MLLP-VRAIN UPV system for the IWSLT 2026 Simultaneous Speech Translation task

Jorge Iranzo-Sánchez*, Gerard Mas-Mollà*
Adrià Giménez, Jorge Civera, Albert Sanchis, Alfons Juan
Machine Learning and Language Processing, VRAIN, Universitat Politècnica de València
{jorirsan,gemamol}@upv.es

Abstract

This work describes the participation of the MLLP-VRAIN research group in the shared task of the IWSLT 2026 Simultaneous Speech Translation track. Our submission utilizes the recently released Parakeet and Qwen 3.5 models to create a robust, cascaded solution for long-form SimulST through the use of adaptive "black-box" policies. We explore relaxations of these policies to achieve better quality-latency trade-offs. Compared to last year, we participate on all language directions. In addition to this, for the En→De, It, Zh directions we also participate in this year's new context track employing a combination of ASR word-boosting and a RAG mechanism of offline pre-translated exemplars to guide generation and enrich our system with domain-specific context. Finally, we provide a detailed latency analysis of our system. Compared to last year, results on the MCIF En→De test set shows a substantial quality improvement of +5.82 XCOMET-XL. Our context track processing further improves performance by +1.03.

1 Introduction

In this paper we describe the participation of the MLLP-VRAIN research group in the shared tasks of the 23th International Conference on Spoken Language Translation (IWSLT) (Adelani et al., 2026). Building on our previous participation on last year IWSLT SimulST Track (Iranzo-Sánchez et al., 2025b), we focus on cascaded solutions for SimulST. This choice is motivated by the IWSLT results from last year (Abdulmumin et al., 2025), where cascaded systems achieved the best performance, as well as the strong results reported on the recently introduced Hearing2Translate (Papi et al., 2025) benchmark and the flexibility that cascaded approaches give us in the choice of our components. We also believe that the creation of a strong cascaded system may show which audio and text components are optimal for the creation of a derived SpeechLLM further down the line. Figure 1 shows the overall architecture of our system which will be described in more detail in the following sections.

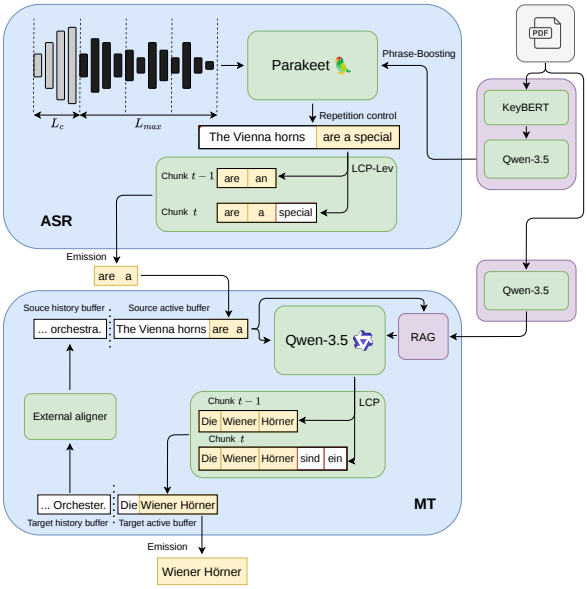

Figure 1: System diagram of our cascaded system for the SimulST track

This year, we participate in all language directions and latency regimes. Furthermore, we participate in the newly introduced extra context track for the En→De, It, Zh directions, proposing mechanisms to leverage this context for both the ASR and MT components. The evaluation metrics we use are XCOMET-XL (Guerreiro et al., 2024) and chrF (Popović, 2015; Macháček et al., 2023) for translation quality and LongYAAL (Polák et al., 2026) for latency.

2 Surface Level Black Box Policies for SimulST

Black-box emission policies are a well-established approach in SimulST. They do not rely on direct access to internal model information and can therefore be applied to any offline model without addi-

Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026), pages 212–226

tional training, covering both fixed policies such as Wait-k (Ma et al., 2019) and Hold-n (Liu et al., 2020a) and adaptive ones such as Longest Common Prefix (LCP) (Liu et al., 2020b). Notably, (Mas-Mollà et al.) recently showed that the combination of offline SOTA ASR models with black box policies achieved competitive results on streaming ASR benchmarks compared to more specialized solutions. Recent winners of past editions of IWSLT have also demonstrated the effectiveness of LCP for the creation of SOTA SimulST systems (Polák et al., 2022, 2023; Macháček and Polák, 2025).

One of the most widely used adaptive policies is the previously mentioned LCP, which accepts the longest common prefix across consecutive model generations as a valid output. However, in practice, LCP often causes high latency spikes when the system has a high degree of oscillation between tokens across generations. However, in many cases, these oscillations have little impact on the final quality. Last year, we relaxed our ASR LCP policy to account for this phenomenon by accepting tokens based on a Levenshtein distance threshold, a method we refer to as LACP (Iranzo-Sánchez et al., 2025b). We propose that this relaxation can be taken further to achieve lower latency at a modest quality trade-off. We refer to this further relaxation as *Soft LCP* (SLCP). SLCP is motivated by two observations. First, the Ratcliff/Obershelp (RO) pattern recognition algorithm (Ratcliff et al., 1988) could provide a more suitable string similarity measure than Levenshtein distance for this task. Second, there frequently exist "anchor" tokens that remain stable across generations even when surrounding tokens vary slightly without affecting the final quality. Figure 2 illustrates in more detail the SLCP policy. The idea is to identify "anchor" tokens via RO and greedily accept all preceding tokens, propagating committed output more frequently than regular LCP would allow. We define γ as the maximum allowable gap (in tokens) between unstable tokens for anchor propagation, and σ as the minimum similarity score for a token to qualify as an anchor. For all experiments in this work, we set $\gamma = 3$ and $\sigma = 0.6$.

3 ASR Component

Speech Foundational Models Our ASR component was chosen based on the results of public ASR systems on the benchmarks available at the HuggingFace Open ASR Leaderboard (Srivastav et al., 2026). Based on the language pairs considered in the competition and attending to our computing limitations, we needed a lightweight multilingual model capable of producing high quality transcriptions under streaming conditions. We finally selected Parakeet (Sekoyan et al., 2025)[1] as our ASR component. Our decision of using Parakeet can be explained by two main reasons. The first reason is that, as a multilingual model, it supports Czech ASR, and so it allows us to participate in the Cs→En direction. The second reason is that it is a lightweight model with only 0.6B parameters, and as such it allows the usage of heavier LLMs as MT systems. Apart from that, since Parakeet already achieves competitive results compared to other ASR systems on the development data provided by the organizers, we did not perform any finetuning process to the model.

