ASKQE: Question Answering as Automatic Evaluation for Machine Translation

Dayeon Ki[✦] **Kevin Duh**[✿] **Marine Carpuat**[✦]
[✦]University of Maryland [✿]Johns Hopkins University
{dayeonki,marine}@umd.edu kevinduh@cs.jhu.edu

Abstract

How can a monolingual English speaker determine whether an automatic translation in French is good enough to be shared? Existing MT error detection and quality estimation (QE) techniques do not address this practical scenario. We introduce ASKQE, a question generation and answering framework designed to detect critical MT errors and provide actionable feedback, helping users decide whether to accept or reject MT outputs even without the knowledge of the target language. Using CONTRATICO, a dataset of contrastive synthetic MT errors in the COVID-19 domain, we explore design choices for ASKQE and develop an optimized version relying on LLAMA-3 70B and entailed facts to guide question generation. We evaluate the resulting system on the BIOMQM dataset of naturally occurring MT errors, where ASKQE has higher Kendall's τ correlation and decision accuracy with human ratings compared to other QE metrics.[1]

1 Introduction

How can a monolingual English speaker determine whether an automatic translation of COVID-19 protocol in French is good enough to be shared? In such high-stakes settings, inaccurate translations can lead to confusion, conversation breakdowns (Yamashita et al., 2009), and even life-threatening risks (Berger, 2017; Vieira et al., 2021; Mehandru et al., 2022). For instance, even rare inaccurate translations in clinical instructions can pose significant risks to patient health (Khoong et al., 2019). However, accurately assessing Machine Translation (MT) quality is significantly more challenging for monolingual speakers than for bilinguals, as they rely on translations in a language they do not understand (Mehandru et al., 2023). To bridge this gap, we need Quality Estimation (QE) feedback to help users assess MT quality (Han et al., 2021).

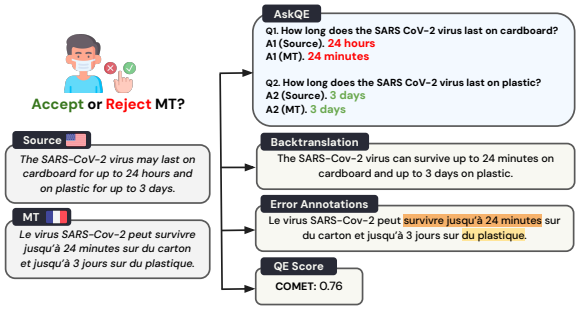


Figure 1: The goal of ASKQE is to generate QA pairs that identify critical translation errors and help monolingual source speakers decide whether to accept or reject MT.

Research in this space has largely relied on QE metrics that provide segment-level holistic assessments of MT quality, but these can be difficult to interpret and fail to explain how mistranslations impact users. Most QE metrics are primarily trained to produce either a single scalar score (Fernandes et al., 2023; Fu et al., 2024a) or error annotations that highlight problematic spans in the target language (Kocmi and Federmann, 2023; Guerreiro et al., 2024; Lu et al., 2024). This indicates that there is currently no feedback to use at the right granularity to support decision-making for users who do not understand the target language (Zouhar and Bojar, 2020). These challenges lead to a key question: How can we identify critical translation errors and provide actionable feedback to help monolingual source speakers decide whether to accept or reject MT in high-stakes contexts?

We hypothesize that asking and answering questions about MT outputs is particularly well suited to address this question by providing a mechanism for explainable MT evaluation. It has been studied and shown effective for assessing factual consistency in related generation tasks such as summarization (Durmus et al., 2020; Wang et al., 2020; Zhong et al., 2022; Deutsch et al., 2021; Fabbri et al., 2022). For MT, questions and answers can provide functional explanations of MT quality, highlight-

[1]We release our code and dataset at https://github.com/dayeonki/askqe.

Findings of the Association for Computational Linguistics: ACL 2025, pages 17478–17515

ing the functional consequences of potential errors rather than providing a mechanistic explanation of what is wrong (Lombrozo and Wilkenfeld, 2019). This approach also aligns with the view of explanations as social, facilitating a knowledge transfer as part of an interaction where the user can weigh the evidence provided in the context of their beliefs (Miller, 2019). Question and answers also have the potential to integrate well with techniques people use to estimate whether their interlocuator understood what was said, such as the teach-back techniques that physicians use in cross-lingual communication settings (Mehandru et al., 2022).

To explore this direction, we introduce **AskQE**, a question generation and answering framework based on the idea that a translation is unreliable if key questions about the source text yield different answers when derived from the source or the backtranslated MT. As illustrated in the example in Figure 1, the monolingual English speaker can see that Q1 is answered incorrectly while Q2 is answered correctly. AskQE consists of two key components: **1)** Question Generation (**QG**) conditioned on the source sentence and the entailed facts extracted from it (§3.1) and **2)** Question Answering (**QA**) based on the source and the backtranslated MT (§3.2) as in Figure 2.

First, we validate our approach and explore design choices using ContraTICO, a controlled synthetic dataset. We simulate MT errors by perturbing translations in the TICO-19 dataset (Anastasopoulos et al., 2020) across five language pairs (English to Spanish, French, Hindi, Tagalog, and Chinese), creating the ContraTICO dataset (§4). We test different variations of models and information given during QG, and propose an optimized version with LLaMA-3 70B (Grattafiori et al., 2024) and entailed facts to guide QG (§6.1). We then show that AskQE effectively distinguishes minor errors from critical ones (§6.2) and aligns well with established QE metrics (§6.3). Second, we show that our findings generalize to naturally occurring MT errors and additional language pairs in BioMQM (Zouhar et al., 2024), achieving comparable Kendall's τ correlation with human judgments (§7.1). Given AskQE's sensitivity to error severity and correlation with human judgments, we hypothesize that it provides actionable findings, which motivates our decision making simulation experiments. Here, we demonstrate that using AskQE feedback achieves higher decision accuracy than other QE metrics (§7.2).

2 Background & Related Work

2.1 QA as (MT) Evaluation

Using Question Answering (QA) for evaluation has been predominantly studied in the context of summarization, where questions are generated from summaries, and answer pairs from the summary and the source document are compared to assess factual consistency between the original document and its summary (Durmus et al., 2020; Wang et al., 2020; Zhong et al., 2022; Deutsch et al., 2021; Fabbri et al., 2022). In parallel, Question Generation (QG) models have been explored to improve question quality, primarily through pre-training language models (Riabi et al., 2021; Shakeri et al., 2021; Dugan et al., 2022). More recently, Fu et al. (2024b) shows that using Large Language Models (LLMs) such as GPT-4 (OpenAI et al., 2024) generate higher-quality questions compared to traditional QG approaches using pre-trained models.

On the other hand, the use of QA for MT evaluation has primarily focused on manual evaluation at the system level. Early approaches employ reading comprehension tests to assess the informativeness and usefulness of MT outputs (Tomita et al., 1993; Fuji, 1999; Fuji et al., 2001) or their readability (Jones et al., 2005b,a, 2007). Berka et al. (2011) introduce yes/no type questions for manual MT evaluation in the English-Czech language pair, finding that different MT systems produce outputs with varying answer accuracy. Weiss and Ahrenberg (2012) extend this approach to Polish-English translations, showing that MT outputs with more errors lead to lower answer accuracy. More recent studies focus on using QA for automatic MT evaluation (Sugiyama et al., 2015; Scarton and Specia, 2016; Han et al., 2022; Krubiński et al., 2021). The closest work to ours is MTEQA (Krubiński et al., 2021), which extracts answer spans and generates questions from the reference translation and compare answer overlap derived from the reference and MT. However, their approach relies on predefined answer spans and reference translations. In a contemporaneous pre-print, Fernandes et al. (2025) show that QA-based evaluation outperforms both neural and LLM-based metrics for ranking paragraph-level translations. We depart from this by exploring the potential of QG/QA framework for more fine-grained, sentence-level MT evaluation with varying levels of error severity.

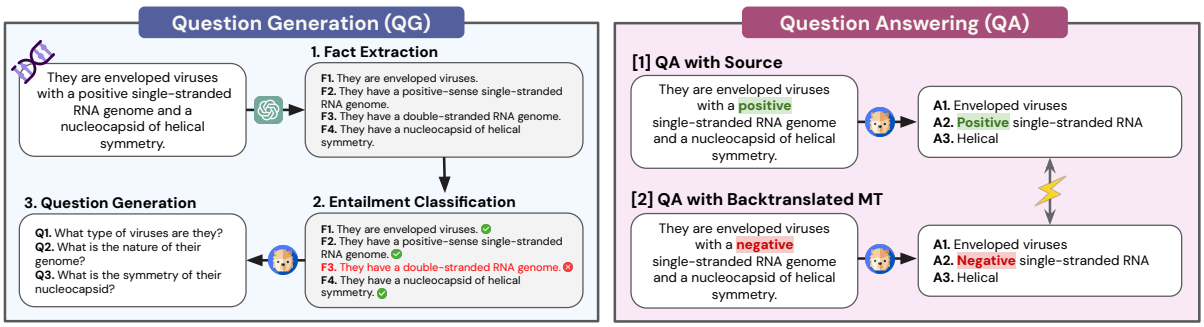


Figure 2: Overview of ASKQE. Given source sentence, we **1) QG:** Generate questions based on the source and entailed atomic facts extracted from it, and **2) QA:** Answer each question using the source sentence and the backtranslated MT. Discrepancies in the answers are indicators of potential errors in the translation. The same model is used for both QG and QA.

2.2 Quality Estimation

Quality Estimation (QE) is the task of automatically assessing MT output quality without relying on human references (Specia et al., 2018). This form of feedback is particularly important for monolingual source speakers, who often lack both target language proficiency and domain expertise to evaluate MT quality as effectively as professional translators or bilinguals (Mehandru et al., 2023). Most existing QE research has focused on training models to produce a single scalar score (Fernandes et al., 2023; Fu et al., 2024a) or error annotations (Kocmi and Federmann, 2023; Guerreiro et al., 2024; Lu et al., 2024) which output error spans in the target language. In parallel, the explainable QE shared task (Fomicheva et al., 2021) frames translation error identification as an explainable QE task, where sentence-level quality judgments are explained by highlighting error-inducing words in the MT (Eksi et al., 2021; Rubino et al., 2021) or providing textual explanations of metric outputs (Fomicheva et al., 2022; Jiang et al., 2024; Lu et al., 2024). While we align with these efforts in pursuing more explainable evaluation metrics, we diverge from prior work by focusing on source side annotation through a QG/QA framework.

We compare our method against three established QE metrics: **1)** XCOMET-QE (Guerreiro et al., 2024), **2)** METRICX-QE (Juraska et al., 2024), and **3)** BT-score (Agrawal et al., 2022). Both XCOMET-QE and METRICX-QE evaluate the source and MT output, with XCOMET-QE generating segment-level scores and error annotations and METRICX-QE generating scores. BT-Score assess the similarity between the source and backtranslated MT output using BERTSCORE (Zhang et al., 2020) as the MT metric.

3 ASKQE

We present an overview of ASKQE in Figure 2. We first generate a list of questions conditioned on the source (§3.1), generate answers for each question based on the source or the backtranslated MT output (§3.2), and compute the answer overlap (§3.3). All prompts are in Appendix A.

3.1 Question Generation (QG)

Given a source sentence X_{src}, we generate a set of questions Q_{src} that can be answered based on the sentence. Before generating questions, we extract information from X_{src} on what to ask questions about and incorporate it as additional context in the prompt. Specifically, to ensure comprehensive coverage of the information from X_{src}, we implement a two-step natural language inference (NLI) pipeline (Stacey et al., 2024): **1) Fact extraction**, where we prompt GPT-4o[2] to extract atomic facts that can be inferred from the source sentence; **2) Entailment classification**, where we use an off-the-shelf NLI classifier[3] to assess the binary entailment relationship (entailed or contradictory) between each extracted fact (as the hypothesis) and X_{src} (as the premise). We discard facts labeled as contradictory, potentially indicating that they cannot be reliably inferred from X_{src}.[4] Finally, we prompt an LLM to generate questions given X_{src} and the filtered set of entailed atomic facts. Details on other tested QG variants beyond the NLI pipeline are provided in Appendix D.2.

[2] https://openai.com/index/hello-gpt-4o/

[3] https://huggingface.co/potsawee/deberta-v3-large-mnli

[4] On average, each instance yields 3.61 facts, with 3.08 retained after entailment classification. We show the effect of entailment classification in Appendix D.3.

3.2 Question Answering (QA)

We generate answers for each question in Q_{src} using two different contexts: source sentence and the backtranslated MT output.

QA with Source. We provide the source sentence X_{src} and each question in Q_{src} as context and prompt an LLM to generate *reference* answers A_{src}, serving as the ground truth for evaluation.

QA with Backtranslated MT. Comparing answers derived from the source (A_{src}) and the MT output Y_{tgt} requires a cross-lingual QA system. However, cross-lingual QA systems may be less accurate than English QA systems, leading to potential disagreements in answer pairs that stem from differences between QA systems rather than translation errors. To mitigate this, we use a monolingual English QA system and instead rely on backtranslation to obtain an English representation of the MT output Y_{bt}. While backtranslation may introduce some noise, we hypothesize that with a high-quality MT system, it is unlikely to mask errors present in the original MT. At worst, it may introduce new errors, potentially making the system overly cautious rather than overlooking critical errors. To this end, we generate *predicted* answers A_{bt} from Y_{bt}.[5]

3.3 AskQE Outputs

Given the AskQE framework, we can present the QE information in multiple ways. On the one hand, we could simply list all the questions and answer pairs or only those where the answers differ. On the other hand, we could use these pairs to compute a score. While other variants exist, a full exploration is left for future work. In this study, we validate AskQE using a two-step process: **1)** measuring answer overlap between A_{src} and A_{bt} using a similarity metric, and **2)** aggregating question-answer similarities into a segment-level metric. For **1)**, following Krubiński et al. (2021), we explore several similarity metrics commonly used in QA and MT evaluation. We consider both string-comparison metrics for lexical overlap and neural metric for understanding semantic similarity:

- **Word-level F1** and **Exact Match (EM)**, following prior QA evaluation works (Rajpurkar et al., 2016; Chen et al., 2019; Durmus et al., 2020).
- **BLEU** (Papineni et al., 2002) and **chrF** (Popović, 2015), widely used in MT evaluation.
- **SentenceBERT** for wordpiece-level embedding similarities (Reimers and Gurevych, 2019).

For **2)**, we compute the final AskQE score for a given translation Y_{tgt} by averaging the similarity scores $D(\cdot, \cdot)$ across all N generated questions:

$$\text{AskQE}(Y_{\text{tgt}}) = \sum_{i=1}^{N} \frac{D(A_{\text{src}}, A_{\text{bt}})}{N} \quad (1)$$

4 ContraTICO: Controlled Error Generation by Perturbation

To simulate high-stakes settings, we use TICO-19 (Anastasopoulos et al., 2020), a MT dataset containing COVID-19-related content translated from English into 36 languages.[6] Since the original TICO-19 dataset only provides English source and reference translations in the target language, we construct a dataset with eight synthetic perturbations across five language pairs, ContraTICO, to assess the impact of AskQE design in a controlled setting. Formally, given a reference translation Y_{ref}, we prompt GPT-4o to perturb Y_{ref} with specific error, which results in a perturbed translation Y_{tgt}. We define eight linguistic perturbations, categorized by their level of severity into either Minor or Critical, based on the potential implications of the translation error in practice (Freitag et al., 2021). Typological errors, for instance, have minimal influence on MT quality (Sai et al., 2021), whereas errors that alter the semantics of the translation, such as those introducing misleading information, have a significantly greater impact (Karpinska et al., 2022). All eight perturbations are applied to each reference translation. We show detailed examples for each perturbation in Appendix Table 16.[7]

Minor. These errors do not lead to loss of meaning but introduce small inaccuracies or stylistic inconsistencies that might marginally affect clarity. We carefully design perturbations as minimal pairs with the reference translation (i.e., differ in only one specific aspect) (Warstadt et al., 2020). We define five types of minor perturbations:

- **Spelling:** Misspell one to two words.
- **Word Order:** Reorder words in the sentence.
- **Synonym:** Replace one word to its synonym.

[5] We show that the English QA system outperforms its cross-lingual counterpart in Appendix C.3.

[6] Detailed dataset statistics, categorized by data source, are outlined in Appendix Table 19.

[7] The overlap ratio of perturbed translations between perturbations is reported in Appendix Figure 6.

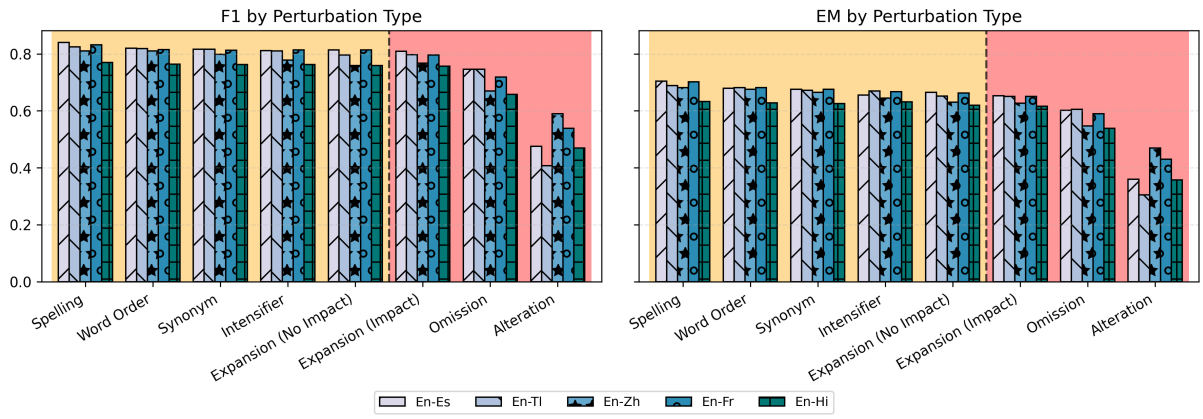


Figure 3: ASKQE (LLAMA-3 70B with NLI) using F1 and EM metric. In each subplot, the x-axis represents perturbation types, while the y-axis denotes the corresponding scores. Due to space constraints, we report only the two metrics common in QA research (Rajpurkar et al., 2016; Deutsch et al., 2021). Full results for all metrics and models are provided in Appendix D.1.

- **Intensifier:** Modify the intensity of an adjective or an adverb (e.g., *small* to *very small*).
- **Expansion (No Impact):** Expand a word or phrase by adding contextually implied details without introducing new meaning.

Critical. These errors significantly changes the original meaning and usually appear in a highly visible or important part of the content. We define three types of critical perturbations:

- **Expansion (Impact):** Expand a word or phrase by introducing new meaning.
- **Omission:** Omit a word or phrase.
- **Alteration:** Alter a word or phrase by changing its original meaning.

To validate the impact of different perturbations, we show that XCOMET-QE (Guerreiro et al., 2024) and cosine similarity scores between the original source X_{src} and the perturbed translation Y_{tgt} decrease as severity increases in Appendix C.1.

5 Experiment Setup

5.1 Dataset

We use CONTRATICO and BIOMQM (Zouhar et al., 2024) as our testbed.[8] BIOMQM is a biomedical domain MT dataset with error annotations by professional translators based on the multidimensional quality metrics (MQM) (Freitag et al., 2021).[9] For CONTRATICO, we evaluate whether ASKQE effectively detects errors, is more sensitive to critical over minor errors, and compare outputs to existing QE methods (§6). With the BIOMQM dataset, we compare ASKQE outputs with human error annotations and benchmark its performance against existing QE methods (§7).

Language Pairs. To align with practical settings, we select language pairs that are high in demand in the United States healthcare system (Khoong et al., 2019; AMN Healthcare Language Services, 2021; Diamond et al., 2019; Taira et al., 2021). For CONTRATICO, we use English-Spanish (EN-ES), French (EN-FR), Hindi (EN-HI), Tagalog (EN-TL), and Chinese (EN-ZH) and for BIOMQM, we use English-German (EN-DE), Spanish (EN-ES), French (EN-FR), Russian (EN-RU), and Chinese (EN-ZH). We fix English as the source language to reflect real-world scenarios where non-English monolinguals rely on translated COVID-19 content originally written in English.

5.2 Models

QG/QA. We benchmark five English-centric, open-weights LLMs across model sizes and families: LLAMA-3 8B and 70B (Grattafiori et al., 2024), GEMMA-2 9B and 27B (Team et al., 2024), and YI-1.5 9B (AI et al., 2024). Since all prompting in the ASKQE pipeline is conducted in English, we hypothesize that English-centric models are better suited for this task than multilingual models.[10]

Backtranslation. We use the Google Translate API for backtranslation due to its efficiency in both time and computation. While backtranslation can introduce noise and has been shown to be an un-

[8] Note that we do not consider general MT benchmarks, since our focus is specifically on high-stakes contexts.

[9] Detailed dataset statistics are in Appendix G.1.

[10] HuggingFace model names are in Appendix Table 4.

