Automatic Labelling of Speech Translation Errors

Anonymous ACL submission

Abstract

Errors in speech translations reduce trustworthiness of Speech Translation (ST) systems and can have serious consequences. Yet currently there is no established methodology for evaluating confidence and quality estimation of speech translations. To initiate progress in this direction, we propose Speech Translation Error Labelling (STEL). We create an annotation protocol, a small authentic end-to-end evaluation dataset, and we analyse how existing text-only and speech-processing systems perform the STEL task. Our results show that text-only XCOMET and multimodal LLM Qwen2.5-Omni are able to perform the STEL task in roughly half the precision of humans. We also find that direct speech processing is necessary for the STEL task, and that the current text-only and speech-processing systems are complementary in labelling translation-only vs. speech-processing errors in ST.

1 Introduction

Speech Translation (ST) has the potential to reduce language barriers in many everyday situations. To enable its widespread adoption, including in domains where misunderstanding can be critical, such as healthcare, business, or law, it is necessary to make its output more trustworthy. Speech Translation Quality Estimation (ST QE) could be used to highlight potentially incorrectly translated words, which would allow ST users to respond appropriately, e.g. ask for clarification of the highlighted output spans. It could also enable automatic system selection and post-editing both by humans and generative large language models.

While automatic QE systems already exist for text machine translation (Lavie et al., 2025; Zhao et al., 2024; Guerreiro et al., 2024), there is an open question whether automatic ST QE should use ASR transcripts or whether they should process speech directly to prevent compounding of errors and information loss.

Moreover, ST is usually used in specific communication situations that differ from text-only communication (Fantinuoli and Prandi, 2021) for which Error Span Annotations datasets exist (Kocmi et al., 2024b,a, 2025). Speech communication participants typically have a communication goal and a knowledge of background context that often make them less sensitive to subtle grammar flaws, and may prefer fluent translations with hesitations and false starts removed. We need a new evaluation methodology that takes these aspects into account.

We therefore propose the Speech Translation Error Labelling (STEL) task and make the following contributions: (1) we design an annotation protocol focused on end-to-end long-form speech translation (Section 2); (2) we create a test set for four language directions, which required 6.5 hours of annotation effort (Section 3); (3) we demonstrate the feasibility and state of the art of STEL using existing automatic systems, both text-only and speech processing, and (4) we investigate using text-only vs. speech-processing methods for STEL (Section 4). We release the dataset and code publicly.[1]

2 STEL Annotation Protocol

The first contribution of this work is an STEL annotation protocol, which builds upon the existing Error Span Annotation for text MT (Kocmi et al., 2024b), but adds the ST user-centric and communication-oriented perspective (Fantinuoli and Prandi, 2021).

In our annotation protocol we assign four severities of errors to erroneous subsequences of translations, called *spans*. *Critical* errors are mistranslations that change the meaning of the sentence and cannot be corrected by the listener. A mistranslation of a name or term that could be inferred by a knowledgeable listener is a *minor* error. Subtle, e.g. grammar errors, that may only affect comfort but

[1] Anonymous preview: anonymous.4open.science/r/STEL

not meaning, are *negligible* errors. *Redundancies* are, for example, verbosely translated repetitions, false starts, filler words, etc. While redundancies may or may not be translation errors, they typically divert focus from the communicative goal.

In addition to error span detection and categories, STEL also includes segment-level DA ranging from 0-100. It is an auxiliary task which may be helpful for training automatic STEL systems, similarly as in text MT ESA.

3 STEL Dataset Creation

The second contribution of this work is a STEL dataset primarily intended for evaluation of automatic ST error labelling systems.

We selected four language directions that represent authentic use cases: Czech (Cs) to English (En), and English to Czech, German (De), and Hebrew (He). We reused documents from existing ST test sets, namely the Cs→En Robothon debate from the ELITR Test Set (Ansari et al., 2021), and one randomly selected talk from the ACL6060 test set (Salesky et al., 2023) for En→{Cs, De, He}.

We selected ST candidate systems as representatives of state of the art from three typically used groups of models that are publicly available for self-hosting: (1) **cascades of ASR and LLM for MT** because they tend to be top-performing (Papi et al., 2025), but may be large and costly; (2) **an end-to-end speech-to-text model** that tends to be small and high-performing; (3) **a simultaneous system** that processes incremental input in low latency, allowing real-time interaction. Since all these systems either include ASR component in the cascade, or use a multi-task models, we used their ASR transcripts for automatic labelling in Section 4. Details about the used ST systems are in Appendix B.

We translated the recordings with ST systems, and applied mWERSegmenter (Post and Hoang, 2025) to segment and align the candidate translation to the sentence segmentation of the reference translation that is available in the reused test sets. Since Hebrew and Czech translations are not available in ACL6060, we used manual segmentation and alignment for Hebrew and machine translated source transcript for mWERSegmenter for Czech. We segmented the recordings without omissions to allow the annotators to take into account the inter-sentence context.

We collected STEL annotations in Pearmut (Zouhar and Kocmi, 2026), which is a flexible web application for multi-lingual evaluation campaigns. We asked three volunteers, native Czech, German, and Hebrew speakers, who are English L2 speakers, to provide STEL annotations. We instructed them to consider typical audience, their background knowledge, and their communication goal. In total, they worked 6.5 hours and annotated 329 sentence-like segments in 32 minutes of audio. Tab. 1 summarizes the size of the annotated data. We present the exact instructions and an overview of the counts and categories of annotated spans in Section C in Fig. 1 and Tab. 4 respectively.

Language directions	**# Seg.**	**Dur.**	**Work** [m.]
Cs→En	169	14:06	126
En→Cs (twice)	37	3:45	51 (+41)
En→De (twice)	37	3:45	49 (+43)
En→He	86	10:34	78
Total:	329	32:10	388

Table 1: Annotation statistics. For En→Cs and En→De, we annotated the same data twice. For En→He, we have two ST candidates, for the others we have three.

