Going Beyond Your Expectations in Latency Metrics for Simultaneous Speech Translation

Anonymous ACL submission

Abstract

Current evaluation practices in Simultaneous Speech Translation (SimulST) systems typically involve segmenting the input audio and corresponding translations, calculating quality and latency metrics for each segment, and averaging the results. Although this approach may provide a reliable estimation of translation quality, it can lead to misleading values of latency metrics due to an inherent assumption that average latency values are good enough estimators of SimulST systems' response time. However, our detailed analysis of latency evaluations for state-of-the-art SimulST systems demonstrates that latency distributions are often skewed and subject to extreme variations. As a result, the mean in latency metrics fails to capture these anomalies, potentially masking the lack of robustness in some systems and metrics. In this paper, a thorough analysis of the results of systems submitted to recent editions of the IWSLT simultaneous track is provided to support our hypothesis and alternative ways to report latency metrics are proposed in order to provide a better understanding of SimulST systems' latency.

1 Introduction

In recent years, there has been a growing demand for real-world applications that use Simultaneous Speech Translation (SimulST) systems to provide real-time translation across languages. Current use cases include live broadcasts of news, lectures, and debates, where the continuous audio stream mainly consists of spoken speech. These applications require systems that do not only deliver high-quality translations consistently, but also maintain low latency to ensure effective communication and keep the audience engaged with the audiovisual content.

Current evaluation of latency in SimulST relies on automatic or reference segmentation of datasets (Di Gangi et al., 2019; Wang et al., 2020) to split the input audio and its translations, computing metrics for each segment, and averaging the results. However, this latency estimation for SimulST systems has significant limitations (Iranzo-Sánchez et al., 2021; Papi et al., 2024), and reported latency figures may differ from the actual behavior of systems in a real-world scenario.

Beyond the segmentation-related issues identified in these previous work, we argue that a major cause of the observed discrepancies may be due to the exclusive reliance on the mean when reporting results. While mean latencies allow to simplify system comparison to speed up their development, we hypothesize that by relying solely on the mean, we may be overlooking spurious or faulty system behaviors, as well as anomalies in the current latency metrics. While the presence of outliers is relatively common when evaluating machine learning systems with any metric, their significance and impact in latency evaluation of SimulST systems are currently being greatly underestimated. Thus, mean values of latency metrics may result in misleading conclusions when comparing SimulST systems.

In this paper we demonstrate that latency metrics, as currently reported by their average values, are not a sufficiently accurate characterization of SimulST systems' response time in the presence of latency distributions that do not follow a normal distribution. To this purpose, latency metrics of systems submitted to recent editions of the IWSLT SimulST track were thoroughly analysed. Our findings reveal that average latency metrics can mask undesirable systems' behavior, potentially resulting in misleading conclusions. This highlights the need for more robust evaluation methods for latency in SimulST systems. Our contributions are summarized as follows:

- We performed a detailed analysis of recent SimulST systems submitted to IWSLT in terms of latency metrics.

- This analysis demonstrates the limitations of current latency metrics as reported by their mean in order to detect undesirable SimulST systems' response time, preventing a fair comparison across systems.
- We report a series of latency phenomena that must be considered and gauged when evaluating SimulST systems to guarantee their consistent response time.
- We propose the usage of a series of descriptive statistics that provide a more robust overview of SimulST systems' response time and allows for a more holistic comparison.

2 Related Work

The evaluation of SimulST systems is performed in two dimensions: translation quality and latency. While translation quality is typically evaluated by using conventional translation metrics such as BLEU (Papineni et al., 2002; Post, 2018) and COMET (Rei et al., 2020, 2022; Guerreiro et al., 2024), multiple metrics have been developed for measuring the latency of SimulST systems. Earlier proposed metrics such as Average Proportion (AP) (Cho and Esipova, 2016) and Consecutive Wait Length (CW) (Gu et al., 2017) have been mostly superseded in usage by Average Latency (AL) (Ma et al., 2019) and proposed variants such as Differentiable Average Lagging (DAL) (Cherry and Foster, 2019) and Length-Adaptive Average Lagging (LAAL) (Papi et al., 2022) which try to remedy several limitations in the original AL metric definition. Another more recent metric which has received a fair amount of adoption is Average Token Delay (ATD) (Kano et al., 2023a,b), which tries to fix several limitations underlining AL based metrics. Additionally, Wein et al. (2024) and Makinae et al. (2024) have proposed metrics tailored towards the evaluation of the quality and latency of translations closer to human interpretation.

As characterized in Iranzo-Sánchez et al. (2021), current latency measures for SimulST can be defined as a normalisation of a latency cost (in terms of words or milliseconds) required to generate a translation $\hat{\boldsymbol{y}}$ provided a source sentence $\boldsymbol{x}$ and its corresponding reference translation $\boldsymbol{y}$

$$L(\boldsymbol{x}, \boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{1}{Z(\boldsymbol{x}, \hat{\boldsymbol{y}})} \sum_i C_i(\boldsymbol{x}, \boldsymbol{y}, \hat{\boldsymbol{y}}) \quad (1)$$

with Z being a normalisation function, i an index over the target positions and C_i a cost function for each target position i. Depending on the latency metric, C_i is defined as

$$C_i(\boldsymbol{x}, \boldsymbol{y}, \hat{\boldsymbol{y}}) = \begin{cases} g(i) - \frac{i-1}{\hat{\gamma}} & \text{AL} \\ g(i) - \frac{i-1}{\gamma} & \text{LAAL} \\ g'(i) - \frac{i-1}{\hat{\gamma}} & \text{DAL} \\ T(\hat{y}_i) - T(x_{a(i)}) & \text{ATD} \end{cases} \quad (2)$$

with

$$g'(i) = \max \begin{cases} g(i) \\ g'(i-1) + \frac{1}{\hat{\gamma}} \end{cases} \quad (3)$$

and

$$a(i) = \min \begin{cases} g(i) \\ i - d(i) \end{cases} \quad (4)$$

$$d(i) = (i-1) - a(i-1). \quad (5)$$

where $g(i)$ is the number of tokens or milliseconds read when a token is written at position i, and $\hat{\gamma} = \frac{|\hat{\boldsymbol{y}}|}{|\boldsymbol{x}|}$ and $\gamma = \frac{\max(|\hat{\boldsymbol{y}}|, |\boldsymbol{y}|)}{|\boldsymbol{x}|}$ are target-to-source length ratios. On the other hand, $T(\cdot)$ represents the ending time of each input or output token, $a(i)$ the index of the input token corresponding to $\hat{y}_i$ and $d(i)$ measures the difference between previous input-translation prefix pairs $(\boldsymbol{x}_{\leq g(i)}, \hat{\boldsymbol{y}}_{\leq i})$. The normalisation function Z depends on the metric according to

$$Z(\boldsymbol{x}, \hat{\boldsymbol{y}}) = \begin{cases} \underset{i:g(i)=|\boldsymbol{x}|}{\text{argmin}} \; i & \text{AL, LAAL} \\ |\hat{\boldsymbol{y}}| & \text{DAL, ATD} \end{cases} \quad (6)$$

Papi et al. (2024) analyzed the behaviour of current SimulST systems surveying the current literature to define standardized terminology and taxonomy across different SimulST papers, while also identifying overlooked challenges in current SimulST systems and recommendations for future work in the field. Related to this, Xu et al. (2024) identified how current computationally-aware metrics are incorrectly calculated in current standard SimulST evaluation toolkits. Finally, Macháček et al. (2023a) showed how MT quality metrics have good correlation with Human Ratings for IWSLT 2022 and Sperber et al. (2024) analyzed the IWSLT 2023 evaluation campaign across different aspects.

3 Limitations of mean latencies

In this section, the limitations of mean latencies are illustrated with a simplified example computed on real data from IWSLT competitions to show how

	Mean	P90	P95	P99	Max
System A	2.4	4.2	5.3	7.5	31.3
System B	2.4	3.5	3.9	4.8	10.4

Table 1: Comparison of latency mean, percentiles 90%, 95% and 99%, and maximum value for two SimulST systems.

high latencies may be underestimated in SimulST. Table 1 shows a latency comparison in seconds between two SimulST systems falling in a low latency band and with similar translation quality. Latency figures for the mean, percentile scores for 90%, 95% and 99%, and the maximum value are reported. As observed, the mean latency of both, system A and B, is 2.4 seconds. However, system A provides 10% of their translations with a latency between 4.2 and 31.3 seconds, while system B does in the range from 3.5 to 10.4 seconds.

If we were to pick between the two systems considering conventional latency metrics based on the mean, system A would be considered as good as system B. However, by looking into latency distribution across samples, the values of system A on approximately 10% of the samples significantly differs from the mean more than system B. By deploying system A instead of B in a real streaming ST scenario, the probability that latency spikes appear is high enough to lead to an accumulation of delays that causes desynchronization between the audio stream and the translation text being generated. This behaviour is highly undesirable for the end-user experience, and system A would not be an acceptable choice in a real streaming scenario, while system B with a more consistent latency would have been selected on the basis of its percentile scores.

4 IWLST as a Case Study

To investigate the limitations of the conventional latency metrics illustrated in the simplified example provided above, the latency of SimulST systems on IWSLT evaluation campaigns are analyzed and compared. In this study, the evaluation logs from IWSLT 2022 and 2024 Simultaneous Translation Speech-to-Text tracks (Anastasopoulos et al., 2022; Ahmad et al., 2024) were processed for all available team submissions provided in standard JSON files[1]

[1] https://dl.fbaipublicfiles.com/simultaneous_translation/iwslt2022_simul-s2t_logs.tgz

Year	#L	Task	Tgt	Avg Len	#S	#T
2022	3	Must-C	de	5.79	2580	5
			ja	5.12	2841	3
			zh	5.12	2841	3
		IWSLT	de	6.25	2059	5
			ja	5.38	1768	3
			zh	5.42	2136	3
2024	1	IWSLT	ja	5.92	1570	3

Table 2: Basic information of the evaluation logs from IWSLT 2022 and 2024 SimulST tracks. The number of latency bands, samples and teams are represented by #L, #S and #T, respectively. In addition, tasks, target languages (Tgt), average length in seconds are provided.

of the SimulEval toolkit (Ma et al., 2020).

Table 2 shows a general overview of the evaluation logs involved in the study. The IWSLT 2022 SimulST task featured English (En) as the source language, with three target directions evaluated: German (De), Japanese (Ja), and Mandarin Chinese (Zh). Five teams entered the German track, while three teams do it for both, Japanese and Chinese tracks. For the 2022 edition, three latency bands were defined and systems were classified into low, medium and high latency bands given by the AL metric. For IWSLT 2024, we were only able to get access to En-Ja results where three teams participated under a single latency band. In IWSLT 2024, team names were anonymized as requested by the IWSLT organizers.

Results for IWSLT 2022 shared tasks were available for the MuST-C 2.0 tst-COMMON partition (Di Gangi et al., 2019) and various segmentations of the official test set, for which the reference segmentation of the datasets was selected. For IWSLT 2024, results were only available on the official test sets. We focused on the Speech-to-Text track logs and left out the available Text-to-Text logs from our study.

For the IWSLT 2022, the delays available in the evaluations logs were used to recalculate AL and DAL scores using the latest version of the SimulEval toolkit[2]. In addition, LAAL and ATD scores, which were still not available at the time of the shared task, were calculated. For each team the

[2] SimulEval 1.1.4: commit 536de82. We found slight discrepancies between the metric scores values provided in the original logs and those obtained with this SimulEval version.

corresponding submitted system was selected. For those systems that did not comply with the latency constraints, we treat them as if they were in the nearest latency band. Finally, the CUNI-KIT participation in the En-De track was left out due to tokenization issues, while the Xiaomi participation in the En-Zh track was not available.

5 IWSLT Analysis

First of all, the evaluation logs were checked to detect the possible errors in the dataset, such as mismatched source-target pairs and misaligned segmentations. These errors would jeopardize the analysis of latency solely explained by the performance of SimulST systems. As a result of this initial error analysis, we decided to focus on the official IWSLT 2022 and 2024 test sets, as these proved largely error-free in contrast to the IWSLT 2022 MuST-C sets. Results for the MuST-C partition can be found in Appendix A.