The adaptation of Parakeet to streaming was performed following (Mas-Mollà et al.), where three main components are applied to perform online decoding. First, incremental data ingestion is managed by an acoustic buffer that receives fixed-length chunks of size L_c. By defining a maximum buffer size L_{max}, the input audio buffer behaves as a sliding window, growing cumulatively until L_{max} is reached. From this point on, whenever a new chunk is added, it pushes the oldest chunk out of the buffer. Following this idea, the acoustic input X_t at any decoding step t can be formally expressed as $X_t = [\max(0, t \cdot L_C - L_{max}), t \cdot L_c]$. Then, since the entire acoustic buffer is fed to the model at every decoding step, and since we set that $L_c \ll L_{max}$, the model is forced to process acoustic information that has already been processed in previous decoding steps, thus leading to repetitions in the output transcription. This phenomenon is mitigated by a timestamp-based repetition control. We leverage Parakeet's capacity to predict token durations to keep track of emission times at the output of the model, which allows us to filter any repetition based on the information of previous decoding steps. Finally, once the output has been filtered, it is added to an output buffer governed by an emission policy. Additionally, we used the ALSD++ beam search decoding implementation of NeMo (Grigoryan et al., 2025) with beam size 32.

Streaming ASR with emission policies The candidate policies considered for the ASR model were LCP, LACP and SLCP. Fixed policies such as Wait-

[1]Model: nvidia/parakeet-tdt-0.6b-v3

Gen.	Hypothesis Tokens								
g_1	the	ether	near	Plasencia					
g_2	the	weather	in	Palencia	reminds	me	of	Valencia	and
Out	the	weather	in	Palencia	⟵ *accepted via anchor propagation*				

Figure 2: SLCP anchor propagation example with maximum gap $\gamma = 2$ and $\sigma = 0.6$. Palencia and Valencia are identified as possible **anchor tokens** of Plasencia with scores 0.82 and 0.70 respectively. Valencia is not a valid anchor, since the length of the chunk reminds me of is $> \gamma$, remaining unstable and not being committed. All preceding tokens of Palencia are greedily accepted, emitting the weather in Palencia. If using LCP, only the first token the would have been accepted. Note that weather is also a valid anchor with respect ether.

k and Hold-n were discarded based on the results of preliminary informal experiments, as they showed poorer WER/latency trade-offs, particularly in lower latency configurations. Regarding LACP, we set the Levenshtein threshold to $\tau = 2$ following (Mas-Mollà et al.) and the findings of our participation last year. The candidate policies were tested by sweeping the chunk size L_c from 0.64 to 2.00 seconds and measuring both computational aware and unaware latency values. The latency figures are computed by aligning the output transcripts with an external HMM-based system using the TLK toolkit (del Agua et al., 2014). Then, we compute the latency values using these alignments and the system emission timestamps. All experiments were performed on a node with a NVIDIA RTX 4090 GPU and an Intel Core 10920X CPU. As seen in the results plotted in Figure 3, LCP and LACP stand out in terms of WER, consistently yielding better results than those of SLCP. As for the latency results, LACP and SLCP perform similarly on low latency configurations, with SLCP yielding the best results in high latency configurations. In light of the results, we selected LACP as the best policy, as it achieves competitive WER results while maintaining low latency figures.

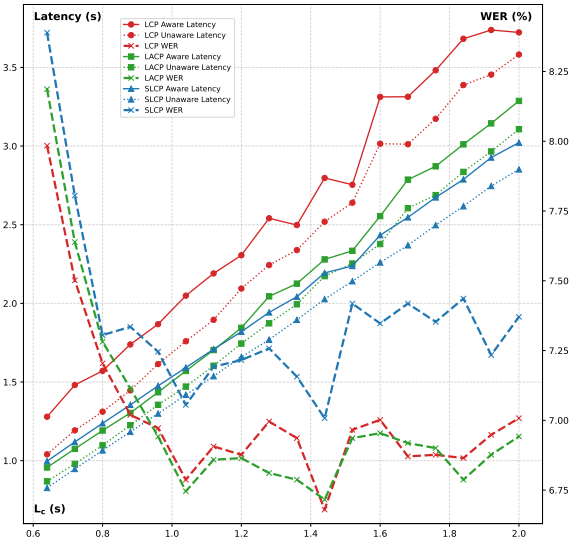

Figure 3: Latency (left) vs. WER (right) trade-off of $L_c = \{0.64, \ldots, 2.00\}$ sweep in MCIF.

4 MT Component

LLMs: newer, bigger, better In the context of SimulST, we are restricted in our choice of foundational models, as we need to be able to run them in real time (RTF <1) to have a true streaming model. Last year we made use of an encoder-decoder approach by using NLLB (Costa-jussà et al., 2022), since we found it to be a good middle-ground in terms of model size, speed and performance. However, recent WMT evaluation have shown the performance of encoder-decoders such as NLLB to be subpar compared to state of the art LLMs for offline MT (Kocmi et al., 2024, 2025). Consequently, we were motivated this year to explore more in depth current LLMs for SimulST, following the plethora of recent work that demonstrate their effectiveness on this task (Koshkin et al., 2024a,b; Raffel et al., 2024; Guo et al., 2025; Cheng et al., 2025).

We conducted an initial survey of publicly available open-weight LLMs and selected several candidates for a preliminary, informal evaluation: HuanYan-MT-1.5 (Zheng et al., 2025), EuroLLM (Ramos et al., 2026) Tower+ (Rei et al., 2025), TranslateGemma (Finkelstein et al., 2026) and Qwen 3.5 (Qwen Team, 2026). Preliminary experiments revealed stability issues in several models of the first model families (e.g., frequent refusals and oscillatory outputs), leading us to focus on TranslateGemma and Qwen 3.5 for further ex-

Model	YAAL ↓		XCOMET ↑	chrF ↑
	CU	CA		
UPV IWSLT25	3.18	3.43	86.85	60.23
TranslateGemma (4B)	3.04	3.20	89.45	57.92
Qwen 3.5 (4B)	2.99	3.33	89.48	58.06
Qwen 3.5 (9B)	2.94	3.39	89.57	58.55
Qwen 3.5 (9B, fp8)	2.92	**3.19**	90.19	57.79
Qwen 3.5 (27B, fp8)	2.84	4.21	90.86	**59.25**
Qwen 3.5 (27B, int4[2])	**2.87**	3.46	**91.09**	58.89

Table 1: MCIF En→De quality–latency results of initial long-form streaming systems. LLMs based models use an emission policy of Hold-3 and for the MT and the ASR component of our last year submission.

ploration as our primary machine translation models. Table 1 presents preliminary XCOMET, chrF and YAAL results for the 4B variants of both used model families, as well as for Qwen3.5 9B and 27B. For the latter two, different quantization methods are also tested to be able to use these model sizes on the <24GB consumer grade GPUs we have available. Results show comparable performance across the two families. This led us to select the Qwen 3.5 family as the backbone of our MT component. This decision was further reinforced upon discovering that TranslateGemma had been trained on a fixed prompt template and thus exhibited poor robustness to prompt variations and external context insertion. Based on the results, we will use the quantized 27B for our final system submission. We also leveraged the 9B-fp8 variant on part of the experimentation.

