

# Leveraging Verbalized Confidence in LLM-as-a-Judge for Automated Translation Quality Evaluation

Richard Servajean, Mohammad Golam Sohrab, Lucie Kunitomo-Jacquín, Matiss Rikters

Artificial Intelligence Research Center (AIRC),

National Institute of Advanced Industrial Science and Technology (AIST),

Tokyo 135-0064, Japan

[richard.servajean@aist.go.jp](mailto:richard.servajean@aist.go.jp), [sohrab.mohammad@aist.go.jp](mailto:sohrab.mohammad@aist.go.jp), [kunitomo-jacquín.lucie@aist.go.jp](mailto:kunitomo-jacquín.lucie@aist.go.jp), [matiss.rikters@aist.go.jp](mailto:matiss.rikters@aist.go.jp)

## Abstract

This paper presents the AIST AIRC team’s contributions to subtasks 1 (segment-level error detection and span annotation), 2 (segment-level quality score prediction) and 3 (detection of error-free segments) of the automated translation quality evaluation systems shared task in WMT 2026. Our approach to address each of the 3 subtasks relies on an LLM-as-a-judge system. Specifically, for subtasks 2, we follow an LLM-as-a-judge approach within the Multidimensional Quality Metrics (MQM) error typology, with the novel addition of prompting the large language model (LLM) to verbalize its confidence in each error type assessment. We then train a Multilayer Perceptron (MLP) Regressor on error severities with and without associated confidence ratings to predict human annotations. First, our results provide further evidence for the relevance of integrating LLMs into automated translation quality evaluation pipelines. Second, this work aims to advance ongoing efforts to make these pipelines uncertainty-aware through the integration of verbalized confidence. Our code is freely available at <https://github.com/sshrichard/WMT-Evaluation-Shared-Task-2026>.

## 1 Introduction

In the last few years, artificial intelligence (AI) systems have been applied to numerous fields such as biology (Jumper et al., 2021) and mathematics (Feng et al., 2026). In particular, since the release of GPT-3.5 in 2022 (Schulman et al., 2022), large language models (LLMs) have seen rapid and widespread adoption. In this context, a straightforward method to address any given task consists in directly prompting an LLM to perform it, an approach commonly referred to as LLM-as-a-judge (Gu et al., 2026). The LLM-as-a-judge approach has been followed in many areas, including the field of translation evaluation (Kocmi and Federmann, 2023b,a; Fernandes et al., 2023; Lu et al., 2024,

2025; Sindhuja et al., 2025; Qian et al., 2024). Accordingly, for subtask 1, we simply prompt an LLM to perform segment-level error detection and span annotation. Similarly, within subtasks 1 and 2, we follow previous work, such as Lu et al. (2024), where they implement the Multidimensional Quality Metrics (MQM; Freitag et al., 2021) framework with LLM judges. More precisely, in the current work, we prompt an LLM to assess the severity of different error types drawn from the MQM typology, with the following twist: rather than prompting the LLM solely to assess error severity, we also prompt it to report (i.e., verbalize) its confidence in each error type assessment.

*Verbalized confidence*, also referred to as *verbalized uncertainty* (Groot and Valdenegro-Toro, 2024), is one among a broad range of measures in the field of LLM uncertainty quantification (Shorinwa et al., 2025). It has also become an important focus of research on the metacognitive capabilities of AI systems (Servajean and Servajean, 2026; Liu et al., 2026a). These measures are typically evaluated on their ability to distinguish correct from incorrect responses. Importantly, it is well-established that taking into account uncertainty signals is crucial in decision-making contexts (Li and Steyvers, 2025). As an illustrative example, in a high risk, high uncertainty scenario, suspending judgment and prioritizing further data collection can help avoid committing to a decision with potentially catastrophic consequences. With regard to automatic translation evaluation, the integration of confidence signals into evaluation pipelines has recently attracted growing interest (Glushkova et al., 2021; Zerva et al., 2022). Hence, in the present paper, we aim to advance ongoing efforts to make translation quality evaluation pipelines uncertainty-aware by investigating whether taking into account verbalized confidence make LLM-based translation evaluation better.

## 2 Related Work

**Subtask 1.** Yeom et al. (2025) introduced Tagged Span Annotation (TSA), an annotation scheme designed to extract span-level information by proposing a translation error span detection system that leverages OpenAI o3<sup>1</sup>, a state-of-the-art proprietary-based reasoning LLM, for the WMT’25 MT evaluation span-level error detection task. Building on this idea, we use an LLM with static span tagging schema.

**Subtasks 2 and 3.** Machine translation evaluation is inherently affected by multiple sources of uncertainty. As discussed by Glushkova et al. (2021), uncertainty may arise from noisy human judgments, limited inter-annotator agreement, imperfect or insufficient reference translations, complex translations, and domain mismatch between evaluation data and the model’s training domain. These factors motivate the development of uncertainty-aware evaluation methods (Beck et al., 2016; Fomicheva et al., 2020; Zerva et al., 2022; Zerva and Martins, 2024; Marashian et al., 2026; Wu and Monz, 2025).

Notably, for the WMT’25 MT evaluation task, Wu and Monz (2025) leveraged the average log-likelihood assigned by an LLM as a measure of translation quality, and incorporated a calibration step in their pipeline that optimizes the Pearson correlation between this likelihood and an external quality metric to further improve its correlation with human annotations. They reported encouraging performance, sometimes exceeding that of some state-of-the-art systems.

