VOICE ACTING ARENA: EVALUATING EXPRESSIVE SPEECH ACROSS ACTING CHALLENGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Expressive text-to-speech models increasingly follow instructions, yet authentic voice acting remains difficult to compare: saying the right words is not the same as performing a scene. We introduce the *Voice Acting Arena*, a human-preference evaluation for complex acting directions. Its challenges span 40 fine-grained emotion targets, selected vocal-delivery directions, and multi-step scenes across ten acting domains. A language-model director translates each challenge into the controls supported by a particular speech system; listeners then compare anonymized takes on overall acting, instruction fit, authenticity, and, when requested, a vocal reaction. In a first round of 560 listener votes on eleven model entries, Gemini 3.1 Flash TTS leads on every question. Our *Humaneness Voice* baseline, released with commercially usable open weights, sits in the middle: with ten-candidate selection its rating cannot be separated from OpenAI GPT Audio 1.5, DramaBox, or Cartesia Sonic 3.5. To curate its training data and rank candidate takes, we train a *Genuineness Score* for spontaneous, authentic-sounding speech and a *Vocal Burst Blending Score* for how naturally laughs, sighs, and similar bursts fit the surrounding speech. A first listener check supports the genuineness score only weakly and does not yet support the blending score, so neither is treated as a measure of acting. We release the challenge design, evaluation protocol, bilingual data views, and baseline components, and keep live descriptive results separate from a frozen benchmark.

1 INTRODUCTION

A performer can say every word correctly and still miss the scene. A character may thank someone while keeping professional composure over suppressed hurt, or shift from irritation to relief within one sentence. The words are easy to transcribe; the acting is in timing, breath, emphasis, and the change of intention. A laugh can be recognizable yet sound pasted onto the line around it. Such a nonverbal reaction is a *vocal burst*; its fit with neighboring speech matters as much as its presence. These distinctions matter in conversation, care and support interfaces, games, film, and advertising. Perceived genuineness may help a conversational voice avoid an uncanny performance; scene-specific instruction following matters whenever an author or director needs a particular arc. Neither implies that a generated voice has feelings, nor that perceived sincerity reveals a human speaker's honesty.

Calm under pressure
Scene. A dispatcher hears a missing child on the radio; her terrified parent is listening.
Direction. Let relief show in one breath, but stay markedly restrained: measured pace, narrow pitch, attentive rather than sleepy.
Words. "I hear her. She's answering us. Stay with me while the team reaches her."

A reunion with unfinished hurt
Scene. A daughter finds her estranged father backstage after her debut.
Direction. Begin guarded; let a small, disbelieving laugh escape after "You came," then let gratitude emerge through the hurt.
Words. "You came. I told myself I didn't care if you watched. I kept looking for you anyway."

Figure 1: Two newly written Arena examples. One tests composure in a frightening moment; the other moves from guarded hurt to gratitude through a brief laugh. Both fix the spoken words and specify their delivery. Neither reproduces a held-out task. Twelve more appear in Appendix B.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Evaluation has lagged behind that shift from reading to acting. Word error rate asks whether the script survived. Broad emotion classification can miss subtext, mixtures, and transitions. A mean opinion score may conflate sound quality, attractive timbre, direction following, and acting. We therefore make the *scene* the unit of evaluation. A model receives the same fixed words and an acting challenge; human listeners compare finished takes and judge different aspects of the performance separately. This follows the logic of pairwise system arenas (Chiang et al., 2024) while making the prompt landscape and performance axes explicit for speech.

Our first contribution is a documented 374-task challenge design containing 240 emotion tasks, 34 vocal-delivery tasks, and 100 multi-step scenes. Its EmoNet labels go beyond basic emotions; selected VoiceNet directions describe how speech is delivered; scenes exercise subtext, physical constraints, social context, and change over time. The second contribution is an Arena protocol that adapts each challenge to a model's documented interface before a blinded A/B comparison, then asks about overall acting, instruction fit, authenticity, and a requested vocal reaction. We report Bradley–Terry ratings with bootstrap intervals for eleven entries from a first round of 560 listener votes. Third, we release a separate, broader pool of 19,247 public acting briefs for training and scene generation, while keeping Arena stimuli held out. Finally, we provide Humaneness Voice, an inspectable open-weight baseline, together with bilingual annotated data views and a candidate-selection recipe. Its perceptual scores are tools inside the baseline, not substitutes for listener judgment.

The paper first describes the Arena landscape and comparison protocol, then the baseline, released data and separate public challenge pool, and finally the live results and limitations. The appendix documents aggregate Arena construction, public data provenance, scorer training, adapter and prompting studies, duration controls, layer-forensics analyses, and unsuccessful interventions. The held-out Arena scenes are not printed there. This division keeps the central question visible: can a speech system perform a direction convincingly, rather than merely synthesize a plausible-sounding line?

2 A CHALLENGE LANDSCAPE FOR VOICE ACTING

The Arena begins with a task, not an audio clip. A task specifies a scene context, a take-level performance direction, and the exact words to speak. Its challenge may be an emotion, a delivery attribute, or an integrated acting scene. This design tests more than whether a system can produce a generic "sad" or "excited" voice: it asks whether a change in emotion, intention, pace, or physical state is audible at the right moment. The archived v0.2 registry fixes 374 such tasks for the design described here; coverage by taxonomy does not by itself establish equal item difficulty.

Table 1: Construction of the archived 374-task v0.2 registry. Counts describe authored tasks, not distinct situations or completed human judgments.

Track	Coverage rule	Targets	Tasks
Fine-grained emotion	40 EmoNet emotion categories; two strengths and three different scenarios for each	40	240
Vocal delivery	17 selected VoiceNet high/low directions from 13 dimensions; two strengths each	17	34
Acting scenes	Ten authored scenes in each of ten domains, with a fixed script and performance arc	100	100
Total	Three complementary ways to probe expressive control	214	374

From taxonomies to playable directions. EmoNet names 40 fine-grained emotions beyond a small basic-emotion set (Schuhmann et al., 2025b;a). VoiceNet (Anonymous, 2026e) is a separate vocabulary of audible delivery properties, such as arousal, tension, emphasis, vulnerability, and speaking style; our tools use 57 of its 59 dimensions (Appendix H.3). The Arena selects 17 directional targets, not the entire VoiceNet rubric. For *low arousal*, the voice should have restrained energy; *intense* says that this restraint should be unmistakable, not that the actor should sound louder or more excited. In the constructed dispatcher example in Figure 1, measured pace and limited pitch movement realize the low-arousal target despite an urgent situation. A moderate-strength version would make the same restraint subtler. Emotion tasks vary their scenarios so that a label is not

synonymous with a particular script. The separate 289-item Situation Taxonomy describes a wider space of physical and social conditions, but the Arena builder does not attach its IDs to these tasks. We do not claim a systematic sample from all 289 situations.

