On the Fallacy of Global Token Perplexity in Spoken Language Model Evaluation

Chan-Jan Hsu[‡], **Liang-Hsuan Tseng**[♭], **Yi-Cheng Lin**[♭], **Yen-Chun Kuo**[♭]
Ju-Chieh Chou[♮], **Kai-Wei Chang**[♣], **Hung-yi Lee**[♭], **Carlos Busso**[‡]
[‡]Carnegie Mellon University, [♭]National Taiwan University,
[♮]Toyota Technological Institute at Chicago, [♣]Massachusetts Institute of Technology
chanjanh@andrew.cmu.edu, busso@cmu.edu

Abstract

Generative spoken language models pretrained on large-scale raw audio can continue a speech prompt with coherent content while preserving attributes such as speaker identity and emotion, making them promising foundation models for spoken dialogue. In prior literature, these models are often evaluated using "global token perplexity," which directly transfers the text perplexity formulation to speech tokens. However, this practice overlooks fundamental differences between speech and text, potentially underestimating important speech characteristics. In this work, we propose a set of likelihood-based and generative-based evaluation methods as alternatives to naive global token perplexity. We show that these evaluations more faithfully reflect perceived generation quality, as evidenced by stronger correlations with human-rated mean opinion scores (MOS). Under these new metrics, the relative performance landscape of spoken language models shifts substantially, revealing a much smaller gap between the best-performing model and the human topline. Together, these results suggest that appropriate evaluation is critical for accurately assessing progress in spoken language modeling.[1]

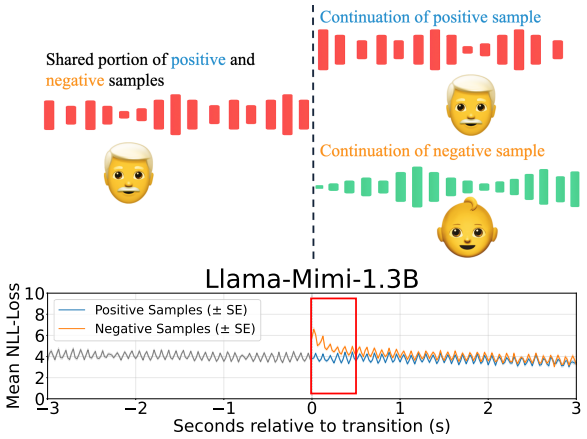

Figure 1: **Acoustic discontinuity disturbs negative log-likelihood loss (NLL) responses locally.** Top: SALMon samples consist of a shared prompt and a separate continuation, where positive samples maintain acoustic consistency, and negative samples contain abrupt acoustic transitions. Bottom: NLL response of Llama-Mimi-1.3B with standard error margins. Response on negative samples show localized spike within a short temporal window after the transition in contrast to the positive sample. Global token perplexity aggregates likelihood contributions outside this localized region (Sec. 2.2), making it susceptible to long-range loss volatility, thereby motivating our localized and normalized evaluation methods (Sec. 3.1, 3.2)

1 Introduction

Recent years have witnessed the emergence of assistant-style spoken dialogue systems that interact with users through speech (Open-AI; Chu et al., 2024; Défossez et al., 2024). Analogous to text language models, generative speech modeling is commonly formulated as a sequence-to-sequence task (Sugiura et al., 2025; Chang et al., 2024). In this paradigm, speech waveforms are first discretized into sequences of tokens (Défossez et al., 2024; An et al., 2024); model predictions are then performed in the token space (Sugiura et al., 2025; Tseng et al., 2025); finally, the resulting audio is reconstructed from the predicted tokens (Du et al., 2024; Ju et al., 2024). We refer to this modeling framework as a *spoken language model (SLM)*, following the terminology of Arora et al. (2025). Similar to the development of text-based large language models (LLMs), SLMs are typically trained in two stages: large-scale pretraining on unlabeled speech data, followed by supervised fine-tuning on task- or instruction-specific datasets.

During the *pre-training stage*, the model acquires foundational knowledge from raw data, which is directly reflected in its generation capabilities. Consequently, the quality of the pre-trained model lays the foundation for downstream perfor-

[1]Code and data open-sourced at https://github.com/Lab-MSP/SpeechPerplexity

Findings of the Association for Computational Linguistics: ACL 2026, pages 37000–37020

Figure 2: NLL loss response of various models on SALMon samples with standard error margins. High-scoring models on SALMon (e.g., Flow-SLM) exhibit localized NLL spikes for negative samples within a short temporal window after the transition. This behavior is less apparent in lower-performing models such as GSLM.

mance (Raffel et al., 2020; Hassid et al., 2024). However, evaluating the generation quality of pre-trained models remains challenging, as assessing whether a generated output is plausible or coherent is inherently subjective.

In text modeling, a common practice is to adopt *perplexity*, a likelihood-based metric that scores model predictions on textual token sequences. Likewise, pre-trained SLMs are often evaluated by computing perplexity over an entire sequence of discrete speech tokens. We refer to this metric as **"global token perplexity,"** which quantifies the sequence-level likelihood of a speech signal in the discrete token space. In text modeling benchmarks, perplexity is typically interpreted in a comparative setting (Brown et al., 2020), where each benchmark entry contrasts positive samples (i.e., fluent, coherent text) with negative samples (i.e., syntactically or semantically corrupted text). The model's ability to assign higher likelihood to positive samples suggests a systematic preference for well-formed outputs during generation. The principles in text modeling have been carried over to the benchmarking of pretrained SLMs, where sWUGGY (Dunbar et al., 2021), sBLIMP (Dunbar et al., 2021), and SALMon (Maimon et al., 2025) likewise adopt a comparative evaluation design. Each benchmark's contrastive pairs highlight various aspects of speech, from context to acoustics.

While effective, directly computing global token-level perplexity may overlook characteristics unique to speech, risking misalignment with human perception. From a cognitive perspective, textual and acoustic generation processes demand different attentional spans. The *Dependency Locality Theory* (Gibson, 1998, 2000) and *Surprisal Theory* (Levy, 2008; Smith and Levy, 2013) suggest that generating and evaluating coherent text relies on tracking dependencies over long contexts, thus requiring *long-range* attentiveness. In contrast, non-semantic acoustic features evolve continuously over time and are regularized by the *Gradualness of Change* (Bregman, 1993), which favors *short-span* conditioning. Likelihood patterns in Figure 1 and 2 echo these theoretical insights. The perplexity difference between positive and negative samples, when present, is short-spanned and concentrated near the onset of divergence.

In this work, we introduce novel evaluation protocols of SLMs that address fundamental asymmetries between speech and text modeling and place greater emphasis on local context sensitivity. Specifically, we propose two families of evaluation methods and re-evaluate SLMs on SALMon. The first family, **likelihood-based methods**, reformulates perplexity through localization and normalization, preserving the likelihood-based evaluation paradigm while emphasizing local context sensitivity (Sec. 3.1, Sec. 5.1). The second family consists of **generation-based methods**, where we evaluate on actual generations produced by the SLM, since successful generation inherently demands sensitivity to local context (Sec. 3.2, Sec. 5.2). New scores from these evaluation protocols yield substantially different conclusions from those based on naive global token perplexity.

To assess perception-faithfulness of the proposed methods, we conduct human-subject rating experiments on SLM-generated samples and collect mean opinion scores (MOS), which serve as the gold-standard reference for SLM performance. Correlation analyses (Sec. 5.3) show that our proposed methods align more closely with human ratings (Pearson correlation: $0.62 \rightarrow 0.73$, Spearman correlation: $0.65 \rightarrow 0.74$), thereby establishing a new paradigm for evaluating spoken language models. Our evaluation reshapes the performance landscape of SLMs: when re-evaluated, the best-performing model closes 84.1% of the gap to the human topline on SALMon, surprisingly achieving a new SOTA.

2 Background

2.1 Evaluation of Pretrained Spoken Language Models

For pretrained SLMs, likelihood-based evaluation provides a principled way to probe speech consistency across multiple dimensions. Prior work has largely emphasized the content aspect, where sWUGGY and sBLIMP (Dunbar et al., 2021) assess semantic coherence and lexical well-formedness. With automatic speech recognition (ASR), text-level perplexity can also be computed (Lakhotia et al., 2021; Hassid et al., 2024; Wu et al., 2023). More recently, paralinguistic modeling capabilities of pretrained SLMs are receiving growing attention, and SALMon (Maimon et al., 2025) is established to measure the consistency of acoustic attributes. These include speaker identity, sentiment, background, and room conditions. In this paper, we focus on improving the evaluation of pretrained SLMs specifically on acoustic attributes.

2.2 Global Token Perplexity in Likelihood Modeling

Likelihood modeling offers a principled approach to evaluate sequence models. In conventional practice, likelihood modeling is assessed via *perplexity*, where higher perplexity indicates lower likelihood under the model. Given a sequence $\mathbf{s}$, perplexity is defined as the exponential of the negative log-likelihood loss (NLL): $\mathrm{PPL}(\mathbf{s}) = \exp(\mathrm{NLL}(\mathbf{s}))$. With this transformation, we will henceforth report perplexity in terms of NLL. In the context of SLMs, perplexity is computed over sequences of *discrete speech tokens* (Maimon et al., 2025; Sugiura et al., 2025; Chou et al., 2025). We refer to this formulation as *global token perplexity*. Formally, given a sequence $s = (x_1, \ldots, x_T)$, the $\mathrm{NLL}_{\mathrm{global}}$ formula yields

$$\mathrm{NLL}_{\mathrm{global}}(s) = \frac{1}{T} \sum_{t=1}^{T} -\log p(x_t \mid x_{<t}) \quad (1)$$

Modern discrete speech units often consist of multiple codebooks, as in hierarchical residual vector quantization (RVQ) architectures (Zeghidour et al., 2021; Wang et al., 2023a; Défossez et al., 2022, 2024) or are multi-channeled, as in joint speech-text modeling (Tseng et al., 2025). In such settings, we flatten the multi-channel likelihood outputs into a single serialized stream. This formulation treats all tokens across channels and time steps as equally informative, yielding a unified, holistic likelihood estimate for the entire speech signal.

With this definition, adopting perplexity to benchmarking is straightforward. In SALMon, each sample consists of a pair of sequences (s_p, s_n), corresponding to the positive (preferred) sequence and the negative (dispreferred) sequence. A perception-faithful model is expected to assign lower perplexity to s_p than to s_n. This comparison is equivalent in the NLL space as NLL is a strictly monotonic transformation of perplexity:

$$\mathrm{NLL}(s_p) < \mathrm{NLL}(s_n) \quad (2)$$

2.3 Likelihood Modeling Calibration

A fundamental principle of likelihood modeling is that a model's predicted probabilities should align with the frequencies of occurrence of the global distribution (Kadavath et al., 2022; Ulmer et al., 2022). When raw prediction scores deviate, calibration methods can be introduced to increase alignment. Certain calibration approaches, such as Platt scaling (Platt, 1999; Guo et al., 2017), are monotonic and rank preserving. In contrast, other methods aim to alter the model's selection, including option debiasing (Brown et al., 2020), option finetuning (Guo et al., 2017), and decoding-time interventions (Chuang et al., 2024). In this work, we extend calibration efforts to SLMs by proposing localization and normalization methods that better align with the characteristics of speech.

3 Proposed Method

Using global token perplexity to assess acoustic consistency can lead to a measurement fallacy, where the resulting scores do not correlate well with human assessments. Empirically, prior works have shown that perplexity in speech is often disproportionately influenced by semantic factors (Maimon and Adi, 2023; Sicherman and Adi, 2023; Polyak et al., 2021), limiting the expressiveness of acoustic features. These semantic contributions constitute a substantial part of global token perplexity, introducing noise into the measurement and causing instability during sample-wise comparisons. SALMon reduced semantic volatility by "enforcing the same speech context between comparisons," but strict attribute-wise independence is neither well defined nor practically attainable in speech. We therefore derive alternative perplexity variants that focus on modeling the target at-

tribute—the axis along which the positive and negative speech samples are contrasted.

3.1 Proposed Likelihood-Based Evaluation

We start from Equation 1 and derive windowed, localized and normalized variants of token perplexity that focus on better capturing the response of the SLM on acoustic discontinuity. The windowed perplexity variant replaces global aggregation over the full NLL array with a predefined short temporal window that slides over the speech sequence, producing a sequence of windowed perplexity values. Then, we take the maximum among these values to obtain the perplexity spike as a measure of local anomaly, allowing comparisons across samples to determine which sample exhibits a stronger local irregularity.

$$\mathrm{NLL}_{\text{windowed}}(\mathbf{s}) = \max_i \frac{1}{\delta} \sum_{t=i}^{i+\delta-1} \big(-\log p(x_t \mid x_{<t})\big) \tag{3}$$

where $1 \leq i \leq T - \delta + 1$ indexes all valid starting positions of a length-δ window in the token sequence. Since it is the temporal window that is fixed across all models, the corresponding token-based value of δ is derived accordingly and therefore differs across models; in some cases, it may even vary dynamically within a model.

