ASKQE: Question Answering as Automatic Evaluation for Machine Translation

Andrea Cauda s343386 | Roberto Cozzone s336155 | Pietro Giancristofaro s341870 | Davide Tonetti s334297 | Antonio Visciglia s346837

***Abstract*—Reference-based metrics for Machine Translation (MT) offer limited semantic coverage and interpretability, especially in biomedical settings. We study AskQE, a question-answering (QA) formulation of MT evaluation, and assess whether it remains effective in reproducible, low-cost configurations based on open-source LLMs. We replicate the AskQE pipeline replacing proprietary models with three compact instruction-tuned open-source LLMs, Meta-Llama-3.1-8B-Instruct, Qwen-2.5-7B-Instruct, and Gemma-2-9B-IT, runnable on consumer hardware. We evaluate on two benchmarks: ContraTICO, a controlled synthetic test suite with eight perturbation types at two severity levels, and BioMQM, a real-world biomedical MT dataset annotated under the MQM framework. Beyond single-model baselines, we propose The Champions Trio, a Multi-LLM ensemble that selects consensus answers via semantic centroid aggregation, stabilizing segment-level evaluation signals. The ensemble achieves a decision accuracy of 67.1% on ContraTICO (+15.0 pp over Qwen) and 73.4% on BioMQM, outperforming the 63.77% reported by the original SBERT-based AskQE approach. To move beyond scalar scores, we introduce an LLM-as-a-Judge module (Qwen2.5-7B) that assigns structured error categories from a predefined taxonomy and produces a natural-language explanation for each translation, making evaluation transparent and actionable. The judge is validated on ContraTICO before application to biomedical data. Overall, our study demonstrates that AskQE-style evaluation can be made cheaper, open, and more interpretable while retaining competitive performance against human judgments. Code available at: https://github.com/gitandrehub/AskQE-dnlp.**

I. Problem Statement

The automatic evaluation of Machine Translation (MT) systems remains a central challenge in Natural Language Processing. Traditional reference-based metrics such as XCOMET-QE [8], MetricX-QE [7], or BT-Score [6], while efficient and widely adopted, have well-known limitations: they correlate imperfectly with human judgments, struggle with semantic adequacy, and provide little to no interpretability regarding the nature and severity of translation errors. These shortcomings become even more evident in specialized domains, such as biomedical translation, where minor lexical variations may be acceptable while subtle semantic errors can be critical.

Recent approaches have proposed leveraging large language models (LLMs) to perform MT evaluation through Question Answering (QA), framing translation quality assessment as the ability to answer meaning-preserving questions derived from the source text. AskQE (Question Answering as Evaluation) [5] represents a significant step in this direction, demonstrating strong correlation with human judgments and improved error interpretability. However, the original framework relies heavily on large proprietary models (e.g., GPT-4o), which introduces substantial limitations in terms of reproducibility, cost, accessibility, and sustainability, especially in academic or low-resource settings.

Beyond this, even when AskQE produces a quality score, it offers no explanation of *why* a translation is penalised: the scalar output indicates that semantic inconsistencies exist but provides no information on their nature or severity, leaving the evaluation inherently opaque.

This work addresses both limitations. First, we investigate whether AskQE retains its effectiveness when proprietary LLMs are replaced with resource-efficient open-source models runnable on consumer hardware, then we introduce an ensemble strategy based on semantic centroid selection to stabilize evaluation signals against variability in individual model outputs. Second, we extend the framework toward interpretable evaluation by integrating a dedicated LLM-as-a-Judge module that goes beyond scalar scores: for each translation, the judge assigns a category from a structured error taxonomy and produces a concise natural-language explanation of the detected error, making the evaluation transparent and actionable.

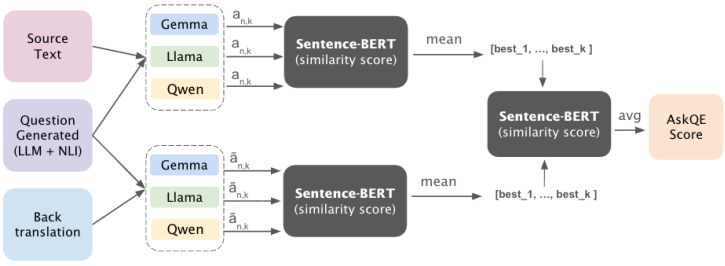

Fig. 1. Functional diagram of the AskQE-style evaluation pipeline. The QA backbone generates answers using three open-source LLMs which can be evaluated independently or combined via semantic centroid aggregation (Extension I).