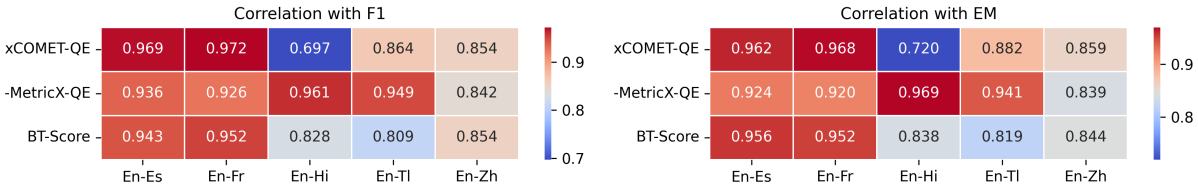


Figure 4: Pearson correlation coefficients between the F1 and EM metrics of AskQE and three established QE metrics: xCOMET-QE, MetricX-QE, and BT-Score. We take negatives of MetricX-QE. AskQE exhibits strong correlations with all three QE metrics, confirming its effectiveness as a QE measure. Full numerical results are provided in Appendix E.1.

reliable standalone measure of translation quality (Agrawal et al., 2022), we believe that it minimizes noise relative to other options, as supported by: **1)** a reasonable average xCOMET-QE score (0.748) between the perturbed translation Y_{tgt} and its corresponding backtranslation Y_{bt} (Appendix C.2), and **2)** superior performance compared to a cross-lingual QA system, particularly in distinguishing minor from critical errors (Appendix C.3).[11]

6 Results on ContraTICO

We first optimize the AskQE metric by evaluating five LLMs across three QG variants and selecting Llama-3 70b (Grattafiori et al., 2024) with NLI as the best-performing configuration (§6.1). Next, we validate this setup by demonstrating that metric scores decrease for more severe perturbations (§6.2) and exhibits strong correlations with established QE metrics (§6.3).

6.1 Optimizing AskQE

In Appendix Tables 25 to 29, we present detailed results of AskQE by evaluating across five LLMs across three QG variants (Appendix D.2), resulting in a total of 15 configurations. We identify the best-performing configuration based on the following criteria: **1)** Minimal AskQE metric score differences *within* the same error severity level (Minor, Critical); **2)** Large AskQE metric score differences *between* different error severity levels; **3)** High Correlations with existing QE metrics (Appendix E.1); **4)** Strong performance in desiderata evaluation, confirming the quality of the generated questions (Appendix F). Each configuration is ranked across these axes, and we select Llama-3 70b with NLI as the best-performing method.

6.2 AskQE can Detect Critical MT Errors

We first validate whether the optimized AskQE with Llama-3 70b can effectively differentiate between different levels of perturbation (Minor, Critical). As shown in Figure 3, AskQE consistently assigns lower scores to critical errors compared to minor ones. Among the eight linguistic perturbations examined, spelling errors yield the highest AskQE scores (Avg. F1: 0.815, Avg. EM: 0.682), while alterations show the lowest scores (Avg. F1: 0.496, Avg. EM: 0.384). This trend is consistent across all five language pairs and five metrics used for computing AskQE. In Appendix Tables 20 and 21, we provide qualitative examples for each perturbation level for the En-Es and En-Zh language pair.

6.3 AskQE Correlates Well with QE Metrics

To validate that AskQE can function as other established QE metrics, we measure Pearson correlation coefficients between AskQE and three existing QE metrics: **1)** xCOMET-QE, **2)** MetricX-QE, and **3)** BT-score (§2.2). As shown in Figure 4, both F1 and EM scores of AskQE exhibit strong, statistically significant correlations with all three QE metrics. The average correlations are 0.871 and 0.878 with xCOMET-QE, -0.923 and -0.919 with MetricX-QE, and 0.877 and 0.882 with BT-Score for F1 and EM, respectively. We report the raw scores for each QE metric in Appendix E.2.

Overall, our empirical results in the simulated setting are promising indicators that AskQE effectively detects critical translation errors (§6.2) and correlates well with established QE metrics (§6.3).

6.4 Question Analysis

Type Categorization. We categorize the types of questions generated during QG according to the 10 pragmatic function types defined by Cao and Wang (2021), using a few-shot LLM prompting classifier.[12] We present the distribution of question types, along with their definitions and examples in

[11] Additionally, we report the Google Translate quality of MT pair in the TICO-19 dataset in Appendix B.

[12] Following Trienes et al. (2024), we prompt GPT-4o with annotation guidelines and few-shot examples from Cao and Wang (2021) (Appendix I).

Question Type	% Q	Examples
Verification. Asking for the truthfulness of an event or a concept.	15.7	• Are many international borders closed? • Is Europe the new epicenter of the pandemic?
Disjunctive. Asking for the true one given multiple events or concepts, where comparison among options is not needed.	0.71	• Is COVID-19 caused by a virus or bacteria? • Should I wear a mask indoors or outdoors?
Concept. Asking for a definition of an event or a concept.	23.6	• What kind of restrictions have countries imposed on arriving travelers? • What is another name for paracetamol?
Extent. Asking for the extent or quantity of an event or a concept.	24.6	• How many COVID-19 cases were reported today? • In how many countries has the healthcare system been stretched?
Example. Asking for example(s) or instance(s) of an event or a concept.	1.88	• What types of events are being canceled worldwide? • What are the sources for further information on the coronavirus outbreak?
Comparison. Asking for comparison among multiple events or concepts.	1.68	• How do countries differ in their testing practices? • What is the difference between formal education and informal education?
Cause. Asking for the cause or reason for an event or a concept.	14.4	• What is the cause of COVID-19? • What is the cause of the severe shortage of test kits in many countries?
Consequence. Asking for the consequences or results of an event.	2.72	• What are serious complications of the disease? • What happens to the people who were sitting near an infected person?
Procedural. Asking for the procedures, tools, or methods by which a certain outcome is achieved.	14.0	• What hygiene practices should be followed on a plane? • What steps is the Office of Emergency Services directed to take?
Judgmental. Asking for the opinions of the answerer's own.	0.66	• Should hourly data be our current approach? • Can you still experience events without traveling?

Table 1: Example questions from ASKQE classified according to the question taxonomy from Cao and Wang (2021).

Table 1. The two most common question types are **1) Extent:** questions asking about the extent or quantity of an event or a concept (24.6%), and **2) Concept:** questions seeking the definition of an event or a concept (23.6%). This distribution suggests a strong domain effect, as the TICO-19 dataset primarily consists of COVID-19-related articles and public announcements (Appendix Table 19). These texts frequently report numerical data (e.g., case numbers, percentages) and define key medical and epidemiological terms, naturally leading to a higher proportion of Extent (e.g., "*How many COVID-19 cases were reported today*?") and Concept (e.g., "*What is another name for paracetamol*?") questions. Conversely, there are fewer Judgmental (0.66%) and Disjunctive (0.71%) questions, which aligns with the fact that news articles and official announcements tend to be objective and neutral, rather than expressing subjective opinions or presenting binary choices.

Quality. We further conduct a fine-grained analysis to quantitatively assess the quality of the questions generated by ASKQE. To this end, we define five quality desiderata that measures the correctness, diversity, readability, and answerability of the questions, as detailed in Appendix Table 23. On average, 3.37 questions are generated per instance[13], with no cases of empty sets or duplicate questions. The diversity of questions within each instance, measured by the average SENTENCEBERT similarity (Reimers and Gurevych, 2019), is 0.634. Readability, measured using the Flesch Reading Ease score (Flesch, 1948), is 65.89 with fairly standard level despite the presence of domain-specific terminology. The average answerability score of each question with respect to the source sentence, measured by SELFCHECKGPT (Manakul et al., 2023), is 92.92, indicating that the generated questions remain highly faithful to the source context.

[13]The average instance length is 24.79 words.

7 Results on BIOMQM

Our initial experiments on the CONTRATICO dataset were conducted in a controlled setting. We extend this analysis with the BIOMQM dataset with more naturally occurring translation errors and additional language pairs.

7.1 ASKQE Generalizes to BIOMQM

Error Detection. Since BIOMQM dataset contains multiple errors (1.95 on average) per segment, we assign the error severity of an MT output based on its highest level error (e.g., if the MT contains both critical and minor errors, it is categorized as critical). As shown in Table 2, the average ASKQE scores with all metrics progressively decreases as error severity increases from **Neutral** to **Critical**. On average, 3.40 questions are generated per instance, with an average diversity score of 0.656 and an answerability score of 91.72, measured using the same metrics as in Section 6.4. We further confirm that this trend holds across individual language pairs, as detailed in Appendix G.2. In sum, our findings confirm that ASKQE effectively distinguishes between different error severity levels

Severity	F1	EM	CHRF	BLEU	SBERT
Neutral	0.736	0.426	78.41	61.63	0.830
Minor	0.720	0.414	76.24	58.44	0.825
Major	0.700	0.406	75.27	56.59	0.823
Critical	0.688	0.382	74.32	54.75	0.819

Table 2: ASKQE scores using different metrics evaluated on BIOMQM per error severity. **SBERT**: SENTENCEBERT.

even with more naturally occurring MT errors.

Correlation. Following Zouhar et al. (2024), we measure segment-level Kendall's τ correlation between QE metrics and human judgments. We use the professional error annotations from BIOMQM, and compute a human judgment score based on the schema from Freitag et al. (2021). We compare the same neural QE metrics against ASKQE as in our CONTRATICO evaluation (§6.3): **1)** Segment-level scores from XCOMET-QE, **2)** METRICX-QE, and **3)** BT-Score. As shown in Figure 5, ASKQE with SENTENCEBERT achieves the highest correlation with human judgments (0.265). This further suggests that ASKQE aligns well with human evaluations of MT quality, motivating its use in a more human-centered application – guiding decisions on whether to accept or reject MT output (§7.2).

7.2 Decision Making Simulation: How Actionable is ASKQE Feedback?

Experiment Setup. We simulate a real-world scenario in which individuals decide whether to accept or reject an MT output Y_{tgt} based on specific QE feedback, as illustrated in Figure 1. To formalize this binary decision-making process, we define a segment-level decision threshold, which depends on the type of QE feedback. When feedback consists of error annotations, such as BIOMQM human judgments or MQM annotations from XCOMET-QE, we apply the following rule, where e represents the highest error severity level:

$$\text{Decision}(Y_{\text{tgt}}) = \begin{cases} \text{Accept}, & \text{if } e \in \{\text{Neutral}, \text{Minor}\} \\ \text{Reject}, & \text{otherwise} \end{cases} \quad (2)$$

Conversely, when feedback is given as a single scalar score, such as outputs from ASKQE, segment scores from XCOMET-QE (DA), METRICX-QE, or BT-Score, we fit a two-component Gaussian Mixture Model (GMM) for each QE metric. This model clusters N segments into "accept" or "reject" groups based on score distribution.[14] We then compare the predicted segment-level decision labels $\hat{l}$ from each QE metric to the BIOMQM human judgment labels l and compute decision accuracy:

$$\text{Decision Acc.(QE)} = \frac{\sum_{N}^{i=1} \mathbb{1}(l_i = \hat{l}_i)}{N} \quad (3)$$

Additionally, for ASKQE, we introduce a simple baseline where decisions are based on the number of mismatches between reference and predicted answer pairs. Specifically, an MT output is rejected if the mismatch count exceeds a predefined threshold.

[14] Detailed description of the process is in Appendix G.3.

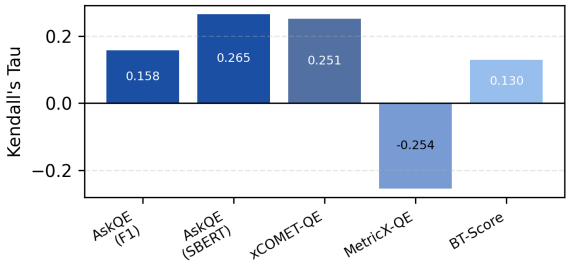


Figure 5: Segment-level correlation (Kendall's τ) between QE metrics and human judgments in the BIOMQM dataset.

QE Metric	Decision Acc. (%)
ASKQE (F1)	**75.75**
ASKQE (SBERT)	63.77
ASKQE (Num. Mismatch=1)	42.16
ASKQE (Num. Mismatch=3)	68.37
ASKQE (Num. Mismatch=5)	74.19
XCOMET-QE (DA)	63.77
XCOMET-QE (MQM)	61.00
METRICX-QE*	59.57
BT-Score	68.85

Table 3: Decision accuracy (%) of each QE metric in the BIOMQM human simulation experiment. Best score in **bold**. *: We subtract the original METRICX-QE score from 25 to align is interpretation with other metrics.

Results. As shown in Table 3, ASKQE achieves decision accuracies comparable to or higher than other QE baselines. Qualitatively, ASKQE using F1 mostly disagrees with human ratings on translations with major linguistic convention errors, such as spelling or mistranslation errors, and disagrees with other QE metrics on minor linguistic convention errors. Compared to other QE metrics, ASKQE is less effective at detecting spelling errors but more effective at identifying translations with stylistic issues (e.g., non-fluent) or mistranslations, leading to the highest decision accuracy (75.75%). We show detailed qualitative analysis in Appendix G.4. Overall, these results highlight ASKQE's potential not only for MT quality assessment but also

as actionable feedback to support real-world decision making in high-stakes contexts.

8 Conclusion

In this work, we introduce ASKQE, a question generation and answering framework designed to identify critical translation errors and provide actionable feedback for users relying on MT outputs in a language they do not understand. Using the CONTRATICO dataset (§4) across five language pairs, we validate the effectiveness of our method for critical error detection (§6.2) and correlation with established QE metrics (§6.3). Analysis of generated questions shows that most focus on extent or definition while remaining faithful to the source context (§6.4). We further show that ASKQE generalizes well to realistic scenarios with naturally occurring translation errors in the BIOMQM dataset, achieving stronger correlations with human judgments (§7.1) and improving MT acceptance decisions compared to other QE metrics (§7.2).

These findings highlight the promise of QG/QA framework for MT quality assessment and QE feedback, calling for future work to explore optimal strategies such as enhancing context integration during QG and expanding to multi-turn QA.

9 Limitations

Our experiment setup is as comprehensive as our computational budget allows, while we could not cover every possible variant. The scope of our study is limited to out-of-English language pairs, as generating questions with LLMs has been more extensively studied in English (Li and Zhang, 2024; Fu et al., 2024b), and using English as the source language benefits performance from its prevalence in LLM training data (Grattafiori et al., 2024). This leaves open questions on how to design an optimal QG/QA framework by exploring various combinations of of LLMs, language pairs, and contextual inputs for QG, which we leave for future work.

As part of the construction process of the CONTRATICO dataset, we introduce several type of perturbations that adds new information to the content of the source. As shown in Figure 3, ASKQE can detect critical perturbation that introduces a new meaning (Expansion (Impact)) but finds it difficult to detect errors that only modify the intensity of an adjective or an adverb (Intensifier) or add contextually implied details without introducing new meaning (Expansion (No Impact)).

Acknowledgments

We thank the anonymous reviewers and the members of the CLIP lab at University of Maryland for their constructive feedback. This work was supported in part by the Human Language Technology Center of Excellence at Johns Hopkins University, by NSF Fairness in AI Grant 2147292, and by the Institute for Trustworthy AI in Law and Society (TRAILS), which is supported by the National Science Foundation under Award No. 2229885. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of NSF or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Sweta Agrawal, Nikita Mehandru, Niloufar Salehi, and Marine Carpuat. 2022. Quality estimation via back-translation at the WMT 2022 quality estimation task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 593–596, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. Yi: Open foundation models by 01.ai. *Preprint*, arXiv:2403.04652.

AMN Healthcare Language Services. 2021. 2021 healthcare world language index: A national and state-by-state listing of the most frequently spoken languages other than english in hospital, medical group, and community health center-based patient encounters.

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitriy Genzel, Franscisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur. 2020. TICO-19: the translation initiative for COvid-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Yotam Berger. 2017. Israel arrests palestinian because facebook translated 'good morning' to 'attack them'. *Ha'aretz*, 22.

Jan Berka, Ondrej Bojar, et al. 2011. Quiz-based evaluation of machine translation. *The Prague Bulletin of Mathematical Linguistics*, 95:77.

Claire Bonial, Olga Babko-Malaya, Jinho D. Choi, Jena Hwang, and Martha Palmer. 2010. *PropBank Annotation Guidelines*.

Shuyang Cao and Lu Wang. 2021. Controllable open-ended question generation with a new question type ontology. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6424–6439, Online. Association for Computational Linguistics.

Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Evaluating question answering evaluation. In *Proceedings of the 2nd workshop on machine reading for question answering*, pages 119–124.

Daniel Deutsch, Tania Bedrax-Weiss, and Dan Roth. 2021. Towards question-answering as an automatic metric for evaluating the content quality of a summary. *Transactions of the Association for Computational Linguistics*, 9:774–789.

Lisa Diamond, Karen Izquierdo, Dana Canfield, Konstantina Matsoukas, and Francesca Gany. 2019. A systematic review of the impact of patient–physician non-english language concordance on quality of care and outcomes. *Journal of General Internal Medicine*, 34(8):1591–1606.

Liam Dugan, Eleni Miltsakaki, Shriyash Upadhyay, Etan Ginsberg, Hannah Gonzalez, DaHyeon Choi, Chuning Yuan, and Chris Callison-Burch. 2022. A feasibility study of answer-agnostic question generation for education. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1919–1926, Dublin, Ireland. Association for Computational Linguistics.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Melda Eksi, Erik Gelbing, Jonathan Stieber, and Chi Viet Vu. 2021. Explaining errors in machine translation with absolute gradient ensembles. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 238–249, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alexander 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, Seattle, United States. Association for Computational Linguistics.

Patrick Fernandes, Sweta Agrawal, Emmanouil Zaranis, André F. T. Martins, and Graham Neubig. 2025. Do llms understand your translations? evaluating paragraph-level mt with question answering. *Preprint*, arXiv:2504.07583.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André Martins, Graham Neubig, Ankush Garg, Jonathan Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1066–1083, Singapore. Association for Computational Linguistics.

Rudolf Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.

Marina Fomicheva, Piyawat Lertvittayakumjorn, Wei Zhao, Steffen Eger, and Yang Gao. 2021. The Eval4NLP shared task on explainable quality estimation: Overview and results. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 165–178, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marina Fomicheva, Lucia Specia, and Nikolaos Aletras. 2022. Translation error detection as rationale extraction. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4148–4159, Dublin, Ireland. Association for Computational Linguistics.

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

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024a. GPTScore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576, Mexico City, Mexico. Association for Computational Linguistics.

Weiping Fu, Bifan Wei, Jianxiang Hu, Zhongmin Cai, and Jun Liu. 2024b. QGEval: Benchmarking multi-dimensional evaluation for question generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11783–11803, Miami, Florida, USA. Association for Computational Linguistics.

Masaru Fuji. 1999. Evaluation experiment for reading comprehension of machine translation outputs.

In *Proceedings of Machine Translation Summit VII*, pages 285–289.

Masaru Fuji, Nobutoshi Hatanaka, Etsuo Ito, Shinichiro Kamei, Hiroyuki Kumai, Tatsuya Sukehiro, Takehiko Yoshimi, and Hitoshi Isahara. 2001. Evaluation method for determining groups of users who find mt "useful". In *Proceedings of Machine Translation Summit VIII*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang,

Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

HyoJung Han, Marine Carpuat, and Jordan Boyd-Graber. 2022. SimQA: Detecting simultaneous MT errors through word-by-word question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5598–5616, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lifeng Han, Alan Smeaton, and Gareth Jones. 2021. Translation quality assessment: A brief survey on manual and automatic methods. In *Proceedings for the First Workshop on Modelling Translation: Translatology in the Digital Age*, pages 15–33, online. Association for Computational Linguistics.

Dongfu Jiang, Yishan Li, Ge Zhang, Wenhao Huang, Bill Yuchen Lin, and Wenhu Chen. 2024. TIGERScore: Towards building explainable metric for all text generation tasks. *Transactions on Machine Learning Research*.

Douglas Jones, Edward Gibson, Wade Shen, Neil Granoien, Martha Herzog, Douglas Reynolds, and Clifford Weinstein. 2005a. Measuring human readability of machine generated text: three case studies in speech recognition and machine translation. In *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 5, pages v–1009. IEEE.

Douglas Jones, Martha Herzog, Hussny Ibrahim, Arvind Jairam, Wade Shen, Edward Gibson, and Michael Emonts. 2007. Ilr-based mt comprehension test with multi-level questions. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 77–80.

Douglas Jones, Wade Shen, Neil Granoien, Martha Herzog, and Clifford Weinstein. 2005b. Measuring translation quality by testing english speakers with a new defense language proficiency test for arabic. In *Proceedings of the 2005 International Conference on Intelligence Analysis*, pages 2–6.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Marzena Karpinska, Nishant Raj, Katherine Thai, Yixiao Song, Ankita Gupta, and Mohit Iyyer. 2022. DEMETR: Diagnosing evaluation metrics for translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9540–9561, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Elaine C. Khoong, Eric Steinbrook, Cortlyn Brown, and Alicia Fernandez. 2019. Assessing the use of google translate for spanish and chinese translations of emergency department discharge instructions. *JAMA Internal Medicine*, 179(4):580–582.

