 2022; Blain et al., 2023). Similarly, for the MQM data, we encouraged participants to refer to data from previous editions that

[6] https://github.com/Unbabel/COMET .

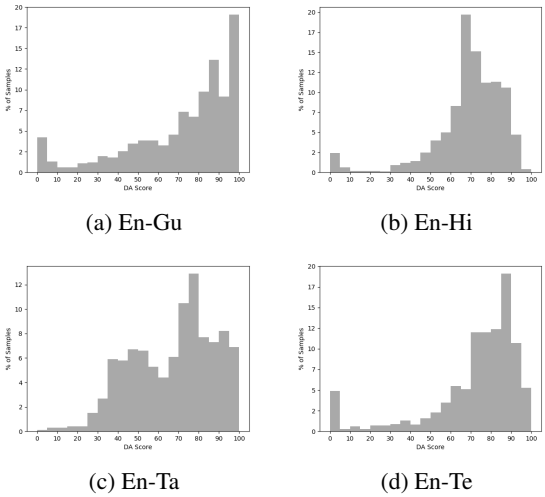
cover translation into German (En-De), Russian (En-Ru), Hebrew (En-He) and out of Chinese (Zh-En) (Freitag et al., 2021a,b), as well as the Indic-MT eval dataset (Sai B et al., 2023). However, we emphasise that in this edition, we introduce no new training data, treating the translations into Spanish (En-Es) and Hindi (En-Hi) as zero-shot tasks, and only En-De as supervised.

3.2 Direct Assessment (DA) Data

For all language pairs, the data provided is selected from publicly available resources.

We expand the Indic language pairs introduced in previous years, providing new unseen test sets of approx 1K segments each for Hindi (Hi; 1000 segments) and Gujarati (Gu; 1012 segments) as target languages from the Indo-Aryan language family as well as Tamil (Ta; 1000 segments) and Telugu (Te; 1000 segments) from the Dravidian language family. Following the previous edition, 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[7] (General, Healthcare and Legal domain; *ii*) IITB English-Hindi parallel corpus[8] (Kunchukuttan et al., 2018) (Culture/Tourism domain), and parallel segments scraped from NPTEL[9]; and *iii*) SpokenTutorials[10] (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 filter parallel segments using LaBSE, and select source sentences with varying token lengths, while the translation was obtained using $1.3B$ parameter NLLB model (Costa-jussà et al., 2022), as discussed in (Blain et al., 2023). 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. We provide the distribution of DA scores for each language pair in Figure 2, where we can see that for all language pairs, we have similar distributions skewed towards high-quality scores. We can also observe that for Tamil, we have fewer segments of very low quality (DA ≤ 20), but instead, we have larger counts of segments of moderate quality ($20 \leq \text{DA} \leq 60$).



(a) En-Gu (b) En-Hi

(c) En-Ta (d) En-Te

Figure 2: Distribution of DA scores for the Indic language pairs.

3.3 MQM Data

As **test data**, we annotated new evaluation sets for three language directions: English-German (En-De), English-Spanish (En-Es) and English-Hindi (En-Hi). The evaluation sets were annotated by professional translators following a MQM typology (Burchardt, 2013) and specific guidelines[11].

The documents used for the evaluation sets are shared with the WMT General MT task and follow the same distribution of domains in that data (*e.g.*, news, social, literary and speech). The full documents were translated using the $54B$ parameters NLLB model (Team et al., 2022)[12] without sentence splitting. We subsequently split segments for annotation and annotated a total of 1511 segments for each translation direction.

The test data distribution according to error severities is shown in Figure 3. The NLLB model used to translate the evaluation sets is clearly stronger for En-De, with less than 100 major and minor errors for each content type. The distribu-

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

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

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

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

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

[12] Model identifier FACEBOOK/NLLB-MOE-54B

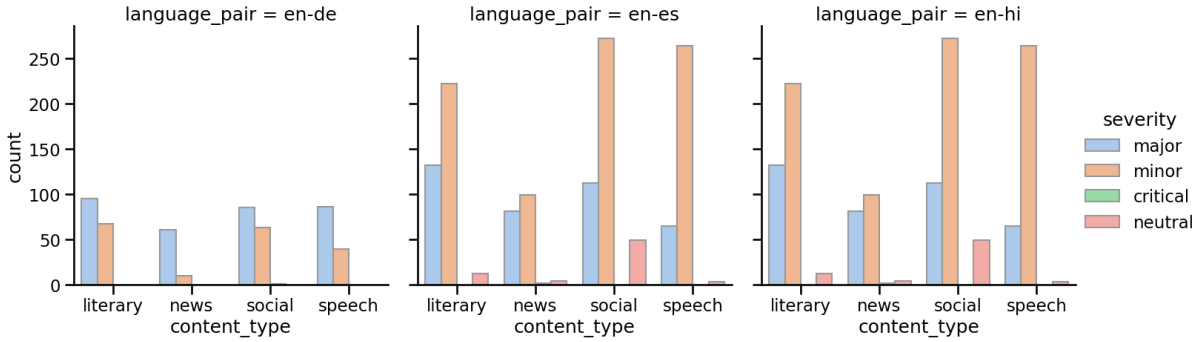


Figure 3: Distribution of error severities across language pairs and domains/content types.

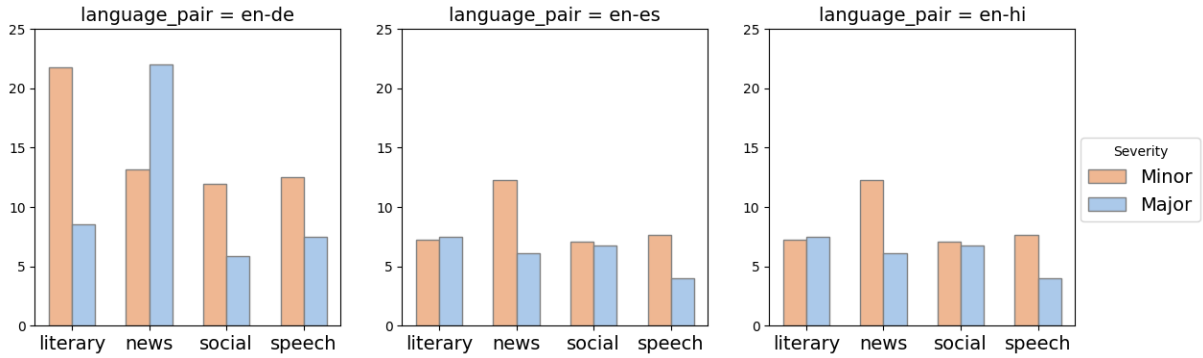


Figure 4: Distribution of average length (character count) for different severities across language pairs and domains/content types.

tion of major and minor errors changes drastically for En-Es and En-Hi, in particular the number of minor errors for the literary, social and speech domains, with more than 200 minor errors each. In addition, we can see that we have fewer errors for the news domain across all three language pairs, both in terms of minor and major errors. Contrary to frequency, however, Figure 4 shows that error spans identified for En-De are significantly longer on average for both identified error categories.

3.4 QE-APE Data

This year we introduce two new language pairs for the APE task: English-Hindi (*En-Hi*) and English-Tamil (*En-Ta*). For each language pair, the train, dev, and test sets respectively consist of $7,000$, $1,000$, and $1,000$ (*source*, *target*, *human post-edit*) triplets, where:

- The source (SRC) is an English sentence;
- The target (TGT) is a Hindi/Tamil translation of the source produced by a generic, black-box NMT system unknown to participants.
- The human post-edit (PE) is a manually revised version of the target, which was produced by native Hindi/Tamil speakers.

The English-Hindi train, dev, and test sets span culture, education, health, tourism, and general domains. Similarly, English-Tamil APE datasets contain sentences from legal, literacy, reviews, subtitles, news, health, and general domains.

We also provide a corpus of artificially generated data as additional training material. It consists of 2.5 million triplets for each language pair derived from the Anuvaad parallel corpus. Specifically, the source, target, and post-edit instances of this synthetic corpus are respectively obtained by combining: i) the original English source sentence from the Anuvaad corpus, ii) its automatic translation into Marathi, iii) the original Marathi target sentence from the Anuvaad corpus. Furthermore, we provide the DA scores for all samples in both train and dev sets. Additionally, the participants were encouraged to use the DA data released in the earlier iteration of the QE shared task for these language pairs.