5.1 Violin plots

Figure 1 shows from top to bottom latency sample distributions as violin plots for AL, DAL, LAAL and ATD, respectively. For each latency metric the three teams participating in the official IWSLT 2024 English-Japanese task are displayed. Each violin plot also represents the mean (orange dot), the median (white bar) when not overlapped with the mean, and the range from the first to the third quartile (horizontal bar). When comparing these four latency metrics, it can be observed that AL, DAL and LAAL exhibit similar shape distributions, while ATD distributions are clearly different from the rest. However, LAAL and DAL distributions in contrast to the AL distribution, stay in the positive range avoiding negative delays. As expected, latency distributions for all the three teams possess right long tails for high latencies that translates into a certain degree of right-skewness. However, right-skewness in Team 1 and Team 2 is aggravated compared to Team 3, observing mean values clearly falling on the right-hand side of median values. Based on these observations, we decided to focus our analysis on the latency metric LAAL, as this does not significantly impact the overall conclusions. Figures for the rest of latency metrics are available in Appendix B.

Figure 2 shows LAAL distributions for the IWLST 2022 team participations for English into German (top), Mandarin Chinese (middle) and

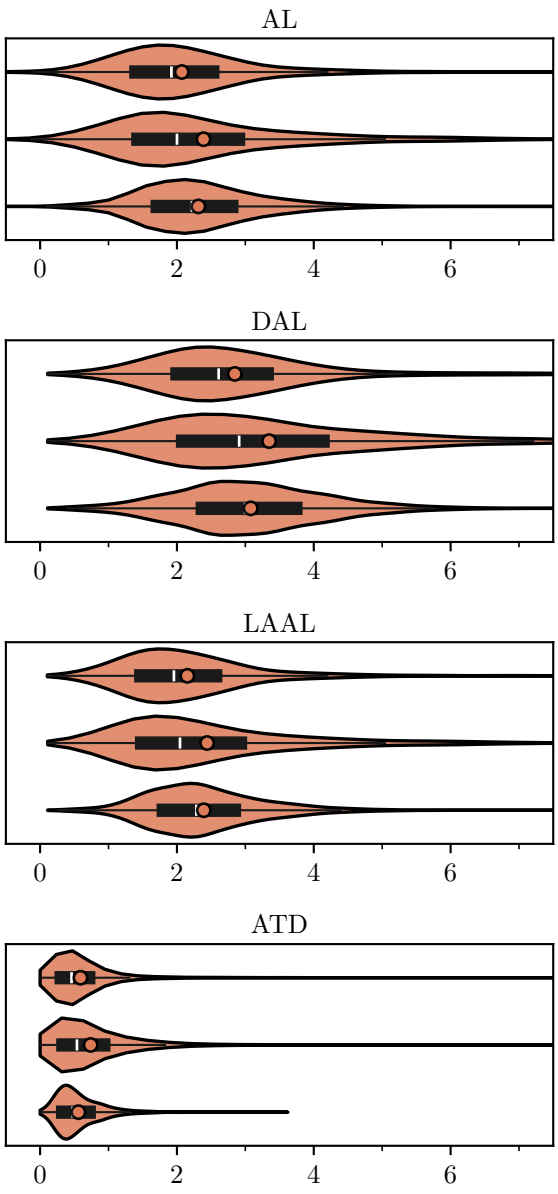
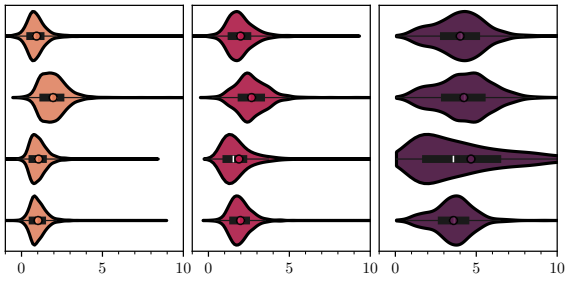
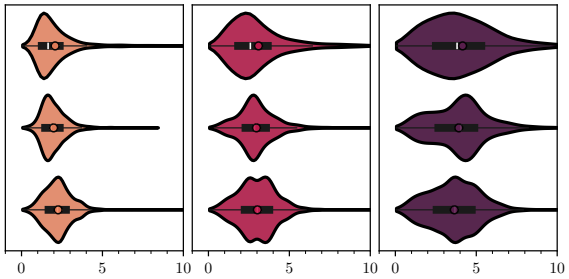


Figure 1: AL, DAL, LAAL and ATD latency distributions for the IWSLT 2024 En-Ja task represented from top to bottom as violin plots for the three teams. Long tails extending beyond a 8-second delay were cropped for clearer visualisation.

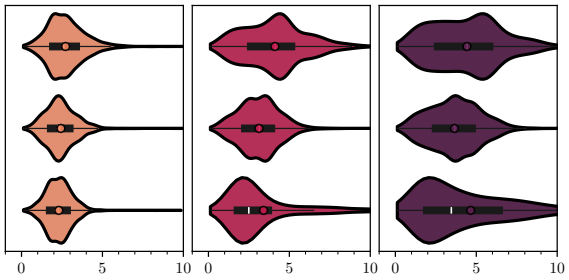
Japanese (bottom), across low (left), medium (center) and high (right) latency bands. Compared to the IWSLT 2024 latency distributions in Figure 1, differences between systems are more pronounced across all participations in IWSLT 2022, with systems seemingly following more unique distribution shapes. However, most systems exhibit similar mean values with slight latency differences of a few tenths of a second that, as shown, concealed widely different latency patterns. In addition, drastic changes in distribution shapes can be observed



(a) Top to bottom: FBK, HW-TSC, NAIST, UPV.



(b) Top to bottom: AISP_SJTU, CUNI-KIT, HW-TSC.



(c) Top to bottom: CUNI-KIT, HW-TSC, NAIST.

Figure 2: LAAL distributions for the IWSLT2022 team participations for English into German (top), Mandarin Chinese (middle) and Japanese (bottom), across low (left), medium (center) and high (right) latency bands.

in all languages across teams and latencies bands. For example, latency distributions in the NAIST En-De 2022 systems for low and medium latency bands significantly differ from that for the high latency band. In general, long right tails representing high latencies are observed for all models in a similar way to the IWSLT 2024 systems. However, these right tails are specially long in IWSLT 2022 systems for the high latency band when compared to those in the low and medium latency bands.

5.2 Normal probability plots

From the latency distributions in Figures 1 and 2, it can be observed that the shapes and right tails of these distributions significantly deviate from those expected in a normal distribution. Therefore, the mean may not properly capture the expected latency of these systems whose latency distributions move away from normality (Sainani, 2012). Thus, we assessed the degree of normality of the latency distributions as a proxy of how reliable the mean is as a estimation of the system's response time.

Shapiro–Wilk tests for normality (Shapiro and Wilk, 1965) were performed for all systems and latency metrics. In all tests the null hypothesis with p-values below $\alpha = 0.01$ was rejected. To graphically represent how the latency distributions deviate from normal distributions, normal probability plots (Dodge, 2008) were generated from the evaluation logs. Normal probability plots are a variant of Q-Q plots (Wilk and Gnanadesikan, 1968) in which observed values of our data sample are displayed with respect to the quantiles obtained from a normal distribution, typically $\mathcal{N}(0, 1)$. More precisely, in our case sample-level latencies (y-axis) are sorted from lowest to highest and distributed along the percentiles (x-axis) of a $\mathcal{N}(0, 1)$. In this way, the sample-level latency at the 50th percentile is the median value leaving 50% of the data samples below, that is, data samples with a lower latency. Intuitively, if our data samples comes from a normal distribution, the resulting plotted points would closely follow a straight line. Data samples featuring a high curvature with distribution edges diverging from a straight line denote latency distributions away from normality.

Figure 3 shows the normal probability plot for the English-Japanese IWLST 2024 systems. As explained above, sample-level latencies are represented in the y-axis, along, in this case, percentiles of a $\mathcal{N}(0, 1)$ are displayed in the x-axis. As observed, Team 1 and Team 2 latency distributions turn away from normality more clearly than Team 3 whose latency distribution approximately follows a straight line. On the one hand, as already observed in Figure 1, percentile values above 80%-90% for Team 1 and Team 2 suffer from considerable longer right tails than Team 3. On the other hand, normal probability plots also allow to easily compare systems on the left side of the tail with Team 3 low latencies being higher than those of the other teams.

To further illustrate the capability of normal probability plots as a visual aid to compare systems and capture latency distributions, Figure 4 shows sample-level latencies for AISP_SJTU (top) and HW-TSC (bottom) IWSLT 2022 En-Zh systems. In this figure reference lines were plotted representing the ideal expected percentiles that would be obtained from a normal distribution with the observed mean and standard deviation of each sys-

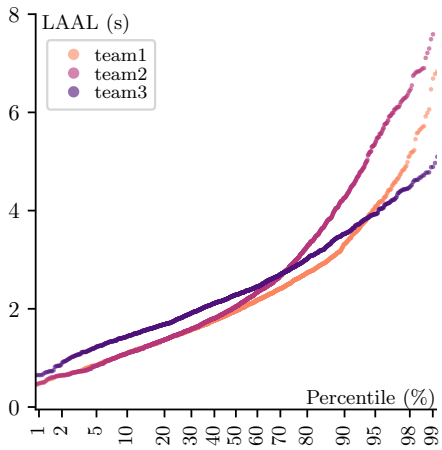


Figure 3: LAAL normal probability plot for IWSLT 2024 English-Japanese teams representing sample-level latencies (y-axis) w.r.t. percentiles (x-axis) of a $\mathcal{N}(0, 1)$.

tem. While HW-TSC's latency distribution in the low (orange), medium (magenta) and high (blue) latency bands seem to follow considerable normal distributions, AISP_SJTU's latency distribution in the same bands tend to show a steep slope towards the right tails, denoting a higher frequency of samples with increased latencies. Normal probability plots for other teams involved in the IWSLT 2022 task are available in Appendix C.

Having shown the capability of normal probability plots to compare latency distributions across systems, Table 3 shows a complementary view of latency distributions to the normal probability plot in Figure 3. More precisely, Table 3 shows from left to right for each team in the English-Japanese IWSLT 2024 task: BLEU score, and LAAL mean (M), median (mdn), percentiles 90%, 95% and 99%, and maximum value. This table is an extension of Table 1 provided in the simplified example of Section 3, corresponding Team 2 and Team 3 to System A and System B, respectively. As observed, percentiles 90%, 95% and 99% allow to characterize systems' high latencies, while all three systems having similar mean and median values. Team 2 exhibits the highest BLEU score at the cost of samples with higher latency ending up with a sample of up to 31 seconds. In this case, one could consider that Team 3, while slightly behind in terms of translation quality to that of Team 2, can be considered a better system due to its consistent lower latency towards the right tail of the distributions.

Similarly to Table 3, Table 4 shows from top to bottom, low, medium and latency bands in the English-Chinese IWSLT 2022 task, and reporting

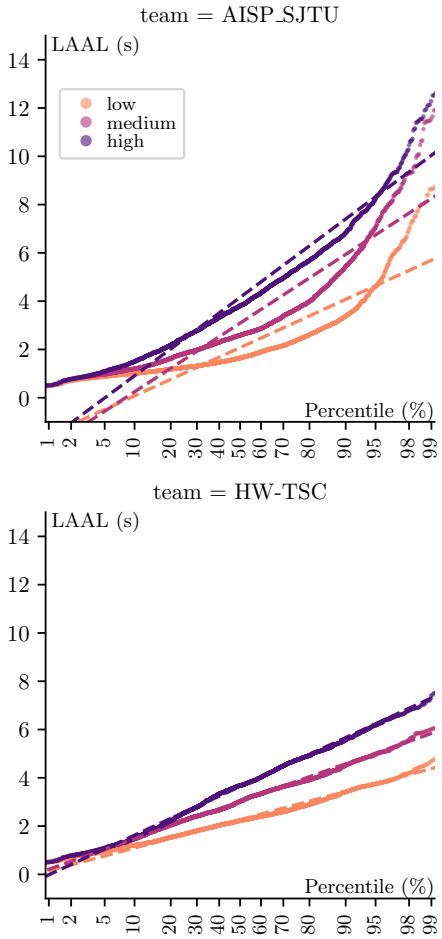


Figure 4: LAAL normal probability plot for IWSLT 2022 En-Zh AISP_SJTU (top) and HW-TSC (bottom) participations representing sample-level latencies (y-axis) w.r.t percentiles (x-axis) of a $\mathcal{N}(0, 1)$.

from left to right: BLEU score, and LAAL mean (M), median (mdn), percentiles 90%, 95% and 99%, and maximum value. As shown, across all bands the AISP_SJTU system achieves the highest BLEU scores and fairly low mean and median LAAL values ranging from 2.0 to 4.1 seconds. However, its latency for percentiles 90%, 95% and 99% are significantly higher than those of the other two teams, consistently suffering across latency bands from considerable worst cases with deltas ranging from 4.0 to 5.5 seconds between percentiles 95% and 99%, and from 10.1 to 19.7 seconds between 99% and the maximum value. Similar tables for English-German and English-Japanese IWSLT 2022 are available in Appendix D.