Integrated scenes. The 100 scene challenges are evenly distributed across Emergency, Intimacy, Conflict, Comedy, Suspense, Public, Embodied, Fantasy, Everyday, and Transformation. They include audition-like constraints: conceal hurt beneath politeness, change status during a confrontation, keep a line intelligible while breathless, or turn a laugh into grief. Each has an original fixed three- or four-sentence script of 20–80 words. Requested nonverbal events are stored separately with placement information rather than written as words to speak; 27 of the 374 tasks require such a reaction. This preserves a common spoken target while allowing each model's interface to express the event appropriately. Appendix A documents the registry construction and limitations, without publishing held-out prompts or scripts.

The builder merges the authored task parts and checks IDs, counts, schema, sentence and word limits, and near-duplicate text. These checks make the registry internally consistent; they are distinct from listener judgments of how clearly and convincingly a task can be performed. We therefore keep the 374-task design count separate from the completed-comparison count in Section 3.

3 From a challenge to a blinded comparison

Model-specific direction. A common acting brief is not a common executable prompt. Systems differ in whether they accept a reference voice, global style text, local inline tags, timing controls, or a separate nonverbal-event field. In the Arena workflow, GLM-5.3-Flash (Z.ai, 2026) reads the task and each target system's README or interface instructions, then recasts the acting direction into that system's supported prompt form. The fixed spoken script is preserved. A requested laugh or pause can be expressed through a documented control, while dialogue changes and model-specific revisions after listening are excluded by the protocol. This translation layer is part of the evaluation treatment, not a property of the speech model alone.

The complete path is *task* $\rightarrow$ *model-specific prompt* $\rightarrow$ *generated audio* $\rightarrow$ *anonymized A/B task* $\rightarrow$ *axis-specific listener choice*. We keep the original task visible for instruction-fit judgments, rather than asking listeners to judge a model's private prompt. Generated audio is the object of comparison; the language-model director does not vote. A system without a requested control can still attempt a take. Model versions, voice policies, decoding, failed generations, and generation budgets are therefore relevant evaluation settings, not incidental implementation details.

Table 2: The Arena's four separate A/B judgments. Listeners may choose a tie on any axis; the reaction question appears only when the scene requests one.

Judgment		Question and evidence
01	**Overall acting**	Which take is the more convincing finished performance? Judge its choices and emotional arc as a whole.
02	**Instruction fit**	Which take better realizes the scene's emotion, intensity, subtext, pace, transitions, and requested action?
03	**Authenticity**	Which delivery feels more genuine through timing, prosody, breath, hesitation, and integrated emotion?
04	**Vocal reaction**	When requested, which take better places and blends a laugh, sob, gasp, scream, or similar event?

Listener procedure. The displayed rubric asks a listener to read the scene and script and hear both full clips. They answer overall preference first, then reset and judge each remaining axis independently. A tie is allowed where no advantage is meaningful. Low confidence is local to one judgment; changed words, truncation, unintelligibility, corruption, and spoken stage directions are flagged rather than silently traded against other strengths. Close calls can be replayed. Accepted comparisons are paid at 15 cents after a human quality review; the Arena interface records, but does not send, payments. These safeguards do not replace inter-rater reliability estimates or an audit of listening behavior.

Evaluation data. Section 8 uses the vote-level export of 25 September 2026. All 560 submitted votes passed quality review. Seven paid raters contributed 352 votes and 63 community volunteers 208. Most tasks were judged once (456 of 507), so listener agreement can only be measured at the system level. Only 46 votes fall on 35 tasks of the archived 374-task registry. The other 514 use tasks from a larger pool whose identifiers follow the same emotion, vocal-delivery, and acting tracks. We fit one Bradley–Terry model per axis, with ties as half a win; Appendix A gives the estimator, the listener pool, and robustness checks.

4 HUMANENESS VOICE: AN OPEN-WEIGHT BASELINE

The Arena is independent of our generator. To make its challenges actionable with an inspectable system, we built Humaneness Voice as a modular baseline. A scene contains fixed words, context, and acting direction; the system returns audio that should preserve the words and perform the direction. Its speech-model lineage begins with the Apache-2.0 MOSS-TTS-Local-Transformer-v1.5 (Gong et al., 2026). The released weights from our third supervised fine-tuning round (SFT3) are under CC-BY-4.0, a commercially usable open-weight license; the recipe is documented in Appendix G. This does *not* make every underlying training audio file or external scoring dependency uniformly permissive.

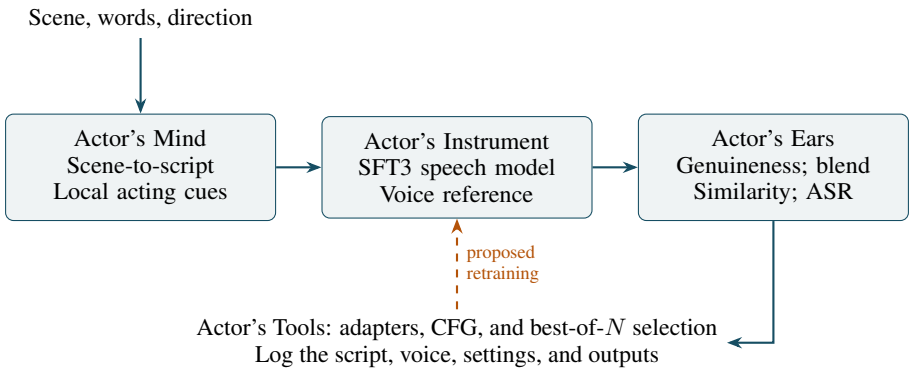

Figure 2: The baseline's mind interprets a scene, its instrument generates takes, its ears inspect them, and its tools alter generation or choose among takes. The dashed retraining path is a proposed future experiment.