Paul Kingsbury and Martha Palmer. 2002. From TreeBank to PropBank. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

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, Singapore. Association for Computational Linguistics.

Mateusz Krubiński, Erfan Ghadery, Marie-Francine Moens, and Pavel Pecina. 2021. Just ask! evaluating machine translation by asking and answering questions. In *Proceedings of the Sixth Conference on Machine Translation*, pages 495–506, Online. Association for Computational Linguistics.

Kunze Li and Yu Zhang. 2024. Planning first, question second: An LLM-guided method for controllable question generation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4715–4729, Bangkok, Thailand. Association for Computational Linguistics.

Tania Lombrozo and Daniel A. Wilkenfeld. 2019. *Mechanistic versus functional understanding*, chapter 11. New York, NY: Oxford University Press.

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, Bangkok, Thailand. Association for Computational Linguistics.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

Nikita Mehandru, Sweta Agrawal, Yimin Xiao, Ge Gao, Elaine Khoong, Marine Carpuat, and Niloufar Salehi. 2023. Physician detection of clinical harm in machine translation: Quality estimation aids in reliance and backtranslation identifies critical errors. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11633–11647, Singapore. Association for Computational Linguistics.

Nikita Mehandru, Samantha Robertson, and Niloufar Salehi. 2022. Reliable and safe use of machine translation in medical settings. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 2016–2025, New York, NY, USA. Association for Computing Machinery.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Pow-

ell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

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.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Arij Riabi, Thomas Scialom, Rachel Keraron, Benoît Sagot, Djamé Seddah, and Jacopo Staiano. 2021. Synthetic data augmentation for zero-shot cross-lingual question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7016–7030, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Raphael Rubino, Atsushi Fujita, and Benjamin Marie. 2021. Error identification for machine translation with metric embedding and attention. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 146–156, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ananya B. Sai, Tanay Dixit, Dev Yashpal Sheth, Sreyas Mohan, and Mitesh M. Khapra. 2021. Perturbation CheckLists for evaluating NLG evaluation metrics. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7219–7234, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Carolina Scarton and Lucia Specia. 2016. A reading comprehension corpus for machine translation evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3652–3658, Portorož, Slovenia. European Language Resources Association (ELRA).

Siamak Shakeri, Noah Constant, Mihir Kale, and Linting Xue. 2021. Towards zero-shot multilingual synthetic question and answer generation for cross-lingual reading comprehension. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 35–45, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Lucia Specia, Carolina Scarton, and Gustavo Henrique Paetzold. 2018. *Quality estimation for machine translation*. Morgan & Claypool Publishers.

Joe Stacey, Pasquale Minervini, Haim Dubossarsky, Oana-Maria Camburu, and Marek Rei. 2024. Atomic inference for NLI with generated facts as atoms. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10188–10204, Miami, Florida, USA. Association for Computational Linguistics.

Kyoshiro Sugiyama, Masahiro Mizukami, Graham Neubig, Koichiro Yoshino, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2015. An investigation of machine translation evaluation metrics in cross-lingual question answering. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 442–449, Lisbon, Portugal. Association for Computational Linguistics.

Breena R. Taira, Vanessa Kreger, Aristides Orue, and Lisa C. Diamond. 2021. A pragmatic assessment of google translate for emergency department instructions. *Journal of General Internal Medicine*, 36(11):3361–3365.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Masaru Tomita, Masako Shirai, Junya Tsutsumi, Miki Matsumura, and Yuki. 1993. Evaluation of MT systems by TOEFL. In *Proceedings of the Fifth Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, Kyoto, Japan.

Jan Trienes, Sebastian Joseph, Jörg Schlötterer, Christin Seifert, Kyle Lo, Wei Xu, Byron Wallace, and Junyi Jessy Li. 2024. InfoLossQA: Characterizing and recovering information loss in text simplification. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4263–4294, Bangkok, Thailand. Association for Computational Linguistics.

Lucas Nunes Vieira, Minako O'Hagan, and Carol O'Sullivan. 2021. Understanding the societal impacts of machine translation: A critical review of the literature on medical and legal use cases. *Information, Communication & Society*, 24(11):1515–1532.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Sandra Weiss and Lars Ahrenberg. 2012. Error profiling for evaluation of machine-translated text: a polish-english case study. In *LREC*, pages 1764–1770.

Naomi Yamashita, Rieko Inaba, Hideaki Kuzuoka, and Toru Ishida. 2009. Difficulties in establishing common ground in multiparty groups using machine translation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, page 679–688, New York, NY, USA. Association for Computing Machinery.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Vilém Zouhar and Ondřej Bojar. 2020. Outbound translation user interface ptakopět: A pilot study. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6967–6975, Marseille, France. European Language Resources Association.

Vilém Zouhar, Shuoyang Ding, Anna Currey, Tatyana Badeka, Jenyuan Wang, and Brian Thompson. 2024. Fine-tuned machine translation metrics struggle in unseen domains. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 488–500, Bangkok, Thailand. Association for Computational Linguistics.

A Prompts

A.1 Perturbation

We show prompt templates for each perturbation types for English-Spanish (EN-ES) language pair. For other target languages, the 2-shot examples are replaced accordingly.

Prompt A.1.1: Perturbation (Synonym)

Task: You will be given a `{target lang}` sentence. Your goal is to perturb the sentence by replacing one or two words (noun, verb, adjective or adverb) to its synonym. Please make sure it does not changes the meaning in a significant way. Output only the perturbed `{target lang}` sentence without giving any additional explanation.

*** Example Starts ***
Original: Puede tratarse de un miembro de la secretaría o del personal clínico, según el protocolo de cada consultorio.
Perturbed: Puede tratarse de un miembro de la secretaría o del personal médico, según el protocolo de cada consultorio.

Original: Además, reclutaremos nuevos consultorios de monitoreo.
Perturbed: Asimismo, contrataremos nuevos consultorios de seguimiento.
*** Example Ends ***

Original: `{original}`
Perturbed:

Prompt A.1.2: Perturbation (Word Order)

Task: You will be given a `{target lang}` sentence. Your goal is to perturb the sentence by changing the word order. Please make sure it does not changes the meaning in a significant way. Output only the perturbed `{target lang}` sentence without giving any additional explanation.

*** Example Starts ***
Original: Puede tratarse de un miembro de la secretaría o del personal clínico, según el protocolo de cada consultorio.
Perturbed: Puede tratarse de un miembro del personal clínico o de la secretaría, según el protocolo de cada consultorio.

Original: Desarrollaremos un observatorio para presentar los datos a nivel nacional, así como un tablero de control para hacer observaciones a los consultorios sobre la calidad de sus datos y la recolección de muestras virológicas y serológicas.
Perturbed: Se desarrollará un observatorio para presentar los datos a nivel nacional, junto con un tablero de control destinado a proporcionar observaciones a los consultorios sobre la calidad de los datos y la recolección de muestras virológicas y serológicas.
*** Example Ends ***

Original: `{original}`
Perturbed:

Prompt A.1.3: Perturbation (Spelling)

Task: You will be given a `{target lang}` sentence. Your goal is to perturb the sentence by misspelling one or two words. The misspelled words should be important in the sentence. Output only the perturbed `{target lang}` sentence without giving any additional explanation.

*** Example Starts ***
Original: Puede tratarse de un miembro de la secretaría o del personal clínico, según el protocolo de cada consultorio.
Perturbed: Puede tratarse de un miembro de la secretaría o del persnal clínico, según el protcolo de cada consultorio.

Original: Además, reclutaremos nuevos consultorios de monitoreo.
Perturbed: Además, reclutaremos nuevos cosultorios de monitoreo.
*** Example Ends ***

Original: `{original}`
Perturbed:

Prompt A.1.4: Perturbation (Intensifier)

Task: You will be given a `{target lang}` sentence. Your goal is to perturb the sentence by adding one or two words that changes the intensity of the existing word. Please make sure that the added word does not disturb the grammaticality of the sentence. Output only the perturbed `{target lang}` sentence without giving any additional explanation.

*** Example Starts ***
Original: Los síntomas comunes incluyen fiebre, tos seca y fatiga.
Perturbed: Los síntomas comunes incluyen fiebre alta, tos seca y fatiga.

Original: La dificultad para respirar, el dolor de garganta, los dolores de cabeza, las molestias corporales o la producción de esputo se encuentran entre los otros síntomas menos comunes.
Perturbed: La dificultad para respirar severa, el dolor de garganta, los dolores de cabeza, las molestias corporales o la producción de esputo se encuentran entre los otros síntomas menos comunes.
*** Example Ends ***

Original: `{original}`
Perturbed:

Prompt A.1.5: Perturbation (Expansion (No Impact))

Task: You will be given a `{target lang}` sentence. Your goal is to perturb the sentence by adding one or two words in the sentence. Please make sure that the added word does not disturb the grammaticality of the sentence and does not changes the meaning in a significant way. The added words would add more context that was already obvious from the sentence. Output only the perturbed `{target lang}` sentence without giving any additional explanation.

*** Example Starts ***
Original: si crees que tus síntomas o problemas justifican un examen más detallado.
Perturbed: si crees que tus síntomas o problemas justifican un examen médico más detallado.

Original: En caso de respuesta afirmativa a estas preguntas de filtro, se debe pedir al paciente que no acuda al consultorio y que siga el esquema de PHE en su lugar.
Perturbed: En caso de respuesta afirmativa a estas preguntas de filtro, se debe pedir al paciente adulto que no acuda al consultorio y que siga el esquema de PHE en su lugar.
*** Example Ends ***

Original: `{original}`
Perturbed:

Prompt A.1.6: Perturbation (Expansion (Impact))

Task: You will be given a `{target lang}` sentence. Your goal is to perturb the sentence by adding words in the sentence. Please make sure that the added word does not disturb the grammaticality of the sentence but should change the meaning in a significant way. Output only the perturbed `{target lang}` sentence without giving any additional explanation.

*** Example Starts ***
Original: Los síntomas comunes incluyen fiebre, tos seca y fatiga.
Perturbed: Los síntomas comunes incluyen fiebre y dolores musculares, tos seca y fatiga.

Original: La dificultad para respirar, el dolor de garganta, los dolores de cabeza, las molestias corporales o la producción de esputo se encuentran entre los otros síntomas menos comunes.
Perturbed: La dificultad para respirar, el dolor de garganta, los dolores de cabeza, las molestias corporales, la producción de esputo y trastornos digestivos se encuentran entre los otros síntomas menos comunes.
*** Example Ends ***

Original: `{original}`
Perturbed:

Prompt A.1.7: Perturbation (Omission)

Task: You will be given a `{target lang}` sentence. Your goal is to perturb the sentence by removing information from the sentence. Remove only one or two words from the sentence. Please make sure that the removed information does not disturb the grammaticality of the sentence but should change the meaning in a significant way. Output only the perturbed `{target lang}` sentence without giving any additional explanation.

*** Example Starts ***
Original: Los síntomas comunes incluyen fiebre, tos seca y fatiga.
Perturbed: Los síntomas comunes incluyen fatiga y fatiga.

Original: Se están realizando investigaciones sobre una vacuna o un tratamiento antiviral específico.
Perturbed: Se están realizando investigaciones sobre un tratamiento antiviral específico.
*** Example Ends ***

Original: `{original}`
Perturbed:

Prompt A.1.8: Perturbation (Alteration)

Task: You will be given a `{target lang}` sentence. Your goal is to perturb the sentence by changing the affirmative sentence into negation (vice versa) or changing one word (noun, verb, adjective or adverb) to its antonym. Please make sure that the perturbation does not disturb the grammaticality of the sentence but should change the meaning in a significant way. Output only the perturbed `{target lang}` sentence without giving any additional explanation.

*** Example Starts ***
Original: No logró aliviar el dolor con medicamentos que la competencia prohíbe a los participantes.
Perturbed: No logró aliviar el placer con medicamentos que la competencia prohíbe a los participantes.

Original: El mes pasado, un comité presidencial recomendó la renuncia del antiguo CEP como parte de medidas para llevar al país a nuevas elecciones.
Perturbed: El mes pasado, un comité presidencial no recomendó la renuncia del antiguo CEP como parte de medidas para llevar al país a nuevas elecciones.
*** Example Ends ***

Original: `{original}`
Perturbed:

Prompt A.2.2: Question Generation (NLI)

Task: You will be given an English sentence and a list of atomic facts, which are short sentences conveying one piece of information. Your goal is to generate a list of relevant questions based on the sentence. Output the list of questions in Python list format without giving any additional explanation.

*** Example Starts ***
Sentence: but if you have the cough
Atomic facts: ["You have the cough."]
Questions: ["What condition is being referred to?"]

Sentence: The number of accessory proteins and their function is unique depending on the specific coronavirus.
Atomic facts: ["The number of accessory proteins is unique depending on the specific coronavirus.", "The function of accessory proteins is unique depending on the specific coronavirus."]
Questions: ["What is unique depending on the specific coronavirus?", "What is unique about the function of accessory proteins?"]
*** Example Ends ***

Sentence: `{sentence}`
Atomic facts: `{atomic facts}`
Questions:

A.2 Question Generation (QG)

Prompt A.2.1: Question Generation (Vanilla)

Task: You will be given an English sentence. Your goal is to generate a list of relevant questions based on the sentence. Output only the list of questions in Python list format without giving any additional explanation.

*** Example Starts ***
Sentence: but if you have the cough
Questions: ["What condition is being referred to?"]

Sentence: The number of accessory proteins and their function is unique depending on the specific coronavirus.
Questions: ["What is unique depending on the specific coronavirus?", "What is unique about the function of accessory proteins?"]
*** Example Ends ***

Sentence: `{sentence}`
Questions:

Prompt A.2.3: Question Generation (SRL)

Task: You will be given an English sentence and a dictionary of semantic roles in the sentence. Your goal is to generate a list of relevant questions based on the sentence. Output the list of questions in Python list format without giving any additional explanation.

*** Example Starts ***
Sentence: but if you have the cough
Semantic roles: {'Verb': 'have', 'ARG0': 'you', 'ARG1': 'the cough'}
Questions: ["What condition is being referred to?"]

Sentence: The number of accessory proteins and their function is unique depending on the specific coronavirus.
Semantic roles: {'Verb': 'is', 'ARG1': 'The number of accessory proteins and their function', 'MNR': 'unique', 'TMP': 'depending on the specific coronavirus'}
Questions: ["What is unique depending on the specific coronavirus?", "What is unique about the function of accessory proteins?"]
*** Example Ends ***

Sentence: `{sentence}`
Semantic roles: `{semantic roles}`
Questions:

Model	HuggingFace Model Name
GEMMA-2 9B	`google/gemma-2-9b-it`
GEMMA-2 27B	`google/gemma-2-27b-it`
LLAMA-3 8B	`meta-llama/Llama-3.1-8B-Instruct`
LLAMA-3 70B	`meta-llama/Llama-3.1-70B-Instruct`
YI-1.5 9B	`01-ai/Yi-1.5-9B-Chat`

Table 4: HuggingFace model names for all tested LLMs.

Language Pair	XCOMET-QE
EN-ES	0.886
EN-FR	0.855
EN-HI	0.564
EN-TL	0.704
EN-ZH	0.730

Table 5: Google Translate quality for the TICO-19 dataset. We measure the XCOMET-QE scores between the source sentence and the reference translation.

A.3 Question Answering (QA)

Prompt A.3: Question Answering

Task: You will be given an English sentence and a list of relevant questions. Your goal is to generate a list of answers to the questions based on the sentence. Output only the list of answers in Python list format without giving any additional explanation.

*** Example Starts ***
Sentence: and does this pain move from your chest?
Questions: ["What moves from your chest?", "Where does the pain move from?"]
Answers: ["The pain", "Your chest"]

Sentence: Diabetes mellitus (784, 10.9%), chronic lung disease (656, 9.2%), and cardiovascular disease (647, 9.0%) were the most frequently reported conditions among all cases.
Questions: ["What were the most frequently reported conditions among all cases?", "What percentage of cases reported diabetes mellitus?", "What percentage of cases reported chronic lung disease?", "What percentage of cases reported cardiovascular disease?"]
Answers: ["Diabetes mellitus, chronic lung disease, and cardiovascular disease", "10.9%", "9.2%", "9.0%"]
*** Example Ends ***

Sentence: `{sentence}`
Questions: `{questions}`
Answers:

B Translation Quality of TICO-19

As shown in Table 5, we measure Google Translate quality of source sentences in the TICO-19 dataset by computing XCOMET-QE scores between the source and reference translation. On average, translation quality is lower for lower-resource languages (EN-HI and EN-TL) than for higher-resource languages (EN-ES, EN-FR, and EN-ZH).

C Effectiveness of ASKQE Design

C.1 Perturbation

We evaluate perturbation effectiveness by measuring XCOMET-QE (Guerreiro et al., 2024) and cosine similarity scores between the original source X_{src} and the perturbed translation Y_{tgt} in Table 17 (*right*). Lower metric scores for more severe perturbations provide empirical support for the validity of our perturbation strategy.

C.2 Backtranslation Quality

We evaluate the XCOMET-QE scores between the perturbed translation Y_{tgt} and its corresponding backtranslation Y_{bt}, as shown in Table 17 (*left*). Higher scores indicate better backtranslation quality. On average, XCOMET-QE scores are 0.757, 0.782, 0.671, 0.762, and 0.767 for EN-ES, EN-FR, EN-HI, EN-TL, and EN-ZH, respectively, resulting in an overall average of 0.748, which suggests a reasonable level of backtranslation quality.

C.3 Comparison to Cross-lingual QA System

We compared the monolingual English QA system used in ASKQE to a cross-lingual QA system that bypasses backtranslation and directly generates answers using the original source X_{src} and the perturbed translation Y_{tgt}. Since the outputs are in different languages, we use a neural metric SENTENCEBERT instead of string-based comparison metrics. As shown in Table 18, our English QA system outperforms the cross-lingual QA system in differentiating between minor and critical errors, for all language pairs. These results further support the importance of incorporating backtranslation into our pipeline.

C.4 Impact of Backtranslation on Critical Errors

We hypothesize that with a high-quality MT system, it is unlikely to mask errors present in the original MT. To test this, we conduct an error analysis to examine whether any critical errors in the original MT go undetected because they are erased during backtranslation. For each language pair, we prompt GPT-4 with the reference translation Y_{ref}, the perturbed translation Y_{tgt}, and its backtranslation Y_{bt}, asking whether the specific error introduced in the perturbed translation remains identifiable in the backtranslation. As shown in Table 6, the average percentage of backtranslations in which the critical

Language pair	Error Type	Identified (%)
EN-ES	Expansion (Impact)	90.42
	Omission	86.86
	Alteration	87.02
EN-FR	Expansion (Impact)	88.88
	Omission	89.94
	Alteration	91.25
EN-HI	Expansion (Impact)	86.32
	Omission	87.78
	Alteration	93.85
EN-TL	Expansion (Impact)	87.65
	Omission	82.78
	Alteration	86.11
EN-ZH	Expansion (Impact)	85.55
	Omission	85.68
	Alteration	92.67

Table 6: Error analysis on the impact of backtranslation on critical errors per language pair.

error remains detectable is high: 87.76% for Expansion (No Impact), 86.61% for Omission, and 90.18% for Alteration errors.

D Details of ASKQE

D.1 Detailed Results

From Tables 25 to 29, we present detailed results of ASKQE using different metrics: F1, EM, CHRF, BLEU, and SENTENCEBERT per language pair, severity level, and perturbation type.

D.2 Variants of QG

We decompose the standard QG process and extract relevant information from the source sentence to enhance the prompt with additional context. We introduce three variants: **1)** Vanilla, **2)** NLI, and **3)** SRL. When no supplementary information is provided beyond the source sentence, we refer to this approach as **Vanilla**. The **NLI**-based process is detailed in Section 3.1. Further, we incorporate semantic role labeling (**SRL**) to generate questions that target the most important parts of the source sentence. We prompt GPT-4o to annotate semantic roles in the source following the PropBank framework (Kingsbury and Palmer, 2002; Palmer et al., 2005), covering both the core and non-core roles (Bonial et al., 2010). We then prompt an LLM to generate questions given the source that focus on the extracted semantic roles. On average, 1.56 core roles are labeled per instance. Detailed distribution of non-core roles are outlined in Table 7.

Role	Description	Count
TMP	Temporal	303
LOC	Locative	227
MNR	Manner	223
PRP	Purpose or Reason	127
MOD	Modality	116
CAU	Causal	47
COM	Comitative	29
DIR	Directional	27
GOL	Goal	14
EXT	Extent	14
NEG	Negation	7
CON	Conditional	1

Table 7: Dataset distribution of non-core roles in SRL.

D.3 Effect of Entailment Classification

As detailed in §3.1, each instance in the CONTRATICO dataset yields an average of 3.61 facts, of which 3.08 are retained after entailment filtering, using a DEBERTA-V3 L (340M) NLI classifier. This component is lightweight, requiring 01:13 (MM:SS) on a single NVIDIA RTX A5000 GPU to process 971 English sentences. To assess the impact of skipping this step, we conduct an ablation where all generated facts are used directly during QA without NLI filtering. As shown in Table 8, removing the entailment classification step consistently lowers average F1 and EM scores and reduces the score gap between minor and critical errors. This demonstrates that entailment filtering enhances both overall QA accuracy and sensitivity to error severity, justifying its inclusion despite minimal computational cost.

