

Article

Beyond Transcript Alignment: Diagnosing Paralinguistic Information Flow in Frozen Speech-to-LLM Adapters

Nurgali Kadyrbek *  and Madina Mansurova 

Department of AI & Big Data, Faculty of Information Technologies, Al-Farabi Kazakh National University, Almaty 050040, Kazakhstan; madina.mansurova@kaznu.edu.kz

* Correspondence: kadyrbek.nurgali@kaznu.kz

Abstract

Frozen speech-to-LLM systems train an adapter between a frozen audio encoder and a text LLM. We test whether adapters preserve sentence stress beyond transcripts and whether the LLM uses it. In a five-seed Qwen3-8B/WavLM baseline, transcript alignment lowers linear adapter Probe-K below the text-only K_T baseline (0.211 vs. 0.290). The R1.8 configuration improves MLP-2 Probe-K from 0.245 to 0.306 (paired $p = 0.012$) and passes three controls, but because warmup and augmentation also change, the gain is not attributed to $\mathcal{L}_{cf}$ alone; its crossing of the linear K_T floor is neither statistically established nor capacity-matched. Response Probe-G remains 0.512 in both cohorts; single-seed LoRA and styled-teacher pilots do not improve it. A post hoc trace through all 36 Qwen blocks finds persistent absolute stress decodability in speech-slot states, but no reliable R1.8-over-R0 advantage at any state and no speech-conditioned answer margin; an explicit text tag instead yields a +6.40-nat margin. Attention differences depend on slot-length normalization, and late left/system concentration is shared across modalities rather than speech-specific. Thus, the tested system remains an end-to-end negative: adapter decodability does not imply causal response use. Scope is limited to sentence stress, mostly synthetic voices, one encoder, and one LLM.

Keywords: speech-language models; frozen-frozen adapter training; paralinguistic preservation; counterfactual learning; sentence stress; LoRA adaptation

1. Introduction

Speech-language models (speech-LLMs) augment text large language models with audio understanding by inserting a continuous-prefix *adapter* between a speech encoder and the LLM. In the frozen–frozen regime (both encoder and LLM kept frozen, only the adapter trained), the dominant training recipe is *transcript-conditioned behavior alignment* [1–9]: the adapter is optimized to make the LLM produce the same response distribution under the audio prefix as it would under the corresponding text prefix.

A growing body of work questions whether this recipe yields a system that *listens* or merely *transcribes*. Billa [10] used logit-lens and LEACE diagnostics to show that the literal transcript emerges from a speech-LLM’s hidden states and that text representations are causally necessary; speech-LLMs in deployed configurations behave as cascades. Chen et al. [11] report consistent lexical dominance across six state-of-the-art audio-LLMs on cue-conflict emotion stimuli. Billa [12] measured a *Text Dominance Ratio* of 10–26 $\times$ on Gemini-2.0-Flash and GPT-4o under audio–text conflict: audio-LLMs follow conflicting text far more than they follow conflicting text-vs-text under otherwise identical conditions.



Academic Editor: Fabrizio Messina

Received: 4 May 2026

Revised: 27 July 2026

Accepted: 28 July 2026

Published: 30 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Yosha et al. [13] show that current speech-LLMs perform poorly on sentence-stress reasoning, a task where the transcript alone is genuinely insufficient.

This paper asks a sharper question: *Can the complete R1.8 training configuration improve stress decodability at the adapter representation, and if it can, does the frozen LLM consume the resulting signal at the response time?* We separate these questions empirically with two complementary probes (illustrated in Figure 1). **Probe-K** reads the adapter’s internal representation and asks whether stress information is preserved there; it is a representational test on the audio path. **Probe-G** reads the frozen LLM’s response and asks whether the LLM acts on that information when generating an answer; it is a consumer-side test on the audio-to-response path. We use both probes on same-transcript counterfactual stimuli (audio pairs that share text but differ in which word is stressed), run a four-part controlled sweep under a locked protocol, and triangulate against three independent falsification controls. Probe-K and Probe-G can dissociate: a high-Probe-K/low-Probe-G outcome means *the information is in the adapter but the LLM does not use it*. This dissociation does not imply that the adapter is lossless; rather, it bounds a response-level handoff problem for whatever residual paralinguistic signal survives adapter training. A post hoc internal analysis added in Stage 8 then follows this signal through every Qwen3-8B block. It finds no abrupt loss of absolute speech-slot decodability, but also no statistically reliable propagation of the R1.8-over-R0 advantage and no speech-conditioned answer margin.

The intervention trajectory is complete rather than prospective. Stages 2–4 establish adapter-side loss and partial representational recovery. Stage 6 then tests whether minimal query/value LoRA can make the LLM consume the residual signal; rank 4 preserves the oracle check but does not improve Probe-G, whereas rank 8 degrades that check. Because this single-seed pilot did not improve Probe-G, Stage 7 changes the alignment target instead of adding further LLM capacity: the teacher receives the transcript plus an explicit stress tag. This styled-teacher pilot also lands in the R0 Probe-G equivalence band (0.5158), while direct text-tag references reach 0.7901 and 0.8020. This study therefore ends with an end-to-end negative result with explicit scope rather than an untested Stage 7 proposal. Stage 8 sharpens that result: the explicit text-tag control becomes strongly decision-relevant late in Qwen3-8B, whereas none of the speech conditions do, and the attention audit does not support a speech-specific attention-sink account.

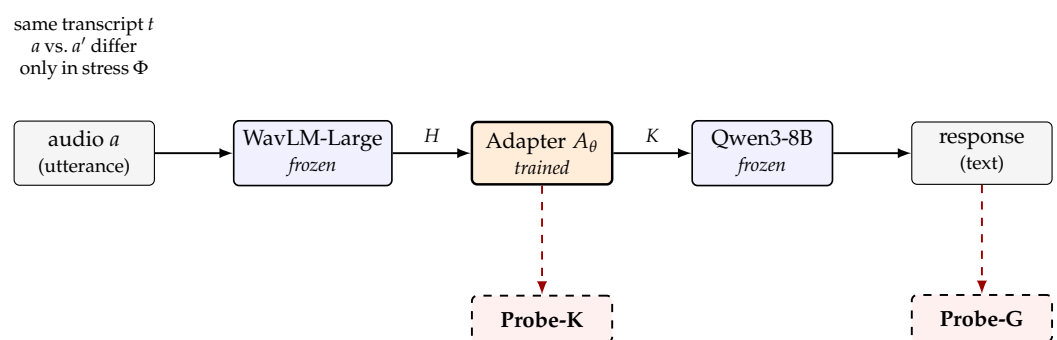


Figure 1. Probe-K/Probe-G diagnostic on the frozen–frozen speech-to-LLM pipeline. Frozen modules are blue, the trainable adapter A_θ is orange; and dashed red boxes are diagnostic probes. Solid arrows show the forward information flow, whereas dashed red arrows show the diagnostic readout paths. Two same-transcript audio inputs a, a' that differ only in which word is stressed are passed through the same pipeline. **Probe-K** (representational) reads the adapter output $K = A_\theta(H)$ and asks whether the stress label Φ is decodable there. **Probe-G** (consumer) reads the frozen LLM’s response and asks whether the answer changes in the correct direction. The two probes can dissociate: high-K/low-G bounds the failure to the path between the adapter output and the generated response for the residual signal present in K . Stage 8 extends this endpoint diagnostic with block-wise hidden-state and attention measurements.

1.1. Contributions

This paper makes five empirical contributions, each anchored in a stage of the sweep:

- **Transcript alignment can reduce paralinguistic signal.** A frozen–frozen BLSP-class baseline replicated over five seeds on Qwen3-8B-Instruct + WavLM-Large layer 16 produces a linear Probe-K of 0.211 ± 0.013 , below the K_T text-only baseline of 0.290 on every seed. Thus, in this task, transcript-conditioned behavior alignment removes stress-decodable structure that is still weakly present in the transcript representation.
- **The complete R1.8 training configuration improves nonlinear Probe-K, with important statistical and capacity bounds.** A correctness-aware single-token margin loss $\mathcal{L}_{cf}$ with $\lambda = 5$ matched to the $\mathcal{L}_{BLSP}$ gradient norm produces a 5-seed MLP-2 Probe-K mean of 0.306 ± 0.044 (3/5 seeds above the pre-specified linear K_T floor). The paired MLP-2 improvement over R0 is significant ($t(4) = 4.39$, $p = 0.012$), but the numerical crossing of the linear K_T floor is not significant as a one-sample seed test ($p = 0.460$) and is not a capacity-matched comparison to an MLP-2 text probe.
- **Response-level use does not improve in the tested configuration.** Probe- $G_{neutral}$ is 0.512 in both R0 and R1.8 at cohort precision (paired $p = 1.0$). Minimal LoRA on Qwen3-8B’s query and value projections W^Q/W^V (q_proj/v_proj in HuggingFace PEFT) and a styled-teacher intervention are reported as single-seed diagnostics: neither improves Probe-G under the evaluated settings, while K_T^{style} and Cascade-T+L show that the same LLM can consume the equivalent paralinguistic tag in direct text form (0.79+/0.80+ on neutral).
- **An internal trace separates decodability from model use.** A seed-matched, post hoc diagnostic follows stress through all 36 frozen Qwen3-8B blocks. Absolute stress remains decodable from speech-slot states, but the simultaneous R1.8-minus-R0 band excludes zero at no state, and the model’s final answer margins remain indistinguishable from zero for R0, R1.8, and Stage 7. The explicit text-tag control instead reaches a +6.40-nat margin. This rules out an abrupt decodability-erasure layer in the observed trajectories while leaving causal routing mechanisms unresolved.
- **Controls separate audio-derived signal from shortcuts.** The Stage 4 control suite (text-only adapter via fixed projection P ; audio-decorrelated cf-pairs; per-seed shortcut probes against a pre-trained raw-encoder baseline) checks whether the modest representational gain can be explained by optimizer regularization, transcript leakage, or TTS/domain shortcuts. Every control passes its locked threshold, supporting an audio-derived interpretation of the R1.8 MLP-2 gain.

1.2. Roadmap

Section 2 situates the work in speech-to-LLM adaptation, speech encoder probing, and classical acoustic modeling. Section 3 describes the fixed architecture and training configuration, datasets, counterfactual-pair construction, the R1.8 loss family, the Stage 4 falsification harness, and the Stage 8 internal trace. Section 4 reports per-stage findings, leading with the controls. Section 5 interprets the response-level gap and bounds the claim. Section 6 states the take-home and future-work axes.

2. Related Work

2.1. Frozen–Frozen Speech-to-LLM Adapters

The BLSP family [1–3] establishes behavior alignment as the dominant training objective in the frozen–frozen regime. Wang et al. [1] train a lightweight modality adapter so that the LLM emits the same continuation under audio and transcript inputs, the $\mathcal{L}_{task} + \mathcal{L}_{KL}$ objective our R0 cohort replicates. Wang et al. [2] extend this with a continuous integrate-and-fire alignment and a Partial-LoRA (PLoRA) variant on the LLM side; Wang et al. [3]

add a second emotion-alignment stage using SER labels. AlignFormer [4] achieves near-100% instruction-following with a CTC + dynamic-window QFormer adapter trained on ASR data. Wav2Prompt [5] uses CIF-based continuous prompts to outperform ASR-LLM cascades on speech translation. SLAM-ASR [9] demonstrates that a single trainable linear projector between frozen encoder and frozen LLM already reaches SOTA on LibriSpeech.

The most direct competitor on the same task is StresSLM (introduced in Yosha et al. [13] alongside the StressTest benchmark): a fine-tuned Qwen2Audio-7B-Instruct using LoRA on q_proj/v_proj *plus encoder training* on Stress-17K. StresSLM achieves 86.2% on SSR (vs. human ceiling 92.6%/96.0% majority-vote, and best baseline 77.9% Gemini-2.5-Pro) and 87.0 SSD F1 (vs. WhiStress specialist 88.5%). StresSLM does not test a frozen-encoder + frozen-LLM baseline. Our work fills that gap: Stage 2 R0 + Stage 3.8 R1.8 establish what the strictly frozen–frozen recipe achieves on the same data, and Stage 6 isolates the consumer-side LoRA contribution under the frozen-encoder + frozen-adapter constraint that StresSLM does not impose.

The DeSTA lineage [6–8] pursues a complementary path: *caption-based* alignment where descriptive speech captions teach the SLM to interpret both linguistic and non-linguistic features. In our Section 3 decomposition, caption-based supervision injects s_{para}^{vis} via the text target; our R1.8 configuration instead targets s_{para}^{res} using a counterfactual objective together with its specified schedule and augmentation pipeline. SALMONN [14] jointly trains the audio encoders with the LLM and is therefore outside of the frozen–frozen regime our paper studies.

2.2. Cascade-Equivalence Diagnosis

Billa [10] introduced *matched-backbone testing* and used logit-lens + LEACE to show that the literal transcript emerges from speech-LLM hidden states and that text representations are causally necessary. Performance advantages over ASR-LLM cascades reverse by up to 7.6% at 0 dB SNR. Billa [12] extended the same author’s investigation to multi-lingual modality arbitration; the *Text Dominance Ratio* on Gemini-2.0-Flash and GPT-4o is 10–26× higher than the text-vs-text-arbitration baseline, suggesting an asymmetry in arbitration accessibility independent of information content. Chen et al. [11] show that six SOTA audio-LLMs predict “neutral” on emotion when lexical cues are neutral or absent and fail under cue conflict. Yosha et al. [13] introduce the StressTest benchmark and the Stress-17K-raw training set; the finetuned StresSLM model (with *trained encoder*) generalizes to real recordings.

Our work occupies a different point in the design space than these diagnostic papers: rather than diagnosing failures of deployed audio-LLMs, we test whether the complete R1.8 configuration applied to a frozen–frozen pipeline can improve adapter representation and whether the frozen LLM consumes any recovered signal. Our Probe-K diagnostic operates on the adapter output one layer earlier than the LLM-internal analyses of Billa [10]; our Probe-G + oracle diagnostic measures consumer-side use directly.

2.3. SpeechEmotionLlama and the Role of Explicit Text-Target Alignment

Kang et al. [15] train a speech encoder (*the encoder is trained*; the LLM remains frozen) to produce token embeddings that align the LLM’s responses to expressive speech with its responses to *semantically equivalent text conditioned on speaking style*. Their result, “frozen large language models can perceive paralinguistic aspects of speech”, has shaped the subsequent literature’s intuition about how to bridge the cascade-equivalence wall.