Mind and instrument. An LLM director turns the brief into a schema-checked script with global delivery, sentence- or word-group cues, pauses, durations, and requested vocal events. The server derives the spoken transcript from the same script and maps semantic instructions to the model's timed tokens. The fine-tuned 4.55B parameter instrument then generates English or German takes. Local timing matters: a small pause inside a sentence can make an emotional turn audible, though the effect of any prompt convention needs listener testing. WikiSkills records prompt and tool recipes, including failures (Tang et al., 2026a; Anonymous, 2026c). The Arena's GLM prompt translator (Section 3) is a distinct evaluation component: it adapts a common task to *each* contender, not just ours.

Ears and tools. Two small predictors on a frozen VoiceCLAP-Commercial embedding estimate perceived genuineness and how naturally a vocal burst blends into the speech around it (Anonymous, 2026d;b;a). The latter concerns the emotional and prosodic continuity *before and after* a laugh, sob, gasp, or similar event, not merely whether the event exists. VoiceCLAP audio–description similarity and ASR word error enter the audited ranker; Empathic Insight Voice Plus and VoiceNet help the director and diagnostics. Specialist LoRAs and optional CFG are tools, but their effects are not attributed to the SFT3 weights alone. The exact reward is given in Section 7; scorer-training and control ablations are in the appendix.

The implemented loop performs one constrained directing pass, candidate generation, and optional best-of-N choice. It does not implement autonomous multi-round rehearsal.

Codec speech models and instruction-controlled TTS already generate expressive voices (Chen et al., 2025; Gong et al., 2026; Guo et al., 2023; Yang et al., 2024; Zhou et al., 2024); speech role-playing systems also combine reasoning and performance controls (Tang et al., 2026b). Our claim is narrower: a documented script, tool recipe, candidate selector, and failure log are connected to a multi-axis human evaluation. The Arena remains useful if a stronger generator replaces this baseline.

5 BILINGUAL DATA FOR REPRODUCIBLE FOLLOW-UP WORK

Large expressive corpora such as Expresso and EARS make spontaneous styles and nonverbal speech accessible (Nguyen et al., 2023; Richter et al., 2024). We release English and German annotated views intended to make future voice-acting training and evaluation less dependent on undisclosed corpora. Records can include transcripts, timing, codec tokens, emotion and delivery attributes, and perceptual scores. Real and synthetic sources are both represented. These are resources for future controlled studies, not all direct inputs to the Humaneness Voice checkpoint.

Table 3: Annotated audio inventory by source and access terms (10 September snapshot). The research subtotal is 103,912 published view-hours under custom noncommercial research-access terms, not CC-BY-NC. CC-BY-4.0 on annotations does not license the underlying audio. Profile views and source performances overlap; rows must not be summed as unique speech. Mean duration is derived from rounded hours. Source licences and training views are detailed in Appendix E.7.

Dataset source / view	Records	Hours	Mean (s)	Audio terms
Public: Emolia, Common Voice, Kartoffelphon, MLS	91,936,029	228,749	8.96	CC-BY-4.0 / CC0 by source
Public: EuroSpeech	9,119,466	37,916	14.97	Parliament-specific terms
Public: Voice-profile VC candidates	6,508,056	17,016	9.41	Synthetic derivative; check source terms
Public branch subtotal	107,563,551	283,681	–	–
Research: Podcast	26,606,892	96,838	13.10	Custom NC research; upstream undeclared
Research: Snippets	2,633,037	4,875	6.67	Custom NC research; upstream undeclared
Research: Eva snippets	500,007	2,199	15.83	Custom NC research; mixed provenance
Research subtotal: custom NC use	29,739,936	103,912	–	–
Synthetic voice-profile takes Generated takes and identity-repair rerenders (original + repair)	28,212,933	71,056	9.07	CC-BY-4.0
Selected voice-converted profile takes VC + SIDON restoration (vc_sidon)	20,078,645	54,811	9.83	CC-BY-4.0

What the hour counts mean. The public and gated research branches report 283,681 and 103,912 hours. The research branch contains 96,838 podcast hours selected from publicly indexed podcast episode links and processed with ASR, emotion, and quality annotation; 4,875 hours of filtered, enhanced AudioSnippets-derived speech; and 2,199 hours of differently selected reference snippets. A UID audit traces 98.56% of sampled reference snippets to the podcast branch, so these are not independent AudioSnippets hours. All three views remain under custom gated noncommercial research terms because source-audio rights are not established by downstream annotation licences (Appendix E.7). The public branch contains 228,749 hours from sources declaring CC-BY-4.0 or CC0, 37,916 EuroSpeech hours with parliament-specific terms, and 17,016 synthetic voice-profile voice-conversion (VC) candidate hours whose derivative provenance needs checking. The public and research branches plus the default profile view arithmetically sum to 458,649 *repository-view* hours, not deduplicated training hours. Identity-repair rerenders replace failed original takes; VC changes voice timbre and a subsequent SIDON pass restores the converted audio. The public VC candidates overlap source performances in the profile release; its separate 54,811-hour selected VC-and-restoration view is not added again. Annotated acting scenes lack verified published duration. Thus "over 400,000 hours" describes reported views, not a uniformly commercially reusable corpus.

From a compact fine-tuning set to a scaling resource. Humaneness Voice was fine-tuned on selected expressive examples rather than on the entire annotated inventory. Building that selection

required collecting, scoring, and filtering a much wider English and German pool. We release the resulting scaling ladder so that this larger resource can support future *from-scratch* voice-acting models and controlled studies of training scale. Table 4 lists cumulative, nested training views from 100 to 98,920.1 measured hours, largest first. Each smaller view is a subset of the larger views; a separate 80.1-hour holdout is excluded throughout. Within language and real/synthetic cells, selection reserves half the volume for 40 EmoNet categories and half for 57 VoiceNet dimensions across ten value bins. Genuineness, burst-blend, and target-attribute scores guide selection (Appendix E).

Table 4: Released German–English scaling ladder, descending by measured cumulative training duration. Core, Expanded, and Broad are descriptive labels for stable loader configurations, not additional disjoint datasets. The 80.1-hour holdout is separate.