While we focus on evaluating full translations, Marashian et al. (2026) focused on various *per-word* and *per-token* confidence measures. They especially compared methods based on verbalized confidence to methods based on internal confidence (token probabilities). Verbalized and internal methods performed comparably when evaluated on calibration and error detection. Yet, they found little to no correlation between verbalized and internal measures.

In this paper, we aim to leverage the capabilities of LLMs to explicitly assess and verbalize their own confidence.

<sup>1</sup><https://openai.com/index/introducing-o3-and-o4-mini/> (accessed 11 September 2026)

## 3 Method

### 3.1 LLM for Segment-level Error Detection and Span Annotation

We tackle the segment-level error detection and span annotation (subtask 1) by employing Qwen/Qwen3-32B<sup>2</sup>, leveraging prompt-based in-context learning for inference. The model is loaded from the Hugging Face hub. Due to time constraints to address this subtask, we mainly focus on prompt-based approaches that exploit in-context learning by providing instructions and examples rather than relying on task-specific training.

**Prompt Engineering for Error Span Detection.** Prompt engineering has become one of the most widely adopted approaches for adapting general-purpose LLMs to many domain specific tasks including text summarization, peer-review generation, and code generation. Notably, this approach does not require modifying model parameters. In a few-shot setting for translation error span detection, models are directly instructed to identify translation errors using a prompt that includes several annotated examples of error spans to guide the structure of the output.

For a given source and target languages, the task uses an error-span annotation scheme where each error span is wrapped in a static tag `<ES>.</ES>` that allows multiple errors to be annotated simultaneously within the same sentence. Therefore, each error span consists of start-token marker and an end-token marker.

For a desired JSON response format, the target schema is illustrated through few-shot examples. An example of the instructions and corresponding JSON output is provided below:

#### Instruction

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

#### EVALUATION GUIDELINES:

- Be thorough but precise: Only mark errors when you are confident they are incorrect
- Consider context and domain: Some variations may be acceptable depending on context
- Distinguish between errors and acceptable alternatives: Multiple valid translations may exist
- Focus on clear, objective errors rather than subjective preferences
- Verify each potential error against the source text before marking
- When in doubt, err on the side of not marking an error

<sup>2</sup><https://huggingface.co/Qwen/Qwen3-32B> (accessed 11 September 2026)

**Error Categories:**

- Accuracy: addition, mistranslation, omission, untranslated text
- Fluency: character encoding, grammar, inconsistency, punctuation, register, spelling
- Style: awkward phrasing
- Terminology: inappropriate for context, inconsistent use
- Other: non-translation, other issues

**Severity Classification:**

- Major: Errors that impact meaning or usability but do not render the text unusable
- Minor: Errors that do not impact meaning or usability

**Omission Classification:**

- Null: No Errors detected in the annotated text
- Major: Errors that appears in the error category as accuracy error like mistranslation, omission, untranslated text
- Minor: Errors that do not impact meaning or usability

**CRITICAL OUTPUT REQUIREMENTS:**

- Mark errors only when you have clear evidence they are incorrect
- Consider whether alternative translations could be equally valid
- Apply strict standards: better to miss a minor error than create a false positive
- Wrap error spans in the translation with tags: `<ES></ES>`
- NO comments, explanations, or additional text
- \*\*Mark only the minimal substring that contains the clear error; do not include extra context\*\*

**English**

Talks have resumed in Vienna to try to revive the nuclear pact, with both sides trying to gauge the prospects of success after the latest exchanges in the stop-start negotiations.

**Czech**

Ve Vídni se ve Vídni obnovily rozhovory o oživení jaderného paktu, přičemž obě partaje se snaží posoudit vyhlídky na úspěch po posledních výměnách v jednáních.

**Output**

```
{
  "annotated_translation": "Ve Vídni se <ES>ve Vídni</ES> obnovily rozhovory o oživení jaderného paktu, přičemž obě <ES>partaje</ES> se snaží posoudit vyhlídky na úspěch po posledních výměnách v<ES></ES> jednáních.",
  "errors": [
    {
      "severity": "major",
      "category": "accuracy/addition"
    },
    {
      "severity": "minor",
      "category": "terminology/inappropriate for context"
    },
    {
      "severity": "major",
      "category": "accuracy/omission"
    }
  ],
  "omission": "major"
}
```

## 3.2 Pipeline for Segment-Level Quality Score Prediction and Detection of Error-Free Segments

Here, we introduce our pipeline for subtask 2 (segment-level quality score prediction) and 3 (detection of error-free segments), in which an LLM is asked to assess the severity of different error types and to verbalize its confidence.

### 3.2.1 Verbalized Confidence in LLM-as-a-Judge

For subtask 2 (segment-level quality score prediction), given a source and a translation, we first assess whether the translation target language coincides indeed with that of the translation using GlotLID model (Kargaran et al., 2023)<sup>3</sup>.

If the assessed translation language is the correct one, we prompt an LLM (we tried *Qwen3.6-27B*<sup>4</sup> [with the recommended temperature of 1.0 and without reasoning], *Qwen3.8-27B-FP8*<sup>5</sup> [with a temperature of 0.7 and without reasoning; it was not included in our submission], and *Babel-83B-chat*<sup>6</sup> [with a temperature of 0.0]) with the source, the translation, their languages, seven error type definitions corresponding to the “seven high-level error type dimensions”<sup>7</sup> within MQM error typology (Lommel et al., 2024), and ask it to assess the severity of each error type by choosing among Not applicable, Neutral, Minor, Major, and Critical. Finally, the model is instructed (within the same prompt) to rate its confidence in each error type assessment (a confidence scale is provided [e.g., 0.6-0.8: “Highly likely”]; see the prompt available on GitHub for further details).

