StressTest: Can YOUR Speech LM Handle the Stress?

Anonymous ACL submission

Abstract

Sentence stress refers to emphasis on words within a spoken utterance to highlight or contrast an idea. It is often used to imply an underlying intention not explicitly stated. Recent speech-aware language models (SLMs) have enabled direct audio processing, allowing models to access the full richness of speech to perform audio reasoning tasks such as spoken question answering. Despite the crucial role of sentence stress in shaping meaning and intent, it remains largely overlooked in evaluation and development of SLMs. We address this gap by introducing *StressTest*, a benchmark designed to evaluate models' ability to distinguish between meanings of speech based on the stress pattern. We evaluate leading SLMs, and find that despite their overall capabilities, they perform poorly on such tasks. Hence, we propose a novel data generation pipeline, and create Stress-17k, a training set that simulates change of meaning implied by stress variation. Results suggest, that our finetuned model, StressSLM, generalizes well to real recordings and notably outperforms existing SLMs on sentence stress reasoning and detection. Models, code and data will be publicly available, samples - https://anonymous.4open.science/w/strsstst.

1 Introduction

Large language models (LLMs) have revolutionized language processing, enabling new forms of human-computer interaction (Minaee et al., 2025; Grattafiori et al., 2024). Further research explored integration of other modalities into LLMs. Notably, incorporation of speech and audio into LLMs has gained traction, equipping models with the ability to listen, speak, and reason about audio (Arora et al., 2025; Ghosh et al., 2025; Maimon et al., 2025a).

An elementary approach of integrating speech into LLMs follows a cascade paradigm (Ji et al., 2024), where audio is transcribed by an automatic speech recognition (ASR) system, and the resulting text is processed by an LM. While somewhat effective, this approach falls short of capturing the full expressive range of spoken language. Speech carries rich paralinguistic cues such as emotion, speaker identity, and prosodic characteristics that are often lost in transcription (Wilson and Wharton, 2006; Van Heuven, 2018). One key aspect of prosody is sentence stress, which refers to the emphasis placed on particular words or phrases within a sentence to highlight an idea or to contrast another (Bolinger, 1972). For example, the sentence *"I didn't say she stole the money"* can express dramatically different meanings, based on the stressed words, for the same written-form.

Recent research on SLMs aim to address these limitations by enabling models to process audio directly, bypassing transcription and allowing them to access acoustic information (Arora et al., 2025). Leading models in this space have demonstrated impressive performance across a wide range of speech tasks (Chu et al., 2024; Tang et al., 2024). Nevertheless, sentence stress has received limited attention in the evaluation and development of SLMs, despite its vital role in expressing the speaker's intent and meaning. We argue that interpreting sentence stress requires the listener to reason about the intended meaning based on stress placement, which can often be inferred even without explicit context.

In this work, we address this gap by introducing *StressTest*, a comprehensive benchmark designed to evaluate a model's ability to distinguish spoken sentence meanings based on different stress patterns. A visual example can be seen in Figure 1. We then evaluate leading SLMs on our benchmark to quantify their capacity for stress based reasoning. Additionally, we introduce a novel synthetic dataset generation pipeline, which produces the Stress-17k dataset. We empirically show that finetuning a model using Stress-17k leads to enhanced ability to detect and model sentence stress on real-world recordings. Through extensive empirical evaluation and ablation studies, we demonstrate that our

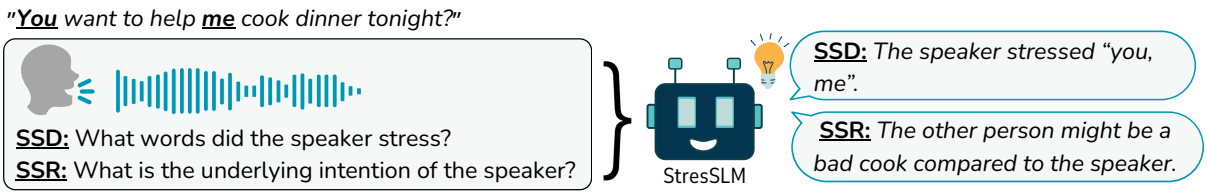

Figure 1: *StressTest* provides samples that can be understood differently based on stress. We consider sentence stress detection (SSD) and sentence stress reasoning (SSR). StresSLM detects stress and reasons about the meaning.

model, StresSLM, significantly outperforms existing models in both stress detection and reasoning, with minimal performance drop on original tasks.

Our contributions: (i) We propose *StressTest*, a novel benchmark for evaluating sentence stress understanding in SLMs, and further extend it with existing stress-labeled data; (ii) We analyze performance of several leading SLMs on the benchmark, and show that they fail to detect stress and reason about its meaning; (iii) We propose a data synthesis and filtering pipeline and demonstrate its effectiveness for training an SLM for stress understanding.

2 Background

Theoretical views of sentence stress. Theoretical accounts of sentence stress fall into two perspectives (Ladd, 2008): (i) The phonological view that treats normal stress as a default prosodic pattern governed by syntax, formalized in the Nuclear Stress Rule (Chomsky and Halle, 1968). This approach is not tied to meaning; (ii) The semantic view, which considers stress as context-dependent, reflecting the speaker's intent to mark focus, contrast, new information, or emphasis (Bolinger, 1972). In this work, we adopt the latter, asking whether SLMs can infer speaker-intended meaning from stress.

Within the semantic view, sentence stress can be grouped into four overlapping categories: (i) *Contrastive Stress*, marking opposition (Bolinger, 1961; Dretske, 1972; Boer, 1979); (ii) *Emphatic Stress*, amplifying or diminishing intensity, often with emotional or scalar meaning (Szwedek, 1986); (iii) *New Information Stress*, signaling the introduction of novel or unexpected content (Bolinger, 1958; Szwedek, 1986); and (iv) *Focus Stress*, highlighting the most relevant element for discourse (Szwedek, 1986; Ladd, 2008). Examples can be see in Table 1.

Realization of sentence stress in speech. A complementary line of research has focused on how stress is realized in a speech signal. These studies investigate the acoustic aspects that contribute to the production and perception of stress. The most widely agreed-upon prosodic features associated with stress are pitch (f0), loudness, duration and vowel quality (Van Heuven, 2018; Silipo and Greenberg, 2000).

3 Benchmarking stress understanding

In benchmarking stress understanding in SLMs, we focus on two tasks: (i) *Sentence Stress Reasoning* (SSR), evaluates whether an SLM can infer the speaker's intended meaning from the speech alone. This task goes beyond recognizing what was said, it requires understanding how it was said, focusing on how stress influences the interpretation of an utterance; (ii) *Sentence Stress Detection* (SSD), evaluates the model's ability to identify stressed word(s), given the ground truth transcription. This simpler task provides a controlled assessment of the model's sensitivity to acoustic prominence. Note that, in contrast to SSR which is a novel task, SSD is established in literature. We include it as a complementary task to SSR.

3.1 Datasets

StressTest is a dataset of texts recorded in two (or more) stress patterns, implying different meanings. It is recorded by a single professional actor. Exact details about the number of recordings per text, and recording setup can be found in Appendix A.

As not all texts can naturally accommodate multiple stress patterns, text sentences have to be collected specifically for this purpose. Therefore, text samples were curated through a rigorous manual process. Annotators who are fluent in English were instructed to suggest sentences that could have at least two different meanings based on a stressed word or words. We provide annotators with several examples from literature for inspiration, and they could use semi-automatic tools for proposing new examples (e.g. LLMs), but had to manually write and verify the samples. We also request they do not use exact templates of the examples, even if they use

Stress Type	Description	Stressed speech (intention)
Contrastive	Demonstrates contrast with another option.	*"I didn't take your book."* (vs. someone else) *"I didn't take your book."* (vs. something else)
Emphatic	Amplifies or diminishes the intensity of a concept.	*"They loved how you treated her dog."* (You really exceeded their expectations)
New-Information	Marks a surprising or novel content in the discourse.	*"He's actually moving to New York."* (surprising since its far from his current home)
Focus	General-purpose mechanism for highlighting key elements.	*"I enjoy the taste of espresso at sunrise."* (It's about that particular time)

Table 1: Examples of sentence stress categories from *StressTest*.

similar stress types. These samples were verified for agreement with at least one further annotator. In cases where the stressed words were agreed upon, but the meaning's text description needed refinement this was done by a third annotator. Finally, annotators were requested to mark samples with corrupted audio for removal. A breakdown of the various stress patterns in *StressTest*, along with their frequencies, is shown in Figure 3 in Appendix A.