To get an idea of the task difficulty, we focused on three aspects of the released data, which provided us with information about the possibility of learning useful correction patterns during APE

	Lang.	Domain	MT type	RR_src	RR_tgt	RR_pe	Basel. BLEU	Basel. TER	δ TER
2015	en-es	News	PBSMT	2.9	3.31	3.08	n/a	23.84	+0.31
2016	en-de	IT	PBSMT	6.62	8.84	8.24	62.11	24.76	-3.24
2017	en-de	IT	PBSMT	7.22	9.53	8.95	62.49	24.48	-4.88
2017	de-en	Medical	PBSMT	5.22	6.84	6.29	79.54	15.55	-0.26
2018	en-de	IT	PBSMT	7.14	9.47	8.93	62.99	24.24	-6.24
2018	en-de	IT	NMT	7.11	9.44	8.94	74.73	16.84	-0.38
2019	en-de	IT	NMT	7.11	9.44	8.94	74.73	16.84	-0.78
2019	en-ru	IT	NMT	18.25	14.78	13.24	76.20	16.16	+0.43
2020	en-de	Wiki	NMT	0.65	0.82	0.66	50.21	31.56	-11.35
2020	en-zh	Wiki	NMT	0.81	1.27	1.2	23.12	59.49	-12.13
2021	en-de	Wiki	NMT	0.73	0.78	0.76	71.07	18.05	-0.77
2022	en-mr	health/tourism/news	NMT	1.46	0.89	0.72	67.55	20.28	-3.49
2023	en-mr	health/tourism/news	NMT	1.85	1.24	1.12	70.66	26.60	+1.13
2024	en-hi	health/tourism/news	NMT	2.7	3.55	3.32	39.28	46.36	-19.29
2024	en-ta	health/tourism/news	NMT	1.97	1.49	1.1	70.16	24.71	-0.47

Table 2: Basic information about the APE shared task data released since 2015- languages, domain, type of MT technology, repetition rate and initial translation quality (TER/BLEU of TGT). The last column (δ TER) indicates, for each evaluation round, the difference in TER between the baseline (*i.e.,* the "*do-nothing*" system) and the top-ranked official submission.

model training and successfully applying them at test time. These are: *i)* repetition rate, *ii)* MT quality, and *iii)* TER distribution in the test set. For the sake of comparison across the nine rounds of the APE task (2015–2023), Table 2 reports, for each dataset, information about the first two aspects. The third aspect, however, will be discussed by referring to Figure 5 and Figure 6.

3.4.1 Repetition Rate

The repetition rate (RR), measures the repetitiveness inside a text by looking at the rate of non-singleton n-gram types ($n = 1...4$) and combining them using the geometric mean. Larger values indicate a higher text repetitiveness that may suggest a higher chance of learning from the training set correction patterns that are also applicable to the test set. However, over the years, the influence of repetition rate in the data on system performance was found to be marginal.[13]

As shown in Table 2, in this edition, the RR for English-Hindi ranges between 2.7-3.3, and for English-Tamil RR ranges between 1.1-2.0. This difference may contribute to motivating the significantly different APE results observed for the two languages, as evidenced by a substantial TER reduction for English-Hindi (-19.29 "δ TER") compared to the "do-nothing" the baseline (see §4.3). Reviewing previous rounds of the APE task, however, suggests that RR remains only a partially informative indicator of task difficulty due to its variable correlation with final results, which may also depend on other factors or on the interaction of multiple factors that are yet to be fully understood.

3.4.2 MT Quality

Another complexity indicator is MT quality, which is the initial quality of the machine-translated (TGT) texts to be corrected. We measure it by computing the TER (↓) and BLEU (↑) scores (Basel. TER/BLEU rows in Table 2) using the human post-edits as reference. In principle, higher quality of the original translations leaves the APE systems with less room for improvement since they have, at the same time, less to learn during training and less to correct at the test stage. On one side, training on good (or near-perfect) automatic translations can drastically reduce the number of learned correction patterns. On the other side, testing on similarly good translations can *i)* drastically reduce the number of corrections required and the applicability of the learned patterns, and *ii)* increase the chance of introducing errors, especially when post-editing near-perfect translations. The findings of all previous rounds of the task support this observation, which is corroborated by the high correlation (>0.78) between the initial MT quality ("Basel. TER" in Table 2) and the TER difference between the baseline and the top-ranked submission ("δ TER" in Table 2).

As discussed in Section 6.3, this year seems to confirm the trends observed in the past. For English-Hindi, the baseline TER is quite high (46.36 points), leaving more room for improvement.

[13]The analyses carried out over the years produced mixed outcomes, with impressive final results obtained in spite of low repetition rates (Chatterjee et al., 2020) and vice-versa (Chatterjee et al., 2018a, 2019; Akhbardeh et al., 2021).

Whereas English-Tamil falls in medium-high difficulty (20.0<TER<25.0), making the task more challenging. The final gains ("δ TER" in Table 2) confirm the correlation between the quality of the initial translations and the actual potential of APE.

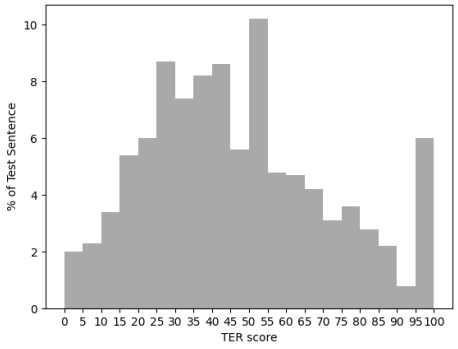


Figure 5: TER distribution in the APE 2024 English-Hindi test set.

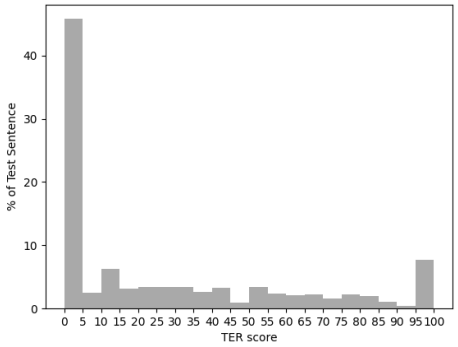


Figure 6: TER distribution in the APE 2024 English-Tamil test set.

3.4.3 TER Distribution

A third complexity indicator is the TER distribution (computed against human references) for the translations present in the test sets. Although TER distribution and MT quality can be seen as two sides of the same coin, it's worth remarking that, even at the same level of overall quality, more/less peaked distributions can result in very different testing conditions. Indeed, as shown by previous analyses, harder rounds of the task were typically characterised by TER distributions particularly skewed towards low values (*i.e.,* a larger percentage of test items having a TER between 0 and 10). On one side, the higher the proportion of (near-)perfect test instances requiring few edits or no corrections at all, the higher the probability that APE systems will perform unnecessary corrections penalised by automatic evaluation metrics. On the other side, less skewed distributions can be expected to be easier to handle as they give automatic systems larger room for improvement (*i.e.,* more test items requiring - at least minimal - revision). In the lack of more focused analyses on this aspect, we can hypothesise that in ideal conditions from the APE standpoint, the peak of the distribution would be observed for "post-editable" translations containing enough errors that leave some margin for focused corrections but not too many errors to be so unintelligible to require a whole re-translation from scratch.[14]

As shown in Figure 5, for English-Hindi the TER distribution follows more or less uniform distribution. The distribution is not too skewed towards near-perfect translation (which would have made it harder to further improve), nor towards the higher end of TER (which would have made it harder to learn error-correction patterns due to too noisy data). These characteristics make it easier to improve translation, which is reflected in the final evaluation results. On the other hand, as shown in Figure 6, for English-Tamil the TER distribution is highly skewed towards near-perfect translations. Around half of the test set falls in 0-5 TER points, making it prone to over-correction, which can be penalised by automatic evaluation metrics. This characteristic makes the English-Tamil test set much more challenging when it comes to gaining further translation quality improvements.

4 Baselines

In this edition, we opted to use publicly available, existing models without further tuning. Hence, we use a more unified architecture for Tasks 1 and 2, where all models use a large XLM-RoBERTa pre-trained encoder without additional language tuning (see also Appendix A for hyperparameter details). The specific hyperparameters used are presented in Table 7. For Task 3, we opted for a simple "do nothing" approach as discussed in Section 4.3.

4.1 Task 1: Quality Estimation

For the **sentence-level** sub-task, we opted for using CometKiwi 2022 (Rei et al., 2022) which was trained on data from the Metrics and QE shared tasks (combining data from previous years up to

[14]For instance, based on the empirical findings reported in (Turchi et al., 2013), TER=0.4 is the threshold that, for human post-editors, separates the "post-editable" translations from those that require complete rewriting from scratch.

2022). Models are publicly available for download[15].

4.2 Task 2: Fine-grained Error Detection

For **Task 2** we also used a CometKiwi model, specifically one trained on the multi-task setting, to produce both sentence-level scores and word-level quality estimates. The model, trained on 2022 QE data is publicly available. [16] The word-level estimates are in the form of OK/BAD tags, and for this reason it is necessary to convert the original output to the one required by the Task 2 format. As such we process the word-level predictions as follows:

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

4.3 Task 3: QE-informed APE

The official baseline results for **Task 3** are the TER/BLEU/chrF/COMET scores calculated by comparing the raw MT output with human post-edits. This corresponds to the score achieved by a "do-nothing" APE system that leaves all the test segments unmodified.

5 Participants

In this section, we present 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 in, and point to relevant publications, if any.

Unbabel (T1; all): The submission for Task 1 follows their work from the previous competition (Rei et al., 2023), which corresponds to an ensemble of multiple checkpoints for the sentence-level subtask, using a weighted averaging of the predicted scores, optimised by language pair. The emphasis is on scaling the size of the pre-trained encoder from InfoXLM to XLM-R XL and XXL.

Pister Labs (T1; all): The team opted for an approach where they generated a set of reading comprehension questions and scored each hypothetical translation by evaluating how well it could answer the comprehension question when compared with the reference translation. The overall score for a hypothetical is then a simple average across the questions asked of it. Answers are generated by providing the question and the hypothetical translation to Llama3.1-8B (Dubey et al., 2024). The initial set of reading comprehension questions is generated through few-shot prompting of Llama3.1-70B, and evaluating results on a subsample of 100 training En-De translation pairs with Llama3.1-70B. The four questions with the highest Spearman correlation were then used for final testing. To improve question generation quality, they use techniques from OpenAI and Anthropic's prompting guides, as well as the self-consistency technique.