As a consistency check and to obtain an upper bound for automatic systems, we asked the same En→Cs and En→De annotators to annotate the same data again for the second time one month after their first annotation. We observe relatively high agreement of 72% for the En→Cs annotator, and 56 % for the En→De annotator. We explain the gap from the ideal 100% agreement by a combination of factors. First, by ambiguity of natural language, which can not be avoided, but can be handled e.g. by using multiple independent annotations as references. The second group of factors include the low experience of annotators, their limited attention, and by annotation guidelines being not specific in corner cases. These factors can be improved in additional work, which is beyond the available capacity. Nevertheless, we conclude that our annotations are reasonable for demonstrating the systems evaluation and for providing evidence for our research questions.

4 Automatic Labelling

Our third contribution is an experimental study of existing automatic systems that can perform STEL, aiming to answer: (1) How well is it possible to perform STEL with the current automatic systems? (2) Do the automatic STEL systems benefit from having access to the source audio?

	Cs→En			En→Cs			En→De			En→He		
	F1 w	F1 uw	Kend.	F1 w	F1 uw	Kend.	F1 w	F1 uw	Kend.	F1 w	F1 uw	Kend.
XCOMET^{ASR}	**18.4**$_{+0.9}$	**22.3**$_{+0.8}$	**23.5**$_{-1.7}$	**31.8** $_{0.0}$	**38.7**$_{+0.4}$	11.5$_{+8.7}$	20.5$_{+0.9}$	**30.5**$_{-2.0}$	26.1$_{+10.1}$	1.0$_{+9.2}$	1.8$_{+12.5}$	1.5$_{+22.3}$
Qwen$^{ASR+audio}$	14.3$_{+0.6}$	18.1$_{-0.4}$	12.6$_{+4.5}$	24.0$_{+1.3}$	29.9$_{+2.6}$	2.1$_{+6.7}$	**27.8**$_{+0.2}$	30.1$_{+0.4}$	32.5$_{+20.7}$	**13.2**$_{-1.5}$	**17.5**$_{-2.1}$	41.0$_{-1.2}$
Qwen^{ASR}	15.0$_{+1.2}$	17.5$_{+0.4}$	13.3$_{-5.3}$	23.1$_{-2.4}$	29.7$_{-5.0}$	2.9$_{-15.3}$	21.4$_{-2.6}$	24.2$_{-2.0}$	25.0$_{-27.3}$	9.6$_{-0.4}$	14.7$_{+0.3}$	−3.0$_{-44.7}$
Qwen^{audio}	11.8	14.7	3.5	27.1	30.4	**13.2**	23.4	26.9	43.0	11.0	12.8	**41.6**
2nd annotation	–	–	–	72.8	71.9	51.4	52.5	55.5	63.1	–	–	–

Table 2: Automatic STEL results. The meta-evaluation scores are micro weighted and unweighted F1 in % (the higher, the better) and Kendall's τ multiplied by 100 (range -100 to 100, the higher, the better). Subscripts indicate the difference when using gold transcripts.

4.1 Automatic STEL Systems

We evaluate representatives of two state-of-the-art systems that can perform STEL in many language directions including the ones in this study: a specialized text-MT QE model, and a general-purpose multimodal LLM. Inference is performed on a single NVIDIA A100-SXM4-40GB GPU. Prompts and inference parameters are in Appendix D.

XCOMET `XCOMET-XL`[2] (Guerreiro et al., 2024) is a state-of-the-art reference-free QE model trained specifically to detect translation errors in text-MT. Given a source sentence and a candidate translation, it produces a sentence-level quality score together with a list of error spans. Since we primarily focus on the real-life use case, where the gold source transcript is unavailable, we process XCOMET with an ASR transcript. The variant with gold transcript is only for contrastive analysis.

Qwen `Qwen2.5-Omni-7B`[3] (Xu et al., 2025) is a state-of-the-art multimodal LLM capable of processing both text and audio in many languages. We prompt it to produce a sentence-level quality score together with error spans in the target text. We experimented with three parameters: modality, spans with/without severities of errors, and using the context of adjacent segments. However, we primarily report the variant with an ASR transcript and audio, with severities, and no adjacent context.

Since neither XCOMET nor Qwen, unlike the ST systems, provide their own segmentation of the long-form source of multiple consecutive sentences into their individually processed units, we follow the common ST practice of using the "gold" source audio segmentation from the test sets. We leave the more realistic segmentation to future work.

[2] Unbabel/XCOMET-XL
[3] Qwen/Qwen2.5-Omni-7B

4.2 Meta-Evaluation

We evaluate the STEL systems against human annotations along two dimensions: span detection quality and score-level correlation.

Span detection. We measure how well predicted error spans overlap with human-annotated spans using character-level F1, following the WMT25 Error Span Annotation (ESA) evaluation protocol (Lavie et al., 2025). For each sample, we convert predicted and reference spans to sets of character indices and we report *unweighted micro F1* over their overlap. We also report the *severity-weighted micro F1*, following Lavie et al. (2025): same-severity overlaps receive full credit, while cross-severity overlaps receive half credit. Since XCOMET predicts three (Critical, Minor, Major) severity labels, we treat a predicted *Minor* as correct for gold *Minor* or *Negligible*, and a predicted *Major* or *Critical* as correct for gold *Major* or *Redundancy*.

Score correlation. We measure correlation to DA scores using Kendall's τ (Kendall, 1945).