Severity	Perturbation	F1 (w/o)	F1 (w/)	EM (w/o)	EM (w/)
Minor	Spelling	0.819	0.840	0.675	0.704
	Word Order	0.811	0.820	0.634	0.679
	Synonym	0.809	0.816	0.635	0.676
	Intensifier	0.798	0.812	0.621	0.655
	Expansion (No Impact)	0.790	0.814	0.622	0.665
Critical	Expansion (Impact)	0.776	0.809	0.618	0.653
	Omission	0.743	0.746	0.603	0.602
	Alteration	0.468	0.475	0.329	0.360

Table 8: Average F1 and EM scores with **(w/)** and without **(w/o)** entailment classification step during the NLI-based question generation process.

E Correlation Analysis

E.1 Detailed Results

As a sanity check for ASKQE, we conduct a correlation analysis with three established QE metrics: **1)** XCOMET-QE (Guerreiro et al., 2024) and **2)** METRICX-QE (Juraska et al., 2024) between the source X_{src} and the perturbed MT output Y_{tgt}, and

3) BT-score (Agrawal et al., 2022) between the source X_{src} and the backtranslated MT output Y_{bt}. We use BERTSCORE (Zhang et al., 2020) as the MT metric for **3)**. As shown in Tables 30 to 34, we observe that ASKQE exhibit strong, statistically significant correlations with all three QE baselines, supporting the validity of our approach. Among the 15 LLM configurations, LLAMA-3 70B with NLI achieves the highest average correlation.

E.2 QE Metric Scores

We report the raw scores for each QE metrics (XCOMET-QE, METRICX-QE, BT-Score) in Table 22 for each language pair and perturbation. We observe that the QE metric scores decrease as the severity level increases, showing similar trends as ASKQE scores.

F Desiderata Evaluation

We perform a fine-grained analysis of each LLM configuration to better understand the quality of the generated questions from ASKQE. A detailed description of each desideratum is provided in Table 23, and results across all configurations are reported in Table 24. On average, we observe that 3.15 questions are generated per instance, with an answerability rate of 89.51%, indicating that the questions are both faithful and answerable using the source sentence. Based on these findings, along with the metric scores of ASKQE, we select LLAMA-3 70B with NLI as our best-performing method, which shows the highest diversity and answerability scores.

G Details of BioMQM

G.1 Dataset Statistics

We use BIOMQM (Zouhar et al., 2024) as our testbed, a biomedical domain MT dataset containing abstracts of scientific publication and medical texts with error annotations by professional translators. We detail dataset statistics of BIOMQM per error category and subcategory in Table 9, per language pair and error severity in Table 10.

We preprocess the development split by filtering instances where: **1)** the source sentence contains no errors and **2)** the source language is English. This results in a total of 5,216 instances with an average instance length is 22.45 words across five language pairs: English-German (EN-DE), English-Spanish (EN-ES), English-French (EN-FR), English-Russian (EN-RU), and English-Chinese (EN-ZH).

Category	Subcategory	Count
Linguistic Conventions	Grammar	555
	Spelling	1575
	Punctuation	393
	Register	11
Accuracy	Mistranslation	1297
	Untranslated	228
	Addition	44
Terminology	Inconsistent	134
	Wrong term	771
Locale Conventions	Date Format	9
	Number Format	38
	Measurement Format	76
Style	Non Fluent	1447
Unintelligible	-	88
Other	-	71
Total		**6737**

Table 9: Dataset statistics for BIOMQM (Zouhar et al., 2024) per error category and subcategories.

G.2 Detailed Results

We present detailed results of ASKQE evaluated on BIOMQM by language pair and error severity in Table 11. To ensure fair comparisons, we match the number of instances per row to the minimum available count. Given the limited number of **Critical** errors, we merge **Major** and **Critical** errors (**C+M**) for analysis. Our findings from CONTRATICO generalize to BIOMQM, confirming that ASKQE detects errors across different severity levels.

G.3 Human Simulation: GMM

For each QE metric baseline that outputs a single scalar score, we fit a two-component Gaussian Mixture Model (GMM).[15] GMM is a probabilistic clustering model that assumes data points are generated from a mixture of Gaussian distributions, assigning a probability to each data point for belonging to a specific cluster. We hypothesize two clusters: **1)** high-quality translations with higher scores and **2)** low-quality translations with lower scores. Each segment in the BIOMQM dataset is assigned a probability of belonging to the low-score cluster (rejection). Figure 7 illustrates the distribution of QE metric scores relative to their probability of rejection.

[15] https://scikit-learn.org/stable/modules/mixture.html

Language pair	Error Severity	Count
EN-DE	Neutral	50
	Minor	572
	Major	1520
	Critical	13
EN-ES	Neutral	169
	Minor	1376
	Major	74
	Critical	94
EN-FR	Neutral	14
	Minor	597
	Major	140
	Critical	11
EN-RU	Neutral	216
	Minor	205
	Major	74
	Critical	20
EN-ZH	Neutral	61
	Minor	1210
	Major	318
	Critical	3
Total		**6737**

Table 10: Dataset statistics for BIOMQM (Zouhar et al., 2024) per language pair and error severity level.

G.4 Detailed Qualitative Analysis

In this section, we present the detailed qualitative findings from the human simulation study as briefly discussed in §7.2. As shown in Tables 12 and 13, ASKQE (F1) tends to disagree with human ratings primarily on major errors in terms of error severity, particularly spelling errors, followed by mistranslation errors in terms of error type. Further in Table 14, we compare ASKQE to other QE metrics by highlighting the top-3 error types where ASKQE is most effective and the bottom-3 where it is least effective. We find that AskQE is less effective at detecting spelling errors, but more effective at identifying issues related to style and mistranslation.

H Computational & Time Efficiency

We compare the average computational, time, and cost efficiency for computing ASKQE and baseline QE methods, as shown in Table 15, on 971 EN-ES sentence pairs from CONTRATICO. Notably, ASKQE using LLAMA-3 70B achieves faster runtime than XCOMET-QE, and fact extraction step costs less than $1 using GPT-4o due to the short average sentence length (24.79 words). We also propose two lower-cost variants of ASKQE: **1)** Using smaller models for question answering (§D.1) and **2)** Substituting GPT-4o with lighter variants for question generation (§D.2). Having established this, future works could focus on optimizing and deploying such a tool, particularly in user-facing applications. ASKQE could be deployed selectively, for instance using a standard segment-level QE score as a first pass and apply ASKQE's fine-grained evaluation only to translations that fall below a certain quality threshold.

Language	Severity	F1	EM	CHRF	BLEU	SBERT
EN-DE	Neutral	0.791	0.500	85.30	69.96	0.900
	Minor	0.785	0.529	82.92	68.26	0.854
	C+M	0.721	0.431	77.94	60.94	0.848
EN-ES	Neutral	0.785	0.549	81.16	69.00	0.853
	Minor	0.735	0.431	79.12	60.98	0.827
	C+M	0.728	0.418	77.33	61.01	0.825
EN-FR	Neutral	0.843	0.567	87.51	62.93	0.829
	Minor	0.721	0.486	77.25	62.90	0.825
	C+M	0.695	0.389	74.99	57.93	0.605
EN-RU	Neutral	0.694	0.340	73.59	54.35	0.821
	Minor	0.703	0.363	74.08	56.32	0.698
	C+M	0.648	0.238	67.92	47.37	0.603
EN-ZH	Neutral	0.705	0.404	75.60	56.13	0.812
	Minor	0.661	0.317	71.40	51.00	0.808
	C+M	0.618	0.296	67.07	48.15	0.790

Table 11: Disaggregated results of ASKQE evaluated on BIOMQM dataset per language pair and error severity. Note that we combine **Major** and **Critical** errors since the number of comparisons are very limited for **Critical** (shown as **C+M**).

Error Severity	Disagreement Count
Neutral	70
Minor	751
Major	2102
Critical	138

Table 12: Disagreement count for human ratings and ASKQE (F1) evaluated on BIOMQM dataset per error severity.

Category	Subcategory	Disagreement Count
Linguistic Conventions	Grammar	153
	Spelling	1073
	Punctuation	87
Accuracy	Mistranslation	641
	Untranslated	72
	Addition	11
	Unspecified	88
Terminology	Inconsistent	44
	Wrong term	344
Locale Conventions	Date Format	4
	Number Format	20
	Measurement Format	14
Style	Non Fluent	490
Other	-	20

Table 13: Disagreement count for human ratings and ASKQE (F1) evaluated on BIOMQM dataset per error category and subcategories.

QE Metric	Top-3	Bottom-3
xCOMET-QE (DA)	(1) Accuracy / Mistranslation / 401 (2) Style / Non-fluent / 248 (3) Linguistic conventions / Spelling / 237	(1) Linguistic conventions / Spelling / 344 (2) Accuracy / Mistranslation / 275 (3) Style / Non-fluent / 87
xCOMET-QE (MQM)	(1) Accuracy / Mistranslation / 421 (2) Style / Non-fluent / 194 (3) Linguistic conventions / Spelling / 156	(1) Linguistic conventions / Spelling / 249 (2) Accuracy / Mistranslation / 81 (3) Linguistic conventions / Punctuation / 73
MetricX-QE	(1) Style / Non-fluent / 537 (2) Accuracy / Mistranslation / 419 (3) Linguistic conventions / Spelling / 301	(1) Linguistic conventions / Spelling / 373 (2) Accuracy / Mistranslation / 245 (3) Style / Non-fluent / 226
BT-Score	(1) Accuracy / Mistranslation / 197 (2) Style / Non-fluent / 128 (3) Linguistic conventions / Spelling / 82	(1) Linguistic conventions / Spelling / 218 (2) Accuracy / Mistranslation / 172 (3) Style / Non-fluent / 72

Table 14: Top-3 and Bottom-3 error types between AskQE and other QE metrics. **Top-3:** AskQE is most effective; **Bottom-3:** AskQE is least effective. Each cell presents in the format as error category/subcategory/count.

Method	Model (size)	Computation (GPU)	Avg. Time (*hh:mm*)	Avg. Cost ($)
xCOMET-QE	xCOMET-XL (3B)	1 × NVIDIA RTX A5000	02:06	0
MetricX-QE	MetricX-24 XL (3B)	1 × NVIDIA RTX A5000	00:02	0
BT-Score	DeBERTa-v3 L (340M)	1 × NVIDIA RTX A5000	00:02	0
AskQE	GPT-4o for fact extraction	-	00:20	0.979
AskQE	LLaMA-3 7B	8 × NVIDIA RTX A5000	01:16	0

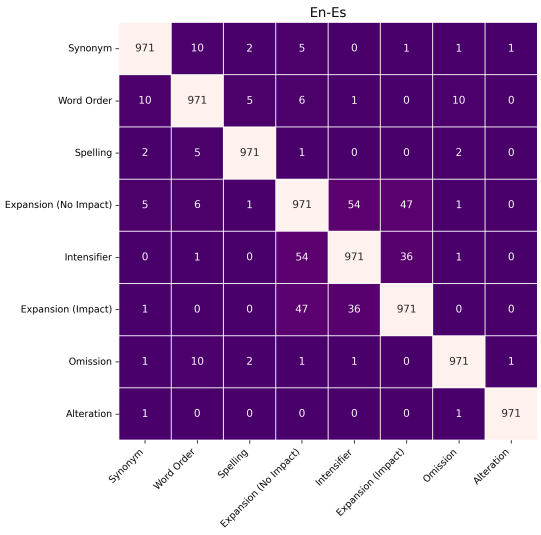
Table 15: Average computational, time, and cost efficiency for tested methods. GPT-4o pricing is based on current OpenAI API rates at https://openai.com/api/pricing/.

Severity	Perturbation	Example
Minor	Spelling	**Original:** Assurez-vous que toutes les informations et tous les conseils que vous obtenez ont été confirmés par des médecins et scientifiques de bonne réputation. **Perturbed:** Assurez-vous que toutes les informations et tous les conseils que vous obtenez ont été confirmés par des *médins* et *scientfiques* de bonne réputation.
	Word Order	**Original:** Dans la mesure du possible, évitez les endroits très fréquentés. **Perturbed:** Dans la mesure du possible, *les endroits très fréquentés doivent être évités.*
	Synonym	**Original:** mais j'ai le rhume des foins aussi **Perturbed**: mais j'ai le rhume des *prés* aussi
	Intensifier	**Original:** Le 11-mars-2020, l'Organisation mondiale de la santé déclarait la maladie à coronavirus-2019 (COVID-19) comme étant une pandémie. **Perturbed:** Le 11 mars 2020, l'Organisation mondiale de la santé déclarait la *grave* maladie à coronavirus 2019 (COVID-19) comme étant une pandémie.
	Expansion (No Impact)	**Original:** si vous pensez que vos symptômes ou problèmes justifient un examen plus approfondi. **Perturbed:** si vous pensez que vos symptômes ou problèmes justifient un examen *médical* plus approfondi.
Critical	Expansion (Impact)	**Original:** Les symptômes courants comprennent la fièvre, une toux sèche et la fatigue. **Perturbed**: Les symptômes courants comprennent la fièvre et des douleurs musculaires, *une toux sèche et la fatigue*.
	Omission	**Original:** Des recherches sur *un vaccin ou* un traitement antiviral spécifique sont en cours. **Perturbed**: Des recherches sur un traitement antiviral spécifique sont en cours.
	Alteration	**Original:** Il n'a pas réussi à soulager la douleur avec des médicaments que la compétition interdit aux concurrents de prendre. **Perturbed**: Il n'a pas réussi à soulager le *plaisir* avec des médicaments, que la compétition interdit aux concurrents de prendre.

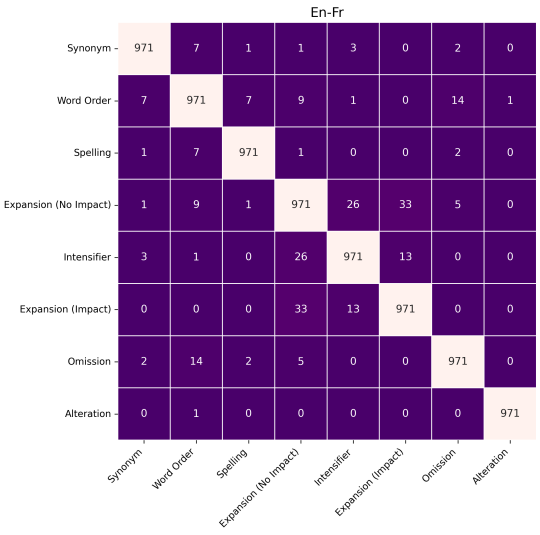
Table 16: Detailed examples for each perturbation type, divided by severity (Minor, Critical) for French (Fr) as target language. Perturbed parts of the reference translation (**Original**) are in *italic* in **Perturbed**.

Language Pair	Severity	Perturbation	xCOMET(X_{src}, Y_{tgt})	SIM(X_{src}, Y_{tgt})	xCOMET(Y_{tgt}, Y_{bt})
EN-ES	Original	-	0.762	0.908	-
	Minor	Spelling	0.758	0.900	0.766
		Word Order	0.738	0.899	0.756
		Synonym	0.725	0.897	0.749
		Intensifier	0.740	0.886	0.760
		Expansion (No Impact)	0.748	0.886	0.760
	Critical	Expansion (Impact)	0.708	0.868	0.766
		Omission	0.619	0.877	0.764
		Alteration	0.473	0.863	0.728
EN-FR	Original	-	0.786	0.897	-
	Minor	Spelling	0.784	0.891	0.789
		Word Order	0.766	0.889	0.784
		Synonym	0.751	0.886	0.775
		Intensifier	0.765	0.880	0.787
		Expansion (No Impact)	0.774	0.878	0.791
	Critical	Expansion (Impact)	0.736	0.858	0.790
		Omission	0.635	0.865	0.784
		Alteration	0.547	0.842	0.756
EN-HI	Original	-	0.632	0.870	-
	Minor	Spelling	0.630	0.858	0.677
		Word Order	0.609	0.849	0.669
		Synonym	0.619	0.864	0.671
		Intensifier	0.624	0.852	0.675
		Expansion (No Impact)	0.617	0.849	0.674
	Critical	Expansion (Impact)	0.574	0.826	0.672
		Omission	0.348	0.821	0.686
		Alteration	0.459	0.835	0.642
EN-TL	Original	-	0.725	0.898	-
	Minor	Spelling	0.739	0.887	0.770
		Word Order	0.722	0.878	0.764
		Synonym	0.714	0.887	0.755
		Intensifier	0.726	0.873	0.768
		Expansion (No Impact)	0.716	0.882	0.762
	Critical	Expansion (Impact)	0.693	0.851	0.771
		Omission	0.609	0.848	0.766
		Alteration	0.581	0.858	0.743
EN-ZH	Original	-	0.767	0.868	-
	Minor	Spelling	0.764	0.861	0.774
		Word Order	0.726	0.854	0.763
		Synonym	0.724	0.851	0.756
		Intensifier	0.763	0.857	0.776
		Expansion (No Impact)	0.749	0.853	0.772
	Critical	Expansion (Impact)	0.718	0.840	0.778
		Omission	0.613	0.815	0.772
		Alteration	0.583	0.829	0.741

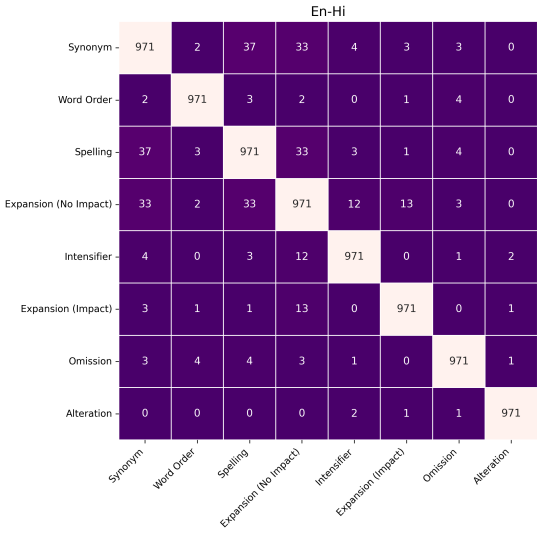
Table 17: Perturbation and backtranslation quality measured per language pair. **xCOMET(X_{src}, Y_{tgt}):** xCOMET-QE scores between the original source and the perturbed translation (↓ as more severe perturbation). **SIM(X_{src}, Y_{tgt}):** Cosine similarity scores between the original source and the perturbed translation (↓ as more severe perturbation). **xCOMET(Y_{tgt}, Y_{bt}):** xCOMET-QE scores between the perturbed translation and the respective backtranslation (↑). The first two metrics evaluate the perturbation effectiveness, while the last one assesses backtranslation quality. For Original, we replace Y_{tgt} to Y_{ref}.



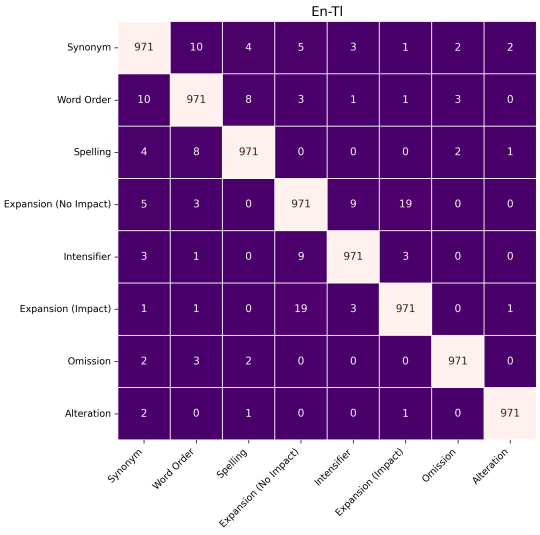
(a) EN-ES



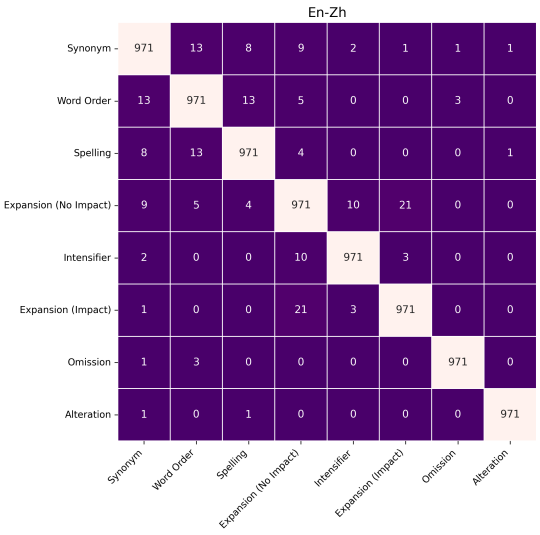
(b) EN-FR



(c) EN-HI



(d) EN-TL



(e) EN-ZH

Figure 6: Overlap ratio between perturbed translations from different perturbation types. We show that there are mostly no overlaps.with one exception of Expansion (No Impact) and Intensifier.