HW-TSC (T1; En-Hi, En-Ta, En-Te, En-Gu): The team employed the CROSS-QE approach (Li et al., 2023) as the basis for further tuning and opted for tuning separate models for each language pair. They used encoder-based models, experimenting with different encoders, which were trained on different combinations of source and translation vectors as input. They focused on improving model performance both in terms of training by employing different data augmentation methods and in terms of inference, exploring better strategies for ensembling checkpoints. In terms of data augmentation, they use a combination of LLMs with specific prompts to generate pseudo-data as well as text editing methods. [17]

HW-TSC (T2; all): The team employs a combination of LLMs, hypothesising that the reasoning abilities of large models may be helpful in the fine-grained task. They use the TowerInstruct-7B-v0.2 (Alves et al., 2024) model and the GPT-4o-mini (Islam and Moushi, 2024) model, using prompt engineering and in-context learning to obtain the predictions. Additionally, they employ data augmentation techniques mentioned for Task 1 and find that they can rely on pseudo-data for tuning the models. [18]

[15] https://huggingface.co/Unbabel/wmt22-cometkiwi-da

[16] https://huggingface.co/Unbabel/WMT24-QE-task2-baseline

[17] We consider submissions from users s50042889 and zhaoxf4 mentioned in the results page as one submission

[18] We consider submissions from users zhuming, zhaoxf4 and mengyao mentioned in the results page as one submission

TMU-HIT (T1; En-Hi, En-Ta, En-Te, En-Gu): The team submitted predictions that rely on LLMs, inspired by (Liu et al., 2023; Enomoto et al., 2024). They designed custom prompts for quality estimation and employed GPT-4o mini (Achiam et al., 2023) to sample assessment scores multiple times using the same prompt. They then experimented with combining the generated scores to compute the final score using either their average or their weighted sum, employing the generation probabilities as weights for the latter. They conducted evaluation experiments in both zero-shot and three-shot settings. Further, they also attempted fine-tuning GPT-4o mini using the training data released for the WMT23 Machine Translation task (Kocmi et al., 2023).

HW-TSC (T3; all): (Yu et al., 2024) The team explored two distinct approaches for developing APE systems. For the En-Hi pair, they leveraged the Llama3-8B-Instruct model through continual pre-training on the collected data and then supervised fine-tuning it on the real APE data. For the En-Ta pair, they trained a transformer model from scratch, first focusing on the MT (Machine Translation) task using web-collected data, followed by training on APE data. External MT candidates were incorporated during the training to boost performance further. To prevent over-correction, Sentence-level QE models were employed to select between MT and APE outputs. Both users (**HW-TSC_yjwsss** and **HW-TSC_zhaoxf4**) from this team made the same submissions for En-Ta, but different submissions for En-Hi.

IT-Unbabel (T3; all): IT-Unbabel submission leveraged xTower (Treviso et al., 2024), a model built on top of TowerLLM (Alves et al., 2024), which is designed to provide free-text explanations for translation errors to guide the generation of an improved translation. The system was trained on material that includes the xTower dataset (GPT-4 generated explanations for translation correction), TowerBlocks, and additional training datasets provided by the WMT24[19] organizers for English-Hindi and English-Tamil, augmented with error span annotations from xCOMET (Guerreiro et al., 2023). A hybrid approach is used to dynamically select between the original translation and the corrected version produced by the xTower model using a quality estimation model.

[19]https://www2.statmt.org/wmt24/

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 described in the Task 1 overview (§2.1.1), sentence-level submissions are evaluated against the true z-normalised sentence scores using Spearman's rank correlation coefficient ρ along with the following secondary metrics: Pearson's correlation coefficient, r, and Kendall's τ. Nonetheless, the final ranking between systems is calculated using the primary metric only (Spearman's ρ). Statistical significance was computed using William's test. The results are shown in Table 3.

Looking at the obtained scores, we observe an overall performance increase for the sentence-level scores compared to previous years for all language pairs (that have been previously tested) except for En-Ta, where we observe a small drop. We note, that while the domains and sources in the En-De MQM test-set are different, all DA test-sets are drawn from the same sources and observe similar score distributions to previous years, thus facilitating comparisons.

It should be noted that there is no clear winner across language pairs. Instead, different systems rank first for each language.

6.2 Task 2

For Task 2, the submissions are scored using the F1-score, computed at character level for the annotated error spans, as described in Section 2.2. 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 4.

This year, the fine-grained annotation task (Task 2) had a lower participation rate compared to the

		Multidimensional Quality Metric (MQM)			Direct Assessment (DA)			
Model	Multi	En-De	En-Es	En-Hi	En-Hi	En-Gu	En-Te	En-Ta
Unbabel	0.553	0.512 †	0.345 †	0.412	0.714	0.703 †	0.510 †	0.675 †
Pister Labs	0.452	0.513 †	0.242	0.363	0.564	0.587	0.379	0.478
HW-TSC	-	-	-	-	0.719 †	0.757 †	0.482 †	0.683†
TMU-HIT	-	-	-	-	0.739 †	0.713	0.482	0.603
BASELINE	0.520	0.514 †	0.340 †	0.441 †	0.678	0.661	0.414	0.592

Table 3: Spearman correlation for the official submissions to WMT24 Quality Estimation **Task 1 Sentence-level**. Baseline systems are highlighted in grey. For each language pair, results marked with † correspond to the winning submissions, as they are not significantly outperformed by any other system according to the Williams Significance Test (Williams, 1959).

.

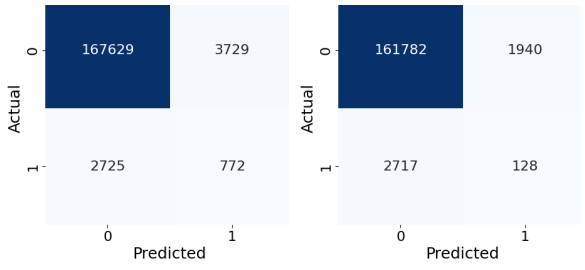
	Multidimensional Quality Metric (MQM)			
Model	Multi	En-De	En-Es	En-Hi
BASELINE	0.278	0.192 †	0.161 †	0.481 †
HW-TSC	0.227	0.178	0.151	0.362

Table 4: F1-score for the official submissions to WMT24 Quality Estimation **Task 2 Error Span Detection**. Baseline systems are highlighted in grey. For each language pair, results marked with † correspond to the best system (not significantly outperformed by any other system) according to randomized paired t-test.

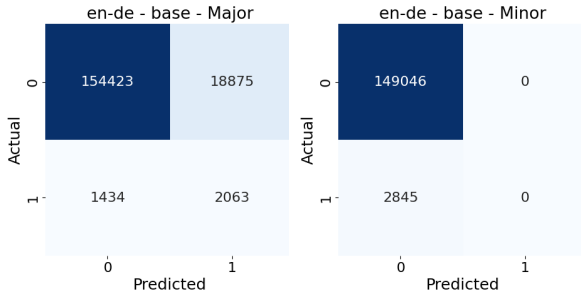
previous edition, and we can also see that the obtained scores remained particularly low, indicating that the task remains challenging and difficult to address.

Specifically, if we focus on confusion matrices shown in Figure 7 for the submission received, we can see that the Baseline is over-predicting *Major* error spans, which gives a slight advantage regarding the F1 score since it leads to higher recall. This finding is consistent with higher precision obtained by HW-TSC submission as seen in the Appendix C, Table 17. We provide the confusion matrices for all language pairs in Appendix E.

Despite this, it is important to note that the methods submitted for Task 1 still seem to benefit from a multi-task approach that considers word-level information. Taking both these observations into account and looking towards future editions, it might be useful to redesign the task, aiming either at a different span representation that would perhaps attempt a better normalisation over different span lengths or deviate from the character level representation. Another alternative view would be to encourage methods that use error spans to support or interpret sentence-level quality (Leiter et al., 2023) or concentrate only on specific error types.



(a) HWTSC - Major (b) HWTSC -Minor



(c) Baseline - Major (d) Baseline - Minor

Figure 7: Confusion matrices for Task 2 English-German, comparing Minor and Major predictions between the Baseline system and the HWTSC one.

6.3 Task 3

6.3.1 Automatic Evaluation

Automatic Post-Editing evaluation results are shown in Table 5. The submitted runs are ranked based on the average TER (case-sensitive) computed using human post-edits of the MT segments as a reference, which is the APE task's primary evaluation metric. To provide a broader view of the systems' performance, BLEU, chrF, and COMET results computed using the same references are also reported. As can be seen from the table, all submissions for English-Hindi outperform the baseline by a significant margin across all metrics, with TER reductions that are always statistically significant. the baseline. The best system is able to improve trans-

		TER	BLEU	CHRF	COMET
En-Hi	IT-Unbabel	**27.08**	58.38	73.45	0.8646
	HW-TSC_yjwsss	**30.37**	54.50	71.06	0.8514
	HW-TSC_zhaoxf4	**31.32**	52.74	69.83	0.8517
	BASELINE (MT)	46.36	39.28	59.48	0.8084
En-Ta	HW-TSC	24.24	69.64	82.36	0.9186
	IT-Unbabel	24.54	70.05	82.30	0.9163
	BASELINE (MT)	24.71	70.16	81.80	0.9137

Table 5: Official results for the WMT24 Quality Estimation **Task 3 QE-informed APE** English-Hindi and English-Tamil shared task – average TER (↓), BLEU (↑), chrF (↑), COMET (↑). Statistical significance test is computed for the primary metric (TER) *wrt.* the baseline and the significant results are highlighted in bold. Baseline systems are highlighted in grey.