Our Stage 7 *styled-teacher* experiment is designed to test this intervention under the constraint that Kang et al. [15] relax: *the encoder is also frozen*. The teacher’s text input becomes transcript + “[stress on word: <word>]” on counterfactual-pair members;

everything else is bit-identical to our R1.8 cohort. The intervention does not bridge in this pilot: Probe- G_{neutral} lands at 0.516, inside the R0 equivalence band, while K_T^{styled} shows the frozen LLM *can* consume the tag at 0.79+ when given direct text access. This narrows the empirical conclusion: under frozen encoder + frozen LLM, the alignment-target intervention alone is not sufficient in the evaluated setting; encoder flexibility (which Kang et al. [15] retain and our setup deliberately denies) may be one important reason their configuration crosses the gap. **To our knowledge, no published work isolates the frozen-encoder ablation of the Kang et al. [15] recipe;** our Stage 7 provides a first pilot ablation. We note that Section 4.4 of Kang et al. [15] reports linear probes on temporally mean-pooled speech tokens for a 15-way SER task (their baselines 25.08%/31.20% vs. SpeechEmotionLlama variants 68.15%/73.15%/78.40%). Our Probe-K methodology adopts that exact probing protocol and extends it to a 13-class sentence-stress label under the frozen-encoder + frozen-LLM regime, with the additional K_T transcript-embedding baseline that their analysis does not include.

Ryumina et al. [16] pursue a related design for paralinguistic recognition: a Wav2Vec2.0 audio encoder is coupled with a Jina-v3 text encoder via a BiFormer cross-modal fusion module on the cross-lingual MELD/RESO corpora, with a frozen Qwen3-4B used for LLM-based label smoothing. Like SpeechEmotionLlama, this design grants the LLM access to text-side abstractions of paralinguistic signal (here, emotion labels); unlike either approach, the LLM operates outside the response loop as an external label augments rather than as the response head, so the audio-to-LLM-state interface that our Stage 7 isolates is not exercised in their setup.

2.4. LoRA in Cross-Modal Adaptation

LoRA [17] is the standard tool for parameter-efficient adaptation of frozen LLMs and underlies a substantial fraction of the audio-LLM literature (e.g., the PLoRA variant of Wang et al. [2]). Concurrent independent work [18] reports that LoRA at certain ranks induces representational drift and changes attention patterns, a finding consistent with our Stage 6 rank 8 oracle degradation (−5.94 pp drift on the Probe-G-oracle re-confirm). Billa [12] notes that fine-tuning ablation suggests “arbitration behavior depends more on LLM reasoning than on the audio input path alone.” Our minimal-LoRA sweep at ranks 4 and 8 over $q_{\text{proj}}/v_{\text{proj}}$ on Qwen3-8B explicitly tests whether minimal consumer adaptation can improve response-level use in this setting; it does not in the single-seed pilots. The closest adjacent setup is StresSLM [13], which pairs LoRA on W^Q/W^V with encoder training and reaches 86.2% SSR on the same data; our Stage 6 isolates the LoRA contribution under the frozen-encoder + frozen-adapter constraint and adds a pre-specified oracle-stability criterion (Probe-G-oracle re-confirm within ± 2 pp of the Stage 1a anchor 0.7871). The rank-8 oracle degradation (−5.94 pp) is consistent with the LoRA rank trade-off pattern documented by Rathore et al. [18].

2.5. Multi-Task Gradient Diagnostics

Du et al. [19] propose using cosine similarity between gradients as an adaptive weight to detect whether an auxiliary loss is helpful. Wang et al. [20] extend this to multilingual MT optimization and introduce GradVac. Yu et al. [21] propose PCGrad, projecting each task’s gradient onto the normal plane of conflicting task gradients. Our Stage 3.7 cosine-to-update diagnostic that led to dropping $\mathcal{L}_{\text{NCE-cond}}$ in Stage 3.8 is the operational form of Du et al. [19]’s rule under the additional constraint of uniform gradient clipping (cosine alignment is scale-invariant under uniform clipping). We treat this finding as an engineering note acknowledging the precedent (Section 4.8) rather than a methodology contribution.

2.6. Speech Encoder Probing

Pasad et al. [22]’s layer-wise analysis of wav2vec 2.0 motivates our raw Probe-K layer sweep on WavLM-Large [23]. SUPERB [24] establishes the lightweight-head-on-frozen representation evaluation protocol that our Probe-K extends. Bentum et al. [25] probe wav2vec 2.0 embeddings for binary stressed/unstressed syllable classification across five languages with diagnostic classifiers; this is the closest prior *stress-axis* probing of self-supervised speech encoders on raw encoder representations. Our Probe-K differs in operating on trained-adaptor outputs and on a 13-class word-position stress label rather than a binary syllable-level distinction. WhiStress [26] provides the canonical sentence-stress annotation pipeline used by Stress-17K-raw and the WhiStress-style specialist label that constructs our Cascade-T+L baseline.

2.7. Continuous Acoustic Representations and Vocoder-Based Modeling

Before current speech-to-LLM adapters, acoustic modeling and voice conversion work already treated paralinguistic variation as information carried by continuous speech features rather than by transcripts alone. For example, Al-Radhi et al. [27] integrate a continuous vocoder into a DNN-based statistical voice-conversion framework, converting continuous F_0 , maximum voiced frequency, and spectral features to improve voice-conversion quality. This line of work is important for our framing because it makes explicit that source/prosodic and spectral features can remain meaningful modeling targets after neural mapping. Our study asks a different question at a later interface: When a frozen speech encoder already encodes stress-related acoustic variation, does a trainable continuous-prefix adaptor preserve that variation, and can a frozen LLM use it without converting it into discrete text tags?

3. Materials and Methods

3.1. Terminology

For non-specialist readers, we briefly recap the four most-used terms. **Frozen–frozen** means both the audio encoder and the LLM are kept at their pre-trained weights and only the small intermediate adaptor is trained; this is the regime studied throughout the paper. **Counterfactual pair** (*cf-pair*) is two audio recordings of the same sentence whose only intentional difference is which word carries primary stress (e.g., “SHE called him” vs. “she called HIM”); *cf-pairs* let us isolate transcript-invisible information. **Continuous-prefix injection** is the standard way frozen–frozen adapters connect to a frozen LLM: the adaptor outputs continuous embedding vectors that are concatenated with the LLM’s input embedding sequence (rather than passed as decoded text tokens). K_T (**text-only baseline**) is the linear-probe accuracy of recovering stress from the frozen LLM’s hidden states when given only the bare transcript as input; it is the floor an audio adaptor must clear to add value over reading the transcript. $K_T^{styeled}$ is the same probe when the transcript is augmented with an explicit text tag like “[stress on word: <word>]”; it represents what the frozen LLM *can* consume when the paralinguistic content is given in text form.

Table 1 summarizes the compact notation used throughout the manuscript. The stage names are audit identifiers from the locked experimental log, not additional model components.

Table 1. Notation and stage labels used in the manuscript.

Term	Plain-Language Meaning	Role in the Paper
R0	BLSP-style baseline adapter	Transcript-conditioned behavior alignment only.
R1.8	Counterfactual adapter	R0 objective plus $\mathcal{L}_{cf}$, $\mathcal{L}_{artifact}$, and $\mathcal{L}_{cond-pred}$ with $\mathcal{L}_{NCE-cond}$ removed.
Probe-K	Representation probe	Classifier on pooled adapter outputs; asks whether stress position is decodable before the LLM response.
Probe-G	Response probe	Constrained-answer scoring on the LLM response; asks whether the response changes with the stressed word.
K_T	Bare-transcript text baseline	Linear stress probe on frozen LLM states from transcript-only input.
K_T^{styled}	Styled text baseline	Same LLM given an explicit stress tag in text form.
Cascade-T, Cascade-T+L	Cascade references	ASR-transcript baselines without, and with, the explicit stress tag.
Stages 2, 3.8, 4, 6, 7, 8	Audit-log stage IDs	Respectively, baseline, counterfactual cohort, falsification controls, LoRA pilot, styled-teacher pilot, and frozen-consumer internal analysis.

3.2. Fixed Architecture and Training Stack

All experiments use the configuration in Table 2, specified in `configs/locked_cells.yaml` and unchanged across Stages 2–7. The encoder is WavLM-Large [23] layer 16, frozen; layer choice is the highest-LINEAR-Probe-K cell on Stress-17K probe-eval (held-out transcripts; Section 3.4), beating WS{8..16} by +0.008 and L12 by +0.015. The LLM is Qwen3-8B-Instruct [28], frozen, with `enable_thinking=False` so the assistant turn carries the empty `<think>\n\n</think>` block (constraint-answer log-probabilities are unstable without this flag; signed margin shrinks $\sim 3\times$). The adapter is a 2-layer MLP with $4\times$ strided 1D conv stem, RMSNorm, and a learned modality token; trainable parameters total 84M. The initial design target was ≤ 60 M parameters. The arithmetic consequence of MLP hidden = $2 \cdot d_{LLM} = 8192$ with $d_{LLM} = 4096$ exceeds this by 40%; we accepted the deviation. The adapter inserts via `inputs_embeds` in a ChatML template with `<|vision_start|>/<|vision_end|>` bookends inherited from the Qwen2-VL family.

Table 2. Fixed architecture and pre-specified hyperparameters. Every Stage 2–7 cell uses the same encoder, adapter, LLM, optimizer, schedule, and augmentation pipeline.

Component	Setting	Source
Speech encoder	WavLM-Large layer 16, frozen ($d_{enc} = 1024$, 50 Hz)	[23]
Adapter architecture	Conv1d (1024 $\rightarrow$ 4096, k = 4, s = 4) $\rightarrow$ MLP-2 (4096 $\rightarrow$ 8192 $\rightarrow$ 4096, GeLU, residual) $\rightarrow$ RMSNorm $\rightarrow$ +modality-token	
Adapter trainable params	84 M	
Adapter init (last linear) std	0.000345 ($\approx$ measured embed std / $\sqrt{d_{LLM}}$)	
LLM	Qwen3-8B-Instruct, frozen, <code>enable_thinking = False</code> , $d_{LLM} = 4096$	[28]
LLM injection	<code>inputs_embeds</code> with ChatML + <code>< vision_start >/< vision_end ></code> bookends	
Optimizer	AdamW lr = 5×10^{-5} , weight_decay = 0.01	
Schedule	400 steps; warmup 300 (R1.8+) or 500/250 (R0); cosine to 10^{-6}	
Effective batch	32 (microbatch 1 + grad-accum 32)	
Grad clip	max_norm = 1.0	
Precision	encoder + LLM bf16 frozen; adapter fp32 trainable; loss/optim fp32	
Augmentation	<code>aggressive_stage3_config</code> : bandlimit, reverb, noise, mu-law, jitter, codec	
Dependencies	<code>transformers==4.51.3</code> , <code>torch==2.6.0+cu124</code> , <code>peft==0.15.2</code>	

The adapter input comprises WavLM hidden states [16] at 50 Hz, the conv stem down-samples to 12.5 Hz, and the MLP + RMSNorm produces continuous prefix embeddings at $d_{\text{LLM}} = 4096$. Adapter init: last-linear std = $0.000345 \approx \text{measured embed std} / \sqrt{d_{\text{LLM}}}$ where the measured Qwen3-8B `embed_tokens.weight.std()` is 0.02205. The resulting $\|A(H)\| / \|\text{embed_row}\|$ ratio stays in $[0.5, 2.0]$ throughout training (mean 1.38 ± 0.02 at end-of-training across the R0 cohort).

Optimizer: AdamW $\text{lr} = 5 \times 10^{-5}$, `weight_decay` 0.01, cosine decay to 1×10^{-6} after warmup. Schedule: 400 steps with warmup 300 (R1.8) or 500/250 (R0 seeds 1234/1235–1238). Effective batch 32 via microbatch 1 + grad-accum 32. Precision: encoder + LLM bf16 frozen; adapter fp32 trainable; loss/optim fp32. Gradient clip `max_norm` = 1.0. Augmentation: `bandlimit` + synthetic reverb + additive noise + `mu-law` + `resample-jitter` under `aggressive_stage3_config` for Stages 3+ (`apply_prob` = 1.0, `bandlimit` ≤ 5 kHz, SNR floor 0 dB, RMS-norm 0.05, `ffmpeg` codec $\in \{\text{mu_law, mp3_64, mp3_128, opus_24, opus_48}\}$).

3.3. Compute and Inference Footprint

All training runs were executed on a single NVIDIA A100-class GPU (NVIDIA Corporation, Santa Clara, CA, USA); the public reproduction scripts require approximately 24 GB VRAM. The R0 baseline costs roughly 30 min per seed, while the R1.8 counterfactual cohort costs roughly 37 min per seed. The Stage 6 LoRA pilots cost 43–44 min each, and the Stage 7 styled-teacher pilot costs 36.7 min. The full released exploration, including failed pilots, controls, evaluations, and repeated diagnostics, is approximately 250 GPU-h. At inference time, the model adds WavLM-Large plus the 84M-parameter adapter before the frozen Qwen3-8B decoder. The adapter checkpoint is about 0.36 GB per seed in the released HuggingFace bundle; the optional LoRA pilots add 1.92 M parameters at rank 4 or 3.83 M at rank 8. We report this footprint rather than optimized latency because the released code prioritizes reproducibility and diagnostic logging over deployment-time batching.

Stage 8 is a frozen-model diagnostic rather than a training intervention. Its extraction and probe-fitting jobs used up to six NVIDIA H200 GPUs in parallel and approximately one aggregate GPU-h: hidden-state extraction required about 3–4 min per condition, the 37-state probe fits about 7 min per condition, and the attention audit about 5 min in total. These figures exclude model download and environment setup.

3.4. Datasets and Partition

We use four datasets, summarized in Table 3.

Table 3. Datasets and partition. Stress-17K-raw probe-eval is the held-out development partition under a deterministic SHA-256 hash; StressPresso is reserved for reporting and is not used for model selection.

Dataset	Rows Used	Role	Partition Strategy	License
Stress-17K-raw probe-train	3629 (822 transcript IDs)	cf-pairs Φ training	SHA-256-hashed 80/20 keyed BTA-2026-05-02	CC-BY-NC-4.0
Stress-17K-raw probe-eval	861 (197 transcript IDs)	Probe-K selector (held-out transcripts)	same hash	CC-BY-NC-4.0
StressPresso test	202 items	Probe-G eval; held out from selection	report-only	CC-BY-NC-4.0
LibriSpeech-clean-100	28,539	audio-domain mix	subset (6000 sampled)	CC-BY-4.0
Expresso-read	5799	artifact-matched cf-pairs	matched (speaker, text)	CC-BY-NC-4.0

Table 3. Cont.

