Beyond AskQE: Summarization and MT-QE via NLI, Multi-Turn and Hallucination Detection

Daniele Amato
Politecnico di Torino
s334211@studenti.polito.it

Alessia Ciccaglione
Politecnico di Torino
s344499@studenti.polito.it

Sonia Foco
Politecnico di Torino
s343043@studenti.polito.it

Ilaria Parodi
Politecnico di Torino
s339184@studenti.polito.it

***Abstract*—AskQE leverages question generation and answering to evaluate MT quality by comparing answers derived from source sentences and their backtranslated outputs, enabling monolingual quality assessment. However, the original framework has limitations in handling semantic nuances, detecting hallucinations, and extending beyond MT tasks. We introduce three extensions: NLI-based answer comparison for factual consistency, multi-turn questioning for iterative semantic probing, and backtranslation-based QA for hallucination detection. We also adapt the framework to summarization evaluation in biomedical contexts. Experiments on BioMQM and PubMed datasets demonstrate improvements in precision and semantic coverage compared to the original framework. The code is available at: https://github.com/AlessiaCicca/Beyond-AskQE**

I. Problem Statement

Monolingual source speakers cannot effectively assess MT quality in languages they do not understand, yet existing QE methods fail to address this challenge. Current approaches produce either difficult-to-interpret scalar scores or target-language error annotations that remain inaccessible to users without target language knowledge. This gap is critical in high-stakes settings like healthcare, where users need actionable, source-language feedback to decide whether to accept or reject translations that could have serious consequences. The AskQE framework addresses this by generating questions from source sentences and comparing answers from the source versus back-translated MT output. The framework has several limitations. It struggles with semantic nuance handling, often treating semantically similar but factually distinct answers as equivalent, especially in specialized domains. Its evaluation is shallow, missing errors that only appear through deeper probing. It also has difficulty detecting hallucinations, failing to identify when translations introduce incorrect information, particularly when backtranslations propagate these issues. Lastly, its scope is limited to machine translation evaluation and cannot be applied to other generation tasks where quality estimation is crucial.

We address these limitations through four extensions: (A) NLI-based answer comparison to assign entailment scores for robust factual consistency assessment, (B) multi-turn questioning to probe deeper into content through iterative follow-up questions, (C) backtranslation-based QA for explicit hallucination detection, and (D) adaptation to summarization evaluation to measure information preservation in biomedical documents, enabling assessment of whether summaries adequately capture essential content or whether consulting the full text is necessary.

II. Methodology

AskQE framework

AskQE, as introduced in [1], operates through a two-stage pipeline: Question Generation (QG) and Question Answering (QA). Given a source sentence X_{src}, AskQE can optionally extract atomic facts and filter them via NLI-based entailment classification to retain only the facts entailed by the source. Alternatively, as used in most extensions and the baseline, an LLM generates questions Q_{src} conditioned solely on X_{src}, without the need for fact extraction. For each question $q \in Q_{src}$, the system generates two answers: reference answers A_{src} by answering using X_{src} as context, and candidate answers A_{bt} using the backtranslated MT output Y_{bt} as context. Backtranslation converts the target-language MT (Y_{tgt}) into source-language text, enabling monolingual evaluation. Translation quality is assessed by comparing answer pairs using similarity metrics, with lower scores indicating potential translation errors.

$$\text{AskQE}(Y_{tgt}) = \frac{1}{N}\sum_{i=1}^{N} \rho(A^i_{src}, A^i_{bt})$$

where ρ is the similarity metric and $N = |Q_{src}|$.

A. NLI-Based Factual Consistency Metric

As a first extension, we extend the AskQE quality estimation framework by incorporating a factual consistency metric inspired by the Q² evaluation paradigm [6]. In this setting, we leverage Natural Language Inference (NLI) to assess the alignment between answers derived from the source sentence and those obtained from the backtranslated sentence. This provides a robust comparison that is less sensitive to lexical variations and enhances factual consistency evaluation.

The extension follows the original steps of the AskQE pipeline for both QG and QA, and for each answer pair A_{src}, A_{bt} corresponding to a question Q_{src}, an NLI pipeline is applied. In particular, we use the concatenation of Q_{src} and A_{src} as the premise and the concatenation Q_{src} and A_{bt} as the hypothesis, as shown in this example:

Premise ($Q_{src} + A_{src}$): What does peri-operative chemotherapy refer to? Chemotherapy administered before and *during* surgery.

Hypothesis ($Q_{src} + A_{bt}$): What does peri-operative chemotherapy refer to? Chemotherapy administered before and *after* surgery.