Cumulative view (loader config)	Measured hours	Clips
Broad (`stage100`)	98,920.1	34,112,433
Expanded (`stage50`)	49,569.2	17,135,901
Core (`tier7`)	20,001.5	6,666,998
Tier 6 (`tier6`)	10,000.7	3,425,099
Tier 5 (`tier5`)	5,000.4	1,764,990
Tier 4 (`tier4`)	2,500.1	912,789
Tier 3 (`tier3`)	1,000.0	382,768
Tier 2 (`tier2`)	500.1	197,662
Tier 1 (`tier1`)	249.9	100,858
Tier 0 (`tier0`)	100.0	41,771

What the ladder enables. The timed-script corpus contains 3,147,802 rows, of which 398,282 selected expressive rows entered the final SFT3 fine-tuning round (Appendix E). The much larger ladder is a separate dataset contribution: researchers can train a new speech backbone at multiple scales, refine it for acting, and evaluate it in the Arena under the same task rubric. Its nested construction also permits broad-to-selective curricula without changing the holdout. The largest released training view measures 98,920.1 hours, 1,082 below its nominal 100,000-hour target. Each record retains its source rights; training under commercial-use constraints requires a rights-filtered view with freshly measured duration.

6 An open challenge space for data generation

Evaluation requires protected test scenes; training requires a broad supply of situations that give a voice a reason to change. We therefore release an Acting Challenge Dataset of 19,247 English *performance briefs*, separate from the Arena's private task instructions and spoken scripts. A brief describes an actor's problem rather than a fixed utterance: conceal jealousy in a professional exchange, speak through exertion, or let relief interrupt anger. A director can turn a brief into new dialogue and model-specific controls, then generate several takes for supervised training, candidate selection, or internal evaluation. Publishing the brief space makes such data generation easier to vary systematically without exposing the Arena test items. The situations may also inspire visual scenes, although this paper evaluates speech alone.

The seed pool contains 278 audition-style briefs from an earlier expressive-acting dataset associated with EmoNet-Voice (Schuhmann et al., 2025a), 1,000 synthetic thematic scenarios, and 200 VoiceNet vocal-attribute prompts. Five contextual variations of each of these 1,478 seeds add 7,390 briefs. A separate route translates actor-training concepts, techniques, and exercises from a frozen 806-item index of Eric Morris's acting method into 4,030 original scene directions. Physical and social situations contribute 5,749 briefs, while 600 extreme-physical briefs probe large changes in breath, exertion, pain, temperature, or startle. The repository retains provenance and a family-balanced sampling view. These are 19,247 usable briefs, not 19,247 independent human-authored ideas. Appendix F explains the construction and the boundary between the public resource and the held-out Arena.

7 TRAINING AND SELECTING THE BASELINE

The Humaneness Voice instrument is a fine-tuned MOSS-TTS Local Transformer lineage, not a model trained from scratch on the full resource inventory. Earlier voice-acting and timed-format stages lead to SFT3; its final full-parameter round used 398,282 selected expressive rows for two epochs and 712 updates (Appendix G). A larger timed-script corpus has 3,147,802 rows, but the final round is not a claim of training on every row. The model supports reference voices and local duration and delivery markup. A director can place an emotional cue or short pause before a word group, allowing a sentence to change intent instead of carrying one global style label. The 12-codebook, 48-kHz codec runs at 12.5 frames per second; precise obedience to every internal duration instruction is not established. Architecture, stage lineage, timing tests, and adapter recipes are detailed in Appendix D.

Perceptual selection. The two VoiceCLAP-Commercial heads estimate genuineness G and vocal-burst integration B. They were trained on Gemini 3.1 Pro weak labels from several TTS datasets; their published held-out checks measure agreement with those labels rather than human perceptual validity. A first matched-pair listener snapshot gives only partial support (Appendix A, Table A.3). The genuineness head picks the listener's choice in 39 of 61 decisive pairs (64%), a margin that rests largely on one of five raters. The blend head matches in 31 of 67 (46%), which is no better than chance. In generation, the selector also uses audio–description similarity C and ASR word error W. For N takes of the *same* script, it applies within-set normalization and chooses

$$R_0(y_i) = \big[\widetilde{G}(y_i) + \widetilde{B}(y_i) + 2\widetilde{C}(y_i)\big]\big[1 - \min\{1, \max\{0, W_i\}\}\big], \qquad y^\star = \arg\max_{i \le N} R_0(y_i). \quad (1)$$

A constant term across a candidate set is assigned 0.5. An optional revision adds evidence for a specifically requested reaction near its planned position. That event detector is different from B: a laugh can be present and still blend poorly. Although emotion and delivery predictors help with direction and diagnosis, no separate target-emotion score appears in Equation 1. Thus the ranker can overlook a subtle emotional turn even when its chosen audio sounds credible. Appendix L specifies the implementation and its diagnostic controls.

Interpretation of pass@10. Best-of-ten expands the search from one to ten takes and selects with Equation 1. The Arena reports both generation budgets, making their rating gap an end-to-end result for the system. Because the two entries share only 25 tasks and meet different opponents, the gap does not isolate the effect of a particular reward term. It does show how the baseline performs when its instrument and perceptual selector are used together.

8 WHAT THE LIVE ARENA CURRENTLY SHOWS

Table 5 and Figure 3 report Bradley–Terry ratings fitted to the vote-level export of 25 September 2026: 560 pairwise votes by 70 listeners on 507 tasks, with each of the 55 system pairs compared 6 to 18 times. Ratings use the Elo scale, where 1000 is the geometric-mean entry and a 100-point gap means a 64% expected preference. Ties count as half a win for each side. Rating intervals resample votes. Rank ranges resample listeners, which accounts for the concentration of votes: four listeners supplied 57% of them. The reaction axis uses only the 177 votes on tasks that requested a vocal burst.

The leader. Gemini 3.1 Flash TTS leads on all four questions. It wins 70 of 85 overall comparisons, loses 10, ranks first in every listener resample, and stays first when the four most active listeners are removed (Appendix A). Its margin is larger among paid listeners (1399) than among community volunteers (1161), whose ratings do not separate it from Breeze, ElevenLabs, or OpenAI GPT Audio 1.5. The training data, compute, voice choice, and prompting of the entries are not controlled, so a ranking cannot tell us why Gemini leads. We have no evidence attributing the result to any specific Google data source.

Where the baseline sits. Humaneness Voice with ten-candidate selection rates 980 overall. Cartesia Sonic 3.5, DramaBox, and OpenAI GPT Audio 1.5 rate between 986 and 999 with almost identical intervals, and the listener-resampled ranks of all four fall between third and eighth. The baseline is fourth on authenticity (1008) and fifth on instruction fit (1000), Against Gemini, ElevenLabs, and Breeze its model-implied win probabilities are 0.14, 0.31, and 0.37, and each listener-bootstrap

Table 5: Voice Acting Arena ratings from the 25 September 2026 export. Entries are sorted by overall rating. n is the number of comparisons; Win % is the overall win rate with ties as half. The overall interval is a 95% vote-bootstrap interval, and Rank is the 95% range over listener resamples. Fit, Authentic, and Reaction give ratings on the other axes; Figure 3 shows their intervals. Humaneness Voice appears as "MOSS Voice Acting 4.55B" in the Arena.