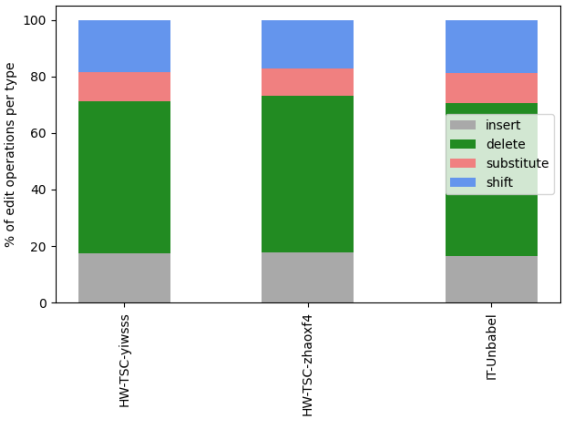


Figure 8: Distribution of edit operations (insertions, deletions, substitutions and shifts) performed by the three primary submissions to the WMT24 APE English-Hindi shared task.

lation quality by nearly 20.0 TER points. However, for English-Tamil, we observe that while all submissions performed slightly better than the baseline in terms of absolute scores across all metrics except BLEU, none of the systems show statistically significant gains compared to the baseline. As discussed in Section 3.4, this can be attributed to the combined effect of less repetitive data (between 1.1-2.0) compared to English-Hindi (between 2.7-3.3) and a stronger baseline (24.7 vs 46.4 TER), leaving less room for improvement.

6.3.2 Analysis: Systems' Behaviour

Modified, improved and deteriorated sentences. To better understand the behaviour of each APE system, we now turn toward the changes made by each system to the test instances. To this end, Table 6 shows, for each submitted run, the number of modified, improved and deteriorated sentences, as well as the overall system's precision (*i.e.,* the proportion of improved sentences out of the total number of modified instances for which improvement/deterioration is observed). It's worth noting that, as in the previous rounds, the number of sentences modified by each system is higher than the sum of the improved and the deteriorated ones. This difference is represented by modified sentences for which the corrections do not yield any TER variations.

As can be seen from Table 6, for English-Hindi, all submissions perform aggressive post-editing, with the top submission modifying 96.5% of the translations, where most of the modifications lead to improving the translation quality with a precision score of 84.56%. In contrast, for English-Tamil, all submissions adopt a conservative approach, limiting edits to 3.8%-4.8% of the test set. This aligns with our previous observations on task difficulty, driven by the higher MT baseline and the skewed TER distribution, with samples concentrated in the near-perfect translation range. In this challenging scenario, all submissions are able to improve the majority of modified translations with a precision score between 54%-59%.

Edit operations. Similar to previous rounds, we analysed systems' behaviour also in terms of the distribution of edit operations (insertions, deletions, substitutions and shifts) done by each system. This fine-grained analysis of how systems corrected the test set instances is obtained by computing the TER between the original MT output and the output of each primary submission taken as reference. As shown in Figures 8 and 9, similar to last year, differences in systems' behaviour are minimal. All of them are characterised by a large number of deletions, followed by insertions, shifts and substitutions. For English-Tamil, we observe a relatively lower proportion of shifts and substitutions compared to English-Hindi. This might indicate that English-Tamil might have more diverse APE outputs, which might be more challenging to evaluate with reference-based automatic metrics.

7 Evaluation on challenge sets

We received two submissions that we could evaluate on challenge sets: Pister Lab's submission, based on prompting Llama 3.1, and Unbabel's, based on CometKiwi. In Figure 10, we report the percentage of samples where the hyp translation is

	Systems	Modified	Improved	Deteriorated	Prec.
En-Hi	IT-Unbabel	965 (96.5%)	756 (78.35%)	138 (14.30%)	84.56
	HW-TSC_yjwsss	952 (95.2%)	688 (72.27%)	171 (17.96%)	80.09
	HW-TSC_zhaoxf4	665 (66.5%)	532 (80.00%)	85 (12.78%)	86.22
En-Ta	HW-TSC	48 (4.8%)	25 (52.08%)	18 (37.50%)	58.14
	IT-Unbabel	38 (3.8%)	19 (50.00%)	16 (42.11%)	54.29

Table 6: Number (raw and proportion) of test sentences modified, improved and deteriorated by each run submitted to the APE 2024 English-Hindi and English-Tamil sub-task. The "Prec." column shows systems' precision as the ratio between the number of improved sentences and the number of modified instances for which improvement/deterioration is observed (*i.e.,* Improved + Deteriorated).

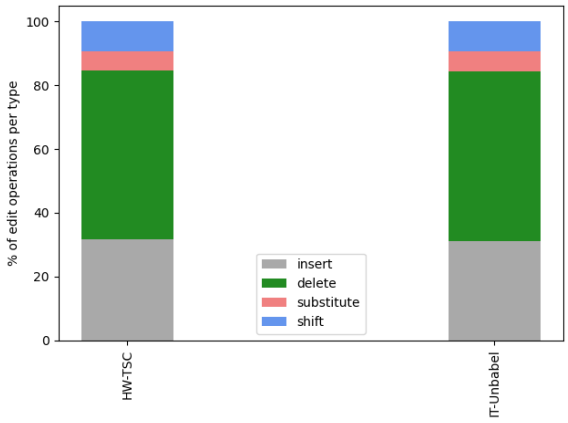


Figure 9: Distribution of edit operations (insertions, deletions, substitutions and shifts) performed by the three primary submissions to the WMT24 APE English-Tamil shared task.

scored higher, lower, or is tied to the con hypothesis.[20] Please refer to Section 2.2 for details on constructing these translation pairs for each phenomenon.

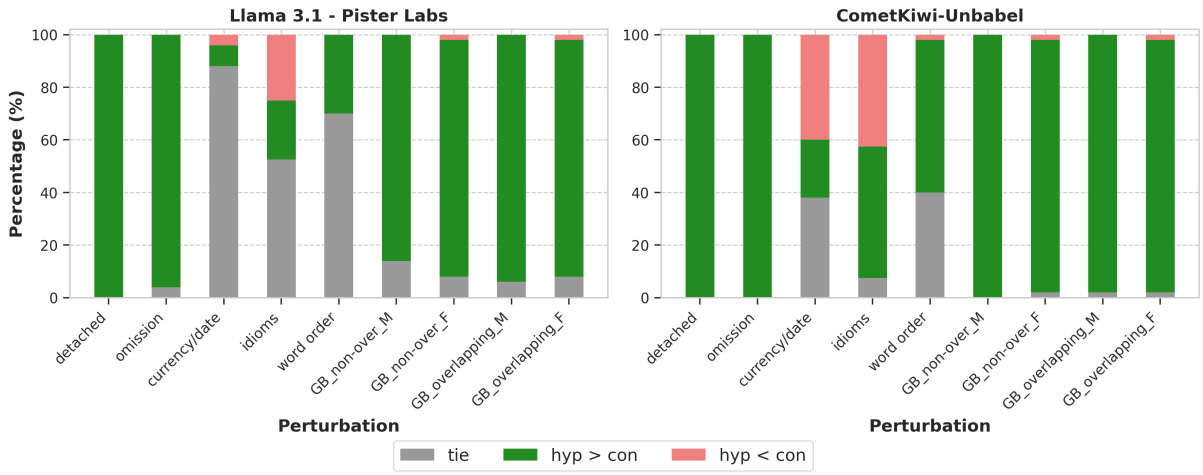
Detached translations and omissions Out of all the phenomena studied, these two constitute the most critical errors. It is thus highly encouraging that both models perform perfectly across the two language pairs in consistently scoring the correct hyp translation higher than the erroneous con translation.

Currency and date formatting This category reveals interesting differences between the two models. Llama 3.1 shows a high tie rate, indicating it often does not distinguish between original and localized formats. This suggests a more neutral stance towards formatting choices. In contrast, CometKiwi is more sensitive to these formats, behaving less predictably. Although, in most cases, it either prefers the source format or is indifferent to the localized format, there are some cases, in particular for en-es translations, where it does prefer the localized format that does not lexically match that found in the source text.