Output: *Contradiction*

Then, we define the answer-pair score as follows: a value of 1 is assigned in the case of entailment, and 0 in the case of contradiction or when the QA model produces no answer. In the neutral case, we retain the original AskQE F1 score between A_{src} and A_{bt}.

B. AskQE Multi-Turn

As a further extension to the AskQE framework, we introduce a multi-turn evaluation approach (Figure 1) aimed at providing a more comprehensive assessment of translation quality.

In our multi-turn framework, both Q_{src} and A_{src}, along with the complete source text X_{src}, serve as inputs for a sequential evaluation where each round of questioning builds on previous responses, allowing for a more thorough and context-aware assessment of the translation's meaning and coherence.

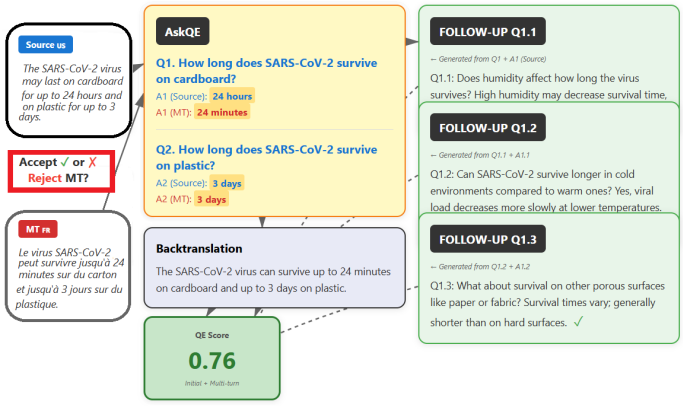

Fig. 1. Overview of the multi-turn AskQE evaluation process with iterative follow-up question generation.

We leverage a large language model to generate coherent, contextually relevant follow-up questions (Appendix Section B) based on the initial answers A_{src}, the original questions Q_{src}, and the complete source text X_{src}. Once the follow-up questions FQ_{src} are generated, we produce answers for both the source and back-translated texts, denoted as FA_{src} and FA_{bt}, respectively. The answer FA_{src} and follow-up question FQ_{src} then serve as inputs for the next evaluation round. This process continues for a predefined number of turns, refining the system's understanding of translation quality and contextual integrity. The cascading approach enables deeper semantic analysis, uncovering discrepancies missed in single-turn evaluations, especially in complex domains like biomedical translations.

The output of this process generates for each source question Q_{src} a series of follow-up questions across N_{turn} iterations, resulting in a total of $|Q_{\text{src}}| \times (N_{\text{turn}} + 1)$ questions when combined with the original single-turn approach. To aggregate this information into a single, cohesive evaluation, we compute the similarity between each follow-up question FQ_{src} and the original source question Q_{src}, as well as between FQ_{src} and the original source answer A_{arc}. Additionally, we compute the similarity between the follow-up responses FA_{src} and FA_{bt}. These similarities are measured using several evaluation metrics, which capture both the semantic alignment and the accuracy of the generated responses.

Subsequently, we integrate the AskQE results and multi-turn results into a unified scoring system. For each question and turn, a weighted average is calculated across the three QA-based metric pairs using the following formula:

$$0.3\left(\frac{\rho(Q_{\text{src}}, FQ_{\text{src}}) + \rho(A_{\text{src}}, FQ_{\text{src}})}{2}\right) + 0.7\rho(FQ_{\text{src}}, FA_{\text{bt}})$$

In this formula, ρ represents a measure of similarity,which will be defined in Experiments. The weighted average produces a single score for each question-turn combination, reflecting the relevance and consistency of the interactions. The average of these scores is then calculated across all turns to yield a unified score. Finally, these multi-turn results are combined with the original AskQE results by averaging the corresponding scores, providing a robust evaluation that considers both accuracy and semantic consistency across multiple layers of questioning.

C. Backtranslation-Based QA for Hallucination Detection

As a complementary extension, we invert the AskQE paradigm to explicitly detect hallucinations introduced during translation. Instead of generating questions from the source X_{src}, we generate them from the backtranslation Y_{bt} and validate whether the queried information is supported by the source.