Based on the LLM output, we construct a  $7 \times 4 + 1 = 29$ -long feature vector, consisting of seven 4-dimensional blocks (one per error type) plus an additional entry indicating whether a language mismatch was detected between the translation and the target language (0 or 1). Each of the seven 4-dimensional block encodes:

<sup>3</sup><https://huggingface.co/cis-lmu/glotlid> (accessed 8 September 2026)

<sup>4</sup><https://huggingface.co/Qwen/Qwen3.6-27B> (accessed 8 September 2026)

<sup>5</sup><https://huggingface.co/Qwen/Qwen3.8-27B-FP8> (accessed 8 September 2026)

<sup>6</sup><https://huggingface.co/Tower-Babel/Babel-83B-Chat> (accessed 8 September 2026)

<sup>7</sup>For definitions of the seven high-level MQM error dimensions, see <https://themqm.org/the-mqm-typology/> (accessed 8 September 2026).

| Subtask | ID     | CS-DE  | CS-UK  | CS-VI  | ZH-JA  | EN-AR  | EN-HY  | EN-BE  | EN-ZH  | EN-ZH (TW) | EN-CS  | EN-ET  | EN-DE  | EN-IS  | EN-ID  | EN-KK  | EN-IT (Ladin) | EN-IT (Ligurian) | EN-RU  | EN-SE  | EN-TH  | EN-UK  |
|---------|--------|--------|--------|--------|--------|--------|--------|--------|--------|------------|--------|--------|--------|--------|--------|--------|---------------|------------------|--------|--------|--------|--------|
| 1       | 875417 | 0.2161 | n/a    | 0.2166 | 0.2287 | 0.1861 | n/a    | n/a    | n/a    | n/a        | n/a    | 0.2075 | n/a    | 0.1881 | 0.2414 | n/a    | n/a           | n/a              | n/a    | n/a    | n/a    | n/a    |
| 1       | 875796 | 0.2155 | n/a    | 0.2161 | 0.2283 | 0.1853 | n/a    | n/a    | n/a    | n/a        | n/a    | 0.2066 | n/a    | 0.1878 | 0.2412 | n/a    | n/a           | n/a              | n/a    | n/a    | n/a    | n/a    |
| 2       | 871400 | 0.6068 | 0.7376 | 0.7085 | 0.5960 | 0.3385 | 0.8471 | 0.7100 | 0.4322 | 0.3716     | 0.6536 | 0.6703 | 0.7123 | 0.7544 | 0.4260 | 0.7545 | 0.2791        | 0.5306           | 0.7556 | 0.5970 | 0.5184 | 0.7283 |
| 2       | 871401 | 0.6005 | 0.7337 | 0.6909 | 0.5683 | 0.333  | 0.848  | 0.7166 | 0.4260 | 0.359      | 0.6455 | 0.6964 | 0.7077 | 0.7355 | 0.3987 | 0.7531 | 0.2735        | 0.5363           | 0.7573 | 0.5987 | 0.5717 | 0.7371 |
| 3       | 871721 | 0.5013 | 0.5603 | 0.5041 | 0.4300 | 0.3424 | 0.4805 | 0.4785 | 0.6233 | 0.4101     | 0.5988 | 0.4676 | 0.5691 | 0.4758 | 0.5438 | 0.5064 | 0.1806        | 0.3345           | 0.6355 |        | 0.0467 | 0.6256 |
| 3       | 874374 | 0.3725 | 0.3466 | 0.3979 | 0.3679 | 0.3983 | 0.2338 | 0.4022 | 0.4666 | 0.4447     | 0.4659 | 0.2819 | 0.4681 | 0.2153 | 0.3891 | 0.3918 | 0.1715        | 0.2572           | 0.5056 |        | 0.4875 | 0.4659 |

Table 1: Results for all three subtasks available at Codabench<sup>8</sup>, where our predictions are compared to the pseudo-gold-standard scores (no further details given on the official website<sup>9</sup>). The metrics for subtasks 1, 2, and 3 are the F-score (Perrella et al., 2026), the correlation (organizers did not specify which), and the Matthews correlation coefficient, respectively. The submission IDs are given in the second column. For subtask 1, submissions 875417 and 875796 correspond, respectively, to the settings in which omission severity was either assessed directly by the model or determined using a simple rule-based approach that checks the different errors associated with omissions (see the example in the Method section). Within subtask 2, submissions 871400 and 871401 correspond to the results obtained with *Qwen3.6-27B*, including and without including confidence ratings, respectively. Within subtask 3, submissions 871721 and 874374 correspond to the results obtained with *Qwen3.6-27B* and *Babel-83B-chat*, respectively. Empty entries (see the EN-SE column) mean that this language pair was excluded, while n/a means that we did not address this language pair.

1. Whether the considered error type is applicable or not (0 or 1),
2. Its severity taking value 0 (Neutral), 1 (Minor), 5 (Major) or 25 (Critical), where we followed standard MQM scoring model,
3. The model’s confidence (float between 0 and 1), and
4. Whether a format issue occurred (0 if no issue occurred, 1 as soon as a parsing issue occurred in a single value or in the entire output). Usually, no parsing issue occurs, but this may depend on the model and on whether a strict output schema is enforced.