Dataset	Rows Used	Role	Partition Strategy	License
$n_{\max_classes}$ (full Probe-K axis)	13	—	—	—
Full-axis chance	≈ 0.077	—	—	—
Multi-stress filtered (dropped)	1221	filtered out before partition	—	—

Stress-17K-raw is partitioned 80/20 by deterministic SHA-256 hash of transcription_id keyed BTA-2026-05-02. After filtering multi-stress and out-of-range items (1221 dropped), 4490 rows remain: probe-train 3629 rows/822 transcript IDs; probe-eval 861 rows/197 transcript IDs ($n_{\max_classes} = 13$, full-axis chance ≈ 0.077). StressPresso ($n = 202$) is held out from selection and used for Probe-G evaluation only. LibriSpeech-clean-100 (HuggingFace openslr/librispeech_asr, config clean, split train.100) provides 28K samples for the audio-domain mix; Espresso-read (HuggingFace ylacombe/espresso, read split) provides matched (speaker, text)-style pairs for the artifact-matched negative cf-pairs.

3.5. Counterfactual Pair Construction

We construct two cf-pair files (Section 3.6): a primary file cf_pairs_train.jsonl of 3666 pairs in which two same-transcript audios from Stress-17K-raw differ only in the stress index Φ , and an artifact file cf_pairs_artifact.jsonl of 6846 pairs in which Φ is unchanged but auxiliary surface features differ (speaker, style). The primary file feeds $\mathcal{L}_{cf}$; the artifact file feeds $\mathcal{L}_{artifact}$ (Section 3.6).

For Stage 4 Control B (Section 3.7), we additionally construct cf_pairs_train_shuffled.jsonl: cf-pairs in which audio is drawn from a *different* transcript per pair member while the (transcript, Φ) label structure is preserved.

3.6. The R1.8 Loss Family

The full loss for the R1.8 cohort is

$$\mathcal{L}_{R1.8} = \mathcal{L}_{BLSP} + \lambda_{cf} \mathcal{L}_{cf} + \lambda_{artifact} \mathcal{L}_{artifact} + \lambda_{cond} \mathcal{L}_{cond-pred} \quad (1)$$

with $\lambda_{cf} = 5$, $\lambda_{artifact} = 1$, $\lambda_{cond} = 0.5$, $\lambda_{NCE} = 0$ (Section 4.8 explains why the InfoNCE term is dropped).

$$\mathcal{L}_{BLSP} = \mathcal{L}_{task} + \lambda_{KL} \cdot \text{KL}(G(A(H), p) \parallel G(K_T, p)) \quad (2)$$

$$\mathcal{L}_{cf} = -\mathbb{E}_{(a,a'),p} \left[\log \frac{e^{s_G(A(H(a)), y_{\Phi(a)})}}{e^{s_G(A(H(a)), y_{\Phi(a)})} + e^{s_G(A(H(a')), y_{\Phi(a')})}} \right] \quad (3)$$

$$\mathcal{L}_{artifact} = \mathbb{E}_{(a,a''),p} [\max(0, \text{KL}(G(A(H(a)), p) \parallel G(A(H(a'')), p)) - \delta_{artifact})] \quad (4)$$

where $s_G(K, y)$ is the LLM's mean-log-prob score for answer y given input K and probe prompt p ; $y_{\Phi(a)}$ is the correct answer for stimulus a and $y_{\Phi(a')}$ the counterfactual answer; a'' is the artifact-matched negative (same Φ , different auxiliary features); $\delta_{artifact} = 0.10$. The $\mathcal{L}_{cond-pred}$ term is the cross-entropy of a small prediction head C_ϕ on top of pool($A(H)$) + restricted- T features against the Φ label; we use bag-of-words restricted- T rather than MiniLM to avoid lexical leakage. $\mathcal{L}_{cf}$'s correctness-aware design is the key update from divergence-only counterfactual losses; it rewards the LLM for assigning more mass to the $\Phi(a)$ -correct answer than to the $\Phi(a')$ -counterfactual answer, not for divergence in arbitrary directions.

3.7. Falsification Controls (Stage 4)

The R1.8 cohort representational gain (Probe-K MLP-2 cohort 0.306, clearing the K_T floor at 0.290) is small in absolute terms. Without controls, it could reflect (i) optimizer regularization (audio-shaped data acts as a regularizer; the loss is what helps); (ii) shortcut features (the adapter retains TTS-detector or domain-classifier biases unrelated to Φ); or (iii) chance variation across seed runs. Stage 4 runs three falsification controls.

3.7.1. Control A: Text-Only Adapter (A_{textK})

Replace the audio path with $P(\text{embed_tokens}(\text{transcript}))$ where P is a frozen $4096 \rightarrow 1024$ linear projection fit on 500 Stress-17K probe-train pool(K_T) $\rightarrow$ pool(H) pairs (numpy lstsq, train MSE 0.090, train cos-sim 0.999). Train under R1.8 loss (Equation (1)) with bit-identical other hyperparameters. Three seeds (1234, 1235, 1236) are included. **Pre-specified criterion:** MLP-2 cohort ≤ 0.31 (K_T floor $0.290 + 0.02$). If A_{textK} reaches the R1.8 representational level, the loss acts as optimizer regularization, not as audio-feature preservation.

3.7.2. Control B: Audio-Decorrelated Adapter ($A_{\text{R1.8-shuffle}}$)

Use `cf_pairs_train_shuffled.jsonl` (3666 pairs; audio drawn from a different transcript per pair member; transcript + Φ labels intact). Train under R1.8 loss with bit-identical hyperparameters. Three seeds are included. **Pre-specified criterion:** MLP-2 cohort ≤ 0.275 (R0 cohort MLP-2 $0.245 + 0.030$). If shuffled adapter reproduces R1.8's gain, transcript-side leakage explains the result.

3.7.3. Shortcut Probes (Per-Seed Cohort Scan)

Train and evaluate shortcut classifiers (tts_vs_real 4-class on {nova, echo, real₁, real₂}; domain binary on {TTS, real}) on the R1.8 cohort's $A_{\text{R1.8}}(H)$ pooled features. **Pre-specified criterion:** per-seed absolute drop ≥ 0.10 vs. the raw-encoder baseline (raw H tts_vs_real 0.89, domain 0.99 measured under `aggressive_stage3_config` in Stage 3.0.5). The cohort criterion is met when at least 4/5 seeds clear the per-seed threshold.

3.8. Evaluation Protocol

Probe-K (representational): Linear and 2-layer MLP heads on pool($A(H)$), fitted with stratified train/eval partition under the deterministic 80/20 hash. Reported metrics: `eval_acc_full` (the primary model-selection metric; argmax over $n_{\text{max_classes}} = 13$ positions), `eval_acc_within` (argmax over per-transcript actually-observed positions), `sp_full` (StressPresso held-out; report-only), `robust_fp16` and `robust_sigma_01` (additive Gaussian $\sigma = 0.1$ on cached pooled features). LINEAR is the pre-specified primary probe; MLP-2 is the higher-capacity diagnostic.

Probe-G (consumer): Mean-log-prob constrained-answer scoring over candidates "A"/"B" on StressPresso ($n = 202$) with delimiter `\nAnswer:.` Twenty paraphrases per item (10 neutral, 10 explicit-cue prompts). Bootstrap 95% CI by transcription_id clusters (1000 iters). Reported: total accuracy, `accuracy_neutral` (load-bearing), and `accuracy_explicit`.

Probe-G-oracle (consistency check): Same task with the stressed word marked `[[word]]` in the user message. Run on Day 1 (Stage 1a); Qwen3-1.7B accuracy 0.728 failed the 0.75 floor and forced the SWITCH_8B decision. Qwen3-8B accuracy 0.7871 (borderline pass against the 0.80 ideal) was locked. The same number re-confirms to four decimal places at every cohort and pilot, serving as a pipeline-integrity check.

Domain, style, and lexical-leakage diagnostics (Stage 3+): `C.1_within_domain` stratifies Probe-K accuracy by source subset (nova/echo/stresspresso); `C.2_cross_domain_transfer`

measures in-domain vs. cross-domain Probe-K and the ratio. C.3_domain_stratified_probe_g measures per-speaker uplift. C.4_domain_probe_on_adapter re-trains a domain probe on the adapter output. B3_style_probe re-trains a style classifier on the adapter output. T_only trains C_ϕ on restricted- T alone (must stay <0.30 , the lexical contamination floor).

Three-Tier Diagnostic Protocol

The combination—paralinguistic-axis Probe-K on pooled adapter output (precedented by Kang et al. [15], Section 4.4, for SER and by Bentum et al. [25] for binary syllable stress on raw wav2vec 2.0), Probe-G on consumer behavior, and deterministic Probe-G-oracle—re-confirms a cross-stage consistency check and composes a 3-tier protocol on the same task with the additional K_T transcript-embedding baseline (which the prior probing literature does not include). It complements the consumer-side metrics of Yosha et al. [13]. Probe-K itself operates one layer upstream of the LLM-internals diagnostics (LEACE + logit-lens of Billa [10], who probes at LLM hidden-state layers $\{0, 4, 8, 12, 16, 20, 24, 28, 31\}$ with acoustic-energy/pitch and character-CTC probes, not paralinguistic-axis classification on adapter output). To our knowledge, this specific 3-tier composition with cross-stage exact-match consistency checks (every cohort + pilot re-confirms the Stage 1a Probe-G-oracle to four decimal places) has not been operationalized in prior frozen–frozen audio-LLM work. Stage 8 then extends the representation–response comparison through all Qwen3-8B blocks without changing the trained models.

3.9. Internal Information-Flow Analysis of the Frozen Consumer (Stage 8)

The endpoint diagnostics above do not show whether stress decodability changes abruptly inside the frozen LLM. We therefore conducted a post hoc internal analysis that traces the stress label Φ through Qwen3-8B (revision b968826) while keeping WavLM, the adapter, and Qwen frozen. Only lightweight diagnostic probes are fit on cached states; this stage does not train or modify a speech-to-LLM model.

3.9.1. States and Conditions

A forward pass with `output_hidden_states = True` yields 37 states the residual-stream input $E_0 = \text{hs}[0]$ and states $B_k = \text{hs}[k]$ after blocks $k = 1, \dots, 36$; B_{36} includes the model’s final normalization. We compare five conditions under the same ChatML prompt and `<|vision_start|>/<|vision_end|>` slot bookends: seed-1234 R0, R1.8, and Stage 7 speech adapters; a direct-text positive control containing the gold transcript plus [stress on word: <word>] (K_T^{styled}); and a bare transcript negative text control ($K_T\text{-plain}$). Stage 6 is not included because its LoRA weights change Qwen and would no longer diagnose the same frozen consumer. The primary prompt is the first locked neutral paraphrase, “What does the speaker most likely mean?”.

3.9.2. Layer-Wise Readouts

At every state we retain the mean and last-token vector of the injected slot and the vector at the final prompt position immediately before the candidate answer. On the fixed Stress-17K-raw split (3629 train and 861 evaluation rows; 822/197 disjoint transcription_ids; 13 classes), we fit the same linear and 256-unit MLP-2 classifiers at each state. The primary trajectory is full-axis MLP-2 accuracy on the slot mean; full-axis linear, within-transcript, slot-last, and answer-state results are released as robustness readouts. Probe optimization uses the locked 80-epoch schedule (learning rate 10^{-3} , batch 256, weight decay 10^{-4} , seed 1234). Probe seeds 1235 and 1236 are additionally refit at seven spaced states; they are optimization replications, not independent model seeds. As an extraction invariant, the pre-Qwen pooled-adapter probes must reproduce the released seed-1234 MLP-2 values within 0.005 (R0 0.2822, R1.8 0.3124, Stage 7 0.3066).

3.9.3. Answer-Margin Logit Lens

At each answer state we apply Qwen’s final normalization and language-model head and record the log-probability difference between the single-token candidates “A” and “B”, oriented toward the correct option. At B_{36} , this calculation must match the ordinary forward pass within 10^{-4} ; the observed maximum difference is zero. Intermediate margins are diagnostic and are not interpreted as generated outputs from a truncated model.

3.9.4. Attention Audit and Statistics

On all 202 StressPresso items (96 transcript clusters), a separate eager forward pass records answer-query attention to the injected slot and to prompt regions at every block. Speech slots average 104.1 tokens, compared with 14.9 for K_T^{styled} and 7.8 for K_T^{plain} , so we report both total slot mass and mass per slot token; neither normalization alone is treated as a causal measure. We also report attention to the broad left/system and right/question regions and to the two slot bookends. Attention weights are descriptive.

The sampling unit for uncertainty is `transcription_id`. Layer-wise probe differences use a paired cluster bootstrap (2000 iterations, seed 0), with pointwise 95% intervals and a simultaneous 95% band for the primary R1.8-minus-R0 trajectory based on the studentized maximum absolute statistic. StressPresso margins and attention comparisons use 1000 transcript-cluster bootstrap iterations and Benjamini–Hochberg adjustment across 36 blocks. Complete trajectories are reported; no block is selected post hoc.

3.10. Locked Sequential Experimental Protocol

All training decisions for Stages 2–7 are pinned in `configs/locked_cells.yaml`, which uses an append-only per-stage structure (each stage’s section was added when that stage closed; threshold values appear inline in the stage’s `branch_thresholds` key). The pinned thresholds (+0.025 response-improvement criterion, +0.020 representation-only criterion, ± 0.015 equivalence band, ± 2 pp oracle drift, K_T floor 0.290, desc-only floor 0.361) were not modified post hoc; the cross-stage temporal ordering is verifiable via `training_logs/stage*_results.md` mtimes released alongside the artifact bundle. Stage 8 is separately labeled as a post-review, post hoc diagnostic with its resolved settings appended to the same file. `paper/numbers_audit.json` confirms 54/54 original claims, and the Stage 8 audit confirms 50/50 additional claims within tolerance. This is a locked sequential training protocol plus a labeled post hoc analysis, not an externally timestamped immutable preregistration.

4. Results

We lead with the falsification controls (Stage 4) so the positive representational result rests on validated evidence. Per W6, the Stage 4 control narrative precedes the R0/R1.8 cohort comparisons in the prose; the table sequence follows the chronological stage order for reader convenience.

Table 4 provides the cohort-level summary across all stages; subsequent subsections elaborate per stage. The following figures summarize the experimental results. Figures 2 and 3 report the Stage 8 internal trace and attention audit; Figures 4 and 5 summarize response- and representation-level outcomes; Figures 6 and 7 show the gradient diagnostics; Figure 8 summarizes the falsification controls; Figure 9 reports the LoRA pilot; Figure 10 compares the Stage 7 reference baselines; and Figure 11 summarizes the complete experimental trajectory.

Table 4. Cohort-level Probe-G + Probe-K across all stages. Reference baselines K_T /Cascade-T/ K_T^{styled} /Cascade-T+L/ Probe-G-oracle are deterministic (single value across seeds).