Given X_{src} and its backtranslation Y_{bt}, the pipeline proceeds as follows: (1) generate 3-4 questions Q_{bt} from Y_{bt} (Appendix, Section C-A); (2) for each $q \in Q_{bt}$, extract answers A_{bt} (from Y_{bt}) and A_{src} (from X_{src}) using the QA prompt in Appendix, Section C-A; (3) compute the Unanswerable Content Rate (UCR):

$$\text{UCR} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}[A^i_{\text{src}} = \text{“No Answer”} \wedge A^i_{\text{bt}} \neq \text{“No Answer”}]$$

where high UCR indicates that Y_{bt} contains facts not grounded in X_{src}. This metric provides a direct signal of hallucination: if the backtranslation can answer a question but the source cannot, the information must have been introduced during translation (Example 1, Appendix Section C-B); (4) for non-empty answer pairs, compute BERTScore similarity and flag pairs below an arbitrary threshold as semantic mismatches; (5) generate 1-3 binary verification questions from each A_{bt} (Appendix, Section C-A) and answer them using X_{src} (Appendix, Section C-A): a hallucination is detected when X_{src} answers “No” to a claim supported by Y_{bt}. This inversion provides a complementary signal to standard AskQE:

while the original pipeline measures information *preservation* (source→BT), our approach measures information *addition* (BT→source).

D. Summarization with AskQE

As final extension, we adapt the AskQE framework to the task of summarization.

Although related question-based evaluation approaches have already been proposed, including QAGS [1], our framework differs from QAGS in a fundamental and complementary way. QAGS generates questions from the summary and evaluates whether the information introduced by the summary is supported by the source document, answering to: "If the summary states X, is X supported by the document?

In contrast, we generate questions from the original document and evaluate whether its information is preserved in the summary. Formally, our approach addresses the question: "If the document states X, is X preserved in the summary?"

The document is treated as ground truth, and the summary is evaluated based on the proportion of source-grounded information that remains answerable. Since summaries are inherently compressive, they are not expected to preserve all source facts. Quantifying the proportion of document-grounded information that remains answerable in the summary, we obtain a structured measure of information coverage.
The evaluation pipeline consists in the following steps:

1) The original document D_{src} is decomposed into a set of atomic facts.
2) A NLI filtering step is applied to retain only atomic facts that are entailed by the source.
3) A set of questions is generated from the filtered facts, ensuring that questions are grounded in factual content present in the document.
4) Each question Q_{src} is answered using both the document D_{src} and the summary S_{sum}, producing respectively a reference answer A_{src} and a candidate answer A_{sum}.
5) The two answers sets are compared to measure information preservation.

To ensure methodological robustness, we introduced task-specific prompt refinements. The QG prompt was adapted to align with the document-level summarization setting. During preliminary experiments, we observed that in-context examples in the QA prompt occasionally induced pattern-copying behavior, causing the model to restate the question, as consequence we removed the example and used stricter output constraints (Appendix D).

III. Experiments

All MT experiments use NLLB-200 for translation and backtranslation [3], representing the state-of-the-art in multilingual neural machine translation, except for the results in Section III-A, which use Google Translate for comparative purposes with the original AskQE settings. We employ Qwen2.5-3B-Instruct and Qwen2.5-7B-Instruct [7] for perturbation, atomic facts extraction, question generation and answering, selected for their optimal trade-off between computational efficiency and generation quality. RoBERTa-large-MNLI [8] is used for NLI-based fact filtering. Across all extensions, the QA prompt is modified to explicitly instruct the model to return "no answer" when information is absent, enabling robust hallucination detection and missing information identification. Translation quality is assessed using standard QE metrics including F1, BLEU, ChrF, Exact Match (EM), and SBERT similarity for text comparison.

A. Baseline and NLI-Based Metric Evaluation

To effectively compare the performance of our proposed NLI-based metric with respect to the other metrics proposed in the original AskQE framwork, we evaluated our approach on the BioMQM dataset [2]. In particular, we examined translation from *English* sentences to *German*, *Spanish*, *French*, *Russian* and *Chinese*.

To assess the method's effectiveness, we simulated a *real-world scenario* on BioMQM where human evaluators decide whether to *accept* or *reject* machine-translated output. Following the original approach, translations are *rejected* in case of critical or major errors when human annotations are available. When feedback is a **single scalar score**, we use a two-component *Gaussian Mixture Model (GMM)* that clusters segments into *"Accept"* or *"Reject"* groups and automatically determines the optimal decision threshold. We evaluated *decision accuracy* by comparing predicted labels $\hat{l}$ with *human judgment labels* l from the *BIOMQM dataset*.

We compared our approach with the original AskQE framework on the same dataset, as shown in Table I (refer to Tables V and VI in the Appendix for more detailed results). Results show a significant improvement in decision accuracy with the NLI-based metric, which reaches 63.97% in the baseline and 60.96% in the modified version, outperforming the other metrics. This demonstrates that NLI-based metrics offer more precise classification of acceptable and rejectable translations, aligning better with human judgment. The results obtained with the baseline and the modified prompt forcing the "No Answer" are comparable, with only slight variations across the metrics.