5.3 Over-wait

When analysing high latency samples on the right end of the distribution, a considerable amount of

team	BLEU	M	mdn	p90	p95	p99	max
Team 1	12.1	2.1	1.9	3.3	4.0	6.5	14.6
Team 2	19.3	2.4	2.0	4.2	5.3	7.5	31.3
Team 3	17.9	2.3	2.2	3.5	3.9	4.8	10.4

Table 3: BLEU scores and LAAL mean (M), median (mdn), percentiles 90%, 95% and 99%, and maximum value in the English-Japanese IWSLT 2024 task.

team	BLEU	M	mdn	p90	p95	p99	max
Low							
AISP_SJTU	30.7	2.0	1.6	3.3	4.5	8.6	18.5
CUNI-KIT	26.7	1.9	1.8	2.9	3.3	4.6	8.4
HW-TSC	19.1	2.2	2.2	3.4	3.7	4.6	12.2
Medium							
AISP_SJTU	31.2	3.0	2.5	5.4	7.0	11.5	27.5
CUNI-KIT	27.0	2.9	2.8	4.4	5.1	6.6	10.4
HW-TSC	26.0	3.0	3.0	4.5	4.9	6.0	11.4
High							
AISP_SJTU	32.0	4.1	3.8	6.8	8.2	12.2	32.1
CUNI-KIT	27.2	3.9	4.0	6.0	6.9	8.8	12.4
HW-TSC	27.6	3.6	3.6	5.5	6.1	7.3	12.5

Table 4: BLEU scores and LAAL mean (M), median (mdn), percentiles 90%, 95% and 99%, and maximum value in the English-Chinese IWSLT 2022 task for low (top), medium (middle), high (bottom) latency bands.

long samples are detected in which some SimulST systems exhibit a degenerated behavior waiting approximately until the end of the input to generate the translation. We refer to this phenomenon as the *over-wait* of a system. In other words, the ratio between the latency score and the length of the input tends to one. While this behavior is to expected to appear in short samples, it is extremely undesirable in the case of long samples.

Figure 5 illustrates the phenomenon of over-wait using sample-level latencies from the low (top), medium (middle) and high (bottom) latency bands generated by the AISP_SJTU system participating in the English-Chinese IWSLT 2022 task. As shown, each sample is plotted according to its source length (x-axis) and its latency (y-axis), with points falling on the diagonal indicating that the system did not write any token until reading the complete input. Lighter and darker colors correspond to ratios closer and further to one, respectively. As expected, short samples tend to accumulate ratios close to one, while it is not so frequent for long samples, but yet significant.

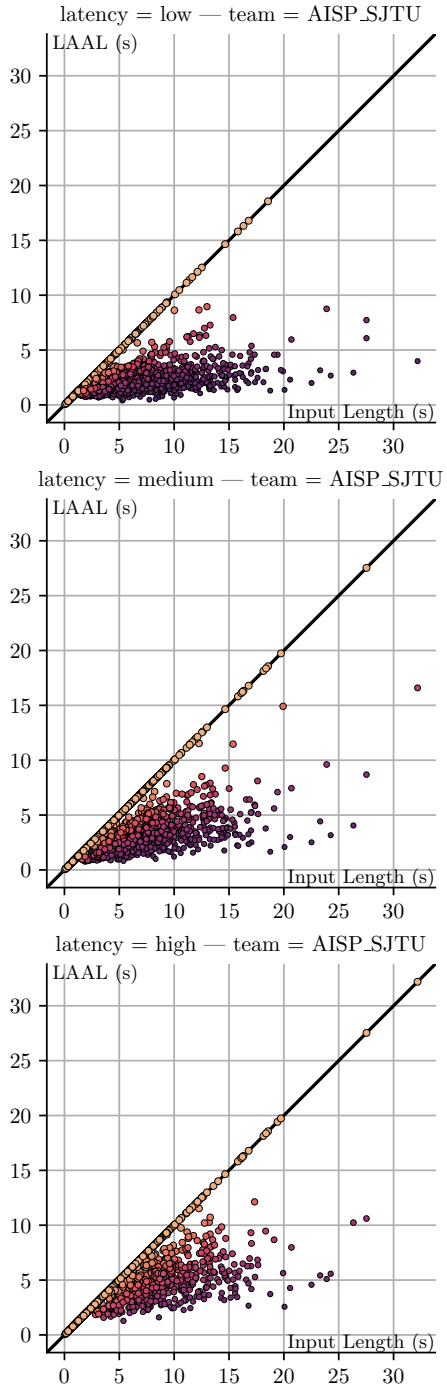


Figure 5: Sample-level latencies versus input length in low (top), medium (middle) and high (bottom) latency bands for the participation of the AISP_SJTU team in the English-Chinese IWSLT 2022 task.

To characterize over-wait, let us define OW_t^r as the percentage of samples whose duration is higher than t and the ratio between their latency score and their input length exceeds r. Table 5 reports over-wait OW_5^r with $r \in \{0.75, 0.85, 0.95, 1.00\}$

	r			
Lat. band	0.75	0.85	0.95	1.00
low	6.5	6.5	6.2	6.2
medium	17.0	16.0	15.7	15.7
high	48.3	38.7	33.1	32.6

Table 5: Over-wait (%) considering samples longer than 5 seconds with ratio $r \in \{0.75, 0.85, 0.95, 1.00\}$ in the low, medium and high latency bands for the AISP_SJTU team in the English-Chinese IWSLT 2022 task.

in terms of LAAL in the low, medium and high latency bands for the AISP_SJTU team in the English-Chinese IWSLT 2022 task. As expected, over-wait increases as we move from low over medium to high latency band. In the latter band, this means that in approximately one third of the samples the system waited for the end of the input to generate the full translation, behaving as a conventional offline translation system. The computation of over-wait for SimulST systems allows to easily detect this undesirable behavior in simultaneous translation. Over-wait figures for the rest of IWSLT 2022 systems are available in Appendix E.

6 Recommendations

In the previous section, a series of tools for descriptive statistics have been presented and illustrated on real data to study and characterize the latency of SimulST systems. The results presented in the previous section has allowed us to reflect on effective ways to report latency for SimulST systems in order to gain insight into their actual behavior in a real scenario.

First of all, as already encouraged in the evaluation of translation quality (Post, 2018; Zouhar et al., 2024), it is strongly recommended to report, in addition to the evaluation tool, the software version of the tool in order to guarantee reproducibility.

Our analysis shows how reporting a measure of central tendency such as the mean is not able to properly capture the underlying latency of a SimulST system and misleading comparisons across systems could be drawn. For this reason, it is convenient to provide descriptive statistics that offer an overall view of the system latency beyond the mean. In this sense, violin and specially normal probability plots were found significantly useful and enlightening to consistently compare systems and detect undesirable system latencies. In addition, normal probability plots proved to be an effective tool for assessing the normality of latency distributions, while also capturing other key descriptive statistics such as skewness and kurtosis, as well as the differences in percentile values across different systems. Complementarily, figures reporting latencies for higher percentiles, along with mean and median, are also recommended to prove the robustness of the system latency. Finally, over-wait scores are very valuable to identify the percentage of samples in which a SimulST system is exhibiting a degenerated offline behavior.

7 Conclusions

In this paper, we have critically examined the current practices for evaluating latency in SimulST systems, focusing on the limitations of relying solely on mean latency metrics. Through a detailed analysis of systems submitted to recent editions of the IWSLT SimulST track, mean latency metrics have demonstrated to fail to provide a complete view of SimulST systems' response time, particularly in the presence of skewed latency distributions and high latency values. Our findings reveal that the mean latency can mask unacceptable latency values which are critical for understanding the performance of these systems in a real scenario.

Alternative methods have been proposed for reporting descriptive statistics of latency metrics, emphasizing the importance of considering the entire latency distribution rather than just the mean value. Specifically, violin and specially normal probability plots were recommended to graphically report latency values per percentile in order to provide a more comprehensive view of the system behavior.

Our analysis underscores the need for more robust evaluation practices in SimulST research. We strongly believe that by adopting the recommendations outlined in this paper, researchers and practitioners can gain a deeper understanding of system performance, leading to more reliable and consistent SimulST systems in real-world applications.

Future work should continue to explore the development of new metrics and evaluation methodologies that better align with the challenges of real-time translation, ensuring that SimulST systems meet the demands of end-users in dynamic and continuous speech scenarios. In particular, the more realistic stream-level latency metrics (Iranzo-Sánchez et al., 2021) must be revisited taking into account the lessons learned in this work.

8 Limitations

In this article we have restricted ourselves to the usage of non-computationally aware metrics to simplify the resulting analysis and to avoid possible inconsistencies such as those indicated in Sperber et al. (2024). Findings in non-computationally aware metrics can be easily extrapolated to computationally aware measures, since the former can be understood as a best case scenario for the latter. In addition, this study was performed on a limited subset of languages and models obtained from the past IWSLT editions. A more extensive study with larger models (Macháček et al., 2023b; Communication et al., 2023; Labiausse et al., 2025), languages directions and datasets may provide deeper insight on distinct aspects of latency evaluation.

References

Ibrahim Said Ahmad, Antonios Anastasopoulos, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, William Chen, Qianqian Dong, Marcello Federico, Barry Haddow, Dávid Javorský, Mateusz Krubiński, Tsz Kin Lam, Xutai Ma, Prashant Mathur, Evgeny Matusov, Chandresh Maurya, John McCrae, Kenton Murray, Satoshi Nakamura, Matteo Negri, Jan Niehues, Xing Niu, Atul Kr. Ojha, John Ortega, Sara Papi, Peter Polák, Adam Pospíšil, Pavel Pecina, Elizabeth Salesky, Nivedita Sethiya, Balaram Sarkar, Jiatong Shi, Claytone Sikasote, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Brian Thompson, Alex Waibel, Shinji Watanabe, Patrick Wilken, Petr Zemánek, and Rodolfo Zevallos. 2024. FINDINGS OF THE IWSLT 2024 EVALUATION CAMPAIGN. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 1–11, Bangkok, Thailand (in-person and online). Association for Computational Linguistics.

Antonios Anastasopoulos, Loïc Barrault, Luisa Bentivogli, Marcely Zanon Boito, Ondřej Bojar, Roldano Cattoni, Anna Currey, Georgiana Dinu, Kevin Duh, Maha Elbayad, Clara Emmanuel, Yannick Estève, Marcello Federico, Christian Federmann, Souhir Gahbiche, Hongyu Gong, Roman Grundkiewicz, Barry Haddow, Benjamin Hsu, Dávid Javorský, Věra Kloudová, Surafel Lakew, Xutai Ma, Prashant Mathur, Paul McNamee, Kenton Murray, Maria Nădejde, Satoshi Nakamura, Matteo Negri, Jan Niehues, Xing Niu, John Ortega, Juan Pino, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Yogesh Virkar, Alexander Waibel, Changhan Wang, and Shinji Watanabe. 2022. Findings of the IWSLT 2022 evaluation campaign. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 98–157, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Colin Cherry and George Foster. 2019. Thinking Slow about Latency Evaluation for Simultaneous Machine Translation. *Preprint*, arXiv:1906.00048.

Kyunghyun Cho and Masha Esipova. 2016. Can neural machine translation do simultaneous translation? *Preprint*, arXiv:1606.02012.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, John Hoffman, Min-Jae Hwang, Hirofumi Inaguma, Christopher Klaiber, Ilia Kulikov, Pengwei Li, Daniel Licht, Jean Maillard, Ruslan Mavlyutov, Alice Rakotoarison, Kaushik Ram Sadagopan, Abinesh Ramakrishnan, Tuan Tran, Guillaume Wenzek, Yilin Yang, Ethan Ye, Ivan Evtimov, Pierre Fernandez, Cynthia Gao, Prangthip Hansanti, Elahe Kalbassi, Amanda Kallet, Artyom Kozhevnikov, Gabriel Mejia Gonzalez, Robin San Roman, Christophe Touret, Corinne Wong, Carleigh Wood, Bokai Yu, Pierre Andrews, Can Balioglu, Peng-Jen Chen, Marta R. Costa-jussà, Maha Elbayad, Hongyu Gong, Francisco Guzmán, Kevin Heffernan, Somya Jain, Justine Kao, Ann Lee, Xutai Ma, Alex Mourachko, Benjamin Peloquin, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Anna Sun, Paden Tomasello, Changhan Wang, Jeff Wang, Skyler Wang, and Mary Williamson. 2023. Seamless: Multilingual Expressive and Streaming Speech Translation. *Preprint*, arXiv:2312.05187.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a Multilingual Speech Translation Corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2012–2017, Minneapolis, Minnesota. Association for Computational Linguistics.