Metric	R0 Cohort 5 Seeds	R1.8 Cohort 5 Seeds	Stage 7 (Styled, Single Seed)	Δ R1.8 vs. R0
Probe-G total (mean $\pm \sigma$)	0.6306 $\pm$ 0.0095	0.6326 $\pm$ 0.0061	0.6245	+0.0020
Probe-G _{neutral} (mean $\pm \sigma$)	0.5122 $\pm$ 0.0108	0.5122 $\pm$ 0.0039	0.5158	+0.0000
Probe-G _{explicit} (mean $\pm \sigma$)	0.7491 $\pm$ 0.0120	0.7531 $\pm$ 0.0138	0.7332	+0.0040
Probe-G signed margin total (mean)	+2.135 $\pm$ 0.225	+2.040 $\pm$ 0.225	+1.929	−0.095
Probe-K linear eval_full (mean $\pm \sigma$)	0.2105 $\pm$ 0.0128	0.2256 $\pm$ 0.0118	0.2199	+0.0151
Probe-K MLP-2 eval_full (mean $\pm \sigma$)	0.2446 $\pm$ 0.0339	0.3059 $\pm$ 0.0436	0.3066	+0.0613
R1.8 per-seed (1234–1238)	—	{0.312, 0.246, 0.343, 0.350, 0.279} (linear K_T floor 0.290; 3/5 above)	—	—
Probe-K MLP-2 within (mean)	—	0.7052	0.7218	—
K_T neutral (deterministic)	0.5064	0.5064	0.5064	—
Cascade-T neutral (deterministic)	0.5069	0.5069	0.5069	—
Probe-G-oracle re-confirm (det)	0.7871	0.7871	0.7871	0.0 drift
C.1 within-domain PASS rate	— (Stage 2 schema)	5/5 seeds	PASS (single)	—
C.2 ratio cross/in (cohort mean)	—	0.910	0.970	—
C.3 stratified PASS rate	—	2/5 seeds	FAIL (single)	—
C.4 dom_acc adapter (det)	—	0.7625	0.7625	—
B3 style probe (cohort mean)	—	0.418	0.4167	—
T-only contamination (cohort mean)	—	0.170	0.164	<0.30 floor

4.1. Seed-Level Statistical Interpretation

For the five matched seeds, R1.8 improves MLP-2 Probe-K over R0 by +0.0613 absolute (paired two-sided $t(4) = 4.39$, $p = 0.012$, 95% CI [0.0226, 0.1001]). The linear Probe-K gain is smaller and does not reach $p < 0.05$ (+0.0151, $t(4) = 2.36$, $p = 0.078$, 95% CI [−0.0027, 0.0328]). The R1.8 MLP-2 mean is numerically above the pre-specified linear K_T floor of 0.290, but this one-sample crossing is not statistically established across seeds ($t(4) = 0.82$, $p = 0.460$, 95% CI for the difference [−0.0382, 0.0700]). Probe-G neutral is unchanged between R0 and R1.8 (paired $p = 1.0$). We therefore interpret the result as a control-supported nonlinear representational improvement over R0, not as a statistically confirmed or capacity-matched superiority over the text-only baseline.

4.2. Baseline Behavior Alignment Reduces Probe-K Signal (Stage 2 R0 Cohort)

The 5-seed R0 cohort (BLSP-class adapter under $\mathcal{L}_{\text{BLSP}}$ alone) produces a cohort linear Probe-K of 0.211 ± 0.013 , strictly below the K_T baseline 0.290 on every seed (per-seed values {0.209, 0.221, 0.214, 0.189, 0.220}). Probe-G_{neutral} is 0.5122 ± 0.0108 , within ± 0.01 of K_T (0.5064) and Cascade-T (0.5069): cascade-equivalent at the consumer level on the neutral subset. The signed margin (total) is +2.135 $\pm$ 0.225nat but driven entirely by the explicit-cue half (0.7491 vs. neutral 0.5122). The five-seed Probe-K range is 0.031 on linear and 0.086 on MLP-2: the BLSP fixed point is empirically tight.

This is a stricter version of the cascade-equivalence prediction: A_{R0} is *strictly worse* than transcript-only at preserving stress-position signal in its latent form. Behavior alignment under $\mathcal{L}_{\text{BLSP}}$ alone removes residual structure that the transcript preserves. For compactness, we call this pattern *sub-cascade collapse*: the adapter falls below a transcript-only baseline on Probe-K. The structural intervention in Stage 3 must therefore clear two floors: the K_T Probe-K baseline (not just the A_{R0} baseline) and the Cascade-T Probe-G_{neutral} baseline (by ≥ 0.025 for the pre-specified response-improvement criterion).

4.3. Falsification Controls Validate the Structural-Training Gain (Stage 4)

We report Stage 4 first because every subsequent positive claim about R1.8 depends on these controls clearing. Table 5 summarizes.

Table 5. Stage 4 falsification controls on the R1.8 representational finding. Three independent gates clear with substantial margins; the R1.8 MLP-2 improvement is supported as audio-derived rather than as transcript leakage or shortcut learning.

Control/Probe	Cohort # Seeds	MLP-2 Mean $\pm \sigma$	Linear Mean $\pm \sigma$	Pre-Specified Criterion	Criterion Met?
R1.8 reference	5	0.3059 $\pm$ 0.0436	0.2256 $\pm$ 0.0118	—	—
Control A (A_{textK} , text-only) gap vs. R1.8 cohort MLP-2	3	0.2122 $\pm$ 0.0203 −0.0937	0.2292 $\pm$ 0.0094 —	MLP-2 $\leq$ 0.31 —	Yes —
Control B (audio-decorrelated, $A_{\text{R1.8-shuffle}}$) gap vs. R1.8 cohort MLP-2	3	0.2033 $\pm$ 0.0112 −0.1026	0.2083 $\pm$ 0.0159 —	MLP-2 $\leq$ 0.275 —	Yes —
Shortcut probes on R1.8 cohort $A_{\text{R1.8}}(H)$ (5 seeds)					
tts_vs_real cohort mean $\pm \sigma$	5	—	0.6450 $\pm$ 0.0565	per-seed drop $\geq$ 0.10	Yes (5/5)
raw H baseline (Stage 3.0.5)		—	0.89	—	—
drop		—	−0.245	—	—
domain cohort mean (det)	5	—	0.7625	per-seed drop $\geq$ 0.10	Yes (5/5)
raw H baseline		—	0.99	—	—
drop		—	−0.228	—	—
All three control families met their pre-specified criteria.					

4.3.1. Control A Results: Text-Only Adapter (A_{textK})

Three-seed MLP-2 cohort 0.2122 $\pm$ 0.0203. Gap to R1.8 cohort MLP-2 = −0.094 absolute. Linear A_{textK} at 0.229 is $\approx$ R1.8 linear 0.226 — the audio-derived gain is concentrated in the nonlinear MLP-2 probe. **Criterion met** (0.2122 < 0.31 threshold).

4.3.2. Control B Results: Audio-Decorrelated Adapter ($A_{\text{R1.8-shuffle}}$)

Three-seed MLP-2 cohort 0.2033 $\pm$ 0.0112. Gap to R1.8 = −0.103. Linear 0.208 is below R0 linear 0.211, when audio is decorrelated from labels, the adapter cannot encode Φ at all. **Criterion met** (0.2033 < 0.275).

4.3.3. Shortcut Probes

tts_vs_real cohort 0.6450 $\pm$ 0.0565 vs. raw H baseline 0.89, drop −0.245. Domain cohort 0.7625 (deterministic) vs. raw H baseline 0.99, drop −0.228. All five seeds clear the per-seed 0.10-absolute-drop threshold. **Criterion met** (5/5).

The composite Stage 4 result supports an audio-derived interpretation of the representational gain against three independent falsification gates.

4.4. The Complete R1.8 Configuration Improves Nonlinear Probe-K (Stage 3.8 Cohort)

The 5-seed R1.8 cohort (Equation (1); $\lambda_{\text{cf}} = 5$ matched to $\mathcal{L}_{\text{BLSP}}$ gradient norm) produces the following:

- Linear Probe-K cohort 0.2256 $\pm$ 0.0118 (Δ vs. R0 + 0.0151); the first crossing of the R0 cohort baseline observed in the sequential R-series experiments.
- **MLP-2 Probe-K cohort 0.3059 $\pm$ 0.0436, numerically above the pre-specified linear K_T floor of 0.290 at the cohort mean.** Per-seed values are {0.312, 0.246, 0.343, 0.350, 0.279}:

3/5 seeds clear the K_T floor individually; 2/5 (seeds 1235 and 1238) fall below it. The cohort-mean crossing is therefore driven by the three high-margin seeds and should be read as a directional cohort result with substantial seed variance, not as per-seed reproducibility or as a statistically established crossing of the K_T floor. MLP-2 within-transcript cohort is 0.7052.

- Probe-G total 0.6326 ± 0.0061 ($\Delta R_0 + 0.0020$), Probe-G_{neutral} 0.5122 ± 0.0039 ($\Delta R_0 + 0.0000$), Probe-G_{explicit} 0.7531 ± 0.0138 ($\Delta R_0 + 0.0040$). **The consumer-side response stays at the R0 baseline** (Section 4.5).
- C.1 within-domain Probe-K stratification PASSES on 5/5 seeds (R0/R1/R1.6 prior trajectories: 0/5). C.2 ratio cross-domain transfer cohort mean 0.910 (R0/R1/R1.6 prior ~ 0.7). The relative cross-domain ratio is the highest in this sweep.
- Per-seed $\mathcal{L}_{cf}$ final $\in [0.475, 0.629]$ (all below uniform $\log 2 = 0.693$); $\mathcal{L}_{cf}$ cosine to total update mean $\in [0.889, 0.985]$ across seeds; main clip rate 1.000 (all 400 steps clip-bound) on every seed.

The linear Probe-K Δ of +0.0151 misses the pre-specified representation-only +0.020 threshold by 0.5 pp; we report this verbatim. The cohort range on linear is 0.031, exceeding the 0.025 tightness criterion. The appropriately limited representational claim is therefore that the complete R1.8 training configuration produces a statistically supported MLP-2 improvement over R0 and passes the audio-derived controls. Because R1.8 also changes the warmup schedule and augmentation pipeline, this result is not attributed to $\mathcal{L}_{cf}$ alone. We do not claim generalized linear gain, statistically established superiority over K_T , or superiority over a capacity-matched MLP-2 text baseline at the cohort level. Stage 8 adds a post hoc capacity-matched diagnostic for the fixed seed-1234 models: at E_0 , the same MLP-2 schedule gives 0.314 for R1.8, 0.279 for R0, and 0.264 for K_T -plain. This single-model comparison is consistent with an audio-derived difference, but it does not replace a multi-seed capacity-matched text baseline for the headline cohort claim.

4.5. Response-Level Use Remains Unchanged in the Tested Configuration

The R1.8 cohort Probe-G_{neutral} matches the R0 baseline at reported precision ($\Delta +0.0000$ within $\sigma_R = 0.0039$ across five seeds), and Probe-G_{explicit} moves only +0.004. The frozen Qwen3-8B does not detectably consume the audio-derived signal at the response axis. We test two interventions designed to bridge this gap. Because both intervention results are single-seed diagnostics, we use them to bound the tested configurations rather than to rule out the intervention families.

4.5.1. Stage 6 Minimal LoRA on Qwen3-8B

Table 6 summarizes the two-rank sweep. With the R1.8 adapter and C_ϕ heads frozen, only LoRA matrices on q_proj/v_proj (72 modules across 36 transformer blocks) update. Rank 8 ($\alpha = 32$, $\alpha/r = 4$, 3.83 M LoRA params) drifts the Probe-G-oracle re-confirm from 0.7871 (Stage 1a anchor) to 0.7277, a -5.94 pp drift outside the pre-specified ± 2 pp tolerance: the oracle consistency check fails. Probe-G_{neutral} is 0.4827, -0.029 pp below R0. Rank 4 ($\alpha = 16$, 1.92 M LoRA params) preserves the oracle within +0.99 pp (0.7970, well within the ± 2 pp tolerance), but Probe-G_{neutral} lands at 0.4926, 1.96 pp below R0 and just outside the $R_0 \pm 0.015$ equivalence band on the negative side. In this single-seed pilot, the lower-drift setting did not improve the audio-conditioned response metric.

Table 6. Stage 6 minimal-LoRA two-rank sweep. In single-seed runs, rank 8 degrades the oracle consistency check; rank 4 preserves the oracle but does not improve audio-conditional Probe-G.

Metric	R0 Cohort 5 Seeds	R1.8 Cohort 5 Seeds	LoRA r4 Seed 1234	LoRA r8 Seed 1234
LoRA params (M)	—	—	1.92	3.83
LoRA α/r	—	—	4	4
target_modules	—	—	q_proj/v_proj $\times$ 36 blocks (72 modules)	same
Probe-G-oracle re-confirm	0.7871	0.7871	0.7970	0.7277
oracle drift (pp vs. Stage 1a)	0.00	0.00	+0.99	−5.94
Probe-G _{neutral}	0.5122	0.5122	0.4926	0.4827
Δ vs. R0	0	+0.0000	−0.0196	−0.0295
Probe-G _{total}	0.6306	0.6326	0.6277	0.5938
Probe-G _{explicit}	0.7491	0.7531	0.7629	0.7050
Probe-K linear eval_full	0.2105	0.2256	0.2265	0.2276
Probe-K MLP-2 eval_full	0.2446	0.3059	0.3120	0.3124
Outcome	—	—	no Probe-G improvement	oracle consistency check failed

4.5.2. Why Stage 7 Follows Stage 6

Stage 6 tests a consumer-side bridge while keeping the recovered R1.8 adapter representation fixed. Its two outcomes delimit the next step: the conservative rank 4 setting preserves general task behavior but does not improve the audio-conditioned decision, while the larger rank 8 setting changes general behavior enough to fail the oracle consistency check. Stage 7 therefore tests a different hypothesis without increasing consumer-side capacity: perhaps the transcript-only teacher never asks the continuous audio prefix to encode stress in the semantic form expected by the frozen LLM. The styled teacher makes that semantic target explicit while preserving the same frozen encoder, frozen LLM, optimization schedule, and response task.

4.5.3. Stage 7 Styled Teacher (Alignment-Target Intervention)

Single-axis change from R1.8: the $\mathcal{L}_{\text{BLSP}}$ teacher’s text input becomes transcript + “[stress on word: <word>]” on counterfactual-pair members and stress_individual rows; LibriSpeech and Espresso (no stress label) keep plain transcript; everything else bit-identical (same λ schedule, same augmentation, same 400 steps). This is the alignment-target intervention from Kang et al. [15] under our frozen-encoder constraint.