Language	Severity	Perturbation	English QA	Cross-lingual QA
EN-ES	Minor	Spelling	0.916	0.263
		Word Order	0.904	0.254
		Synonym	0.904	0.255
		Intensifier	0.900	0.251
		Expansion (No Impact)	0.897	0.247
	Critical	Expansion (Impact)	0.896	0.241
		Omission	0.864	0.244
		Alteration	0.762	0.232
EN-FR	Minor	Spelling	0.907	0.248
		Word Order	0.902	0.239
		Synonym	0.892	0.242
		Intensifier	0.897	0.234
		Expansion (No Impact)	0.903	0.232
	Critical	Expansion (Impact)	0.890	0.226
		Omission	0.863	0.228
		Alteration	0.729	0.228
EN-HI	Minor	Spelling	0.907	0.167
		Word Order	0.866	0.156
		Synonym	0.879	0.138
		Intensifier	0.893	0.141
		Expansion (No Impact)	0.901	0.139
	Critical	Expansion (Impact)	0.879	0.134
		Omission	0.825	0.132
		Alteration	0.809	0.136
EN-TL	Minor	Spelling	0.906	0.254
		Word Order	0.902	0.238
		Synonym	0.900	0.232
		Intensifier	0.900	0.227
		Expansion (No Impact)	0.898	0.230
	Critical	Expansion (Impact)	0.888	0.228
		Omission	0.847	0.211
		Alteration	0.778	0.209
EN-ZH	Minor	Spelling	0.880	0.158
		Word Order	0.879	0.162
		Synonym	0.876	0.165
		Intensifier	0.879	0.156
		Expansion (No Impact)	0.876	0.157
	Critical	Expansion (Impact)	0.872	0.147
		Omission	0.817	0.130
		Alteration	0.758	0.138

Table 18: ASKQE scores using SENTENCEBERT for English and cross-lingual QA system.

Source	Domain	# Sent	Avg. Length
CMU	Medical, Conversational	379	8.86
PubMed	Medical, Scientific (COVID-19 related articles)	263	22.98
Wikinews	News (COVID-19 related news articles)	21	20.10
Wikivoyage	Travel (Summary of travel restrictions)	243	18.61
Wikipedia	General (COVID-19 related articles)	383	23.57
Wikisource	Announcements (Government/organization announcements)	24	29.25
Total		**971**	**24.79**

Table 19: Dataset statistics for TICO-19 evaluation dataset (Anastasopoulos et al., 2020). We categorize the development split according to data source. **# Sent:** Number of sentences; **Avg. Length:** Average length of sentences.

Severity	Perturbation	X_{src}	Y_{tgt}	Q_{src}	A_{src}	A_{bt}
Minor	Spelling	it is right in the center of my chest	es justo en el centro de mi puedo	['Where is it in relation to your chest?', 'Is it exactly in the center of your chest?']	['In the center of my chest', 'Yes']	['In the center of my can', 'No']
	Word Order	Disinfect the surfaces with a suitable disinfectant, such as diluted household bleach.	Desinfecte las superficies con lejía de uso doméstico diluida, un desinfectante adecuado.	['What disinfectant should be used to disinfect the surfaces?', 'What should be used to disinfect the surfaces?']	['A suitable disinfectant', 'A suitable disinfectant, such as diluted household bleach']	['Diluted household bleach', 'Diluted household bleach']
	Synonym	Remove it from the back, throw it away, and then wash your hands.	Quíteselo por la parte trasera, deséchelo y límpiese las manos.	['What should you do after throwing it away?', 'What should you do after removing it from the back?']	['Wash your hands', 'Throw it away']	['Clean your hands', 'Discard it']
	Intensifier	or if you have high blood pressure	o si tiene presión arterial muy alta	['What might does he have?']	['high blood pressure']	['very high blood pressure']
	Expansion (No Impact)	With the COVID-19 outbreak, PHE and RCGP RSC have adapted existing influenza surveillance to monitor the spread of COVID-19 in the community, and this protocol sets out the basis for that collaboration.	Con el brote de la COVID-19, PHE y el RSC del RCGP han adaptado la actual vigilancia de la influenza para controlar la propagación de la COVID-19 en la comunidad, y este protocolo conjunto sienta las bases para esa colaboración.	['What is the basis for the collaboration between PHE and RCGP RSC?', 'Why have PHE and RCGP RSC adapted existing influenza surveillance?', 'What is the purpose of adapting existing influenza surveillance?', 'What is the collaboration between PHE and RCGP RSC about?']	['This protocol', 'The COVID-19 outbreak', 'To monitor the spread of COVID-19 in the community', 'The spread of COVID-19 in the community']	['The joint protocol', 'The outbreak of COVID-19', 'To monitor the spread of COVID-19 in the community', 'Monitoring the spread of COVID-19 in the community']
Critical	Expansion (Impact)	mandatory real-name registration for public transit cards	registro obligatorio con el nombre real y la dirección para tarjetas de transporte público	['Is real-name registration mandatory for public transit cards?', 'What is mandatory for public transit cards?']	['Yes', 'Real-name registration']	['Yes', 'Registration with real name and address']
	Omission	The genome size for coronaviruses ranges from 26.4 to 31.7 kilobases.	El tamaño del genoma de los coronavirus varía de 31,7 kilobases.	['What is the range of genome size for coronaviruses?', 'What is the lower limit of genome size for coronaviruses?', 'What is the upper limit of genome size for coronaviruses?']	['26.4 to 31.7 kilobases', '26.4 kilobases', '31.7 kilobases']	['31.7 kilobases', '31.7 kilobases', '31.7 kilobases']
	Alteration	Among these 7,162 cases, 2,692 (37.6%) patients had one or more underlying health condition or risk factor, and 4,470 (62.4%) had none of these conditions reported.	Entre estos 7162 casos, 2692 (37,6 %) pacientes no tenían una o más afecciones médicas subyacentes o factores de riesgo, y 4470 (62,4 %) tenían alguna de estas afecciones informadas.	['How many cases are there?', 'How many patients had one or more underlying health condition or risk factor?', 'What percentage of patients had one or more underlying health condition or risk factor?', 'How many patients had none of these conditions reported?', 'What percentage of patients had none of these conditions reported?']	['7,162', '2,692', '37.6%', '4,470', '62.4%']	['7162', '4470', '62.4%', '2692', '37.6%']

Table 20: Qualitative Examples of questions and answer pairs generated by ASKQE for EN-ES language pair. X_{src}: Source sentence; Y_{tgt}: Perturbed MT output; Q_{src}: Questions generated from source; A_{src}: Answers generated from source; A_{bt}: Answers generated from the backtranslated MT output.

Severity	Perturbation	X_{src}	Y_{tgt}	Q_{src}	A_{src}	A_{bt}
Minor	Spelling	Thailand is using an app and SIM cards for all travelers to enforce their quarantine.	泰国利用手机应用程序和SIN 卡对所有游客实施隔离。	['What is Thailand using for all travelers?', 'What is Thailand using to enforce travelers' quarantine?']	['An app and SIM cards', 'An app and SIM cards']	['A mobile app and SIN cards', 'A mobile app and SIN cards']
	Word Order	Some patients have very mild symptoms, similar to a cold.	症状较轻的有些患者，类似感冒。	['What symptoms do some patients have?', 'What are these symptoms similar to?']	['Very mild symptoms', 'A cold']	['Milder symptoms', 'A cold']
	Synonym	Phase II trials are used to establish an initial reading of efficacy and further explore safety in small numbers of people having the disease targeted by the NCE.	II 期试验用于确定疗效的初始读数，并进一步探索对于一些NCE 目标人群的安全性。	['What is the purpose of Phase II trials?', 'What is further explored in Phase II trials?', 'How many people are involved in Phase II trials?']	['To establish an initial reading of efficacy and further explore safety', 'Safety', 'Small numbers']	['An initial readout of efficacy and to further explore safety', 'Safety', 'Unknown']
	Intensifier	Many of these underlying health conditions are common in the United States: based on self-reported 2018 data, the prevalence of diagnosed diabetes among U.S. adults was 10.1%, and the U.S. age-adjusted prevalence of all types of heart disease (excluding hypertension without other heart disease) was 10.6% in 2017.	许多这些基础疾病在美国很常见：根据2018 年自我报告的数据，美国成年人中诊断为高糖尿病的患病率为10.1%，美国所有类型心脏病（不伴其他心脏病的高血压除外）的年龄调整患病率在2017 年为10.6%。	['What are the underlying health conditions that are common in the United States?', 'What percentage of U.S. adults have diagnosed diabetes based on self-reported 2018 data?', 'What was the age-adjusted prevalence of all types of heart disease in the U.S. in 2017?']	['Diabetes and heart disease', '10.1%', '10.6%']	['Diagnosed high diabetes and all types of heart disease', '10.1%', '10.6%']
	Expansion (No Impact)	now i send you an image	现在我发给您一张详细图像	['What do you send?', 'When do you send the image?']	['An image', 'Now']	['A detailed image', 'Now']
Critical	Expansion (Impact)	mandatory real-name registration for public transit cards	registro obligatorio con el nombre real y la dirección para tarjetas de transporte público	['Is real-name registration mandatory for public transit cards?', 'What is mandatory for public transit cards?']	['Yes', 'Real-name registration']	['Yes', 'Registration with real name and address']
	Omission	Others in early-stage Phase II trials or numerous treatment candidates in Phase I trials, are also excluded.	早期II 期试验的其他药物或I 期试验中的大量候选药物也不含在此列。	['Who are excluded from early-stage Phase II trials?', 'Who are excluded from Phase I trials?']	['Others', 'Numerous treatment candidates']	['Other drugs', 'Large numbers of drug candidates']
	Alteration	Phylogentically, mouse hepatitis virus (Murine coronavirus), which infects the mouse's liver and the central nervous system, is related to human coronavirus OC43 and bovine coronavirus.	从生物系统上来说，感染小鼠肝脏和中枢神经系统的小鼠肝炎病毒（鼠冠状病毒），与人冠状病毒OC43 和猪冠状病毒有关。	['What is the relationship between mouse hepatitis virus and human coronavirus OC43?', 'What is the relationship between mouse hepatitis virus and bovine coronavirus?', 'What organs does mouse hepatitis virus infect in a mouse?']	['Mouse hepatitis virus is phylogenetically related to human coronavirus OC43', 'Mouse hepatitis virus is phylogenetically related to bovine coronavirus', 'The liver and the central nervous system"]	['Mouse hepatitis virus is related to human coronavirus OC43', 'There is no relationship mentioned between mouse hepatitis virus and bovine coronavirus', 'The liver and central nervous system']

Table 21: Qualitative Examples of questions and answer pairs generated by ASKQE for EN-ZH language pair. X_{src}: Source sentence; Y_{tgt}: Perturbed MT output; Q_{src}: Questions generated from source; A_{src}: Answers generated from source; A_{bt}: Answers generated from the backtranslated MT output.

Language	Severity	Perturbation	XCOMET-QE (↑)	METRICX-QE (↓)	BT-Score (↑)
EN-ES	Minor	Spelling	0.926	1.832	0.924
		Word Order	0.910	2.044	0.916
		Synonym	0.882	3.325	0.925
		Intensifier	0.908	2.175	0.917
		Expansion (No Impact)	0.904	2.394	0.920
	Critical	Expansion (Impact)	0.885	2.519	0.905
		Omission	0.870	3.462	0.901
		Alteration	0.712	6.510	0.871
EN-FR	Minor	Spelling	0.910	2.020	0.926
		Word Order	0.889	2.324	0.912
		Synonym	0.845	3.682	0.917
		Intensifier	0.889	2.320	0.920
		Expansion (No Impact)	0.873	2.676	0.920
	Critical	Expansion (Impact)	0.842	2.897	0.904
		Omission	0.830	3.730	0.903
		Alteration	0.573	6.718	0.857
EN-HI	Minor	Spelling	0.630	3.334	0.924
		Word Order	0.631	3.946	0.902
		Synonym	0.611	3.237	0.927
		Intensifier	0.595	3.359	0.921
		Expansion (No Impact)	0.571	3.605	0.920
	Critical	Expansion (Impact)	0.534	4.141	0.899
		Omission	0.517	4.599	0.884
		Alteration	0.491	5.391	0.890
EN-TL	Minor	Spelling	0.771	2.548	0.939
		Word Order	0.730	2.830	0.918
		Synonym	0.738	3.463	0.943
		Intensifier	0.734	2.804	0.928
		Expansion (No Impact)	0.723	3.274	0.933
	Critical	Expansion (Impact)	0.687	3.336	0.911
		Omission	0.705	4.085	0.902
		Alteration	0.593	5.887	0.893
EN-ZH	Minor	Spelling	0.839	1.626	0.894
		Word Order	0.818	2.113	0.877
		Synonym	0.802	2.929	0.874
		Intensifier	0.794	1.741	0.887
		Expansion (No Impact)	0.782	1.823	0.887
	Critical	Expansion (Impact)	0.781	1.899	0.874
		Omission	0.768	2.428	0.863
		Alteration	0.666	3.856	0.850

Table 22: Average metric scores for each of three QE metric: XCOMET-QE, METRICX-QE, BT-Score.

Desiderata	Level	Description
Empty	I	Number of empty questions.
Duplicate	I	Number of duplicated questions.
Diversity	I	Output diversity of questions measured by average Sentence-BERT similarity (Reimers and Gurevych, 2019).
Answerability	Q	Answerability of question given the source measured by SelfCheckGPT (Manakul et al., 2023).
Readability	Q	Readability of question measured by Flesch Reading Ease score (Flesch, 1948).
Answerability	Q	Answerability of question given the source measured by SelfCheckGPT (Manakul et al., 2023).

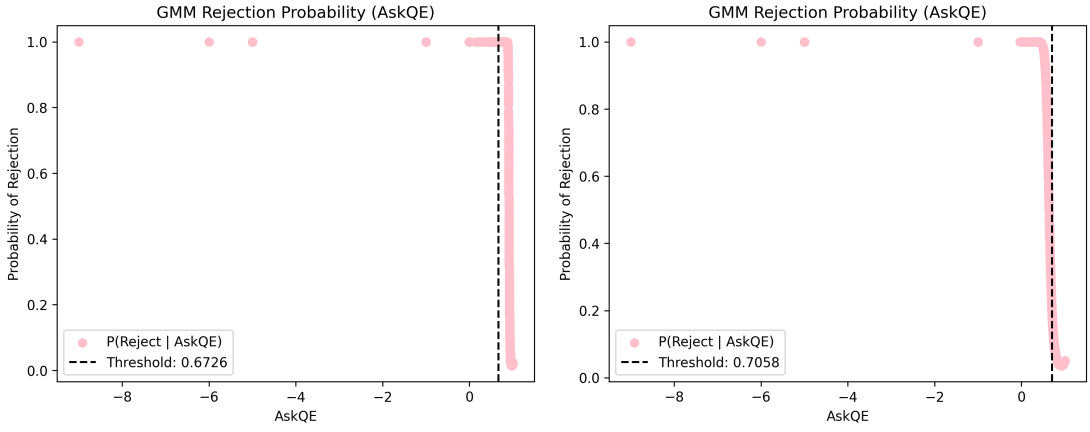
Table 23: Six quality desiderata. **Level:** Level of measurement (**I:** Instance, **Q:** Question-level).

Model	Variant	Avg.# Q	Empty (↓)	Duplicate (↓)	Diversity (↑)	Answerability (↑)	Readability (↓)
GEMMA-2 9B	Vanilla	3.04	0	0	0.534	88.70	66.44
	NLI	2.92	0	0	0.559	90.33	68.28
	SRL	2.86	0	0	0.559	91.17	66.87
GEMMA-2 27B	Vanilla	2.82	0	0	0.539	90.25	68.25
	NLI	2.43	0	1	0.531	91.19	69.41
	SRL	2.81	0	0	0.500	90.27	71.57
LLAMA-3 8B	Vanilla	2.18	0	0	0.597	84.88	69.73
	NLI	4.21	0	1	0.633	90.19	**60.92**
	SRL	4.21	0	1	0.575	87.32	66.57
LLAMA-3 70B	Vanilla	5.03	0	0	0.520	86.14	68.98
	NLI	3.37	0	0	**0.634**	**92.92**	65.89
	SRL	5.05	0	1	0.574	88.31	66.78
YI-1.5 9B	Vanilla	2.82	0	0	0.569	90.42	63.96
	NLI	2.78	0	2	0.586	89.97	66.09
	SRL	3.33	0	0	0.586	90.59	63.62

Table 24: Desiderata evaluation for each LLM configuration (**Model, Variant**). LLAMA-3 70B with NLI has the highest diversity and answerability score.