In SALMon, each positive–negative pair shares a common prefix, which allows us to capitalize on the diverging timeframe as additional information. For each pair (s_p, s_n), we extract the longest common prefix as the prompt (S), and the remainder of the sequence forms the positive (P) and negative (N) responses, respectively, analogous to QA-style setups in NLP benchmarks (Brown et al., 2020; Hendrycks et al.). Concretely, we write $s_p = S \frown P$ and $s_n = S \frown N$, where $\frown$ denotes sequence concatenation. The localized variant of token perplexity only accounts for information in a localized window of length δ starting from the timeframe where the speech prompt ends (t_p):

$$\mathrm{NLL}_{\text{localized}}(\mathbf{s}) = \tfrac{1}{\delta} \textstyle\sum_{t=t_p}^{t_p+\delta-1} \big(-\log p(x_t \mid x_{<t})\big) \tag{4}$$

We also consider normalization, which adjusts each response probability by factoring out its prompt-free probability (Brown et al., 2020). Normalization applies to both global and localized perplexity by choosing the corresponding window $\Delta \in \{T, \delta\}$, where T represents a value sufficiently large that it is ultimately bounded by the sequence length. The normalized perplexity aggregates normalized probabilities at each time step.

$$\mathrm{NLL}_{\text{normalized}}(\mathbf{s}) = \tfrac{1}{\Delta} \textstyle\sum_{t=t_p}^{t_p+\Delta-1} \big(-\log \frac{p(x_t|x_{<t})}{p(x_t|x_{t_p:<t})}\big) \tag{5}$$

3.2 Proposed Generation-Based Evaluation

The evaluation methods in Sections 2.2 and 3.1 compute the likelihood of the samples rather than letting the SLM continue the speech. Here, we directly evaluate on continuations of SLMs given a speech prompt S, which provide multiple benefits. With real continuations, it is possible to conduct human evaluations to obtain mean opinion scores (MOS), which provide a perception-faithful estimate of model quality and can serve as the reference for the model's *true continuation performance*. Second, by approximating human judgements with scores from model-as-a-judge, we obtain another evaluator candidate to compete with global token perplexity.

Given a speech prompt S, we define a continuation G sampled from model M by

$$G = M(\cdot \mid S). \tag{6}$$

For human evaluations, annotators assign a quality score between 1 and 5 based on how good the generated continuation is relative to the positive continuation reference (P).

For scoring continuations with model-as-a-judge, success can be determined more straightforwardly using a contrastive criterion: a continuation is deemed correct if it is closer to the gold positive continuation than to the negative one.

$$d(G, P) < d(G, N) \tag{7}$$

where $d(\cdot)$ denotes a distance function. The following section describes the procedure for selecting a qualified model to serve as an automatic judge, as well as the corresponding evaluation strategy to assess SLM-generated continuations (Sec. 3.3).

3.3 Model-as-a-Judge for Scoring SLM Continuations

To select an appropriate judge J, we require (i) a labeled set with known correctness and (ii) a model that assigns a distance score to a prompt–response pair. Fortunately, the shared prompt S and paired responses (P, N) in each contrastive

example (s_p, s_n) provides a natural labeled set with the following objective:

$$d(S, P) < d(S, N). \tag{8}$$

We explore using embedding models E as judge candidates, leveraging their inherent distance metric, taking inspiration from retrieval systems (Feng et al., 2022):

$$\mathrm{d}(A, B) = 1 - \cos\big(E(A), E(B)\big) \tag{9}$$

where $\cos(\cdot, \cdot)$ denotes cosine similarity and $E(\cdot)$ denotes the forward pass through the embedding model. Plugging Equation 9 into Equation 7 yields the explicit sample-wise objective for choosing E as a qualified judge.

$$\cos\big(E(S), E(P)\big) > \cos\big(E(S), E(N)\big) \tag{10}$$

The aggregation of correct predictions over the development set yields the accuracy. An ideal judge would achieve a perfect score; in practice, a more realistic qualification threshold is the human accuracy on the same benchmark. We begin with a comprehensive set of pretrained embedding models and narrow it to the best-performing model (over the qualification threshold), which we adopt as the final judge (J). The judge benchmarks continuations G from SLM following Equation 7:

$$\cos\big(J(G), J(P)\big) > \cos\big(J(G), J(N)\big), \tag{11}$$

where aggregation over the whole benchmark yields the accuracy score of the evaluated SLM.

4 Experimental Setup

We adopt SALMon for all evaluations. SALMon includes 6 subsets that measure acoustic consistency in gender, speaker identity, sentiment, two background conditions, and room attributes. Each data point consists of a positive and a negative sample, where the negative sample contains an inconsistency in one of the attributes. Our evaluations of SLMs cover GSLM (Lakhotia et al., 2021), TWIST (Hassid et al., 2024), pGSLM (Kharitonov et al., 2021), Spirit-LM (Nguyen et al., 2024), TASTE (Tseng et al., 2025), Flow-SLM (Chou et al., 2025), and Llama-Mimi (Sugiura et al., 2025). We measure their performance under original and proposed methods. The localized window δ is set as 0.5s.

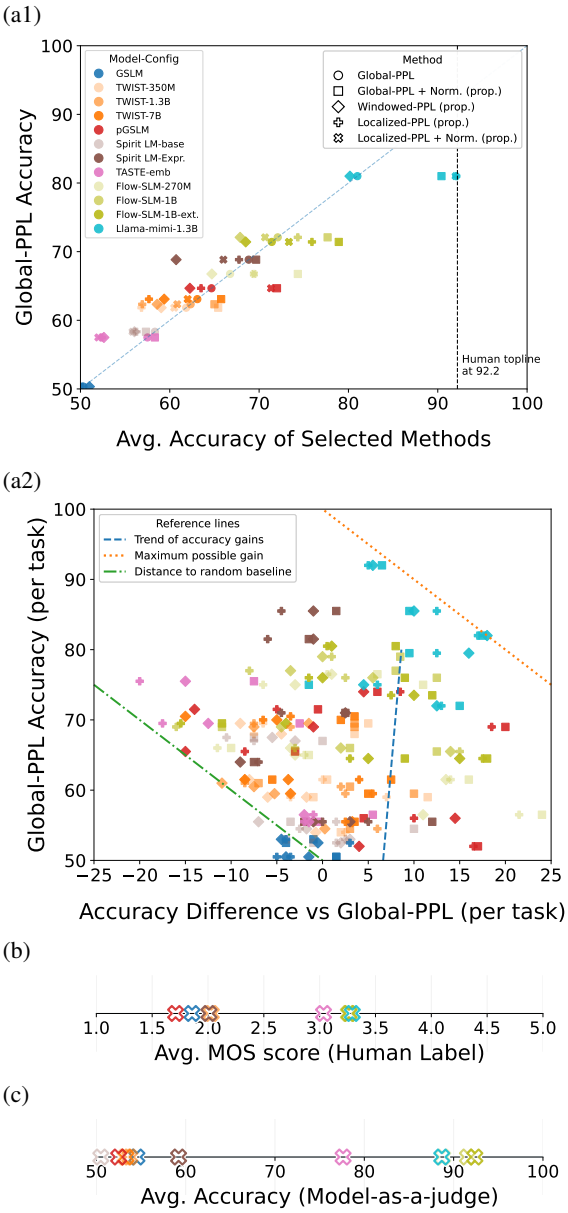

Figure 3: **Overall performance of spoken language models on consistency tasks.** The x-axis reports model accuracy under different evaluators: (a1) alternative likelihood estimators, (b) MOS, and (c) embedding-as-a-judge, where model color codes are shown in (a1) and shared among all plots. In (a1), we correlate scores from proposed methods against those from global token perplexity (`Global-PPL`); the horizontal spread highlights the discrepancy across evaluation methods. The alternative methods rate strong models more favorably than `Global-PPL`, substantially closing the gap to the human topline. In (a2), we correlate deviations from the proposed methods against `Global-PPL` scores. Deviations generally become larger at higher `Global-PPL` performance (blue), until it saturates due to the maximum performance ceiling (orange). Negative deviations exhibit a similar trend in absolute magnitude, though this is less surprising since they are soft-bounded by distance to random baseline (green).

Table 1: Generation performance of SLMs judged by human ratings (MOS scores) with the model's associated rank.

	MOS Scores Evaluation							
	Sentiment ↑	Speaker ↑	Gender ↑	Bg (domain) ↑	Bg (rand.) ↑	Room ↑	Avg ↑	Rank
GSLM	1.88 ± 1.06	1.94 ± 1.13	2.76 ± 1.61	1.38 ± 0.85	1.36 ± 0.86	1.82 ± 1.00	1.86 ± 0.45	7
TWIST-1.3B	1.91 ± 1.08	2.04 ± 1.18	2.73 ± 1.55	1.62 ± 1.07	1.59 ± 0.99	2.29 ± 1.21	2.03 ± 0.49	5
pGSLM	1.86 ± 1.05	1.76 ± 1.05	2.38 ± 1.28	1.34 ± 0.78	1.28 ± 0.71	1.65 ± 0.97	1.71 ± 0.40	8
Spirit-LM-Expr.	3.41 ± 1.49	1.98 ± 1.14	2.63 ± 1.49	1.27 ± 0.73	1.19 ± 0.51	1.58 ± 0.98	2.01 ± 0.46	6
TASTE-emb.	3.68 ± 1.40	4.37 ± 1.02	4.63 ± 0.93	1.64 ± 1.16	1.60 ± 1.04	2.29 ± 1.27	3.03 ± 0.47	4
Flow-SLM-1B	3.86 ± 1.27	4.21 ± 1.13	4.47 ± 0.96	1.89 ± 1.08	1.86 ± 1.12	3.25 ± 1.34	3.26 ± 0.47	3
Flow-SLM-1B-Ext.	3.80 ± 1.31	4.20 ± 1.10	4.52 ± 0.94	1.98 ± 1.13	2.00 ± 1.23	3.08 ± 1.41	3.26 ± 0.49	2
Llama-Mimi-1.3B	3.78 ± 1.31	4.14 ± 1.16	4.32 ± 1.10	2.20 ± 1.30	2.21 ± 1.29	3.11 ± 1.41	3.29 ± 0.52	1

Table 2: The best-performing embedding model E on each task provides a viable judge J for evaluating continuation performance. In addition to surpassing the human baseline, four of the six models achieve near perfect performance.

	Sentiment ↑	Speaker ↑	Gender ↑	Bg (domain) ↑	Bg (rand.) ↑	Room ↑
Selected embedding model	TITANET	TITANET	TITANET	HuBERT-large-audioset	HuBERT-large-audioset	wav2vec2-large-audioset
Classifier performance	**99.5**	**100.0**	**100.0**	**86.5**	**97.5**	**100.0**
Human performance	97.2	91.5	98.6	83.1	88.7	94.4

We use the Prolific service to obtain MOS scores. We evaluate 50 samples over 8 models, which yields 400 generations to be evaluated. Each generation is independently assessed by five annotators on a five-point Likert scale. The annotators are proficient English speakers, and they are fairly compensated for their time. See Appendix B.1 for the annotation guidelines provided to annotators.

For model-as-a-judge, we consider a diverse pool of embedding models trained with different objectives and datasets, including TITANET (Koluguri et al., 2022), CAM++ (Wang et al., 2023b), CLAP (Elizalde et al., 2023), and AudioSet-trained models (La Quatra et al., 2024). We use SALMon prompts S as the dev set to select the best judge for each subset, and use the judge to obtain SLM performance scores from continuations G.

5 Experimental Results

5.1 Localized and Normalized Likelihood-Based Evaluation

We first examine how the proposed likelihood-based estimators reshape the performance landscape of spoken language models. Figure 3 (a1) shows the average accuracy over all consistency benchmarks for each SLM–method combination, plotted against the corresponding accuracy measured by conventional global token perplexity. From the plot, it can be seen that the horizontal spread is quite large, indicating systematic disagreement between perplexity methods. The degree of disagreement is quantified in Figure 3 (a2), which reveals a positive association between disagreement and model competence (measured by `Global-PPL` performance). The linearly regressed positive deviation (blue line) starts at $+6.63\%$ at around 50% accuracy, and increases to $+8.62\%$ at 80% accuracy, which is substantial.

A closer inspection shows that the shift is predominantly one-sided for each SLM configuration, reflecting a stable bias toward either over- or underestimation. These patterns are tightly linked to the underlying token type. Our proposed methods consistently assign lower scores to HuBERT-based SLMs (GSLM, TWIST, SpiritLM), while in contrast yielding higher scores for Mimi-based models (Flow-SLM, Llama-Mimi). pGSLM is the lone outlier among HuBERT-based models, likely due to its distinctive auxiliary training objective. In contrast, model families trained with the same recipe but varying only in scale exhibit little behavioral difference. We report the full set of scores in Appendix C.1.

5.2 Generation-Based Evaluation

We conduct experiments on actual model continuations to gain a better understanding of the generative abilities of SLMs.