Probe-G_{neutral} pilot result: 0.5158, inside the $R_0 \pm 0.015 = [0.4972, 0.5272]$ equivalence band. Oracle re-confirm 0.7871 (zero drift to four decimal places). Under the pre-specified single-seed stopping criterion, the pilot is reported as an equivalence-band result and the cohort run is not triggered. We flag this as a one-seed observation; replication at ≥ 3 seeds is left to follow-up work and is needed before the single-seed null can be treated as a population-level conclusion. Probe-K representational gain is preserved (single-seed MLP-2 0.3066, still $> K_T$ 0.290); C.1 spread is the tightest yet observed (0.021 vs. R1.8 cohort 0.051); and C.2 ratio is also tighter than the cohort (0.970 vs. R1.8 cohort 0.910). The styled-teacher signal makes the adapter representation more consistent across domains, while Probe-G_{neutral} does not move beyond the equivalence band.

The Stage 7 reference baselines anchor what the LLM can consume in text form (Table 7). K_T^{styled} (gold transcript + tag in audio slot) reaches Probe-G_{neutral} 0.7901; Cascade-T+L (Whisper transcript + tag in audio slot) reaches 0.8020. The frozen Qwen3-8B is therefore

capable of consuming the bracketed paralinguistic tag at 0.79+ on neutral when given direct text access. Under our frozen-encoder + frozen-LLM setup with continuous-prefix injection, this capability is not elicited via the trained audio path even when $\mathcal{L}_{\text{BLSP}}$'s text teacher distills the tag-conditioned distribution.

Distillation gap: K_T^{styled} Probe- G_{neutral} 0.7901 minus Stage 7 audio adapter 0.5158 = 0.2743 absolute (27.4 percentage points). The student speech branch matches the teacher's response distribution on training data ($\mathcal{L}_{\text{KL}}$ converges normally; final value 1.53 vs. R1.8 cohort 1.49, indistinguishable), but at inference time the audio slot fails to elicit the same tag-conditioned representation in the frozen LLM that the text slot elicits. Section 5 interprets the mechanism.

Stage 7 therefore answers the question raised by Stage 6, but negatively for this evaluated pilot: changing the teacher's semantic target preserves the adapter-side gain and improves domain consistency, yet it does not transfer the stress distinction into the generated response. This is not evidence that styled-teacher alignment fails in general; it is evidence that response-token KL distillation alone is insufficient in the reported frozen-encoder, frozen-LLM, single-seed configuration.

Table 7. Stage 7 reference baselines. The frozen Qwen3-8B can consume the bracketed stress tag at 0.79+ on neutral when given direct text access; the trained audio adapter under styled-teacher distillation lands in the R0 equivalence band.

Variant	Probe-G Total	Probe- G_{neutral}	Probe- G_{explicit}
R0 cohort (Stage 2 BLSP)	0.6306	0.5122	0.7491
R1.8 cohort (Stage 3.8)	0.6326	0.5122	0.7531
Stage 7 styled-teacher (single seed)	0.6245	0.5158	0.7332
Cascade-T (Whisper transcript → frozen Qwen3-8B)	0.6574	0.5069	0.8079
K_T (gold transcript → frozen Qwen3-8B)	0.6488	0.5064	0.7911
K_T^{styled} (gold transcript + tag in audio slot)	0.8025	0.7901	0.8149
Cascade-T+L (Whisper transcript + tag in audio slot)	0.8129	0.8020	0.8238
Probe-G-oracle (tag in user message text)	0.7871	—	—
Distillation gap (K_T^{styled} neutral – Stage 7 audio neutral)	—	0.2743	—

4.6. Internal Trace: Stress Remains Decodable But is Not Read Out (Stage 8)

Stage 8 examines one seed-matched model from each speech condition and the two text controls; it is an internal diagnostic, not an additional cohort. The pre-Qwen pooled-adapter probes reproduce the released seed-1234 MLP-2 values (R0 0.2822, R1.8 0.3124, Stage 7 0.3066), the Stress-17K split has no shared transcription IDs, all slot-boundary checks pass, and the B_{36} answer margins equal the ordinary forward calculation exactly.

4.6.1. Absolute Slot Decodability Does Not Collapse

Figure 2A follows full-axis MLP-2 accuracy at the injected slot. R0 rises from 0.279 at E_0 to 0.425 at B_{36} (peak 0.443 at B_{11}); R1.8 rises from 0.314 to 0.433 (peak 0.455 at B_{26}); and Stage 7 rises from 0.312 to 0.467. The K_T^{styled} positive control peaks at 0.632, whereas K_T -plain remains between 0.229 and 0.287. These absolute trajectories show no abrupt block at which speech-slot stress decodability disappears. They do not, however, establish propagation of the R1.8-over-R0 gain: the simultaneous 95% band for R1.8 minus R0 excludes zero at 0/37 states (Figure 2D). Thus, Stage 8 finds *no discrete erasure of absolute speech-slot decodability*, but it also finds no reliable internal R1.8-over-R0 advantage.

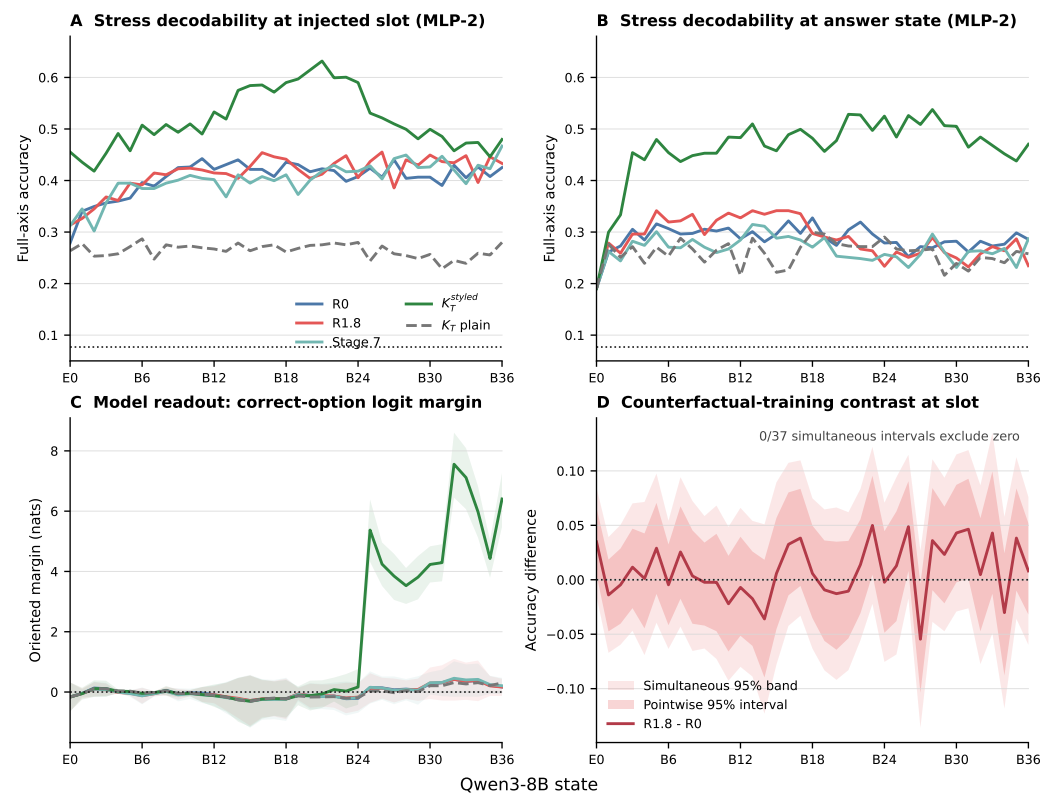


Figure 2. Internal stress-information trace through frozen Qwen3-8B. (A) Full-axis MLP-2 stress decodability at the injected-slot mean remains above the bare-transcript control for the speech conditions rather than collapsing at one block. (B) A fitted answer-state probe remains distinct from (C) the model’s own correct-option logit margin, which becomes large only for the explicit text-tag control. Shading in (C) denotes 95% transcript-cluster bootstrap intervals for R1.8, Stage 7, and K_T^{styled} . (D) The R1.8-minus-R0 slot-probe contrast with pointwise and simultaneous 95% bands; no simultaneous interval excludes zero. Dotted black horizontal lines mark the 1/13 chance level in panels (A,B) and zero margin/difference in panels (C,D). Probe accuracy is diagnostic and does not establish causal use.

4.6.2. A Fitted Answer-State Probe Is Not the Model’s Readout

At B_{36} , full-axis MLP-2 accuracy at the answer state is 0.286 for R0, 0.235 for R1.8, 0.287 for Stage 7, 0.470 for K_T^{styled} , and 0.258 for K_T -plain (Figure 2B). Because the plain-transcript control is itself above the 1/13 full-axis chance level and R1.8 is not higher than R0, these fitted-probe scores are not evidence that the R1.8-specific factor has reached a decision-relevant subspace.

The model’s own oriented “A”/“B” margin makes the distinction clear (Figure 2C). At B_{36} it is +0.168 for R0 (95% CI [−0.071, 0.388]), +0.165 for R1.8 ([−0.134, 0.439]), and +0.209 for Stage 7 ([−0.069, 0.470]). All three intervals include zero. The K_T^{styled} positive control instead reaches +6.405 nats ([5.475, 7.272]), diverging after B_{24} . The R1.8-minus-R0 final-margin difference is −0.002 with Benjamini–Hochberg-adjusted $p = 0.978$. Accordingly, the analysis does not identify a layer at which an R1.8 advantage is discarded; it shows that no reliable advantage appears in the model’s answer readout at any block.

4.6.3. Attention Provides a Normalization-Dependent Routing Correlate

Figure 3 reports answer-query attention on StressPresso. Per slot token, K_T^{styled} receives more attention than R1.8 and Stage 7 at all 36 blocks (paired cluster bootstrap, Benjamini–Hochberg-adjusted $p < 0.05$). Total slot mass often has the opposite ordering because the speech slots are approximately seven times longer. The broad left/system prompt

region receives a late-block mean of about 0.72 in every condition, while the two bookend tokens receive little mass. The late prompt concentration is therefore not speech-specific and does not by itself explain the speech–text decision gap. We interpret the attention result only as a descriptive, length-sensitive routing correlate; it does not establish an attention-sink mechanism.

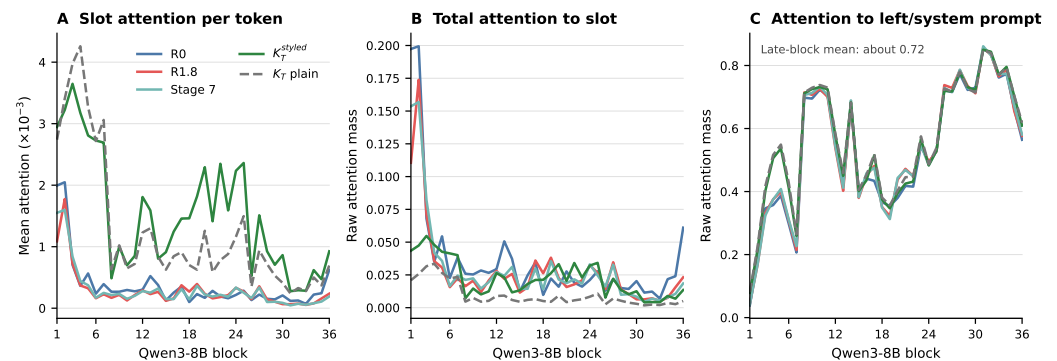


Figure 3. Descriptive answer-query attention by prompt region on StressPresso. (A) Attention per slot token is higher for the short text slots; (B) total slot mass shows why the conclusion depends on length normalization; and (C) late attention to the broad left/system prompt region is similar across all conditions. Speech slots average 104.1 tokens, versus 14.9 for K_T^{styled} and 7.8 for K_T -plain. Attention weights are not a causal attribution.

4.7. Qualitative Error Analysis

The dominant error pattern is response invariance under same-transcript stress changes. In the task, a pair such as “the BLUE rocket” versus “the blue ROCKET” should flip the preferred answer even though the transcript is unchanged. Probe-K shows that R1.8 makes some of this contrast decodable in the adapter representation, but Probe-G_{neutral} remains at the same cohort value as R0. Thus, the typical response-level failure is not a malformed answer or an inability to follow the A/B format; it is that the response behaves as if the stressed-word contrasts were not available.

Three observations support this interpretation. First, explicit-cue prompts reach much higher accuracy than neutral prompts for the same audio path (about 0.75 explicit vs. 0.51 neutral), indicating that the response head can solve the forced-choice task when the cue is linguistically available. Second, direct text access to the stress tag reaches 0.79–0.80 on the neutral setting (K_T^{styled} and Cascade-T+L), so the frozen LLM can use the relevant semantic distinction in text form. Third, real-speech transfer remains weak: the R1.8 cross-domain StressPresso Probe-K absolute values are only 0.188–0.233 across seeds, below the desc-only reference floor of 0.361, and the speaker-stratified Probe-G diagnostic passes on only 2/5 seeds. These patterns are consistent with a narrow residual adapter signal that is partly audio-derived but not robustly organized for response-level use, especially on natural speech.

4.8. Per-Component Gradient Analysis: $\mathcal{L}_{\text{NCE-cond}}$ Orthogonality

Stage 3.6 grad-norm logging revealed that $\mathcal{L}_{\text{BLSP}}$ dominates the structural terms by $5 \times -370 \times$ under the initial baseline λ schedule (mean $\|\nabla \mathcal{L}_{\text{BLSP}}\| = 141.26$, $\|\nabla \mathcal{L}_{\text{cf}}\| = 28.12$, $\|\nabla \mathcal{L}_{\text{cond-pred}}\| = 0.38$, $\|\nabla \mathcal{L}_{\text{NCE-cond}}\| = 6.06$). Stage 3.7 rebalanced $\lambda_{\text{cf}} = 1 \rightarrow 5$ and $\lambda_{\text{NCE}} = 0.25 \rightarrow 3$ —a $5 \times$ and $12 \times$ bump, respectively. The $\mathcal{L}_{\text{cf}}$ rebalance produced the first measurable Probe-K gain in any R1 cell. The $\mathcal{L}_{\text{NCE-cond}}$ rebalance did not: the loss stayed at uniform (log 2).

The diagnostic explained this was the *cosine-to-total-update* measurement (Figure 6). Across eight diagnostic events spanning the 400 training steps, $\mathcal{L}_{\text{cf}}$ ’s gradient direction

reliably aligns with the total update at cosine 0.95–1.00 from step ~ 100 , while $\mathcal{L}_{\text{NCE-cond}}$'s gradient direction remains at cosine 0.05–0.16 throughout. Cosine alignment is scale-invariant under uniform gradient clipping (every step is clip-bound at the $\text{max_norm} = 1.0$ threshold; main clip rate = 1.000), so λ tuning cannot recover an orthogonal-direction loss. We dropped λ_{NCE} to 0 in Stage 3.8.