By default, all values are set to 0. For example, in case of an error type not applicable, the corresponding severity entry is set to 0.

Then, we stacked all vectors corresponding to all translations from the file `wmt25-genmt-humeval.jsonl`<sup>10</sup>, and trained a Multilayer Perceptron (MLP) Regressor to predict the corresponding human annotations. We considered an MLP with two hidden layers (100 and 50 ReLU units), with scikit-learn’s default settings, for up to 5000 epochs with the default convergence tolerance.

<sup>8</sup><https://www.codabench.org/profiles/user/metricstask/>

<sup>9</sup><https://www2.statmt.org/wmt26/> (accessed 10 September 2026)

<sup>10</sup>Available at <https://github.com/wmt-conference/wmt25-general-mt/tree/main> (accessed 11 September 2026)

### 3.2.2 LLM-as-a-Judge for Error-Free Translation

For subtask 3 (detection of error-free segments), a segment is considered error-free if, in subtask 2, the LLM judge considered that all error types were either not applicable or had a neutral severity (see Section 3.2.1). By default, parsing issues cause the translation to be classified as *not* error-free.

## 4 Experiments

Our implementation is based on Hugging Face transformers. During the model implementation, we use the publicly available previous year WMT ESA dataset; where the models are initialized from the publicly available checkpoints that we mentioned in the Method section. For model inference, a local machine with four NVIDIA RTX 6000 Ada GPUs with 48GB GDDR6 memory is used along with a cluster node with eight NVIDIA H200 GPUs with 141GB of HBM3e memory. To address segment-level error detection and span annotation for the official test set, we only considered eight prioritized language pairs due to time constraints. The running time for each language pair were around 30 hours where we disable the chain-of-thought thinking to non-thinking mode. Besides, we set `max_new_tokens` to 1024, which controls the maximum number of tokens generated by the model.

## 5 Results

The results obtained on the initial submission in which our predictions were compared against

pseudo-gold-standard scores are given in Table 1. For subtask 2, the correlation between human annotations and our predictions are given in Table 2. When considering the overall correlations, we can see that the impact of including confidence ratings in the regression is unclear; extremely small differences in correlation (of the order of  $10^{-2}$  or  $10^{-3}$ ) end up statistically significant (the 95% confidence interval of the paired bootstrap distribution with  $n = 1,000$  of the differences in correlation excludes zero) in both directions. For a similar analysis, note that we also considered all translations from 1,000 source segments randomly drawn from the 2025 data (see the Method section) regardless of the source and translation languages, with an MLP regressor trained on *Qwen3.6-27B* assessments of the translations from a disjoint set of 1,000 source segments. With confidence ratings included in the regression, Pearson, Spearman and Kendall correlations with human scores were 0.5501, 0.4791 and 0.3459 respectively; without them, 0.5348, 0.4722 and 0.3405 (all  $p < 0.001$ ). A paired bootstrap ( $n = 10,000$ ) demonstrates that the improvement when including confidence ratings is significant for all three coefficients. However, the improvement is very small. Finally, for comparison, note that CometKiwi (Rei et al., 2022)<sup>11</sup> evaluated on the same translations yielded Pearson, Spearman, and Kendall correlations of 0.4347, 0.3308, and 0.2280, respectively (all  $p < 0.001$ ).

Even though Table 1 and Table 2 cannot be perfectly compared as the considered translations differ (many translations from `mteval-test26.jsonl.gz` are not in `annotations_preliminary.json`; see captions for further details), we still notice significant differences. A possible explanation is that there is a varying number of human annotations per translation in `annotations_preliminary.json`, so that the presence of segments with very few annotations or even a single annotation may add noise that would degrade the correlation with our predictions, compared to the case where they are evaluated against pseudo-gold-standard scores. Similarly important differences can be noticed between Table 1 and Table 3 for subtask 3. A striking difference concerns the language pairs EN-IT (Ladin) and EN-IT (Ligurian), which were not included in the training set (i.e., last year data;

see the Methods); while the Matthew correlation coefficients are near zero when our predictions are evaluated against human annotations, they remain positive when evaluated against the pseudo-gold-standard scores, although the correlations are still very moderate.

## 6 Discussion

In this paper presenting our contribution to subtasks 1 (segment-level error detection and span annotation), 2 (segment-level quality score prediction) and 3 (detection of error-free segments) of the automated translation quality evaluation systems shared task in WMT 2026, we considered an LLM-as-a-judge approach.

For subtask 1, we adopted a prompt-based in-context learning approach using few-shot examples rather than relying on task-specific training. We believe that task-specific training across all language pairs using data from previous years could further improve segment-level error detection and span annotation performance. In addition, to reduce generation time, we disabled chain-of-thought reasoning and used the model in non-thinking mode. Future work could therefore explore combining task-specific training with chain-of-thought prompting at inference time to further improve performance on subtask 1.

For subtasks 2 and 3, we followed the MQM framework, with the novel addition of prompting the LLM to verbalize its confidence in each error type assessment. More specifically, we trained a Multilayer Perceptron Regressor on error severities with and without associated confidence ratings, generated by an LLM, to predict human annotations. We find no clear evidence that including confidence ratings improves the predictions. There exist many different measures of uncertainty (Shorinwa et al., 2025). Future work could integrate alternative uncertainty measures into our pipeline and compare their effects on performance. Additionally, other regression methods could be explored, as could explicit strategies for combining error type severities with their associated confidence scores, such as a weighted sum.