MT Buffer control Last year, we finetuned our MT component to emit "sentinel" tokens following Iranzo-Sánchez et al. (2024) which served to indicate a target side end of sentence. This would then trigger a history buffer update that would identify the corresponding source-size end of sentence by obtaining an alignment through the usage of cross-attention maps as proxy alignments (Li et al., 2019). This year, we get rid of this mechanism. This decision is based on two observations. First, in our previous system, where the source and output streams were cased (unlike the original paper, which assumed lowercase), the system typically generated a sentinel token after a strong punctuation mark (`!?.`) was emitted. Thus, if we can identify when this punctuation is generated, we can directly use it to trigger the alignment mechanism. Second, while we use LLMs, we still require source-target alignments. To our knowledge, obtaining reliable alignments from text-based LLMs using attention maps remains an open question. Although some works suggest that alignments can be achieved through a discrete or generative approach by self-prompting the model (Mao and Yu, 2024), we decided to use a lightweight external aligner, similar to how external CTC alignments are used in ASR to obtain timestamps. This avoids potential model hallucinations during the alignment task. In our experimentation we run SimAlign (Jalili Sabet et al., 2020) with XLM-Roberta Base (∼125M) as a backend model (Conneau et al., 2020) quantized to int8 and running on CPU to minimize GPU memory usage and computing overhead. We set the maximum history buffer to 20 sentences or 1024 words (or characters for En→Zh) and eject the oldest sentence when this limit is surpassed in either the source or target buffer.

Decoding Due to model size and to keep computational costs at a reasonable range, we decided to use greedy search for our MT component as the decoding algorithm. We also explored the use of Minimum Bayes Risk (MBR) decoding for both ASR and MT component; however, mixed results led us to discard this approach in our final submissions. The results of this exploration are detailed in Appendix A.

Prevention of Catastrophic Failure We identify two rare cases in which the MT system fails catastrophically and cannot recover. In certain configurations, early termination may occur within a document: the system stops emitting tokens for the current segment due to an overconfident prediction that produces strong punctuation. This, in turn, causes the EOS token to dominate subsequent emissions, even if the source stream continues to grow. To mitigate this issue, we allow the system to rewrite the last two previously emitted tokens when such a condition is detected. With this mechanism in place, we no longer observe this phenomenon, and the introduced flickering remains minimal. In addition, we observe occasional oscillatory hallucinations. To address these, we adopt the temperature fallback mechanism proposed in Whisper (Radford et al., 2023). Unlike the original work, we only trigger the temperature fallback if the gzip compression ratio of the emitted tokens exceeds 2.4.

[2]Model: Intel/Qwen3.5-2B-int4-AutoRound

5 Cascaded system

Optimal buffer sizes Figure 4 presents results for all language directions of MCIF, comparing the usage of LACP/SLCP in the ASR component and LCP/SLCP in MT. Focusing first on the ASR component, we observe that LACP and SLCP yield very similar latency and quality across different values of L_c, with LACP showing a slight overall advantage. For the MT component, comparing LCP and SLCP, we find that SLCP can achieve a considerable reduction in average YAAL, with latency improvements of approximately 0.3–1 seconds compared to LCP. However, these gains come at the cost of noticeable drops in XCOMET, particularly at smaller L_c values. In last year's evaluation, our system achieved significantly lower latency than the winning system, but at the expense of translation quality. Human evaluation, however, showed a clear preference for the higher-latency system. This suggests that in our case, mid-range chunk size configurations with LCP may be preferable to lower-latency SLCP alternatives by human preference. Based on these observations, we ultimately adopt LACP for our ASR systems and LCP for our MT system, leaving further investigation of SLCP for more latency-constrained scenarios. Regarding the acoustic chunk size, we select $L_c = 1.04$s for our high latency configurations in all language directions, as the quality seems to peak around this value.

Re-translation for Low Latency As shown in Figure 4, for all our tested configurations there is no models for which the YAAL <2 seconds restrictions for the low latency track is valid. As such, we adopt a simple mask-k re-translation based approach (Arivazhagan et al., 2020a,b) on the MT component. By applying this technique, we are able to reduce YAAL of systems with low L_c values to participate in this latency regime. More specifically, at each step, we take the non committed suffix resulting from the LCP based policy, remove the last k tokens and consider the resulting output suffix as a "speculative" emission which is not committed to the internal translation buffer, and for which tokens can be overwritten in the next generation. Figure 5 shows both computational aware and unaware YAAL-Normalized Erasure trade-off across all language directions for Qwen 3.5 variant, $L_c = 0.64$ and $k \in \{0, \ldots, 3\}$. Based on the results, we select $k = 2$ as the optimal choice, as it obtains the best generalizable latency-flickering

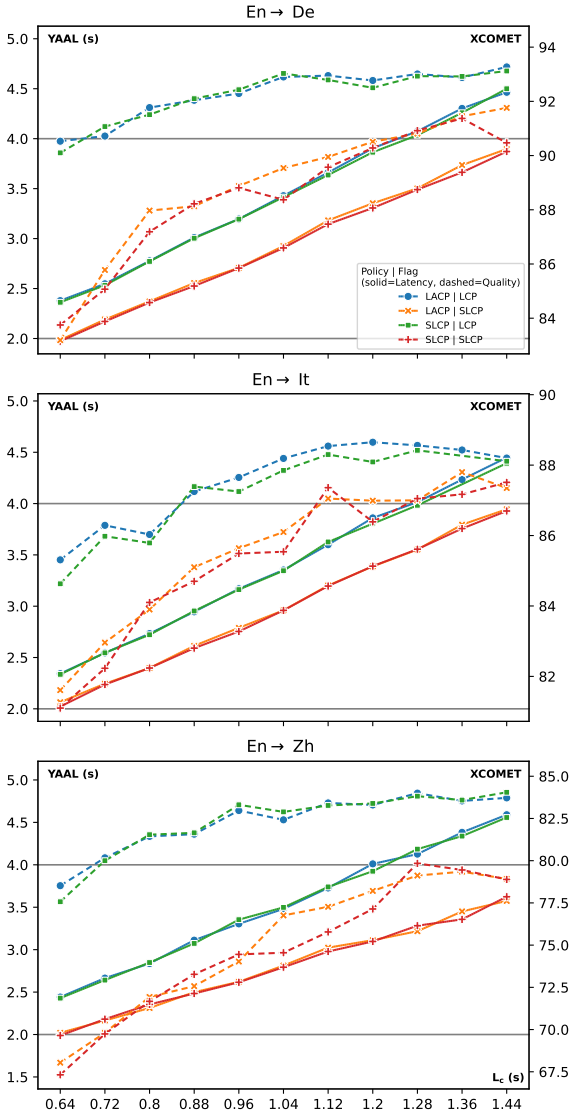

Figure 4: YAAL (left) vs. XCOMET (right) trade-off of $L_c \in \{0.64, ..., 1.44\}$ sweep on the MCIF (Papi et al., 2026) IWSLT 2026 test set with Qwen 27B.

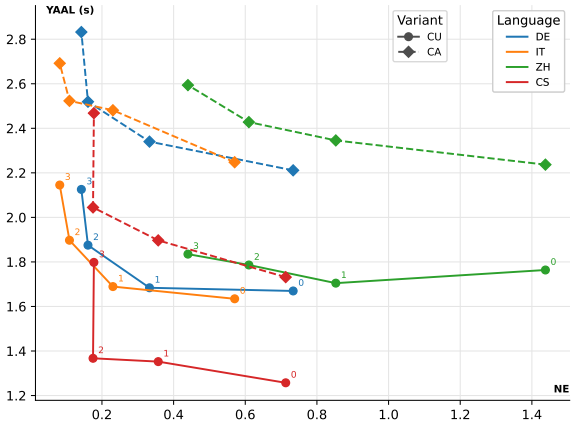

Figure 5: Normalized erasure (x-axis) vs. YAAL (y-axis) trade-off for Mask-k with $k \in \{0, \ldots, 3\}$ on the LCP speculative emission using Qwen3.5 9B with $L_c = 0.64$s

ratio across all languages directions and fix $L_c = 0.64$ for our low latency systems. [3]

6 Context Track

We also participate in this year extra context track which allows participants to use additional information from associated paper PDFs.