4.3 Results

Tab. 2 shows results of automatic systems using ASR, contrasted to the same systems using gold transcripts and the second human annotation. In general, we observe that **automatic systems are able to perform the STEL task with approximately half the quality of humans** (e.g. En→Cs F1 unweighted XCOMET 38.7 vs. human 71.9.). Both XCOMET and Qwen perform better in span identification than in assigning the severity labels (unweighted F1 scores are higher than weighted). XCOMET outperforms Qwen in three language directions, but not in En→He, which may be caused by low amount of Hebrew training data in COMET, which leads to nearly no span predictions and unweighted F1 only 1.8, compared to 14.3 with gold transcript. The gold transcripts improve the error

Avg. of all langs.	F1 w wer>0	F1 w wer=0	F1 uw wer>0	F1 uw wer=0
XCOMET^{ASR}	10.78	**9.95**	15.08	**12.24**
Qwen$^{ASR+audio}$	12.98	8.80	**15.96**	11.23
Qwen^{ASR}	11.84	7.03	15.24	9.16
Qwen^{audio}	**13.11**	7.04	15.15	9.00

Table 3: STEL scores for speech-processing errors (wer>0) vs. translation-only errors (wer=0) for the realistic STEL systems. Maximum scores in bold.

span labelling over the same model setups with ASR, probably due to higher quality over ASR, but only by little margin of 0 to 3.6, which suggests that most of the STEL task complexity is in other challenges than speech processing, for instance, translation modelling, not using inter-segment context, or segmentation to processing units. Exceptions are En→He, and Cs→En unweighted F1 for Qwen, but not weighted, which may be caused by lower quality of the En→Cs gold transcript.

We also observe that the systems are able to provide segment-level DA with varying quality over language directions and models. Comparisons of Kendall's τ for Qwen on En→De vs. En→Cs (32.5 vs. 2.1) may indicate Qwen limitations in generalization, multitasking, or prompting, or the quality of DA annotations over annotators.

4.4 Speech vs. Text Translation Errors

Next, we focus on the second research question, on speech vs. text-only processing in the automatic STEL systems. We presume two groups of ST errors by their origin: in ASR (or in the speech processing component of end-to-end ST) vs. translation-only errors. We therefore split the annotated error spans into these groups, depending whether the corresponding ASR spans (detected with AwesomeAlign, Dou and Neubig, 2021, after tokenization with Moses Koehn et al., 2007) are correct (WER=0 when compared to the gold transcript), or include an error (WER>0). We observe that these two groups have nearly the same size.

Tab. 3 summarizes the scores for both groups. We observe that translation-only errors are better detected by XCOMET, while speech-processing errors are better detected by Qwen using either audio (weighted F1), or ASR and audio (unweighted F1). We conclude that **speech-processing is a necessary component of the STEL methods,** and that the **current text-only and speech-processing STEL systems are complementary**.

5 Recommendations

In our experiments, we explored two STEL systems that used the internal ASR transcripts of the evaluated ST systems. While the text-based XCOMET is cheaper to run and excels at predicting translation-only errors, Qwen is more robust to ASR errors because it can leverage the speech input. Based on our experience, we would recommend choosing systems for future evaluations based on the language pairs under consideration, taking into account the amount of target-language text seen during training of XCOMET/Qwen, and the accuracy of the ASR system in the source language.

One approach to making STEL predictions more accurate is to reduce the gap to gold transcripts by improving the accuracy of the ASR systems used for STEL. This could be achieved by using a strong external ASR system, which would also allow black-box evaluation of ST systems that do not produce ASR transcripts. The external ASR system could also be further improved by using ASR post-correction. We plan to evaluate using the strong external ASR system in future work.

The last criterion to consider is the severity taxonomy and how easily it can be customised. Since XCOMET cannot be easily adapted to predict new severity classes, it may be better to use speech LLMs, which can predict new severity labels with in-context learning.

6 Conclusion

In this paper, we introduced the Speech Translation Error Labelling task (STEL), which is crucial for making speech translation systems more trustworthy, especially in settings where errors can lead to critical consequences. We proposed the STEL annotation protocol for annotating spans of incorrect translations and assigning severity to these errors. We created and annotated a small evaluation dataset for four language directions, and we analysed the performance of two existing automatic systems on this task. Our results show that it is possible to label speech translation errors automatically, although current methods have limitations and achieve roughly half the performance of humans. We provide an evidence that future better performing methods may need to process speech directly, rather than relying on ASR, and that the current text-only and speech-processing systems are complementary in labelling translation-only vs. speech-processing ST errors.

7 Limitations

One limitation of this work is the limited coverage of language directions, domains, and speech documents – only two documents, 14 and 10 minutes long. Other domains and documents may include phenomena that are more challenging for speech processing, such as speaker characteristics (gender, age, accents, idiolect, dialect), emotion, etc.; or phenomena more challenging for translation, such as specific terminology, context, ambiguities, rare words, etc.

We warn against considering every annotated or omitted error span as ground truth without considering the limiting factors: the number and availability of annotators – one annotator per language direction, 6.5 hours of annotation effors in total — the limited experience in language annotations, and low variability of annotators. The annotated dataset is primarily intended for measuring progress in development of automatic STEL systems by aggregating scores across the whole dataset.

There was also limited investigation of the optimal ST and STEL systems and their parameters.

8 Ethical Considerations

The annotators participated voluntarily. No personal data were collected nor published. The participating annotators were informed openly about their participations and planned used of their outcomes. No ethics board was involved, as no regulation of any institutions or country where this work was carried out requires it for this kind of work.

In this paper, AI assistants were used for assistance with the coding and formatting of tables. The authors reviewed their outputs.

All artifacts (datasets, models, software) were used in accordance with their licence and permitted use. The created artifact is equipped with limitations section and primarily intended use.

References

Ebrahim Ansari, Ondřej Bojar, Barry Haddow, and Mohammad Mahmoudi. 2021. SLTEV: Comprehensive evaluation of spoken language translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 71–79, Online. Association for Computational Linguistics.

Mauro Cettolo, Marco Gaido, Matteo Negri, Sara Papi, and Luisa Bentivogli. 2025. How to evaluate speech translation with source-aware neural mt metrics. *ArXiv*, abs/2511.03295.

Mingda Chen, Paul-Ambroise Duquenne, Pierre Andrews, Justine Kao, Alexandre Mourachko, Holger Schwenk, and Marta R. Costa-jussà. 2023. BLASER: A text-free speech-to-speech translation evaluation metric. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9064–9079, Toronto, Canada. Association for Computational Linguistics.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online. Association for Computational Linguistics.