5.2.1 MOS Evaluations by Human Labelers

MOS evaluation results are presented in Figure 3 (b). Llama-Mimi obtains the top score of 3.29, followed by Flow-SLM, while models using HuBERT tokens (GSLM, TWIST, Spirit-LM, pGSLM) perform much worse. Full model-by-task results are reported in Table 1, which presents the quantitative

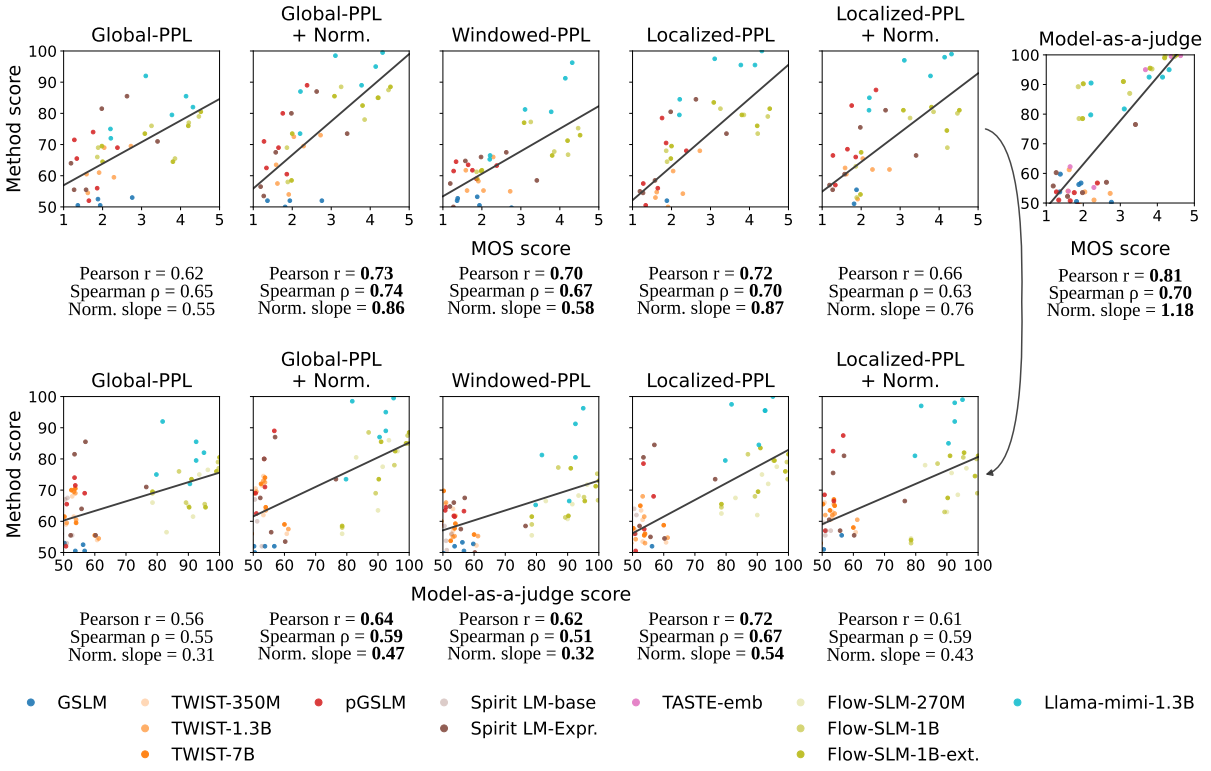

Figure 4: Correlation between perplexity evaluation methods vs golden labels provided by either MOS scores (top), or model-as-a-judge proxies (bottom) on the SALMon benchmark. Compared with using global perplexity as the aggregator (leftmost), windowing, normalization, and localization are more effective alternative operators and show stronger correlation with MOS scores. Using model-as-a-judge on generations (top right) also exhibits higher correlation with the MOS scores.

MOS scores across models and tasks. The results suggest that stronger models improve primarily on speech-centric attributes (e.g., sentiment, speaker identity, and gender), while substantial room for improvement remains in modeling ambience-related information.

5.2.2 Model-as-a-Judge for Measuring Generation Consistency

Identification of suitable judge models. Table 2 shows the top performing embedding model on the SALMon dev set along with its accuracy scores, following Equation 10. Results reveal that the selected embedding models consistently outperform human performance, and even reach ceiling-level performance ($> 99\%$) in four out of six cases. Collectively, these results support the credibility of model-as-a-judge for the SALMon task using a combination of TITANET, HuBERT-large-audioset, and wav2vec2-large-audioset. Detailed model-by-task results are provided in Appendix C.2.

Generation consistency evaluated by qualified judge models. Figure 3 (c) shows the result of evaluating model generations on SALMon using the selected judge models. Consistent with findings in likelihood estimators and MOS scores, the speech tokenizer is the most dominant factor for the performance difference. Most HuBERT-based models struggle to retain speech properties during continuation, obtaining performance close to random choice. On the other end of the spectrum, Flow-SLM and Llama-mimi exhibit strong performance, with scores in the vicinity of the human topline. Finally, TASTE generations perform relatively well, highlighting the importance of adopting a speaker vector during token-to-speech conversion. In Appendix C.2, we present verbose task-wise results, which further suggest that continuation failures principally arise from inadequate information being preserved during the speech encoding phase.

5.3 Correlation between methods

We now have four likelihood-based evaluators and a generation-based evaluator scored with an embedding judge model. To determine the metrics that are more faithful to human perception for this task, we correlate them to the "true continuation

Table 3: Kendall τ correlation between evaluator scores and human MOS scores across the 6 acoustic-consistency splits on SALMon, based on the rankings of the 7 models present in the top row of Figure 4.

Method	Sentiment	Speaker	Gender	Bg (domain)	Bg (rand.)	Room	Avg
Global-PPL	0.524	0.333	0.143	0.619	0.333	**0.714**	0.444
+ Normalization	0.524	0.451	0.048	0.524	0.619	**0.714**	0.480
Windowed-PPL	0.524	0.429	0.238	0.619	0.586	0.619	0.502
Localized-PPL	0.488	0.048	0.195	0.714	0.619	0.333	0.400
+ Normalization	0.429	-0.048	-0.098	0.238	0.619	0.524	0.277
Model-as-a-judge	**0.683**	**0.878**	**0.524**	**0.878**	**0.651**	0.429	**0.674**

performance," provided by the MOS scores. The top row of Figure 4 is a comprehensive display of these correlations. Global token perplexity sets the baseline with Pearson score of 0.62 and Spearman of 0.65. Localized perplexity achieves higher scores of Pearson of 0.72 and Spearman of 0.70, which are further surpassed by normalized perplexity, where both Pearson and Spearman correlations improve to 0.73 and 0.74 respectively. Continuations scored by embedding judges obtain the highest Pearson score overall at 0.81, with a slightly lower Spearman score of 0.70. The high Pearson correlation is best accounted for by the tighter dispersion of points in the high-performance regime. In addition, the regression slopes for normalized and localized perplexity are substantially closer to 1[2], indicating that they not only improve relative scoring, but also produce more accurate absolute scores by better matching the target scale. These results indicate that, on acoustic consistency benchmarks such as SALMon, evaluating SLMs with our proposed methods yields judgments that are better aligned with human perception than those obtained using global token perplexity. In the bottom row of Figure 4, we replicate the analysis using the embedding model-as-a-judge scores as a proxy for the MOS scores given its high correlation, to conduct correlations with likelihood-methods on all models. The results reinforce the conclusion that windowed, normalized, and localized methods correlate better to true generation performance.

Next, we examine how well evaluator rankings correlate with MOS scores for each individual split. In Table 3, we report the Kendall τ scores on the ranking of the 7 models that are universally present in the top row of Figure 4. The results show that normalized and windowed perplexity, as well as model-as-a-judge scores, yield substantially stronger rank agreement with MOS across acoustic-consistency tasks, leading to higher average Kendall's τ overall. In particular, the overall Kendall's τ increases from 0.444 to 0.480 with normalized global perplexity, to 0.502 with windowed perplexity, and to 0.674 with model-as-a-judge evaluation.

Collectively, these results establish a principled and scalable evaluation framework for continuation quality. Among likelihood-based methods, normalized perplexity is preferred over global token perplexity because it shows stronger correlation with MOS and better preserves model rankings. With normalized perplexity as the evaluation metric, the performance landscape of SLMs is completely reshaped. Most notably, the best-performing model Llama-Mimi improves from 80.92 to 90.42, closing 84.1% of the gap to the human topline on SALMon and achieving a new state of the art. While generation-based model-as-a-judge evaluation is not universally superior to normalized perplexity in terms of correlation with the MOS scores, it remains valuable for assessing SLMs in cases where token generation does not fully capture audio generation quality, as in TASTE. Despite these advancements, human inspection of the generated samples reveals that these SLMs still have substantial room for improvement in handling complex speech signals, underscoring the need for more rigorous benchmarks.

5.4 A Tale of Two Models: How Loss Composition Shapes Performance

Building on our earlier discussion that acoustic quality can be decomposed into interpretable axes (e.g., speaker- and background-related attributes), we expect the model's NLL loss to be governed by an analogous set of separable components. Breaking the loss down by axis would quantify each attribute's contribution and its effect on performance. Nonetheless, an explicit axis-wise decomposition of the loss remains difficult in practice, because

[2]The optimal slope is obtained by linearly anchoring the random-performance baseline, with accuracy = 50, to MOS = 1, and the perfect-performance ceiling, with accuracy = 100, to MOS = 5.

Table 4: Shapley value decompositions for Spirit-LM-Expressive and Llama-Mimi over token types (HuBERT Φ_H, pitch Φ_P, style Φ_S) and layer groups (Φ_0–Φ_3). The Shapley values of the primary tokens for the two models (Φ_H and Φ_0) shift in opposite directions under localization and normalization.

Model	Window	Term	Original	Norm.
Spirit-LM Expr.	Global	ϕ_H	+9.6	+7.6
		ϕ_P	+9.5	+13.1
		ϕ_S	-0.3	-1.0
	Localized ($t = 0.5$s)	ϕ_H	+9.6	+5.3
		ϕ_P	+9.4	+13.0
		ϕ_S	-1.2	-2.3
Llama-Mimi	Global	ϕ_0	+4.4	+5.2
		ϕ_1	+6.2	+9.0
		ϕ_2	+12.2	+15.8
		ϕ_3	+8.1	+10.5
	Localized ($t = 0.5$s)	ϕ_0	+7.6	+7.8
		ϕ_1	+8.9	+7.6
		ϕ_2	+14.5	+16.4
		ϕ_3	+11.1	+10.1

the relevant factors are entangled within a single embedding or even within individual tokens. Fortunately, we can consider Spirit-LM-Expressive and Llama-Mimi as analytical lenses for this study, since their token inventories are inherently functionally distinct. Spirit-LM-Expressive comprises three token types during sequence construction: HuBERT, pitch, and style; whereas Llama-Mimi employs tokens from different RVQ layers, which similarly exhibit functional heterogeneity. Their original works (Sugiura et al., 2025; Nguyen et al., 2024) already categorized these token types as reflecting different balances of semantic versus acoustic utility. Using these models as representative cases, we contrast their token-type contribution profiles to account for the opposite effects of our alternative methods in Llama-Mimi and Spirit-LM-Expressive. We perform a comprehensive combinatorial ablation across token types, enabling Shapley value analysis (Shapley et al., 1953) to quantify each token type's marginal contribution.

Table 4 exhibits the Shapley contributions of each token type. For Llama-Mimi, tokens from different residual layers all contribute positively, with the largest contribution coming from residual layer 1 (ϕ_1). Contributions of individual layers are further amplified in normalized and localized settings. On average, normalized settings improved +2.4 points, and localized settings improved +2.8 points spread evenly across token types. This improvement clearly carries over to the final accuracy score, where the normalized method achieves a gain of +9.42 points and the localized method achieves a gain of +11.08 points.

The case of Spirit-LM-Expressive is markedly different from Llama-Mimi. In Table 4, HuBERT tokens contribute to acoustic tasks more than the other tokens combined, despite being labeled as a "semantic" token type. Even with localization, the HuBERT Shapley value remains unchanged, suggesting that a non-trivial portion of the acoustic inconsistencies captured by the SLM is distributed throughout the full sequence. Furthermore, normalization leads to an additional decrease in the HuBERT tokens' Shapley values, indicating a strong entanglement between semantic and acoustic information in Spirit-LM-Expressive's HuBERT tokens. As a result, reducing semantic influence also diminishes acoustic information. As reflected in the accuracy scores, these alternative methods do not improve performance on the SALMon benchmark. The increase in Shapley values for pitch tokens is offset by the decrease in Shapley values for HuBERT tokens, which remain equally crucial for capturing acoustic information.

As we uncover the role of HuBERT tokens in Spirit-LM-Expressive, it becomes clearer why models that adopt HuBERT tokens (GSLM, TWIST, pGSLM, Spirit-LM) fail to translate high classification scores into strong continuation performance as measured by MOS. Strong global-perplexity performance in HuBERT-based modeling appears to depend substantially on long-range dependencies to judge correctness, whereas continuation quality must be established locally and cannot rely on future context. Our proposed evaluation penalizes this behavior, resulting in better agreement with MOS-based continuation outcomes.