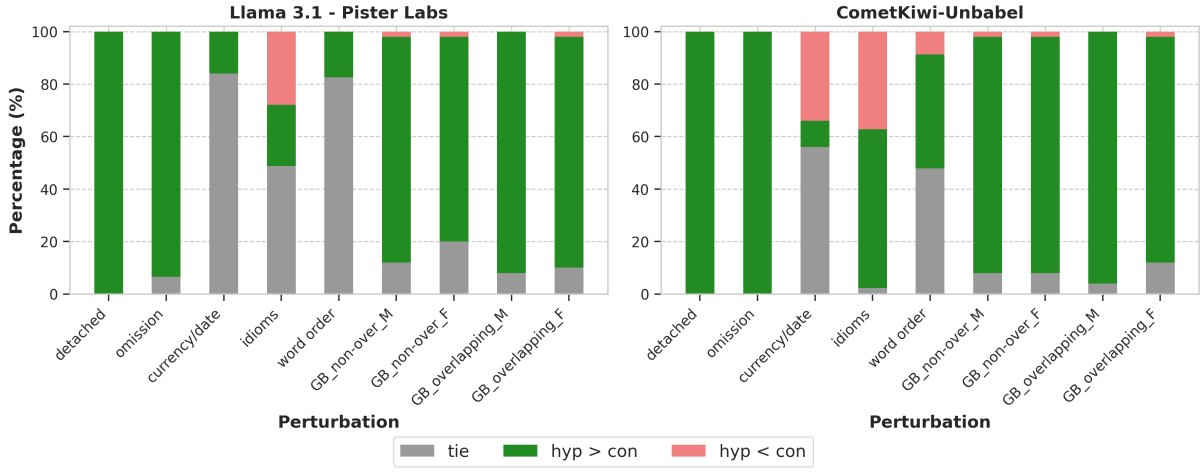
Idioms Llama 3.1 predominantly shows ties or a slight preference for non-literal, idiomatic renderings (hyp) that accurately convey the meaning in the target language. In contrast, CometKiwi's behavior is more varied and, perhaps surprisingly, often favors literal translations (con) even when they may not preserve the source text's meaning in the target language. This tendency towards literalness can be quite problematic in the context of idioms and other figurative texts, where meaning often diverges from word-for-word translations. One potential way to alleviate these trends is to train neural metrics with more diverse data that includes idiomatic and figurative language to improve their robustness.

Word order Here, Llama 3.1 shows a high rate of ties, suggesting that, similarly to what we found for the currency and date formatting phenomenon, it does not distinguish between monotonic translations that closely follow the source sentence order and non-monotonic translations that rearrange words while preserving meaning. This suggests that Llama 3.1's scoring may be more tied to the overall meaning of the translation. In contrast, CometKiwi demonstrates more preference for monotonic translations (hyp) across both language pairs, particularly for en-de. As such, CometKiwi appears to be more sensitive to word order, poten-

[20] Inspired by the analysis in Kocmi et al. (2024), we consider a tie with CometKiwi when the absolute difference between the scores of the hyp and con hypotheses is lower or equal to 0.1 points. For the Llama-based submission, for its more coarse-grained scoring range (more akin to a categorical distribution), we consider a tie when both translations receive the same score.



(a) en-de



(b) en-es

Figure 10: Share of instances in challenge sets where participant systems ranked the hypothesis translation higher than (green), lower than (salmon), or equal to (grey) the contrast. Results on en-de (top) and en-es (bottom).

tially favouring translations that maintain a structure closer to the source text. As a learned metric, this behaviour might be attributed to CometKiwi's training data, which may have contained more monotonic translations (more common among classical encoder-decoder NMT models that constitute most of the translations that the model has seen during training) than paraphrastic or non-literal ones (more prevalent among the more novel LLM-based translation approaches (Raunak et al., 2023)).

Gender subset In most instances, both systems score the hypothesis with the correct gender inflection higher. However, we noticed that some cases have ties, which we consider as errors: the model does not capture the difference in gender forms and wrongly assigns equal scores to the hypothesis and the contrast. Expectedly, this phenomenon is more present in Pister Lab's scores, as Llama 3.1 tends to assign more coarse-grained assessments. In analyzing sources with non-overlapping content, Llama 3.1 exhibits a higher frequency of errors for male sources in `en-de` translation while demonstrating increased error rates for female sources in `en-es`. Conversely, CometKiwi maintains a comparable error rate across genders in both language pairs, with an elevated error rate in `en-es` translation overall. When examining sources with identical content differentiated only by gender (categorized as "overlapping"), we observed higher errors for female sources across all configurations, except for CometKiwi's performance in `en-es`.

Closing remarks Our analysis of Llama 3.1 and CometKiwi on various challenge sets reveals distinct behaviours and potential areas for improvement. Both models excel at identifying critical errors like detached translations and omissions.

However, they differ in their handling of formatting, idioms, and word order, with Llama 3.1—perhaps for the more discrete nature of its quality assessments—often showing neutrality (manifested through a large number of ties) and CometKiwi demonstrating more varied preferences, some of which are problematic (e.g., preference towards literalness in the translation of idiomatic expressions). Gender-related evaluations suggest potential biases in both systems, mainly due to scoring masculine and feminine gender inflections equally despite only one being correct. When controlling for the source content, we notice more errors for the instances mentioning a feminine referent in specific contexts. These findings indicate that both models display gender-dependent behaviour in source processing, warranting further investigation into potential model biases.

8 Discussion

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

Large language models in Quality Estimation In this edition, we observed an increased use of LLMs, not only in order to generate pseudo-data for training or as a complementary system –which was the trend in the previous year– but rather as the primary model to address a task. Indeed, across tasks, it was possible to observe the performance of encoder-based models that follow the predictor-estimator architecture (Kim et al., 2017), as well as models that relied on large decoder-based approaches, where the emphasis was more on prompt engineering or instruction tuning. This is in line with recent works (Huang et al., 2023; Fernandes et al., 2023a; Kocmi and Federmann, 2023; Vu et al., 2024; Hada et al., 2024) that suggest that multilingual LLMs can be prompted to predict the quality of a translation, given some tuning or in-context learning.

Looking at the results for Tasks 1 and 2, however, we can see that the methods that rely on LLMs are still outperformed by predictor-estimator-based systems, especially when it comes to predicting sentence-level scores. One key disparity, in this case, relates to the fact that methods relying on scores generated by such models lack the granularity of predictor-estimator architectures that treat the QE task as a regression and, hence, can differentiate better between different translations and quality levels. Instead, LLMs tend to default to a smaller range of values (as we can also see in the ties detected in the analysis of Section 7 and Figures 10a and 10b. However, we can see that the LLM-based methods are closing the gap in terms of performance when compared to the predictor-estimator-based model for Task 2, which involves error detection. More importantly, LLM-based approaches perform on par and even outperform other methods for Task 3, which focuses on translation correction (APE). Thus, it seems that in the MT evaluation and correction family of tasks, there is potential for both LLMs and "traditional" neural systems. Potentially, more hybrid methods, i.e. methods that employ sentence-level quality scores predicted from encoder-based models to inform LLM decisions on error detection and correction, would lead to improved performance and could take the lead in future editions for the shared task.

Role of QE signals in APE Both participants in Task 3 used QE information to perform APE in alignment with the task objectives. Their approaches share similarities, as they both involve a final QE-driven selection step to choose between the original MT output and the generated APE hypothesis. One participant (HW-TSC) exploited QE information only for this final selection step, while the other (IT-Unbabel) integrated the two technologies more tightly by generating APE outputs with an LLM informed by free-text explanations for translation errors, which can be considered as proxies for QE predictions. Overall, despite being obtained with different degrees of QE integration, the evaluation results reinforce previous findings regarding the effectiveness of combined QEAPE and approaches for enhancing MT output (Chatterjee et al., 2017; Deoghare et al., 2023).

9 Conclusions

This year's edition of the QE Shared Task introduced two key new elements besides fresh test sets: (1) A new task on QE-informed APE, motivating participants to consider the QE scores to improve the generated MT corrections and (2) an updated challenge set for En-De and En-Es language pairs to help analyse the behaviour and robustness of submitted models for different phenomena such as gender bias, idiomatic expressions, handling of numerical entities, hallucinations, and word order changes.

We found that overall QE performance is consis-

tently high across languages on the sentence level. Still, there is ample room for improvement regarding fine-grained error span detection. The addition of quality informed APE sub-task made it easier for participants to leverage their QE system for the APE task, achieving significant gains for *en-hi* and marginal (non-significant) gains on *en-ta* language pairs. In addition, we found that approaches that employ LLMs still have some way to go in competing on correlations with human scores at the sentence level but can provide competitive solutions for error span detection and QE-informed APE tasks.

In future iterations, we aim to redefine meaningful fine-grained QE tasks, targeting attainable error detection that can help detect critical errors, explain predicted quality, and better inform APE systems. Additionally, we intend to expand further the provided resources to aid the finer grained analysis of model behaviour, as it was discussed in Section 7.

10 Ethical Considerations

Post-editing, MQM, and DA annotations in this paper are carried out by professional translators. They are all paid at professional rates. In creating the gender subset, we drew examples from MT-GenEval (Currey et al., 2022), a corpus where gender is treated as a binary variable. We recognize that gender identities exist on a spectrum, going beyond just the masculine-feminine dichotomy. Our intention is to expand the evaluation of gender-related aspects to include more inclusive forms of machine translation.

Organisers from Unbabel and IT have submitted to this task without using prior access to test sets or 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, by the Portuguese Recovery and Resilience Plan through project C645008882-00000055 (Center for Responsible AI), and by the project Meetween (My Personal AI Mediator for Virtual MEETtings BetWEEN People – grant agreement No 101135798).

We thank the annotation agencies Zibanka Media Services *Pvt. Ltd.* and Techliebe for working with us towards annotating DA data for Indic language pairs. Part of this work was supported by European Association for Machine Translation (EAMT) sponsored Indic Languages QE annotation project at the University of Surrey (UoS/RN580).

References

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

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Duarte M Alves, José Pombal, Nuno M Guerreiro, Pedro H Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, et al. 2024. Tower: An open multilingual large language model for translation-related tasks. *arXiv preprint arXiv:2402.17733*.