An important limitation is that last year annotations, on which we performed the MLP regressions, were obtained within the ESA protocol (Kocmi et al., 2024), while 2026 data uses cESA protocol (Zouhar et al., 2026). In the latter case, annotators have access to multiple translations of the

<sup>11</sup><https://huggingface.co/Unbabel/wmt22-cometkiwi-da> (accessed 11 September 2026)

| Language pair    | Qwen3.8-27B |        |        |               |        |        | Qwen3.6-27B |        |        |               |        |        | Babel-83B-chat |        |        |               |        |        |
|------------------|-------------|--------|--------|---------------|--------|--------|-------------|--------|--------|---------------|--------|--------|----------------|--------|--------|---------------|--------|--------|
|                  | With conf.  |        |        | Without conf. |        |        | With conf.  |        |        | Without conf. |        |        | With conf.     |        |        | Without conf. |        |        |
|                  | $r$         | $\rho$ | $\tau$ | $r$           | $\rho$ | $\tau$ | $r$         | $\rho$ | $\tau$ | $r$           | $\rho$ | $\tau$ | $r$            | $\rho$ | $\tau$ | $r$           | $\rho$ | $\tau$ |
| CS-DE            | 0.429       | 0.364  | 0.261  | 0.435         | 0.360  | 0.262  | 0.365       | 0.314  | 0.225  | 0.356         | 0.317  | 0.227  | 0.303          | 0.198  | 0.160  | 0.259         | 0.100  | 0.083  |
| CS-UK            | 0.523       | 0.386  | 0.276  | 0.551         | 0.385  | 0.281  | 0.525       | 0.372  | 0.267  | 0.527         | 0.387  | 0.280  | 0.422          | 0.239  | 0.192  | 0.414         | 0.202  | 0.167  |
| CS-VI            | 0.497       | 0.454  | 0.329  | 0.516         | 0.465  | 0.337  | 0.494       | 0.410  | 0.296  | 0.498         | 0.422  | 0.305  | 0.399          | 0.257  | 0.203  | 0.413         | 0.221  | 0.178  |
| ZH-JA            | 0.481       | 0.463  | 0.339  | 0.491         | 0.484  | 0.358  | 0.463       | 0.462  | 0.339  | 0.474         | 0.468  | 0.345  | 0.361          | 0.258  | 0.206  | 0.365         | 0.252  | 0.205  |
| EN-AR            | 0.191       | 0.187  | 0.149  | 0.193         | 0.193  | 0.155  | 0.177       | 0.186  | 0.148  | 0.184         | 0.199  | 0.159  | 0.182          | 0.186  | 0.153  | 0.173         | 0.170  | 0.144  |
| EN-HY            | 0.599       | 0.529  | 0.384  | 0.613         | 0.544  | 0.397  | 0.604       | 0.554  | 0.404  | 0.622         | 0.583  | 0.428  | 0.479          | 0.365  | 0.286  | 0.480         | 0.385  | 0.310  |
| EN-BE            | 0.536       | 0.533  | 0.398  | 0.537         | 0.528  | 0.398  | 0.519       | 0.530  | 0.397  | 0.526         | 0.533  | 0.402  | 0.350          | 0.338  | 0.269  | 0.330         | 0.330  | 0.269  |
| EN-ZH            | 0.192       | 0.260  | 0.191  | 0.177         | 0.253  | 0.187  | 0.197       | 0.317  | 0.234  | 0.201         | 0.319  | 0.238  | 0.097          | 0.146  | 0.116  | 0.076         | 0.102  | 0.083  |
| EN-ZH (Taiwan)   | 0.162       | 0.270  | 0.192  | 0.166         | 0.293  | 0.213  | 0.155       | 0.283  | 0.205  | 0.163         | 0.291  | 0.211  | 0.097          | 0.191  | 0.151  | 0.070         | 0.122  | 0.099  |
| EN-CS            | 0.432       | 0.462  | 0.339  | 0.424         | 0.448  | 0.332  | 0.420       | 0.472  | 0.345  | 0.432         | 0.482  | 0.355  | 0.299          | 0.270  | 0.215  | 0.278         | 0.251  | 0.204  |
| EN-ET            | 0.479       | 0.459  | 0.339  | 0.502         | 0.474  | 0.352  | 0.488       | 0.476  | 0.350  | 0.486         | 0.484  | 0.358  | 0.337          | 0.287  | 0.225  | 0.317         | 0.256  | 0.205  |
| EN-DE            | 0.432       | 0.483  | 0.354  | 0.410         | 0.466  | 0.346  | 0.369       | 0.433  | 0.315  | 0.389         | 0.451  | 0.332  | 0.273          | 0.281  | 0.229  | 0.246         | 0.260  | 0.215  |
| EN-IS            | 0.540       | 0.491  | 0.361  | 0.554         | 0.517  | 0.382  | 0.580       | 0.550  | 0.407  | 0.601         | 0.565  | 0.420  | 0.379          | 0.322  | 0.257  | 0.372         | 0.328  | 0.269  |
| EN-ID            | 0.361       | 0.371  | 0.267  | 0.357         | 0.367  | 0.267  | 0.322       | 0.356  | 0.255  | 0.339         | 0.370  | 0.268  | 0.292          | 0.309  | 0.243  | 0.264         | 0.281  | 0.226  |
| EN-KK            | 0.561       | 0.483  | 0.352  | 0.574         | 0.492  | 0.363  | 0.549       | 0.498  | 0.364  | 0.573         | 0.519  | 0.382  | 0.427          | 0.308  | 0.238  | 0.422         | 0.290  | 0.231  |
| EN-IT (Ladin)    | 0.137       | 0.168  | 0.133  | 0.141         | 0.175  | 0.139  | 0.127       | 0.181  | 0.142  | 0.137         | 0.184  | 0.145  | 0.142          | 0.182  | 0.145  | 0.139         | 0.182  | 0.146  |
| EN-IT (Ligurian) | 0.239       | 0.284  | 0.206  | 0.245         | 0.293  | 0.214  | 0.228       | 0.281  | 0.203  | 0.226         | 0.274  | 0.201  | 0.254          | 0.287  | 0.221  | 0.245         | 0.285  | 0.227  |
| EN-RU            | 0.565       | 0.514  | 0.377  | 0.570         | 0.501  | 0.371  | 0.540       | 0.495  | 0.361  | 0.554         | 0.509  | 0.374  | 0.480          | 0.363  | 0.292  | 0.448         | 0.355  | 0.289  |
| EN-SE            | 0.749       | 0.710  | 0.531  | 0.757         | 0.711  | 0.527  | 0.669       | 0.692  | 0.511  | 0.696         | 0.675  | 0.507  | 0.488          | 0.498  | 0.375  | 0.509         | 0.513  | 0.396  |
| EN-TH            | 0.425       | 0.384  | 0.278  | 0.506         | 0.437  | 0.320  | 0.450       | 0.381  | 0.277  | 0.411         | 0.382  | 0.281  | 0.520          | 0.463  | 0.380  | 0.475         | 0.418  | 0.346  |
| EN-UK            | 0.561       | 0.483  | 0.351  | 0.570         | 0.472  | 0.348  | 0.569       | 0.494  | 0.360  | 0.572         | 0.497  | 0.364  | 0.480          | 0.348  | 0.277  | 0.464         | 0.342  | 0.276  |
| All              | 0.420       | 0.414  | 0.301  | 0.424         | 0.410  | 0.300  | 0.403       | 0.394  | 0.286  | 0.408         | 0.406  | 0.296  | 0.372          | 0.361  | 0.282  | 0.353         | 0.354  | 0.283  |