This finding is the operational form of the cosine-similarity rule [19] under uniform clipping (Section 2.5); we treat it as an engineering note, not a methodology contribution. PCGrad [21] or GradVac [20] would in principle recover the orthogonal $\mathcal{L}_{\text{NCE-cond}}$ direction by per-component projection; we did not run this experiment.

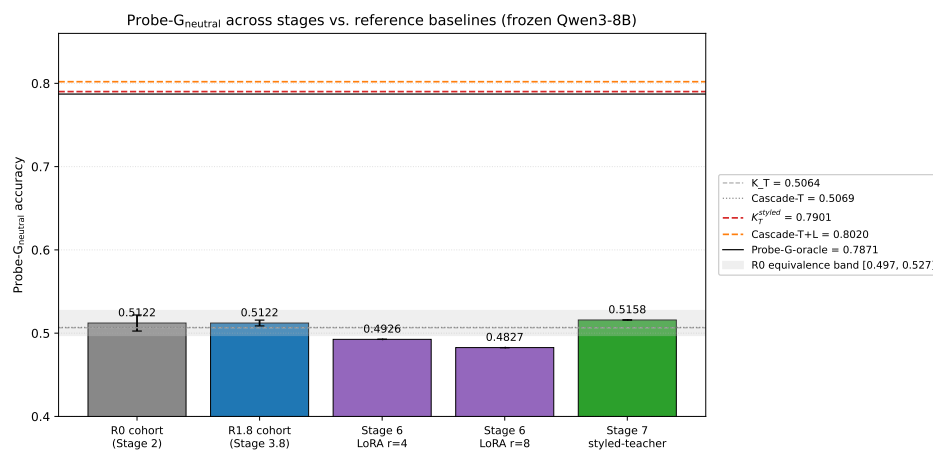


Figure 4. Probe-G_{neutral} across stages and reference baselines. Direct text access to the stress tag reaches 0.79+, while the audio-path adapters remain near the R0 baseline. Error bars show σ across five seeds for R0/R1.8 and per-run bootstrap CIs for single-seed pilots. Gray and blue bars denote the R0 and R1.8 cohorts, purple bars the two Stage 6 LoRA pilots, and green the Stage 7 styled-teacher pilot; the horizontal lines and gray band denote the labeled text/cascade references and R0 equivalence band.

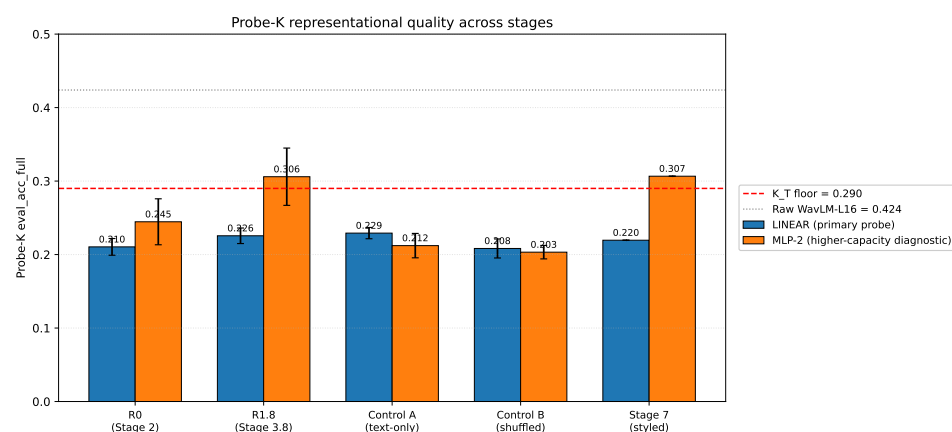


Figure 5. Probe-K representational quality across stages. The R1.8 cohort improves MLP-2 Probe-K over R0, while the linear gain remains small. The K_T line is the pre-specified linear text-only floor, so the MLP-2 comparison is not capacity-matched. Blue and orange bars denote the linear and MLP-2 probes; the dashed red line marks the K_T floor, and the dotted gray line marks the raw WavLM-L16 reference.

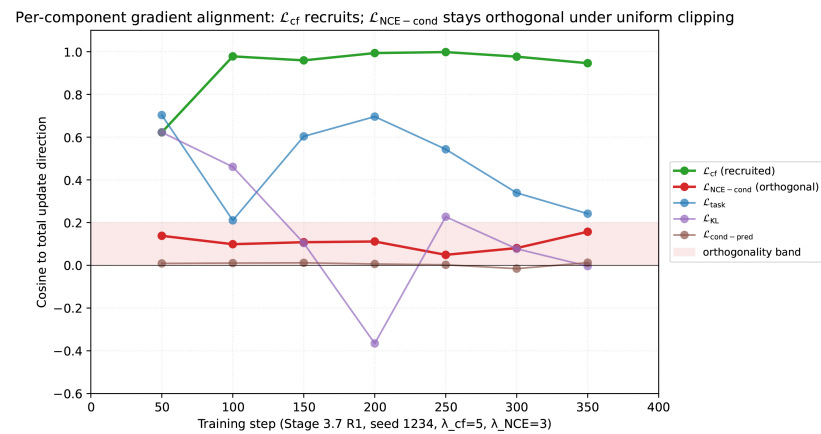


Figure 6. Per-component cosine-to-total-update evolution during Stage 3.7 training (seed 1234). $\mathcal{L}_{cf}$ aligns with the total update after warmup, whereas $\mathcal{L}_{NCE-cond}$ remains nearly orthogonal. This motivated removing $\mathcal{L}_{NCE-cond}$ from R1.8.

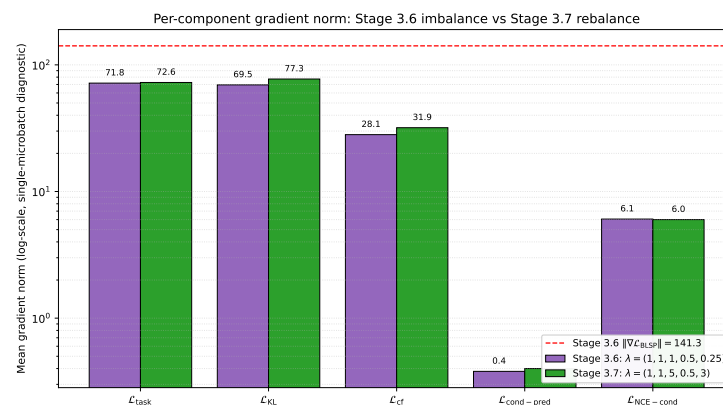


Figure 7. Per-component gradient norm during Stage 3.6 R1 training. At the initial schedule, $\mathcal{L}_{BLSP}$ dominates the structural losses. Increasing λ_{cf} brings $\mathcal{L}_{cf}$ into a comparable gradient-norm range and is followed by the first measurable Probe-K gain.

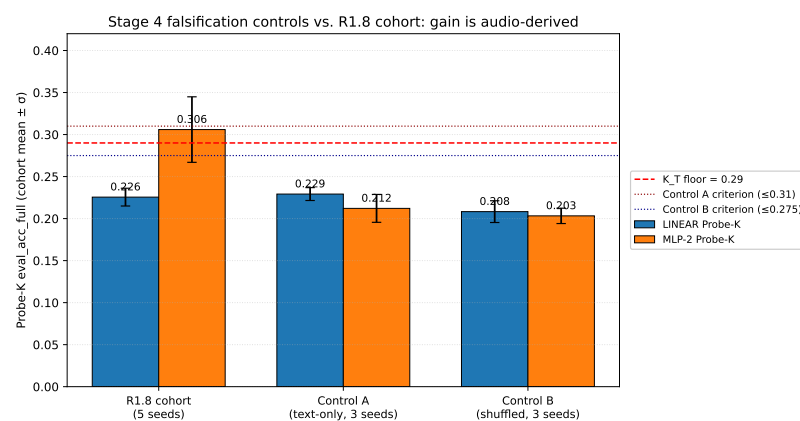


Figure 8. Stage 4 falsification controls on the R1.8 MLP-2 cohort result. Control A (text-only A_{textK}) MLP-2 cohort 0.212 $\ll$ R1.8 cohort 0.306. Control B (audio-decorrelated) MLP-2 cohort 0.203 $\ll$ R1.8. Both controls clear the pre-specified criteria with substantial margin.

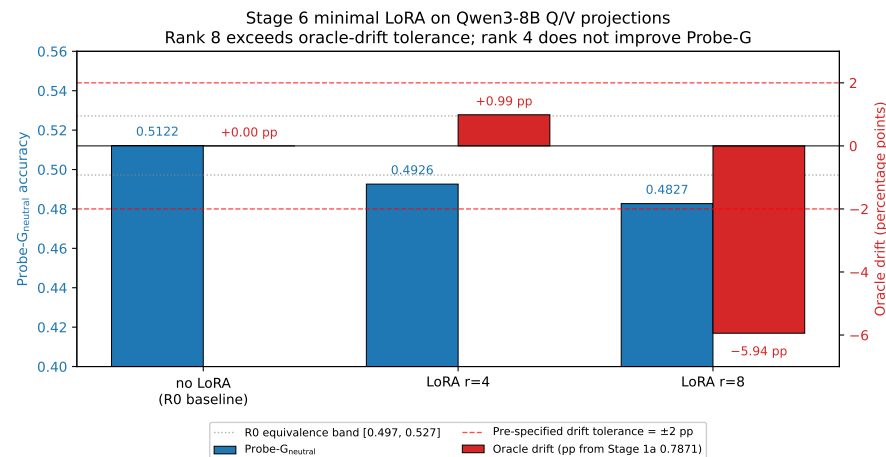


Figure 9. Stage 6 minimal-LoRA on Qwen3-8B q_{proj}/v_{proj} . In single-seed pilots, rank 8 degrades the oracle consistency check, while rank 4 preserves the oracle but does not improve Probe-G_{neutral}.

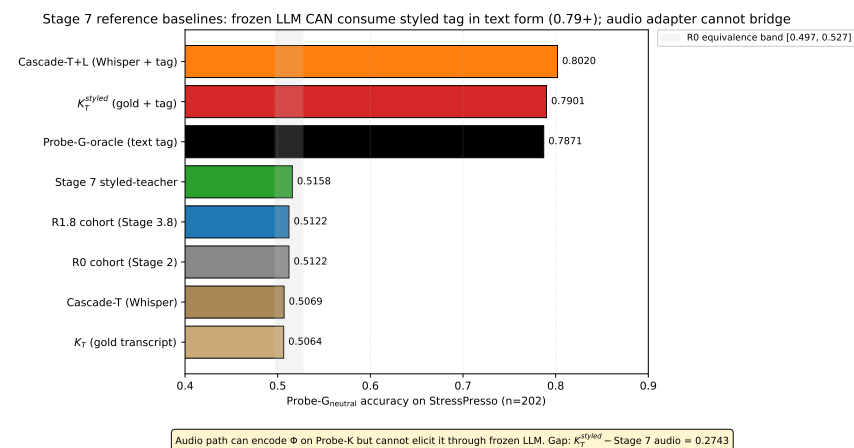


Figure 10. Stage 7 reference baselines. The frozen Qwen3-8B consumes the styled paralinguistic tag at 0.79+ on neutral when given direct text access (K_T^{style} 0.7901; Cascade-T+L 0.8020; Probe-G-oracle 0.7871), but the audio-path adapter trained against the styled teacher under $\mathcal{L}_{KL}$ distillation lands at 0.516. Gap: 27.4 percentage points versus K_T^{style} .

Beyond Transcript Alignment — 4-act experimental trajectory

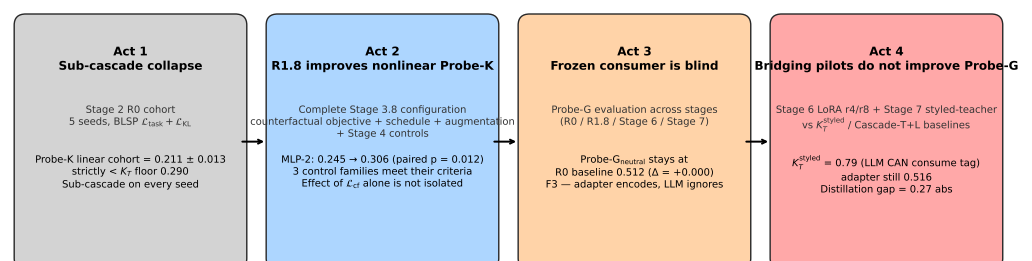


Figure 11. Four-part experimental trajectory. Stage 2 shows that transcript alignment reduces Probe-K. The complete Stage 3.8 R1.8 configuration improves nonlinear Probe-K. Stages 6 and 7 test two response-side bridges; in the evaluated pilots, Probe-G remains near R0 while direct text-tag baselines show the LLM can use the information in text form.

5. Discussion

5.1. Why the $\mathcal{L}_{KL}$ Trajectory Is Uninformative About the Styled Tag

The Stage 7 styled-teacher pilot's $\mathcal{L}_{KL}$ final value is 1.53, indistinguishable from R1.8 cohort 1.49. We hypothesize that at the response axis, the bracketed stress tag at the prefix end is too low-bandwidth in the teacher's logits to provide sufficient gradient signal for the audio path to learn the corresponding tag-conditioned representation: *most response-token predictions are determined by earlier prefix tokens, so KL between the styled-teacher and the unstyled-student stays similar to the unstyled-teacher case*. The proximate mechanism for the Stage 7 equivalence-band result is therefore not inadequacy of the alignment target itself— K_T^{styled} shows the target is informationally sufficient for the LLM's response axis—but inadequacy of *response-axis logit distillation as a transfer channel for tag-conditioned content*. Because Stage 7 is a single-seed pilot, this mechanism should be read as an interpretation of the observed run rather than as a population-level conclusion.

5.2. The Linear-vs-MLP-2 Asymmetry

The R1.8 cohort linear Probe-K Δ vs. R0 is +0.015; the MLP-2 Δ is +0.061. Stage 4 Control A linear is 0.229 ($\approx$ R1.8 linear 0.226), but Control A MLP-2 is 0.212 (vs. R1.8 MLP-2 0.306, gap -0.094). The audio-derived gain is concentrated to the nonlinear probe.