Frederic Blain, Chrysoula Zerva, Ricardo Rei, Nuno M. Guerreiro, Diptesh Kanojia, José G. C. de Souza, Beatriz Silva, Tânia Vaz, Yan Jingxuan, Fatemeh Azadi, Constantin Orasan, and André Martins. 2023. Findings of the WMT 2023 shared task on quality estimation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 629–653, Singapore. Association for Computational Linguistics.

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

Rajen Chatterjee, Christian Federmann, Matteo Negri, and Marco Turchi. 2019. Findings of the WMT 2019

shared task on automatic post-editing. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 11–28, Florence, Italy. Association for Computational Linguistics.

Rajen Chatterjee, Markus Freitag, Matteo Negri, and Marco Turchi. 2020. Findings of the WMT 2020 shared task on automatic post-editing. In *Proceedings of the Fifth Conference on Machine Translation*, pages 646–659, Online. Association for Computational Linguistics.

Rajen Chatterjee, Matteo Negri, Raphael Rubino, and Marco Turchi. 2018a. Findings of the WMT 2018 shared task on automatic post-editing. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 710–725, Belgium, Brussels. Association for Computational Linguistics.

Rajen Chatterjee, Matteo Negri, Marco Turchi, Frédéric Blain, and Lucia Specia. 2018b. Combining quality estimation and automatic post-editing to enhance machine translation output. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 26–38.

Rajen Chatterjee, Matteo Negri, Marco Turchi, Marcello Federico, Lucia Specia, and Frédéric Blain. 2017. Guiding neural machine translation decoding with external knowledge. In *Proceedings of the Second Conference on Machine Translation, Volume 1: Research Papers*, Copenhagen, Denmark. Association for Computational Linguistics.

Rajen Chatterjee, Marion Weller, Matteo Negri, and Marco Turchi. 2015. Exploring the planet of the APEs: a comparative study of state-of-the-art methods for MT automatic post-editing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 156–161, Beijing, China. Association for Computational Linguistics.

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*.

Anna Currey, Maria Nadejde, Raghavendra Reddy Pappagari, Mia Mayer, Stanislas Lauly, Xing Niu, Benjamin Hsu, and Georgiana Dinu. 2022. MT-GenEval: A counterfactual and contextual dataset for evaluating gender accuracy in machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4287–4299, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sourabh Deoghare, Diptesh Kanojia, Fred Blain, Tharindu Ranasinghe, and Pushpak Bhattacharyya. 2023. Quality estimation-assisted automatic post-editing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1686–1698, Singapore. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Taisei Enomoto, Hwichan Kim, Tosho Hirasawa, Yoshinari Nagai, Ayako Sato, Kyotaro Nakajima, and Mamoru Komachi. 2024. TMU-HIT at MLSP 2024: How well can GPT-4 tackle multilingual lexical simplification? In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 590–598, Mexico City, Mexico. Association for Computational Linguistics.

Alexander Richard Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. Qafacteval: Improved qa-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André Martins, Graham Neubig, Ankush Garg, Jonathan Clark, Markus Freitag, and Orhan Firat. 2023a. 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*, pages 1066–1083, Singapore. 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. 2023b. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. *arXiv preprint arXiv:2308.07286*.

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.

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.

Nuno M Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André FT Martins. 2023. xcomet: Transparent machine translation evaluation through fine-grained error detection. *arXiv preprint arXiv:2310.10482*.

Rishav Hada, Varun Gumma, Adrian Wynter, Harshita Diddee, Mohamed Ahmed, Monojit Choudhury, Kalika Bali, and Sunayana Sitaram. 2024. Are large language model-based evaluators the solution to scaling up multilingual evaluation? In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1051–1070.

Hui Huang, Shuangzhi Wu, Xinnian Liang, Bing Wang, Yanrui Shi, Peihao Wu, Muyun Yang, and Tiejun Zhao. 2023. Towards making the most of llm for translation quality estimation. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 375–386. Springer.

Raisa Islam and Owana Marzia Moushi. 2024. Gpt-4o: The cutting-edge advancement in multimodal llm. *Authorea Preprints*.

Hyun Kim, Jong-Hyeok Lee, and Seung-Hoon Na. 2017. Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation. In *Proceedings of the Second Conference on Machine Translation*, pages 562–568, Copenhagen, Denmark. Association for Computational Linguistics.

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*, pages 1–42, Singapore. 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*, pages 768–775.

Tom Kocmi, Vilém Zouhar, Christian Federmann, and Matt Post. 2024. Navigating the metrics maze: Reconciling score magnitudes and accuracies.

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).

Christoph Leiter, Juri Opitz, Daniel Deutsch, Yang Gao, Rotem Dror, and Steffen Eger. 2023. The Eval4NLP 2023 shared task on prompting large language models as explainable metrics. In *Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems*, pages 117–138, Bali, Indonesia. Association for Computational Linguistics.

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.

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*, pages 2511–2522, 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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Vikas Raunak, Arul Menezes, Matt Post, and Hany Hassan Awadalla. 2023. Do gpts produce less literal translations?

Ricardo Rei, Nuno M Guerreiro, Daan van Stigt, Marcos Treviso, Luísa Coheur, José GC de Souza, André FT Martins, et al. 2023. Scaling up cometkiwi: Unbabel-ist 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848.

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. 2022. Cometkiwi: Ist-unbabel 2022 submission for the quality estimation shared task. *WMT 2022*, page 634.

Ananya Sai B, Tanay Dixit, Vignesh Nagarajan, Anoop Kunchukuttan, Pratyush Kumar, Mitesh M. Khapra, and Raj Dabre. 2023. IndicMT eval: A dataset to meta-evaluate machine translation metrics for Indian languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14210–14228, Toronto, Canada. Association for Computational Linguistics.

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

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.

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.

Marcos Treviso, Nuno M Guerreiro, Sweta Agrawal, Ricardo Rei, José Pombal, Tania Vaz, Helena Wu, Beatriz Silva, Daan van Stigt, and André FT Martins. 2024. xtower: A multilingual llm for explaining and correcting translation errors. *arXiv preprint arXiv:2406.19482*.

Marco Turchi, Matteo Negri, and Marcello Federico. 2013. Coping with the subjectivity of human judgements in MT quality estimation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 240–251, Sofia, Bulgaria. Association for Computational Linguistics.

Tu Vu, Kalpesh Krishna, Salaheddin Alzubi, Chris Tar, Manaal Faruqui, and Yun-Hsuan Sung. 2024. Foundational autoraters: Taming large language models for better automatic evaluation. *arXiv e-prints*, pages arXiv–2407.

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

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.

Jiawei Yu, Xiaofeng Zhao, Min Zhang, Yanqing Zhao, Yuang Li, Chang Su, Xiaosong Qiao, Miaomiao Ma, and Hao Yang. 2024. Hw-tsc's participation in the wmt 2024 qeape task. *In Proceedings of the Ninth Conference on Machine Translation (WMT24)*.

Emmanouil Zaranis, Giuseppe Attanasio, Sweta Agrawal, and André F. T. Martins. 2024. Watching the watchers: Exposing gender disparities in machine translation quality estimation.

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 and Task 2 Quality Estimation

Hyper-parameter	**T1 Sentence-level** COMETKIWI-DA-22	**T2 Fine-grained** COMETKIWI-MULTITASK-22
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	3e-06 and 1e-05
Batch size	4	4
Loss function	MSE and CE	MSE and CE
Dropout	0.15	0.1
FP precision	32	32
Feed-Forward hidden units	[2048, 1024]	[3072, 1024]
Word weights	[0.3, 0.7]	[0.1, 0.9]
Feed-Forward activation	Tanh	–
Language prefix	False	False

Table 7: Hyper-parameters of both the CometKiwi models used as baselines for Task 1 Quality Estimation.

B Official Results of the WMT24 Quality Estimation Task 1 Sentence-level

Tables 8, 9, 10, 11, 12, 13, 14 and 15 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	0.553	0.438	0.410
BASELINE	0.520	0.474	0.382
Pister Labs	0.452	0.378	0.354

Table 8: Official results of the WMT24 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
BASELINE •	0.514	0.050	0.397	2,260,734,705	569,330,715	1
Pister Labs •	0.513	0.114	0.455	1,400,000,000	70,000,000,000	1
Unbabel •	0.512	0.037	0.393	42,868,104,221	10,716,932,147	6

Table 9: Official results of the WMT24 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 •	0.345	0.116	0.257	42,868,104,221	10,716,932,147	6
BASELINE •	0.340	0.197	0.253	2,260,734,705	569,330,715	1
Pister Labs	0.282	0.104	0.215	1,400,000,000	70,000,000,000	1

Table 10: Official results of the WMT24 Quality Estimation Task 1 Sentence-level for **English-Spanish (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
BASELINE •	0.441	0.223	0.328	2,260,734,705	569,330,715	1
Unbabel	0.412	0.065	0.318	42,868,104,221	10,716,932,147	6
Pister Labs	0.363	0.142	0.300	1,400,000,000	70,000,000,000	1

Table 11: Official results of the WMT24 Quality Estimation Task 1 Sentence-level for **English-Hindi (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
TMU-HIT •	0.739	0.760	0.547	-	-	1
HW-TSC •	0.719	0.783	0.531	2,387,827,161	596,896,035	8
Unbabel	0.714	0.679	0.524	42,868,104,221	10,716,932,147	6
BASELINE	0.678	0.771	0.497	2,260,734,705	569,330,715	1
Pister Labs	0.564	0.536	0.443	1,400,000,000	70,000,000,000	1

Table 12: Official results of the WMT24 Quality Estimation Task 1 Sentence-level for **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
TMU-HIT •	0.713	0.808	0.531	-	-	1
Unbabel •	0.703	0.751	0.514	42,868,104,221	10,716,932,147	6
HW-TSC	0.686	0.757	0.500	2,387,827,161	596,896,035	8
BASELINE	0.661	0.776	0.486	2,260,734,705	569,330,715	1
Pister Labs	0.587	0.716	0.366	1,400,000,000	70,000,000,000	1

Table 13: Official results of the WMT24 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.