StressPresso. We further extend our evaluation by annotating a subset of Expresso (Nguyen et al., 2023) — a real-speech stress-labeled dataset. We denote this subset as *StressPresso*. As Expresso is an expressive speech dataset, we consider neutral-emotion samples only, to decouple emotion from stress understanding. A major distinction from *StressTest* lies in the annotation process. *StressPresso* was annotated post-hoc, with annotators inferring underlying intentions from existing audio, whereas *StressTest* was annotated with an intended meaning and then recorded accordingly. The rest of the annotation procedure was the same. Another difference is that *StressPresso* has four speakers, two male and two female. Further data statistics can be found in Appendix A.

Importantly, *StressPresso* serves as a complementary evaluation set to the primary benchmark. This is due to inherent limitations of the post-hoc annotation process. This limitation motivated the design of *StressTest*, where recordings were explicitly produced to match intended meanings, ensuring alignment between speech and annotation. At the same time, *StressPresso* contains multiple speakers and diverse recording conditions, making it valuable for examining model generalization across voices and environments. As such, results on *StressPresso* should be recognized as complementary evidence of broader generalization beyond *StressTest*. We encourage readers to listen to examples of both datasets in the project page.

Overall, we define each dataset as $\mathcal{D} = \{x_1, ..., x_n\}$ and each sample $x \in \mathcal{D}$ consists of: $x = (a, t, s, A, l)$, where: (i) t is the transcription; (ii) s is a binary vector of the stressed words of the transcription t, where 1 marks a word as stressed; (iii) A is a set of two possible interpretations of the transcription t, both possible given a different stress pattern; (iv) $l \in A$ marks the correct interpretation; and finally, (v) a is the speech sample.

3.2 Evaluation procedure

Given a sample x, we evaluate SLMs' ability to perform stress reasoning and stress detection by constructing task-specific prompts. The full set of evaluation prompts used is detailed in Appendix D.

For the sentence stress reasoning (SSR) task, the model is provided with the audio and is instructed to select the most likely meaning from a set of possible interpretations. Performance is measured by accuracy. As prompt adherence varies notably across models, we follow the LLM-as-a-judge (Gu et al., 2025) approach, using gpt-4o (Hurst et al., 2024), to interpret model's output. This judgment is then used to compute the final evaluation metric consistently for all evaluated models. Given a model $\mathcal{M}$, dataset $\mathcal{D}$, a prompt $\mathcal{P}$ and a judge $\mathcal{J}$ the SSR accuracy is measured as follows:

$$\text{SSR}(\mathcal{M}, \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \mathbb{I}\{\mathcal{J}(\mathcal{M}(a, \mathcal{P}(A))) = l\}.$$

For sentence stress detection (SSD), the model receives both the audio a and the ground truth transcription t and is instructed to identify stressed word(s) in the utterance. Then, a judge interprets the model's output and we compute precision, recall, and F1 scores based on the predicted stressed words.

Category	Model	*Trans.*	*Stress*	*Audio*	SSR ↑	
					StressTest	*StressPresso*
Text LLM	gpt-4o (Hurst et al., 2024)	✓	✓	✗	**86.2**	**83.6**
	gpt-4o-mini (Hurst et al., 2024)	✓	✓	✗	79.3	80.1
	Llama-3.1-8B-Instruct (Grattafiori et al., 2024)	✓	✓	✗	73.3	81.1
	Qwen2-7B-Instruct (Yang et al., 2024a)	✓	✓	✗	67.8	74.2
	Qwen-7B-Chat (Bai and et al., 2023)	✓	✓	✗	61.4	64.8
Cascade Pipeline	WhiStress → gpt-4o	✗	✗	✓	**82.5**	**79.7**
	WhiStress → gpt-4o-mini	✗	✗	✓	76.6	73.2
	WhiStress → Llama-3.1-8B-Instruct	✗	✗	✓	66.9	76.2
	WhiStress → Qwen2-7B-Instruct	✗	✗	✓	63.7	69.3
	WhiStress → Qwen-7B-Chat	✗	✗	✓	55.5	62.3
SLM	Gemini-2.5-Pro (Comanici et al., 2025)	✗	✗	✓	77.9	72.7
	gpt-4o-audio (Hurst et al., 2024)	✗	✗	✓	58.7	64.8
	Qwen2Audio-7B-Instruct (Chu et al., 2024)	✗	✗	✓	56.4	51.4
	SALMONN (Tang et al., 2024)	✗	✗	✓	56.8	52.4
	LLaMA-Omni (Fang et al., 2025)	✗	✗	✓	53.6	52.9
	Phi-4-multimodal-instruct (Microsoft et al., 2025)	✗	✗	✓	53.2	50.4
	StressSLM (ours)	✗	✗	✓	**86.2**	**87.6**
Human	Annotators	✗	✗	✓	**92.6 (96.0)**	**89.6 (96.0)**

Table 2: SSR performance on *StressTest* with different inputs - ground truth transcriptions, stress labels, and the raw audio. We also report human performance and majority vote of three annotators.

3.3 Benchmarking results

Equipped with a method to evaluate sentence stress modeling, we benchmark leading SLMs on *StressTest* and *StressPresso*. We consider Qwen2Audio-7B-Instruct (Chu et al., 2024), SALMONN (Tang et al., 2024), LLaMA-Omni (Fang et al., 2025), Phi-4-multimodal-instruct (Microsoft et al., 2025), gpt-4o-audio (Hurst et al., 2024) and Gemini-2.5-Pro (Comanici et al., 2025). In addition, we conduct a human study to validate that our core SSR task is relatively easy and straightforward for human listeners. We randomly sample 100 samples from each dataset, sufficient to estimate human-level performance. We ask annotators to answer the same multiple-choice questions. Each sample is independently labeled by three annotators. We report both overall accuracy and majority-vote accuracy of three annotators in Table 2 (bottom cell). The annotation system and protocol can be viewed in Appendix A, Figure 4.

Results suggest that leading SLMs struggle to infer the intended meaning conveyed through stress patterns, most achieving near-random performance, while Gemini-2.5-Pro achieves relatively strong results of 77.9 and 72.7 on *StressTest* and *StressPresso* respectively. In contrast, human annotators achieve near-perfect scores, with 96.0% majority-vote accuracy and overall accuracies of 92.6% on *StressTest* and 89.6% on *StressPresso*.

4 Synthetic data generation

In order to improve model performance on sentence stress modeling and reasoning, we present a synthetic data generation pipeline. With it, we create Stress-17k, a training set aimed at enhancing performance on sentence stress understanding tasks. The main premise behind this approach is that once data is generated with sufficient diversity and quality, finetuning SLMs on this dataset will generalize to real-world recordings. The data generation methodology is illustrated in Figure 2 and is divided into four components: (i) Text sample generation; (ii) Stressed speech synthesis; (iii) Stress verification; (iv) Training tasks definition. Overall, the training set amounts to ∼17K audio samples, of which ∼4.5K are automatically verified. Additional data statistics are available in Appendix B.1.

Text sample generation. We first generate texts, stress patterns, and their interpretations. Note that, as discussed in Section 3, not all texts can have different meanings based on emphasis, thus we explicitly create them to that end. The texts are generated through a sequential agentic process, with gpt-4o (Hurst et al., 2024) as the agent, using the CrewAI framework (Moura et al., 2023).