TABLE I
ACCURACY COMPARISON WITH NLI-BASED METRIC: BASELINE (ASKQE ORIGINAL FRAMEWORK) VS. NO-ANSWER (MODIFIED PROMPT FORCING "NO ANSWER")

Dec. Accuracy	SBERT	BLEU	ChrF	F1	NLI
Baseline	49.95%	52.07%	44.94%	55.13%	**63.97%**
No-Answer	50.29%	50.29%	43.46%	53.13%	**60.96%**

B. Multi-Turn Evaluation

To enhance multi-turn evaluation, we use long and informative texts that allow for more detailed questions and deeper semantic assessment. We began with three-sentence source texts from PubMed [4] and BioMQM [2] datasets, which provide complex biomedical content for thorough translation

evaluation. On PubMed, we analyzed translation quality from *English* to *Italian*, *French*, and *Spanish*, assessing the impact of perturbations on source texts to evaluate the system's robustness under critical semantic modifications. Following the original work, the perturbation strategies simulate critical translation errors, including **expansion impact** (adding new meaning), **omission** (removing words or phrases), and **alteration** (changing the meaning of words or phrases). The results from multi-turn evaluation are generally comparable to the original evaluation, with improvements in specific cases, such as higher ChrF scores for French-Alteration (Table VIII in the Appendix). Although metrics like F1, BLEU, and EM may be lower in some cases due to the increased number of comparisons, this reflects a more rigorous evaluation process rather than decreased system quality (Table VII in the Appendix).

TABLE II
ACCURACY COMPARISON: SINGLE-TURN VS. MULTI-TURN ACROSS DIFFERENT METRICS.

Dec. Accuracy	SBERT	BLEU	ChrF	F1
Single-Turn	55.96%	49.22%	48.19%	51.81%
Multi-Turn	**56.99%**	48.70%	**51.81%**	50.78%

The simulated real-world scenario defined in Section III-A is also applied to assess the effectiveness of this extension. The *multi-turn evaluation* showed improvements in handling perturbations and ensuring consistency across evaluation rounds. For instance, *SBERT* exhibited increased *accuracy* (from 55.96% to 56.99%), indicating better semantic capture. Similarly, *ChrF* scores improved (from 48.19% to 51.81%), showing better character-level n-gram alignment. The *decision accuracy* values for each *QE metric*, shown in Table II, demonstrate that multi-turn evaluation provides consistent improvements in the *accept/reject* process. Detailed *precision*, *recall* and *F1 score* values, with notable improvements also in the other metrics, are reported in Appendix B (Table X).

C. Hallucination Detection Evaluation

We evaluated the extension on 50 sentences from ContraTICo with **expansion impact** perturbations (English→Spanish). The Yes/No verification step is particularly effective because it targets facts at a finer granularity, making hallucination detection more likely: a "No" answer from the source unambiguously indicates that a fact stated in the backtranslation is unsupported (Example 3, Appendix Section C-B). BERTScore filtering, with the threshold set to $\tau = 0.6$, confirmed that low-scoring pairs (roughly 20% of the answered questions) often correspond to genuine hallucinations rather than paraphrasing (Example 2, Appendix Section C-B).

TABLE III
HALLUCINATION DETECTION RATES

Method	Questions	Sentences
UCR	12.12%	32.00%
Yes/No Verification	22.84%	56.25%

D. Summarization results

We evaluate our approach on a subset of the PubMed dataset [4], which contains biomedical articles paired with gold abstracts. In our setting, only the first 800 tokens of each article are considered. Processing full-length biomedical articles was not feasible due to computational constraints that would require splitting documents into multiple overlapping segments for atomic fact extraction and question generation, potentially introducing misalignment between facts and questions and biasing coverage estimation. Restricting the input to a fixed-length segment ensures methodological consistency and a coherent evaluation context.

Summaries are automatically generated from the truncated documents using Distilbart-cnn-12-6 [9], avoiding comparison against gold abstracts that may contain information not present in the truncated input. For the NLI step the entailment threshold was fixed at 0.5.

TABLE IV
METRICS FOR SUMMARIZATION

Metric	F1	EM	ChrF	BLEU	SBERT
Avg Score	0.2334	0.1586	26.2753	19.8980	0.3139

The results in Table IV indicate limited information coverage between source documents and generated summaries. This is consistent with the inherently compressive nature of abstractive summarization, where substantial information reduction is expected. The relatively low Exact Match score compared to F1 suggests that answers extracted from summaries rarely coincide verbatim with those from the source, even when partial lexical overlap exists, reflecting the paraphrastic nature of abstractive summarization. Overall, the proposed framework provides a graded measurement of information preservation, allowing us to quantify the extent to which source content is retained in the summary.