6 Conclusion

In this work, we revisit the use of global token perplexity for evaluating SLMs and highlight normalization, localization, and generation-based evaluation techniques as alternative methods that better capture key characteristics of speech. Correlations with MOS indicate that our proposed methods better reflect human perception. Under this re-evaluation, the previously best-performing model closes 84.1% of the gap to the human topline on SALMon. Together, these findings reshape the SLM performance landscape and establish a new evaluation paradigm for future studies.

Limitations

We propose novel evaluation methods as alternatives to conventional global token perplexity. However, since these methods are still applied to existing benchmarks, their scope remains inherently constrained by the limitations of those benchmarks. For instance, SALMon does not systematically probe compounded variations (e.g., speaker changes under noisy background conditions), which restricts our ability to characterize SLM performance in such settings—even when using our improved evaluators. In addition, we focused our discussion on acoustic continuity, which constitutes a substantial and distinctive aspect of speech. For other dimensions, such as semantics, global perplexity may still be, and very likely remains, the most appropriate approach. Nonetheless, the broader notion of "speech perplexity" warrants careful scrutiny, as different aspects of speech are inherently entangled.

References

Keyu An, Qian Chen, Chong Deng, Zhihao Du, Changfeng Gao, Zhifu Gao, Yue Gu, Ting He, Hangrui Hu, Kai Hu, and 1 others. 2024. Funaudiollm: Voice understanding and generation foundation models for natural interaction between humans and llms. *arXiv preprint arXiv:2407.04051*.

Siddhant Arora, Kai-Wei Chang, Chung-Ming Chien, Yifan Peng, Haibin Wu, Yossi Adi, Emmanuel Dupoux, Hung yi Lee, Karen Livescu, and Shinji Watanabe. 2025. On the landscape of spoken language models: A comprehensive survey. *Transactions on Machine Learning Research*.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. 2022. data2vec: A general framework for self-supervised learning in speech, vision and language. *Preprint*, arXiv:2202.03555.

Albert S. Bregman. 1993. 2 auditory scene analysis: hearing in complex environments. In *Thinking in Sound: The Cognitive Psychology of Human Audition*. Oxford University Press.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

Kai-Wei Chang, Haibin Wu, Yu-Kai Wang, Yuan-Kuei Wu, Hua Shen, Wei-Cheng Tseng, Iu-thing Kang, Shang-Wen Li, and Hung-yi Lee. 2024. SpeechPrompt: Prompting speech language models for speech processing tasks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T. Tan, and Haizhou Li. 2024. VoiceBench: Benchmarking LLM-based voice assistants. *Preprint*, arXiv:2410.17196.

Ju-Chieh Chou, Jiawei Zhou, and Karen Livescu. 2025. Flow-slm: Joint learning of linguistic and acoustic information for spoken language modeling. *arXiv preprint arXiv:2508.09350*.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report. *Preprint*, arXiv:2407.10759.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. 2024. DoLa: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations*.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2022. High fidelity neural audio compression. *Transactions on Machine Learning Research*.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: a speech-text foundation model for real-time dialogue. *Preprint*, arXiv:2410.00037.

Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, Zhifu Gao, and Zhijie Yan. 2024. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *Preprint*, arXiv:2407.05407.

Ewan Dunbar, Mathieu Bernard, Nicolas Hamilakis, Tu Anh Nguyen, Maureen de Seyssel, Patricia Rozé, Morgane Rivière, Eugene Kharitonov, and Emmanuel Dupoux. 2021. The zero resource speech challenge 2021: Spoken language modelling. In *Proc. Interspeech*, pages 1574–1578.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. CLAP: Learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891.

Kuofeng Gao, Shu-Tao Xia, Ke Xu, Philip Torr, and Jindong Gu. 2025. Benchmarking open-ended audio dialogue understanding for large audio-language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4763–4784.

Edward Gibson. 1998. Linguistic complexity: locality of syntactic dependencies. *Cognition*, 68(1):1–76.

Edward Gibson. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. In Alec P. Marantz, Yasushi Miyashita, and Wayne O'Neil, editors, *Image, Language, Brain: Papers from the First Mind Articulation Project Symposium*, pages 95–126. MIT Press, Cambridge, MA.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Michael Hassid and 1 others. 2024. Textually pretrained speech language models. *Advances in Neural Information Processing Systems*, 36.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Yixuan Hou, Heyang Liu, Yuhao Wang, Ziyang Cheng, Ronghua Wu, Qunshan Gu, Yanfeng Wang, and Yu Wang. 2025. SOVA-Bench: Benchmarking the speech conversation ability for LLM-based voice assistant. In *Proc. Interspeech*, pages 5713–5717.

Chan-Jan Hsu, Yi-Cheng Lin, Chia-Chun Lin, Wei-Chih Chen, Ho Lam Chung, Chen-An Li, Yi-Chang Chen, Chien-Yu Yu, Ming-Ji Lee, Chien-Cheng Chen, Ru-Heng Huang, Hung yi Lee, and Da-Shan Shiu. 2025. Breezyvoice: Adapting TTS for taiwanese mandarin with enhanced polyphone disambiguation – challenges and insights. *Preprint*, arXiv:2501.17790.

Wei-Ning Hsu, Anuroop Sriram, Alexei Baevski, Tatiana Likhomanenko, Qiantong Xu, Vineel Pratap, Jacob Kahn, Ann Lee, Ronan Collobert, Gabriel Synnaeve, and Michael Auli. 2021. Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training. *Preprint*, arXiv:2104.01027.

Chien-yu Huang, Ke-Han Lu, Shih-Heng Wang, Chi-Yuan Hsiao, Chun-Yi Kuan, Haibin Wu, Siddhant Arora, Kai-Wei Chang, Jiatong Shi, Yifan Peng, and 1 others. 2024. Dynamic-superb: Towards a dynamic, collaborative, and comprehensive instruction-tuning benchmark for speech. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12136–12140. IEEE.

Chien-yu Huang and 1 others. 2025. Dynamic-SUPERB phase-2: A collaboratively expanding benchmark for measuring the capabilities of spoken language models with 180 tasks. In *The Thirteenth International Conference on Learning Representations*.

Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Yanqing Liu, Yichong Leng, Kaitao Song, Siliang Tang, and 1 others. 2024. Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. *CoRR*.

Saurav Kadavath, Tom Conerly, Amanda Askell, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Eugene Kharitonov, Ann Lee, Adam Polyak, Yossi Adi, Jade Copet, Kushal Lakhotia, Tu Anh Nguyen, Morgane Riviere, Abdelrahman Mohamed, Emmanuel Dupoux, and Wei-Ning Hsu. 2022. Text-free prosody-aware generative spoken language modeling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8666–8681, Dublin, Ireland. Association for Computational Linguistics.

Eugene Kharitonov and 1 others. 2021. Text-free prosody-aware generative spoken language modeling. *arXiv preprint arXiv:2109.03264*.

Nithin Rao Koluguri, Taejin Park, and Boris Ginsburg. 2022. Titanet: Neural model for speaker representation with 1d depth-wise separable convolutions and global context. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 8102–8106. IEEE.

Moreno La Quatra, Alkis Koudounas, Lorenzo Vaiani, Elena Baralis, Luca Cagliero, Paolo Garza, and Sabato Marco Siniscalchi. 2024. Benchmarking representations for speech, music, and acoustic events. In *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 505–509.

Kushal Lakhotia and 1 others. 2021. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Gallil Maimon and Yossi Adi. 2023. Speaking style conversion in the waveform domain using discrete

self-supervised units. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Gallil Maimon, Amit Roth, and Yossi Adi. 2025. SALMon: A suite for acoustic language model evaluation. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. Voxceleb: A large-scale speaker identification dataset. In *Proc. Interspeech*, pages 2616–2620.

Tu Anh Nguyen and 1 others. 2024. Spirit-lm: Interleaved spoken and written language model. *arXiv preprint arXiv:2402.05755*.

Open-AI. Gpt-4o. https://openai.com/index/gpt-4o-system-card/. Accessed: Jul 2024.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

John C. Platt. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74. MIT Press.

Adam Polyak and 1 others. 2021. Speech resynthesis from discrete disentangled self-supervised representations. *arXiv preprint arXiv:2104.00355*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Lloyd S Shapley and 1 others. 1953. A value for n-person games.

Amitay Sicherman and Yossi Adi. 2023. Analysing discrete self supervised speech representation for spoken language modeling. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Nathaniel J Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Issa Sugiura, Shuhei Kurita, Yusuke Oda, and Ryuichiro Higashinaka. 2025. Llama-mimi: Speech language models with interleaved semantic and acoustic tokens. *arXiv preprint arXiv:2509.14882*.

Hsiang-Sheng Tsai, Heng-Jui Chang, Wen-Chin Huang, Zili Huang, Kushal Lakhotia, Shu-wen Yang, Shuyan Dong, Andy Liu, Cheng-I Lai, Jiatong Shi, Xuankai Chang, Phil Hall, Hsuan-Jui Chen, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-yi Lee. 2022. SUPERB-SG: Enhanced speech processing universal PERformance benchmark for semantic and generative capabilities. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8479–8492, Dublin, Ireland. Association for Computational Linguistics.

Liang-Hsuan Tseng, Yi-Chang Chen, Kuan-Yi Lee, Da-Shan Shiu, and Hung-yi Lee. 2025. Taste: Text-aligned speech tokenization and embedding for spoken language modeling. *arXiv preprint arXiv:2504.07053*.

Yuan Tseng, Layne Berry, Yi-Ting Chen, I-Hsiang Chiu, Hsuan-Hao Lin, Max Liu, Puyuan Peng, Yi-Jen Shih, Hung-Yu Wang, Haibin Wu, Po-Yao Huang, Chun-Mao Lai, Shang-Wen Li, David Harwath, Yu Tsao, Shinji Watanabe, Abdelrahman Mohamed, Chi-Luen Feng, and Hung-yi Lee. 2024. AV-SUPERB: A multi-task evaluation benchmark for audio-visual representation models. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Also available as arXiv:2309.10787.

Dennis Ulmer, Jes Frellsen, and Christian Hardmeier. 2022. Exploring predictive uncertainty and calibration in NLP: A study on the impact of method & data scarcity. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2707–2735, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, and 1 others. 2023a. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*.

Hui Wang, Siqi Zheng, Yafeng Chen, Luyao Cheng, and Qian Chen. 2023b. Cam++: A fast and efficient network for speaker verification using context-aware masking. In *INTERSPEECH*.

Pete Warden. 2018. Speech commands: A dataset for limited-vocabulary speech recognition. *Preprint*, arXiv:1804.03209.

Haibin Wu, Kai-Wei Chang, Yuan-Kuei Wu, and Hung-yi Lee. 2023. SpeechGen: Unlocking the generative power of speech language models with prompts. *arXiv preprint arXiv:2306.02207*.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and Jingren Zhou. 2024. AIR-bench: Benchmarking large audio-language models via generative comprehension. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1979–1998, Bangkok, Thailand. Association for Computational Linguistics.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Kotik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-yi Lee. 2021. SUPERB: Speech processing universal PERformance benchmark. In *Proc. Interspeech*, pages 1194–1198.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2021. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507.

A Further Evaluations of Spoken Language Models

While *universal task robustness* serves as the prevailing objective, evaluation paradigms often progress in a way that mirrors the model's training stages. In the main text, we have shown that for pretrained SLMs, likelihood-based evaluation is prominent. For finetuned models, task-based evaluation metrics pinpoint utility, such as speech recognition (Panayotov et al., 2015), emotion recognition (Busso et al., 2008), keyword spotting (Warden, 2018), and speaker identification (Nagrani et al., 2017). Aggregated suites such as the SUPERB series (Yang et al., 2021; Tseng et al., 2024; Tsai et al., 2022; Huang et al., 2024, 2025) broaden coverage, though their constituent tasks remain predefined. Prompt-following benchmarks built around arbitrary natural-language instructions highlight general intelligence of finetuned SLMs (Yang et al., 2024; Gao et al., 2025; Hou et al., 2025; Chen et al., 2024).