Claudio Fantinuoli and Bianca Prandi. 2021. Towards the evaluation of automatic simultaneous speech translation from a communicative perspective. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 245–254, Bangkok, Thailand (online). Association for Computational Linguistics.

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, and 542 others. 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, Kevin Duh, and Marine Carpuat. 2024. SpeechQE: Estimating the quality of direct speech translation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21852–21867, Miami, Florida, USA. Association for Computational Linguistics.

Maurice George Kendall. 1945. The treatment of ties in ranking problems. *Biometrika*, 33(3):239–251.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Christof Monz, Kenton Murray, and 10 others. 2025. Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*, pages 355–413, Suzhou, China. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, and 3 others. 2024a. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024b. Error span annotation: A balanced approach for human evaluation of machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1440–1453, Miami, Florida, USA. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Alon Lavie, Greg Hanneman, Sweta Agrawal, Diptesh Kanojia, Chi-Kiu Lo, Vilém Zouhar, Frederic Blain, Chrysoula Zerva, Eleftherios Avramidis, Sourabh Deoghare, Archchana Sindhujan, Jiayi Wang, David Ifeoluwa Adelani, Brian Thompson, Tom Kocmi, Markus Freitag, and Daniel Deutsch. 2025. Findings of the WMT25 shared task on automated translation evaluation systems: Linguistic diversity is challenging and references still help. In *Proceedings of the Tenth Conference on Machine Translation*, pages 436–483, Suzhou, China. Association for Computational Linguistics.

Dominik Macháček, Ondřej Bojar, and Raj Dabre. 2023. MT metrics correlate with human ratings of simultaneous speech translation. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 169–179, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Dominik Macháček and Peter Polák. 2025. Simultaneous translation with offline speech and LLM models in CUNI submission to IWSLT 2025. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 389–398, Vienna, Austria (in-person and online). Association for Computational Linguistics.

Sara Papi, Javier Garcia Gilabert, Zachary Hopton, Vilém Zouhar, Carlos Escolano, Gerard I. Gállego, Jorge Iranzo-Sánchez, Ahrii Kim, Dominik Macháček, Patricia Schmidtova, and Maike Züfle. 2025. Hearing to translate: The effectiveness of speech modality integration into llms. *Preprint*, arXiv:2512.16378.

Matt Post and Hieu Hoang. 2025. Effects of automatic alignment on speech translation metrics. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 84–92, Vienna, Austria (in-person and online). Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, João Alves, Pedro Teixeirinha, Amin Farajian, and André F. T. Martins. 2025. Tower+: Bridging generality and translation specialization in multilingual llms. *Preprint*, arXiv:2506.17080.

Elizabeth Salesky, Kareem Darwish, Mohamed Al-Badrashiny, Mona Diab, and Jan Niehues. 2023. Evaluating multilingual speech translation under realistic conditions with resegmentation and terminology. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 62–78, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Monica Sekoyan, Nithin Rao Koluguri, Nune Tadevosyan, Piotr Zelasko, Travis Bartley, Nikolay Karpov, Jagadeesh Balam, and Boris Ginsburg. 2025. Canary-1b-v2 & parakeet-tdt-0.6b-v3: Efficient and high-performance models for multilingual asr and ast. *Preprint*, arXiv:2509.14128.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025. Qwen2.5-omni technical report. *Preprint*, arXiv:2503.20215.

Haofei Zhao, Yilun Liu, Shimin Tao, Weibin Meng, Yimeng Chen, Xiang Geng, Chang Su, Min Zhang, and Hao Yang. 2024. From handcrafted features to llms: A brief survey for machine translation quality estimation. *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–10.

Vilém Zouhar, Tom Kocmi, and Mrinmaya Sachan. 2025. AI-assisted human evaluation of machine

translation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4936–4950, Albuquerque, New Mexico. Association for Computational Linguistics.

Vilém Zouhar and Tom Kocmi. 2026. Pearmut: Human evaluation of translation made trivial. *Preprint*, arXiv:2601.02933.

A Related Work

Quality Estimation is a broad area of methods for simulating or facilitating costly human quality assessments. Existing work on ST QE focuses on ST metrics – the cases when reference translation is available (Cettolo et al., 2025; Macháček et al., 2023; Post and Hoang, 2025), or on segment-level Direct Assessment (Han et al., 2024; Chen et al., 2023), which is less informative than error span labelling, but possible to perform because there exist evaluation data. To our knowledge, no evaluation data exist for ST error span labelling.

Kocmi et al. (2024b) propose Error Span Annotation, a task in which human annotators highlight spans, erroneous subsequences of candidate translations, and assign them severity classes. They also add segment-level Direct Assessment (DA) quality scores. Error Span Annotation is proved to be easy to learn by annotators, and informative enough to evaluate candidate translation systems. It is widely used in text MT evaluation campaigns (WMT, Kocmi et al., 2024a, 2025), and as a task for automatic systems (Zouhar et al., 2025). There are supervised neural models, such as XCOMET-XL (Guerreiro et al., 2024) for text MT, that are designed for Error Span Annotation and DA. These tasks can also be performed by general-purpose Large Language Models, such as the multimodal Qwen2.5-Omni (Xu et al., 2025), but ST error span labelling with LLMs has not yet been studied.

However, Error Span Annotation is primarily designed for text MT. As Fantinuoli and Prandi (2021) propose, ST is typically supposed to facilitate a specific communication goal of participants at a real communicative event. They propose a user-centric and communication-oriented ST evaluation, which we follow up in this work.

B ST Model Details

The selected models are the following:

1. **Cascades of ASR and LLM for MT:** We select Whisper-large-v3 (Radford et al., 2023) as ASR and Gemma3:12b-it-qat as MT (Team et al., 2025) for all directions except for En→He where LLama3:70b (Grattafiori et al., 2024) scores better.

   We used WhisperX (`https://github.com/m-bain/whisperX`).

   We used Gemma3 on complete WhisperX segments that fit into 3000 characters. Llama3 processed the whole document in one step.