IV. CONCLUSION

Our work extends the AskQE framework by addressing key limitations in semantic sensitivity, hallucination detection, and task generalization. The integration of NLI enhances factual robustness by moving beyond surface similarity toward entailment-based comparison. The multi-turn extension transforms evaluation into a dynamic process, uncovering discrepancies that single-turn evaluation may overlook. The hallucination module shifts the focus from information preservation to information addition, capturing errors that would otherwise remain undetected. Finally, adapting the framework to summarization shows that question-based evaluation can quantify information coverage. However, these improvements have a computational cost, especially for multi-turn evaluation, which increases latency and resource requirements.

Overall, our results suggest that question-based evaluation frameworks offers a flexible and extensible paradigm for assessing generation quality across multiple dimensions.

REFERENCES

[1] Alex Wang, Kyunghyun Cho, and Mike Lewis. "Asking and Answering Questions to Evaluate the Factual Consistency of Summaries," arXiv preprint arXiv:2004.04228, 2020. Available at https://arxiv.org/abs/2004.04228.

[2] Vilém Zouhar, Shuoyang Ding, Anna Currey, Tatyana Badeka, Jenyuan Wang, and Brian Thompson, *Fine-Tuned Machine Translation Metrics Struggle in Unseen Domains*, arXiv preprint arXiv:2306.07899, 2024, https://arxiv.org/abs/2402.18747.

[3] Facebook AI Research, *No Language Left Behind (NLLB) 200 3.3B model*, 2023. Available at: https://huggingface.co/facebook/nllb-200-3.3B.

[4] Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian, *A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents*, Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), New Orleans, Louisiana, 2018, Association for Computational Linguistics, pp. 615–621, https://aclanthology.org/N18-2097, 10.18653/v1/N18-2097.

[5] Dayeon Ki, Kevin Duh, and Marine Carpuat. "AskQE: Question Answering as Automatic Evaluation for Machine Translation." arXiv preprint arXiv:2504.11582, 2025. Available at https://arxiv.org/abs/2504.11582.

[6] Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend "Q^2: Evaluating Factual Consistency in Knowledge-Grounded Dialogues via Question Generation and Question Answering" https://aclanthology.org/2021.emnlp-main.619/.

[7] Qwen Team, *Qwen2.5: A Party of Foundation Models*, Available at https://qwenlm.github.io/blog/qwen2.5/, September 2024.

[8] Facebook AI, RoBERTa-large-MNLI, Available at https://huggingface.co/FacebookAI/roberta-large-mnli, September 2024.

[9] Distilbart-cnn-12-6, Available at https://huggingface.co/sshleifer/distilbart-cnn-12-6.

APPENDIX

The QA prompt used in this work to explicitly instructs the model to return 'No answer' when the requested information is absent or unavailable.

Prompt: 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. If there is no answer in the sentence answer with 'No answer'. Output only the list of answers in Python list format without giving any additional explanation. Do not output as code format (python).

***** Example Starts *****
Sentence: and does this pain move from your chest?
Questions: ["What moves from your chest?", "Where does the pain move from?", "When did the pain start?"]
Answers: ["The pain", "Your chest", "No Answer"]

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?", "Which conditions were reported with a frequency of 10.9%, 9.2%, and 9.0%, respectively?", "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", "Diabetes mellitus (10.9%), chronic lung disease (9.2%), and cardiovascular disease (9.0%)", "10.9%", "9.2%", "9.0%"]

***** Example Ends *****

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

APPENDIX A
NLI-BASED FACTUAL CONSISTENCY METRIC

TABLE V
RESULTS OF ASKQE ORIGINAL FRAMEWORK (BASELINE) EVALUATED ON BIOMQM DATASET PER LANGUAGE PAIR AND ERROR SEVERITY.

Language	Severity	F1	EM	CHRF	BLEU	SBERT	NLI
En-De	No Error	0.6883	0.5368	74.73	63.55	0.8355	0.8775
	Neutral	0.7856	0.6333	83.35	71.97	0.8658	0.8818
	Minor	0.6677	0.4542	73.86	58.21	0.8266	0.8718
	C+M	0.6577	0.4426	74.47	58.04	0.8428	0.8899
En-Es	No Error	0.7608	0.5593	82.32	69.43	0.8866	0.9237
	Neutral	0.7427	0.6088	80.20	68.50	0.8581	0.9410
	Minor	0.7554	0.5264	80.48	67.63	0.8698	0.9047
	C+M	0.7502	0.4833	79.72	66.12	0.8759	0.9181
En-Fr	No Error	0.7393	0.5871	79.70	67.20	0.8744	0.9219
	Neutral	0.5444	0.0000	55.57	26.93	0.8096	1.0000
	Minor	0.6908	0.4595	74.11	60.07	0.8512	0.9212
	C+M	0.7236	0.4955	77.32	63.56	0.8414	0.8444
En-Ru	No Error	0.7511	0.4545	77.63	62.51	0.8788	0.9420
	Neutral	0.7404	0.3454	76.32	56.76	0.8800	0.9510
	Minor	0.7842	0.5011	79.69	67.08	0.8937	0.9337
	C+M	0.6387	0.4034	71.02	56.03	0.7752	0.7808
En-Zh-CN	No Error	0.6730	0.4437	74.93	58.72	0.8591	0.9011
	Neutral	0.6259	0.4105	76.37	53.64	0.8445	0.9031
	Minor	0.6456	0.3072	70.45	51.94	0.8086	0.8463
	C+M	0.5917	0.3059	67.57	47.99	0.7928	0.8200