Table 2: **Subtask 2.** Segment-level correlation between the predictions from each model and the human scores obtained within the cESA protocol (Zouhar et al., 2026), per language pair and overall (last row). When multiple human annotations are assigned to a given translation, we take the average.  $r$  is Pearson,  $\rho$  Spearman and  $\tau$  Kendall’s tau. All  $p < 0.001$ . The cESA scores come from the released file annotations\_preliminary.json<sup>12</sup>. We only consider translation segments that are concerned by the predictions from our submission, i.e. those from mteval-test26.jsonl.gz<sup>13</sup>.

same document at the same time, facilitating “more consistent and efficient judgments” (Zouhar et al., 2026). This may introduce differences between the training and testing targets.

Beyond confidence ratings, additional information could have been incorporated as regression features, such as segment length or the source and target languages (encoded as categorical variables) as there exists cross-lingual differences in human annotations (Magdy et al., 2026; Liu et al., 2026b).

Also, note that we only considered evaluating models that are publicly available at no cost or free trial to help with reproducibility, such as *Qwen3-32B*, *Qwen3.6-27B*, *Qwen3.8-27B-FP8*, and *Babel-83B-chat*. The choice of the model and parameters like the reasoning budget when applicable are likely to have a significant impact on performance.

Finally, it is worth noting that incorporating AI confidence into pipelines is becoming an increasingly prominent topic across research communities in AI.

<sup>12</sup>Available at <https://github.com/wmt-conference/wmt26-general-mt/releases/tag/humeval> (accessed 9 September 2026)

<sup>13</sup>Available at <https://github.com/wmt-conference/wmt26-mteval> (accessed 11 September 2026)

## Acknowledgments

This paper is based on results obtained from AIST policy-based budget project “R&D on Generative AI Foundation Models for the Physical Domain,” and a project, JPNP25006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

## Limitations

In this work, we only considered evaluating open-weight models and open-access systems that are publicly available at no cost or free trial to help with reproducibility. Our code is freely available on GitHub under a permissive license at <https://github.com/sshrichard/WMT-Evaluation-Shared-Task-2026>.