2. **End-to-end speech-to-text:** Canary-v2-1B (Sekoyan et al., 2025) for all language directions except for En→He, for which there is no publicly available end-to-end ST. We used the script for chunked inference: `https://github.com/NVIDIA-NeMo/NeMo/blob/main/examples/asr/asr_chunked_inference/aed/speech_to_text_aed_chunked_infer.py`.

3. **Simultaneous systems:** We use SimulStreaming `https://github.com/ufal/SimulStreaming` (Macháček and Polák, 2025) that integrates end-to-end speech-to-text Whisper-large-v3 for Cs→En, and Whisper-large-v3 as En ASR and Tower+-9B (Rei et al., 2025) for En→Cs and En→De, and Gemma2-9B-it for En→He.

We selected the ST systems as representatives of state of the art for given language directions from the three typically used groups of models that are publicly available for self-hosting on an affordable hardware, where possible. We therefore selected models that have maximum 12B parameters and can be deployed on one NVIDIA A40 GPU, with an exception of Llama3:70b, which we used via a self-hosted OpenWebUI service that used 80GB H100 GPU in the backend.

C Dataset and Human Annotation Details

See Tab. 4 and Fig. 1.

D STEL Model Details

XCOMET-XL. `XCOMET-XL`[4] is run with batch size 8 on a single NVIDIA A100-SXM4-40GB GPU.

[4] 🤗 Unbabel/XCOMET-XL

system	Cs→En					En→Cs					En→De					En→He				
	C	M	N	R	T	C	M	N	R	T	C	M	N	R	T	C	M	N	R	T
cascade	8	24	18	20	70	20	18	5	2	45	12	9	0	1	22	2	14	11	0	27
end2end	34	51	25	19	129	24	27	3	1	55	17	14	2	5	38	-	-	-	-	-
simultaneous	42	69	20	18	149	25	23	13	6	67	28	14	3	5	50	8	40	48	18	114

Table 4: Counts of annotated error spans and categories (Critical, Minor, Negligible, Redundancy, Total).

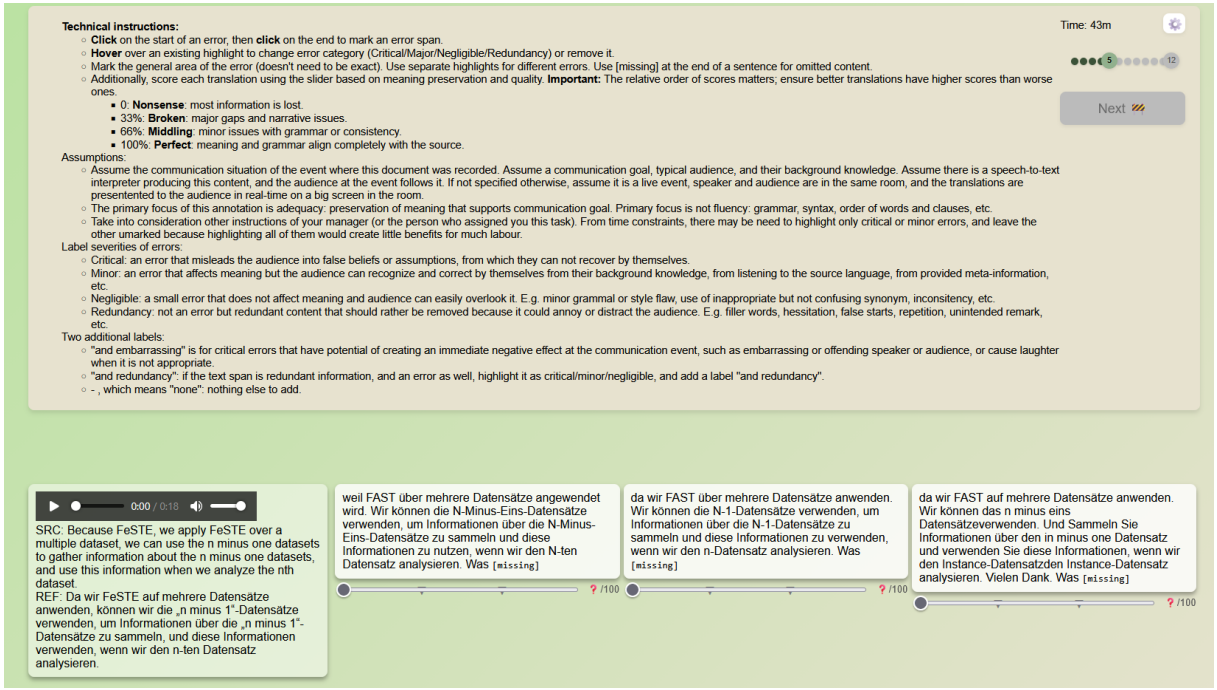

Figure 1: Screenshot of the annotation platform with instructions for the annotators. It is a version of Pearmut (Zouhar and Kocmi, 2026) that we extended with our STEL protocol.

Qwen2.5-Omni. `Qwen2.5-Omni-7B`[5] uses greedy decoding and `max_new_tokens=512`. Each call consists of a system prompt and a modality-specific user message. The prompts are listed in Fig. 3 for prompts with context and Fig. 2.

D.1 Qwen Parameter Selection

For inspecting results in Section 4.3, we selected the Qwen parameters in the following way: (1) we excluded the less realistic setups with gold transcripts; (2) with severities of errors, because it is our primary task. Secondary task is error detection without classifying the severities. (3) We averaged the scores for all language directions, and selected the micro F1 weighted top performing variant of source modality and context length.

Tab. 5 shows the scores. The selected setup is Qwen with ASR and audio, and with no adjacent context.

[5] 🤗 Qwen/Qwen2.5-Omni-7B

E Detailed Results

Tabs. 6 and 7 summarize scores of all inspected model variants, sorted by **F1 uw** or by Kendall's τ.