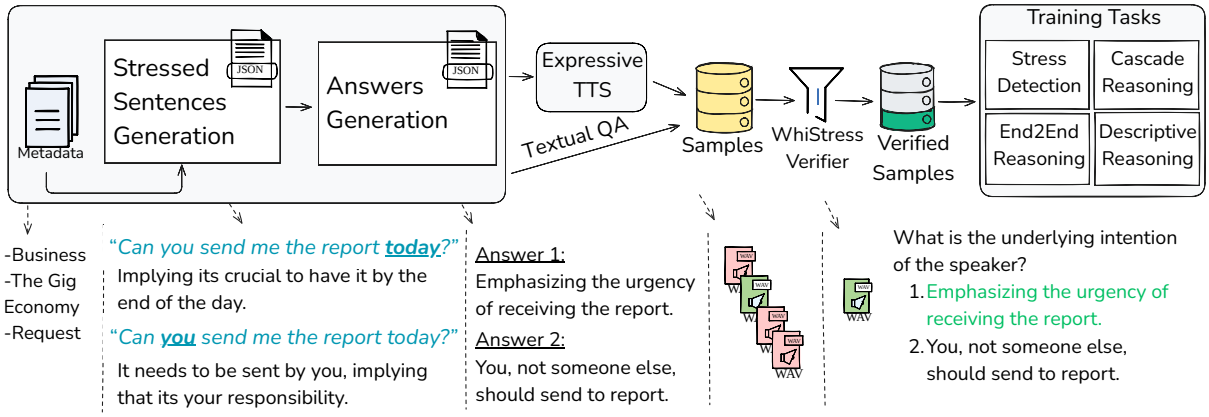

Figure 2: An illustrative example of the synthetic training data generation process.

This sequential process is comprised of two parts: (i) Prompt the agent to create a sentence that can be understood differently according to the stressed words. In order to increase diversity, and avoid repetitions, the instruction is given a domain, topic and sentence type from a list. Following notations from Section 3.1, after this phase we have two samples with the same transcription: $x_1 = (t, s_1, d_1), x_2 = (t, s_2, d_2)$ where for $i \in \{1, 2\}$, d_i describes the underlying meaning of each interpretation implied by the stress, s_i; (ii) As d can be lengthy, we aim to shorten it. Given the input from (i), we prompt the agent to create a set of answers A, each summarizing the corresponding interpretation's description d. Hence, $h_i \in A$ can seen as a concise version of d_i. Then, we end with two samples of the form $x = (t, s, d, A, l)$ where $l \in A$ is the target answer out of two possible answers corresponding to t. By that, we can ask a model what is the speaker's intention while providing two feasible answers. Prompts and metadata can be seen in Appendix D.5.

Stressed speech synthesis. Given text samples with known words to stress to convey a desired meaning, we use the OpenAI text-to-speech (OpenAI, 2023) to generate stressed speech. We find that marking the stressed words with enclosing asterisks leads to them being synthesized as stressed. Our preliminary results suggest, this approach leads to more natural voice compared to editing prosodic features directly. For each text stress pattern, we generate two audio samples using randomly selected male or female speakers. This results in four audio samples per transcription t: two reflecting one stress pattern and two reflecting a different one. Finally, each sample is of the form $x = (a, t, s, d, A, l)$ where a marks the synthesized speech.

Stress verification. Despite using expressive TTS, our studies reveal frequent stress errors, such as misplaced or missing emphasis. To address this, we use WhiStress (Yosha et al., 2025), which predicts transcriptions and stressed words, allowing us to filter out incorrectly stressed samples. This yields a higher-quality subset (see Appendix B.2) and improves performance, particularly under a curriculum setup (see Section 5.5).

Training task definition. We aim to improve SLM performance on SSR and SSD. To this end, each sample $x = (a, t, s, d, A, l)$ is utilized in four tasks, with a corresponding prompt: (i) *Sentence stress detection*: The model is requested to identify the stressed words s, given the ground truth transcription t; (ii) *End-to-end reasoning*: The model is asked to choose the most likely underlying meaning out of A according to the audio, directly responding the correct answer $l \in A$; (iii) *Elaborate reasoning*: Similar to the above, this task aims to train the model on SSR. However, the model is instructed to first elaborate on the interpretation's meaning and then choose the most likely answer. In practice, we use the description d as the model's elaboration prefix before the final answer l; (iv) *Cascade reasoning*: This task targets both SSD and SSR by requesting the model to output the transcription t with the emphasized words s and then output the correct answer l. We hypothesize that this variation allows the model to better connect sentence stress and implied underlying meaning. Notice, we copute the loss with regards to the entire answer (including elaborations). The training prompts are given in Appendix D.4.

Model	Expresso			*StressPresso*			*StressTest*		
	P.	*R.*	*F1*	*P.*	*R.*	*F1*	*P.*	*R.*	*F1*
WhiStress	**57.3**	**86.3**	**68.9**	**74.0**	**95.7**	**83.5**	**88.8**	**88.1**	**88.5**
Gemini-2.5-Pro	26.1	**82.1**	39.6	27.5	77.8	40.7	37.5	77.1	50.5
gpt-4o-audio	23.6	66.1	34.7	23.8	81.6	36.9	33.1	52.1	40.5
SALMONN	13.2	45.5	20.5	14.5	55.1	23.0	19.1	29.5	23.2
LLaMA-Omni	18.7	58.2	28.3	17.0	40.0	23.8	24.1	47.6	32.0
Phi-4-mm-instruct	22.5	37.5	28.2	18.6	42.9	26.0	19.9	32.8	24.7
Qwen2Audio-7B-Instruct	34.2	30.6	32.3	17.1	50.0	25.5	24.6	46.2	32.1
StresSLM (ours)	**51.8**	68.6	**59.1**	**77.1**	**84.4**	**80.6**	**89.8**	**84.5**	**87.0**

Table 3: Sentence stress detection (SSD) performance across three datasets. We compare the results of StresSLM to other SLMs, and also to WhiStress, which is trained solely for this task. We report Precision (P), Recall (R) and F1.

5 Experiments and results

We study the efficacy of our synthetic training data, Stress-17k, on SSR and SSD. We further conduct analysis to ablate the impact of pipeline components - stress verifier, trained encoder and training tasks.

We finetune Qwen2Audio-7B-Instruct (Chu et al., 2024) using LoRA adapters (Hu et al., 2021) on the query and value projections with the synthetic training dataset generated by our proposed pipeline. To prevent overfitting to stress-focused tasks, we also add samples from the original tasks which the base model was trained on, namely, Libri-Light (Kahn et al., 2020) for automatic-speech-recognition (ASR) and MELD (Poria et al., 2019) for speech-emotion-recognition (SER). We ensure that the total audio duration of these auxiliary tasks approximately matches our verified training subset.

During training, we employ a staged training approach, where the model is first finetuned on the full Stress-17k (both verified and unverified) for one epoch, then finetuned on a high-quality subset for another. We find this two-stage strategy effective in balancing the performance on both SSR and SSD tasks. Hyperparameters and implementation details are reported in Appendix B.3.

5.1 Sentence stress detection

One may argue that an essential prerequisite for understanding the meaning conveyed by sentence stress is the ability to accurately detect it. For this purpose, we start by evaluating the ability of SLMs on the SSD task. We evaluate models on *StressTest*, *StressPresso* and Expresso (Nguyen et al., 2023) benchmarks. To align with previous work (Yosha et al., 2025; de Seyssel et al., 2024) we use samples with at least one stressed word of speakers ex01 and ex02. Results in Table 3 show that current SLMs struggle with detecting stress, reaching at most F1 of 50.5. In contrast, StresSLM achieves F1 scores above 80.6 on our proposed datasets, and 59.1 on Expresso, marking a substantial improvement.

5.2 Sentence stress reasoning

We evaluate SSR performance across different input settings in both speech-aware and text-only LMs. Since SLMs are an extension of text LMs, we analyze the scenario where both transcription and stressed words are given as input. We consider three primary settings: (i) an oracle configuration in which both ground truth transcription and stressed words are given to an LM; (ii) a cascade pipeline in which these inputs are predicted by WhiStress; (iii) a fully end-to-end setting, where SLMs receive only the raw audio and directly predict the underlying meaning. Results are summarized in Table 2, and additional settings tested are available in Table 10.

The proposed model demonstrates strong results on SSR, exceeding evaluated SLMs and cascade models that receive audio only as input, 86.2, 87.6 vs. 82.5, 79.7 on *StressTest* and *StressPresso* respectively, when comparing to WhiStress followed by gpt-4o. This demonstrates a possible benefit to the direct approach vs a cascade, even against a much larger and stronger text model. It further shows the efficacy of the data synthesis pipeline.