B Verbose Evaluation Settings

B.1 MOS Evaluation Prompt

The annotators are given the following prompt:

> You will hear a generated continuation that extends from an audio prompt. Compare the two audio clips only on the specific target **feature** that varies between samples, focus only on the similarity of that feature, and assign a similarity score from 1 to 5.
>
> A score of 5 indicates perfect match on the feature (gender/speaker/sentiment/background, etc.). The two audios are indistinguishable on that feature. Naturally, two identical audios will score 5 on any given feature.
>
> A score of 1 indicates complete mismatch on the feature (gender/speaker/sentiment/background, etc.). The two audios are easily distinguishable on that feature. Naturally, if the target feature is missing entirely, the score is unequivocally 1.
>
> Target feature is considered missing when:
>
> The audio is completely silent. The attribute is gender / speaker / sentiment but there is no speech (sound made by humans). The attribute is background but there is only human speech and no other acoustic source. Use the guideline below for sample comparisons and guidance for scores 2–4.
>
> Scoring guidelines:
>
> 5 – Indistinguishable from the prompt on the target attribute.
>
> 4 – Only distinguishable with close attention to small sections; casual listening still feels nearly identical.
>
> 3 – Distinguishable, but most attribute traits still feel similar.
>
> 2 – Clearly distinguishable with only minor attribute overlap.
>
> 1 – Totally distinguishable; no attribute similarity at all.
>
> (Examples for each score)

C Verbose Evaluation Results

C.1 Likelihood-Based Evaluations

Table 5 shows full evaluation results on SALMon across spoken language models. Generally, alternative likelihood-based evaluation methods assign lower scores to HuBERT-based models (GSLM, TWIST, SpiritLM) and higher scores to Mimi-based models (Flow-SLM, Llama-Mimi). For the top-performing models, our proposed evaluation methods occasionally deliver results that place model performance above the human topline. Notably, these cases are concentrated around speaker information.

For semantic-acoustic alignment tasks, there is not a common speech prompt, hence only Global-PPL and Windowed-PPL evaluation are supported. Results show consistent low accuracies ($< 60\%$), regardless of the model used or the methods selected. This agreement confirms that alignment is likely a trait that is not picked up by current SLMs, motivating future work on this direction.

Table 5: Comparison of spoken language model performance on the SALMon benchmark, including GSLM (Kharitonov et al., 2022), TWIST (Hassid et al., 2024), pGSLM (Kharitonov et al., 2021), Spirit-LM (Nguyen et al., 2024), TASTE (Tseng et al., 2025), Flow-SLM (Chou et al., 2025), and Llama-mimi (Sugiura et al., 2025). For each model, we report task accuracies measured with global token perplexity and our proposed likelihood evaluators. Comparing accuracies within each model shows that the perceived performance difference between methods is quite huge. As a result, these evaluation protocols yield substantially different conclusions on the SLM performance landscape. Bold number highlight evaluation results that surpass the human topline performance by (Maimon et al., 2025).

			Acoustic Consistency						**Semantic-Acoustic Alignment**	
Model	**Config**	**Method**	Sentiment ↑	Speaker ↑	Gender ↑	Bg (domain) ↑	Bg (rand.) ↑	Room ↑	Sentiment ↑	Background ↑
GSLM	–	Global-PPL	52.5	50.5	53.0	47.5	50.5	48.0	55.0	52.5
		+ Normalization (prop.)	48.5	52.0	52.0	45.0	52.0	50.0	-	-
		Windowed-PPL (prop.)	53.2	50.5	49.8	52.8	52.0	47.8	56.0	43.0
		Localized-PPL (prop.)	52.0	49.0	48.5	47.5	46.5	42.5	-	-
		+ Normalization (prop.)	55.5	47.0	49.0	46.5	45.5	51.0	-	-
TWIST	350M	Global-PPL	59.0	69.5	68.0	54.0	61.5	59.0	51.0	56.5
		+ Normalization (prop.)	58.0	74.5	71.5	56.0	62.0	70.5	-	-
		Windowed-PPL (prop.)	59.8	57.0	55.8	52.2	64.0	65.5	51.5	58.0
		Localized-PPL (prop.)	53.0	60.5	63.5	53.2	53.5	57.2	-	-
		+ Normalization (prop.)	55.5	63.0	60.5	56.5	61.5	66.0	-	-
	1.3B	Global-PPL	61.0	69.0	69.5	54.5	60.5	59.5	52.5	57.0
		+ Normalization (prop.)	54.0	72.5	73.0	57.5	63.5	69.5	-	-
		Windowed-PPL (prop.)	58.2	55.2	55.0	55.2	61.8	66.0	54.0	55.0
		Localized-PPL (prop.)	50.0	61.5	68.0	54.8	53.0	54.2	-	-
		+ Normalization (prop.)	52.5	65.5	62.0	60.5	62.5	62.0	-	-
	7B	Global-PPL	61.5	70.5	70.0	55.5	59.5	61.5	51.0	53.5
		+ Normalization (prop.)	56.0	74.0	72.0	59.0	64.5	69.0	-	-
		Windowed-PPL (prop.)	60.0	57.8	53.8	55.8	59.2	69.8	51.5	56.5
		Localized-PPL (prop.)	53.0	55.5	65.0	58.8	56.0	57.8	-	-
		+ Normalization (prop.)	53.0	67.0	63.5	58.0	65.0	65.5	-	-
pGSLM	–	Global-PPL	56.0	74.0	69.0	65.5	71.5	52.0	53.0	54.0
		+ Normalization (prop.)	60.5	80.0	89.0	62.5	71.0	69.0	-	-
		Windowed-PPL (prop.)	59.0	61.8	63.2	64.5	61.5	63.5	49.0	52.5
		Localized-PPL (prop.)	70.5	78.5	68.0	50.5	57.5	56.0	-	-
		+ Normalization (prop.)	66.0	82.5	87.5	57.0	66.5	68.5	-	-
Spirit LM	base	Global-PPL	52.5	67.5	67.0	53.0	55.5	54.5	47.5	51.0
		+ Normalization (prop.)	51.5	60.0	67.0	49.0	52.0	64.5	-	-
		Windowed-PPL (prop.)	48.8	50.8	58.5	53.8	56.0	68.5	46.0	60.8
		Localized-PPL (prop.)	54.5	62.0	64.0	56.0	48.5	52.8	-	-
		+ Normalization (prop.)	54.0	57.0	59.5	55.5	57.0	52.0	-	-
	Expr.	Global-PPL	71.0	81.5	85.5	55.5	64.0	55.5	53.0	59.5
		+ Normalization (prop.)	73.5	80.0	87.0	53.5	56.5	67.5	-	-
		Windowed-PPL (prop.)	58.5	66.0	67.5	50.0	57.5	64.8	53.0	55.0
		Localized-PPL (prop.)	73.5	80.5	84.5	54.5	55.0	58.5	-	-
		+ Normalization (prop.)	66.5	75.5	81.0	55.5	57.0	60.5	-	-
TASTE	emb	Global-PPL	56.5	69.5	75.5	39.5	48.5	55.5	54.0	42.0
		+ Normalization (prop.)	62.0	67.0	68.0	49.0	49.0	55.0	-	-
		Windowed-PPL (prop.)	57.2	52.2	57.5	48.2	49.5	51.0	51.5	48.0
		Localized-PPL (prop.)	54.5	57.0	60.5	44.5	43.8	54.0	-	-
		+ Normalization (prop.)	55.5	52.0	55.5	47.0	48.5	53.5	-	-
Flow-SLM	270M	Global-PPL	61.5	75.0	76.5	66.0	65.0	56.5	59.0	53.5
		+ Normalization (prop.)	77.5	86.0	82.5	56.0	63.5	80.5	-	-
		Windowed-PPL (prop.)	65.5	68.2	71.8	55.5	60.8	66.5	54.5	55.0
		Localized-PPL (prop.)	75.0	72.0	77.5	62.5	62.0	67.5	-	-
		+ Normalization (prop.)	75.5	68.5	77.0	54.5	63.0	78.0	-	-
	1B	Global-PPL	65.5	77.0	79.0	69.0	66.0	76.0	58.5	54.5
		+ Normalization (prop.)	78.0	85.0	87.5	58.0	69.0	88.5	-	-
		Windowed-PPL (prop.)	67.8	66.8	75.2	59.0	61.2	77.2	58.0	53.5
		Localized-PPL (prop.)	79.5	73.5	79.0	64.5	68.0	81.5	-	-
		+ Normalization (prop.)	80.5	69.0	80.0	53.0	63.0	78.5	-	-
	1B-ext.	Global-PPL	64.5	76.0	80.5	69.5	64.5	73.5	57.5	53.5
		+ Normalization (prop.)	82.5	85.0	88.5	58.5	73.5	85.5	-	-
		Windowed-PPL (prop.)	66.5	71.2	73.0	61.5	61.8	77.0	52.0	57.0
		Localized-PPL (prop.)	79.5	76.0	81.5	65.5	69.5	83.5	-	-
		+ Normalization (prop.)	82.0	74.5	81.0	54.0	67.5	81.0	-	-
Llama-mimi	1.3B	Global-PPL	79.5	85.5	82.0	75.0	72.0	92.0	53.0	49.0
		+ Normalization (prop.)	89.0	**95.0**	**99.5**	73.5	87.0	**98.5**	-	-
		Windowed-PPL (prop.)	80.5	91.2	96.2	65.2	66.5	81.2	51.0	53.5
		Localized-PPL (prop.)	95.5	**95.5**	**100.0**	79.5	84.5	**97.5**	-	-
		+ Normalization (prop.)	92.0	**98.0**	**99.0**	81.0	85.0	**97.0**	-	-
Human Topline (Measured by (Maimon et al., 2025))										
Human	–	–	97.2	91.5	98.6	83.1	88.7	94.4	93.3	95.7

C.2 Model-as-a-Judge for Measuring Generation Consistency

Table 6 exhibits task-wise performance on embedding judge candidates. From the results, we observe that no single embedding model aces all six attributes, motivating future work on more generalizable audio embedding approaches. In addition, these results indicate that acoustic features (speaker features, background features) are effectively time-invariant at the resolution probed by current evaluation protocols. Such stability endorses a prevalent architectural choice of conducting speech synthesis conditioned on a constant residual acoustic embedding. Notably, CAM++ serves as a highly versatile, general-purpose acoustic representation model and is widely adopted in generative speech systems (Tseng et al., 2025; Du et al., 2024; Hsu et al., 2025). Related design choices also appear in other spoken-language models with speech generation (Nguyen et al., 2024; Lakhotia et al., 2021; Hassid et al., 2024) and conversational speech frameworks (Défossez et al., 2024).

Table 7 shows the results of SLM performance evaluated by selected judge models. Similar to the conclusions made in the proposed likelihood estimation methods, evaluating on true continuations shows scores are in the vicinity of the human topline for top performing models, especially on traits related to human speech (sentiment, speaker, gender).

Whereas Llama-Mimi integrates a deeper hierarchy of Mimi token layers into its speech modeling pipeline, Flow-SLM deploys a more intricate flow-matching speech decoder, which may account for its elevated scores on certain subtasks measuring directly in the speech.

Overall, HuBERT-based models perform close to chance level. A notable exception is Spirit-LM-Expressive, which includes additional pitch and style tokens. Experiment shows that this additional information is best reflected in sentiment performance, reaching 72%.

A closer examination of reconstruction and continuation performance reveals that the failure in continuation arises from the reconstructed audio lacking the relevant content to begin with. In most cases, reconstructions achieve near-chance scores mirroring their continuation counterparts, which is strikingly poor given that evaluation on the original audio yields near-ceiling performance on the benchmark. Manual inspection of the audio reveals that the audio is indeed greatly distorted compared to the original sample, where semantic pronunciations are greatly preserved but speaker and background information collapses. This example illustrates a key limitation of global uncertainty measures. Even when they showed moderate performance, such performance failed to generalize to continuation performance that requires local acoustic fidelity at each generation step.

C.3 Shapley Analysis

Table 8 shows token-type ablations and Shapley attributions for Spirit-LM-Expressive (HuBERT H, pitch P, style S) and Llama-Mimi (layers 0–3), under four evaluation settings. The top half of each panel reports the observed accuracies, and the bottom half reports the corresponding Shapley values ϕ computed from those accuracies using a null baseline of 50% on every task. For Spirit-LM-Expressive, localization alone leaves the average HuBERT and pitch attributions nearly unchanged ($\phi_H : 9.6 \rightarrow 9.6$, $\phi_P : 9.5 \rightarrow 9.4$), whereas normalization shifts attribution away from HuBERT ($\phi_H : 9.6 \rightarrow 7.6$) and toward pitch ($\phi_P : 9.5 \rightarrow 13.1$). For Llama-Mimi, both localization and normalization generally increase Shapley values, with the largest gains concentrated in the middle layers: under normalization, ϕ_2 rises from 12.2 to 15.8 and ϕ_3 from 8.1 to 10.5; under localization, ϕ_2 rises to 14.5 and ϕ_3 to 11.1. These attribution shifts are most pronounced on speaker-related attributes, including sentiment, speaker identity, and gender.

C.4 Loss Response Figures

Figures 5 through 12 show a per-task breakdown of the models' NLL-loss responses. The degree of separability between the positive and negative NLL responses in these plots largely correlates with the resulting accuracy. Consistent with this, speaker-related attributes (sentiment, speaker identity, and gender) exhibit larger separations.

Table 6: Embedding model performance on SALMon, where the accuracy is aggregated over $d(S, P) < d(S, N)$.