Yadolah Dodge. 2008. *Normal Probability Plot*, pages 382–385. Springer New York, New York, NY.

Jiatao Gu, Graham Neubig, Kyunghyun Cho, and Victor O.K. Li. 2017. Learning to translate in real-time with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1053–1062, Valencia, Spain. Association for Computational Linguistics.

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.

Javier Iranzo-Sánchez, Jorge Civera Saiz, and Alfons Juan. 2021. Stream-level latency evaluation for simultaneous machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 664–670, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Javier Iranzo-Sánchez, Jorge Civera Saiz, and Alfons Juan. 2021. Stream-level Latency Evaluation for Simultaneous Machine Translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 664–670, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yasumasa Kano, Katsuhito Sudoh, and Satoshi Nakamura. 2023a. Average Token Delay: A Duration-aware Latency Metric for Simultaneous Translation. *Preprint*, arXiv:2311.14353.

Yasumasa Kano, Katsuhito Sudoh, and Satoshi Nakamura. 2023b. Average token delay: A latency metric for simultaneous translation. In *24th Annual Conference of the International Speech Communication Association, Interspeech 2023, Dublin, Ireland, August 20-24, 2023*, pages 4469–4473.

Tom Labiausse, Laurent Mazaré, Edouard Grave, Patrick Pérez, Alexandre Défossez, and Neil Zeghidour. 2025. High-fidelity simultaneous speech-to-speech translation. *Preprint*, arXiv:2502.03382.

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.

Xutai Ma, Mohammad Javad Dousti, Changhan Wang, Jiatao Gu, and Juan Pino. 2020. SIMULEVAL: An evaluation toolkit for simultaneous translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 144–150, Online. Association for Computational Linguistics.

Dominik Macháček, Ondřej Bojar, and Raj Dabre. 2023a. 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, Raj Dabre, and Ondřej Bojar. 2023b. Turning whisper into real-time transcription system. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 17–24, Bali, Indonesia. Association for Computational Linguistics.

Mana Makinae, Katsuhito Sudoh, Mararu Yamada, and Satoshi Nakamura. 2024. A Word Order Synchronization Metric for Evaluating Simultaneous Interpretation and Translation. *Preprint*, arXiv:2407.06650.

Sara Papi, Marco Gaido, Matteo Negri, and Marco Turchi. 2022. Over-generation cannot be rewarded: Length-adaptive average lagging for simultaneous speech translation. In *Proceedings of the Third Workshop on Automatic Simultaneous Translation*, pages 12–17, Online. Association for Computational Linguistics.

Sara Papi, Peter Polak, Ondřej Bojar, and Dominik Macháček. 2024. How "Real" is Your Real-Time Simultaneous Speech-to-Text Translation System? *Preprint*, arXiv:2412.18495.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Kristin L. Sainani. 2012. Dealing with non-normal data. *PM&R*, 4(12):1001–1005.

S. S. Shapiro and M. B. Wilk. 1965. An analysis of variance test for normality (complete samples)†. *Biometrika*, 52(3-4):591–611.

Matthias Sperber, Ondřej Bojar, Barry Haddow, Dávid Javorský, Xutai Ma, Matteo Negri, Jan Niehues, Peter Polák, Elizabeth Salesky, Katsuhito Sudoh, and Marco Turchi. 2024. Evaluating the IWSLT2023 speech translation tasks: Human annotations, automatic metrics, and segmentation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6484–6495, Torino, Italia. ELRA and ICCL.

Changhan Wang, Juan Pino, Anne Wu, and Jiatao Gu. 2020. CoVoST: A diverse multilingual speech-to-text translation corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4197–4203, Marseille, France. European Language Resources Association.

Shira Wein, Te I, Colin Cherry, Juraj Juraska, Dirk Padfield, and Wolfgang Macherey. 2024. Barriers to effective evaluation of simultaneous interpretation. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 209–219, St. Julian's, Malta. Association for Computational Linguistics.

M. B. Wilk and R. Gnanadesikan. 1968. Probability plotting methods for the analysis of data. *Biometrika*, 55(1):1–17.

Xi Xu, Wenda Xu, Siqi Ouyang, and Lei Li. 2024. CA*: Addressing Evaluation Pitfalls in Computation-Aware Latency for Simultaneous Speech Translation. *Preprint*, arXiv:2410.16011.

Vilém Zouhar, Pinzhen Chen, Tsz Kin Lam, Nikita Moghe, and Barry Haddow. 2024. Pitfalls and outlooks in using COMET. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1272–1288, Miami, Florida, USA. Association for Computational Linguistics.

A IWSLT 2022: MuST-C

Figure 6 shows MuST-C violin plots for LAAL. Figures 7, 8 and 9 show LAAL normal probability plots for En-De, En-Zh and En-Ja of the MuST-C partition from IWSLT 2022. Figure 10, 11 and 12 show LAAL vs Input length over-wait graphs for En-De, En-Zh and En-Ja of MuST-C from IWSLT 2022.

B IWSLT 2022: Official, Other Metrics

Figures 13, 14 and 15 show violin plots for AL, DAL and ATD on the official IWSLT 2022 datasets for all available languages.

C IWSLT 2022: Normal Probability Plots

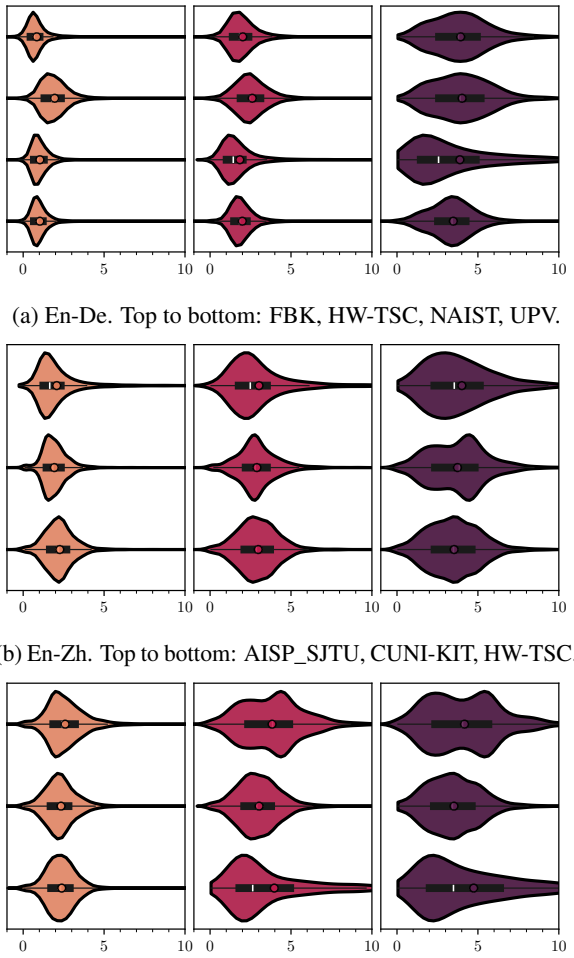
Figures 16, 17 and 18 show LAAL normal probability plots for En-De, En-Zh and En-Ja of IWSLT 2022.

D IWSLT 2022: Additional Tables

Tables 6 and 7 show numerical results for En-De and En-Ja.

E IWSLT 2022-2024: Over-wait

Figure 19, 20 and 21 show LAAL vs Input length over-wait graphs for En-De, En-Zh and En-Ja of IWSLT 2022. Figure 22 shows results for IWSLT 2024 En-Ja.



(a) En-De. Top to bottom: FBK, HW-TSC, NAIST, UPV.

(b) En-Zh. Top to bottom: AISP_SJTU, CUNI-KIT, HW-TSC.

(c) En-Ja. Top to bottom: CUNI-KIT, HW-TSC, NAIST.

Figure 6: LAAL distributions for the IWSLT 2022 Must-C team participations. From left to right, low, medium and high band systems for each language.

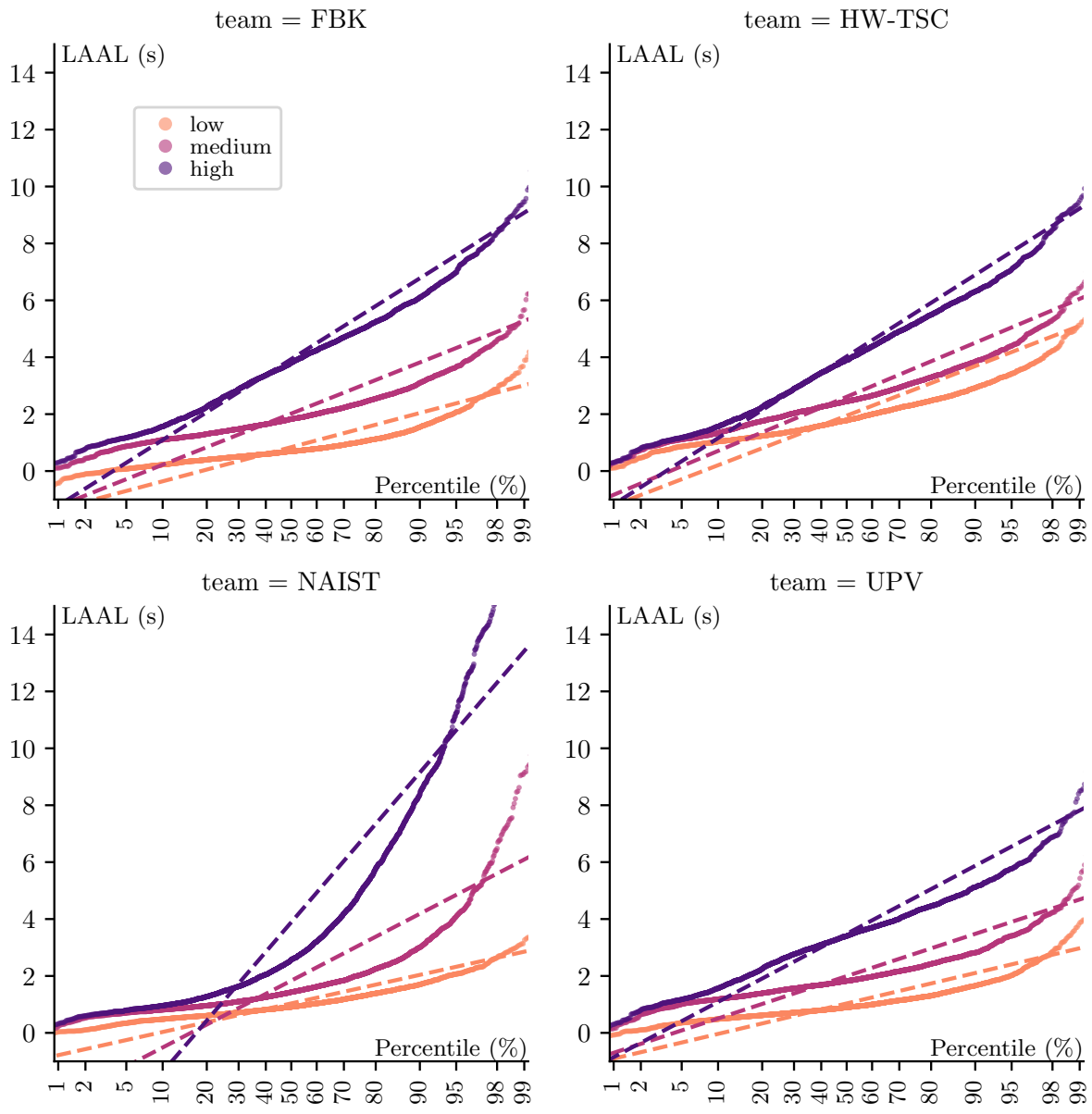
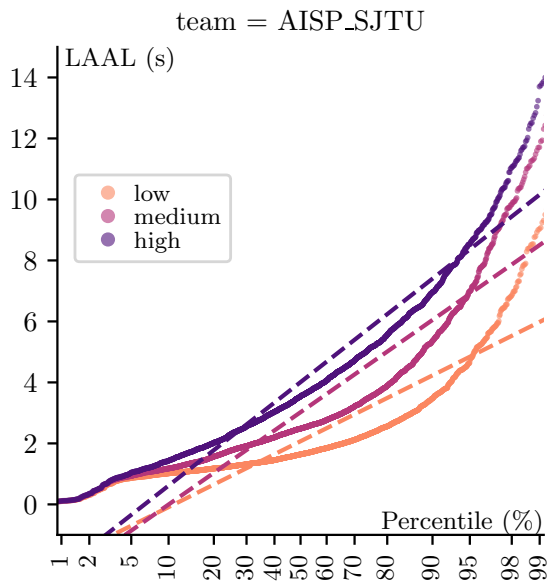
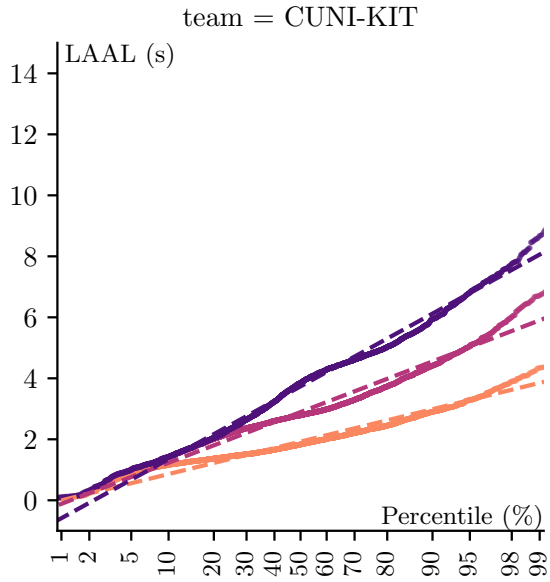