While Gemini-2.5-Pro performs worse than StresSLM on both datasets, it is the only SLM to surpass 70 SSR. We attribute this to its high SSD recall (Table 3), which, combined with the two possible answers given as context, may help the model identify the likely stressed words. Moreover, as shown in Table 10, Gemini notably outperforms

Model	ASR (WER) ↓		SER ↑
	LS *clean / other*	CV *test*	MELD *test*
Qwen2Audio-7B	**1.73 / 4.01**	**8.72**	54.6
Qwen2Audio-7B-Inst.	2.31 / 4.92	11.49	26.4
StressSLM (ours)	2.29 / 4.59	10.49	**57.6**

Table 4: Results on core speech tasks. ASR shown on test split of LibriSpeech (LS) and CommonVoice (CV).

Verifier	# Samples	SSD ↑			SSR ↑
		P.	*R.*	*F1*	*Acc.*
✓	~4K	87.3	76.3	81.4	**79.3**
✗	~17K	87.4	81.9	84.5	76.6
✗→✓	17K→4K	**88.3**	**83.7**	**85.9**	78.4

Table 5: Effect of using WhiStress verifier on model performance on *StressTest* benchmark. We report Precision (P), Recall (R), F1, and Accuracy (Acc).

other LMs when given ground-truth stress labels, showing the strong text reasoning ability.

In assessing sentence stress understanding in text LLMs, we note, a somewhat expected, drop for cascade versions against the oracle stress labels on SSR, for instance 73.3 vs. 66.9 for the Llama3.1 version. This again highlights the error propagation in cascade approaches. Moreover, the results indicate that performance of text LLMs tends to correlate with the model's overall language capabilities—as reflected by their scores on standard text benchmarks—where open-source 7B models lag behind proprietary models like gpt-4o-mini and gpt-4o.

Finally, despite error propagation, the cascade approach outperforms current end-to-end SLMs (with the exception of our approach). This implies that stress information remains underutilized when derived directly from raw audio, thereby highlighting the effectiveness of our pipeline in extracting and leveraging prosodic cues for reasoning.

5.3 *StressTest* vs. *StressPresso*

Overall, results on both datasets show consistent trends, with StressSLM leading on both SSD and SSR (86.2 and 87.6) with Gemini second, and all open source models achieving near random SSR. This demonstrates that finetuning on our generated data not only enhances sentence stress understanding on *StressTest*, but also enables the model to generalize to additional speakers and recording conditions in *StressPresso*.

5.4 Effect on original tasks

To assess whether our approach introduces trade-offs with existing capabilities, we evaluate StressSLM on original tasks used to train the base model. Specifically, we test ASR, over LibriSpeech test sets and CommonVoice, and SER using MELD. This analyzes if our stress-focused training interferes with the SLM's abilities on basic speech understanding tasks. Results are reported in Table 4. We observe no degradation in ASR or SER compared to the Qwen2Audio-7B-Instruct model, indicating that the model retains its speech recognition ability. These findings suggest that our multitask training setup does not inherently conflict with the original objectives, demonstrating the potential to enrich SLMs with sentence stress reasoning capabilities, without compromising on established tasks.

5.5 Ablation study

We analyze the impact of our design choices in training StressSLM, on the performance on *StressTest*. Models are trained on Stress-17k without the ASR and SER rehearsal data. Additionally, all models are trained under the same conditions with a compute budget equivalent to 5 epochs on the verified data subset, the number of steps is smaller when removing training tasks. Further results on *StressPresso*, in Appendix C.2, show similar trends.

Staged training introduces balance. We first evaluate the efficacy of the WhiStress verifier. We train models on all stress-related tasks using a frozen speech encoder and assess performance under different training strategies. Results, presented in Table 5, show that training on the verified subset improves performance on SSR, suggesting that the data selected by WhiStress is beneficial for stress reasoning. However, this comes at the cost of reduced performance on the SSD task, likely due to reduced data diversity or quantity. In contrast, the staged training approach, in which the model is first trained on the full dataset and then fine-tuned on the filtered subset, achieves a better balance, improving SSD performance while slightly reducing SSR.

Speech encoder captures stress. We assess whether stress related information must be explicitly extracted from the audio encoder of StressSLM, or if the model can rely solely on frozen speech representations from the original model. As shown in Table 6, the model benefits from training the

Train Enc.	*Stress Det.*	*Cas. Reas.*	*Elab. Reas.*	SSD ↑ *P.*	*R.*	*F1*	SSR ↑ *Acc.*
✗	✓	✓	✓	88.3	83.7	85.9	78.4
✓	✗	✓	✓	45.5	86.3	59.9	**84.4**
✓	✓	✗	✓	90.4	83.3	86.7	82.5
✓	✓	✓	✗	**94.4**	**87.0**	**90.5**	78.4
✓	✓	✓	✓	92.4	86.3	89.3	83.0

Table 6: Ablating different training choices of StresSLM on stress modeling capabilities, using staged-training. We report Precision (P), Recall (R), F1, and Accuracy (Acc) for encoder training, stress detection, cascade reasoning, and elaborated reasoning.

encoder on both SSR and SSD tasks. This finding is consistent with previous works (Yosha et al., 2025; Pasad et al., 2022) which demonstrated that prosodic features are encoded in different layers of speech representation models, thus, finetuning the encoder enables extracting this stress-related information. Moreover, the results indicate that SLMs finetuned on the SSD task can outperform existing stress detection models such as WhiStress. As shown in Table 3, WhiStress achieves F1 scores of 88.5 on *StressTest*, while StresSLM that was trained solely on stress detection-oriented tasks achieves higher F1 scores of 90.5 (Table 6).

Training tasks balance each other. We analyze the effect of the different tasks introduced in Section 4. To do so, we conduct an ablation by removing one training task at a time (always retaining the end-to-end reasoning task). Results shown in Table 6, show that no combination yields the best performance for all metrics. However, including all tasks produces second-best results for both SSR and SSD, suggesting that a diverse task mixture leads to more balanced training. Interestingly, removing the elaborated explanation task results in a substantial drop in SSR performance, while improving SSD results. This may be due to SSD-oriented data becoming more dominant in the training data, indicating a trade-off in task emphasis.

6 Related work

Sentence stress modeling. Existing work define sentence stress detection as word-level binary classification. Approaches aim to directly detect stress from speech (Silipo and Greenberg, 2000; Mishra et al., 2012) or also include grammar and contextual information (Lin et al., 2020; Lee et al., 2016). Many modern methods use speech representations from pre-trained models (de Seyssel et al., 2024; Yosha et al., 2025). Another effort is developing expressive TTS models that can generate emphasized speech (Stephenson et al., 2022). Other work explore if text LLMs understand the meaning of emphasized words in discourse (Lin and Lee, 2024).

Speech-aware LMs. Recently, integrating speech and audio to LLMs has demonstrated impressive abilities on speech tasks (Arora et al., 2025). These models follow a similar approach - use a pretrained LLM, encode speech to a latent representation with a pretrained encoder, and project it to the LLM embedding space. Qwen2Audio use a Whisper encoder (Radford et al., 2022) and Qwen-7B (Bai and et al., 2023). Other models use additional encoders (Tang et al., 2024), introduce the ability to generate speech (Fang et al., 2025; Xie and Wu, 2024), or add modalities (Microsoft et al., 2025).

Related benchmarks. SLM evaluation spans diverse tasks, e.g. audio question answering (Lipping et al., 2022), speech-to-text-translation (Wang et al., 2020). Several evaluation suites cover many tasks, including prosodic elements (Yang et al., 2024b; Wang et al., 2025; wen Yang et al., 2021; Maimon et al., 2025b; Ma et al., 2025). For stress detection, a few benchmarks have been proposed, relying on synthetic data or existing emotion datasets (Yosha et al., 2025; de Seyssel et al., 2024). Another work includes crowd-sourcing to collect stress annotations in spoken utterances (Morrison et al., 2023). Recently, de Seyssel et al. (2024) introduced a framework to assess if spoken translation models preserve stress. Despite this, there remains a lack of benchmarks to evaluate models' ability to reason about meaning of spoken stress.

7 Conclusion

We introduce *StressTest*, a benchmark for evaluating stress understanding in SLMs, considering two tasks: Sentence Stress Reasoning and Detection. These tasks highlight the important yet underexplored role of stress placement in shaping meaning in spoken language, particularly in the context of speech-aware LMs, exposing a clear gap in existing models. We then propose a novel automatic data synthesis pipeline tailored to stress modeling, by which creating Stress-17k. Finetuning on Stress-17k allows our model, StresSLM, to significantly surpass existing SLMs on both SSR and SSD, while preserving performance on core tasks.