II. Data Description

To ensure a rigorous assessment of our reproduced AskQE framework, we employ a dual-dataset strategy that combines controlled synthetic calibration with real-world validation, allowing us to evaluate the model's ability to assign significantly lower scores to critical perturbations compared to minor ones,

before testing generalization.

We first utilize **ContraTICO**, a synthetic test suite derived from the TICO-19 biomedical corpus. Starting from high-quality human translations, the dataset authors introduced deliberate errors using GPT-4o, yielding 8 perturbation types across two severity levels (minor and critical). This controlled construction allows precise measurement of how sensitively the metric responds to each error type.

Them, we use **BioMQM** to validate the framework on real-world biomedical translation errors. Annotations follow the Multidimensional Quality Metrics (MQM) framework, in which expert human annotators identify error spans and assign quality scores. This dataset provides a reliable human-grounded reference for evaluating the correlation between our metric and expert judgment.

III. Methodology

In our work we re-implements the AskQE backbone using efficient open-source LLMs and extends it with ensemble-based aggregation and interpretable error categorization.

A. Question generation and NLI-based grounding

We reuse the question sets released by the original AskQE framework, generated via a LLaMA-3 70B model guided by an NLI-based grounding pipeline: atomic facts are extracted from the source x_n, filtered by entailment, and used to generate semantic questions Q_n. Reusing these resources avoids additional generation costs and ensures full comparability with the original framework.

B. Problem formulation and QA backbone

We formulate MT evaluation over a dataset

$$\mathcal{D} = \{(x_n, y_n, \tilde{x}_n, Q_n)\}_{n=1}^{N},$$

where x_n is the source segment, y_n a candidate translation, $\tilde{x}_n$ its backtranslation, and $Q_n = \{q_{n,k}\}_{k=1}^{K_n}$ the set of AskQE-generated questions. Given an LLM m and a question $q_{n,k}$, we extract answers conditioned on the source and on the backtranslation,

$$a_{n,k}^{(m)} = f_m(x_n, q_{n,k}), \qquad \tilde{a}_{n,k}^{(m)} = f_m(\tilde{x}_n, q_{n,k}),$$

where $f_m(\cdot)$ denotes instruction-following generation constrained to the provided context.

We instantiate the QA backbone using three open-source instruction-tuned LLMs: Meta-Llama-3.1-8B-Instruct, Gemma-2-9B-IT and Qwen-2.5-7B-Instruct. For each model independently, segment-level quality is computed as

$$AskQE(x_n, y_n) = s_n^{(m)} = \frac{1}{K_n} \sum_{k=1}^{K_n} \phi\big(a_{n,k}^{(m)}, \tilde{a}_{n,k}^{(m)}\big),$$

where $\phi(\cdot,\cdot)$ computes cosine similarity between embeddings of the two answers using the Sentence-BERT model `all-MiniLM-L6-v2` [4].

C. Extension I: Ensemble LLMs for robust agreement

To mitigate variability across individual LLMs and improve robustness when handling ambiguous or underspecified questions, we introduce an ensemble QA module (Fig. 1) that combines model-specific answers into a single, more reliable output. As observed by Li et al. [9], integrating multiple models is motivated by the fact that different architectures and training data can yield complementary answers and feedback, improving the overall robustness of the system.