F Word Alignment

For the ASR to ST candidate alignment in Section 4.4, we used AwesomeAlign (`https://github.com/neulab/awesome-align`) with the model `bert-base-multilingual-cased`.

For tokenization, we used `tokenizer.perl`, version 1.1, from Moses (Koehn et al., 2007), with the known language ID parameter for the processed language.

Avg. of all langs.	**F1 w**	**F1 uw**	Prec.	Rec.	ma F1w	ma F1uw	Pear.	Spear.	**Kend.**
Qwen (ASR+audio)	19.8	23.9	16.6	50.9	17.8	37.6	27.3	25.9	22.0
Qwen (ASR)	17.3	21.5	14.8	50.5	16.1	32.8	8.9	11.6	9.6
Qwen (ASR, ctx=1)	17.4	21.5	14.1	62.6	16.4	30.9	12.1	10.6	9.0
Qwen (audio)	18.3	21.2	14.8	48.6	19.2	37.5	37.4	29.9	25.3
Qwen (ASR, ctx=2)	15.7	20.4	13.5	54.9	15.4	32.0	17.8	18.4	15.9
Qwen (audio, ctx=2)	18.0	18.2	14.9	41.1	16.2	34.1	14.5	15.7	13.2
Qwen (ASR, ctx=5)	15.2	18.2	12.7	42.6	15.4	31.9	6.8	11.1	9.4
Qwen (audio, ctx=1)	17.1	18.0	13.7	49.4	16.9	31.0	18.7	12.6	10.7
Qwen (ASR+audio, ctx=1)	16.2	17.0	13.9	30.6	15.8	37.6	11.8	11.9	10.3
Qwen (ASR+audio, ctx=2)	16.4	16.8	15.2	23.0	15.6	39.9	8.7	10.5	9.2
Qwen (ASR+audio, ctx=5)	14.5	13.9	13.2	16.4	13.8	41.9	-5.5	-7.5	-6.5
Qwen (audio, ctx=5)	14.3	12.3	12.2	26.7	13.2	34.2	0.9	-0.3	-0.1

Table 5: Qwen results, average of all language directions, from which we selected the primary variant. Scores are averaged across all the ST systems, and sorted in decreasing order by micro F1 weighted. Scores description: **F1 w** = micro F1 weighted; **F1 uw** = micro F1 unweighted; Precision; Recall; macro F1 weighted/unweighted; Pearson, Spearman, Kendall. Left-most column shows the parameters, ctx is the number of adjacent segments as context.

Avg. of all langs.	**F1 w**	**F1 uw**	Prec.	Rec.	ma F1w	ma F1uw	Pear.	Spear.	**Kend.**
XCOMET (gold)	20.7	26.2	18.8	57.2	16.0	27.2	27.8	34.9	25.5
Qwen (ASR+audio, no sev)	-	24.7	17.3	49.3	-	39.6	29.0	27.5	23.4
Qwen (gold, no sev)	-	24.4	16.6	55.9	-	39.6	43.6	40.4	33.3
Qwen (gold+audio)	20.0	24.0	16.0	59.0	19.7	36.9	43.0	35.2	29.7
Qwen (ASR+audio)	19.8	23.9	16.6	50.9	17.8	37.6	27.3	25.9	22.0
Qwen (gold, ctx=2, no sev)	-	23.9	16.0	56.5	-	38.9	46.2	40.2	34.0
XCOMET (ASR)	17.9	23.3	17.3	40.8	13.1	33.5	18.1	21.3	15.7
Qwen (gold)	18.3	23.1	15.4	56.6	17.9	35.1	44.1	39.6	32.7
Qwen (gold, ctx=1)	18.3	23.1	15.0	66.3	17.9	32.5	44.0	33.8	28.3
Qwen (gold, ctx=1, no sev)	-	22.7	15.2	60.3	-	34.2	44.9	37.8	31.7
Qwen (ASR, ctx=1, no sev)	-	22.2	15.0	57.5	-	32.4	13.4	13.2	11.2
Qwen (gold, ctx=2)	18.5	22.1	14.5	58.4	18.4	34.2	46.7	40.3	34.4
Qwen (gold+audio, no sev)	-	22.1	15.0	52.1	-	37.7	44.5	37.1	31.4
Qwen (ASR, ctx=2, no sev)	-	22.0	15.2	51.8	-	34.4	14.8	15.1	12.9
Qwen (gold, ctx=5, no sev)	-	21.7	15.1	45.9	-	37.8	40.2	34.0	29.0
Qwen (ASR)	17.3	21.5	14.8	50.5	16.1	32.8	8.9	11.6	9.6
Qwen (ASR, ctx=1)	17.4	21.5	14.1	62.6	16.4	30.9	12.1	10.6	9.0
Qwen (audio)	18.3	21.2	14.8	48.6	19.2	37.5	37.4	29.9	25.3
Qwen (ASR, ctx=2)	15.7	20.4	13.5	54.9	15.4	32.0	17.8	18.4	15.9
Qwen (gold, ctx=5)	16.8	20.4	13.8	50.5	16.2	34.9	39.8	32.9	28.0
Qwen (audio, no sev)	-	20.3	15.2	42.3	-	38.4	38.5	30.8	25.8
Qwen (ASR, no sev)	-	20.0	14.1	44.6	-	32.9	8.2	10.8	8.5
Qwen (ASR, ctx=5, no sev)	-	19.3	14.0	40.0	-	33.2	8.0	9.6	8.2
Qwen (gold+audio, ctx=2)	19.0	19.0	16.2	29.0	18.7	39.0	16.1	19.1	16.2
Qwen (audio, ctx=2)	18.0	18.2	14.9	41.1	16.2	34.1	14.5	15.7	13.2
Qwen (ASR, ctx=5)	15.2	18.2	12.7	42.6	15.4	31.9	6.8	11.1	9.4
Qwen (audio, ctx=1)	17.1	18.0	13.7	49.4	16.9	31.0	18.7	12.6	10.7
Qwen (gold+audio, ctx=1, no sev)	-	17.9	14.2	35.7	-	37.1	26.8	16.3	13.8
Qwen (gold+audio, ctx=1)	17.4	17.8	14.4	36.3	17.6	36.8	25.0	14.9	12.5
Qwen (ASR+audio, ctx=1, no sev)	-	17.5	14.0	28.8	-	39.3	17.8	13.7	11.8
Qwen (audio, ctx=2, no sev)	-	17.1	14.8	34.0	-	35.2	12.7	14.8	12.5
Qwen (ASR+audio, ctx=1)	16.2	17.0	13.9	30.6	15.8	37.6	11.8	11.9	10.3
Qwen (audio, ctx=1, no sev)	-	16.9	12.9	41.8	-	32.1	24.6	19.9	17.2
Qwen (ASR+audio, ctx=2)	16.4	16.8	15.2	23.0	15.6	39.9	8.7	10.5	9.2
Qwen (gold+audio, ctx=2, no sev)	-	16.3	14.4	24.0	-	38.6	18.9	20.8	17.8
Qwen (ASR+audio, ctx=5, no sev)	-	14.9	17.0	17.9	-	40.9	-7.5	-5.8	-4.9
Qwen (gold+audio, ctx=5)	14.3	14.7	13.3	20.3	16.0	42.8	3.8	4.5	4.2
Qwen (gold+audio, ctx=5, no sev)	-	14.0	14.5	19.6	-	41.1	0.7	1.7	1.8
Qwen (ASR+audio, ctx=5)	14.5	13.9	13.2	16.4	13.8	41.9	-5.5	-7.5	-6.5
Qwen (ASR+audio, ctx=2, no sev)	-	13.4	13.5	17.2	-	38.6	8.8	11.0	9.5
Qwen (audio, ctx=5, no sev)	-	13.3	11.9	24.4	-	37.1	-1.7	-5.6	-4.9
Qwen (audio, ctx=5)	14.3	12.3	12.2	26.7	13.2	34.2	0.9	-0.3	-0.1