Limitations

Despite these contributions, our work is currently limited to English and limited speakers. How well stress-based reasoning generalizes across languages and accents remains an open question. In addition, our proposed benchmark is framed as a binary classification task. While this setup is simple and efficient, extensions to multiple-choice or open-ended question formats could capture more complex and diverse interpretations, though open-ended evaluation also introduces challenges of its own. Nonetheless, our findings highlight the value of prosody in speech-language modeling and suggest promising directions for future work in speech understanding and generation that better reflect the nuances of human communication.

References

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. *Preprint*, arXiv:2504.08528.

Jinze Bai and Shuai Bai et al. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

Steven E. Boer. 1979. Meaning and contrastive stress. *The Philosophical Review*, 88(2):263–298.

Dwight Bolinger. 1972. Accent is predictable (if you're a mind-reader). *Language*, 48(3):633–644.

Dwight L. Bolinger. 1958. Stress and information. *American Speech*, 33(1):5–20.

Dwight L. Bolinger. 1961. Contrastive accent and contrastive stress. *Language*, 37(1):83–96.

Noam Chomsky and Morris Halle. 1968. The sound pattern of english.

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.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, and Asaf Aharoni. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Maureen de Seyssel, Antony D'Avirro, Adina Williams, and Emmanuel Dupoux. 2024. Emphassess : a prosodic benchmark on assessing emphasis transfer in speech-to-speech models. *Preprint*, arXiv:2312.14069.

Fred I. Dretske. 1972. Contrastive statements. *The Philosophical Review*, 81(4):411–437.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. 2025. Llama-omni: Seamless speech interaction with large language models. *Preprint*, arXiv:2409.06666.

Sreyan Ghosh, Zhifeng Kong, Sonal Kumar, S Sakshi, Jaehyeon Kim, Wei Ping, Rafael Valle, Dinesh Manocha, and Bryan Catanzaro. 2025. Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities. *Preprint*, arXiv:2503.03983.

Aaron Grattafiori and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Shengpeng Ji, Yifu Chen, Minghui Fang, Jialong Zuo, Jingyu Lu, Hanting Wang, Ziyue Jiang, Long Zhou, Shujie Liu, Xize Cheng, and 1 others. 2024. Wavchat: A survey of spoken dialogue models. *arXiv preprint arXiv:2411.13577*.

J. Kahn, M. Riviere, W. Zheng, E. Kharitonov, Q. Xu, P.E. Mazare, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux. 2020. Libri-light: A benchmark for asr with limited or no supervision. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 7669–7673. IEEE.

Damjan Kalajdzievski. 2023. A rank stabilization scaling factor for fine-tuning with lora. *Preprint*, arXiv:2312.03732.

D. Robert Ladd. 2008. *Intonational Phonology*, 2 edition. Cambridge Studies in Linguistics. Cambridge University Press.

Gary Lee, Ho-Young Lee, Jieun Song, Byeongchang Kim, Sechun Kang, Jinsik Lee, and Hyosung Hwang. 2016. Automatic sentence stress feedback for non-native english learners. *Computer Speech & Language*, 41.

Binghuai Lin, Liyuan Wang, Xiaoli Feng, and Jinsong Zhang. 2020. Joint detection of sentence stress and phrase boundary for prosody. In *Interspeech*.

Guan-Ting Lin and Hung-yi Lee. 2024. Can LLMs understand the implication of emphasized sentences in dialogue? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13391–13401, Miami, Florida, USA. Association for Computational Linguistics.

Samuel Lipping, Parthasaarathy Sudarsanam, Konstantinos Drossos, and Tuomas Virtanen. 2022. Clotho-aqa: A crowdsourced dataset for audio question answering. *Preprint*, arXiv:2204.09634.

Chengqian Ma, Wei Tao, and Yiwen Guo. 2025. C3: A bilingual benchmark for spoken dialogue models exploring challenges in complex conversations. *Preprint*, arXiv:2507.22968.

Gallil Maimon, Michael Hassid, Amit Roth, and Yossi Adi. 2025a. Scaling analysis of interleaved speech-text language models. *arXiv preprint arXiv:2504.02398*.

Gallil Maimon, Amit Roth, and Yossi Adi. 2025b. Salmon: A suite for acoustic language model evaluation. *Preprint*, arXiv:2409.07437.

Microsoft, :, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, and 57 others. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *Preprint*, arXiv:2503.01743.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2025. Large language models: A survey. *Preprint*, arXiv:2402.06196.

Taniya Mishra, Vivek Rangarajan Sridhar, and Alistair Conkie. 2012. Word prominence detection using robust yet simple prosodic features. In *Interspeech 2012*, pages 1864–1867.

Max Morrison, Pranav Pawar, Nathan Pruyne, Jennifer Cole, and Bryan Pardo. 2023. Crowdsourced and automatic speech prominence estimation. *Preprint*, arXiv:2310.08464.

João Moura, Brandon Hancock, and CrewAI Developers. 2023. CrewAI: Fast and Flexible Multi-Agent Automation Framework. https://www.crewai.com.

Tu Anh Nguyen, Wei-Ning Hsu, Antony D'Avirro, Bowen Shi, Itai Gat, Maryam Fazel-Zarani, Tal Remez, Jade Copet, Gabriel Synnaeve, Michael Hassid, Felix Kreuk, Yossi Adi, and Emmanuel Dupoux. 2023. Expresso: A benchmark and analysis of discrete expressive speech resynthesis. *Preprint*, arXiv:2308.05725.

OpenAI. 2023. Text-to-Speech API. https://platform.openai.com/docs/guides/text-to-speech.

Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. 2022. Layer-wise analysis of a self-supervised speech representation model. *Preprint*, arXiv:2107.04734.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *Preprint*, arXiv:1810.02508.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Rosaria Silipo and Steven Greenberg. 2000. Prosodic stress revisited: Reassessing the role of fundamental frequency.

Brooke Stephenson, Laurent Besacier, Laurent Girin, and Thomas Hueber. 2022. Bert, can he predict contrastive focus? predicting and controlling prominence in neural tts using a language model. *Preprint*, arXiv:2207.01718.

Aleksander Szwedek. 1986. *A Linguistic Analysis of Sentence Stress*. [Gunter Narr Verlag, Tubingen].

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2024. Salmonn: Towards generic hearing abilities for large language models. *Preprint*, arXiv:2310.13289.

Vincent Van Heuven. 2018. *Acoustic Correlates and Perceptual Cues of Word and Sentence Stress: Theories, Methods and Data*, pages 15–59.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F. Chen. 2025. Audiobench: A universal benchmark for audio large language models. *Preprint*, arXiv:2406.16020.

Changhan Wang, Anne Wu, and Juan Pino. 2020. Covost 2 and massively multilingual speech-to-text translation. *Preprint*, arXiv:2007.10310.

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, Ko tik 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. *Preprint*, arXiv:2105.01051.

Deirdre Wilson and Tim Wharton. 2006. Relevance and prosody. *Journal of Pragmatics*, 38(10):1559–1579. Special Issue: Prosody and Pragmatics.

Zhifei Xie and Changqiao Wu. 2024. Mini-omni: Language models can hear, talk while thinking in streaming. *Preprint*, arXiv:2408.16725.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024a. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and Jingren Zhou. 2024b. Air-bench: Benchmarking large audio-language models via generative comprehension. *Preprint*, arXiv:2402.07729.

Iddo Yosha, Dorin Shteyman, and Yossi Adi. 2025. Whistress: Enriching transcriptions with sentence stress detection. *Preprint*, arXiv:2505.19103.

A Benchmark information

StressTest comprises 101 unique texts, each recorded with at least two distinct sentence stress patterns, yielding different underlying interpretations of the same utterance. Specifically, 85 sentences have 2 different interpretations, while 16 have 3. Of the interpretations - 170 have a single stressed word, 43 contain 2 stressed words and 5 include 3 distinct stressed words. In total, the number of audio samples is 218. All audio recordings are sampled of 48kHz. To match SLMs requirements, audio recordings were down-sampled to 16kHz. All recordings were approved by the Ethics committee at the university the dataset were recorded, where the professional actor get paid more than 5 times the base salary.