TABLE VI
RESULTS OF ASKQE WITH THE MODIFIED PROMPT "NO ANSWER" EVALUATED ON BIOMQM DATASET PER LANGUAGE PAIR AND ERROR SEVERITY.

Language	Severity	F1	EM	CHRF	BLEU	SBERT	NLI
En-De	No Error	0.7041	0.5201	76.71	64.54	0.8490	0.8797
	Neutral	0.6664	0.4409	74.76	58.77	0.8009	0.8238
	Minor	0.7012	0.5098	76.17	62.76	0.8340	0.8675
	C+M	0.6715	0.4395	75.20	59.32	0.8421	0.8827
En-Es	No Error	0.6998	0.5015	76.76	63.30	0.8500	0.8993
	Neutral	0.6795	0.4431	70.96	57.08	0.7986	0.9154
	Minor	0.7238	0.4637	76.79	62.58	0.8468	0.8674
	C+M	0.6677	0.3835	73.28	55.52	0.8272	0.8385
En-Fr	No Error	0.7063	0.5221	77.95	63.31	0.8507	0.8891
	Neutral	0.5625	0.1250	66.42	39.80	0.7904	0.8750
	Minor	0.6971	0.4307	75.15	59.60	0.8350	0.8950
	C+M	0.6753	0.4634	73.11	59.66	0.8125	0.8460
En-Ru	No Error	0.7004	0.4028	74.05	56.72	0.8533	0.8985
	Neutral	0.6926	0.3096	72.54	50.78	0.8399	0.9182
	Minor	0.7467	0.4843	77.86	63.96	0.8547	0.9095
	C+M	0.6099	0.3281	68.13	51.24	0.7367	0.6964
En-Zh-CN	No Error	0.6500	0.4270	71.14	56.13	0.8255	0.8828
	Neutral	0.6476	0.4396	73.36	55.61	0.8113	0.8507
	Minor	0.6322	0.3139	67.16	51.28	0.8020	0.8395
	C+M	0.5587	0.2676	63.64	45.33	0.7477	0.7920

APPENDIX B
ASKQE MULTI-TURN

Prompt: Multi-Turn Follow-Up Question Generation

Task: You will be given the full original text, a previous question, and the answer to that previous question (derived from the text). Your goal is to generate ONE follow-up question for a multi-turn question answering process. The follow-up question must be answerable using ONLY the given text and must logically follow from the previous question and answer. Output ONLY the question text without any additional explanation.

*** **Rules** ***

- The follow-up question MUST be answerable using ONLY the given text.
- The question MUST logically follow from the previous question and answer.
- Do NOT introduce new entities, facts, or assumptions.
- Do NOT repeat the previous question.
- Ask about a different aspect, detail, or consequence explicitly mentioned in the text.
- If no valid follow-up question can be generated, output exactly: `NO_FOLLOWUP`
- Output ONLY the question text. No explanations.

*** **Input Format** ***

Text: {{text}}
Previous question: {{prev_question}}
Previous answer: {{prev_answer}}
Follow-up question:

TABLE VII
PERFORMANCE METRICS BY LANGUAGE AND PERTURBATION: SINGLE-TURN

Language	Perturbation	F1	EM	CHRF	BLEU	SBERT	Count
En-It	Alteration	0.408	0.239	48.694	33.726	0.582	2437
	Expansion	0.466	0.285	52.943	38.832	0.631	1680
	Omission	0.452	0.279	52.659	37.984	0.620	2434
En-Fr	Alteration	0.409	0.233	48.556	33.835	0.579	2464
	Expansion	0.463	0.258	54.866	38.181	0.657	2473
	Omission	0.452	0.264	53.475	38.016	0.632	2435
En-Es	Alteration	0.412	0.241	48.904	34.583	0.583	2455
	Expansion	0.458	0.267	54.067	38.521	0.635	2463
	Omission	0.502	0.295	57.913	41.927	0.676	2455