Embedding Model Performance						
	Acoustic Consistency					
	Sentiment ↑	Speaker ↑	Gender ↑	Bg (domain) ↑	Bg (rand.) ↑	Room ↑
TITANET (Koluguri et al., 2022)	**99.5**	**100.0**	**100.0**	58.5	70.5	94.0
CAM++ (Wang et al., 2023b)	95.5	**95.5**	**99.0**	69.5	84.5	**94.5**
CLAP (Elizalde et al., 2023)	96.0	91.0	98.5	78.0	**90.0**	**97.0**
wav2vec2-large-audioset (La Quatra et al., 2024)	91.5	70.0	75.5	82.5	**95.5**	**99.0**
HuBERT-large-audioset (La Quatra et al., 2024)	90.0	74.5	74.5	**86.5**	**97.5**	**95.5**
data2vec-audio-large (Baevski et al., 2022)	58.5	67.0	59.5	54.0	45.5	77.5
wavlm-large (Chen et al., 2022)	75.0	66.5	69.5	52.5	56.0	93.0
wav2vec2-large-robust (Hsu et al., 2021)	74.5	58.5	52.5	59.5	67.5	76.5
Selected Classifier	TITANET	TITANET	TITANET	HuBERT-large-audioset	hubert-large-audioset	wav2vec2-large-audioset
Human	97.2	91.5	98.6	83.1	88.7	94.4

Table 7: Generation performance of spoken language models judged by the appropriate model as a judge. Main numbers report performance of continuation samples from the SLM; parenthesized numbers report the performance of reconstructed audio that is generated from the speech tokens of SLMs. Boldface items indicate that the generation performance exceeds the human topline reported by (Maimon et al., 2025).

	Acoustic Consistency					
	Sentiment ↑	Speaker ↑	Gender ↑	Bg (domain) ↑	Bg (rand.) ↑	Room ↑
Judge Model / Evaluated Model	TITANET	TITANET	TITANET	HuBERT-large-audioset	HuBERT-large-audioset	wav2vec2-large-audioset
GSLM	55.0 (57.5)	55.0 (58.5)	50.5 (50.0)	59.5 (60.0)	55.5 (52.0)	53.0 (48.0)
TWIST-350M	48.0 (48.5)	53.0 (53.5)	52.5 (53.0)	61.5 (60.0)	53.0 (53.5)	52.5 (49.0)
TWIST-1.3B	48.0 (50.5)	55.0 (52.5)	54.5 (52.0)	62.5 (60.0)	54.0 (54.0)	53.5 (48.5)
TWIST-7B	48.0 (51.0)	54.5 (53.0)	52.0 (53.0)	59.5 (60.5)	54.0 (54.0)	51.5 (49.0)
pGSLM	46.5 (52.0)	49.5 (57.5)	55.5 (58.0)	51.0 (51.0)	54.0 (53.5)	52.5 (49.0)
Spirit-LM	47.5 (50.0)	52.0 (50.5)	50.5 (50.5)	50.5 (51.5)	52.5 (53.0)	50.0 (47.5)
Spirit-LM-expr.	72.0 (81.0)	54.0 (53.0)	56.0 (58.0)	58.5 (62.0)	55.5 (56.0)	54.5 (50.0)
TASTE-emb.	95.0 (95.0)	**100.0** **(99.0)**	**100.0** **(99.5)**	61.5 (63.0)	54.5 (53.5)	57.5 (53.0)
Flow-SLM-270M	85.0 (98.0)	87.0 (100.0)	92.5 (100.0)	77.0 (80.0)	86.0 (94.0)	69.5 (96.5)
Flow-SLM-1B	92.5 (98.0)	**99.5** **(100.0)**	**99.0** **(100.0)**	77.0 (80.0)	84.5 (94.0)	77.0 (97.0)
Flow-SLM-1B-Ext.	93.0 (98.0)	**98.0** **(100.0)**	**100.0** **(100.0)**	77.0 (80.0)	86.5 (94.0)	85.0 (97.0)
Llama-mimi	93.0 (92.0)	90.0 (95.0)	95.0 (95.0)	81.0 (78.5)	**90.0** **(91.0)**	78.5 (85.0)
Human Topline	97.2	91.5	98.6	83.1	88.7	94.4

Table 8: Token-type ablations and Shapley attributions for Spirit-LM (HuBERT H, pitch P, style S) and Llama-Mimi (layers 0–3), under four evaluation settings that combine estimator locality (Global vs. Localized) and scoring normalization (Original vs. Normalized).

Spirit-LM

Original

Global

H	P	S	Sent.	Spk.	Gen.	BgD.	BgR.	Room	Avg
✓	✓	✓	71.0	81.5	85.5	55.5	64.0	55.5	68.8
✓	✓	–	73.0	81.5	85.0	57.0	63.0	54.5	69.0
✓	–	✓	57.5	67.0	74.0	49.5	58.0	57.0	60.5
–	✓	✓	57.0	70.0	74.0	54.0	56.5	51.0	60.4
✓	–	–	56.5	67.5	73.0	50.5	60.0	54.5	60.3
–	✓	–	56.0	69.5	72.5	54.0	57.0	51.5	60.1
–	–	✓	54.0	48.5	46.0	45.5	49.0	52.0	49.2
	ϕ_H		+10.2	+14.8	+18.2	+1.8	+8.3	+4.3	+9.6
	ϕ_P		+9.8	+17.2	+18.0	+5.8	+6.1	-0.2	+9.5
	ϕ_S		+1.0	-0.5	-0.7	-2.2	-0.4	+1.3	-0.3

Localized ($t = 0.5$s)

H	P	S	Sent.	Spk.	Gen.	BgD.	BgR.	Room	Avg
✓	✓	✓	73.5	80.5	84.5	54.5	55.0	58.5	67.8
✓	✓	–	75.0	81.5	85.5	57.0	55.0	58.5	68.8
✓	–	✓	59.5	67.5	72.0	53.5	50.0	59.5	60.3
–	✓	✓	64.5	68.5	80.0	52.5	46.5	53.0	60.8
✓	–	–	61.0	66.0	71.5	56.5	54.0	60.0	61.5
–	✓	–	62.0	71.0	80.0	52.0	47.0	52.0	60.7
–	–	✓	54.0	43.0	46.2	48.2	47.8	47.2	47.8
	ϕ_H		+9.8	+15.2	+13.9	+4.5	+5.9	+8.3	+9.6
	ϕ_P		+12.7	+18.2	+22.1	+1.8	+0.6	+1.0	+9.4
	ϕ_S		+1.0	-2.8	-1.5	-1.8	-1.5	-0.8	-1.2

Normalized

Global

H	P	S	Sent.	Spk.	Gen.	BgD.	BgR.	Room	Avg
✓	✓	✓	73.5	80.0	87.0	53.5	56.5	67.5	69.7
✓	✓	–	78.5	78.0	89.5	53.0	56.0	72.0	71.2
✓	–	✓	66.0	70.5	71.0	55.5	46.5	52.5	60.3
–	✓	✓	66.0	65.0	78.0	49.5	58.5	69.5	64.4
✓	–	–	64.0	73.5	72.5	50.5	51.0	52.0	60.6
–	✓	–	67.0	67.5	82.0	53.0	61.5	73.0	67.3
–	–	✓	49.0	51.5	52.5	44.5	49.5	53.0	50.0
	ϕ_H		+11.9	+17.8	+14.8	+3.3	-1.8	-0.2	+7.6
	ϕ_P		+13.4	+12.0	+23.1	+1.6	+9.5	+18.8	+13.1
	ϕ_S		-1.8	+0.2	-0.9	-1.4	-1.2	-1.0	-1.0

Localized ($t = 0.5$s)

H	P	S	Sent.	Spk.	Gen.	BgD.	BgR.	Room	Avg
✓	✓	✓	66.5	75.5	81.0	55.5	57.0	60.5	66.0
✓	✓	–	73.0	80.0	84.0	54.0	62.0	63.5	69.4
✓	–	✓	56.0	67.5	58.5	58.5	52.5	53.5	57.8
–	✓	✓	62.0	66.5	83.0	57.0	53.5	62.0	64.0
✓	–	–	54.0	69.5	59.5	55.5	52.5	56.0	57.8
–	✓	–	68.5	70.5	86.0	58.0	56.5	62.0	66.9
–	–	✓	46.0	42.5	42.5	52.8	47.0	57.0	48.0
	ϕ_H		+5.2	+15.2	+4.8	+1.6	+3.8	+1.2	+5.3
	ϕ_P		+15.5	+15.2	+30.3	+2.1	+6.3	+8.4	+13.0
	ϕ_S		-4.2	-5.0	-4.2	+1.7	-3.2	+0.9	-2.3

Llama-Mimi

Original

Global

0	1	2	3	Sent.	Spk.	Gen.	BgD.	BgR.	Room	Avg
✓	✓	✓	✓	79.5	85.5	82.0	75.0	72.0	92.0	81.0
✓	–	–	–	57.0	71.0	65.0	67.5	67.5	71.5	66.6
–	✓	–	–	60.5	70.0	73.5	64.5	62.5	71.5	67.1
–	–	✓	–	75.0	82.5	84.5	68.0	67.0	91.0	78.0
–	–	–	✓	62.5	78.5	77.0	67.0	67.0	81.5	72.2
✓	✓	–	–	72.5	76.5	75.5	67.0	68.5	75.0	72.5
–	✓	✓	–	76.0	80.0	84.0	70.5	69.5	90.5	78.4
–	–	✓	✓	71.0	84.5	85.5	72.5	73.0	96.0	80.4
✓	–	–	✓	61.0	79.0	73.5	73.5	74.0	79.5	73.4
✓	–	✓	–	70.0	78.5	80.5	71.5	70.0	92.5	77.2
–	✓	–	✓	76.0	80.5	81.0	73.0	70.5	80.5	76.9
✓	✓	✓	–	78.0	82.0	80.0	73.5	72.0	91.5	79.5
✓	✓	–	✓	74.0	79.5	77.5	71.0	71.0	80.5	75.6
✓	–	✓	✓	67.0	83.0	80.5	73.5	72.5	95.5	78.7
–	✓	✓	✓	80.5	85.5	85.0	71.0	72.5	93.0	81.2
	ϕ_0			+1.6	+5.5	+1.5	+6.6	+5.8	+5.4	+4.4
	ϕ_1			+10.8	+6.5	+7.7	+4.5	+3.5	+4.4	+6.2
	ϕ_2			+12.0	+12.8	+13.9	+7.2	+6.1	+21.4	+12.2
	ϕ_3			+5.0	+10.8	+8.9	+6.8	+6.6	+10.8	+8.1

Localized ($t = 0.5$s)

0	1	2	3	Sent.	Spk.	Gen.	BgD.	BgR.	Room	Avg
✓	✓	✓	✓	95.5	95.5	100.0	79.5	84.5	97.5	92.1
✓	–	–	–	75.5	83.5	85.0	59.5	64.8	73.0	73.5
–	✓	–	–	78.0	81.5	85.0	59.0	68.0	80.0	75.2
–	–	✓	–	88.5	91.0	93.5	73.0	80.0	92.5	86.4
–	–	–	✓	75.0	84.0	94.0	70.5	75.5	80.0	79.8
✓	✓	–	–	85.5	90.0	92.5	67.0	74.0	87.5	82.8
–	✓	✓	–	90.5	92.5	95.0	73.5	80.5	94.5	87.8
–	–	✓	✓	92.0	91.5	98.0	76.5	83.5	94.5	89.3
✓	–	–	✓	83.0	91.5	95.0	73.5	79.2	82.5	84.1
✓	–	✓	–	89.5	92.5	97.5	74.5	80.0	91.0	87.5
–	✓	–	✓	87.5	90.0	96.5	75.0	77.5	89.0	85.9
✓	✓	✓	–	94.0	93.5	99.0	74.0	81.0	97.5	89.8
✓	✓	–	✓	90.0	96.0	98.5	75.0	81.5	91.0	88.7
✓	–	✓	✓	94.0	94.5	99.5	79.5	82.5	92.5	90.4
–	✓	✓	✓	94.5	95.5	99.0	78.5	84.5	97.5	91.6
	ϕ_0			+8.7	+10.7	+10.7	+4.0	+4.8	+6.7	+7.6
	ϕ_1			+10.6	+10.1	+10.3	+3.5	+6.3	+12.4	+8.9
	ϕ_2			+16.8	+13.4	+14.6	+11.2	+12.7	+18.5	+14.5
	ϕ_3			+9.4	+11.3	+14.4	+10.8	+10.7	+9.9	+11.1