Stage 8 shows why probe capacity must not be equated with model use. Both linear and MLP-2 stress decodability can rise inside the frozen backbone, yet the model's own answer margin remains uninformative. A classifier's ability to separate labels in a hidden state therefore does not imply that the model's learned computation uses the same separating direction. At the same time, the adapter is not lossless: the raw WavLM-Large layer 16 representation selected in Stage 1 reaches a linear Probe-K of 0.424, substantially above both the R0 linear adapter mean (0.211) and the R1.8 linear adapter mean (0.226). The evidence therefore supports a two-part interpretation. First, adapter training still removes a large amount of stress-related information available in the frozen encoder. Second, the residual stress signal that R1.8 makes nonlinearly decodable is not detectably used by the frozen LLM response path. The audio-to-LLM-state alignment under continuous-prefix injection is thus a response-level bottleneck in this configuration but not the only source of information loss.

5.3. Internal Localization: No Erasure Layer and No Routed Advantage

The layer-wise trace changes the localization question. Absolute speech-slot stress decodability does not collapse at any Qwen3-8B block; indeed, it generally increases from E_0 into the early and middle blocks. It would therefore be inaccurate to name a discrete block where all stress information is discarded. At the same time, the R1.8-minus-R0 simultaneous interval includes zero at every state. The R1.8-over-R0 advantage observed in the five-seed adapter cohort is not recovered as a reliable seed-1234 internal advantage.

The answer-side results further separate decodability from use. A fitted classifier can decode labels from answer states, including from the plain transcript control, while Qwen's own correct-option margin remains statistically neutral for every speech condition. The explicit text tag, by contrast, produces a large margin only after block 24. This is a functional localization to a routing/readout mismatch, not a causal localization to a particular head or subspace. Table 8 summarizes what the combined evidence supports and what remains open.

Table 8. Mechanistic bounds after the Stage 8 internal analysis. Layer-wise probing and attention constrain the explanation but do not replace causal intervention inside Qwen3-8B.

Observation	Supported Interpretation	Not Established
Raw WavLM linear Probe-K is 0.424, versus 0.226 after R1.8	Substantial stress information is lost before the LLM	The LLM interface is the only or dominant bottleneck
Absolute speech-slot MLP-2 accuracy rises to 0.425–0.467	No abrupt loss of absolute slot decodability is observed	That the decoded directions causally affect the answer
R1.8–R0 simultaneous band includes zero at all 37 states	No reliable propagation of the seed-1234 R1.8 advantage is detected	Population-level equivalence of internal trajectories
Speech B_{36} margins are near zero; the text tag is +6.405 nats	The frozen LLM uses explicit textual stress but not the tested speech prefixes	A specific causal head, layer, or subspace mechanism
Per-token and total slot attention give different orderings	Attention comparison is sensitive to the sevenfold slot-length difference	A speech-specific attention sink
Stage 6/7 pilots do not improve Probe-G	The tested LoRA and styled-teacher bridges are insufficient	Failure of other targets, ranks, seeds, or adaptation regimes

Taken together, the results are more consistent with a *modality-alignment or readout mismatch* than with generic task incapacity: Qwen3-8B acts on an explicit text tag, whereas none of the continuous speech prefixes produce the corresponding answer margin. This phrase is a functional description, not a claim that we have identified the responsible geometry. The attention audit weakens a simple speech-specific attention-sink explanation. The late left/system concentration is shared by speech and text conditions, while the speech–text slot ordering changes between total and per-token attention. Causal activation patching, head-level interventions, or subspace ablation would be required to determine which internal paths control the decision.

5.4. Value and Boundary of the End-to-End Negative Result

By the practical criterion that the generated answer must change, the tested system does not solve sentence-stress reasoning. We now state this directly. The contribution is diagnostic rather than a claim of deployment improvement: the controlled results show that an increase in adapter decodability is not sufficient evidence of useful speech-LLM behavior. This distinction prevents a Probe-K gain from being reported as end-to-end progress, rules out two specific bridging configurations as explanations for the reported run, and narrows the next design question from “is stress present anywhere?” to “how must an acoustic contrast be aligned so that it causally steers the response?” The negative outcome is therefore scientifically informative, while its immediate practical value remains limited.

5.5. What Our Findings Add to the Cited Literature

We organize the contribution by closely related work to make the delta explicit. vs. StresSLM [13].

StresSLM is the most direct competitor on the same Stress-17K data. It reaches 86.2% SSR by pairing encoder training with LoRA on Qwen2Audio W^Q/W^V . Our work establishes the strictly frozen–frozen lower bound (Stage 2 R0: cohort linear Probe-K 0.211, sub-cascade) and isolates the consumer-side LoRA contribution under the frozen-encoder + frozen-adapter constraint (Stage 6: minimal LoRA does not improve Probe-G in single-seed pilots; rank 8 degrades the oracle consistency check); neither test exists in the StresSLM publication. The frozen–frozen $\rightarrow$ encoder-trained gap on the same dataset is therefore decomposable through our Stages 2 + 6 plus StresSLM as the trained-encoder upper bound.

vs. Kang et al. [15].

They demonstrate that frozen LLMs *can* perceive paralinguistics under *encoder training* and style-conditioned text-target distillation. Our Stage 7 provides a pilot frozen-encoder ablation of their recipe, adopting their alignment-target structure (transcript + bracketed paralinguistic tag) but holding the encoder frozen. In that single-seed pilot, Probe- G_{neutral} remains inside the R0 equivalence band despite K_T^{styled} showing the LLM is capable in text form. To our knowledge, this frozen-encoder ablation is not present in published work; it suggests that encoder-side flexibility may contribute substantially to the positive result in Kang et al. [15].

vs. Billa [10] and Billa [12].

The Cascade-Equivalence Hypothesis establishes that text representations are causally necessary in deployed speech-LLM hidden states via LEACE + logit-lens at the LLM-internals layer. Our Probe-K diagnostic starts one layer upstream, at the adapter output, and Stage 8 follows the stress label through every Qwen3-8B block. The complementary finding is that absolute speech-slot stress remains decodable without becoming decision-relevant, while the seed-matched R1.8-over-R0 advantage is not reliable at any internal state. Unlike LEACE, our internal analysis is correlational and does not establish causal necessity. Billa [12]’s Text Dominance Ratio on closed-source models is corroborative on the consumer-side text bias.

vs. Wang et al. [2] (BLSP-KD’s PLoRA).

BLSP-KD uses PLoRA on the LLM but *also trains the encoder*; its scope is explicitly lexical-content alignment (the authors note paralinguistic preservation as next-step future work). Our Stage 6 takes up this follow-up specifically and tests whether minimal consumer LoRA alone (without encoder training) is sufficient in two single-seed pilot settings; it is not under those settings.

5.6. Limitations

We report the following bounds on the claim:

- **Five-seed cohort.** The R1.8 cohort spans five seeds; cohort σ estimation has limited statistical power. The MLP-2 cohort-mean crossing of the K_T floor is driven by 3/5 seeds (seeds 1234, 1236, 1237 clear; seeds 1235 and 1238 fall below); the Probe- G_{neutral} stay-at-baseline result holds across all five seeds. A larger cohort would be required to bound the per-seed recovery rate with tighter confidence.
- **Probe-capacity mismatch against K_T .** The K_T floor is a pre-specified linear text-only baseline, whereas the strongest R1.8 cohort result uses an MLP-2 probe. Stage 8 adds a capacity-matched seed-1234 diagnostic (R1.8 0.314 vs. K_T -plain 0.264 at E_0), but there is no independently trained text-model cohort from which to estimate model-seed uncertainty. We therefore do not use this post hoc single-model comparison to claim cohort-level superiority over K_T ; the primary positive claim remains the paired MLP-2 improvement over R0 plus the Stage 4 controls.
- **R0/R1.8 schedule and augmentation mismatch.** R1.8 differs from R0 in the counterfactual objective and in the later-stage warmup and augmentation configuration. The Stage 4 text-only and audio-shuffled controls show that the observed MLP-2 gain requires audio-label alignment, but they do not fully isolate $\mathcal{L}_{\text{cf}}$ from all schedule and augmentation changes. A matched augmented R0 would be needed for that narrower causal attribution.
- **StressPresso $n = 202$.** Bootstrap CIs (1000 iters, transcript-id clusters) are reported on every Probe-G number; effect-size considerations bind tighter than n does. We do not run a power analysis.

- **Single LLM (Qwen3-8B).** Cross-LLM modularity (Qwen3-14B, Qwen3-32B, Llama-3.1-8B) is part of the broader research agenda but not tested here. We do not claim our findings generalize across LLM scale or family.
- **Single encoder (WavLM-L16).** STT-pre-output cells (Whisper-large-v3 layer 22 or layer 32; Qwen3-ASR encoder) are deferred to follow-up work. Stage 1c's diagnostic appendix already flags that the bottleneck pattern $\text{WavLM-best} > \text{Whisper-22} > \text{Whisper-32} \approx K_T$ is not observed on Whisper-large-v3.
- **Single LoRA target set and rank range.** q_proj/v_proj at ranks $\{4, 8\}$ only. Alternative target modules (k_proj , o_proj , MLPs) and rank 16 are not tested. Rank 16 was not run after the rank 8 pilot failed the oracle consistency check.
- **Single-seed bridging diagnostics.** The Stage 6 LoRA and Stage 7 styled-teacher results are informative pilot diagnostics, not population-level negative results. We use them to bound the evaluated settings and avoid broader claims that these intervention families cannot work.
- **Single-model internal trace.** Stage 8 compares the seed-1234 R0, R1.8, and Stage 7 checkpoints. Its transcript-cluster bootstrap quantifies item uncertainty for these fixed models, but it does not estimate variability across independently trained model seeds. Probe seeds 1235/1236 test classifier optimization only.
- **Internal probes and attention are correlational.** Stage 8 traces all 36 Qwen3-8B blocks and finds no discrete loss of absolute slot decodability, but it does not use activation patching, head-level intervention, or subspace ablation. The direct text control also differs from speech in modality and slot length. Attention conclusions change between total and per-token normalization, and the broad late left/system concentration is shared across conditions. We therefore do not claim a causal discard layer or a speech-specific attention sink.
- **Stress-17K-raw is OpenAI-TTS-only** (nova/echo voices). Generalization to a diverse-voice training set is not tested. Likewise, sentence stress is only one paralinguistic phenomenon; emotion, speaking style, emphasis beyond the StressTest construction, speaker traits, and other discourse cues may behave differently.
- **Cross-domain absolute floors fail.** The C.2 in-domain and cross-domain absolute Probe-K accuracies stay at ~ 0.20 , below the desc-only baseline of 0.361. The audio-derived Φ -encoding is reliably above the K_T floor in absolute MLP-2 terms on the TTS subsets but does not transfer to real recordings (StressPresso) in absolute magnitude; the relative ratio condition (cohort 0.910, Stage 7 0.970) is what passes.

5.7. Future Work

- **Causal localization in the LLM.** Extend the Stage 8 trace with activation patching, head-level interventions, and subspace ablation on the same counterfactual pairs. These tests could distinguish a causally inactive decoded subspace from attention diversion or a downstream readout mismatch.
- **Encoder-side training** (the SpeechEmotionLlama path [15]). Test whether unfreezing the speech encoder under our R1.8 + styled-teacher regime bridges the consumer-side gap.
- **Discrete prosody tokens** (the ProsodyLM [29]/Spirit-LM lineage). Test whether word-level prosody tokens replace continuous-prefix injection as the audio-to-LLM-state alignment.
- **Multi-stage curriculum.** Train under R1.8 first (frozen $\rightarrow$ encoder fine-tuning $\rightarrow$ minimal LoRA on the LLM). This is hypothesized to bridge under tight oracle constraints.
- **Alternative LoRA target modules and ranks.** Test k_proj , o_proj , and MLPs at ranks $\{4, 8, 16, 32\}$ for direct comparison with the LoRA-rank-trade-off literature [18].

- **PCGrad/GradVac for $\mathcal{L}_{\text{NCE-cond}}$ recovery [20,21].** Test whether per-component gradient projection recovers the orthogonal-direction signal that uniform clipping cannot.
- **Larger LLMs (Qwen3-14B, Qwen3-32B).** Probe-G-oracle is expected to be higher; the $\text{Gap-G} = \text{Probe-K}(M) - \text{Probe-G}(M) / \text{Probe-G-oracle}$ may close with scale.
- **Extension to low-resource languages.** Our group's prior work on Kazakh speech recognition [30] and on small-scale Kazakh language models trained with direct preference optimization [31] provides a natural test bed: under the frozen–frozen recipe, a Kazakh-tuned encoder paired with a Kazakh-aware small LLM would let us measure whether the audio-to-LLM-state bottleneck is recipe-dependent or shifts with encoder/LLM coverage of the target language.

6. Conclusions

We studied whether a frozen–frozen speech-to-LLM adapter can preserve and use sentence-stress information beyond the transcript. The baseline BLSP-style adapter reduces linear Probe-K below the text-only K_T baseline on all five seeds. The complete R1.8 training configuration improves MLP-2 Probe-K over R0 with a significant paired seed test and passes three falsification controls, supporting an audio-derived representational gain. Because R1.8 also changes the warmup schedule and augmentation pipeline, this improvement is not attributed to the counterfactual objective alone. However, the numerical crossing of the pre-specified linear K_T floor is not statistically established across seeds and is not a capacity-matched comparison to an MLP-2 text probe.

At the response level, $\text{Probe-G}_{\text{neutral}}$ remains unchanged between R0 and R1.8. The LoRA and styled-teacher interventions are single-seed diagnostics, so we interpret them only within the evaluated pilot settings: under the evaluated settings, they do not improve Probe-G, while direct text-tag baselines show that the frozen LLM can use the same information in text form. Thus, the evaluated system does not provide an end-to-end solution to sentence-stress reasoning. Its contribution is a controlled diagnosis: representational recovery at the adapter is not sufficient evidence of useful generated behavior. The resulting interpretation is narrower than a claim that the handoff is the only bottleneck. The trained adapter still loses much of the stress information available in raw WavLM, and the residual signal that R1.8 makes nonlinearly decodable is not reliably consumed through continuous-prefix injection. Stage 8 finds that absolute speech-slot stress remains decodable throughout Qwen3-8B, but the seed-matched R1.8-over-R0 advantage is not reliable at any state and no speech condition develops a reliable answer margin. The explicit text tag does. Because total and per-token attention give different speech–text orderings and late left/system concentration is shared across conditions, the analysis does not support a speech-specific attention-sink mechanism.

For frozen–frozen speech-LLM implementations, transcript-only behavior alignment should therefore be treated cautiously when the target behavior depends on paralinguistic information. Future work should test whether encoder-side flexibility [15], discrete prosody tokens [29], or alternative injection mechanisms organize the acoustic signal in a form that the LLM can use.

The artifact-license summary is cited here. Table 9 summarizes the licenses and release status of each reproducibility artifact.