Our ensemble strategy, which we call **"The Champions Trio"**, operates at the answer level via semantic aggregation. For each question $q_{n,k}$, we:

1) Collect the three model-specific answers: $a_{n,k}^{(\text{llama})}$, $a_{n,k}^{(\text{qwen})}$, $a_{n,k}^{(\text{gemma})}$
2) Compute the average pairwise cosine similarities between all answer embeddings:

$$m^* = \arg\max_{m \in \mathcal{M}} \gamma_{n,k}$$

$$\gamma_{n,k} = \tfrac{1}{|\mathcal{M}|-1} \textstyle\sum_{m' \in \mathcal{M} \setminus \{m^*\}} \text{sim}\big(a_{n,k}^{(m^*)}, a_{n,k}^{(m')}\big),$$

   where $\mathcal{M} = \{\text{llama}, \text{qwen}, \text{gemma}\}$ and $\text{sim}(\cdot,\cdot)$ denotes cosine similarity between Sentence-BERT embeddings.
3) Select the answer closest to the semantic centroid as the consensus response:

$$\hat{a}_{n,k} = a_{n,k}^{(m^*)}$$

The same procedure is applied independently to backtranslation answers $\tilde{a}_{n,k}$, yielding ensemble answers $\hat{\tilde{a}}_{n,k}$ and consensus scores $\tilde{\gamma}_{n,k}$.

Finally, the askQE score for the ensemble segment-level quality is computed as:

$$s_n^{(\text{ens})} = \frac{1}{K_n} \sum_{k=1}^{K_n} \phi\big(\hat{a}_{n,k}, \hat{\tilde{a}}_{n,k}\big),$$

using the same Sentence-BERT similarity metric as for single models. Additionally, to detect segments where the source and back-translation answers diverge strongly after consensus selection, which may indicate translation failures or hallucinations, we flag instances where $\phi(\hat{a}_{n,k}, \hat{\tilde{a}}_{n,k}) < 0.3$.

D. Extension II: LLM as a Judge with QA evidence

Our second extension reframes AskQE from scalar quality estimation to structured error categorization, aiming at more interpretable MT evaluation. A judge LLM (`Qwen2.5-7B`) is prompted to assign a single error category to each translation, drawing on QA evidence produced by the backbone pipeline, and to provide a brief explanation of at most 15 words.

For each segment n, we construct a structured prompt consisting of three components: the source sentence x_n, the backtranslation $\tilde{x}_n$, and a QA evidence triplet comprising all questions $q_{n,k}$, source answer $a_{n,k}$ and backtranslation answer $\tilde{a}_{n,k}$. The prompt follows a role-task-definition structure: the model is assigned the role of a professional linguist, tasked

with identifying the specific error type from the discrepancy between source and back-translation, and provided with per-category definitions drawn from the original AskQE error taxonomy:

- **Critical:** *alteration*, *omission*, *expansion_impact*
- **Minor:** *expansion_noimpact*, *intensifier*, *spelling*, *synonym*, *word_order*

The full prompt template is reported in Appendix A.

IV. Experiments and Results

A. Experimental Design

All experiments are conducted in a Python environment on a single NVIDIA GPU L4 (Google Colab). For both datasets, we evaluate on the EN-ES language pair.

We acknowledge that the computational demands of ensemble inference with multiple Large Language Models necessitated a reduction of the dataset size; consequently, we did not utilize the full corpora but instead performed representative subsampling. For the ContraTICO calibration set, we selected 63 distinct source sentences and evaluated all 8 associated perturbation types for each, resulting in a total of 504 controlled test instances. Similarly, for the BioMQM validation set, we sampled 500 segments containing a diverse distribution of error severities.

All models share a common inference configuration: the maximum number of new tokens is set to 20 to prevent verbose or redundant outputs and encourage concise, answer-focused responses. The temperature is set to 0 and `do_sample` is set to `False` to enforce fully deterministic generation, eliminating stochastic variation across runs and ensuring full reproducibility of results.

B. Results

We report results separately for ContraTICO and BioMQM.

1) Results on ContraTICO: We follow the evaluation protocol of the original paper, fitting AskQE scores with a two-component Gaussian Mixture Model (GMM) to classify each source, translation pair as accepted or rejected, and comparing predictions against the ground-truth severity label (critical = reject, minor = accept).

Table I reports mean AskQE scores per perturbation type

Alteration consistently yields the lowest scores across all models (0.619–0.700), confirming that the metric is sensitive to direct meaning changes. Minor types (spelling, synonym, intensifier) cluster above 0.80, reflecting limited semantic distortion. The Champions Trio outperforms both Gemma-2-9B and Llama-3.1-8B across most perturbation types, with consistent gains on minor categories (e.g. $+0.047$ on *intensifier* and $+0.041$ on *expansion_noimpact* vs. the best individual baseline). Qwen2.5-7B achieves higher raw similarity scores across all eight perturbation types. However, the ensemble surpasses Qwen on the most practically relevant metric: decision accuracy (0.671 vs. 0.520, $+0.150$). This result indicates that centroid aggregation produces better-calibrated outputs for binary accept/reject decisions.