Model	**Spearman**	Pearson	Kendall	Disk footprint (B)	# Model params	Ensemble
Unbabel •	0.510	0.719	0.363	42,868,104,221	10,716,932,147	6
HW-TSC •	0.482	0.643	0.340	2,387,827,161	596,896,035	8
TMU-HIT	0.465	0.550	0.329	-	-	1
BASELINE	0.414	0.716	0.294	2,260,734,705	569,330,715	1
Pister Labs	0.379	0.535	0.304	1,400,000,000	70,000,000,000	1

Table 14: Official results of the WMT24 Quality Estimation Task 1 Sentence-level **English-Telugu (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 •	0.683	0.719	0.506	2,387,827,161	596,896,035	8
Unbabel •	0.675	0.702	0.499	42,868,104,221	10,716,932,147	6
TMU-HIT	0.603	0.664	0.445	-	-	1
BASELINE	0.592	0.584	0.419	2,260,734,705	569,330,715	1
Pister Labs	0.478	0.503	0.366	1,400,000,000	70,000,000,000	1

Table 15: Official results of the WMT24 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.

C Official Results of the WMT24 Quality Estimation Task 2 Fine grained Error Detection

Tables 16, 17, 18 and 19 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
BASELINE	0.278	0.220	0.427
HW-TSC	0.227	0.203	0.268

Table 16: Official results of the WMT24 Quality Estimation Task 2 Fine grained Error Detection **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
BASELINE	0.192	0.127	0.394	2,260,743,915	569,309,780	1
HW-TSC	0.178	0.175	0.181	2,409,244,995	2,280,000,000	1

Table 17: Official results of the WMT24 Quality Estimation Task 2 Fine grained Error Detection **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
BASELINE	0.161	0.106	0.337	2,260,743,915	569,309,780	1
HW-TSC	0.151	0.106	0.261	2,409,244,995	2,280,000,000	1

Table 18: Official results of the WMT24 Quality Estimation Task 2 Fine grained Error Detection **English-Spanish (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
BASELINE	0.481	0.428	0.551	2,260,743,915	569,309,780	1
HW-TSC	0.362	0.329	0.401	2,409,244,995	2,280,000,000	1

Table 19: Official results of the WMT24 Quality Estimation Task 2 Fine grained Error Detection **English-Hindi (MQM)**. The winning submission is indicated by a •. Baseline systems are highlighted in grey.

D Official Results of the WMT24 Quality Estimation Task 3 Quality-informed APE

Tables 20 and 21 show the results for all language pairs, ranking participating systems from best to worst using TER as the primary key for each of these cases.

Model	**TER**	BLEU	ChrF	COMET	Disk footprint (B)	# Model params	Ensemble
IT-Unbabel •	27.08	58.38	73.45	0.8646	28,991,029,248	7,000,000,000	1
HW-TSC •	31.32	52.74	69.83	0.8517	1,265,490,783	99,388,416	1
BASELINE	46.36	39.28	59.48	0.8084	-	-	-

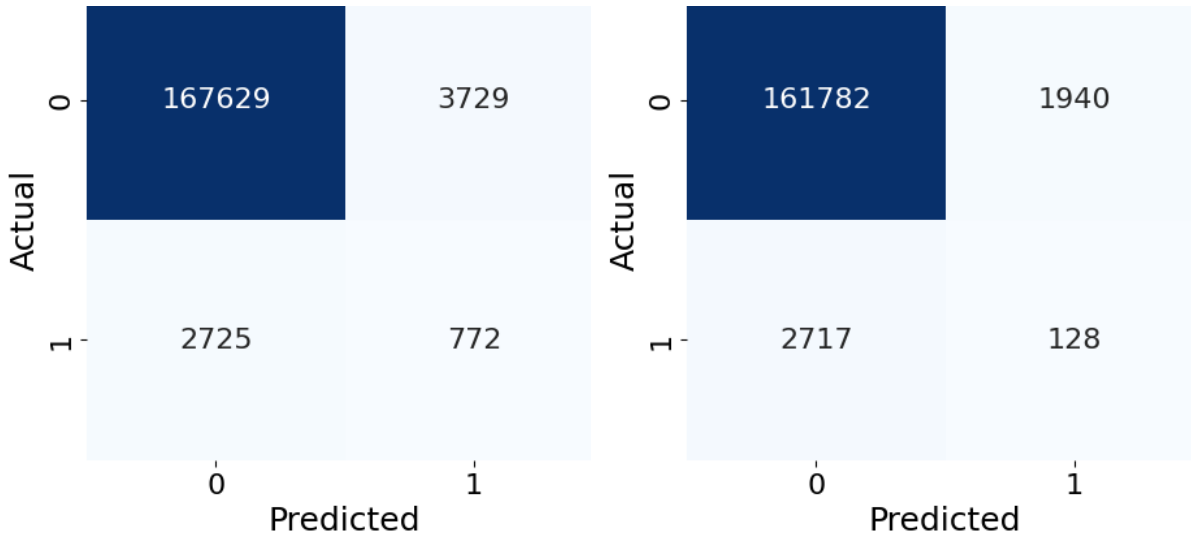
Table 20: Official results of the WMT24 Quality Estimation Task 3 Quality-informed APE **English-Hindi (DA)**. The winning submission is indicated by a •. Baseline systems are highlighted in grey.

Model	**TER**	BLEU	ChrF	COMET	Disk footprint (B)	# Model params	Ensemble
HW-TSC	24.24	69.64	82.36	0.9186	1,265,490,783	99,388,416	1
IT-Unbabel	24.54	70.05	82.30	0.9163	28,991,029,248	7,000,000,000	1
BASELINE	24.71	70.16	81.80	0.9137	-	-	-

Table 21: Official results of the WMT24 Quality Estimation Task 3 Quality-informed APE **English-Tamil (DA)**. The winning submission is indicated by a •. Baseline systems are highlighted in grey.

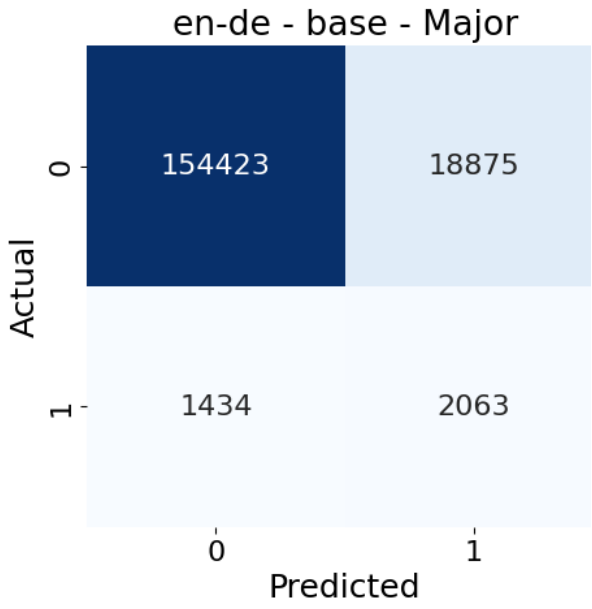
E Confusion Matrices for Task 2

We present below the confusion matrices for Major and Minor error span prediction between HW-TSC and the Baseline, for each language pair. We can see that overall HW-TSC targets precision, being more conservative in error span prediction, while the Baseline model greedily predicts major errors.