			Overall acting				
Model	n	Win %	Elo (95% CI)	Rank	Fit	Authentic	Reaction
Gemini 3.1 Flash TTS	85	85	1292 (1211–1403)	1	1213	1274	1389
ElevenLabs Eleven v3	95	67	1118 (1060–1188)	2–4	1087	1071	1060
Breeze TTS 2 (Voice Design)	100	61	1073 (1010–1141)	2–4	1075	1097	1122
Cartesia Sonic 3.5	108	51	999 (938–1060)	3–8	941	982	960
Resemble AI DramaBox	102	50	987 (924–1048)	4–8	1023	974	1052
OpenAI GPT Audio 1.5	113	49	986 (935–1043)	3–8	998	986	987
Humaneness Voice, pass@10	**102**	**48**	**980 (918–1042)**	**4–8**	**1000**	**1008**	**967**
MiniMax Speech 2.8 HD	110	42	925 (866–978)	5–11	913	903	969
OpenAI GPT-4o mini TTS	116	37	902 (838–959)	7–11	922	908	818
Humaneness Voice, pass@1	95	37	898 (832–958)	7–11	927	946	764
Fish Audio S2.1 Pro	94	29	841 (768–899)	9–11	899	852	912

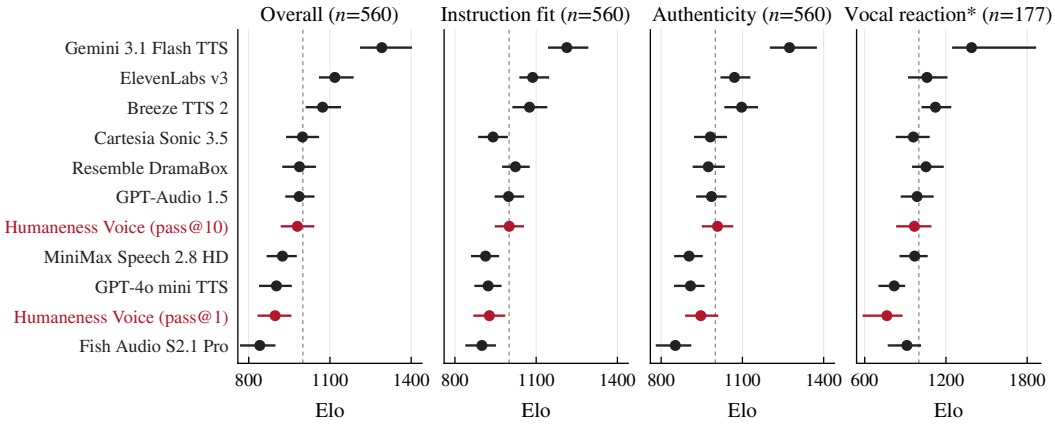

Figure 3: Ratings with 95% vote-bootstrap intervals for each listener question; the dashed line marks the average entry (1000). *Vocal reaction uses only the 177 votes on tasks that request a burst. Humaneness Voice rows are in red.

interval lies below one half. Its checkpoint has commercially usable CC-BY-4.0 weights, whereas Breeze's published weights carry noncommercial restrictions (BreezeBlue, 2026). We did not audit the weight licences of the other entries, and this comparison concerns released weights, not every data view or auxiliary component.

Requested reactions and candidate selection. The dashboard of 24 September divided reaction wins by all comparisons, including tasks that requested no burst, which put every entry between 6% and 26%. Counting only eligible votes, pass@10 wins 14, loses 15, and ties 2 (967). The single-take pass@1 entry is last (764; 4 wins, 20 losses), while Gemini wins 21 of 24. Selecting one take from ten raises the rating on every axis: by 82 points overall, 73 on fit, 62 on authenticity, and 203 on reactions. In the 13 direct pass@10 versus pass@1 votes, pass@10 wins 10 and loses 2 (sign test $p = 0.04$). The overall gap has a listener-bootstrap interval of 16 to 170 but a vote-bootstrap interval of -6 to 167, and the authenticity gap is not separable from zero under either. The two entries share only 25 of their tasks and meet different opponents. The gap therefore describes the deployed selection recipe end to end; it is not a matched ablation of Equation 1.

9 DISCUSSION AND LIMITATIONS

What the Arena reveals. The four judgments separate a preferred performance from one that follows the direction, sounds authentic, or handles a requested nonverbal reaction. The top and bottom of the table are broadly stable across questions, but the middle reorders. DramaBox is fourth on instruction fit and reactions yet seventh on authenticity; Humaneness Voice with selection is seventh overall but fourth on authenticity. A plausible line can still fail a laugh or sob, and the single-take baseline shows this most clearly: it is last on reactions but eighth on fit and authenticity, and selection produces its largest gain on exactly that question. Gemini 3.1 Flash TTS leads everywhere, which shows that strong acting under these directions is already possible, but the ranking does not reveal whether the advantage comes from training data, architecture, voice choice, prompting, or compute.

What remains uncertain. The first round is small and concentrated. Four paid listeners supplied 318 of the 560 votes, and all seven paid listeners are native English-speaking US undergraduates aged 18 to 24. Removing the four most active listeners leaves the top three unchanged but moves OpenAI GPT Audio 1.5 from sixth to fourth and Humaneness Voice pass@10 from seventh to eighth, so middle ranks should be read as ranges. Paid and community listeners agree moderately: the Spearman correlation of their overall ratings is 0.73 (vote-bootstrap interval 0.25 to 0.86) and only 0.48 on authenticity. A shared ranking is not rejected at the 5% level (likelihood-ratio $p = 0.06$, listener permutation $p = 0.09$), yet paid listeners rate Gemini 238 points higher and GPT Audio 1.5 170 points lower than volunteers do. Who listens is part of the measurement. Because each pair of clips was judged once, we cannot estimate item-level reliability, and the recorded confidence field holds the same value in every vote. Most votes use an extended task pool rather than the archived 374-task registry, twenty registry tasks were used in earlier director development (Appendix A), and the task-to-audio join is not yet frozen. The pass@10 and pass@1 entries share only 25 tasks, so their gap is evidence for the deployed recipe, not for the ranking formula or either perceptual head. Those heads are weak on their own terms: genuineness orders listener pairs correctly 64% of the time and blend 46%, on 61 and 67 decisive choices from a handful of raters. The export cannot tell whether the large reaction gain from selection comes from the blend term, the requested-event revision, or the word-error filter. Perceived authenticity may also depend on recording quality, voice familiarity, and genre expectations, and a predictor trained on model labels can inherit its annotator's preferences.