6.1 Word boosting for ASR

For ASR, we leverage the efficient GPU-accelerated Phrase-Boosting (GPU-PB) (Andrusenko et al., 2025) implementation for the Parakeet models backed by Nvidia NeMo to guide ASR generation by shallow fusion from an extracted keyword list from the given PDFs. For keyword list extraction, we make use of a two-step strategy. First, we use KeyBERT (Grootendorst, 2020)[4] to get an initial set of keywords. Then, we reuse the same Qwen 3.5 model used as our MT backbone to refine the keywords extracted in the first step. Also, compared to the organizer's baseline, we make use of the whole document instead of just the title, author's list and abstract section. More specifically, in our extraction pipeline, all paper sections except the references are first extracted and cleaned with some formatting regexes. Then, the full cleaned text is chunked into overlapping segments. These segments are then fed to KeyBERT, which extracts the initial list of keywords. Finally, the keywords are passed to the LLM to be refined. Additionally, we tested two levels of granularity for the application of GPU-PB: dataset and document-level.

Figure 6 shows results sweeping across GPU-PB α interpolation parameter, comparing our keyword extraction method versus the organizers baseline at different levels of granularity. We see that word-boosting at the document level obtains lower WER results overall compared to the dataset level. Furthermore, we consider $\alpha = 0.6$ to be optimal on the MCIF dev set, as it reduces WER results from 7.2 to 6.4. We set this α value for our final model.

6.2 RAG with lexical retrieval for MT

Since the context track PDFs only provide source-language information, we provide the MT model with additional contextual guidance by pretranslating the document at the sentence level, creating an offline translation memory that can be queried at runtime. This component is intended to provide

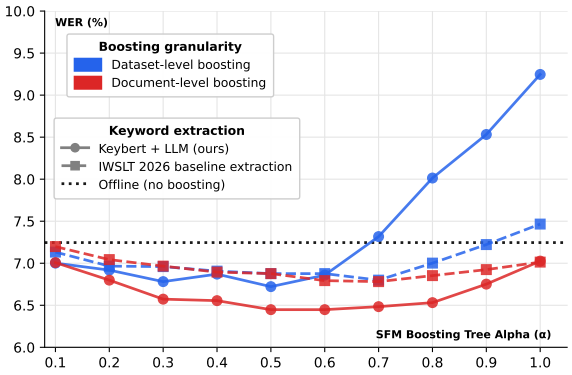

Figure 6: WER (x-axis) vs. SFM boosting tree alpha (α) (y-axis) on the MCIF IWSLT 2026 test set for greedy search and $L_c = 0.96s$.

look-ahead hints about upcoming content, while helping to preserve consistent and accurate terminology. Our hypothesis is that this approach can improve both translation quality and latency, particularly in scenarios where the system lacks full context and may otherwise struggle to disambiguate terms correctly. For our retrieval mechanism, we take the list of source and target sentence translation pairs. Then, before starting decoding, we generate a BM25s (Lù, 2024) index per document with default parameters by creating a lowercased and lex-normalized copy of the source sentences. Then at run time, at each timestep we query the index with the current source sentence content and retrieve the indices of the top-k best matches r_k. We then use r_k to retrieve the best translation pairs and inject them as context into our prompt. Our selection for a lexical based approach is based on the demonstrated effectiveness of using BM25 for domain adaptation in offline translation by Agrawal et al. (2023). In addition to this, the low cost of BM25s allows us to run the query on the CPU and avoid the training and the higher inference cost compared to a neural based solution such as that of RAAST (Luo et al., 2026). We tried three different configurations to inject the retrieved sentences: at the header position before the system prompt, after the source sentence context and before the source sentence context. From these three configurations, we make use of the latter, as we observed that with the other two, the model had a tendency to start hallucinating additional source and target pairs.

Table 2 reports YAAL and XCOMET results obtained by sweeping over different top-k values for MCIF En→De, It, Zh with $L_c = 0.96$, comparing them to base context-less systems and ASR

[3] Arivazhagan et al. (2020b) considers a model to have "few revisions" if NE < 0.2.

[4] Model: sentence-transformers/all-MiniLM-L6-v2

word-boosted ones. As it can be observed, incorporating the RAG mechanism consistently improves XCOMET scores while maintaining latency comparable to both context-free systems and ASR word-boosted baselines. Regarding the number of retrieved exemplars k, the relationship between performance gains and quality varies across systems. Overall, we find in other reduced sweeps of L_c with $k \in \{2, 5\}$ that increasing k beyond two does not tend to lead to further improvements and can even plateau or slightly degrade performance by starting to retrieve irrelevant exemplars for the current active sentence. Based on these findings, we set $r_k = 2$ for all final systems in the context track.

WB	RAG-k	XCOMET ↑			YAAL (s) ↓		
		De	It	Zh	De	It	Zh
✗	✗	92.42	87.77	79.69	3.40	3.34	3.47
✓	✗	92.69	88.02	81.40	3.49	3.37	3.58
✓	1	92.99	88.38	81.69	3.45	3.32	3.52
✓	2	93.01	88.19	82.20	3.41	3.32	3.55
✓	3	92.94	88.70	81.67	3.52	3.40	3.47
✓	4	93.06	88.50	82.27	3.36	3.39	3.66
✓	5	93.28	88.96	82.94	3.52	3.37	3.66

Table 2: Sweep for MT RAG system across k for MCIF set with $L_c = 0.96$ and Qwen 3.5 9B. WB denotes ASR Word Boost; RAG-k indicates retrieved exemplars.

7 On SimulST Latency Scores

True latency, "macro" average latencies and oracle offsets To ensure that our system latencies would have a similar performance in real use cases and reflect user-perceived latency (UPL), we calculate latency scores of our complete pipeline by calculating alignments of our final configurations. For the MT component, we make use of latency metric based on the definition of Polák et al. (2026) by using forced alignment of the audio and source references and then aligning to the translation hypothesis[5]. We refer to this metric as TrueLatency.

During this evaluation process, we identified three problems with current latency metrics. First, SimulST latency metrics are currently reported as a *macro **average** of average token latencies* per sentence. We argue that in practice, this makes latency dependent on reference target segmentation and sentence length, distorting UPL. As an example, Figure 7 shows the source word length distribution of the MCIF dataset, where it can be seen that, by taking the "macro", latency on longer sentences

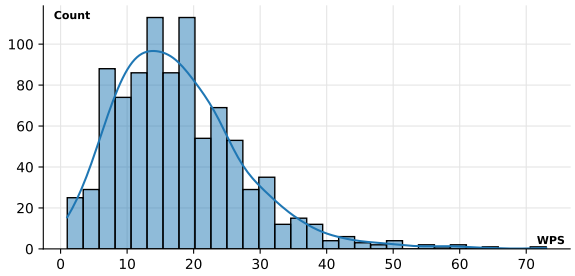

Figure 7: Words per sentence (WPS) histogram of MCIF En→De target.