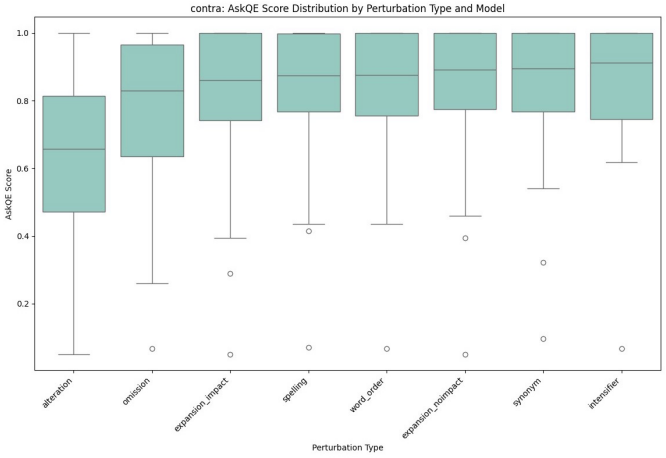

Fig. 2. AskQE score distribution by perturbation type for the Champions Trio ensemble on ContraTICO (EN-ES). Alteration and omission show the lowest medians and widest spread; minor perturbation types cluster above 0.85.

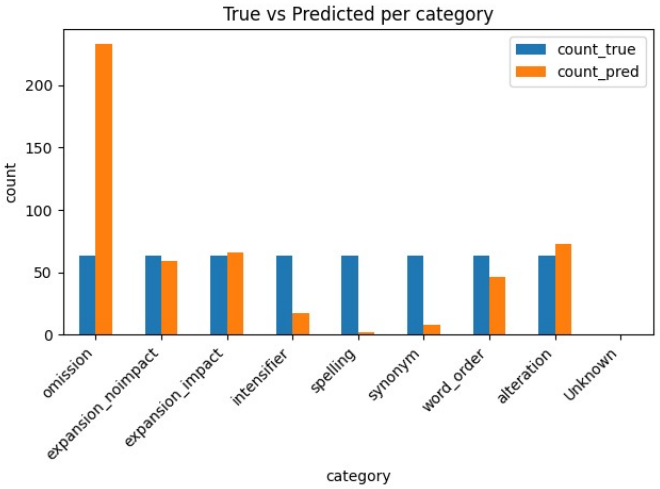

Fig. 3. True vs. predicted category counts for Extension II (Qwen2.5-7B judge) on ContraTICO ($\approx$ 63 instances per category).

Figure 2 confirms the expected quality gradient. *Alteration* scores the lowest (median ≈ 0.66, IQR 0.47–0.81), as direct meaning changes produce strongly divergent backtranslation answers. *Omission* occupies an intermediate position (median ≈ 0.83) with high variance, reflecting that partial omissions are harder to detect than complete ones. All minor perturbation types cluster above 0.85, confirming that meaning-preserving modifications leave the consistency signal largely unaffected. Notably, *expansion_impact*,despite being a critical perturbation type, exhibits a similar score distribution to minor categories (median ≈ 0.86), indicating that the metric fails to penalise this error type adequately.

As shown in Figure 3, the model produces a strong over-prediction bias concentrated on *omission* (232 predicted vs. 63 true, $+169$), which absorbs predictions from multiple under-represented minor categories. By contrast, *expansion_noimpact* (60 vs. 63), *expansion_impact* (65 vs.

TABLE I

MEAN ASKQE SCORES PER PERTURBATION TYPE AND DECISION ACCURACY ON CONTRATICO (EN-ES). *Critical*: ALTERATION, OMISSION, EXPANSION_IMPACT. *Minor*: SPELLING, WORD_ORDER, SYNONYM, INTENSIFIER, EXPANSION_NOIMPACT. BEST RESULTS IN **BOLD**.