TABLE VIII
PERFORMANCE METRICS BY LANGUAGE AND PERTURBATION: MULTI-TURN

Language	Perturbation	F1	EM	CHRF	BLEU	SBERT	Count
En-It	**Alteration**	0.394	0.119	45.979	31.310	0.570	2437
	Expansion	0.447	0.142	50.353	35.456	0.621	1680
	Omission	0.435	0.140	49.946	35.032	0.608	2434
En-Fr	**Alteration**	0.401	0.116	46.605	31.951	0.578	2464
	Expansion	0.448	0.129	51.676	35.553	0.639	2473
	Omission	0.432	0.132	49.917	34.841	0.612	2435
En-Es	**Alteration**	0.392	0.120	45.523	31.536	0.563	2455
	Expansion	0.437	0.134	50.377	35.344	0.613	2468
	Omission	0.478	0.147	54.318	38.448	0.653	2455

TABLE IX
BIOMQM: COMPARISON BETWEEN SINGLE-TURN AND MULTI-TURN EVALUATION

Type	F1	EM	ChrF	BLEU	SBERT	Count
Single-Turn	0.5362	0.3579	59.38	46.39	0.6898	950
Multi-Turn	0.5195	0.1789	57.46	43.42	0.6845	950

TABLE X
COMPARISON OF METRICS BETWEEN SINGLE-TURN AND MULTI-TURN DECISION ACCURACY.

Metric	Type	Accuracy	Precision	Recall	F1 Score
SBERT	Single-Turn	55.96%	0.1299	0.3571	0.1905
	Multi-Turn	56.99%	0.1687	0.5000	0.2523
BLEU	Single-Turn	49.22%	0.1354	0.4643	0.2097
	Multi-Turn	48.70%	0.1553	0.5714	0.2443
ChrF	Single-Turn	48.19%	0.1170	0.3929	0.3929
	Multi-Turn	51.81%	0.1429	0.4643	0.2185
F1	Single-Turn	51.81%	0.1264	0.3929	0.1913
	Multi-Turn	50.78%	0.1616	0.5714	0.2520

APPENDIX C
BACKTRANSLATION-BASED QA FOR HALLUCINATION DETECTION

A. *Prompts*

Prompt: Question Generation from Backtranslation (QG)

Task: You will be given an English sentence. Your goal is to generate a list of NOT LESS THAN 3 AND NOT MORE THAN 4 relevant questions based on the sentence. Every question must be answerable using an exact text span from the sentence. Output only the list of questions in Python list format without giving any additional explanation.

***** Example Starts *****
Sentence: It is not yet known whether the severity or level of control of underlying health conditions affects the risk for severe disease associated with COVID-19.
Questions: ["What is not yet known?", "What might affect the risk for severe disease associated with COVID-19?", "What is associated with severe disease?", "What is the disease mentioned in the sentence?"]

***** Example Ends *****

Sentence: {{sentence}}
Questions:

Prompt: Question Answering (QA)

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. If there is no answer in the sentence answer with 'No answer'. Output only the list of answers in Python list format without giving any additional explanation. Do not output as code format (“‘python“‘).

***** Example Starts *****
Sentence: and does this pain move from your chest?
Questions: [”What moves from your chest?”, ”Where does the pain move from?”, ”When did the pain start?”]
Answers: [”The pain”, ”Your chest”, ”No Answer”]

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?”, ”Which conditions were reported with a frequency of 10.9%, 9.2%, and 9.0%, respectively?”, ”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”, ”Diabetes mellitus (10.9%), chronic lung disease (9.2%), and cardiovascular disease (9.0%)”, ”10.9%”, ”9.2%”, ”9.0%”]

***** Example Ends *****

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

Prompt: Yes/No Verification Question Generation

You are given:

- a question
- an answer

Your task: Generate YES/NO verification questions to check whether the information in the answer is explicitly supported by the ORIGINAL source text.

Rules:

- Output MUST be a valid JSON array: [”Question 1?”, ”Question 2?”, ”Question 3?”]
- Generate 1 to 3 YES/NO questions.
- Each question must be answerable with YES or NO.
- Each question must verify ONE specific fact in the answer.
- Do NOT ask open-ended questions.
- Do NOT add explanations.
- Output ONLY a JSON array of strings.

Examples:
Question: ”What symptoms does the patient have?”
Answer: ”fever and headache”
→ ”The patient has fever and headache?”

Question: ”Which university does he attend?”
Answer: ”University of Tokyo”
→ [”Does he attend the University of Tokyo?”]