Figure 7: IWSLT2022 En-De MuST-C Test set LAAL normal probability plot.





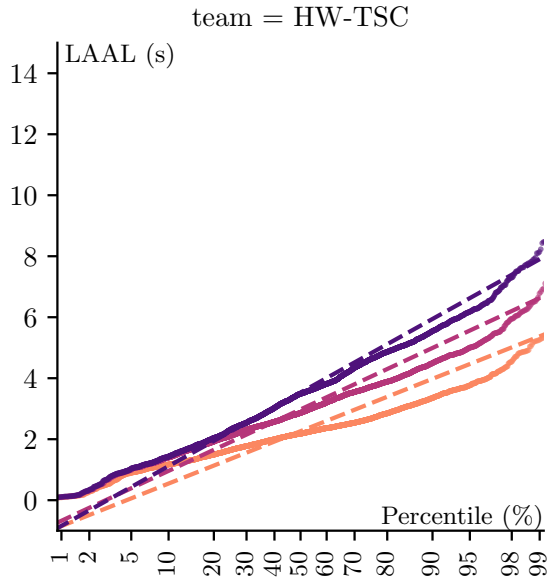
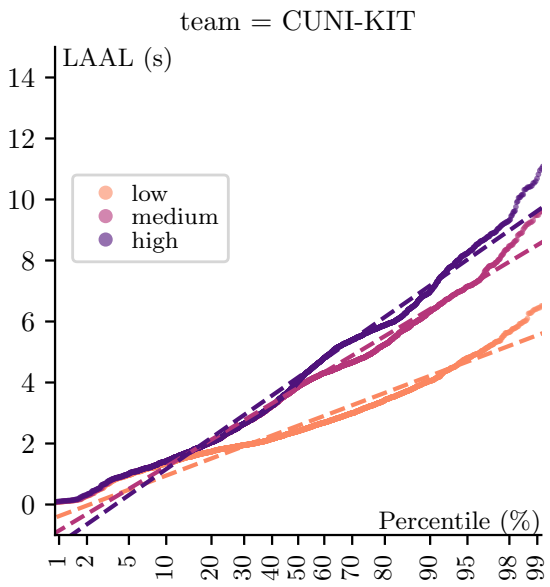
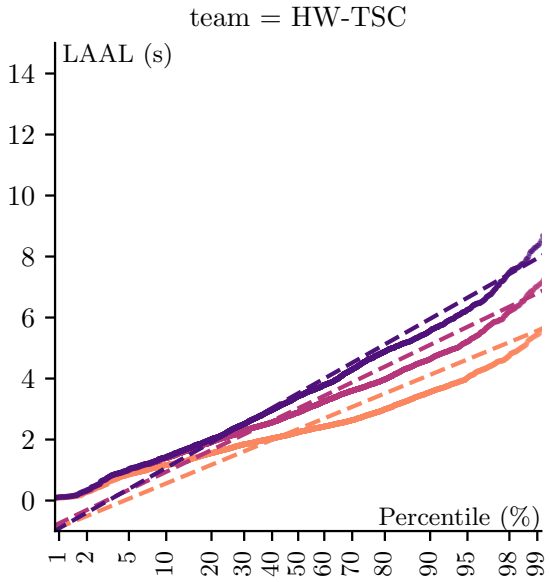


Figure 8: IWSLT2022 En-Zh MuST-C Test set LAAL normal probability plot.





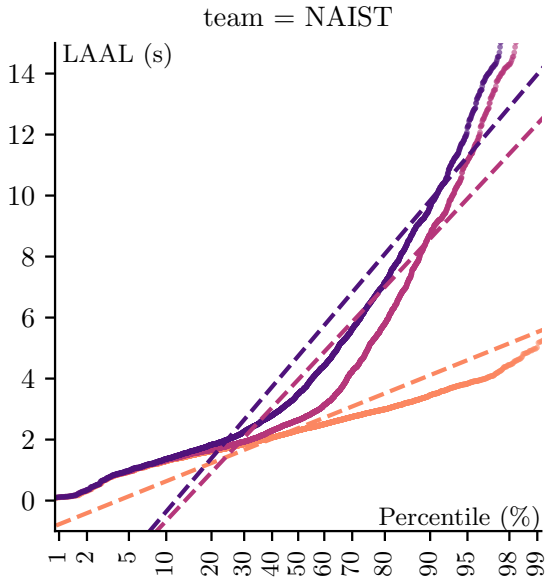


Figure 9: IWSLT2022 En-Ja MuST-C Test set LAAL normal probability plot.

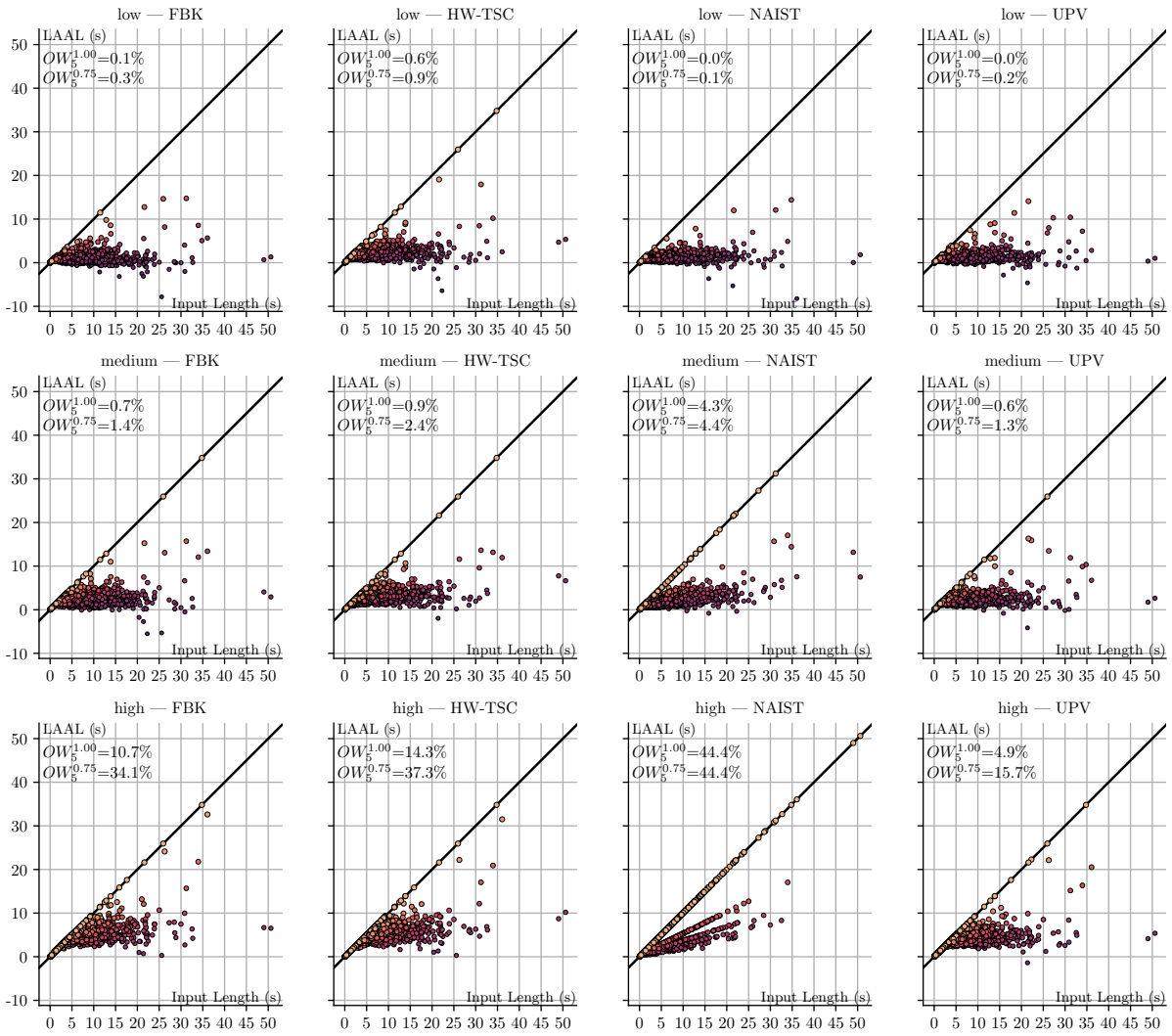


Figure 10: Over-wait graphs for MuST-C 2022 LAAL En-De.

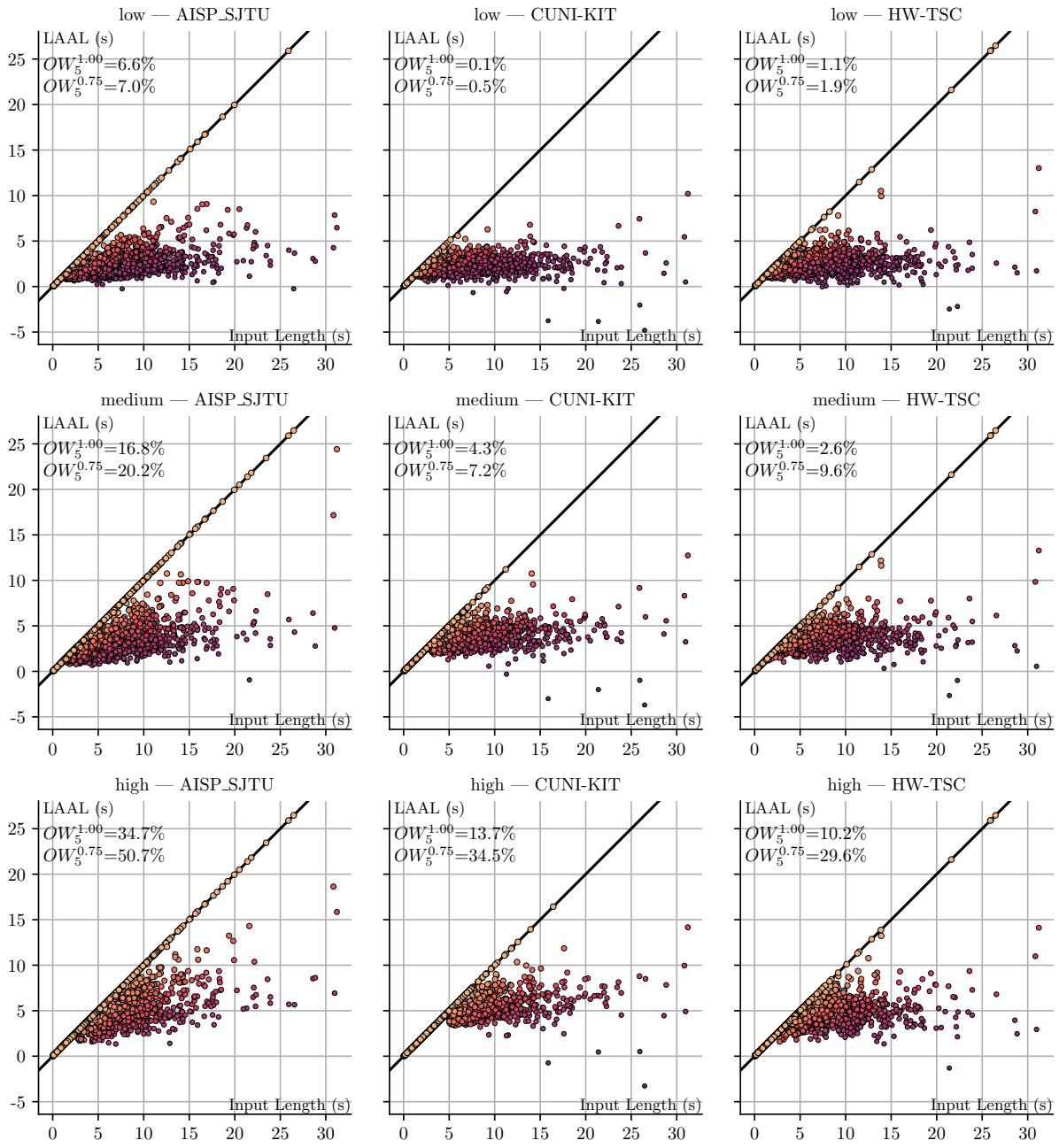


Figure 11: Over-wait graphs for MuST-C 2022 LAAL En-Zh.