A reusable research path. The released Humaneness Voice weights are CC-BY-4.0 and start from Apache-2.0 MOSS-TTS v1.5. The much larger annotated inventory is not a deduplicated, uniformly permissive pretraining corpus; source rights differ and SFT3 used a curated subset. The public Acting Challenge Dataset supplies new performance briefs without publishing held-out Arena scripts, while the nested scaling ladder makes controlled training-scale studies possible. A from-scratch model built from rights-cleared views could then be compared on the same four Arena questions. New systems, including the later Gemini 3.8 Flash TTS family (Google, 2026), belong in that comparison once sufficient Arena judgments are collected; they are not folded into the present table.

Conclusion. Voice acting calls for more than correct words or a single emotion label. This work pairs a taxonomy-guided human comparison with an inspectable open-weight baseline and public resources for building future contenders. The current Arena is a useful measurement instrument, not a certificate that any model has solved genuine, scene-sensitive speech. Keeping its evaluation scenes distinct from data-generation prompts and training audio is the basis for improving both the models and the measurements.

REPRODUCIBILITY STATEMENT

The appendix documents data lineage, training formats, model and scorer versions, control sweeps, negative results, and evaluation protocols, and it includes per-dimension probe and adapter tables. Anonymized copies of the director and inference code, the released models, and the released data views are listed with their access conditions in Appendix M.1. Held-out Arena scripts and participant-level votes are not released during review. The aggregate evidence files, source hashes, and the scripts that regenerate every table and figure will be released with the final version.

ETHICS STATEMENT

Voice cloning and expressive synthesis can support creative work and accessibility, but can also enable impersonation or deceptive presentation. The system name and genuineness score describe perceived delivery, not a speaker's humanity, mental state, honesty, or identity. Users of released audio must follow source-specific rights and access conditions. Dataset annotations are imperfect model predictions and should not be treated as personal or clinical facts. The Arena reports only aggregate judgments in this paper, without participant identities.

Human listeners. Both listener studies are minimal-risk preference tasks: participants hear synthetic or already released speech and choose between two clips. Before the first Arena comparison, every participant reads a participation agreement. It describes the task, states that participation is voluntary and can be stopped at any time, and lists the data stored for quality review: responses, playback coverage and timing, IP address, browser identifier, focus changes, copy attempts, and aggregate keyboard and pointer statistics. It also states that identifiable data are restricted to study administrators and that public reports use aggregate or pseudonymous results. Participants confirm that they are at least 18 years old and consent to participation and data collection. Community volunteers sign in with an existing Hugging Face account and are not paid. Seven paid raters use study credentials and receive 15 cents per comparison accepted in quality review. The study was run by a non-profit research organization that has no institutional review board. It was not submitted for formal ethics review, which the organization's jurisdiction does not require for non-medical studies of this kind. The analyzed exports replace raters with pseudonymous codes and contain no names, account IDs, access codes, IP addresses, browser identifiers, timestamps, or free text; demographics are reported only in aggregate.

AI USE STATEMENT

Generative AI was used to synthesize speech and text, generate weak annotations, and direct model-specific prompts. For the manuscript, the authors used Claude Opus 5.5 (Anthropic) and OpenAI Astra 6 to support source review, literature lookup, English drafting, evidence organization, LaTeX preparation, and figure code. The Arena table and figure are recomputed by the released scripts from the vote-level export. The authors checked the generated text and references and are responsible for all claims and interpretations.

REFERENCES

Anonymous. VoiceCLAP-Commercial Vocal-Burst Blend: Model card, 2026a. URL https://anonymous-hf.com/a/xdsygvc9zcfb/. Anonymized for review. Training-set size, label provenance, and split construction are not specified in the card.

Anonymous. VoiceCLAP-Commercial Genuineness: Model card, 2026b. URL https://anonymous-hf.com/a/km3933c74q97/. Anonymized for review. Training and validation facts checked against the model card; not human validation.

Anonymous. Humaneness Voice: Demo server, director, and wikiskills, 2026c. URL https://anonymous.4open.science/r/I271. Anonymized snapshot of the audited commit aa299f1, September 10, 2026.

Anonymous. VoiceCLAP-Commercial (small): Frozen audio encoder, 2026d. URL https://anonymous-hf.com/a/km3933c74q97/. Anonymized for review. Frozen audio-text backbone, shipped inside the genuineness and vocal-burst blending head repositories.

Anonymous. VoiceNet: A taxonomy of 59 perceptual speech dimensions, 2026e. Manuscript in preparation.

BreezeBlue. Breeze TTS 2: Model and inference code, 2026. URL https://huggingface.co/BreezeBlue/Breeze-TTS-2. Accessed September 10, 2026. Code: https://github.com/breezeblue-ai/breeze-tts.

Sanyuan Chen, Chengyi Wang, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. Neural codec language models are zero-shot text to speech synthesizers. *IEEE Transactions on Audio, Speech and Language Processing*, 33:705–718, 2025. doi: 10.1109/TASLPRO.2025.3530270. arXiv:2301.02111.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 8359–8388, 2024. URL https://proceedings.mlr.press/v235/chiang24b.html.

David C. Funder and Daniel J. Ozer. Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, 2(2):156–168, 2019. doi: 10.1177/2515245919847202. URL https://doi.org/10.1177/2515245919847202.

Gilles E. Gignac and Eva T. Szodorai. Effect size guidelines for individual differences researchers. *Personality and Individual Differences*, 102:74–78, 2016. doi: 10.1016/j.paid.2016.06.069. URL https://doi.org/10.1016/j.paid.2016.06.069.

Yitian Gong, Botian Jiang, Yiwei Zhao, Yucheng Yuan, Kuangwei Chen, Yaozhou Jiang, Cheng Chang, Dong Hong, Mingshu Chen, Ruixiao Li, Yiyang Zhang, Yang Gao, Hanfu Chen, Ke Chen, Songlin Wang, Xiaogui Yang, Yuqian Zhang, Kexin Huang, ZhengYuan Lin, Kang Yu, Ziqi Chen, Jin Wang, Zhaoye Fei, Qinyuan Cheng, Shimin Li, and Xipeng Qiu. MOSS-TTS technical report. *arXiv preprint arXiv:2603.18090*, 2026. URL https://arxiv.org/abs/2603.18090.