Table 6: Scores of all inspected XCOMET and Qwen variants. Averaged across all language directions, sorted by **F1 uw** in descending order. Description of columns as in Tab. 5.

Avg. of all langs.	**F1 w**	**F1 uw**	Prec.	Rec.	ma F1w	ma F1uw	Pear.	Spear.	**Kend.**
Qwen (gold, ctx=2)	18.5	22.1	14.5	58.4	18.4	34.2	46.7	40.3	34.4
Qwen (gold, ctx=2, no sev)	-	23.9	16.0	56.5	-	38.9	46.2	40.2	34.0
Qwen (gold, no sev)	-	24.4	16.6	55.9	-	39.6	43.6	40.4	33.3
Qwen (gold)	18.3	23.1	15.4	56.6	17.9	35.1	44.1	39.6	32.7
Qwen (gold, ctx=1, no sev)	-	22.7	15.2	60.3	-	34.2	44.9	37.8	31.7
Qwen (gold+audio, no sev)	-	22.1	15.0	52.1	-	37.7	44.5	37.1	31.4
Qwen (gold+audio)	20.0	24.0	16.0	59.0	19.7	36.9	43.0	35.2	29.7
Qwen (gold, ctx=5, no sev)	-	21.7	15.1	45.9	-	37.8	40.2	34.0	29.0
Qwen (gold, ctx=1)	18.3	23.1	15.0	66.3	17.9	32.5	44.0	33.8	28.3
Qwen (gold, ctx=5)	16.8	20.4	13.8	50.5	16.2	34.9	39.8	32.9	28.0
Qwen (audio, no sev)	-	20.3	15.2	42.3	-	38.4	38.5	30.8	25.8
XCOMET (gold)	20.7	26.2	18.8	57.2	16.0	27.2	27.8	34.9	25.5
Qwen (audio)	18.3	21.2	14.8	48.6	19.2	37.5	37.4	29.9	25.3
Qwen (ASR+audio, no sev)	-	24.7	17.3	49.3	-	39.6	29.0	27.5	23.4
Qwen (ASR+audio)	19.8	23.9	16.6	50.9	17.8	37.6	27.3	25.9	22.0
Qwen (gold+audio, ctx=2, no sev)	-	16.3	14.4	24.0	-	38.6	18.9	20.8	17.8
Qwen (audio, ctx=1, no sev)	-	16.9	12.9	41.8	-	32.1	24.6	19.9	17.2
Qwen (gold+audio, ctx=2)	19.0	19.0	16.2	29.0	18.7	39.0	16.1	19.1	16.2
Qwen (ASR, ctx=2)	15.7	20.4	13.5	54.9	15.4	32.0	17.8	18.4	15.9
XCOMET (ASR)	17.9	23.3	17.3	40.8	13.1	33.5	18.1	21.3	15.7
Qwen (gold+audio, ctx=1, no sev)	-	17.9	14.2	35.7	-	37.1	26.8	16.3	13.8
Qwen (audio, ctx=2)	18.0	18.2	14.9	41.1	16.2	34.1	14.5	15.7	13.2
Qwen (ASR, ctx=2, no sev)	-	22.0	15.2	51.8	-	34.4	14.8	15.1	12.9
Qwen (audio, ctx=2, no sev)	-	17.1	14.8	34.0	-	35.2	12.7	14.8	12.5
Qwen (gold+audio, ctx=1)	17.4	17.8	14.4	36.3	17.6	36.8	25.0	14.9	12.5
Qwen (ASR+audio, ctx=1, no sev)	-	17.5	14.0	28.8	-	39.3	17.8	13.7	11.8
Qwen (ASR, ctx=1, no sev)	-	22.2	15.0	57.5	-	32.4	13.4	13.2	11.2
Qwen (audio, ctx=1)	17.1	18.0	13.7	49.4	16.9	31.0	18.7	12.6	10.7
Qwen (ASR+audio, ctx=1)	16.2	17.0	13.9	30.6	15.8	37.6	11.8	11.9	10.3
Qwen (ASR)	17.3	21.5	14.8	50.5	16.1	32.8	8.9	11.6	9.6
Qwen (ASR+audio, ctx=2, no sev)	-	13.4	13.5	17.2	-	38.6	8.8	11.0	9.5
Qwen (ASR, ctx=5)	15.2	18.2	12.7	42.6	15.4	31.9	6.8	11.1	9.4
Qwen (ASR+audio, ctx=2)	16.4	16.8	15.2	23.0	15.6	39.9	8.7	10.5	9.2
Qwen (ASR, ctx=1)	17.4	21.5	14.1	62.6	16.4	30.9	12.1	10.6	9.0
Qwen (ASR, no sev)	-	20.0	14.1	44.6	-	32.9	8.2	10.8	8.5
Qwen (ASR, ctx=5, no sev)	-	19.3	14.0	40.0	-	33.2	8.0	9.6	8.2
Qwen (gold+audio, ctx=5)	14.3	14.7	13.3	20.3	16.0	42.8	3.8	4.5	4.2
Qwen (gold+audio, ctx=5, no sev)	-	14.0	14.5	19.6	-	41.1	0.7	1.7	1.8
Qwen (audio, ctx=5)	14.3	12.3	12.2	26.7	13.2	34.2	0.9	-0.3	-0.1
Qwen (audio, ctx=5, no sev)	-	13.3	11.9	24.4	-	37.1	-1.7	-5.6	-4.9
Qwen (ASR+audio, ctx=5, no sev)	-	14.9	17.0	17.9	-	40.9	-7.5	-5.8	-4.9
Qwen (ASR+audio, ctx=5)	14.5	13.9	13.2	16.4	13.8	41.9	-5.5	-7.5	-6.5