Model	Alteration	Exp. Impact	Exp. No-imp.	Intensifier	Omission	Spelling	Synonym	Word Order	Dec. Acc.
Gemma-2-9B	0.627	0.776	0.807	0.812	0.735	0.820	0.823	0.777	0.607
Llama-3.1-8B	0.651	0.786	0.806	0.834	0.761	0.808	0.822	0.800	0.464
Qwen2.5-7B	**0.700**	**0.887**	**0.888**	**0.887**	**0.827**	**0.884**	**0.895**	**0.890**	0.520
Champions Trio	0.619	0.827	0.854	0.859	0.782	0.841	0.850	0.838	**0.671**

TABLE II

CHAMPIONS TRIO RESULTS ON BIOMQM BY SEVERITY. *Disagreement*: SEGMENTS WHERE THE MODEL'S BINARY DECISION DIFFERS FROM HUMAN ANNOTATION. GLOBAL: KENDALL $\tau = 0.0634$ ($p = 0.073$), DECISION ACCURACY (GMM) $= 73.40\%$.

Severity	Mean AskQE	Disagreement
No error	0.914	66/318
Neutral	0.909	2/10
Minor	0.882	46/145
Major	0.928	16/20
Critical	0.860	3/7

63), and *alteration* (72 vs. 63) are predicted with good accuracy. Fine-grained lexical categories remain strongly under-predicted: *spelling* (3), *synonym* (9), and *intensifier* (16) receive far fewer labels than their true frequency, as their surface-level signals are indistinguishable from omission when only one QA triplet is available. Compared to the original paper, our "LLM as a Judge" overcomes the difficulty observed in the original AskQE framework in detecting *expansion_noimpact* errors, while still struggling to identify errors that affect only the intensity of the meaning. Meanwhile, the model adopts a conservative decision threshold, systematically over-predicting critical error types.

2) Results on BioMQM: Table II reports ensemble results disaggregated by human severity level.

The per-severity breakdown exposes a failure that scalar metrics alone would obscure. The most striking is a *severity inversion*: *major* errors achieve the highest mean AskQE score of all classes (0.928), surpassing even error-free segments (0.914), and are misclassified in 80% of cases (16/20). The global Kendall $\tau = 0.063$ ($p = 0.073$) is not statistically significant, indicating that the ensemble scores do not reliably capture the fine-grained quality ranking encoded in human ratings. However, the global decision accuracy of 73.40% is more encouraging and compares favourably against the 63.77% reported by the original AskQE paper for SBERT-based approaches, suggesting that our ensemble is better calibrated for binary accept/reject decisions than for quality ranking. A direct comparison remains limited, as our evaluation is conducted on a smaller sample (500 segments) than the original study due to computational constraints.

V. CONCLUSIONS AND LIMITATIONS

This work demonstrates that the AskQE evaluation framework can be effectively reproduced using compact, open-source LLMs without access to proprietary models, making it accessible in academic and resource-constrained settings. On ContraTICO, semantic similarity scores show a clear quality gradient: alteration errors yield the lowest scores (median ≈ 0.66), while minor perturbation types consistently cluster above 0.85, confirming that the backtranslation-based consistency signal reliably captures coarse-grained meaning distortions. The Champions Trio ensemble, combining Llama-3.1-8B, Qwen-2.5-7B, and Gemma-2-9B via centroid-based answer selection, achieves the best decision accuracy (0.671), outperforming all individual models on binary accept/reject classification, despite Qwen obtaining higher raw similarity scores. This confirms that centroid aggregation yields better-calibrated outputs than any single model alone.On BioMQM, the ensemble reaches 73.4% decision accuracy, surpassing the original AskQE SBERT baseline, though the non-significant Kendall τ (0.063, $p = 0.073$) indicates limited ability to rank translations at a fine-grained level. A notable severity inversion, major errors scoring above error-free segments, reveals a structural weakness of the backtranslation-based approach in real-world biomedical data. This inversion likely reflects the nature of major biomedical errors, which often involve domain-specific terminology or register shifts that do not produce divergent back-translations, rendering the consistency signal blind to this error type.

The LLM-as-a-Judge extension provides structured, interpretable error categorization beyond scalar scores. While structurally grounded categories (alteration, expansion) are predicted accurately, fine-grained lexical types (spelling, synonym, intensifier) remain systematically under-predicted due to insufficient discriminative signal in a single QA triplet.