Google. Gemini API release notes: 22 september 2026, 2026. URL https://ai.google.dev/gemini-api/docs/changelog. Announcement of Gemini 3.8 Flash TTS and Flash-Lite TTS; accessed September 24, 2026.

Zhifang Guo, Yichong Leng, Yihan Wu, Sheng Zhao, and Xu Tan. PromptTTS: Controllable text-to-speech with text descriptions. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10096285. arXiv:2211.12171.

Wen-Chin Huang, Erica Cooper, Yu Tsao, Hsin-Min Wang, Tomoki Toda, and Junichi Yamagishi. The VoiceMOS challenge 2022. In *Interspeech*, pp. 4536–4540, 2022. doi: 10.21437/Interspeech.2022-970. arXiv:2203.11389.

Tu Anh Nguyen, Wei-Ning Hsu, Antony D'Avirro, Bowen Shi, Itai Gat, Maryam Fazel-Zarani, Tal Remez, Jade Copet, Gabriel Synnaeve, Michael Hassid, Felix Kreuk, Yossi Adi, and Emmanuel Dupoux. Expresso: A benchmark and analysis of discrete expressive speech resynthesis. In *Interspeech*, pp. 4823–4827, 2023. doi: 10.21437/Interspeech.2023-1905. URL https://arxiv.org/abs/2308.05725.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 53728–53741, 2023. arXiv:2305.18290.

Julius Richter, Yi-Chiao Wu, Steven Krenn, Simon Welker, Bunlong Lay, Shinji Watanabe, Alexander Richard, and Timo Gerkmann. EARS: An anechoic fullband speech dataset benchmarked for speech enhancement and dereverberation. In *Interspeech*, pp. 4873–4877, 2024. doi: 10.21437/Interspeech.2024-153. URL https://arxiv.org/abs/2406.06185.

Christoph Schuhmann, Robert Kaczmarczyk, Gollam Rabby, Felix Friedrich, Maurice Kraus, Kourosh Nadi, Huu Nguyen, Kristian Kersting, and Sören Auer. EmoNet-Voice: A fine-grained, expert-verified benchmark for speech emotion detection. *arXiv preprint arXiv:2506.09827*, 2025a. URL https://arxiv.org/abs/2506.09827.

Christoph Schuhmann, Robert Kaczmarczyk, Gollam Rabby, Maurice Kraus, Felix Friedrich, Huu Nguyen, Krishna Kalyan, Kourosh Nadi, Kristian Kersting, and Sören Auer. EmoNet-Face: An expert-annotated benchmark for synthetic emotion recognition. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, volume 38, pp. 107134–107169, 2025b. doi: 10.52202/085713-3230. URL https://openreview.net/forum?id=BqVIt5Dxxh.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Liyan Tang, Cyrus Rashtchian, Chun-Sung Ferng, Andrew Tomkins, Da-Cheng Juan, and Tu Vu. WikiSkill: Compiling agent experience into persistent knowledge for skill evolution. *arXiv preprint arXiv:2608.27454*, 2026a. URL https://arxiv.org/abs/2608.27454.

Wenqiu Tang, Zhen Wan, Takahiro Komamizu, and Ichiro Ide. DeSRPA: Decoupled speech role-playing agent via inference-time intervention. In *Interspeech*, pp. 4281–4285, 2026b. URL https://www.isca-archive.org/interspeech_2026/tang26b_interspeech.html.

Dongchao Yang, Songxiang Liu, Rongjie Huang, Chao Weng, and Helen Meng. InstructTTS: Modelling expressive TTS in discrete latent space with natural language style prompt. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2913–2925, 2024. doi: 10.1109/TASLP.2024.3402088. arXiv:2301.13662.

Z.ai. GLM-5.3-Flash: Model card, 2026. URL https://huggingface.co/zai-org/GLM-5.3-Flash. Accessed September 25, 2026.

Yixuan Zhou, Xiaoyu Qin, Zeyu Jin, Shuoyi Zhou, Shun Lei, Songtao Zhou, Zhiyong Wu, and Jia Jia. VoxInstruct: Expressive human instruction-to-speech generation with unified multilingual codec language modelling. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 554–563, 2024. doi: 10.1145/3664647.3681680. arXiv:2408.15676.

APPENDIX: ARENA, DATA, AND VOICE-ACTING STACK

This appendix documents the current Arena snapshot and metric-study status first, then the system, released data, measurement instruments, and successful and unsuccessful interventions. Arena task content remains withheld; only aggregate construction statistics accompany the manuscript. The Arena summary is recomputed from the 25 September 2026 vote-level export, and the scorer check analyzes the 151 responses exported on the same date; both are preliminary. No protected audio or access credentials are included.

Reader's guide. A *take* is one generated audio candidate for a script; *best-of-N* chooses one from N such takes. A *vocal burst* is a nonverbal reaction such as a laugh, sob or gasp. *SFT3* names the third supervised fine-tuning round of the MOSS speech model; a *LoRA* is a small task- or voice-specific weight adapter, *CFG* is generation-time classifier-free guidance, and *DPO* is preference-based fine-tuning. EmoNet and VoiceNet refer to 40 emotion targets and 57 perceptual voice dimensions, respectively. These model-estimated labels and scores are not human ground truth.

A ARENA RESULTS AND HUMAN-STUDY PROTOCOLS

Evidence boundary. Arena results come from the vote-level export of 25 September 2026 (`evidence/arena_2026-09-25/`): 560 submitted votes, all accepted in quality review, together with the scripts that produce every rating, interval, and figure in this appendix. The export removes account identifiers, timestamps, and free text, and it omits the held-out task texts and audio. It supersedes the dashboard percentages of 24 September, which counted reaction wins over all comparisons. For the scorer-validation study we report the stimulus composition and a first descriptive snapshot of 151 listener responses exported on 25 September 2026. That snapshot is not the final validation.