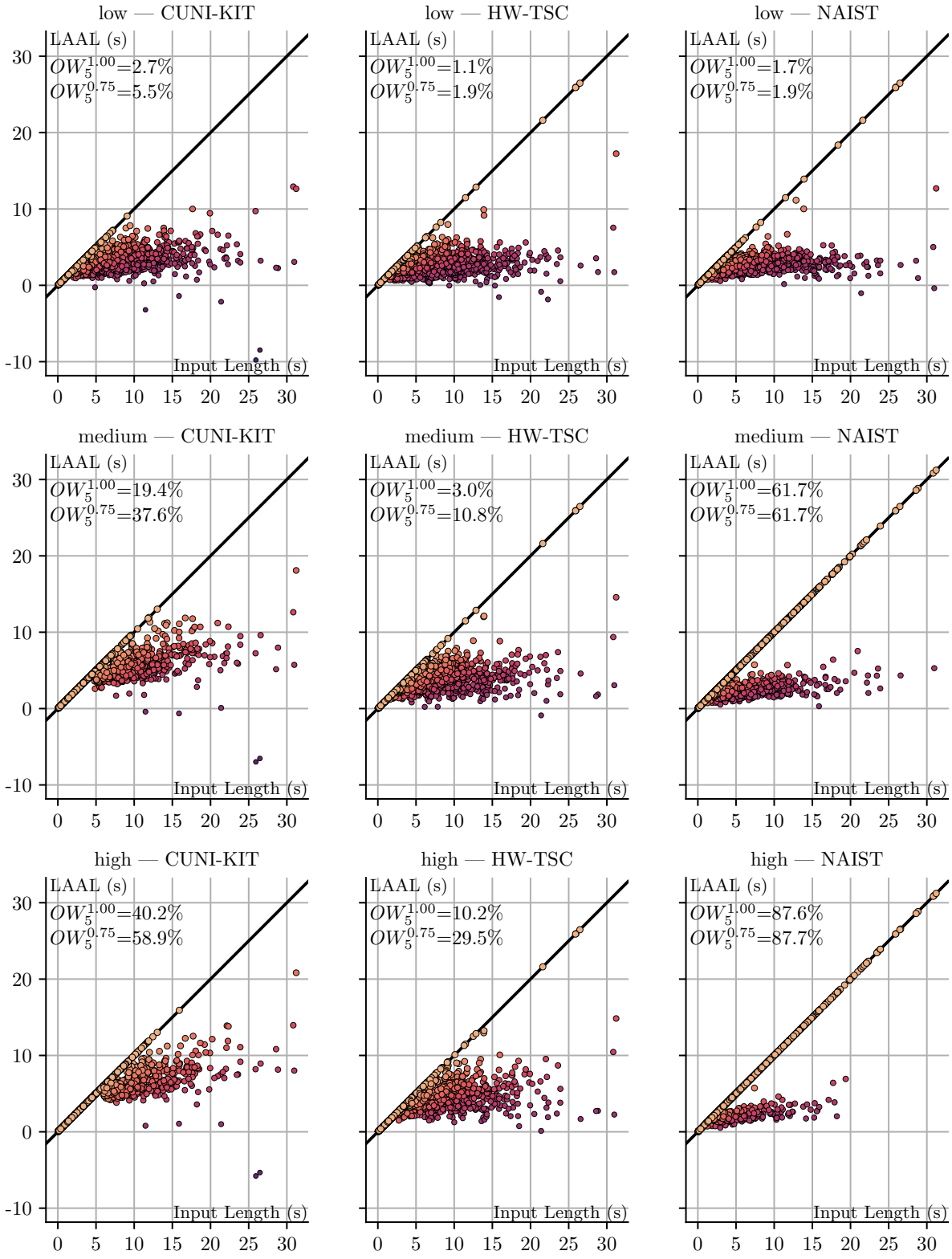
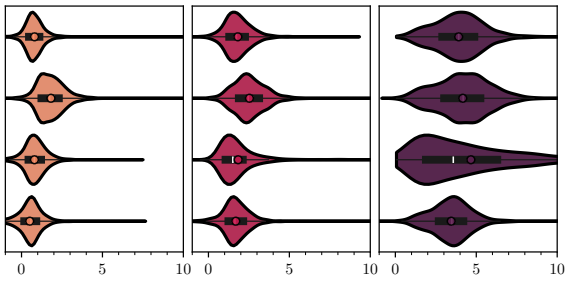
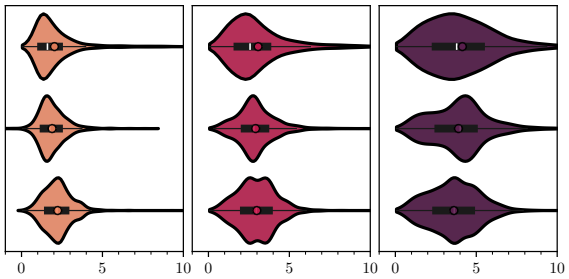


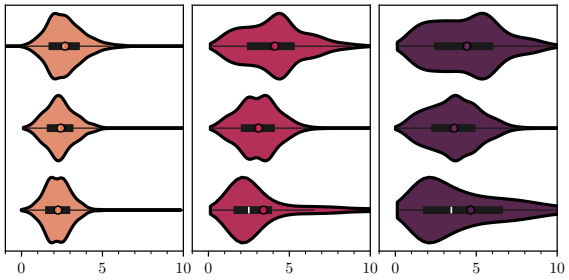
Figure 12: Over-wait graphs for MuST-C 2022 LAAL En-Ja.



(a) En-De. Top to bottom: FBK, HW-TSC, NAIST, UPV.

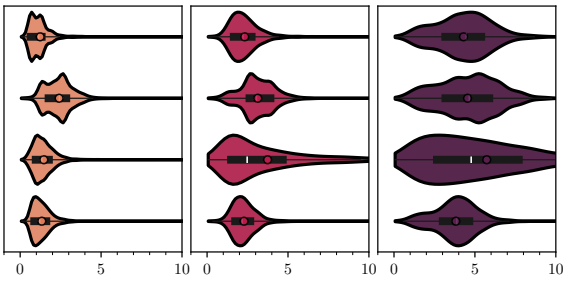


(b) En-Zh. Top to bottom: AISP_SJTU, CUNI-KIT, HW-TSC.

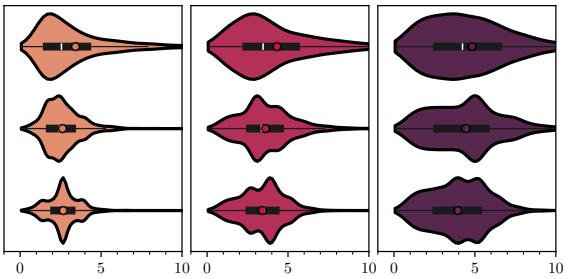


(c) En-Ja. Top to bottom: CUNI-KIT, HW-TSC, NAIST.

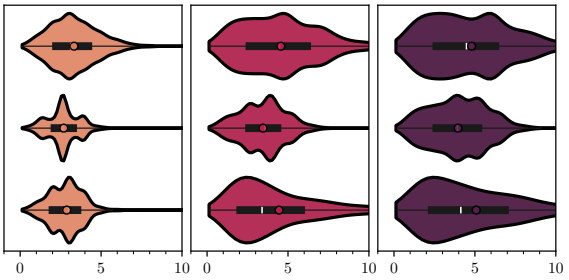
Figure 13: AL distributions for the IWSLT 2022 official test sets team participations. From left to right, low, medium and high band systems for each language.



(a) En-De. Top to bottom: FBK, HW-TSC, NAIST, UPV.

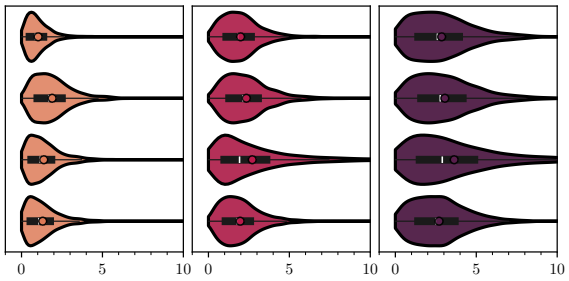


(b) En-Zh. Top to bottom: AISP_SJTU, CUNI-KIT, HW-TSC.

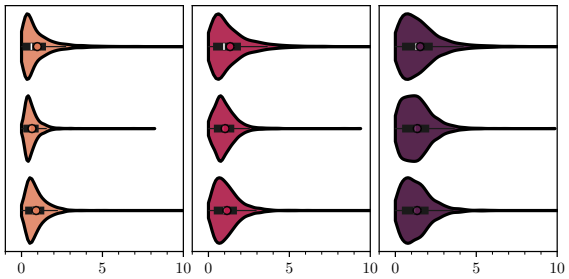


(c) En-Ja. Top to bottom: CUNI-KIT, HW-TSC, NAIST.

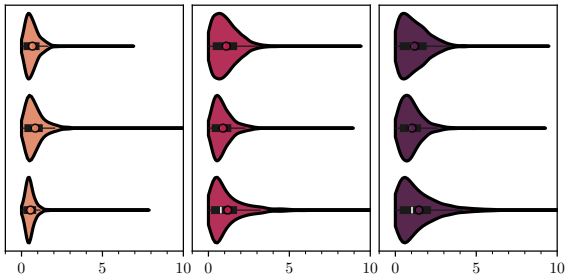
Figure 14: DAL distributions for the IWSLT 2022 official test sets team participations. From left to right, low, medium and high band systems for each language.



(a) En-De. Top to bottom: FBK, HW-TSC, NAIST, UPV.



(b) En-Zh. Top to bottom: AISP_SJTU, CUNI-KIT, HW-TSC.



(c) En-Ja. Top to bottom: CUNI-KIT, HW-TSC, NAIST.