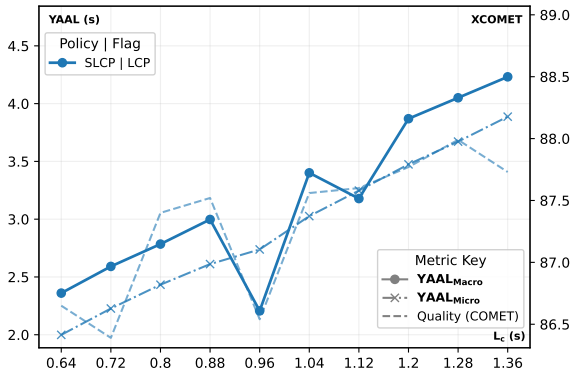

Figure 8: Example of early L_c sweep with Qwen-9B SCLP+LCP for MCIF En→It without whisper-like temperature fallback where $L_c = 0.96$ and $L_c = 1.12$ present early "end of stream" failure cases. The $\mathbf{YAAL_{Macro}}$ gets artificially decreased due the negative latencies, while $\mathbf{YAAL_{Micro}}$ is more robust to this type of noise.

may be under-represented. Our second identified problem lays on the way that AL based metrics calculate word delays with respect to the reference "wait-0" oracle. When adapting text based AL metrics to source speech, the delay for a word is the difference between the emission time and the oracle assigned $(t - 1/r)$, with r being the length ratio between target and source sequences. This can also be interpreted as taking the **start** emission time of equally distributed source words. This mismatches standard ASR latency calculations, which measure the difference between hypothesis and reference end delays [6]. Our final identified problem is that current macro-level AL metrics are highly sensitive to alignment errors. For instance, if a faulty system stops emitting prematurely, the sentence aligner of YAAL may force-align single words from the last sentences to missing sentences, generating extreme

[5]CTC based aligners from WhisperX (Bain et al., 2023) and SimAlign (Jalili Sabet et al., 2020)

[6]For example, see Caiman ASR and UFAL asr_latency script. It is worth noting that contrary to AL based metrics, ATD does take source end emission times on its formulation.

Latency	Context	XCOMET ↑	$\mathbf{YAAL}_{\mathbf{Macro}}$ ↓	$\mathbf{YAAL}_{\mathbf{Micro+EndOffset}}$ ↓			NE ↓
				Mean	P50	P99	
En-De							
LOW	✗	90.26	1.89	$1.51_{+0.02}$	$1.39_{+0.11}$	$4.19_{+1.36}$	0.17
	✓	92.05	1.89	$1.53_{+0.10}$	$1.42_{+0.14}$	$4.20_{+1.50}$	0.21
HIGH	✗	92.67	3.41	$2.99_{+0.15}$	$2.90_{+0.28}$	$6.30_{+0.85}$	0.00
	✓	93.70	3.41	$3.00_{+0.18}$	$2.89_{+0.31}$	$6.57_{+0.66}$	0.00
En-It							
LOW	✗	85.16	1.89	$1.54_{+0.01}$	$1.45_{+0.08}$	$4.02_{+1.37}$	0.15
	✓	87.03	1.90	$1.54_{+0.06}$	$1.44_{+0.10}$	$4.37_{+1.14}$	0.18
HIGH	✗	87.97	3.36	$2.96_{+0.07}$	$2.87_{+0.19}$	$6.23_{+0.96}$	0.00
	✓	89.36	3.42	$3.00_{+0.13}$	$2.91_{+0.20}$	$6.29_{+0.90}$	0.00
En-Zh							
LOW	✗	78.46	1.80	$1.61_{-0.20}$	$1.48_{-0.13}$	$4.78_{+1.54}$	0.38
	✓	82.12	1.82	$1.65_{-0.19}$	$1.49_{-0.09}$	$5.17_{+1.29}$	0.53
HIGH	✗	82.79	3.44	$3.25_{-0.22}$	$3.15_{-0.07}$	$6.77_{+1.28}$	0.00
	✓	84.56	3.55	$3.36_{-0.26}$	$3.23_{-0.11}$	$7.47_{+0.90}$	0.01
Cs-En							
LOW	✗	77.60	1.56	$1.07_{+0.44}$	$1.11_{+0.34}$	$5.34_{+3.70}$	0.18
HIGH	✗	82.77	2.79	$1.99_{+0.61}$	$2.55_{+0.35}$	$7.26_{+3.28}$	0.00

Table 3: Final evaluation results across all language pairs. Subindices of $\mathbf{YAAL}_{\mathbf{Micro+EndOffset}}$ submetrics indicate the corresponding $\Delta(\mathbf{TrueLatency}_{\mathbf{Micro}} - \mathbf{YAAL}_{\mathbf{Micro+EndOffset}})$.NE indicates Normalized Erasure.

negative delays that may distort the system's real latency. We observe that taking the macro average in this cases greatly skews the YAAL scores, while the micro average smooths noisy negative delays, yielding a more realistic latency score for the functional part of the inference. Figure 8 shows an example of this phenomenon of faulty inferences in Qwen3.5 9B. Configurations with $L_c = 0.96$ and $L_c = 1.12$ show how YAAL calculated at the macro level in these cases artificially reduces latency with respect to the expected YAAL score that correlates with the increase of L_c, while YAAL at the micro level properly captures the expected linearity and behavior of the model. In addition to all of this, this negative delay phenomenon can be easily overlooked, as common checks for empty sentence alignments will not report this cases.

8 Final Results

Following the previous section, for our final systems reported in Table 3, in addition to standard macro, start oracle emission YAAL, XCOMET and NE, we report the YAAL average at the micro level with oracle end offsets alongside the corresponding deltas with respect to our calculated TrueLatency. We also report median and p99 following the recommendations of (Iranzo-Sánchez et al., 2025a) to ensure the robustness of our systems and give a better picture of latency distribution beyond the mean. For our final systems, $\mathbf{YAAL}_{\mathbf{Macro}}$ latencies hover the ∼1.9 and ∼3.5 second mark for MCIF directions and 1.5 and 2.8 for Cs→En for the low and high latency regimes. In terms of quality, compared to models configurations of this year organizer baselines on MCIF with similar YAAL scores [7], we obtain substantial improvements, with $\Delta\mathbf{XCOMET}^{Low}_{De,It,Zh} = (+13.5, +16.7, +3.0)$ and $\Delta\mathbf{XCOMET}^{High}_{De,It,Zh} = (+7.6, +9.0, +2.3)$ for low and high latency respectively. Versus the improved context track baselines, we also maintain substantial gains of $\Delta\mathbf{XCOMET}^{Low+Ctx}_{De,It,Zh} = (+14.4, +17.7, +7.2)$ and $\Delta\mathbf{XCOMET}^{High+Ctx}_{De,It,Zh} = (+7.4, +9.7, +3.2)$. We can also observe that our final models' $\mathbf{YAAL}_{\mathbf{Micro+EndOffset}}$ are very similar to the obtained $\mathbf{TrueLatency}_{\mathbf{Micro}}$, alongside reasonable median and p99 values, which lead us to affirm that our final models are robust and their latency will probably reflect real observed UPL. We do note that for Cs →En, bigger Δ gaps appear compared to the MCIF language pairs.