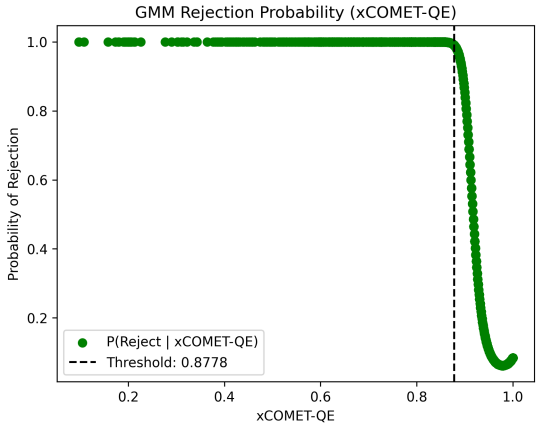
Language	Severity	Perturbation	F1	EM	CHRF	BLEU	SBERT
EN-ES	Minor	Spelling	0.697 / 0.739 / 0.726	0.413 / 0.430 / 0.429	72.89 / 76.58 / 75.80	57.18 / 60.89 / 60.09	0.862 / 0.888 / 0.877
		Word Order	0.678 / 0.715 / 0.712	0.389 / 0.402 / 0.410	71.41 / 74.58 / 74.37	54.60 / 57.64 / 57.73	0.850 / 0.874 / 0.869
		Synonym	0.677 / 0.720 / 0.706	0.388 / 0.399 / 0.406	70.36 / 73.88 / 73.20	55.03 / 58.39 / 57.99	0.848 / 0.874 / 0.864
		Intensifier	0.665 / 0.697 / 0.690	0.341 / 0.358 / 0.358	70.69 / 73.60 / 72.60	51.60 / 54.82 / 53.88	0.845 / 0.867 / 0.858
		Expansion (No Impact)	0.663 / 0.710 / 0.686	0.346 / 0.367 / 0.358	70.46 / 73.99 / 72.95	52.12 / 55.34 / 54.43	0.848 / 0.871 / 0.857
	Critical	Expansion (Impact)	0.645 / 0.689 / 0.670	0.315 / 0.340 / 0.333	68.82 / 72.70 / 70.94	48.79 / 52.38 / 50.92	0.830 / 0.859 / 0.844
		Omission	0.625 / 0.670 / 0.645	0.336 / 0.352 / 0.343	65.93 / 69.53 / 67.44	49.42 / 52.85 / 51.22	0.818 / 0.843 / 0.826
		Alteration	0.577 / 0.605 / 0.597	0.298 / 0.293 / 0.294	62.42 / 64.64 / 63.78	44.83 / 46.64 / 46.31	0.779 / 0.803 / 0.798
EN-FR	Minor	Spelling	0.693 / 0.730 / 0.881	0.410 / 0.415 / 0.420	72.36 / 75.26 / 74.87	56.72 / 59.51 / 59.25	0.859 / 0.864 / 0.876
		Word Order	0.669 / 0.689 / 0.860	0.379 / 0.365 / 0.381	69.54 / 71.56 / 72.46	54.08 / 54.00 / 55.15	0.845 / 0.860 / 0.863
		Synonym	0.668 / 0.702 / 0.864	0.379 / 0.374 / 0.378	70.14 / 71.75 / 71.27	53.32 / 56.09 / 55.77	0.845 / 0.881 / 0.858
		Intensifier	0.667 / 0.699 / 0.868	0.350 / 0.363 / 0.350	70.01 / 72.94 / 72.62	52.17 / 54.79 / 53.83	0.846 / 0.866 / 0.862
		Expansion (No Impact)	0.666 / 0.695 / 0.866	0.367 / 0.351 / 0.358	70.09 / 72.69 / 72.16	52.76 / 54.00 / 53.99	0.845 / 0.868 / 0.863
	Critical	Expansion (Impact)	0.641 / 0.668 / 0.846	0.334 / 0.329 / 0.335	68.34 / 70.84 / 70.59	49.37 / 50.61 / 50.81	0.829 / 0.846 / 0.843
		Omission	0.641 / 0.670 / 0.848	0.346 / 0.345 / 0.342	66.92 / 68.85 / 68.04	50.82 / 52.26 / 51.53	0.831 / 0.848 / 0.840
		Alteration	0.541 / 0.561 / 0.770	0.268 / 0.264 / 0.269	58.17 / 59.87 / 60.75	41.66 / 42.66 / 43.42	0.751 / 0.770 / 0.772
EN-HI	Minor	Spelling	0.704 / 0.737 / 0.732	0.429 / 0.439 / 0.440	74.25 / 76.90 / 76.78	58.99 / 61.96 / 61.37	0.864 / 0.882 / 0.881
		Word Order	0.683 / 0.673 / 0.698	0.387 / 0.375 / 0.385	70.98 / 70.67 / 72.50	55.75 / 53.93 / 56.95	0.821 / 0.836 / 0.841
		Synonym	0.644 / 0.709 / 0.670	0.361 / 0.388 / 0.366	67.84 / 73.42 / 70.62	51.56 / 58.00 / 53.36	0.842 / 0.857 / 0.851
		Intensifier	0.682 / 0.706 / 0.715	0.385 / 0.375 / 0.407	71.87 / 74.57 / 75.12	55.58 / 56.78 / 58.62	0.840 / 0.867 / 0.860
		Expansion (No Impact)	0.666 / 0.718 / 0.692	0.360 / 0.407 / 0.368	70.80 / 75.19 / 73.33	53.20 / 59.10 / 55.54	0.847 / 0.870 / 0.869
	Critical	Expansion (Impact)	0.653 / 0.687 / 0.675	0.351 / 0.366 / 0.357	69.26 / 72.46 / 71.64	51.33 / 53.95 / 53.01	0.825 / 0.845 / 0.843
		Omission	0.605 / 0.634 / 0.613	0.339 / 0.343 / 0.331	63.73 / 66.12 / 64.76	48.51 / 50.91 / 48.91	0.794 / 0.808 / 0.802
		Alteration	0.606 / 0.640 / 0.639	0.327 / 0.331 / 0.335	64.76 / 67.58 / 67.64	48.57 / 51.24 / 51.17	0.795 / 0.818 / 0.819
EN-TL	Minor	Spelling	0.753 / 0.783 / 0.775	0.497 / 0.507 / 0.501	77.31 / 80.11 / 79.80	64.92 / 67.72 / 66.90	0.885 / 0.900 / 0.897
		Word Order	0.718 / 0.729 / 0.747	0.442 / 0.445 / 0.450	73.64 / 75.61 / 76.58	61.20 / 61.23 / 63.71	0.857 / 0.871 / 0.880
		Synonym	0.707 / 0.754 / 0.740	0.440 / 0.457 / 0.454	73.68 / 76.65 / 76.74	59.49 / 64.40 / 62.43	0.860 / 0.881 / 0.881
		Intensifier	0.718 / 0.735 / 0.749	0.417 / 0.401 / 0.430	74.87 / 76.82 / 77.68	59.02 / 60.25 / 61.66	0.864 / 0.881 / 0.883
		Expansion (No Impact)	0.706 / 0.750 / 0.735	0.405 / 0.422 / 0.410	74.48 / 77.55 / 76.99	58.16 / 61.48 / 60.63	0.866 / 0.887 / 0.886
	Critical	Expansion (Impact)	0.690 / 0.729 / 0.725	0.376 / 0.387 / 0.400	72.88 / 76.04 / 75.90	54.95 / 58.17 / 58.37	0.848 / 0.870 / 0.868
		Omission	0.657 / 0.694 / 0.666	0.391 / 0.399 / 0.379	68.09 / 70.29 / 68.70	54.50 / 56.94 / 54.24	0.823 / 0.842 / 0.829
		Alteration	0.620 / 0.648 / 0.657	0.337 / 0.342 / 0.360	65.68 / 68.01 / 69.03	50.27 / 52.64 / 53.70	0.801 / 0.817 / 0.827
EN-ZH	Minor	Spelling	0.592 / 0.613 / 0.620	0.296 / 0.295 / 0.306	64.70 / 65.85 / 66.93	46.12 / 47.42 / 48.43	0.807 / 0.821 / 0.822
		Word Order	0.592 / 0.603 / 0.617	0.298 / 0.294 / 0.311	64.15 / 64.83 / 66.35	45.98 / 46.49 / 48.24	0.804 / 0.812 / 0.819
		Synonym	0.588 / 0.617 / 0.613	0.299 / 0.304 / 0.308	64.10 / 65.93 / 66.58	45.73 / 47.75 / 47.92	0.804 / 0.823 / 0.818
		Intensifier	0.581 / 0.604 / 0.607	0.286 / 0.281 / 0.292	64.07 / 65.67 / 66.52	44.65 / 45.57 / 46.52	0.802 / 0.820 / 0.815
		Expansion (No Impact)	0.579 / 0.611 / 0.603	0.278 / 0.290 / 0.282	63.93 / 65.99 / 65.98	44.16 / 46.53 / 45.79	0.801 / 0.821 / 0.818
	Critical	Expansion (Impact)	0.557 / 0.589 / 0.578	0.252 / 0.265 / 0.263	61.98 / 64.28 / 63.99	40.81 / 43.12 / 42.61	0.783 / 0.806 / 0.797
		Omission	0.537 / 0.562 / 0.554	0.262 / 0.270 / 0.264	58.53 / 60.07 / 59.83	41.30 / 42.93 / 42.35	0.765 / 0.782 / 0.775
		Alteration	0.514 / 0.529 / 0.526	0.233 / 0.229 / 0.224	57.01 / 57.84 / 58.12	38.83 / 39.43 / 39.00	0.747 / 0.763 / 0.761

Table 25: Detailed results of ASKQE using GEMMA-2 9B. Each cell presents scores for Vanilla / SRL / NLI variant. Each metric quantifies the overlap between answers derived from the source sentence (A_{src}) and from the backtranslated MT output (A_{bt}). **SBERT**: SENTENCEBERT.

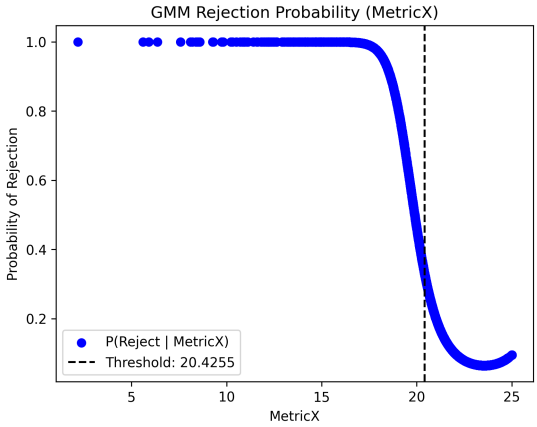


(a) AskQE (F1)

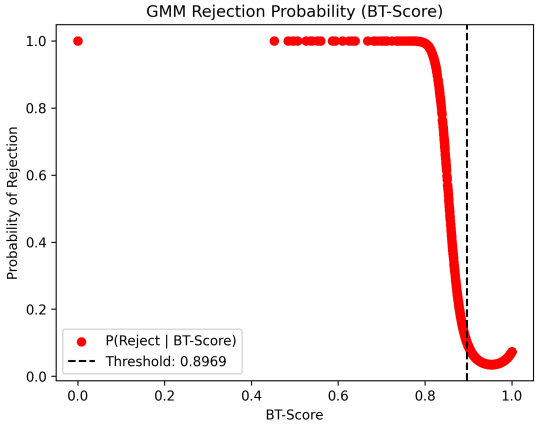
(b) AskQE (SBERT)



(c) xCOMET-QE (DA)



(d) MetricX-QE



(e) BT-Score

Figure 7: Gaussian Mixture Model clustering for each QE metric. x-axis: QE metric scores; y-axis: Probability of rejection.

Language	Severity	Perturbation	F1	EM	CHRF	BLEU	SBERT
EN-ES	Minor	Spelling	0.735 / 0.742 / 0.735	0.434 / 0.425 / 0.409	76.29 / 77.00 / 76.71	60.71 / 60.77 / 59.82	0.886 / 0.897 / 0.895
		Word Order	0.713 / 0.726 / 0.715	0.395 / 0.403 / 0.384	73.70 / 75.50 / 75.16	58.04 / 58.38 / 57.24	0.874 / 0.887 / 0.883
		Synonym	0.711 / 0.723 / 0.713	0.406 / 0.396 / 0.376	74.33 / 74.41 / 74.09	57.55 / 58.67 / 57.37	0.875 / 0.884 / 0.879
		Intensifier	0.701 / 0.714 / 0.704	0.363 / 0.362 / 0.343	73.80 / 74.75 / 74.52	54.79 / 55.46 / 53.92	0.866 / 0.881 / 0.876
		Expansion (No Impact)	0.697 / 0.704 / 0.701	0.362 / 0.354 / 0.341	73.85 / 74.37 / 74.76	55.10 / 54.86 / 54.53	0.868 / 0.881 / 0.877
	Critical	Expansion (Impact)	0.682 / 0.687 / 0.687	0.342 / 0.331 / 0.323	72.22 / 72.75 / 73.06	52.14 / 51.97 / 51.31	0.853 / 0.865 / 0.863
		Omission	0.657 / 0.676 / 0.661	0.347 / 0.353 / 0.334	68.30 / 70.09 / 68.95	51.95 / 53.31 / 51.81	0.834 / 0.851 / 0.841
		Alteration	0.614 / 0.624 / 0.619	0.297 / 0.300 / 0.288	65.52 / 66.82 / 66.67	47.59 / 48.05 / 47.36	0.803 / 0.815 / 0.814
EN-FR	Minor	Spelling	0.720 / 0.727 / 0.726	0.408 / 0.399 / 0.393	74.96 / 75.30 / 75.52	58.63 / 58.82 / 58.38	0.881 / 0.885 / 0.887
		Word Order	0.705 / 0.708 / 0.709	0.381 / 0.377 / 0.372	72.57 / 73.25 / 73.41	56.62 / 55.87 / 55.41	0.869 / 0.874 / 0.876
		Synonym	0.701 / 0.703 / 0.707	0.381 / 0.369 / 0.360	72.88 / 72.19 / 72.81	55.40 / 55.87 / 55.69	0.871 / 0.869 / 0.876
		Intensifier	0.698 / 0.703 / 0.709	0.352 / 0.342 / 0.337	72.93 / 73.53 / 74.03	54.13 / 53.99 / 54.24	0.869 / 0.879 / 0.875
		Expansion (No Impact)	0.693 / 0.703 / 0.697	0.359 / 0.351 / 0.337	73.05 / 73.85 / 73.44	54.32 / 54.68 / 53.63	0.869 / 0.874 / 0.879
	Critical	Expansion (Impact)	0.670 / 0.677 / 0.678	0.337 / 0.325 / 0.320	70.79 / 71.72 / 71.68	50.84 / 51.04 / 50.46	0.844 / 0.854 / 0.856
		Omission	0.656 / 0.669 / 0.659	0.335 / 0.335 / 0.316	67.75 / 68.95 / 68.20	50.98 / 52.13 / 50.46	0.836 / 0.847 / 0.842
		Alteration	0.565 / 0.587 / 0.584	0.264 / 0.270 / 0.258	60.45 / 62.49 / 62.45	42.69 / 44.32 / 43.65	0.765 / 0.784 / 0.783
EN-HI	Minor	Spelling	0.740 / 0.742 / 0.744	0.440 / 0.432 / 0.415	77.81 / 77.96 / 78.08	62.24 / 62.16 / 61.83	0.889 / 0.894 / 0.898
		Word Order	0.713 / 0.708 / 0.714	0.395 / 0.379 / 0.371	74.04 / 73.39 / 74.04	58.60 / 57.64 / 57.66	0.842 / 0.848 / 0.846
		Synonym	0.677 / 0.679 / 0.678	0.371 / 0.369 / 0.350	70.82 / 71.23 / 71.07	53.88 / 54.26 / 53.46	0.864 / 0.863 / 0.865
		Intensifier	0.715 / 0.718 / 0.724	0.403 / 0.394 / 0.388	75.35 / 75.59 / 76.21	58.73 / 58.60 / 58.82	0.870 / 0.873 / 0.875
		Expansion (No Impact)	0.706 / 0.707 / 0.712	0.369 / 0.365 / 0.353	74.89 / 75.02 / 75.21	56.54 / 56.36 / 56.36	0.873 / 0.876 / 0.882
	Critical	Expansion (Impact)	0.687 / 0.690 / 0.699	0.368 / 0.359 / 0.341	72.88 / 73.00 / 73.78	54.38 / 53.85 / 54.14	0.846 / 0.851 / 0.858
		Omission	0.618 / 0.629 / 0.612	0.333 / 0.329 / 0.310	64.84 / 66.15 / 64.60	49.40 / 50.02 / 48.21	0.796 / 0.814 / 0.798
		Alteration	0.638 / 0.658 / 0.657	0.332 / 0.334 / 0.318	67.69 / 69.35 / 69.61	51.17 / 52.41 / 51.97	0.813 / 0.827 / 0.833
EN-TL	Minor	Spelling	0.780 / 0.777 / 0.794	0.497 / 0.477 / 0.484	80.00 / 79.53 / 81.33	67.40 / 66.47 / 67.83	0.899 / 0.901 / 0.910
		Word Order	0.758 / 0.757 / 0.771	0.454 / 0.440 / 0.449	77.04 / 77.15 / 78.47	64.50 / 64.34 / 65.32	0.880 / 0.881 / 0.889
		Synonym	0.742 / 0.737 / 0.745	0.456 / 0.438 / 0.432	76.55 / 76.16 / 77.42	62.54 / 61.48 / 61.84	0.883 / 0.891 / 0.897
		Intensifier	0.750 / 0.747 / 0.767	0.432 / 0.403 / 0.423	77.84 / 77.43 / 79.42	62.13 / 60.72 / 62.85	0.885 / 0.888 / 0.895
		Expansion (No Impact)	0.750 / 0.737 / 0.755	0.426 / 0.388 / 0.395	78.04 / 77.17 / 78.73	62.25 / 59.82 / 61.44	0.886 / 0.892 / 0.901
	Critical	Expansion (Impact)	0.726 / 0.726 / 0.734	0.394 / 0.374 / 0.377	75.67 / 75.92 / 76.78	58.47 / 57.79 / 58.24	0.870 / 0.875 / 0.882
		Omission	0.665 / 0.694 / 0.677	0.382 / 0.386 / 0.372	68.25 / 70.57 / 69.22	54.68 / 56.82 / 55.06	0.818 / 0.846 / 0.830
		Alteration	0.667 / 0.675 / 0.678	0.362 / 0.347 / 0.351	69.60 / 70.86 / 71.17	54.70 / 54.61 / 55.08	0.823 / 0.838 / 0.838
EN-ZH	Minor	Spelling	0.628 / 0.621 / 0.632	0.325 / 0.295 / 0.294	67.21 / 66.27 / 67.81	49.22 / 47.73 / 48.28	0.823 / 0.825 / 0.835
		Word Order	0.620 / 0.617 / 0.626	0.311 / 0.301 / 0.299	66.67 / 66.44 / 67.61	48.51 / 47.66 / 47.99	0.831 / 0.824 / 0.835
		Synonym	0.622 / 0.615 / 0.620	0.308 / 0.292 / 0.281	66.45 / 66.25 / 67.30	48.56 / 47.37 / 47.21	0.828 / 0.830 / 0.840
		Intensifier	0.620 / 0.613 / 0.625	0.305 / 0.278 / 0.282	67.07 / 66.59 / 67.86	47.68 / 46.15 / 46.80	0.820 / 0.830 / 0.840
		Expansion (No Impact)	0.611 / 0.611 / 0.619	0.294 / 0.278 / 0.266	66.44 / 66.60 / 67.58	46.82 / 46.08 / 45.79	0.828 / 0.828 / 0.840
	Critical	Expansion (Impact)	0.595 / 0.590 / 0.599	0.273 / 0.255 / 0.254	65.24 / 64.76 / 66.26	43.94 / 42.60 / 43.07	0.811 / 0.813 / 0.825
		Omission	0.548 / 0.563 / 0.557	0.265 / 0.263 / 0.241	58.80 / 60.17 / 59.62	41.62 / 42.70 / 40.85	0.769 / 0.784 / 0.783
		Alteration	0.550 / 0.544 / 0.545	0.251 / 0.228 / 0.222	59.80 / 59.86 / 60.39	41.48 / 40.12 / 39.74	0.771 / 0.773 / 0.781

Table 26: Detailed results of ASKQE using GEMMA-2 27B. Each cell presents scores for Vanilla / SRL / NLI variant.

Language	Severity	Perturbation	F1	EM	CHRF	BLEU	SBERT
EN-ES	Minor	Spelling	0.696 / 0.703 / 0.715	0.357 / 0.373 / 0.381	72.06 / 72.43 / 73.89	54.32 / 55.68 / 56.73	0.830 / 0.828 / 0.833
		Word Order	0.673 / 0.680 / 0.684	0.326 / 0.346 / 0.352	70.02 / 70.46 / 71.59	50.94 / 52.62 / 52.95	0.832 / 0.831 / 0.839
		Synonym	0.672 / 0.677 / 0.680	0.321 / 0.338 / 0.337	69.38 / 69.47 / 70.04	51.46 / 52.69 / 52.89	0.842 / 0.845 / 0.856
		Intensifier	0.660 / 0.665 / 0.678	0.289 / 0.303 / 0.307	69.32 / 69.24 / 71.02	48.22 / 49.48 / 50.21	0.827 / 0.825 / 0.837
		Expansion (No Impact)	0.650 / 0.662 / 0.664	0.279 / 0.307 / 0.296	68.45 / 69.38 / 70.16	47.82 / 50.07 / 49.36	0.818 / 0.826 / 0.832
	Critical	Expansion (Impact)	0.634 / 0.641 / 0.660	0.263 / 0.275 / 0.289	67.01 / 67.18 / 69.10	44.94 / 46.36 / 47.62	0.805 / 0.806 / 0.823
		Omission	0.611 / 0.638 / 0.637	0.273 / 0.299 / 0.294	63.76 / 65.75 / 65.55	45.15 / 48.22 / 48.04	0.786 / 0.800 / 0.802
		Alteration	0.549 / 0.539 / 0.557	0.210 / 0.209 / 0.218	59.18 / 58.29 / 60.90	38.81 / 38.14 / 39.75	0.733 / 0.723 / 0.744
EN-FR	Minor	Spelling	0.691 / 0.693 / 0.694	0.346 / 0.360 / 0.353	71.22 / 71.56 / 71.91	53.68 / 54.77 / 54.05	0.838 / 0.803 / 0.808
		Word Order	0.668 / 0.671 / 0.672	0.320 / 0.329 / 0.330	69.01 / 68.97 / 69.07	50.10 / 51.92 / 51.35	0.830 / 0.806 / 0.814
		Synonym	0.665 / 0.662 / 0.671	0.311 / 0.317 / 0.329	68.41 / 68.74 / 69.58	50.54 / 50.26 / 50.74	0.826 / 0.820 / 0.830
		Intensifier	0.662 / 0.671 / 0.670	0.284 / 0.303 / 0.298	68.61 / 69.41 / 70.00	48.31 / 50.12 / 49.31	0.823 / 0.800 / 0.812
		Expansion (No Impact)	0.653 / 0.656 / 0.669	0.284 / 0.301 / 0.305	68.24 / 68.63 / 69.95	47.91 / 49.14 / 49.49	0.822 / 0.802 / 0.807
	Critical	Expansion (Impact)	0.628 / 0.630 / 0.643	0.262 / 0.276 / 0.270	66.16 / 66.50 / 67.77	44.64 / 45.58 / 45.67	0.765 / 0.782 / 0.799
		Omission	0.625 / 0.637 / 0.635	0.281 / 0.299 / 0.292	64.05 / 65.47 / 65.57	46.32 / 48.09 / 47.33	0.746 / 0.776 / 0.778
		Alteration	0.516 / 0.501 / 0.517	0.187 / 0.185 / 0.183	56.11 / 54.64 / 55.96	35.72 / 34.92 / 35.71	0.693 / 0.701 / 0.722
EN-HI	Minor	Spelling	0.697 / 0.699 / 0.706	0.369 / 0.383 / 0.381	72.96 / 72.79 / 73.59	55.66 / 56.88 / 56.57	0.796 / 0.763 / 0.768
		Word Order	0.666 / 0.670 / 0.680	0.318 / 0.333 / 0.331	69.03 / 69.21 / 69.83	51.70 / 52.93 / 53.08	0.788 / 0.766 / 0.773
		Synonym	0.643 / 0.644 / 0.644	0.302 / 0.318 / 0.306	67.17 / 67.32 / 67.33	48.25 / 49.39 / 48.48	0.785 / 0.779 / 0.788
		Intensifier	0.674 / 0.681 / 0.686	0.328 / 0.348 / 0.336	70.47 / 70.95 / 71.63	51.86 / 53.63 / 53.09	0.782 / 0.760 / 0.772
		Expansion (No Impact)	0.665 / 0.673 / 0.684	0.303 / 0.321 / 0.317	70.18 / 71.16 / 71.57	50.13 / 51.98 / 51.92	0.781 / 0.761 / 0.767
	Critical	Expansion (Impact)	0.634 / 0.649 / 0.663	0.282 / 0.305 / 0.309	66.78 / 68.11 / 69.83	46.83 / 48.80 / 49.74	0.727 / 0.743 / 0.759
		Omission	0.580 / 0.606 / 0.594	0.265 / 0.289 / 0.280	61.07 / 63.37 / 62.22	43.40 / 46.60 / 45.04	0.708 / 0.737 / 0.739
		Alteration	0.577 / 0.586 / 0.604	0.250 / 0.261 / 0.258	61.88 / 62.65 / 64.18	42.77 / 44.12 / 45.14	0.659 / 0.666 / 0.686
EN-TL	Minor	Spelling	0.748 / 0.746 / 0.759	0.435 / 0.436 / 0.451	76.47 / 76.50 / 77.77	61.91 / 62.19 / 63.44	0.813 / 0.779 / 0.784
		Word Order	0.723 / 0.722 / 0.737	0.393 / 0.401 / 0.414	73.28 / 73.62 / 74.89	59.03 / 59.72 / 60.68	0.805 / 0.782 / 0.789
		Synonym	0.706 / 0.695 / 0.715	0.388 / 0.381 / 0.395	72.56 / 72.07 / 73.66	56.32 / 56.06 / 57.64	0.801 / 0.795 / 0.805
		Intensifier	0.713 / 0.704 / 0.728	0.358 / 0.359 / 0.377	73.70 / 73.07 / 75.44	55.83 / 55.76 / 57.54	0.798 / 0.776 / 0.788
		Expansion (No Impact)	0.699 / 0.698 / 0.716	0.333 / 0.350 / 0.356	72.94 / 72.89 / 74.57	54.11 / 55.15 / 56.13	0.797 / 0.778 / 0.783
	Critical	Expansion (Impact)	0.683 / 0.677 / 0.690	0.326 / 0.321 / 0.325	71.04 / 70.86 / 72.48	51.91 / 52.06 / 52.16	0.742 / 0.759 / 0.775
		Omission	0.633 / 0.650 / 0.637	0.314 / 0.335 / 0.325	64.78 / 66.57 / 65.29	49.11 / 51.36 / 50.18	0.723 / 0.753 / 0.755
		Alteration	0.588 / 0.596 / 0.615	0.255 / 0.272 / 0.277	62.27 / 63.16 / 65.23	44.20 / 45.68 / 46.59	0.672 / 0.680 / 0.700
EN-ZH	Minor	Spelling	0.588 / 0.592 / 0.600	0.244 / 0.255 / 0.262	63.31 / 63.31 / 64.56	42.80 / 43.76 / 44.22	0.871 / 0.869 / 0.875
		Word Order	0.587 / 0.594 / 0.598	0.238 / 0.260 / 0.255	62.51 / 63.06 / 63.86	42.49 / 44.08 / 43.70	0.873 / 0.872 / 0.881
		Synonym	0.579 / 0.588 / 0.594	0.239 / 0.258 / 0.250	62.51 / 62.79 / 64.03	41.82 / 43.39 / 43.43	0.884 / 0.887 / 0.898
		Intensifier	0.585 / 0.588 / 0.598	0.232 / 0.244 / 0.247	63.15 / 63.15 / 64.77	41.61 / 42.54 / 42.90	0.868 / 0.866 / 0.879
		Expansion (No Impact)	0.574 / 0.582 / 0.593	0.219 / 0.235 / 0.237	62.37 / 62.84 / 64.69	40.40 / 41.91 / 42.30	0.859 / 0.868 / 0.873
	Critical	Expansion (Impact)	0.561 / 0.564 / 0.574	0.208 / 0.223 / 0.221	61.21 / 61.31 / 63.12	38.38 / 39.15 / 39.52	0.845 / 0.847 / 0.864
		Omission	0.527 / 0.556 / 0.534	0.206 / 0.243 / 0.225	56.47 / 58.81 / 57.49	37.26 / 40.61 / 38.32	0.825 / 0.840 / 0.842
		Alteration	0.492 / 0.490 / 0.503	0.170 / 0.175 / 0.178	54.69 / 54.23 / 56.03	33.54 / 33.89 / 34.42	0.770 / 0.759 / 0.782