Figure 15: ATD distributions for the IWSLT 2022 official test sets team participations. From left to right, low, medium and high band systems for each language.

team	BLEU	M	p50	p90	p95	p99	max
Low							
FBK	10.2	0.9	0.8	1.6	2.0	3.4	10.6
HW-TSC*	13.9	1.9	1.8	2.9	3.3	4.3	12.5
NAIST	13.4	1.0	0.9	1.7	2.0	2.9	8.4
UPV	16.0	1.0	0.9	1.6	1.9	2.8	8.9
Medium							
FBK	20.1	1.9	1.8	3.0	3.5	4.9	9.3
HW-TSC	19.1	2.6	2.5	3.8	4.3	5.4	10.9
NAIST	15.2	1.8	1.5	3.1	3.8	8.4	18.7
UPV	21.1	1.9	1.8	2.8	3.1	4.3	10.9
High							
FBK	23.6	4.0	4.0	5.8	6.5	8.0	12.0
HW-TSC	19.7	4.2	4.2	6.2	6.9	8.0	13.1
NAIST	15.4	4.6	3.5	9.5	12.5	17.7	28.3
UPV	23.5	3.5	3.6	5.0	5.4	6.8	10.9

Table 6: Metric Values for Official Test Set for IWSLT2022 English to German. The low HW-TSC system did not originally comply with the latency constraints.

team	BLEU	M	mdn	p90	p95	p99	max
Low							
CUNI-KIT	16.5	2.7	2.6	4.2	4.7	5.9	10.0
HW-TSC	5.6	2.4	2.3	3.6	4.1	4.8	12.7
NAIST	8.7	2.3	2.2	3.3	3.6	4.3	9.8
Medium							
CUNI-KIT	16.6	4.1	4.1	6.7	7.5	9.6	13.5
HW-TSC	11.7	3.1	3.1	4.7	5.1	6.2	12.0
NAIST	9.4	3.4	2.4	7.1	9.3	14.6	32.1
High							
CUNI-KIT	16.7	4.4	4.4	7.3	8.3	10.7	18.3
HW-TSC	11.4	3.6	3.6	5.7	6.2	7.3	13.2
NAIST	9.8	4.6	3.4	9.3	12.2	17.0	32.1

Table 7: Metric Values for Official Test Set for IWSLT2022 English to Japanese.

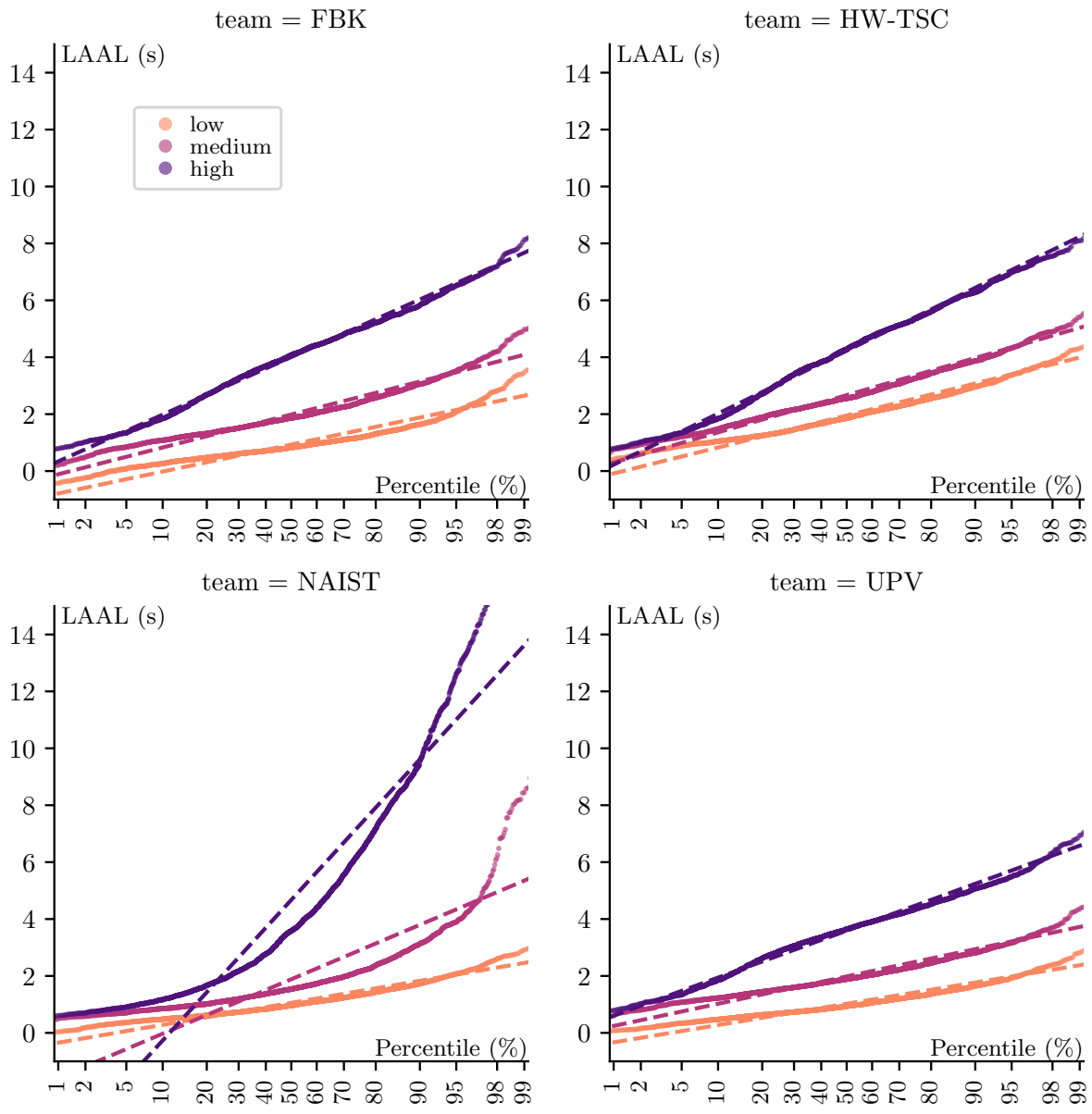
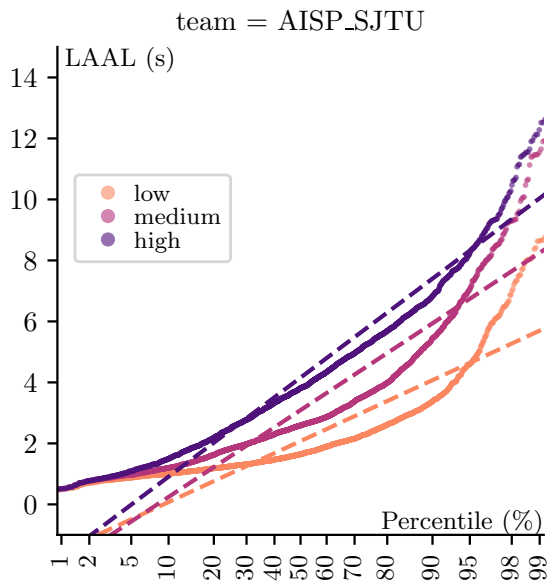
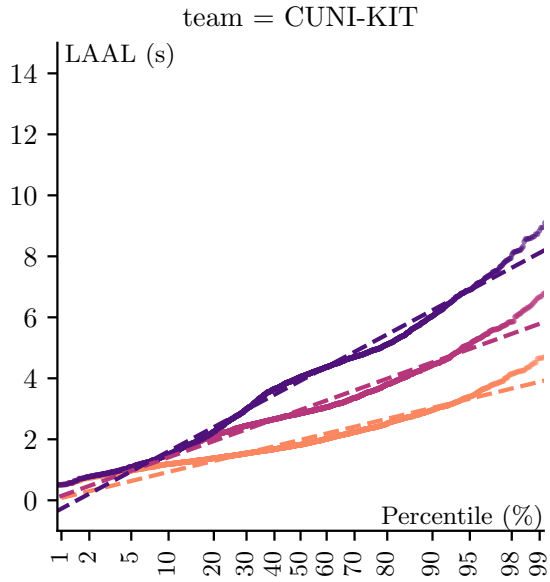


Figure 16: IWSLT2022 En-De Official Test set LAAL normal probability plot.





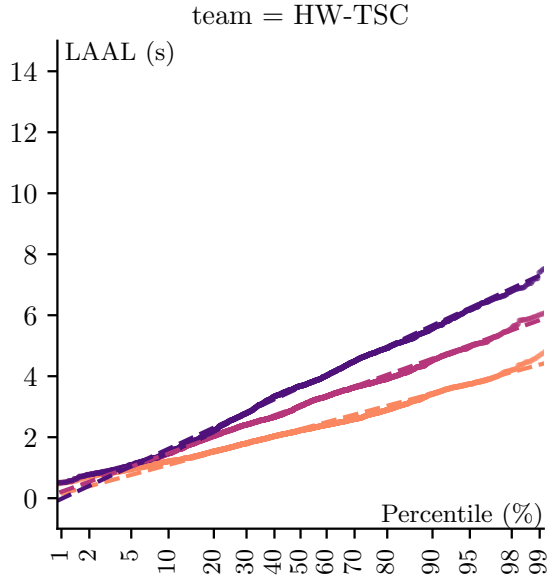
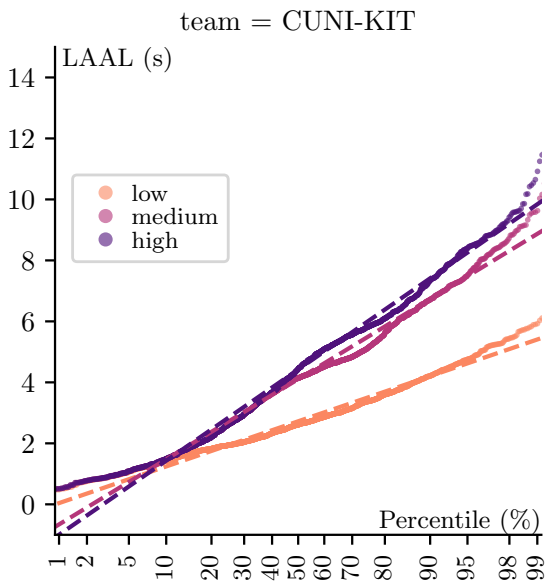
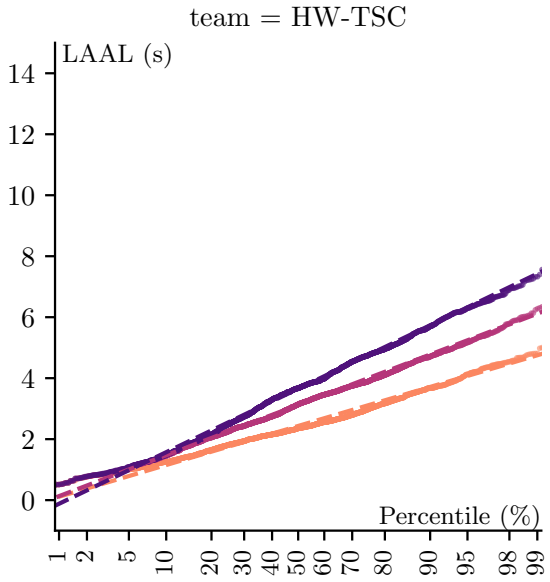


Figure 17: IWSLT2022 En-Zh Official Test set LAAL normal probability plot.





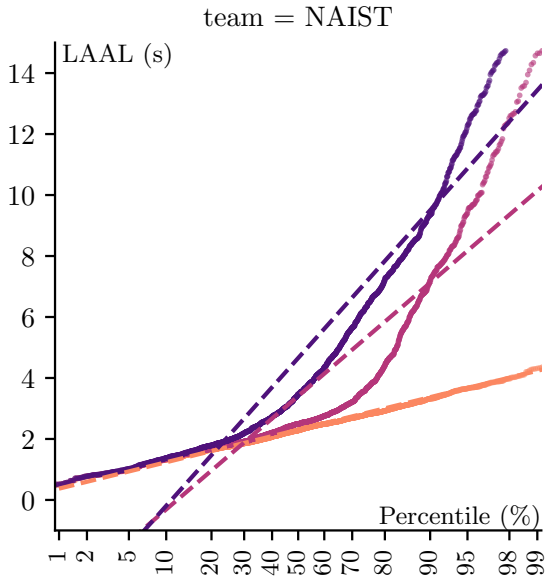


Figure 18: IWSLT2022 En-Ja Official Test set LAAL normal probability plot.