[7] Baseline models with 0.64s and 1.28s chunk size.

9 Limitations

Several limitations of this work should be acknowledged. First, our exploration of relaxed LCP policies was limited due to time constraints. LACP was not evaluated as an MT emission policy, and the sensitivity analysis of SLCP parameters γ and σ were selected on previous small scale experiments. It is possible that per-language tuning of these hyperparameters could yield better latency–quality trade-offs for the remaining language directions, which we leave to future work. Second, also due to time constraints, policy exploration in ASR was entirely conducted in English ASR. Since we transferred the best English configuration to Czech ASR, our hope is that our results for the Czech - English translation pair could be further improved with a language-specific policy exploration. Third, our system is restricted to cascaded architectures. While this choice is empirically motivated by strong results on recent benchmarks, it may forgo potential gains from tighter integration of acoustic and linguistic information. We acknowledge that SpeechLLM-based approaches via a modality adapter (Verdini et al., 2025) represent a promising alternative, and our decision not to explore them here is primarily driven by the limited availability of in-domain training data for this track and the computational cost of bridging the modality gap in such architectures. Finally, the computational budget available to us constrained several design choices. Our experiments were conducted mostly consumer-grade GPUs with at most 24GB of memory, which prevented us from evaluating larger unquantized models and from running MBR decoding at a scale that would be competitive with greedy decoding in terms of real-time factor.

Acknowledgments

We would like to specially thank the IWSLT Simultaneous Speech track organizers, especially Katsuhito Sudoh and Victor Agostinelli, for providing us with last years logs of the SimulST track. We also acknowledge the usage of large language model based tools to assist our writing and proofreading process of this paper. The research leading to these results has received funding from EU4Health Programme 2021–2027 as part of Europe's Beating Cancer Plan under Grant Agreements nos. 101056995 and 101129375; and from the Government of Spain's grant PID2021-122443OB-I00 funded by MICIU/AEI/10.13039/501100011033 and by "ERDF/EU", grant PDC2022-133049-I00 funded by MICIU/AEI/10.13039/501100011033 and by the "European Union NextGenerationEU/PRTR", and grant PRE2022-103662 funded by MICIU/AEI/10.13039/501100011033 and by "ESF+". The authors gratefully acknowledge the financial support of Generalitat Valenciana under project IDIFEDER/2021/059.

References

Idris Abdulmumin, Victor Agostinelli, Tanel Alumäe, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Fethi Bougares, Roldano Cattoni, Mauro Cettolo, Lizhong Chen, William Chen, Raj Dabre, Yannick Estève, Marcello Federico, Mark Fishel, Marco Gaido, Dávid Javorský, Marek Kasztelnik, and 33 others. 2025. Findings of the IWSLT 2025 evaluation campaign. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 412–481, Vienna, Austria (in-person and online). Association for Computational Linguistics.

David Ifeoluwa Adelani, Victor Agostinelli, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Sebastien Bratières, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Lizhong Chen, Marcello Federico, Marco Gaido, Mahendra Gupta, HyoJung Han, Ali Hatami, David Javorský, Yejin Jeon, Marek Kasztelnik, Antoine Laurent, and 33 others. 2026. Speech translation and metrics in 2026: Findings of the iwslt campaign. In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, San Diego, California, US. Association for Computational Linguistics.

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. In-context examples selection for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8857–8873, Toronto, Canada. Association for Computational Linguistics.

Andrei Andrusenko, Vladimir Bataev, Lilit Grigoryan, Vitaly Lavrukhin, and Boris Ginsburg. 2025. Turbobias: Universal asr context-biasing powered by gpu-accelerated phrase-boosting tree. In *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–7.

Naveen Arivazhagan, Colin Cherry, Te I, Wolfgang Macherey, Pallavi Baljekar, and George F. Foster. 2020a. Re-translation strategies for long form, simultaneous, spoken language translation. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, pages 7919–7923.

Naveen Arivazhagan, Colin Cherry, Wolfgang Macherey, and George Foster. 2020b. Re-translation

versus streaming for simultaneous translation. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 220–227, Online. Association for Computational Linguistics.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. In *24th Annual Conference of the International Speech Communication Association, Interspeech 2023, Dublin, Ireland, August 20-24, 2023*, pages 4489–4493.

Shanbo Cheng, Yu Bao, Zhichao Huang, Yu Lu, Ningxin Peng, Lu Xu, Runsheng Yu, Rong Cao, Yujiao Du, Ting Han, Yuxiang Hu, Zeyang Li, Sitong Liu, Shengtao Ma, Shiguang Pan, Jiongchen Xiao, Nuo Xu, Meng Yang, Rong Ye, and 9 others. 2025. Seed liveinterpret 2.0: End-to-end simultaneous speech-to-speech translation with your voice. *CoRR*, abs/2507.17527.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Y. Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, and 19 others. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.

Hiroyuki Deguchi, Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2024. mbrs: A library for minimum Bayes risk decoding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 351–362, Miami, Florida, USA. Association for Computational Linguistics.

M. A. del Agua, A. Giménez, N. Serrano, J. Andrés-Ferrer, J. Civera, A. Sanchis, and A. Juan. 2014. The translectures-upv toolkit. In *Proc. of VIII Jornadas en Tecnología del Habla and IV Iberian SLTech Workshop (IberSpeech 2014)*, Las Palmas de Gran Canaria (Spain).

John DeNero, David Chiang, and Kevin Knight. 2009. Fast consensus decoding over translation forests. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 567–575, Suntec, Singapore. Association for Computational Linguistics.

Bryan Eikema and Wilker Aziz. 2022. Sampling-based approximations to minimum Bayes risk decoding for neural machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10978–10993, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mara Finkelstein, Isaac Caswell, Tobias Domhan, Jan-Thorsten Peter, Juraj Juraska, Parker Riley, Daniel Deutsch, Geza Kovacs, Cole Dilanni, Colin Cherry, Eleftheria Briakou, Elizabeth Nielsen, Jiaming Luo, Kat Black, Ryan Mullins, Sweta Agrawal, Wenda Xu, Erin Kats, Stephane Jaskiewicz, and 2 others. 2026. Translategemma technical report. *CoRR*, abs/2601.09012.

Markus Freitag, Behrooz Ghorbani, and Patrick Fernandes. 2023. Epsilon sampling rocks: Investigating sampling strategies for minimum Bayes risk decoding for machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9198–9209, Singapore. Association for Computational Linguistics.

Vaibhava Goel and William J. Byrne. 2000. Minimum bayes-risk automatic speech recognition. *Comput. Speech Lang.*, 14(2):115–135.

Lilit Grigoryan, Vladimir Bataev, Andrei Andrusenko, Hainan Xu, Vitaly Lavrukhin, and Boris Ginsburg. 2025. Pushing the limits of beam search decoding for transducer-based ASR models. In *26th Annual Conference of the International Speech Communication Association, Interspeech 2025, Rotterdam, The Netherlands, 17-21 August 2025*.