StressPresso contains 202 audio samples, each with a unique stress pattern, derived from 96 distinct texts. Of these, 113 are female recordings (51 from speaker ex02 and 62 from ex04) and 89 are male recordings (31 from ex01 and 58 from ex03).

Recording information. All audio samples were recorded using Shure-SM7 microphone over RME-800 sound card in a professional and acoustically treated studio. All samples were recorded in English by a professional actor, who got paid more than 4 times the minimum wage.

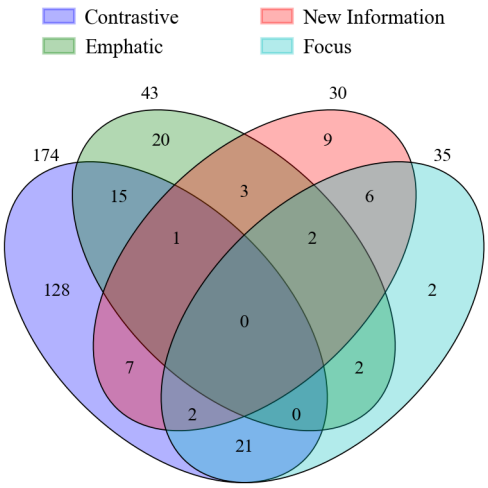

Figure 3: Categorization of sentence stress types in *StressTest*.

Types of sentence stress in *StressTest*. Based on the stress types presented in Section 2, we coarsely categorize and illustrate the different types of sentence stress represented in the *StressTest* samples in Figure 3. As definitions of sentence stress are fluid and not strictly defined, categorization inaccuracies may be present.

Human evaluation. Human annotators where requested to fill forms with 15 samples to evaluate. The form is illustrated in Figure 4.

Statistic	Full	Verified
# Samples (audio)	4400	1311
# Unique Interpretations	2200	931
# Unique Transcriptions	1100	731
# Trans. with ≥2 Interp.	1100	200
Gender Distribution		
Female	2122	597
Male	2278	714

Table 7: Stress-17k statistics before and after verifying with WhiStress.

B Training data and training procedure

B.1 Stress-17k information

Table 7 features statistics about the synthetically generated dataset before and after the use of the WhiStress verifier. The number of samples in the table is multiplied by the number of task templates presented in Section 4 resulting in ~17K samples.

B.2 Stress-17k data validation

We conducted a small-scale human study to validate the synthetic data. We asses the TTS model's ability to generate stressed speech, alignment between stressed speech and textual explanations and the WhiStress verifier effectiveness.

We use SSD as a proxy for measuring stressed TTS quality, and SSR for alignment of the text and TTS generated audio. We randomly sampled 20 recordings from the WhiStress verified subset, and 20 samples from the non-verified subset. Then, each synthetic sample was annotated by 3 different annotators, completing both SSD and SSR tasks. The SSD and SSR are calculated based on the desired TTS stress label vs the annotator stress marking as the target answer. For SSR, we report accuracy across all annotated samples, 60 per subset, as well as majority vote accuracy between the 3 annotators on the 20 samples of each subset. In the SSD task, we report mean scores for 3 annotators where each score was calculated on 20 samples. Results are summarized in Table 8.

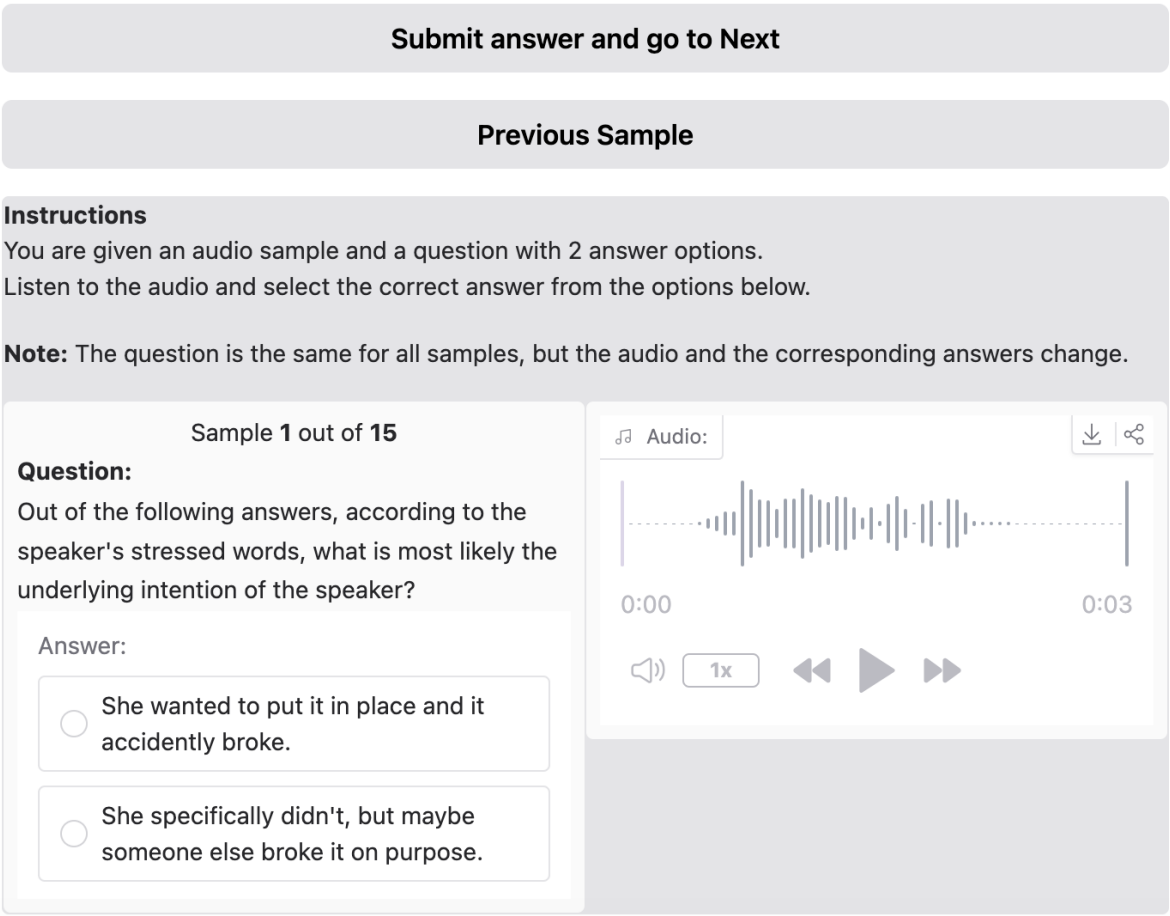

Figure 4: Human evaluation annotation view.

Subset	SSD ↑			SSR ↑
	P.	*R.*	*F1*	*Acc.*
Verified	**72.30**	**86.67**	**78.81**	**83.33** (**85.0**)
Non-verified	47.12	55.35	50.75	63.33 (70.0)

Table 8: Human evaluation results comparing verified and non-verified Stress-17k subsets. SSR accuracy is reported with majority vote in parentheses.

Verified data captures stress better. Results suggest that the verified subset adheres to the stressed synthetic speech more than the non-verified subset, achieving F1 score of 78.81% v.s 50.75% respectively. Additionally, in terms of alignment between the stress and the underlying intention presented in the textual answers of the SSR task, the verified subset results outperform those of the non-verified, with 85% compared to 70%. This shows the efficacy of the WhiStress verifier to capture stress and create a higher quality subset of the training data.

Non-verified data contains meaningful signals. Results indicate that valuable positive signal is present in the non-verified training data for detecting and reasoning about stress as well. This validates the ability of the training data to address our proposed tasks and supports our motivation to first train on large, noisy, data, and then continue training on a smaller but higher quality filtered data.

B.3 Training procedure

During training, gradients are computed only over the model's final answers. Hyperparameters for all models were chosen using grid search, taking the model to reach the best results on the validation set.

Final model. The final model was trained for 1595 steps using a cosine learning rate scheduler with a warm-up ratio of 5%. In the first 1261 steps, we use the entire data set (stage 1), followed by continued training in the verified subset for the remaining steps (stage 2), while preserving the internal state of the scheduler. The model was finetuned on a single NVIDIA L40S GPU. Training hyperparameters are summarized in Table 9.