## Ethics Statement

Our work fully complies with the ACL Code of Ethics<sup>14</sup>. We use only publicly available models/systems and relatively low compute amounts while conducting our experiments to enable reproducibility. We do not perform any studies on other

<sup>14</sup><https://www.aclweb.org/portal/content/acl-code-ethics> (accessed 11 September 2026)

| Language pair    | MCC          |              |                |
|------------------|--------------|--------------|----------------|
|                  | Qwen3.8-27B  | Qwen3.6-27B  | Babel-83B-chat |
| CS-DE            | <b>0.213</b> | 0.121        | 0.106          |
| CS-UK            | <b>0.232</b> | 0.205        | 0.150          |
| CS-VI            | <b>0.192</b> | 0.139        | 0.155          |
| ZH-JA            | <b>0.258</b> | 0.191        | 0.177          |
| EN-AR            | <b>0.138</b> | <b>0.138</b> | 0.132          |
| EN-HY            | <b>0.235</b> | 0.185        | 0.124          |
| EN-BE            | <b>0.354</b> | 0.318        | 0.172          |
| EN-ZH            | <b>0.224</b> | 0.222        | 0.024          |
| EN-ZH (Taiwan)   | <b>0.263</b> | 0.163        | 0.084          |
| EN-CS            | <b>0.339</b> | 0.329        | 0.173          |
| EN-ET            | <b>0.312</b> | 0.302        | 0.136          |
| EN-DE            | <b>0.353</b> | 0.238        | 0.204          |
| EN-IS            | <b>0.334</b> | 0.291        | 0.168          |
| EN-ID            | <b>0.285</b> | 0.220        | 0.220          |
| EN-KK            | <b>0.289</b> | 0.260        | 0.175          |
| EN-IT (Ladin)    | <b>0.042</b> | -0.004       | 0.005          |
| EN-IT (Ligurian) | 0.113        | 0.095        | <b>0.125</b>   |
| EN-RU            | <b>0.337</b> | 0.303        | 0.212          |
| EN-SE            | <b>0.272</b> | 0.043        | 0.114          |
| EN-TH            | 0.153        | 0.132        | <b>0.203</b>   |
| EN-UK            | <b>0.386</b> | 0.328        | 0.260          |
| All              | <b>0.256</b> | 0.221        | 0.173          |

Table 3: **Subtask 3.** Matthews correlation coefficient between the error-free predictions from each model and the human labels, per language pair and overall (last row). A translation is labeled error-free when the human annotation marks no error spans; when multiple human annotations are assigned to a given translation, we take the majority, ties counting as containing errors. Chi-square tests of independence lead to  $p < 0.001$  in all cases, with the following exceptions: for *Babel-83B-chat*, EN-ZH ( $p = 2.50 \times 10^{-1}$ ), EN-IT (Ladin) ( $p = 7.28 \times 10^{-1}$ ), and EN-SE ( $p = 2.09 \times 10^{-3}$ ); for *Qwen3.6-27B*, EN-IT (Ladin) ( $p = 7.50 \times 10^{-1}$ ) and EN-SE ( $p = 2.49 \times 10^{-1}$ ); and for *Qwen3.8-27B-FP8*, EN-IT (Ladin) ( $p = 2.42 \times 10^{-3}$ ). Human annotations come from the released file `annotations_preliminary.json`<sup>15</sup>. We only consider translation segments that are concerned by the predictions from our submission, i.e. those from `mteval-test26.jsonl.gz`<sup>16</sup>.

humans or animals in this research. We use privacy-preserving, locally hosted AI assistants for help with coding data processing scripts necessary for our experiments and evaluation.

## References

- Daniel Beck, Lucia Specia, and Trevor Cohn. 2016. Exploring prediction uncertainty in machine translation quality estimation. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 208–218.
- Tony Feng, Trieu Trinh, Garrett Bingham, Jiwon Kang,<sup>15</sup> Available at <https://github.com/wmt-conference/wmt26-general-mt/releases/tag/humeval> (accessed 9 September 2026)
- <sup>16</sup> Available at <https://github.com/wmt-conference/wmt26-mteval> (accessed 11 September 2026)
- Shengtong Zhang, Sang hyun Kim, Kevin Barreto, Carl Schildkraut, Junehyuk Jung, Jaehyeon Seo, Carlo Pagano, Yuri Chervonyi, Dawsen Hwang, Kaiying Hou, Sergei Gukov, Cheng-Chiang Tsai, Hyunwoo Choi, Youngbeom Jin, Wei-Yuan Li, and 5 others. 2026. Semi-autonomous mathematics discovery with gemini: A case study on the erdős problems. *arXiv preprint arXiv:2601.22401*.
- Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André FT Martins, Graham Neubig, Ankush Garg, Jonathan H Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1066–1083.
- Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555.
- Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.
- Taisiya Glushkova, Chrysoula Zerva, Ricardo Rei, and André FT Martins. 2021. Uncertainty-aware machine translation evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3920–3938.
- Tobias Groot and Matias Valdenegro-Toro. 2024. Overconfidence is key: Verbalized uncertainty evaluation in large language and vision-language models. In *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024)*, pages 145–171.
- Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Zhouchi Lin, Bowen Zhang, Lionel Ni, Wen Gao, Yuanzhuo Wang, and Jian Guo. 2026. *A survey on llm-as-a-judge*. *The Innovation*, 7(6).
- John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislaw Nikolov, Rishub Jain, Jonas Adler, and 15 others. 2021. *Highly accurate protein structure prediction with alphafold*. *Nature*, 596(7873):583–589.
- Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. 2023. *GlottID: Language identification for low-resource languages*.