Maarten Grootendorst. 2020. Keybert: Minimal keyword extraction with bert.

Nuno Miguel Guerreiro, Ricardo Rei, Daan van Stigt, Luísa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet : Transparent machine translation evaluation through fine-grained error detection. *Trans. Assoc. Comput. Linguistics*, 12:979–995.

Shoutao Guo, Shaolei Zhang, Zhengrui Ma, Min Zhang, and Yang Feng. 2025. Agent-simt: Agent-assisted simultaneous translation with large language models. *IEEE Transactions on Audio, Speech and Language Processing*, 33:2074–2083.

John Hewitt, Christopher Manning, and Percy Liang. 2022. Truncation sampling as language model desmoothing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3414–3427, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Javier Iranzo-Sánchez, Jorge Iranzo-Sánchez, Adrià Giménez, Jorge Civera, and Alfons Juan. 2024. Segmentation-free streaming machine translation. *Transactions of the Association for Computational Linguistics*, 12:1104–1121.

Jorge Iranzo-Sánchez, Javier Iranzo-Sánchez, Adrià Giménez, and Jorge Civera. 2025a. Going beyond your expectations in latency metrics for simultaneous

speech translation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18205–18228, Vienna, Austria. Association for Computational Linguistics.

Jorge Iranzo-Sánchez, Javier Iranzo-Sanchez, Adrià Giménez Pastor, Jorge Civera Saiz, and Alfons Juan. 2025b. MLLP-VRAIN UPV system for the IWSLT 2025 simultaneous speech translation translation task. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 340–346, Vienna, Austria (in-person and online). Association for Computational Linguistics.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.

Yuu Jinnai. 2025. Re-evaluating minimum bayes risk decoding for automatic speech recognition. *CoRR*, abs/2510.19471.

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. 2024. 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.

Roman Koshkin, Katsuhito Sudoh, and Satoshi Nakamura. 2024a. LLMs are zero-shot context-aware simultaneous translators. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1207, Miami, Florida, USA. Association for Computational Linguistics.

Roman Koshkin, Katsuhito Sudoh, and Satoshi Nakamura. 2024b. TransLLaMa: LLM-based simultaneous translation system. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 461–476, Miami, Florida, USA. Association for Computational Linguistics.

Shankar Kumar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 169–176, Boston, Massachusetts, USA. Association for Computational Linguistics.

Daniil Larionov, Mikhail Seleznyov, Vasiliy Viskov, Alexander Panchenko, and Steffen Eger. 2024. xCOMET-lite: Bridging the gap between efficiency and quality in learned MT evaluation metrics. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21934–21949, Miami, Florida, USA. Association for Computational Linguistics.

Xintong Li, Guanlin Li, Lemao Liu, Max Meng, and Shuming Shi. 2019. On the word alignment from neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1293–1303, Florence, Italy. Association for Computational Linguistics.

Zhaolin Li, Yining Liu, Danni Liu, Tuan Nam Nguyen, Enes Yavuz Ugan, Tu Anh Dinh, Carlos Mullov, Alexander Waibel, and Jan Niehues. 2025. KIT's low-resource speech translation systems for IWSLT2025: System enhancement with synthetic data and model regularization. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 212–221, Vienna, Austria (in-person and online). Association for Computational Linguistics.

Danni Liu, Gerasimos Spanakis, and Jan Niehues. 2020a. Low-latency sequence-to-sequence speech recognition and translation by partial hypothesis selection. In *21st Annual Conference of the International Speech Communication Association, Interspeech 2020, Virtual Event, Shanghai, China, October 25-29, 2020*, pages 3620–3624.

Danni Liu, Gerasimos Spanakis, and Jan Niehues. 2020b. Low-Latency Sequence-to-Sequence Speech Recognition and Translation by Partial Hypothesis Selection. In *Interspeech 2020*, pages 3620–3624.

Jiaxuan Luo, Siqi Ouyang, and Lei Li. 2026. Rasst: Fast cross-modal retrieval-augmented simultaneous speech translation. *Preprint*, arXiv:2601.22777.

Xing Han Lù. 2024. Bm25s: Orders of magnitude faster lexical search via eager sparse scoring. *Preprint*, arXiv:2407.03618.

Mingbo Ma, Liang Huang, Hao Xiong, Renjie Zheng, Kaibo Liu, Baigong Zheng, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, Hua Wu, and Haifeng Wang. 2019. STACL: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework. In *Proceedings of the 57th Annual Meeting of the Association for*

Computational Linguistics, pages 3025–3036, Florence, Italy. 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.

Zhuoyuan Mao and Yen Yu. 2024. Tuning LLMs with contrastive alignment instructions for machine translation in unseen, low-resource languages. In *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 1–25, Bangkok, Thailand. Association for Computational Linguistics.

Gerard Mas-Mollà, Albert Sanchis, and Alfons Juan. Improving streaming ASR with foundation models using emission policies. Submitted to Interspeech 2026.

Kenton Murray and David Chiang. 2018. Correcting length bias in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 212–223, Brussels, Belgium. 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.

Sara Papi, Maike Züfle, Marco Gaido, Beatrice Savoldi, Danni Liu, Ioannis Douros, Luisa Bentivogli, and Jan Niehues. 2026. MCIF: Multimodal crosslingual instruction-following benchmark from scientific talks. In *The Fourteenth International Conference on Learning Representations*.

Peter Polák, Danni Liu, Ngoc-Quan Pham, Jan Niehues, Alexander Waibel, and Ondřej Bojar. 2023. Towards efficient simultaneous speech translation: CUNI-KIT system for simultaneous track at IWSLT 2023. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 389–396, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Peter Polák, Ngoc-Quan Pham, Tuan Nam Nguyen, Danni Liu, Carlos Mullov, Jan Niehues, Ondřej Bojar, and Alexander Waibel. 2022. CUNI-KIT system for simultaneous speech translation task at IWSLT 2022. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 277–285, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Peter Polák, Sara Papi, Luisa Bentivogli, and Ondřej Bojar. 2026. Better late than never: Meta-evaluation of latency metrics for simultaneous speech-to-text translation. *Preprint*, arXiv:2509.17349.

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.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, pages 28492–28518.

Matthew Raffel, Victor Agostinelli, and Lizhong Chen. 2024. Simultaneous masking, not prompting optimization: A paradigm shift in fine-tuning LLMs for simultaneous translation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18302–18314, Miami, Florida, USA. Association for Computational Linguistics.

Miguel Moura Ramos, Duarte M. Alves, Hippolyte Gisserot-Boukhlef, João Alves, Pedro Henrique Martins, Patrick Fernandes, José Pombal, Nuno M. Guerreiro, Ricardo Rei, Nicolas Boizard, Amin Farajian, Mateusz Klimaszewski, José G. C. de Souza, Barry Haddow, François Yvon, Pierre Colombo, Alexandra Birch, and André F. T. Martins. 2026. Eurollm-22b: Technical report. *Preprint*, arXiv:2602.05879.

John W Ratcliff, David E Metzener, and 1 others. 1988. Pattern matching: The gestalt approach. *Dr. Dobb's Journal*, 13(7):46.