Hyperparameter	Value
Learning Rate	7e-5
Batch Size (per device)	8
Gradient Accumulation Steps	2
LoRA Configuration (Hu et al., 2021)	
Rank r	16
LoRA α	32
Use rslora (Kalajdzievski, 2023)	True
Target Modules	`q,v proj`
LoRA Dropout	0.1

Table 9: Training hyperparameters.

Ablation models. When trained on the verified only and the full datasets the checkpoint reaching the best performance on the validation set is chosen, while in the staged training the best checkpoint in the second stage is considered.

C Additional results

C.1 Extended inputs

In this section we analyze the impact of performance on providing SLMs with the audio as well as ground truth stress and optionally the ground truth transcription. This can be seen as extending table 2, helping understand if the SLMs utilize the audio input, and how much of the error has to do with transcription performance. The prompts used for this analysis are provided in Section D.3 and the results are provided in Table 10.

We test two scenarios: (i) An oracle provides the SLM with ground truth transcription and ground truth stress in addition to the provided audio; and, (ii) An oracle provides solely the ground truth stress while relying on the ability of the SLM to recognize the spoken utterance from the audio.

Results suggest that generally, SLMs slightly improve their results the more information is provided to the model. An exception to this is SALMONN which shows better results when it is only provided stress as opposed to stress and transcription.

Our proposed model, StresSLM, that results with 86.2 and 87.6 on *StressTest* and *StressPresso* using the audio signal alone, on average (86.6) preserves its capabilities on SSR in both scenarios on both datasets. It also demonstrates improvement in the oracle scenario when both transcription and stress are provided, perhaps due to the slight decrease in transcription abilities shown in Table 4. Additionally, gpt-4o-audio, compared to its text-only counterpart in Table 2 that yields 86.2 and 83.6 on *StressTest* and *StressPresso* respectively, shows comparable performance when stress information is provided.

Notably, Gemini outperforms all other models by a wide margin when given ground-truth stress information. This, in contrast to its audio only performance that achieves less than 80.0 SSR. We attribute this gap to the poor precision in detecting stress as shown in Table 3.

Overall, these findings underscore the importance of precision and overall SSD performance for improving SSR.

C.2 Ablation study

We complement the ablation study results presented in Table 5 and Table 6 with results on *StressPresso* benchmark. Results in Table 11 and Table 12 show similar trends to the ablation study in Section 5.5 supporting our design choices.

D Prompts

D.1 Sentence stress reasoning evaluation

Stress reasoning task. The following prompt is the one used to query speech-aware language models to evaluate the performance on the proposed end-to-end sentence stress reasoning task.

Speech-Aware LM SSR Prompt

```
[audio]
Out of the following answers, according to the speaker's stressed words, what is most likely the underlying intention of the speaker?
1. [answer 1]
2. [answer 2]
Answer:
```

Reasoning LM judge. For the LLM-as-judge model we used the following prompt.

LLM-as-judge SSR Prompt

```
<system prompt>
You are a Speech-LM evaluator that helps evaluating models that have trouble in outputting a correct schema for an answer. You are very good at outputting the correct schema according to the instructions.
INSTRUCTIONS: Given a prompt with a
```

Model	Input			SSR ↑	
	Trans.	*Stress*	*Audio*	*StressTest*	*StressPresso*
SLM					
Gemini-2.5-Pro	✓	✓	✓	**98.1**	**96.5**
gpt-4o-audio	✓	✓	✓	87.1	85.6
SALMONN	✓	✓	✓	59.6	62.8
LLaMA-Omni	✓	✓	✓	70.1	71.2
Phi-4-multimodal-instruct	✓	✓	✓	55.5	56.9
Qwen2-Audio-7B-Instruct	✓	✓	✓	60.5	62.3
StresSLM (ours)	✓	✓	✓	83.4	92.5
Gemini-2.5-Pro	✗	✓	✓	**97.2**	**94.5**
gpt-4o-audio	✗	✓	✓	83.9	85.0
SALMONN	✗	✓	✓	66.5	69.3
LLaMA-Omni	✗	✓	✓	62.3	67.8
Phi-4-multimodal-instruct	✗	✓	✓	51.3	55.4
Qwen2-Audio-7B-Instruct	✗	✓	✓	59.1	65.3
StresSLM (ours)	✗	✓	✓	78.8	92.0
Human	✗	✗	✓	92.6 (96.0)	

Table 10: Performance on SSR the task of *StressTest* using different input configurations. The input column demonstrates the information provided to the model in addition to our task's question. This extends upon table 2.

Verifier	# Samples	SSD ↑			SSR ↑
		P.	*R.*	*F1*	*Acc.*
✓	~4K	**81.1**	83.4	82.3	85.5
✗	~17K	71.6	79.7	75.4	82.1
✗→ ✓	17K→4K	78.8	**86.3**	**82.4**	**85.6**

Table 11: Effect of using WhiStress verifier on model performance on *StressPresso*.

Train Enc.	*Stress Det.*	*Cas. Reas.*	*Elab. Exp.*	SSD ↑			SSR ↑
				P.	*R.*	*F1*	*Acc.*
✗	✓	✓	✓	78.8	86.3	82.4	85.6
✓	✗	✓	✓	31.5	**89.6**	46.6	87.6
✓	✓	✗	✓	71.7	82.5	76.7	**88.6**
✓	✓	✓	✗	**80.6**	88.2	**84.2**	88.1
✓	✓	✓	✓	79.9	87.2	83.3	**88.6**

Table 12: Ablating different training choices of StresSLM on stress modeling capabilities, using staged-training on *StressPresso*. We report Precision (P), Recall (R), F1, and Accuracy (Acc).

```
question and possible answers that the
Speech-LM received, and the output the
model emmited, you are required to
output the Speech-LM answer in a fixed
format.
* The output should be aligned with
the Speech-LM output, and should not
include any additional information or
context.
* The output should be a JSON object
with a single key "answer" and a value
that is the number of the correct
answer according to the output of the
Speech-LM.
* The answer should be an integer,
either 1 or 2.
EXAMPLE 1:
INPUT TO Speech-LM: According to the
intonation of the speaker, what is
more probable?
1. Yesterday, someone did not inform
the speaker about the meeting.
2. Someone did not inform the speaker
about the meeting that occurred
yesterday.
OUTPUT FROM Speech-LM: Someone did not
inform the speaker about the meeting
that occurred yesterday. Therefore,
option 2 is more probable than option
1.
YOUR EXPECTED JSON OUTPUT:
{"answer": 2}
```

EXAMPLE 2:
INPUT TO Speech-LM: Question: According to the intonation of the speaker, what is more probable? Possible answers:
1. Yesterday, someone did not inform the speaker about the meeting.
2. Someone did not inform the speaker about the meeting that occurred yesterday.
OUTPUT FROM Speech-LM: Answer: 1. Yesterday, someone did not inform the speaker about the meeting.
YOUR EXPECTED JSON OUTPUT: {"answer": 1}
<user prompt>
INPUT TO Speech-LM:
[input prompt]
OUTPUT FROM Speech-LM:
[speech lm output]
YOUR EXPECTED JSON OUTPUT:

D.2 Sentence stress detection evaluation

Stress detection task. We use the following prompt to asses the *SSD* task.

Speech-Aware LM SSD Prompt

[audio]
The speaker said "[transcription]".
According to the audio, what words did the speaker stress?
Answer format: [stressed_word_1, ...]
Answer:

Detection LM judge. For the LLM-as-judge model we used the following prompt.

LLM-as-judge SSD Prompt

<system prompt> You are a Speech-LM evaluator that helps evaluating models that have trouble in outputting a correct schema for an answer. You are very good at outputting the correct schema according to the instructions.
INSTRUCTIONS: Given a prompt with a question that the Speech-LM received, and the output the model emitted, you are required to output the Speech-LM answer in a fixed format.
* The output should be aligned with the Speech-LM output, and should not include any additional information or context.
* The output should be a JSON object with a single key "answer" and a value that is a list of words according to the output of the Speech-LM.
* The answer should be a list of strings.
* If the model mistakenly outputs two or more words as a single word, you should split them into separate words.
EXAMPLE:
INPUT TO Speech-LM:
The speaker said "What a lovely day we have".
According to the audio, what words did the speaker stress?
Answer format: [stressed_word_1, ...]
Answer:
OUTPUT FROM Speech-LM:
The speaker stressed: ["lovely", "we have"].
YOUR EXPECTED JSON OUTPUT:
{"answer": ["lovely", "we", "have"]}
<user prompt>
INPUT TO Speech-LM:
[input prompt]
OUTPUT FROM Speech-LM:
[speech lm output]
YOUR EXPECTED JSON OUTPUT:

D.3 Reasoning analysis prompts

Transcription and stress as input. The prompt evaluates whether additional context helps sentence stress understanding. In case the evaluated model is an LM, the audio placeholder is not used.