Question: {{question}}
Answer: {{answer}}
Verification questions:

Prompt: Yes/No Question Answering

You are given a sentence and a YES/NO question.

Sentence:
{{sentence}}

Question:
{{question}}

Answer the question using ONLY the information in the sentence.
Answer with:

- "Yes" if the statement is explicitly supported,
- "No" if it is contradicted or not supported or if the information is not mentioned.

Do not add explanations.
Answer with a single token.

Answer:

B. Examples

The following examples illustrate the three detection approaches using real samples from the ContraTICo dataset with *expansion_impact* perturbation (English→Spanish→English):

Example 1: UCR Detection

Source: We have the opportunity to escalate the data extraction to twice weekly if needed.

Perturbed MT: Tenemos la oportunidad de aumentar la extracción de datos al doble por semana en caso de ser necesario *y reducir costos*.

Backtranslation: We have the opportunity to double data extraction per week if necessary *and reduce costs*.

Question (from BT): What can be reduced if necessary?
Answer from BT: Costs
Answer from Source: No Answer

→ **Hallucination detected:** The backtranslation can answer the question about cost reduction, but this information is absent from the source, indicating added content.

Example 2: BERTScore Validation

Source: Seven trials were evaluating repurposed drugs already approved to treat malaria, including four studies on hydroxychloroquine or chloroquine phosphate.

Perturbed MT: Siete ensayos evaluaban fármacos readaptados ya aprobados para el paludismo, incluidos cuatro estudios sobre hidroxicloroquina, fosfato de cloroquina *y otros antipalúdicos*.

Backtranslation: Seven trials evaluated re-adapted and approved drugs for malaria, including four studies on hydroxychloroquine, chlorophyll phosphate *and other anti-malarial drugs*.

Question (from BT): What drugs were studied in four of the trials?
Answer from BT: Hydroxychloroquine, chlorophyll phosphate and other anti-malarial drugs
Answer from Source: Hydroxychloroquine or chloroquine phosphate
BERTScore: 0.5958

→ **Hallucination detected:** The similarity score reveals semantic divergence caused by the added phrase "and other anti-malarial drugs," which expands the scope of the original statement beyond what was explicitly mentioned in the source.

Example 3: Yes/No Verification

Source: The fever started two days ago.

Perturbed MT: La fiebre *alta* empezó hace dos días.

Backtranslation: The *high* fever started two days ago.

Question (from BT): What is the symptom mentioned in the sentence?
Answer from BT: High fever
Answer from Source: The fever

Note: UCR and BERTScore fail to detect this subtle difference.

Verification Question: Is the symptom high fever?
Answer from BT: Yes
Answer from Source: No

→ **Hallucination detected:** The binary verification successfully identifies the semantic difference between "fever" and "high fever" that other methods missed. By decomposing the BT answer into an atomic claim and validating it against the source, the method detects the subtle but potentially significant modifier introduced by the perturbation.

APPENDIX D
SUMMARIZATION WITH ASKQE

This section reports the prompts used for question generation and question answering in the summarization setting.

A. Question Generation Prompt

Prompt: Question Generation (QG)

Task: You will be given an English article 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 atomic facts. Output the list of questions in Python list format without giving any additional explanation. Do not output as code format.

***** Example Starts *****
Article: Tardive dystonia is characterized by sustained involuntary muscle contractions. It occurs in approximately 3% of patients with long-term antipsychotic exposure.
Atomic facts: ["Tardive dystonia is characterized by sustained involuntary muscle contractions.", "Tardive dystonia occurs in approximately 3% of patients with long-term antipsychotic exposure."]
Questions: ["How is tardive dystonia characterized?", "In approximately what percentage of patients with long-term antipsychotic exposure does tardive dystonia occur?"]

***** Example Ends *****

Article: {{article}}
Atomic facts: {{atomic_facts}}
Questions:

B. Question Answering Prompt

Prompt: Question Answering (QA)

Instruction: You are answering factual questions using ONLY the provided text.

Rules:

- Answers must be short spans copied verbatim from the text.
- Do NOT repeat the question.
- Do NOT repeat the article.
- If the answer is not explicitly stated, output exactly: No_Answer.
- Output ONLY a Python list of answers.

Text: {{text}}
Questions: {{questions}}
Answers:

We constrain the QA model to produce extractive span-level answers in order to minimize generative variability and prevent implicit inference. This design ensures that answer comparison reflects explicit textual support rather than paraphrastic reformulation or model hallucination. When the requested information is not explicitly present in the input text, the model is required to output a predefined null token (`No_Answer`), enabling a consistent and interpretable measurement of information coverage.