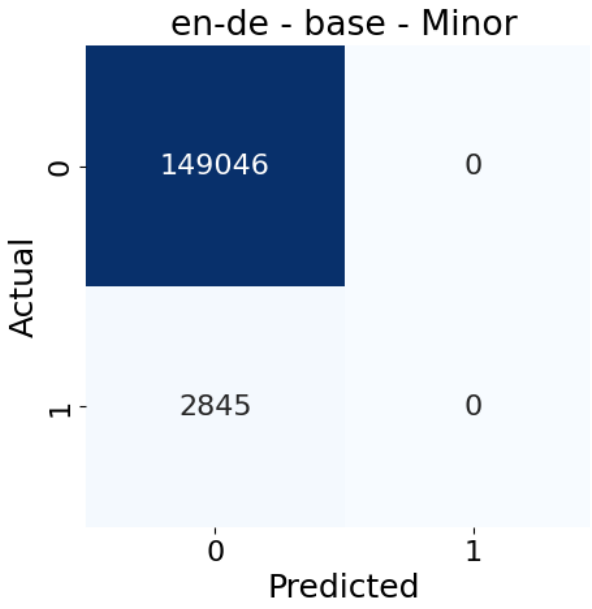


(a) HWTSC - Major

(b) HWTSC -Minor

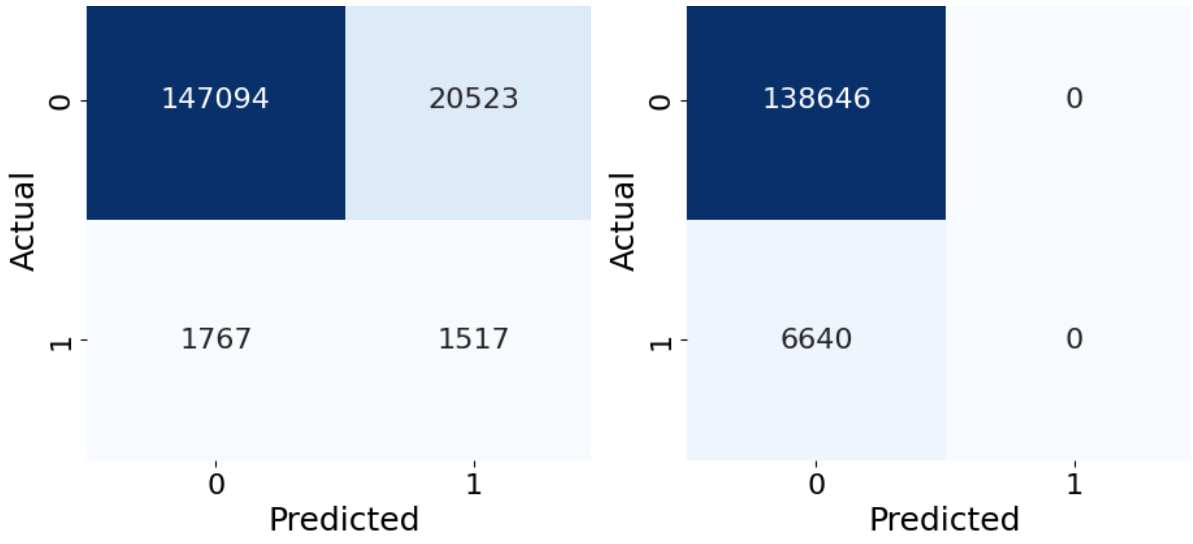


(c) Baseline - Major



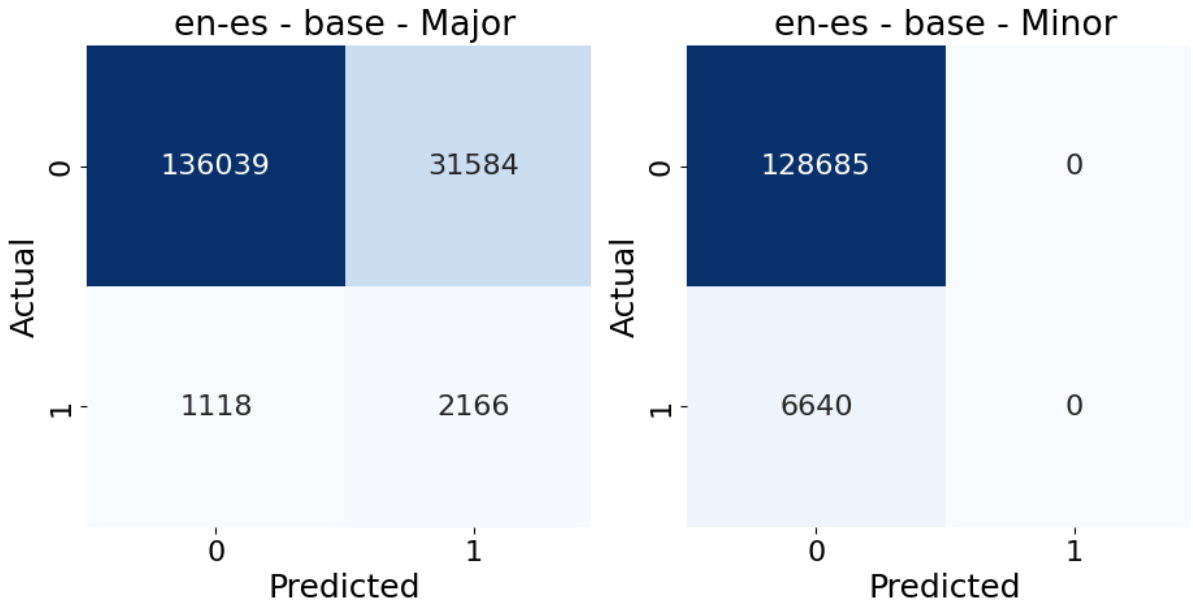
(d) Baseline - Minor

Figure 11: Confusion matrices for Task 2 English-German, comparing Minor and Major predictions between the Baseline system and the HWTSC one.



(a) HWTSC - Major

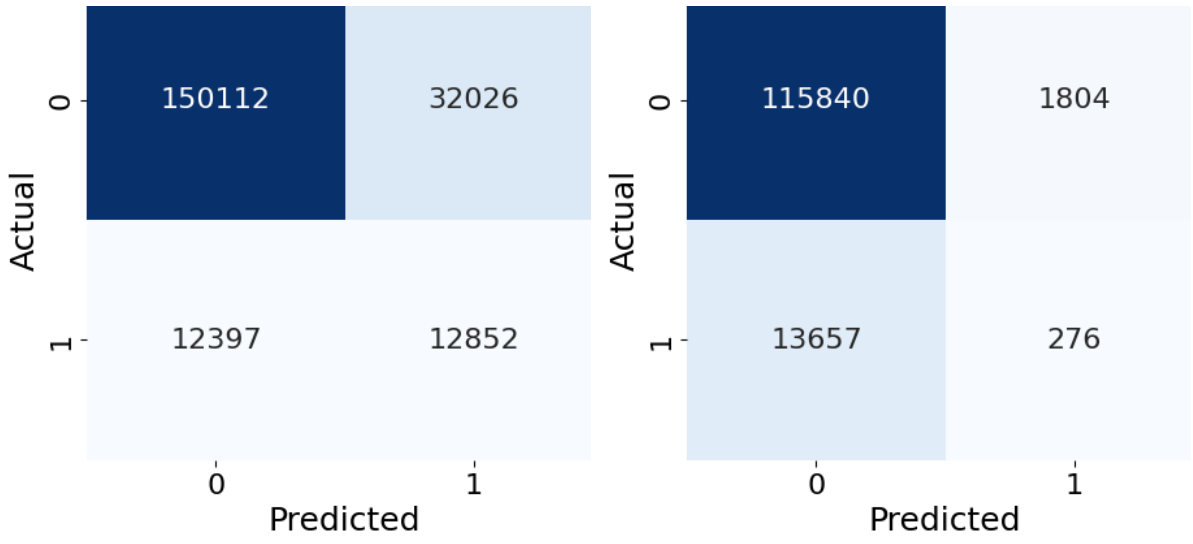
(b) HWTSC -Minor



(c) Baseline - Major

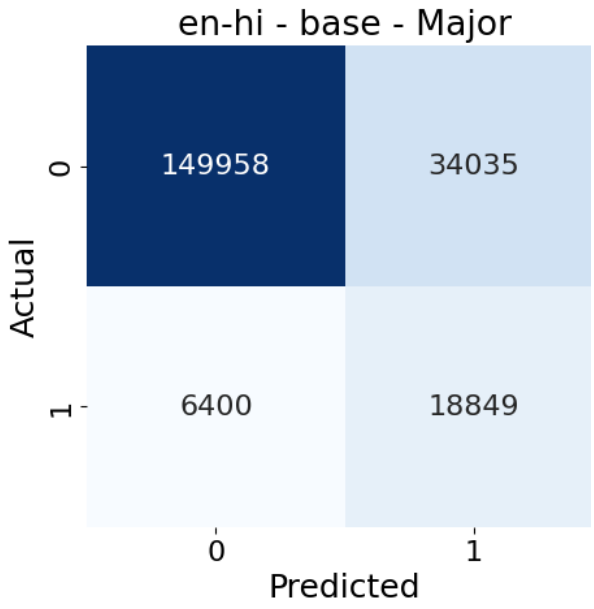
(d) Baseline - Minor

Figure 12: Confusion matrices for Task 2 English-Spanish, comparing Minor and Major predictions between the Baseline system and the HWTSC one.

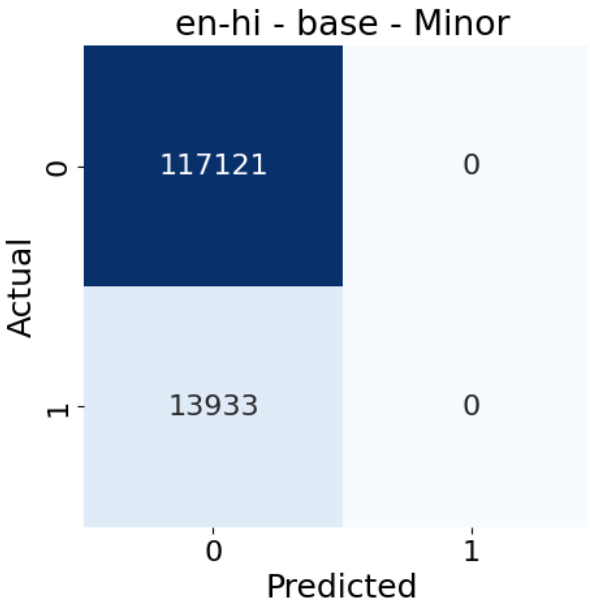


(a) HWTSC - Major

(b) HWTSC -Minor



(c) Baseline - Major



(d) Baseline - Minor

Figure 13: Confusion matrices for Task 2 English-Hindi, comparing Minor and Major predictions between the Baseline system and the HWTSC one.