Normalized

Global

0	1	2	3	Sent.	Spk.	Gen.	BgD.	BgR.	Room	Avg
✓	✓	✓	✓	89.0	95.0	99.5	73.5	87.0	98.5	90.4
✓	–	–	–	55.5	69.0	73.5	60.0	62.0	75.5	65.9
–	✓	–	–	75.5	72.5	77.5	54.5	67.0	80.0	71.2
–	–	✓	–	84.5	89.5	93.0	67.5	70.0	86.5	81.8
–	–	–	✓	72.0	78.0	82.5	64.5	67.5	83.5	74.7
✓	✓	–	–	72.5	78.5	84.5	60.5	72.0	81.5	74.9
–	✓	✓	–	90.5	93.0	98.0	65.0	79.5	94.0	86.7
–	–	✓	✓	84.5	92.5	96.5	70.5	77.0	92.5	85.6
✓	–	–	✓	72.5	83.0	88.0	66.0	74.5	87.0	78.5
✓	–	✓	–	77.5	90.5	93.5	67.5	73.5	89.0	81.9
–	✓	–	✓	82.5	86.0	89.5	63.0	75.5	87.5	80.7
✓	✓	✓	–	83.5	92.5	95.5	69.5	80.5	95.0	86.1
✓	✓	–	✓	79.5	86.0	92.5	65.0	79.0	91.0	82.2
✓	–	✓	✓	82.0	92.0	95.0	71.5	80.0	94.5	85.8
–	✓	✓	✓	89.5	93.5	98.0	71.5	83.0	96.0	88.6
	ϕ_0			-0.6	+6.0	+7.2	+4.2	+5.9	+8.2	+5.2
	ϕ_1			+12.4	+8.6	+10.6	+1.5	+9.8	+11.1	+9.0
	ϕ_2			+17.4	+19.4	+19.2	+10.4	+11.6	+16.5	+15.8
	ϕ_3			+9.7	+11.0	+12.4	+7.3	+9.7	+12.8	+10.5

Localized ($t = 0.5$s)

0	1	2	3	Sent.	Spk.	Gen.	BgD.	BgR.	Room	Avg
✓	✓	✓	✓	92.0	98.0	99.0	81.0	85.0	97.0	92.0
✓	–	–	–	64.0	76.0	86.0	64.0	66.0	71.5	71.2
–	✓	–	–	75.0	75.5	83.5	53.0	61.5	74.0	70.4
–	–	✓	–	84.5	89.5	91.0	67.0	83.5	89.5	84.2
–	–	–	✓	75.0	80.0	88.0	72.0	67.5	75.5	76.3
✓	✓	–	–	76.5	85.5	90.5	66.5	66.0	83.5	78.1
–	✓	✓	–	91.5	94.0	95.5	72.5	81.0	94.5	88.2
–	–	✓	✓	86.5	94.0	96.5	77.0	79.5	93.0	87.8
✓	–	–	✓	76.0	87.5	94.0	75.0	68.5	84.5	80.9
✓	–	✓	–	83.5	92.0	94.0	74.5	83.0	90.5	86.2
–	✓	–	✓	85.0	86.0	94.0	67.0	68.5	85.0	80.9
✓	✓	✓	–	90.0	94.0	98.5	76.0	84.5	95.0	89.7
✓	✓	–	✓	86.5	91.0	98.0	72.0	72.0	90.5	85.0
✓	–	✓	✓	84.5	95.5	97.5	83.0	85.5	94.0	90.0
–	✓	✓	✓	92.0	95.5	98.5	74.0	79.5	96.0	89.2
	ϕ_0			+3.5	+9.3	+11.1	+8.5	+6.9	+7.8	+7.8
	ϕ_1			+12.5	+9.2	+10.9	+0.1	+3.0	+10.1	+7.6
	ϕ_2			+16.4	+17.8	+14.2	+11.5	+19.5	+18.9	+16.4
	ϕ_3			+9.7	+11.6	+12.8	+11.0	+5.5	+10.2	+10.1

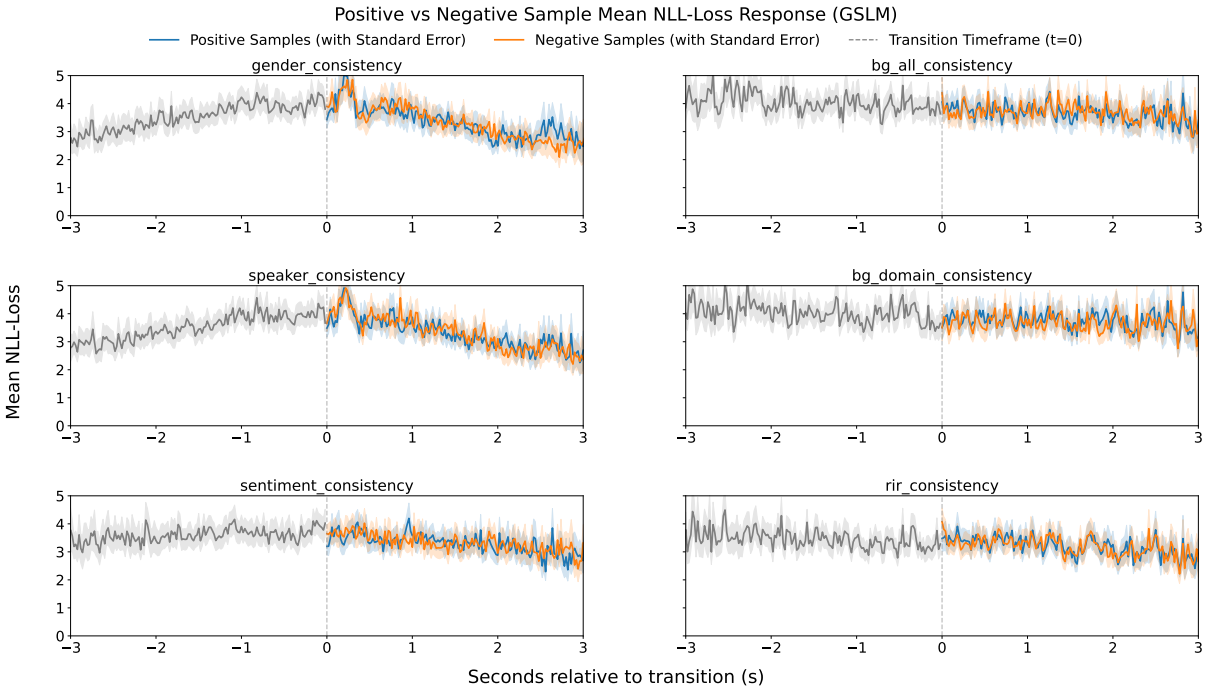

Figure 5: Positive vs. Negative Sample Mean NLL-Loss Response for GSLM across six consistency splits.

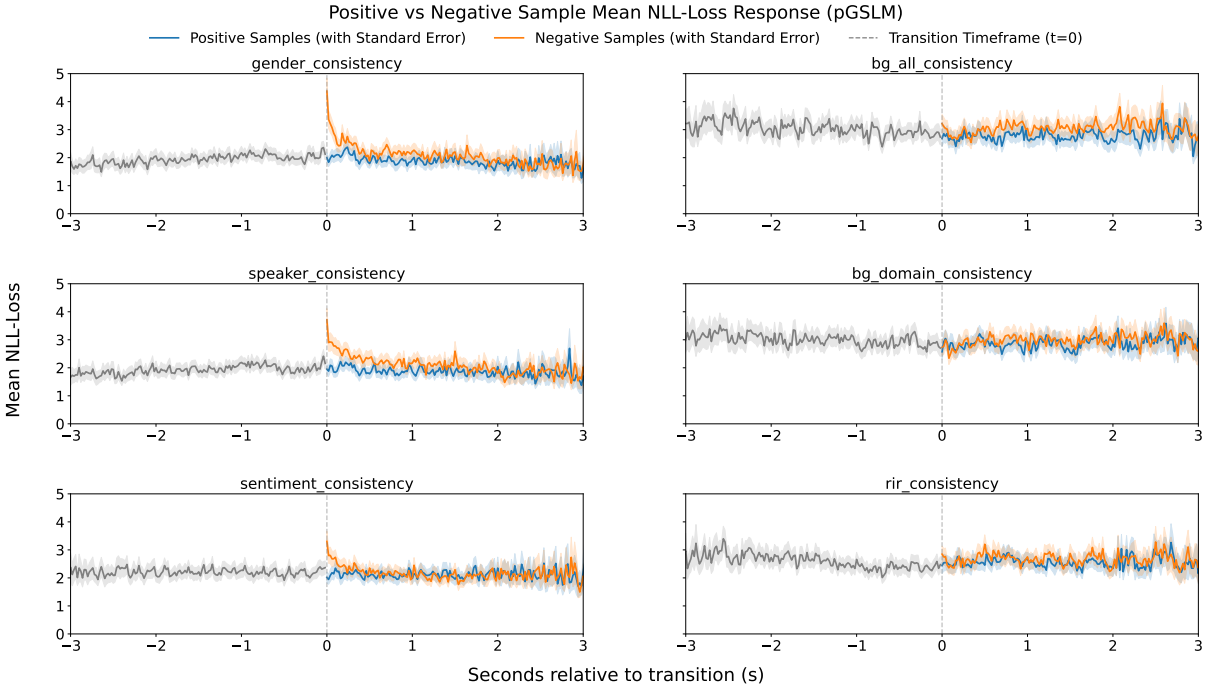

Figure 6: Positive vs. Negative Sample Mean NLL-Loss Response for pGSLM across six consistency splits.

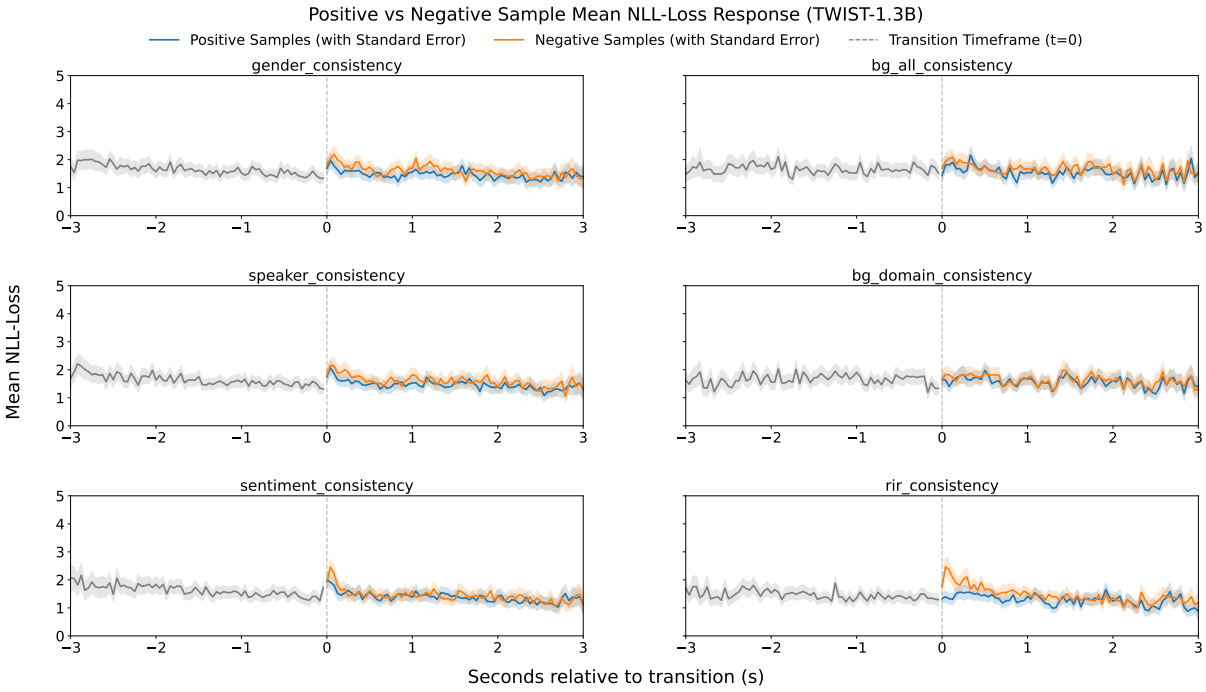

Figure 7: Positive vs. Negative Sample Mean NLL-Loss Response for `TWIST-1.3B` across six consistency splits.

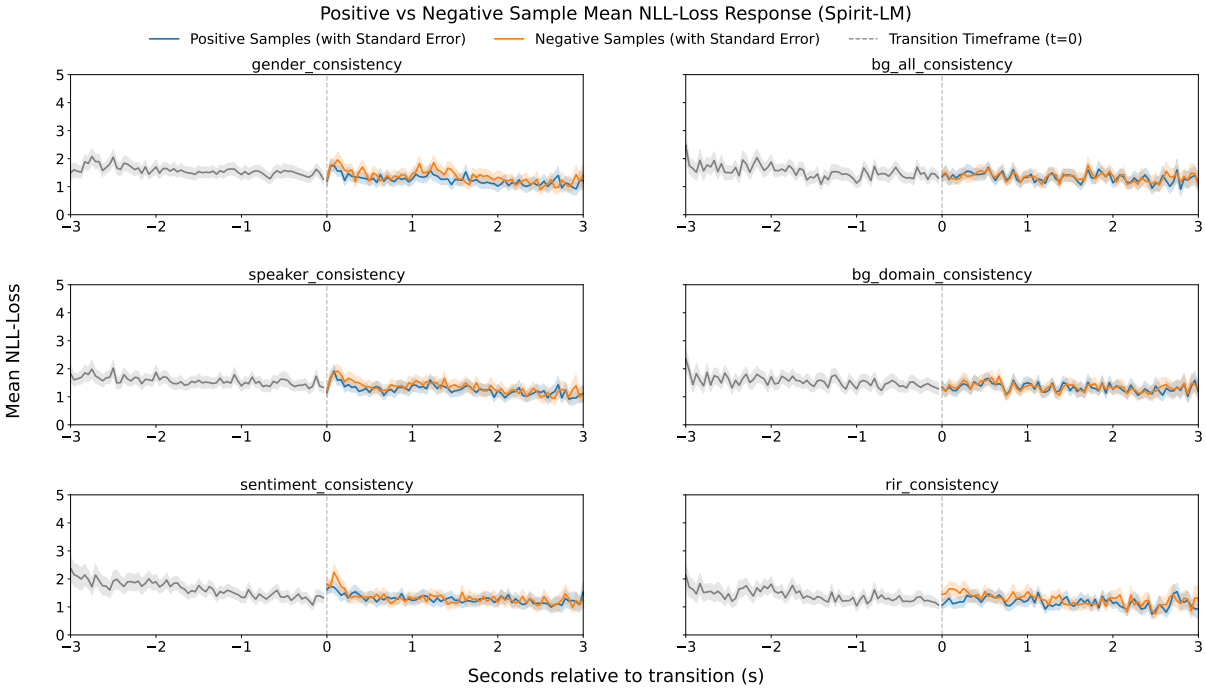

Figure 8: Positive vs. Negative Sample Mean NLL-Loss Response for `Spirit-LM` across six consistency splits.