Table 27: Detailed results of ASKQE using LLAMA-3 8B. Each cell presents scores for Vanilla / SRL / NLI variant.

Language	Severity	Perturbation	F1	EM	CHRF	BLEU	SBERT
EN-ES	Minor	Spelling	0.648 / 0.655 / 0.840	0.341 / 0.338 / 0.704	68.58 / 69.21 / 85.28	51.50 / 51.73 / 78.10	0.809 / 0.818 / 0.916
		Word Order	0.628 / 0.639 / 0.820	0.318 / 0.331 / 0.679	66.88 / 67.73 / 83.43	49.06 / 50.00 / 75.80	0.798 / 0.805 / 0.904
		Synonym	0.627 / 0.628 / 0.816	0.326 / 0.308 / 0.676	66.30 / 66.10 / 82.71	49.64 / 48.71 / 75.63	0.797 / 0.802 / 0.904
		Intensifier	0.631 / 0.626 / 0.812	0.305 / 0.286 / 0.655	67.45 / 66.60 / 82.85	48.39 / 46.74 / 73.96	0.799 / 0.800 / 0.900
		Expansion (No Impact)	0.620 / 0.620 / 0.814	0.298 / 0.284 / 0.665	66.74 / 66.53 / 82.90	47.59 / 46.63 / 74.83	0.801 / 0.801 / 0.897
	Critical	Expansion (Impact)	0.606 / 0.609 / 0.809	0.287 / 0.265 / 0.653	65.37 / 65.31 / 82.46	45.45 / 44.26 / 73.49	0.786 / 0.787 / 0.896
		Omission	0.571 / 0.587 / 0.746	0.279 / 0.280 / 0.602	61.21 / 62.30 / 75.83	44.13 / 44.81 / 68.33	0.760 / 0.765 / 0.864
		Alteration	0.442 / 0.430 / 0.475	0.188 / 0.180 / 0.360	49.31 / 48.21 / 49.77	32.83 / 31.62 / 42.37	0.636 / 0.618 / 0.762
EN-FR	Minor	Spelling	0.635 / 0.645 / 0.825	0.336 / 0.328 / 0.689	67.58 / 68.06 / 83.81	50.56 / 50.80 / 76.66	0.805 / 0.810 / 0.907
		Word Order	0.628 / 0.628 / 0.818	0.323 / 0.310 / 0.681	66.54 / 66.14 / 82.93	48.93 / 48.23 / 75.65	0.802 / 0.798 / 0.902
		Synonym	0.615 / 0.622 / 0.816	0.305 / 0.286 / 0.672	65.18 / 65.98 / 83.05	48.35 / 46.88 / 74.94	0.795 / 0.792 / 0.892
		Intensifier	0.620 / 0.627 / 0.810	0.291 / 0.285 / 0.670	66.10 / 66.35 / 82.66	47.10 / 47.08 / 74.55	0.797 / 0.800 / 0.897
		Expansion (No Impact)	0.613 / 0.619 / 0.796	0.300 / 0.300 / 0.652	65.63 / 64.77 / 80.76	47.21 / 47.97 / 73.44	0.797 / 0.799 / 0.903
	Critical	Expansion (Impact)	0.607 / 0.600 / 0.797	0.288 / 0.277 / 0.650	64.87 / 64.53 / 81.30	45.51 / 44.57 / 72.61	0.783 / 0.776 / 0.890
		Omission	0.585 / 0.591 / 0.746	0.293 / 0.280 / 0.605	61.78 / 61.90 / 75.90	45.22 / 44.99 / 68.46	0.763 / 0.767 / 0.863
		Alteration	0.409 / 0.384 / 0.408	0.175 / 0.159 / 0.305	46.52 / 43.43 / 43.17	30.03 / 27.90 / 36.10	0.609 / 0.575 / 0.729
EN-HI	Minor	Spelling	0.657 / 0.651 / 0.810	0.365 / 0.352 / 0.681	70.23 / 69.62 / 83.24	53.52 / 53.04 / 75.73	0.820 / 0.811 / 0.907
		Word Order	0.637 / 0.637 / 0.810	0.342 / 0.326 / 0.676	67.91 / 68.37 / 82.97	51.12 / 50.59 / 75.29	0.759 / 0.772 / 0.866
		Synonym	0.623 / 0.618 / 0.798	0.327 / 0.308 / 0.665	66.03 / 65.59 / 81.83	50.01 / 48.93 / 73.96	0.783 / 0.780 / 0.879
		Intensifier	0.625 / 0.619 / 0.778	0.316 / 0.295 / 0.645	67.46 / 66.94 / 79.57	49.04 / 48.00 / 72.47	0.794 / 0.792 / 0.893
		Expansion (No Impact)	0.591 / 0.602 / 0.759	0.303 / 0.303 / 0.630	62.89 / 64.30 / 77.92	45.93 / 46.93 / 70.26	0.796 / 0.802 / 0.901
	Critical	Expansion (Impact)	0.605 / 0.607 / 0.767	0.304 / 0.293 / 0.627	65.39 / 65.29 / 79.20	46.80 / 46.48 / 70.14	0.775 / 0.776 / 0.879
		Omission	0.540 / 0.541 / 0.670	0.274 / 0.265 / 0.547	58.08 / 58.37 / 69.29	42.20 / 42.01 / 61.61	0.717 / 0.723 / 0.825
		Alteration	0.516 / 0.492 / 0.590	0.251 / 0.223 / 0.469	56.37 / 53.86 / 61.34	40.07 / 37.60 / 53.87	0.700 / 0.676 / 0.809
EN-TL	Minor	Spelling	0.686 / 0.690 / 0.832	0.384 / 0.389 / 0.702	72.10 / 72.10 / 83.73	56.54 / 57.44 / 78.18	0.827 / 0.823 / 0.906
		Word Order	0.664 / 0.665 / 0.815	0.366 / 0.354 / 0.682	69.68 / 69.31 / 82.54	54.81 / 54.66 / 76.35	0.812 / 0.801 / 0.902
		Synonym	0.661 / 0.652 / 0.813	0.377 / 0.354 / 0.676	69.71 / 68.20 / 82.05	54.19 / 52.62 / 76.19	0.813 / 0.812 / 0.900
		Intensifier	0.665 / 0.667 / 0.814	0.351 / 0.342 / 0.667	70.09 / 70.03 / 82.58	53.18 / 53.00 / 75.35	0.819 / 0.810 / 0.900
		Expansion (No Impact)	0.671 / 0.656 / 0.814	0.358 / 0.323 / 0.663	71.04 / 69.61 / 82.59	54.12 / 51.99 / 75.40	0.813 / 0.814 / 0.898
	Critical	Expansion (Impact)	0.642 / 0.638 / 0.796	0.318 / 0.303 / 0.650	68.66 / 67.97 / 80.89	50.18 / 49.49 / 73.17	0.798 / 0.795 / 0.888
		Omission	0.586 / 0.599 / 0.718	0.305 / 0.309 / 0.590	61.60 / 62.41 / 72.45	46.83 / 47.59 / 66.75	0.743 / 0.751 / 0.847
		Alteration	0.508 / 0.489 / 0.539	0.251 / 0.228 / 0.430	55.66 / 53.31 / 55.57	39.88 / 37.94 / 49.45	0.681 / 0.659 / 0.778
EN-ZH	Minor	Spelling	0.554 / 0.542 / 0.770	0.267 / 0.237 / 0.633	61.02 / 59.26 / 79.05	42.34 / 40.40 / 70.77	0.753 / 0.747 / 0.880
		Word Order	0.546 / 0.543 / 0.764	0.263 / 0.250 / 0.628	60.24 / 59.78 / 78.79	41.60 / 40.91 / 70.29	0.762 / 0.751 / 0.879
		Synonym	0.552 / 0.541 / 0.762	0.266 / 0.237 / 0.626	61.31 / 59.28 / 78.26	42.02 / 40.64 / 70.35	0.754 / 0.753 / 0.876
		Intensifier	0.548 / 0.541 / 0.762	0.257 / 0.237 / 0.631	60.91 / 59.59 / 78.46	41.15 / 40.09 / 70.60	0.763 / 0.755 / 0.879
		Expansion (No Impact)	0.541 / 0.539 / 0.759	0.259 / 0.232 / 0.620	59.77 / 60.01 / 78.13	41.24 / 39.51 / 70.07	0.767 / 0.753 / 0.876
	Critical	Expansion (Impact)	0.534 / 0.528 / 0.757	0.236 / 0.221 / 0.616	59.69 / 58.92 / 78.30	38.93 / 37.69 / 69.01	0.750 / 0.742 / 0.872
		Omission	0.480 / 0.489 / 0.658	0.224 / 0.214 / 0.539	53.18 / 53.45 / 68.28	36.00 / 36.14 / 60.49	0.695 / 0.699 / 0.817
		Alteration	0.420 / 0.394 / 0.470	0.182 / 0.148 / 0.357	47.79 / 45.69 / 49.43	30.98 / 27.86 / 42.00	0.637 / 0.611 / 0.758

Table 28: Detailed results of ASKQE using LLAMA-3 70B. Each cell presents scores for Vanilla / SRL / NLI variant.

Language	Severity	Perturbation	F1	EM	chrF	BLEU	SBERT
EN-ES	Minor	Spelling	0.645 / 0.621 / 0.805	0.339 / 0.299 / 0.674	70.55 / 67.55 / 82.74	52.55 / 49.27 / 75.16	0.824 / 0.804 / 0.891
		Word Order	0.620 / 0.616 / 0.793	0.313 / 0.296 / 0.660	68.10 / 67.09 / 81.57	49.61 / 48.17 / 73.86	0.809 / 0.799 / 0.886
		Synonym	0.617 / 0.614 / 0.783	0.307 / 0.293 / 0.650	67.45 / 65.88 / 80.71	49.47 / 48.51 / 73.09	0.807 / 0.794 / 0.885
		Intensifier	0.615 / 0.601 / 0.798	0.293 / 0.257 / 0.658	67.64 / 65.41 / 82.27	47.89 / 45.42 / 73.99	0.808 / 0.793 / 0.891
		Expansion (No Impact)	0.609 / 0.605 / 0.795	0.294 / 0.265 / 0.655	67.33 / 65.99 / 82.15	47.76 / 46.39 / 73.82	0.806 / 0.792 / 0.892
	Critical	Expansion (Impact)	0.600 / 0.577 / 0.801	0.284 / 0.249 / 0.661	66.65 / 63.40 / 82.03	46.33 / 42.99 / 73.75	0.798 / 0.771 / 0.896
		Omission	0.570 / 0.575 / 0.721	0.272 / 0.258 / 0.594	62.44 / 62.31 / 74.94	44.70 / 44.07 / 66.76	0.770 / 0.770 / 0.855
		Alteration	0.554 / 0.529 / 0.536	0.264 / 0.223 / 0.398	61.54 / 59.03 / 55.66	43.47 / 39.92 / 47.98	0.750 / 0.732 / 0.804
EN-FR	Minor	Spelling	0.644 / 0.611 / 0.808	0.345 / 0.289 / 0.678	69.91 / 66.67 / 83.09	52.63 / 48.09 / 75.38	0.830 / 0.800 / 0.894
		Word Order	0.618 / 0.598 / 0.798	0.300 / 0.276 / 0.676	67.29 / 64.43 / 81.75	48.71 / 46.59 / 74.80	0.808 / 0.790 / 0.894
		Synonym	0.604 / 0.597 / 0.798	0.293 / 0.282 / 0.672	65.45 / 64.78 / 81.91	48.10 / 46.14 / 74.44	0.795 / 0.788 / 0.889
		Intensifier	0.611 / 0.599 / 0.790	0.294 / 0.268 / 0.657	67.68 / 65.17 / 81.84	47.99 / 45.69 / 73.51	0.812 / 0.799 / 0.904
		Expansion (No Impact)	0.617 / 0.607 / 0.793	0.305 / 0.275 / 0.660	68.17 / 65.84 / 81.51	48.99 / 46.76 / 73.60	0.807 / 0.789 / 0.891
	Critical	Expansion (Impact)	0.589 / 0.566 / 0.779	0.271 / 0.244 / 0.646	65.11 / 62.36 / 80.18	45.27 / 42.07 / 71.93	0.787 / 0.763 / 0.881
		Omission	0.582 / 0.568 / 0.746	0.275 / 0.250 / 0.610	63.07 / 61.11 / 76.52	45.04 / 43.06 / 69.11	0.777 / 0.763 / 0.866
		Alteration	0.539 / 0.505 / 0.480	0.249 / 0.218 / 0.371	59.91 / 56.09 / 49.86	41.64 / 38.13 / 43.31	0.744 / 0.712 / 0.776
EN-HI	Minor	Spelling	0.641 / 0.615 / 0.806	0.351 / 0.311 / 0.686	71.28 / 68.15 / 82.93	53.16 / 49.54 / 75.96	0.824 / 0.802 / 0.891
		Word Order	0.615 / 0.597 / 0.769	0.313 / 0.285 / 0.654	67.51 / 65.06 / 80.23	50.06 / 47.21 / 72.37	0.784 / 0.765 / 0.872
		Synonym	0.600 / 0.571 / 0.753	0.308 / 0.270 / 0.640	66.07 / 63.36 / 78.06	47.99 / 44.53 / 70.65	0.794 / 0.779 / 0.879
		Intensifier	0.620 / 0.601 / 0.799	0.310 / 0.291 / 0.680	69.76 / 66.25 / 82.74	50.12 / 47.44 / 74.77	0.808 / 0.788 / 0.897
		Expansion (No Impact)	0.618 / 0.598 / 0.790	0.301 / 0.277 / 0.664	68.95 / 66.42 / 82.00	49.18 / 46.81 / 74.04	0.815 / 0.789 / 0.902
	Critical	Expansion (Impact)	0.596 / 0.567 / 0.761	0.293 / 0.269 / 0.640	66.37 / 63.50 / 79.41	46.56 / 43.97 / 70.59	0.791 / 0.761 / 0.872
		Omission	0.543 / 0.538 / 0.692	0.267 / 0.249 / 0.580	60.63 / 59.73 / 72.32	43.16 / 42.00 / 64.47	0.742 / 0.735 / 0.843
		Alteration	0.581 / 0.551 / 0.652	0.290 / 0.255 / 0.538	65.03 / 61.03 / 67.53	46.63 / 43.17 / 60.51	0.776 / 0.745 / 0.837
EN-TL	Minor	Spelling	0.663 / 0.660 / 0.823	0.355 / 0.348 / 0.695	72.59 / 71.03 / 84.25	54.89 / 54.64 / 77.29	0.835 / 0.820 / 0.911
		Word Order	0.652 / 0.633 / 0.813	0.357 / 0.322 / 0.696	70.24 / 68.05 / 83.03	54.39 / 51.85 / 76.70	0.821 / 0.792 / 0.895
		Synonym	0.648 / 0.624 / 0.799	0.343 / 0.329 / 0.678	70.53 / 67.91 / 82.42	52.74 / 50.94 / 75.43	0.817 / 0.802 / 0.899
		Intensifier	0.641 / 0.637 / 0.817	0.330 / 0.310 / 0.684	70.35 / 69.08 / 83.88	51.85 / 50.79 / 76.62	0.829 / 0.805 / 0.899
		Expansion (No Impact)	0.654 / 0.632 / 0.809	0.337 / 0.309 / 0.670	71.81 / 68.96 / 83.72	53.26 / 50.68 / 75.47	0.821 / 0.806 / 0.905
	Critical	Expansion (Impact)	0.622 / 0.615 / 0.804	0.301 / 0.293 / 0.676	68.96 / 67.23 / 82.54	49.02 / 48.34 / 74.88	0.806 / 0.784 / 0.896
		Omission	0.578 / 0.593 / 0.703	0.300 / 0.292 / 0.580	63.24 / 63.81 / 72.53	46.60 / 47.47 / 65.66	0.757 / 0.766 / 0.852
		Alteration	0.598 / 0.573 / 0.625	0.306 / 0.271 / 0.504	65.72 / 62.59 / 64.52	48.58 / 45.67 / 57.47	0.777 / 0.748 / 0.834
EN-ZH	Minor	Spelling	0.545 / 0.521 / 0.752	0.255 / 0.219 / 0.635	61.75 / 58.97 / 77.93	42.92 / 39.78 / 70.42	0.762 / 0.745 / 0.878
		Word Order	0.544 / 0.520 / 0.730	0.247 / 0.222 / 0.610	61.29 / 58.35 / 76.87	42.35 / 39.50 / 68.02	0.762 / 0.743 / 0.864
		Synonym	0.540 / 0.517 / 0.750	0.253 / 0.221 / 0.636	61.53 / 58.65 / 78.45	42.34 / 39.34 / 70.47	0.760 / 0.745 / 0.871
		Intensifier	0.551 / 0.519 / 0.738	0.263 / 0.212 / 0.617	63.10 / 58.76 / 77.48	43.09 / 38.88 / 68.64	0.766 / 0.739 / 0.875
		Expansion (No Impact)	0.545 / 0.518 / 0.752	0.245 / 0.220 / 0.640	62.26 / 58.24 / 78.95	41.84 / 38.81 / 70.36	0.775 / 0.745 / 0.865
	Critical	Expansion (Impact)	0.519 / 0.488 / 0.748	0.221 / 0.193 / 0.630	59.89 / 55.84 / 77.86	38.49 / 35.32 / 69.24	0.751 / 0.715 / 0.869
		Omission	0.499 / 0.485 / 0.685	0.228 / 0.201 / 0.571	56.80 / 54.83 / 71.11	38.67 / 36.53 / 63.60	0.726 / 0.712 / 0.832
		Alteration	0.500 / 0.472 / 0.543	0.217 / 0.180 / 0.417	57.58 / 54.42 / 56.70	38.14 / 34.93 / 49.21	0.726 / 0.706 / 0.813

Table 29: Detailed results of AskQE using Yi-1.5 9B. Each cell presents scores for Vanilla / SRL / NLI variant.