Several constraints bound the scope of our conclusions: 1) all experiments are conducted on subsampled data (504 ContraTICO instances, 500 BioMQM segments) due to GPU memory and time constraints on a single L4 GPU, limiting direct comparison with the original paper; 2) evaluation is restricted to the EN-ES language pair, and generalization to other language pairs or domains cannot be assumed; 3) question sets are reused from the original AskQE framework rather than regenerated, which, while ensuring comparability, may introduce a dependency on the quality of those questions; 4) the judge's omission bias and inability to distinguish surface-level lexical errors suggest that richer prompting strategies or multi-triplet evidence aggregation should be explored in future work.

REFERENCES

[1] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," in *Proc. ACL*, 2002.

[2] S. Banerjee and A. Lavie, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments," in *Proc. ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, 2005.

[3] M. Popović, "chrF: Character n-gram F-score for Automatic MT Evaluation," in *Proc. WMT*, 2015.

[4] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in *Proc. EMNLP-IJCNLP*, 2019.

[5] D. Ki, K. Duh, and M. Carpuat, "AskQE: Question Answering as Automatic Evaluation for Machine Translation,"

[6] S. Patra and A. Tarsis, "Quality Estimation via Backtranslation," 2024.

[7] J. Juraska, D. Deutsch, M. Finkelstein, M. Freitag, "MetricX-24: The Google submission to the WMT 2024 metrics shared task," 2024.

[8] N. M. Guerreiro, R. Rei, D. van Stigt, L. Coehur, P. Colombo, A. F. T. Martins, "XCOMET: Transparent Machine Translation Evaluation through Fine-Grained Error Detection," 2024.

[9] Y. Li et al., "Composing Ensembles of Pre-trained Models via Iterative Consensus," 2022.

APPENDIX A
PROMPTS FOR THE EXTENSIONS

Prompts used in the experiments

PROMPT_EXTENSION I

```
You are a question answering model.

Rules:
- Use ONLY the information in the sentence.
- If the question can be answered with Yes or No, answer strictly Yes or No.
- If the question asks for a specific span (number, phrase, symptom, etc.), copy it EXACTLY from the
    sentence.
- Do NOT invent information.
- Answer with a SHORT phrase, without explanations or additional text.
- Do not output code.
- Respond with the answer only.

Sentence: {{sentence}}
Question: {{question}}
Answer:
```

PROMPT_EXTENSION II

```
JUDGE_PROMPT = You are a professional linguist evaluating Machine Translation quality using
    Backtranslation.

Task: Analyze the discrepancies between the SOURCE text and the BACKTRANSLATION to identify the
    specific type of translation error.

Definitions:
- Source: The original English text.
- Backtranslation: The translation of the MT output back into English.
- QA Evidence: Questions asked to both texts. Different answers indicate an error.

Error Categories (Choose ONE):
1. omission: Information in Source is missing in Backtranslation.
2. expansion_noimpact: Added detail that does not change the overall meaning; stylistic expansion.
3. expansion_impact: Added content that changes or extends the meaning compared to the source.
4. intensifier: Adds intensity (e.g., severe) that was not in the source, i.e., extra semantic
    content.
5. spelling: Orthographic / surface error.
6. synonym: Lexical substitution.
7. word_order: Directly corresponds to word order / reordering errors.
8. alteration: Stronger change of meaning or contradiction relative to the source.

*** ANALYSIS ***
Source: {{source}}
Backtranslation: {{backtrans}}

QA Evidence (Discrepancies):
{{qa_evidence}}

*** INSTRUCTIONS ***
Write ONLY two lines:
1) Category: <state the category explicitly in the format Category: [CategoryName]>.
2) Explanation: <max 15 words>.


Response:
```

APPENDIX B
JUDGE CLASSIFICATION

Qualitative Example

Metadata

```
row_index: 2
index_db: 1967
true_category: expansion_noimpact
predicted_category: expansion_noimpact
explanation: Added clear does not change the
    core meaning
askqe_score: 1.0
```

Source and Backtranslation

```
Source: i will send you an image
Backtranslation: I will send you a clear image
```