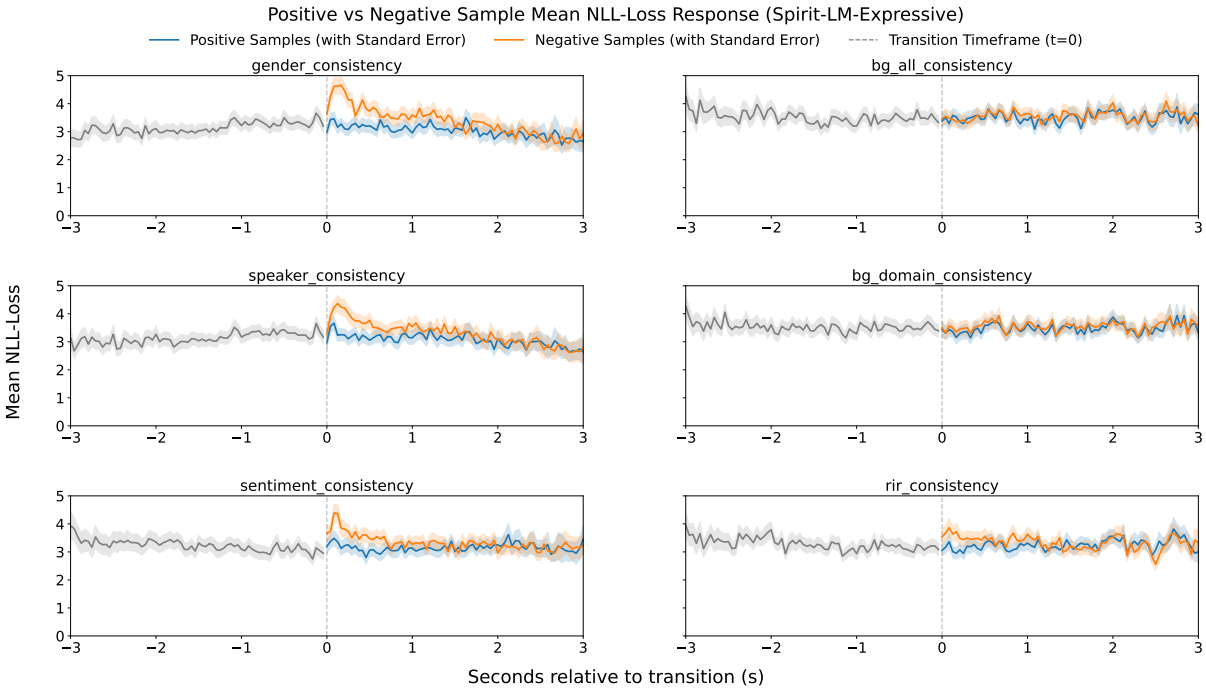

Figure 9: Positive vs. Negative Sample Mean NLL-Loss Response for `Spirit-LM-Expressive` across six consistency splits.

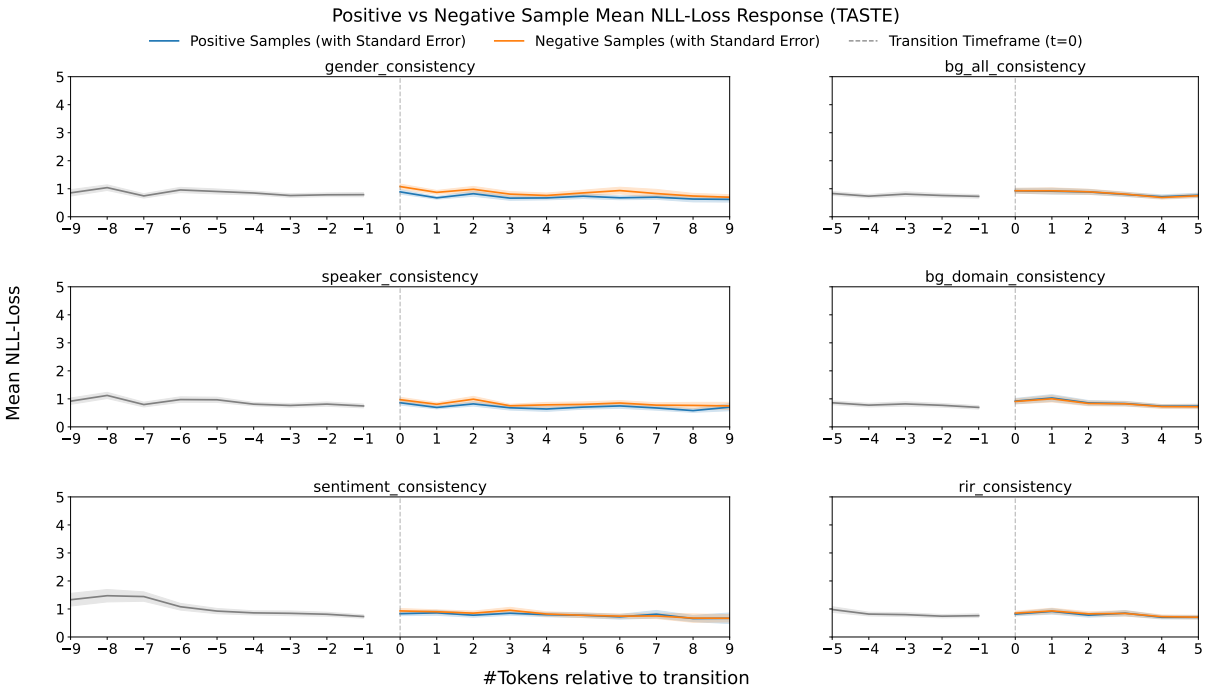

Figure 10: Positive vs. Negative Sample Mean NLL-Loss Response for `TASTE` across six consistency splits. We follow TASTE's audio-text alignment setting and report with textual tokens as the granularity of the x-axis.

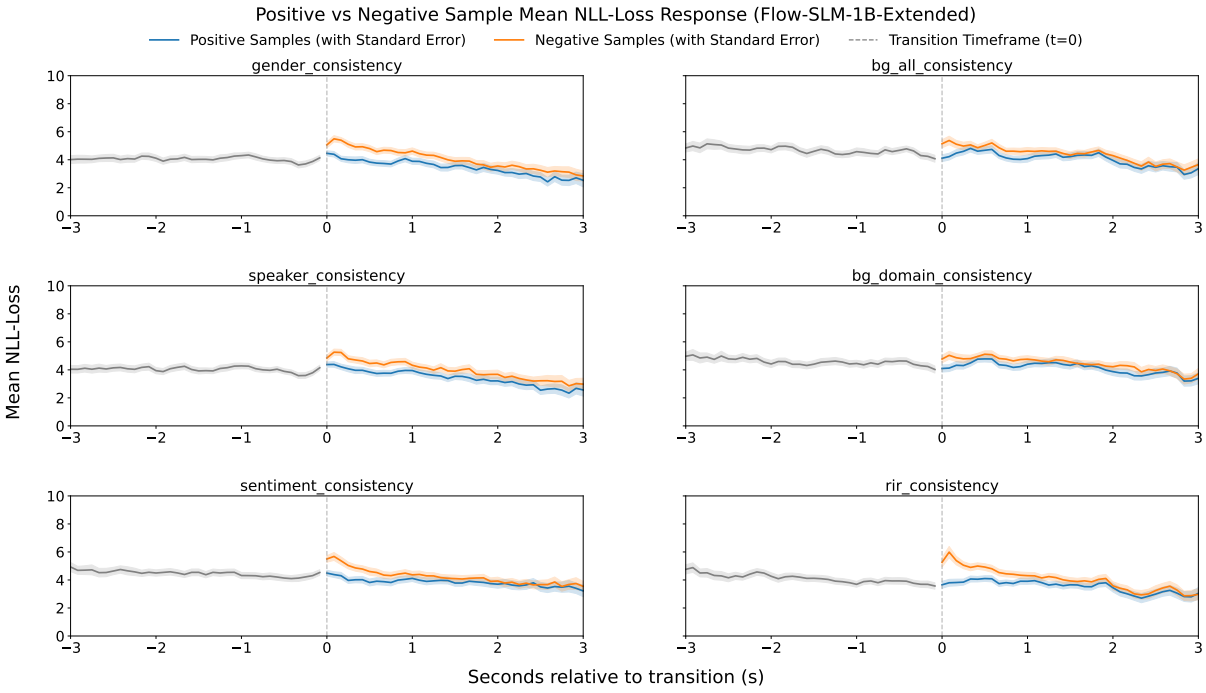

Figure 11: Positive vs. Negative Sample Mean NLL-Loss Response for `Flow-SLM-1B-Extended` across six consistency splits. At the transition timeframe (t=0), each of the category has distinct postive and negative response (clear separation by 95% confidence interval).

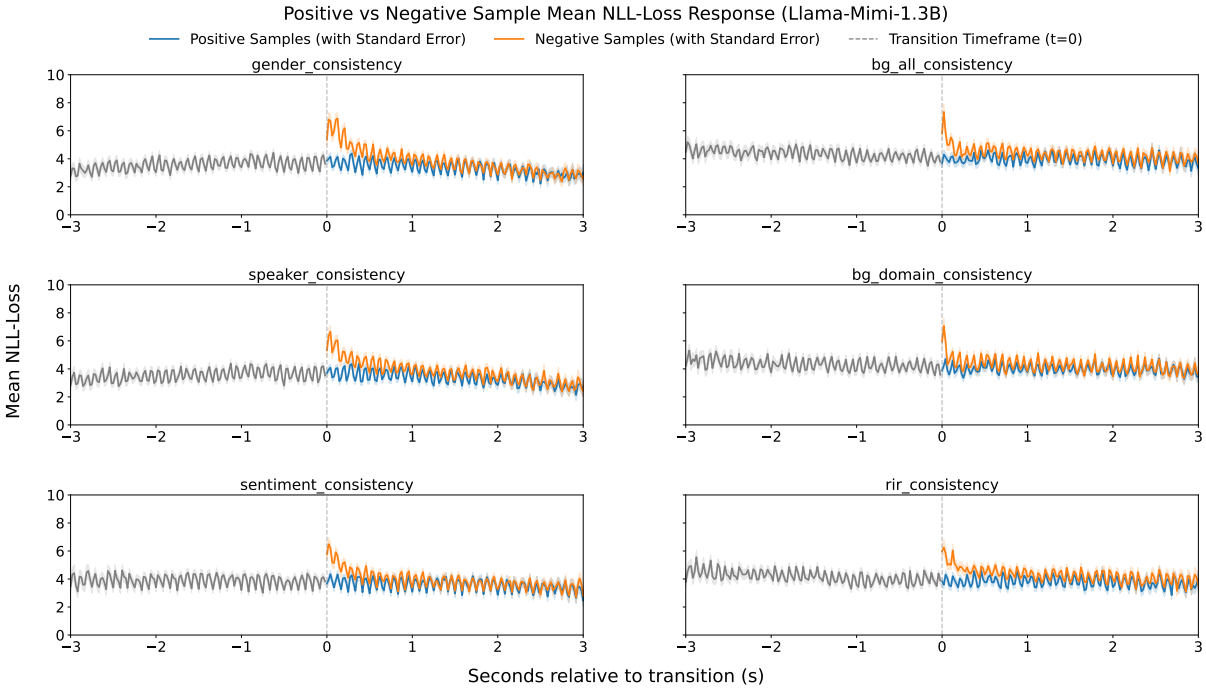

Figure 12: Positive vs. Negative Sample Mean NLL-Loss Response for `Llama-mimi-1.3B` across six consistency splits. At the transition timeframe (t=0), each of the category has distinct postive and negative response (clear separation by 95% confidence interval).