Table 7: Scores of all inspected XCOMET and Qwen variants. Averaged across all language directions, sorted by **Kend.** in descending order. Description of columns as in Tab. 5.

With severity — no context

System: You are a machine translation quality estimator. Given *source description* and its translation, you must:

1. Assign an overall quality score between 0.0 (completely wrong) and 1.0 (perfect).
2. Identify error spans in the translation text with their severity.

Respond ONLY with a valid JSON object in this exact format:
{"score": <float 0.0–1.0>, "error_spans": [{"text": "<exact substring>", "severity": "<minor|major|neutral|redundancy>"}, ...]}
If there are no errors, use an empty list for error_spans. Output nothing else.

User (text): [Target – evaluate this]
Source: *source sentence* Translation: *MT output*

User (audio): [Target – evaluate this]
⟨audio⟩ Translation: *MT output*

User (text+audio): [Target – evaluate this]
⟨audio⟩
Transcript: *source sentence* Translation: *MT output*

Without severity — no context

System: You are a machine translation quality estimator. Given *source description* and its translation, you must:

1. Assign an overall quality score between 0.0 (completely wrong) and 1.0 (perfect).
2. Identify error spans in the translation text.

Respond ONLY with a valid JSON object in this exact format:
{"score": <float 0.0–1.0>, "error_spans": [{"text": "<exact substring>"}, ...]}
If there are no errors, use an empty list for error_spans. Output nothing else.

User (text): [Target – evaluate this]
Source: *source sentence* Translation: *MT output*

User (audio): [Target – evaluate this]
⟨audio⟩ Translation: *MT output*

User (text+audio): [Target – evaluate this]
⟨audio⟩
Transcript: *source sentence* Translation: *MT output*

Figure 2: Prompts for Qwen2.5-Omni-7B without context. *source description* is instantiated as *a source sentence* (text), *a source speech audio* (audio), or *a source speech audio with its transcript* (text+audio).

With severity — with context (n sentences before/after)

System: You are a machine translation quality estimator. Given surrounding context sentences and a target *source description* with its translation, you must:

1. Assign an overall quality score between 0.0 (completely wrong) and 1.0 (perfect).
2. Evaluate ONLY the target translation (marked [Target]). Identify error spans in the translation text with their severity.

Respond ONLY with a valid JSON object in this exact format:
{"score": <float 0.0-1.0>, "error_spans": [{"text": "<exact substring>", "severity": "<minor|major|neutral|redundancy>"}, ...]}
If there are no errors, use an empty list for error_spans. Output nothing else.

User (text): [Context before]
Source: *ctx* src_1 Translation: *ctx* MT_1 $\cdots$
[Target – evaluate this]
Source: *source sentence* Translation: *MT output*
[Context after]
Source: *ctx* src_1 Translation: *ctx* MT_1 $\cdots$

User (audio): [Context before]
⟨audio⟩ Translation: *ctx* MT_1 $\cdots$
[Target – evaluate this]
⟨audio⟩ Translation: *MT output*
[Context after]
⟨audio⟩ Translation: *ctx* MT_1 $\cdots$

User (text+audio): [Context before]
⟨audio⟩ Transcript: *ctx* src_1 Translation: *ctx* MT_1 $\cdots$
[Target – evaluate this]
⟨audio⟩
Transcript: *source sentence* Translation: *MT output*
[Context after]
⟨audio⟩ Transcript: *ctx* src_1 Translation: *ctx* MT_1 $\cdots$

Without severity — with context (n sentences before/after)

System: You are a machine translation quality estimator. Given surrounding context sentences and a target *source description* with its translation, you must:

1. Assign an overall quality score between 0.0 (completely wrong) and 1.0 (perfect).
2. Evaluate ONLY the target translation (marked [Target]). Identify error spans in the translation text.

Respond ONLY with a valid JSON object in this exact format:
{"score": <float 0.0-1.0>, "error_spans": [{"text": "<exact substring>"}, ...]}
If there are no errors, use an empty list for error_spans. Output nothing else.

User messages identical to the with-severity variant above.

Figure 3: Prompts for Qwen2.5-Omni-7B with context. We evaluate context sizes $n \in \{1, 2, 5\}$; the template is the same for all sizes. *source description* is instantiated as *a source sentence* (text), *a source speech audio* (audio), or *a source speech audio with its transcript* (text+audio).