Language	QE Metric	F1	EM	CHRF	BLEU	SBERT
EN-ES	XCOMET-QE	0.862 / 0.892 / 0.849	0.608 / 0.733 / 0.695	0.845 / 0.874 / 0.832	0.749 / 0.794 / 0.751	0.944 / 0.947 / 0.902
	METRICX-QE	-0.819 / -0.846 / -0.812	-0.530 / -0.658 / -0.630	-0.814 / -0.838 / -0.804	-0.681 / -0.724 / -0.689	-0.925 / -0.931 / -0.893
	BT-SCORE	0.970 / 0.973 / 0.955	0.792 / 0.876 / 0.848	0.952 / 0.960 / 0.948	0.911 / 0.932 / 0.907	0.982 / 0.976 / 0.958
EN-FR	XCOMET-QE	0.936 / 0.920 / 0.912	0.832 / 0.804 / 0.787	0.930 / 0.910 / 0.899	0.882 / 0.850 / 0.844	0.990 / 0.939 / 0.987
	METRICX-QE	-0.862 / -0.837 / -0.839	-0.734 / -0.692 / -0.685	-0.864 / -0.842 / -0.841	-0.789 / -0.749 / -0.752	-0.948 / -0.863 / -0.960
	BT-SCORE	0.960 / 0.959 / 0.942	0.879 / 0.869 / 0.836	0.949 / 0.945 / 0.926	0.927 / 0.917 / 0.899	0.990 / 0.966 / 0.992
EN-HI	XCOMET-QE	0.768 / 0.741 / 0.667	0.805 / 0.766 / 0.740	0.710 / 0.668 / 0.612	0.818 / 0.814 / 0.740	0.748 / 0.703 / 0.610
	METRICX-QE	-0.934 / -0.916 / -0.865	-0.853 / -0.884 / -0.833	-0.896 / -0.879 / -0.838	-0.894 / -0.892 / -0.843	-0.898 / -0.858 / -0.809
	BT-SCORE	0.959 / 0.960 / 0.949	0.855 / 0.866 / 0.874	0.950 / 0.940 / 0.924	0.943 / 0.955 / 0.945	0.938 / 0.932 / 0.896
EN-TL	XCOMET-QE	0.795 / 0.818 / 0.697	0.791 / 0.800 / 0.709	0.709 / 0.700 / 0.604	0.815 / 0.818 / 0.708	0.831 / 0.816 / 0.698
	METRICX-QE	-0.821 / -0.834 / -0.805	-0.660 / -0.657 / -0.660	-0.813 / -0.808 / -0.768	-0.743 / -0.742 / -0.732	-0.929 / -0.920 / -0.867
	BT-SCORE	0.936 / 0.929 / 0.916	0.846 / 0.827 / 0.847	0.908 / 0.909 / 0.880	0.931 / 0.936 / 0.930	0.905 / 0.912 / 0.895
EN-ZH	XCOMET-QE	0.729 / 0.812 / 0.728	0.706 / 0.847 / 0.759	0.692 / 0.759 / 0.690	0.675 / 0.785 / 0.704	0.764 / 0.769 / 0.727
	METRICX-QE	-0.658 / -0.755 / -0.656	-0.559 / -0.709 / -0.630	-0.668 / -0.751 / -0.674	-0.536 / -0.654 / -0.578	-0.721 / -0.733 / -0.702
	BT-SCORE	0.867 / 0.915 / 0.863	0.767 / 0.847 / 0.803	0.870 / 0.911 / 0.866	0.786 / 0.859 / 0.805	0.906 / 0.907 / 0.903

Table 30: Correlation analysis of ASKQE using GEMMA-2 9B. Each cell presents Pearson correlation coefficients for Vanilla / SRL / NLI. XCOMET-QE (↑), METRICX-QE (↓): Computed between source X_{src} and perturbed MT output Y_{tgt}; BT-Score (↑): Computed between source X_{src} and backtranslated MT output Y_{bt}. **SBERT**: SENTENCEBERT.

Language	QE Metric	F1	EM	CHRF	BLEU	SBERT
EN-ES	XCOMET-QE	0.843 / 0.864 / 0.851	0.700 / 0.680 / 0.676	0.819 / 0.833 / 0.806	0.746 / 0.748 / 0.737	0.905 / 0.936 / 0.906
	METRICX-QE	-0.801 / -0.816 / -0.812	-0.628 / -0.602 / -0.599	-0.793 / -0.800 / -0.787	-0.680 / -0.678 / -0.669	-0.890 / -0.921 / -0.904
	BT-SCORE	0.963 / 0.965 / 0.963	0.860 / 0.832 / 0.837	0.941 / 0.948 / 0.926	0.914 / 0.897 / 0.902	0.972 / 0.976 / 0.957
EN-FR	XCOMET-QE	0.940 / 0.925 / 0.923	0.843 / 0.804 / 0.799	0.916 / 0.898 / 0.891	0.880 / 0.856 / 0.851	0.970 / 0.983 / 0.966
	METRICX-QE	-0.885 / -0.861 / -0.874	-0.762 / -0.710 / -0.719	-0.870 / -0.849 / -0.851	-0.806 / -0.767 / -0.777	-0.937 / -0.962 / -0.943
	BT-SCORE	0.966 / 0.953 / 0.951	0.884 / 0.843 / 0.834	0.947 / 0.929 / 0.926	0.931 / 0.905 / 0.904	0.990 / 0.991 / 0.986
EN-HI	XCOMET-QE	0.683 / 0.644 / 0.573	0.740 / 0.694 / 0.740	0.626 / 0.591 / 0.527	0.757 / 0.730 / 0.681	0.668 / 0.674 / 0.550
	METRICX-QE	-0.894 / -0.837 / -0.805	-0.870 / -0.829 / -0.858	-0.852 / -0.803 / -0.766	-0.875 / -0.825 / -0.815	-0.862 / -0.833 / -0.754
	BT-SCORE	0.959 / 0.945 / 0.926	0.882 / 0.859 / 0.907	0.937 / 0.915 / 0.899	0.960 / 0.939 / 0.946	0.940 / 0.914 / 0.889
EN-TL	XCOMET-QE	0.658 / 0.773 / 0.670	0.729 / 0.769 / 0.749	0.561 / 0.614 / 0.547	0.683 / 0.777 / 0.696	0.656 / 0.743 / 0.621
	METRICX-QE	-0.772 / -0.801 / -0.758	-0.684 / -0.628 / -0.656	-0.731 / -0.750 / -0.711	-0.707 / -0.694 / -0.685	-0.842 / -0.872 / -0.808
	BT-SCORE	0.942 / 0.955 / 0.952	0.890 / 0.791 / 0.872	0.896 / 0.918 / 0.894	0.953 / 0.922 / 0.953	0.885 / 0.938 / 0.895
EN-ZH	XCOMET-QE	0.602 / 0.743 / 0.716	0.600 / 0.727 / 0.698	0.560 / 0.647 / 0.601	0.575 / 0.700 / 0.665	0.566 / 0.699 / 0.629
	METRICX-QE	-0.616 / -0.711 / -0.716	-0.533 / -0.601 / -0.642	-0.618 / -0.682 / -0.660	-0.525 / -0.587 / -0.608	-0.601 / -0.691 / -0.657
	BT-SCORE	0.835 / 0.902 / 0.909	0.740 / 0.758 / 0.814	0.821 / 0.875 / 0.859	0.791 / 0.817 / 0.848	0.812 / 0.882 / 0.858

Table 31: Correlation analysis of ASKQE using GEMMA-2 27B. Each cell presents Pearson correlation for Vanilla / SRL / NLI.

Language	QE Metric	F1	EM	CHRF	BLEU	SBERT
EN-ES	XCOMET-QE	0.740 / 0.887 / 0.957	0.567 / 0.729 / 0.960	0.716 / 0.844 / 0.962	0.642 / 0.785 / 0.962	0.890 / 0.935 / 0.946
	METRICX-QE	-0.703 / -0.842 / -0.937	-0.502 / -0.668 / -0.934	-0.692 / -0.806 / -0.940	-0.585 / -0.722 / -0.938	-0.899 / -0.909 / -0.950
	BT-SCORE	0.904 / 0.978 / 0.928	0.796 / 0.847 / 0.933	0.877 / 0.956 / 0.932	0.853 / 0.922 / 0.939	0.935 / 0.975 / 0.910
EN-FR	XCOMET-QE	0.802 / 0.916 / 0.984	0.599 / 0.821 / 0.988	0.764 / 0.883 / 0.980	0.713 / 0.846 / 0.986	0.910 / 0.953 / 0.986
	METRICX-QE	-0.719 / -0.856 / -0.945	-0.481 / -0.753 / -0.955	-0.703 / -0.829 / -0.939	-0.619 / -0.774 / -0.948	-0.933 / -0.928 / -0.960
	BT-SCORE	0.861 / 0.963 / 0.965	0.707 / 0.878 / 0.971	0.830 / 0.936 / 0.964	0.804 / 0.918 / 0.970	0.933 / 0.981 / 0.973
EN-HI	XCOMET-QE	0.559 / 0.761 / 0.714	0.563 / 0.739 / 0.729	0.541 / 0.697 / 0.714	0.644 / 0.789 / 0.740	0.499 / 0.720 / 0.675
	METRICX-QE	-0.781 / -0.888 / -0.959	-0.701 / -0.853 / -0.971	-0.743 / -0.870 / -0.962	-0.766 / -0.859 / -0.970	-0.701 / -0.858 / -0.895
	BT-SCORE	0.929 / 0.990 / 0.873	0.818 / 0.920 / 0.871	0.908 / 0.957 / 0.858	0.929 / 0.974 / 0.889	0.860 / 0.943 / 0.877
EN-TL	XCOMET-QE	0.535 / 0.749 / 0.785	0.645 / 0.766 / 0.812	0.453 / 0.672 / 0.775	0.562 / 0.756 / 0.810	0.557 / 0.816 / 0.747
	METRICX-QE	-0.682 / -0.781 / -0.929	-0.608 / -0.700 / -0.943	-0.641 / -0.772 / -0.928	-0.614 / -0.710 / -0.942	-0.771 / -0.892 / -0.919
	BT-SCORE	0.898 / 0.950 / 0.852	0.884 / 0.853 / 0.850	0.859 / 0.922 / 0.843	0.916 / 0.945 / 0.856	0.826 / 0.939 / 0.869
EN-ZH	XCOMET-QE	0.619 / 0.694 / 0.872	0.596 / 0.745 / 0.872	0.552 / 0.601 / 0.860	0.576 / 0.628 / 0.872	0.625 / 0.635 / 0.782
	METRICX-QE	-0.594 / -0.601 / -0.841	-0.473 / -0.621 / -0.832	-0.585 / -0.534 / -0.835	-0.462 / -0.468 / -0.828	-0.674 / -0.549 / -0.784
	BT-SCORE	0.865 / 0.857 / 0.845	0.715 / 0.816 / 0.836	0.826 / 0.809 / 0.845	0.765 / 0.749 / 0.847	0.880 / 0.832 / 0.902

Table 32: Correlation analysis of ASKQE using LLAMA-3 8B. Each cell presents Pearson correlation for Vanilla / SRL / NLI.

Language	QE Metric	F1	EM	chrF	BLEU	SBERT
En-Es	xCOMET-QE	0.954 / 0.956 / 0.969	0.918 / 0.858 / 0.962	0.951 / 0.950 / 0.967	0.937 / 0.914 / 0.967	0.989 / 0.991 / 0.983
	MetricX-QE	-0.919 / -0.913 / -0.936	-0.864 / -0.793 / -0.924	-0.921 / -0.912 / -0.937	-0.888 / -0.855 / -0.930	-0.965 / -0.963 / -0.964
	BT-Score	0.971 / 0.958 / 0.943	0.974 / 0.916 / 0.956	0.963 / 0.955 / 0.942	0.984 / 0.958 / 0.953	0.958 / 0.955 / 0.955
En-Fr	xCOMET-QE	0.969 / 0.968 / 0.972	0.916 / 0.925 / 0.968	0.963 / 0.960 / 0.970	0.956 / 0.952 / 0.972	0.990 / 0.993 / 0.989
	MetricX-QE	-0.917 / -0.908 / -0.926	-0.831 / -0.843 / -0.920	-0.914 / -0.904 / -0.925	-0.888 / -0.878 / -0.922	-0.954 / -0.953 / -0.962
	BT-Score	0.952 / 0.960 / 0.952	0.908 / 0.922 / 0.952	0.952 / 0.953 / 0.951	0.954 / 0.956 / 0.955	0.972 / 0.978 / 0.971
En-Hi	xCOMET-QE	0.749 / 0.719 / 0.697	0.810 / 0.757 / 0.720	0.715 / 0.705 / 0.687	0.814 / 0.782 / 0.727	0.723 / 0.701 / 0.674
	MetricX-QE	-0.963 / -0.969 / -0.961	-0.951 / -0.942 / -0.969	-0.935 / -0.958 / -0.955	-0.962 / -0.975 / -0.973	-0.923 / -0.940 / -0.916
	BT-Score	0.916 / 0.850 / 0.828	0.903 / 0.813 / 0.838	0.908 / 0.851 / 0.820	0.939 / 0.876 / 0.849	0.884 / 0.813 / 0.861
En-Tl	xCOMET-QE	0.823 / 0.883 / 0.864	0.849 / 0.902 / 0.882	0.770 / 0.835 / 0.852	0.855 / 0.904 / 0.881	0.843 / 0.877 / 0.863
	MetricX-QE	-0.919 / -0.925 / -0.949	-0.874 / -0.830 / -0.941	-0.901 / -0.914 / -0.951	-0.905 / -0.888 / -0.948	-0.968 / -0.964 / -0.972
	BT-Score	0.884 / 0.867 / 0.809	0.887 / 0.856 / 0.819	0.877 / 0.871 / 0.805	0.917 / 0.905 / 0.822	0.858 / 0.846 / 0.824
En-Zh	xCOMET-QE	0.793 / 0.866 / 0.854	0.806 / 0.856 / 0.859	0.775 / 0.831 / 0.856	0.796 / 0.857 / 0.858	0.802 / 0.851 / 0.820
	MetricX-QE	-0.799 / -0.833 / -0.842	-0.747 / -0.798 / -0.839	-0.801 / -0.834 / -0.846	-0.765 / -0.787 / -0.838	-0.823 / -0.828 / -0.822
	BT-Score	0.868 / 0.862 / 0.854	0.861 / 0.810 / 0.844	0.870 / 0.869 / 0.851	0.872 / 0.845 / 0.854	0.880 / 0.858 / 0.882

Table 33: Correlation analysis of AskQE using LLaMA-3 70B. Each cell presents Pearson correlation for Vanilla / SRL / NLI.

Language	QE Metric	F1	EM	chrF	BLEU	SBERT
En-Es	xCOMET-QE	0.868 / 0.914 / 0.885	0.734 / 0.794 / 0.755	0.854 / 0.901 / 0.838	0.788 / 0.850 / 0.811	0.912 / 0.939 / 0.911
	MetricX-QE	-0.825 / -0.856 / -0.835	-0.659 / -0.715 / -0.684	-0.820 / -0.851 / -0.802	-0.721 / -0.775 / -0.738	-0.879 / -0.888 / -0.871
	BT-Score	0.968 / 0.974 / 0.959	0.876 / 0.908 / 0.874	0.959 / 0.975 / 0.940	0.931 / 0.950 / 0.929	0.978 / 0.981 / 0.969
En-Fr	xCOMET-QE	0.932 / 0.947 / 0.950	0.833 / 0.870 / 0.882	0.914 / 0.941 / 0.935	0.874 / 0.899 / 0.903	0.874 / 0.941 / 0.912
	MetricX-QE	-0.864 / -0.874 / -0.887	-0.734 / -0.768 / -0.793	-0.857 / -0.875 / -0.879	-0.784 / -0.807 / -0.817	-0.861 / -0.877 / -0.859
	BT-Score	0.954 / 0.967 / 0.968	0.859 / 0.907 / 0.911	0.943 / 0.964 / 0.956	0.913 / 0.938 / 0.939	0.923 / 0.961 / 0.935
En-Hi	xCOMET-QE	0.743 / 0.783 / 0.690	0.800 / 0.815 / 0.792	0.704 / 0.728 / 0.608	0.813 / 0.862 / 0.788	0.747 / 0.740 / 0.704
	MetricX-QE	-0.951 / -0.969 / -0.907	-0.898 / -0.920 / -0.908	-0.916 / -0.928 / -0.851	-0.934 / -0.952 / -0.913	-0.948 / -0.921 / -0.936
	BT-Score	0.953 / 0.931 / 0.950	0.885 / 0.864 / 0.873	0.952 / 0.931 / 0.914	0.955 / 0.934 / 0.959	0.826 / 0.686 / 0.723
En-Tl	xCOMET-QE	0.815 / 0.853 / 0.761	0.828 / 0.847 / 0.816	0.731 / 0.779 / 0.651	0.834 / 0.857 / 0.806	0.854 / 0.908 / 0.863
	MetricX-QE	-0.869 / -0.848 / -0.829	-0.758 / -0.719 / -0.727	-0.838 / -0.836 / -0.784	-0.808 / -0.781 / -0.770	-0.915 / -0.912 / -0.908
	BT-Score	0.933 / 0.949 / 0.955	0.852 / 0.875 / 0.895	0.924 / 0.947 / 0.923	0.935 / 0.941 / 0.952	0.905 / 0.834 / 0.846
En-Zh	xCOMET-QE	0.784 / 0.878 / 0.767	0.756 / 0.848 / 0.813	0.698 / 0.830 / 0.688	0.764 / 0.845 / 0.767	0.786 / 0.825 / 0.777
	MetricX-QE	-0.736 / -0.787 / -0.760	-0.624 / -0.685 / -0.710	-0.695 / -0.785 / -0.734	-0.654 / -0.695 / -0.694	-0.699 / -0.718 / -0.694
	BT-Score	0.892 / 0.867 / 0.900	0.783 / 0.726 / 0.803	0.865 / 0.888 / 0.874	0.845 / 0.811 / 0.857	0.825 / 0.794 / 0.788

Table 34: Correlation analysis of AskQE using Yi-1.5 9B. Each cell presents Pearson correlation for Vanilla / SRL / NLI.

I Question Categorization Prompt

We use the annotation guidelines and few-shot examples from Cao and Wang (2021).

Prompt: Question Categorization

Task: You will be given a question. Your goal is to annotate the question type. The question type reflects the nature of the question. It is NOT determined by the interrogative word of the question. There are 10 question types in total. The definition for each type is shown in the following, along with examples per question type. During annotation, you can label two most-confident types when no clear decision can be made for the most probable type. Output you answer in Python list format without giving any additional explanation.

*** Question Type Starts ***
1. Verification: Asking for the truthfulness of an event or a concept.
- Is Michael Jackson an African American?
- Could stress, anxiety, or worry cause cholesterol levels to rise?

2. Disjunctive: Asking for the true one given multiple events or concepts, where comparison among options is not needed.
- Is Michael Jackson an African American or Latino?
- When you get a spray-on tan does someone put it on you or does a machine do it?

3. Concept: Asking for a definition of an event or a concept.
- Who said the sun never sets on the British empire?
- Where do dolphins have hair at?

4. Extent: Asking for the extent or quantity of an event or a concept.
- How long does gum stay in your system?
- To what extent is the Renewable Fuel Standard accurate nationwide?

5. Example: Asking for example(s) or instance(s) of an event or a concept.
- What are some examples to support or contradict this?
- What countries/regions throughout the world do not celebrate the Christmas holidays?

6. Comparison: Asking for comparison among multiple events or concepts.
- What is the best tinted facial moisturizer?
- In what hilariously inaccurate ways is your job/career portrayed on television or in movies?

7. Cause: Asking for the cause or reason for an event or a concept.
- Why are parents strick on girls than boys?
- What makes nerve agents like 'Novichok' so hard to produce and why can only a handful of laboratories create them?

8. Consequence: Asking for the consequences or results of an event.
- What are the negative consequences for the services if they do not evaluate their programs?
- What would happen if employers violate the legislation?

9. Procedural: Asking for the procedures, tools, or methods by which a certain outcome is achieved.
- How did the Amish resist assimilation into the current social status in the U.S?
- How astronomers detect a nebula when there are no stars illuminating it?

10. Judgmental: Asking for the opinions of the answerer's own.
- Do you think that it's acceptable to call off work for a dying-dead pet?
- How old is too old for a guy to still live with his mother?
*** Question Type Ends ***

Question: `{question}`
Question type(s):