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

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.

Vaibhav Srivastav, Steven Zheng, Eric Bezzam, Eustache Le Bihan, Nithin Koluguri, Piotr Żelasko, Somshubra Majumdar, Adel Moumen, and Sanchit Gandhi. 2026. Open asr leaderboard: Towards reproducible and transparent multilingual and long-form speech recognition evaluation. *Preprint*, arXiv:2510.06961.

Jannis Vamvas and Rico Sennrich. 2024. Linear-time minimum Bayes risk decoding with reference aggregation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 790–801, Bangkok, Thailand. Association for Computational Linguistics.

Francesco Verdini, Pierfrancesco Melucci, Stefano Perna, Francesco Cariaggi, Marco Gaido, Sara Papi, Szymon Mazurek, Marek Kasztelnik, Luisa Bentivogli, Sebastien Bratières, Paolo Merialdo, and Simone Scardapane. 2025. How to Connect Speech Foundation Models and Large Language Models? What Matters and What Does Not. In *Interspeech 2025*, pages 1813–1817.

Wenxuan Wang, Yingxin Zhang, Yifan Jin, Binbin Du, and Yuke Li. 2025. NYA's offline speech translation system for IWSLT 2025. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 206–211, Vienna, Austria (in-person and online). Association for Computational Linguistics.

Yilin Yang, Liang Huang, and Mingbo Ma. 2018. Breaking the beam search curse: A study of (re-)scoring methods and stopping criteria for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3054–3059, Brussels, Belgium. Association for Computational Linguistics.

Mao Zheng, Zheng Li, Tao Chen, Mingyang Song, and Di Wang. 2025. Hy-mt1.5 technical report. *Preprint*, arXiv:2512.24092.

Vilém Zouhar, Maike Züfle, Beni Egressy, Julius Cheng, Mrinmaya Sachan, and Jan Niehues. 2026. Early-exit and instant confidence translation quality estimation. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 55–76, Rabat, Morocco. Association for Computational Linguistics.

A Minimum Bayes Risk Decoding Study

While Minimum Bayes Risk (MBR) decoding was popular during both the statistical ASR era (Goel and Byrne, 2000) and early machine translation research (Kumar and Byrne, 2004), it has recently seen renewed interest in offline MT (Eikema and Aziz, 2022).

This resurgence is largely driven by the emergence of strong neural evaluation metrics, which can outperform standard beam search while avoiding known issues such as the beam search curse (Murray and Chiang, 2018; Yang et al., 2018). To the best of our knowledge, this constitutes the first study exploring the use of MBR in the SimulST setting, as prior work has largely restricted MBR to offline ASR and MT scenarios (Jinnai, 2025; Wang et al., 2025; Li et al., 2025).

We leverage the MBR implementations provided by the `mbrs` library (Deguchi et al., 2024) and experiment with multiple evaluation metrics. Due to computational constraints, we primarily focus on XCOMET-lite (Larionov et al., 2024) and chrF [8]. We also evaluated chrF++ (Popović, 2017), default BLEU via sacreBLEU (Post, 2018), and PartialCOMET (Zouhar et al., 2026) in earlier experiments, but observed similar or worse performance at higher computational cost. For hypothesis generation, we use epsilon sampling (Hewitt et al., 2022; Freitag et al., 2023) with $\epsilon = 0.02$ and $\tau = 1.0$. For all applicable metrics, we make use of Reference Aggregation to speed up MBR (DeNero et al., 2009; Vamvas and Sennrich, 2024)

Table 4 reports results for configurations feasible on a single NVIDIA RTX 4090. For the Qwen models, chrF with $n = 32$ performs comparably to greedy decoding, but at a significantly higher computational cost. This ultimately led us to discard MBR for our final submission.

The table also includes results for the IWSLT 2025 UPV system, where we replace the RALCP policy and force the system to always commit outputs. This setup highlights an interesting property of MBR when adapting offline methods to SimulST: it mitigates hallucinations and reduces the tendency of mode-seeking decoding algorithms to emit empty outputs. While the original submission utilized RALCP to address these issues, removing the emission policy causes both greedy and beam search decoding to produce divergent target content which end up in inference failures. In contrast, MBR naturally prevents this behavior, enabling stable decoding without requiring additional control policies.

[8] https://github.com/jvamvas/fastChrF

We also explored applying offline MBR to the context track. However, generating large numbers of hypotheses (k) resulted in significantly slower decoding, with real-time factors exceeding 1 relative to dataset duration. As a result, we discarded this approach for the context track as well.

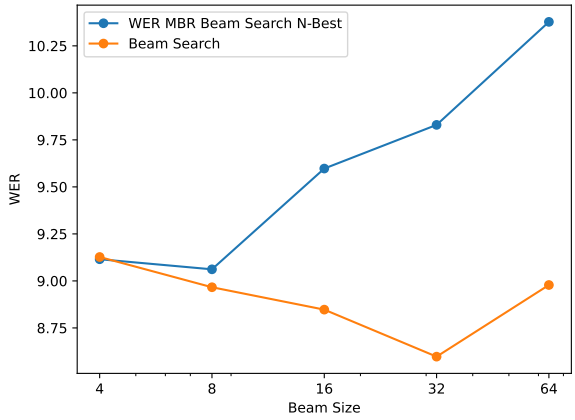

Figure 9: WER of beam search vs. MBR re-ranking of n-best hypotheses for Parakeet with $L_c = 0.96$ on MCIF across different values of n

Finally, we evaluated a standard MBR re-ranking approach for ASR over n-best hypotheses generated via beam search. However, this method consistently yielded negative results, as shown in Figure 9. Overall, we did not adopt MBR in neither the ASR nor the MT component in our final system due to its computational cost and limited benefits, although it remains an interesting direction for future study.

Model	Policy	MBR Metric	Samples	YAAL (s) ↓		XCOMET ↑	chrF ↑	BLEU ↑
				CU	CA			
Qwen 3.5 (4B)	Hold-3	Greedy	1	2.99	3.33	**89.48**	58.06	24.75
		chrF	16	3.05	3.96	87.43	57.15	23.83
			32	3.10	4.52	87.15	58.01	24.66
		XCOMET-lite	8	3.06	3.96	85.33	51.40	17.51
Qwen 3.5 (9B)	Hold-3	Greedy	1	2.94	3.39	**89.57**	58.55	25.58
		chrF	16	2.87	4.22	87.52	57.98	23.18
			32	2.89	4.87	89.55	59.22	24.66
		XCOMET-lite	8	2.82	4.04	88.51	53.14	17.38
IWSLT 25 UPV	Write All	Greedy		✗ Unstable, results in an inference failure				
		Beam Search		✗ Unstable, results in an inference failure				
		chrF	64	1.96	2.43	76.46	56.19	20.98
		chrF++	64	1.99	5.34	76.74	54.90	20.36
		BLEU	64	2.23	4.91	76.70	54.10	22.85
		PartialComet	64	2.22	2.71	76.29	45.85	10.28

Table 4: Comparison of Greedy vs. MBR decoding for Qwen 3.5 variants hold-3 variants and IWSLT 25 UPV without RALCP for MCIF En→De.