- In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 6155–6218, Singapore. Association for Computational Linguistics.
- Tom Kocmi and Christian Federmann. 2023a. GEMBA-MQM: Detecting translation quality error spans with gpt-4. In Proceedings of the Eighth Conference on Machine Translation, pages 768–775.
- Tom Kocmi and Christian Federmann. 2023b. Large language models are state-of-the-art evaluators of translation quality. In Proceedings of the 24th Annual Conference of the European Association for Machine Translation, pages 193–203.
- Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024. Error span annotation: A balanced approach for human evaluation of machine translation. In Proceedings of the Ninth Conference on Machine Translation, pages 1440–1453.
- ZhaoBin Li and Mark Steyvers. 2025. Beyond accuracy: How AI metacognitive sensitivity improves AI-assisted decision making. arXiv preprint arXiv:2507.22365.
- Gabrielle Kaili-May Liu, Areeb Gani, Jacqueline Lu, Jordan Thomas, Mark Steyvers, and Arman Cohen. 2026a. Metacognition in LLMs: Foundations, progress, and opportunities. arXiv preprint arXiv:2607.11881.
- Jingxuan Liu, Zhi Qu, Jin Tei, Hidetaka Kamigaito, Lemao Liu, and Taro Watanabe. 2026b. XQ-MEval: A dataset with cross-lingual parallel quality for benchmarking translation metrics. In Findings of the Association for Computational Linguistics: ACL 2026, pages 37034–37052.
- Arle Lommel, Serge Gladkoff, Alan Melby, Sue Ellen Wright, Ingemar Strandvik, Katerina Gasova, Angelika Vaasa, Andy Benzo, Romina Marazzato Sparano, Monica Foresi, Johani Innis, Lifeng Han, and Goran Nenadic. 2024. The multi-range theory of translation quality measurement: MQM scoring models and statistical quality control. In Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 2: Presentations), pages 75–94, Chicago, USA. Association for Machine Translation in the Americas.
- Qingyu Lu, Liang Ding, Kanjian Zhang, Jinxia Zhang, and Dacheng Tao. 2025. MQM-APE: toward high-quality error annotation predictors with automatic post-editing in LLM translation evaluators. In Proceedings of the 31st International Conference on Computational Linguistics, pages 5570–5587.
- Qingyu Lu, Baopu Qiu, Liang Ding, Kanjian Zhang, Tom Kocmi, and Dacheng Tao. 2024. Error analysis prompting enables human-like translation evaluation in large language models. In Findings of the Association for Computational Linguistics: ACL 2024, pages 8801–8816.
- Samar M Magdy, Fakhraddin Alwajih, Abdellah El Mekki, Wesam El-Sayed, and Muhammad Abdul-Mageed. 2026. LQM: Linguistically motivated multidimensional quality metrics for machine translation. arXiv preprint arXiv:2604.18490.
- Ali Marashian, Alexis Palmer, and Katharina von der Wense. 2026. Speaking in self-assessing tongues: On the verbalized confidence of LLMs in machine translation. arXiv preprint arXiv:2606.17234.
- Stefano Perrella, Eric Morales Agostinho, and Hugo Zaragoza. 2026. Span-level machine translation meta-evaluation. arXiv preprint arXiv:2603.19921.
- Shenbin Qian, Constantin Orasan, Diptesh Kanojia, and Félix Do Carmo. 2024. Are large language models state-of-the-art quality estimators for machine translation of user-generated content? In Proceedings of the Eleventh Workshop on Asian Translation (WAT 2024), pages 45–55.
- Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In Proceedings of the Seventh Conference on Machine Translation (WMT), pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- John Schulman, Barret Zoph, Christina Kim, Jacob Hilton, Jacob Menick, Jiayi Weng, Juan Felipe Ceron Uribe, Liam Fedus, Luke Metz, Michael Pokorny, Rapha Gontijo Lopes, Shengjia Zhao, Arun Vijayvergiya, Eric Sigler, Adam Perelman, Chelsea Voss, Mike Heaton, Joel Parish, Dave Cummings, and 68 others. 2022. Introducing ChatGPT. <https://openai.com/blog/chatgpt>. Accessed: July 9, 2026.
- Richard Servajean and Philippe Servajean. 2026. Measuring the metacognition of AI. arXiv preprint arXiv:2603.29693.
- Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z Ren, and Anirudha Majumdar. 2025. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. ACM Computing Surveys, 58(3):1–38.
- Archchana Sindhujan, Diptesh Kanojia, Constantin Orasan, and Shenbin Qian. 2025. When LLMs struggle: Reference-less translation evaluation for low-resource languages. In Proceedings of the First Workshop on Language Models for Low-Resource Languages, pages 437–459.
- Di Wu and Christof Monz. 2025. UvA-MT at WMT25 evaluation task: LLM uncertainty as a proxy for translation quality. In Proceedings of the Tenth Conference on Machine Translation, pages 974–983.

- Taemin Yeom, Yonghyun Ryu, Yoonjung Choi, and Jinyeong Bak. 2025. [Tagged span annotation for detecting translation errors in reasoning LLMs](#). In Proceedings of the Tenth Conference on Machine Translation, pages 878–886, Suzhou, China. Association for Computational Linguistics.
- Chrysoula Zerva, Taisiya Glushkova, Ricardo Rei, and André FT Martins. 2022. Disentangling uncertainty in machine translation evaluation. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 8622–8641.
- Chrysoula Zerva and André FT Martins. 2024. Conformalizing machine translation evaluation. Transactions of the Association for Computational Linguistics, 12:1460–1478.
- Vilém Zouhar, Roman Grundkiewicz, Sara Rajaei, Parker Riley, Martin Popel, Rachel Bawden, Philipp Koehn, Marine Carpuat, and Tom Kocmi. 2026. Contrastive ESA: Human evaluation of multiple translations at once. arXiv preprint arXiv:2607.26640.