Sentence Stress Reasoning - stress and transcription input

[Audio]
Question:
Given that a speaker said: "[transcription]", and stressed the words: [stressed words].
Out of the following answers, what is most likely the underlying intention of the speaker?
Possible answers:
1. [answer 1]
2. [answer 2]

```
Answer:
```

Stress as input. The prompt evaluates whether only the stressed words helps sentence stress understanding, since ASR is a fundamental task for speech-aware LMs.

Sentence Stress Reasoning - stress input

```
[Audio]
Question:
Out of the following answers, given that the speaker stressed the words: [stressed words].
What is most likely the underlying intention of the speaker?
Possible answers:
1. [answer 1]
2. [answer 2]
Answer:
```

D.4 Training prompts

End-to-end task. The following prompt guide the model to precisely choose the speaker's intended meaning based on stressed words.

End to end reasoning

```
[Audio]
Out of the following answers, according to the speaker's stressed words, what is most likely the underlying intention of the speaker?
1. [answer 1]
2. [answer 2]
Answer:
```

Expected Answer Format

```
[answer label]. [correct answer]
```

Elaborated answer task. The model is required to first explain its reasoning and then answer.

Elaborated Answer Prompt

```
[Audio]
According to the speaker's stressed words, what is the speaker's underlying intention?
1. [answer 1]
2. [answer 2]
Elaborate, then answer in the following way: "answer_number. correct_answer"
```

Expected Answer Format

```
[description]. Therefore, the correct answer is: [answer label]. [correct answer]
```

Cascade reasoning task. This prompt encourages the model to reason based on the stressed words and transcription before answering.

Stress Detection Reasoning Prompt

```
[Audio]
The speaker stressed some words. What is the speaker trying to communicate?
1. [answer 1]
2. [answer 2]
Think about the transcription and the stressed words. Then, answer like this: "answer_number. correct_answer"
```

Expected Answer Format (Format 7)

```
The speaker said "[transcription]" and emphasized "[stressed words]". Therefore, the correct answer is: [answer label]. [correct answer]
```

Stress detection task. This prompt focuses only on detecting which words were emphasized.

Stress detection prompt

```
[Audio]
The speaker said "[transcription]". According to the audio, what words did the speaker stress?
Answer format: [stressed_word_1, ...]
Answer:
```

Expected Answer Format

```
[stressed words]
```

D.5 Data generation pipeline

Text sample generation. The agentic process used to create the textual examples will be open-sourced with the full yml configurations and code. **Metadata.** We use 10 sentence types: statement, question, command, exclamation, request, suggestion, invitation, offer, opinion and warning. Additionally, a domain and a topic are injected into the agent generation prompt out of a list of 32 domains and 110 topics corresponding to the domains, generated by gpt-4o.

Domains and topics

Art & Culture:
-The Renaissance Period
-Street Art and Graffiti
-The Influence of Jazz Music
-The Role of Film in Society
-Modern Architecture Trends
-Art as Political Protest
-The Influence of Surrealism
-Cultural Appropriation in Fashion
-Folk Music and Tradition
-The Globalization of Art Markets
-The Role of Museums in Preserving Culture
-The Influence of Classical Music on Modern Genres
-Digital Art and Its Rising Popularity
-The Evolution of Dance Across Cultures
-The Impact of the Internet on Art Consumption
Business:
-Startup Culture and Innovation
-Corporate Social Responsibility
-Leadership Styles and Management
-The Gig Economy
Economics:
-Income Inequality
-Global Trade Wars
-Cryptocurrency and Digital Economy
-The Future of Work and Automation
Education:
-Online Learning Platforms
-Special Education Needs
-The Role of Arts in Education
-STEM Education for Girls
Engineering:
-Renewable Energy Engineering
-Aerospace Innovations
-Civil Engineering and Urban Planning
-Robotics and Automation
Environment:
-Ocean Pollution and Marine Life
-Conservation of Endangered Species
-Urban Green Spaces
-Sustainable Agriculture Practices
Food & Culinary Arts:
-The Science of Baking
-Global Cuisine Trends
-Farm-to-Table Movement
-Veganism and Plant-Based Diets
Health & Medicine:
-Mental Health Awareness
-The Rise of Telemedicine
-Nutrition and Lifestyle Diseases
-Vaccine Development Processes
History:
-The Industrial Revolution
-Ancient Civilizations and Their Contributions
-World Wars and Their Consequences
-The History of Human Rights
Law:
-Intellectual Property Rights
-The Justice System and Reforms
-International Law and Human Rights
-Cyber Law and Internet Regulations
Literature:
-Dystopian Novels and Society
-The Impact of Shakespeare
-The Evolution of Poetry
-Modern Graphic Novels
Philosophy:
-Existentialism and Modern Thought
-The Ethics of Artificial Intelligence
-Eastern vs. Western Philosophical Traditions
-The Philosophy of Science
Politics:
-Globalization and Nationalism
-Election Systems and Voter Rights
-The Role of the United Nations
-Immigration Policies and Refugee Crisis
Psychology:
-Cognitive Behavioral Therapy
-The Psychology of Social Media
-Child Development Stages
-Emotional Intelligence in Leadership
-The Impact of Stress on Health
Religion:
-Comparative Religion Studies
-Secularism and Society
-Rituals and Traditions
-The Role of Religion in Politics
Science:
-CRISPR and Genetic Engineering
-Climate Change and Its Impact on Biodiversity
-Space Exploration and Mars Colonization
-Nanotechnology in Medicine
Sociology:
-Urbanization and Its Challenges
-The Dynamics of Family Structures
-Gender Roles in Society
-The Sociology of Religion
Sports:
-The Science of Sports Performance
-Gender Equality in Sports
Technology:
-Artificial Intelligence and Ethics
-Quantum Computing Advancements
Travel & Tourism:
-Eco-Tourism and Sustainability
-Cultural Heritage Sites
Anthropology:
-Cultural Practices of Hunter-Gatherer Societies
-The Evolution of Human Speech and Language
Astronomy:
-The Birth and Death of Stars
-The Formation of Black Holes
Cryptography:
-The Evolution of Cryptographic Algorithms
-Quantum Cryptography and Secure Communication
Fashion:
-The Rise of Fast Fashion and Its Environmental Impact
-Fashion Icons Who Changed History
Gaming:
-The Psychology of Video Game Addiction
-The Rise of Esports and Competitive Gaming
Geopolitics:
-The Role of Natural Resources in International Conflict
-Geopolitical Impacts of Climate Change
Mathematics:
-Chaos Theory and its Applications
-The Role of Mathematics in Predictive Modeling
Performing Arts:
-The Evolution of Contemporary Dance
-The Role of Theater in Political Activism
Photography:

```
-The Evolution of Documentary Photography
-The Role of Photography in Social Movements
 Space Exploration:
-The Privatization of Space Travel
-The Role of Satellites in Modern Communication
 Visual Arts:
-The Impact of Technology on Visual Arts
-The Role of Photography in Modern Art
 Urban Studies:
-The Role of Public Transportation in Urban
 Growth
-Urbanization and Its Impact on Housing
```

E Broader impact

As with any advancement in language models, our work carries the potential for both positive and unintended consequences. By equipping models with the ability to interpret the meaning conveyed through stress patterns in speech, we aim to foster more efficient and nuanced communication systems. Such capabilities may particularly benefit populations who rely heavily on prosody for meaning - such as individuals with hearing disabilities, language learners, etc and ultimately contribute to more accessible and context-aware AI systems.

F Usage of LLMs

As part of this research study we use LLMs for two main reasons: (i) Writing helper, e.g., improving writing style; (ii) LLM-as-a-Judge. For writing style we mainly use Chat-GPT, while for LLMs-as-a-Judge we use gpt4o.