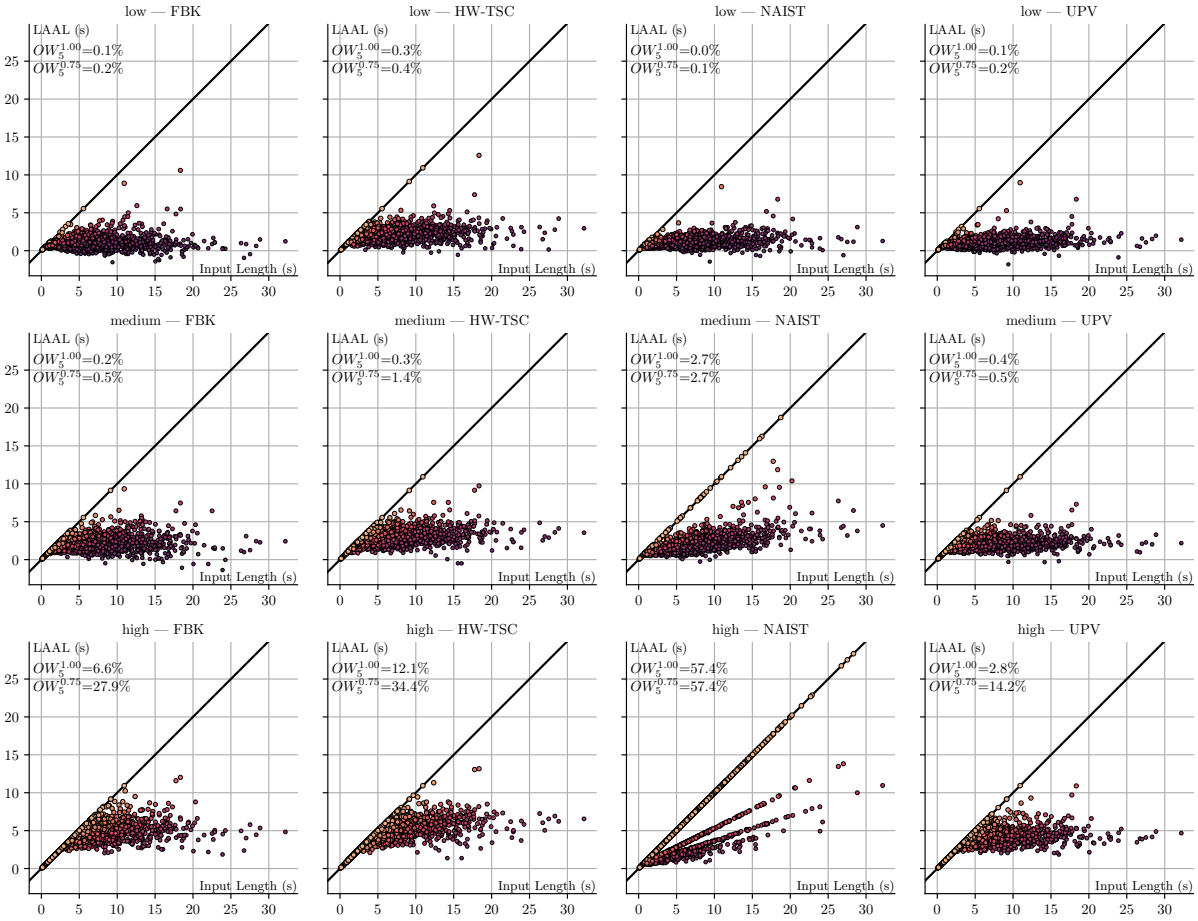


Figure 19: Over-wait graphs for IWSLT 2022 LAAL En-De.

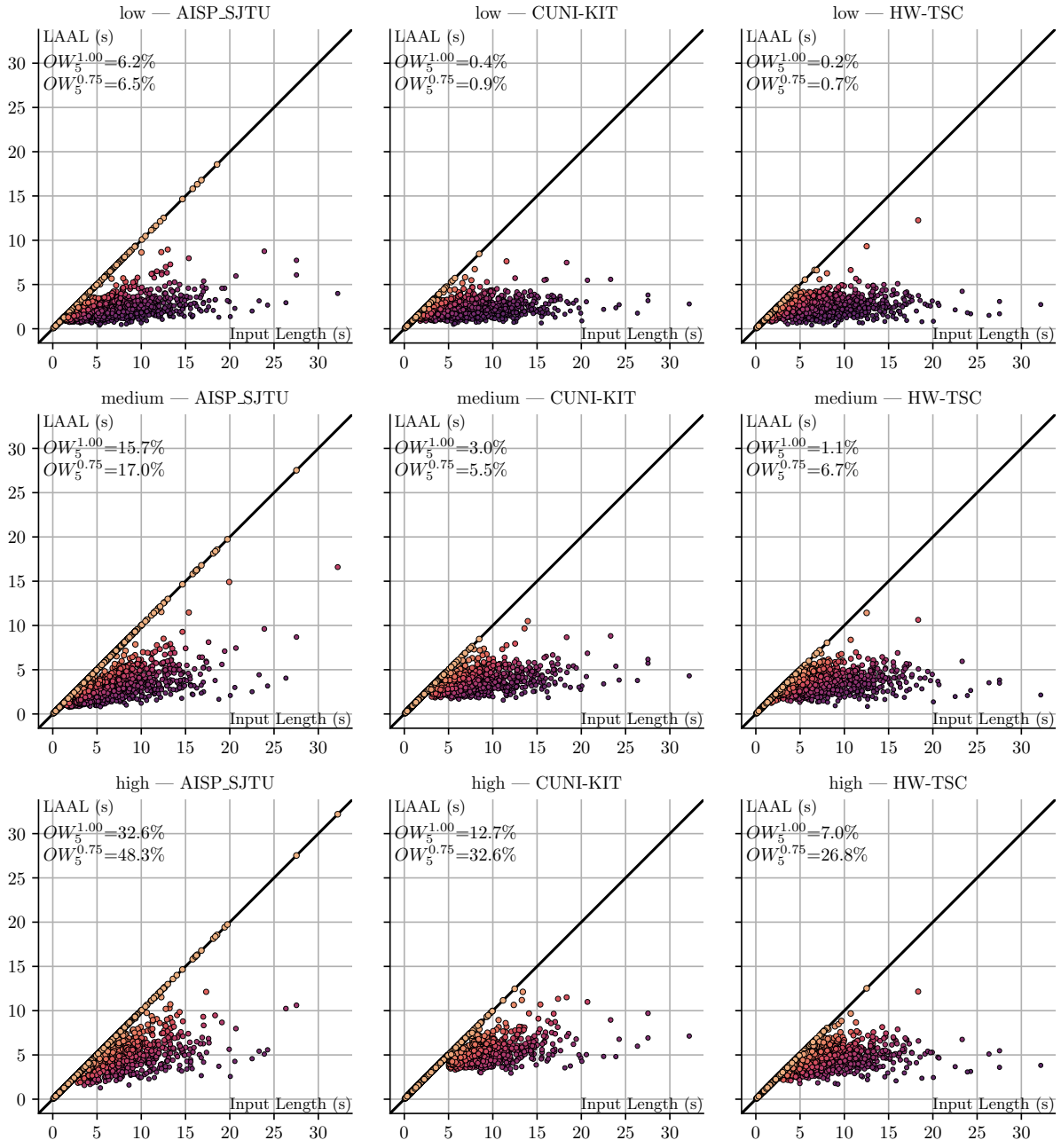


Figure 20: Over-wait graphs for IWSLT 2022 LAAL En-Zh.

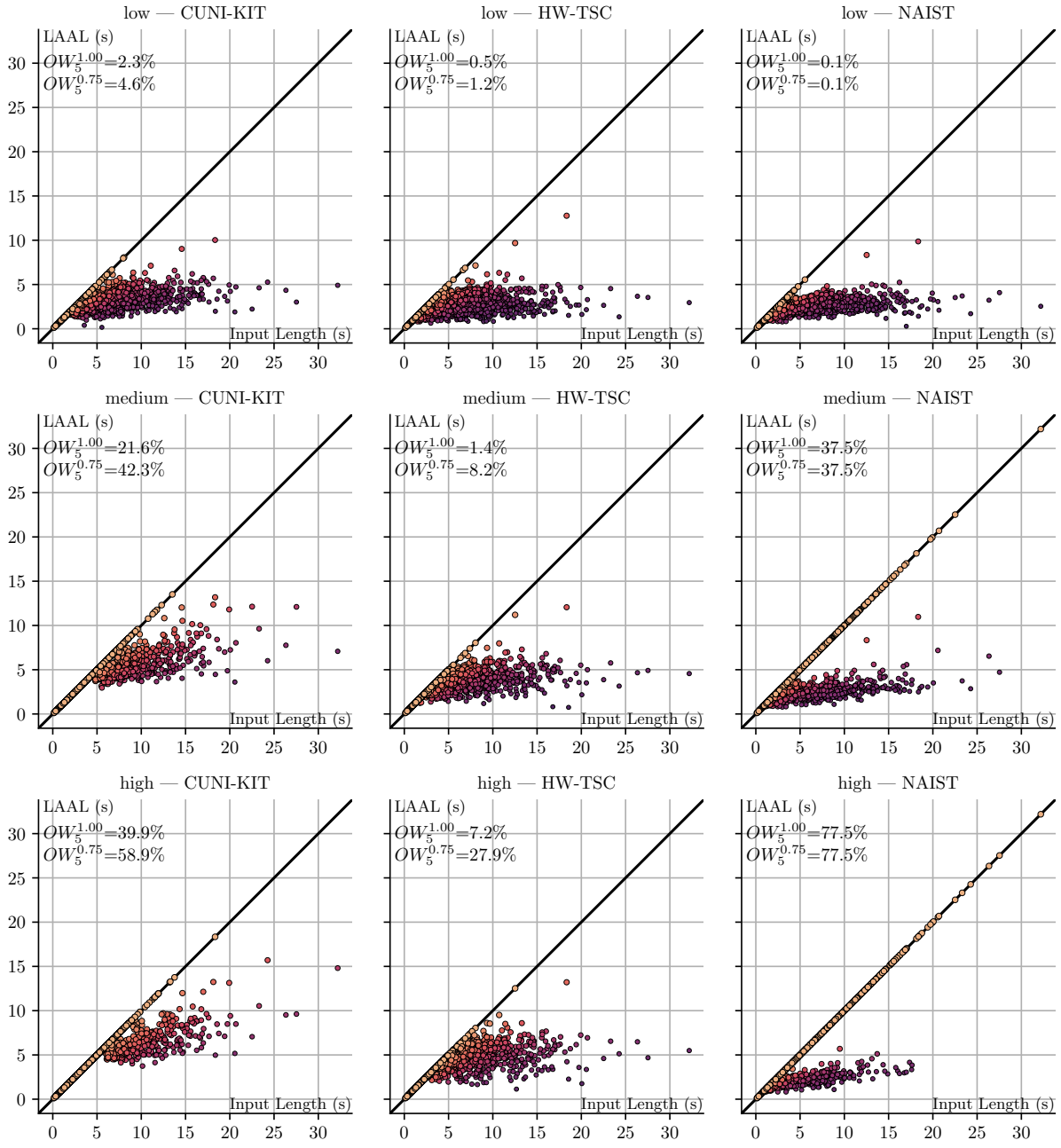


Figure 21: Over-wait graphs for IWSLT 2022 LAAL En-Ja.

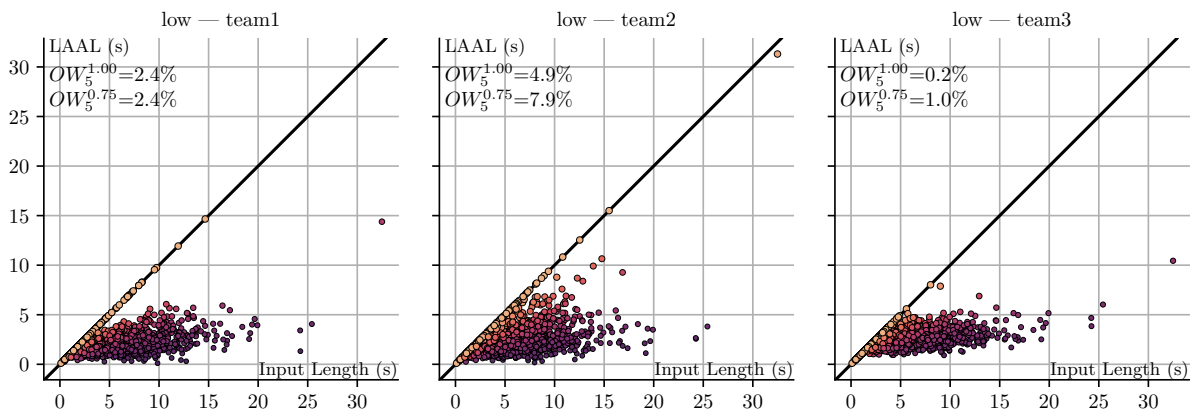


Figure 22: Over-wait graphs for IWSLT 2024 LAAL En-Ja.