Table 9. Artifact license clarity for the released reproducibility bundle.

Artifact	License	Releasable?
Code (training/eval scripts)	BSD-3-Clause	Yes
Adapter checkpoints (R1.8 5 seeds; Stage 6 LoRA r4/r8; Stage 7)	code BSD-3-Clause; weights inherit CC-BY-NC-4.0 from training datasets	Yes for non-commercial research
cf_pairs JSONL files (transcript IDs and Φ indices; no raw audio)	CC-BY-NC-4.0 (derived from Stress-17K-raw)	Yes
Audio data (StressPresso, Stress-17K-raw, Espresso)	per-dataset upstream licenses	No, not redistributed; users access via HuggingFace Hub
Per-seed evaluation logs (summary.json, rows.json)	BSD-3-Clause	Yes
Stage 8 derived probes, attention aggregates, audits, and figure scripts	BSD-3-Clause	Yes; cached hidden states are not duplicated

Author Contributions: Conceptualization, N.K. and M.M.; methodology, N.K.; software, N.K.; validation, N.K. and M.M.; formal analysis, N.K.; investigation, N.K.; data curation, N.K.; writing—original draft preparation, N.K.; writing—review and editing, N.K. and M.M.; visualization, N.K.; supervision, M.M.; project administration, M.M.; funding acquisition, M.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Ministry of Science and Higher Education of the Republic of Kazakhstan, grant number BR24993001, “Creation of a large language model (LLM) to maintain the implementation of Kazakh language and increase the technological progress”.

Institutional Review Board Statement: Not applicable. This work uses only publicly released datasets (StressPresso, Stress-17K-raw, Espresso, LibriSpeech-clean-100) under their stated licenses. No new human-subjects data was collected.

Informed Consent Statement: Not applicable.

Data Availability Statement: The full reproducibility artifact bundle—pinned environment (requirements.txt), locked configuration (configs/locked_cells.yaml), counterfactual-pair JSONL files, adapter checkpoints (R0 cohort 5 seeds, R1.8 cohort 5 seeds, Stage 6 LoRA rank-4/rank-8, Stage 7 styled-teacher), per-seed evaluation JSON files, and a single-command smoke test—is available at <https://github.com/Nurgali-Kadyrbek/frozen-speech-llm-stress> (accessed on 27 July 2026). The Stage 8 revision additionally includes the derived layer-wise trajectories, per-row probe predictions, attention rows and aggregates, checkpoint/model provenance, numerical audit, and F9/F10 generation scripts under results/stage8_internal_analysis and paper/figures/scripts. Adapter checkpoints and the derived cf_pairs dataset are released as a HuggingFace collection at <https://huggingface.co/collections/nur-dev/beyond-transcript-alignment-69f7af0198a059e1d6b9bd5f> (accessed on 27 July 2026). License clarity for released artifacts is summarized in Table 9.

Acknowledgments: The authors gratefully acknowledge support from the Ministry of Science and Higher Education of the Republic of Kazakhstan, and computational resources provided by Al-Farabi Kazakh National University.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Locked-Protocol Disclosure: All training decisions for Stages 2–7 are pinned in configs/locked_cells.yaml, an append-only per-stage configuration file released with the artifact bundle. Each stage’s section was added when that stage’s experiments closed; thresholds and λ schedules are visible in the per-stage branch_thresholds entries and were not modified post hoc. Cross-stage temporal ordering is verifiable from the released training_logs/stage*_results.md file timestamps. The Stage 7 equivalence-band outcome follows the pre-specified stopping criterion for a pilot that did not

improve Probe-G. Stage 8 is explicitly recorded as a post-review, post hoc frozen-model diagnostic rather than part of the original lock. `paper/numbers_audit.json` confirms 54/54 original claims, and the separate Stage 8 audit confirms 50/50 additional claims within tolerance. This disclosure describes a locked sequential audit trail plus a labeled post hoc analysis, not an externally timestamped immutable preregistration.

Ethics/Misuse Statement: Sentence-stress preservation in speech-LLMs could in principle contribute to deception detection or coercion-aware applications. We study perception only, do not produce a deployable model and do not release synthesized voice. Future deployment in any of these adjacent applications should be preceded by deployment-context ethical review.

References

1. Wang, C.; Liao, M.; Huang, Z.; Lu, J.; Wu, J.; Liu, Y.; Zong, C.; Zhang, J. BLSP: Bootstrapping Language-Speech Pre-training via Behavior Alignment of Continuation Writing. *arXiv* **2023**, arXiv:2309.00916. <https://doi.org/10.48550/arXiv.2309.00916>.
2. Wang, C.; Liao, M.; Huang, Z.; Zhang, J. BLSP-KD: Bootstrapping Language-Speech Pre-training via Knowledge Distillation. *arXiv* **2024**, arXiv:2405.19041. <https://doi.org/10.48550/arXiv.2405.19041>.
3. Wang, C.; Liao, M.; Huang, Z.; Wu, J.; Zong, C.; Zhang, J. BLSP-Emo: Towards Empathetic Large Speech-Language Models. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024), Miami, FL, USA, 12–16 November 2024; pp. 19186–19199. <https://doi.org/10.18653/v1/2024.emnlp-main.1070>.
4. Fan, R.; Ren, B.; Hu, Y.; Zhao, R.; Liu, S.; Li, J. AlignFormer: Modality Matching Can Achieve Better Zero-shot Instruction-Following Speech-LLM. *IEEE J. Sel. Top. Signal Process.* **2025**, *19*, 1329–1337. <https://doi.org/10.1109/JSTSP.2025.3588378>.
5. Deng, K.; Sun, G.; Woodland, P. Wav2Prompt: End-to-End Speech Prompt Learning and Task-based Fine-tuning for Text-based LLMs. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Albuquerque, NM, USA, 29 April–4 May 2025; pp. 6940–6956. <https://doi.org/10.18653/v1/2025.naacl-long.354>.
6. Lu, K.H.; Chen, Z.; Fu, S.W.; Huang, H.; Ginsburg, B.; Wang, Y.C.F.; Lee, H.Y. DeSTA: Enhancing Speech Language Models through Descriptive Speech-Text Alignment. In Proceedings of the Interspeech 2024, Kos Island, Greece, 1–5 September 2024; pp. 4159–4163. <https://doi.org/10.21437/Interspeech.2024-457>.
7. Lu, K.H.; Chen, Z.; Fu, S.W.; Yang, C.H.H.; Balam, J.; Ginsburg, B.; Wang, Y.C.F.; Lee, H.Y. DeSTA2: Developing Instruction-Following Speech Language Model Without Speech Instruction-Tuning Data. In Proceedings of the ICASSP 2025–2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 6–11 April 2025; pp. 1–5. <https://doi.org/10.1109/ICASSP49660.2025.10889444>.
8. Lu, K.H.; Chen, Z.; Fu, S.W.; Yang, C.H.H.; Huang, S.F.; Yang, C.; Yu, C.; Chen, C.; Huang, C.; Lin, Y.; et al. DeSTA2.5-Audio: Toward General-Purpose Large Audio Language Model with Self-Generated Cross-Modal Alignment. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2026**, *34*, 2062–2076. <https://doi.org/10.1109/TASLPRO.2026.3675792>.
9. Ma, Z.; Yang, G.; Yang, Y.; Gao, Z.; Wang, J.; Du, Z.; Yu, F.; Chen, Q.; Zheng, S.; Zhang, S.; et al. Speech Recognition Meets Large Language Model: Benchmarking, Models, and Exploration. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2025), Philadelphia, PA, USA, 25 February–4 March 2025; Volume 39, pp. 24840–24848; SLAM-ASR; Issue 23. <https://doi.org/10.1609/aaai.v39i23.34666>.
10. Billa, J. The Cascade Equivalence Hypothesis: When Do Speech LLMs Behave Like ASR→LLM Pipelines? *arXiv* **2026**, arXiv:2602.17598. <https://doi.org/10.48550/arXiv.2602.17598>.
11. Chen, J.; Guo, Z.; Chun, J.; Wang, P.; Perrault, A.; Elsnér, M. Do Audio LLMs Really LISTEN, or Just Transcribe? Measuring Lexical vs. Acoustic Emotion Cues Reliance. In Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), Rabat, Morocco, 24–29 March 2026; pp. 5848–5877. <https://doi.org/10.18653/v1/2026.eacl-long.274>.
12. Billa, J. When Audio-LLMs Don't Listen: A Cross-Linguistic Study of Modality Arbitration. *arXiv* **2026**, arXiv:2602.11488. <https://doi.org/10.48550/arXiv.2602.11488>.
13. Yosha, I.; Maimon, G.; Adi, Y. StressTest: Can YOUR Speech LM Handle the Stress? In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2026, San Diego, CA, USA, 2–7 July 2026; pp. 1264–1285. <https://doi.org/10.18653/v1/2026.findings-acl.64>.
14. Tang, C.; Yu, W.; Sun, G.; Chen, X.; Tan, T.; Li, W.; Lu, L.; Ma, Z.; Zhang, C. SALMONN: Towards Generic Hearing Abilities for Large Language Models. In Proceedings of the 12th International Conference on Learning Representations (ICLR 2024), Vienna, Austria, 7–11 May 2024.

15. Kang, W.; Jia, J.; Wu, C.; Zhou, W.; Lakomkin, E.; Gaur, Y.; Sari, L.; Kim, S.; Li, K.; Mahadeokar, J.; et al. Frozen Large Language Models Can Perceive Paralinguistic Aspects of Speech. In Proceedings of the Interspeech 2025, Rotterdam, The Netherlands, 17–21 August 2025; pp. 4323–4327. <https://doi.org/10.21437/Interspeech.2025-166>.
16. Ryumina, E.; Axyonov, A.; Abdulkadirov, T.; Koryakovskaya, D.; Ryumin, D. Cross-Lingual Bimodal Emotion Recognition with LLM-Based Label Smoothing. *Big Data Cogn. Comput.* **2025**, *9*, 285. <https://doi.org/10.3390/bdcc9110285>.
17. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. In Proceedings of the 10th International Conference on Learning Representations (ICLR 2022), Virtual, 25–29 April 2022.
18. Rathore, D.; Kumar, V.; Bansal, C.; Moitra, A. How Much is Too Much? Exploring LoRA Rank Trade-offs for Retaining Knowledge and Domain Robustness. In Proceedings of the Findings of the Association for Computational Linguistics: IJCNLP-AACL 2025, Mumbai, India, 20–24 December 2025; pp. 1003–1013. <https://doi.org/10.18653/v1/2025.findings-ijcnlp.58>.
19. Du, Y.; Czarnecki, W.M.; Jayakumar, S.M.; Farajtabar, M.; Pascanu, R.; Lakshminarayanan, B. Adapting Auxiliary Losses Using Gradient Similarity. *arXiv* **2018**, arXiv:1812.02224. <https://doi.org/10.48550/arXiv.1812.02224>.
20. Wang, Z.; Tsvetkov, Y.; Firat, O.; Cao, Y. Gradient Vaccine: Investigating and Improving Multi-task Optimization in Massively Multilingual Models. In Proceedings of the 9th International Conference on Learning Representations (ICLR 2021), Virtual, 3–7 May 2021.
21. Yu, T.; Kumar, S.; Gupta, A.; Levine, S.; Hausman, K.; Finn, C. Gradient Surgery for Multi-Task Learning. In Proceedings of the Advances in Neural Information Processing Systems 33 (NeurIPS 2020), Virtual, 6–12 December 2020.
22. Pasad, A.; Chou, J.C.; Livescu, K. Layer-Wise Analysis of a Self-Supervised Speech Representation Model. In Proceedings of the 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Cartagena, Colombia, 13–17 December 2021; pp. 914–921. <https://doi.org/10.1109/ASRU51503.2021.9688093>.
23. Chen, S.; Wang, C.; Chen, Z.; Wu, Y.; Liu, S.; Chen, Z.; Li, J.; Kanda, N.; Yoshioka, T.; Xiao, X.; et al. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. *IEEE J. Sel. Top. Signal Process.* **2022**, *16*, 1505–1518. <https://doi.org/10.1109/JSTSP.2022.3188113>.
24. Yang, S.w.; Chi, P.H.; Chuang, Y.S.; Lai, C.I.J.; Lakhotia, K.; Lin, Y.Y.; Liu, A.T.; Shi, J.; Chang, X.; Lin, G.T.; et al. SUPERB: Speech Processing Universal PERformance Benchmark. In Proceedings of the Interspeech 2021, Brno, Czechia, 30 August–3 September 2021; pp. 1194–1198. <https://doi.org/10.21437/Interspeech.2021-1775>.
25. Bentum, M.; ten Bosch, L.; Lentz, T.O. Word stress in self-supervised speech models: A cross-linguistic comparison. In Proceedings of the Interspeech 2025, Rotterdam, The Netherlands, 17–21 August 2025; pp. 251–255. <https://doi.org/10.21437/Interspeech.2025-106>.
26. Yosha, I.; Shteyman, D.; Adi, Y. WhiStress: Enriching Transcriptions with Sentence Stress Detection. In Proceedings of the Interspeech 2025, Rotterdam, The Netherlands, 17–21 August 2025; pp. 4718–4722. <https://doi.org/10.21437/Interspeech.2025-1475>.
27. Al-Radhi, M.S.; Csapó, T.G.; Németh, G. Continuous vocoder applied in deep neural network based voice conversion. *Multimed. Tools Appl.* **2019**, *78*, 33549–33572. <https://doi.org/10.1007/s11042-019-08198-5>.
28. Yang, A.; Qwen Team. Qwen3 Technical Report. *arXiv* **2025**, arXiv:2505.09388. <https://doi.org/10.48550/arXiv.2505.09388>.
29. Qian, K.; Fan, X.; Ni, J.; Shechtman, S.; Hasegawa-Johnson, M.; Gan, C.; Zhang, Y. ProsodyLM: Uncovering the Emerging Prosody Processing Capabilities in Speech Language Models. In Proceedings of the Second Conference on Language Modeling (COLM 2025), Montreal, QC, Canada, 7–10 October 2025.
30. Kadyrbek, N.; Mansurova, M.; Shomanov, A.; Makharova, G. The Development of a Kazakh Speech Recognition Model Using a Convolutional Neural Network with Fixed Character Level Filters. *Big Data Cogn. Comput.* **2023**, *7*, 132. <https://doi.org/10.3390/bdcc7030132>.
31. Kadyrbek, N.; Tuimebayev, Z.; Mansurova, M.; Viegas, V. The Development of Small-Scale Language Models for Low-Resource Languages, with a Focus on Kazakh and Direct Preference Optimization. *Big Data Cogn. Comput.* **2025**, *9*, 137. <https://doi.org/10.3390/bdcc9050137>.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.