A.1 ARENA EXPORT: COVERAGE, LISTENERS, AND ROBUSTNESS

Coverage. The 560 votes cover 507 tasks; 456 tasks received one vote, 49 two, and 2 three. All 55 pairs of the eleven entries were compared, 6 to 18 times each, so the comparison graph is complete. The reaction question applied to 177 votes. On the overall question 66 votes (11.8%) were ties. Listeners flagged a clip problem in 17 votes: unintelligible speech 9 times, changed words and spoken stage directions 6 times each, truncation 4, and corruption 3; flagged votes are kept. Every vote records at least one play of both clips. The confidence field holds the interface default of 3 in all 560 votes and carries no information. Only 46 votes fall on 35 tasks of the archived v0.2 registry (15 emotion, 21 vocal-delivery, and 10 acting votes). The remaining 514 use tasks with `SYN-` identifiers in the same three tracks (219 emotion, 141 vocal-delivery, 154 acting); their generator and relation to the public Acting Challenge briefs are not documented in the export. The pass@10 and pass@1 entries appear on 94 and 90 tasks, of which 25 are shared. These counts are in `coverage.json`.

Table A.1: Active audio pool shown in the 24 September administrator dashboard and comparison counts in the 25 September export. Available clips are not the number of judgments.

Model	Active clips	Available tasks	Comparisons
Gemini 3.1 Flash TTS	694	694	85
ElevenLabs Eleven v3	697	697	95
Breeze TTS 2 (Voice Design)	607	607	100
Cartesia Sonic 3.5	694	694	108
Resemble AI DramaBox	708	708	102
OpenAI GPT Audio 1.5	696	696	113
Humaneness Voice, pass@10	245	245	102
MiniMax Speech 2.8 HD	697	697	110
OpenAI GPT-4o mini TTS	697	697	116
Humaneness Voice, pass@1	245	245	95
Fish Audio S2.1 Pro	696	696	94

Listeners. Seventy listeners submitted votes (Figure A.1). The seven paid raters are US undergraduates aged 18 to 24 who report native English; they supplied 352 votes, with individual totals of 118, 82, 67, 51, 22, 7, and 5. The 63 community volunteers supplied 208 votes, at most 16 each, and most left the optional profile blank. Community volunteers were not paid; consent and review status are described in the ethics statement.

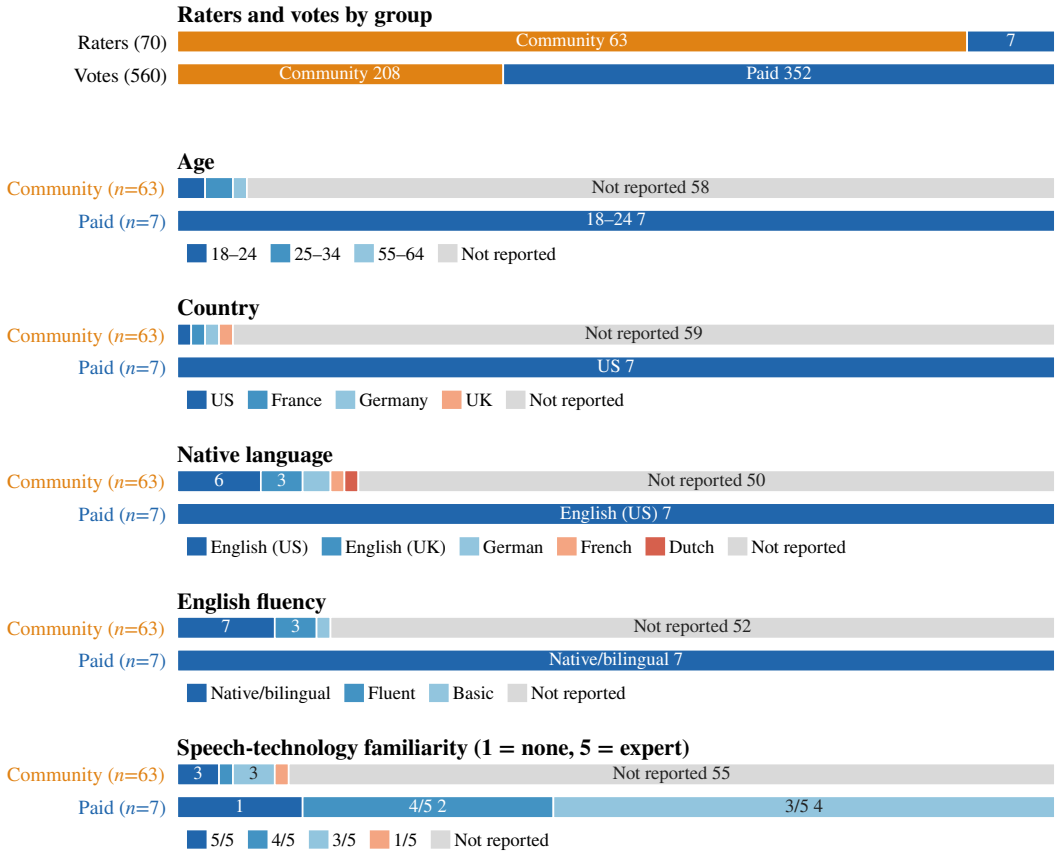

Figure A.1: Composition of listeners with at least one vote. Bars are 100% of each group; numbers are listener counts. The community profile was optional, and only attributes asked of both groups are shown.

Paid and community listeners. Figure A.2 fits the leaderboard separately for each group, and Figure A.3 compares them. The Spearman correlation between group ratings is 0.73 overall (vote-bootstrap interval 0.25 to 0.86), 0.79 for instruction fit, 0.48 for authenticity, and 0.75 for reactions. Split-half correlations on the overall question are 0.72 within the paid votes and 0.71 within the community votes, against 0.59 between a paid half and a community half, so the groups agree less with each other than with themselves. A likelihood-ratio test of one shared overall ranking against two separate rankings gives $\chi^2_{10} = 17.6$ ($p = 0.06$; listener permutation $p = 0.09$). Two per-entry differences have intervals excluding zero: paid listeners rate Gemini 238 points higher and OpenAI GPT Audio 1.5 170 points lower. A model fitted on one group predicts the other group's non-tied overall votes with 59% (paid model on community votes) and 67% accuracy (community model on paid votes).

Heavy listeners. The four most active listeners, all paid, supplied 318 of 560 votes. Figure A.4 compares vote- and listener-bootstrap intervals and refits without those four. Gemini, ElevenLabs, and Breeze keep ranks one to three, and Gemini's overall rating falls from 1292 to 1170. OpenAI GPT Audio 1.5 rises from sixth to fourth (986 to 1081), MiniMax from eighth to sixth, and Fish Audio from eleventh to ninth; Humaneness Voice pass@10 moves from seventh to eighth (980 to

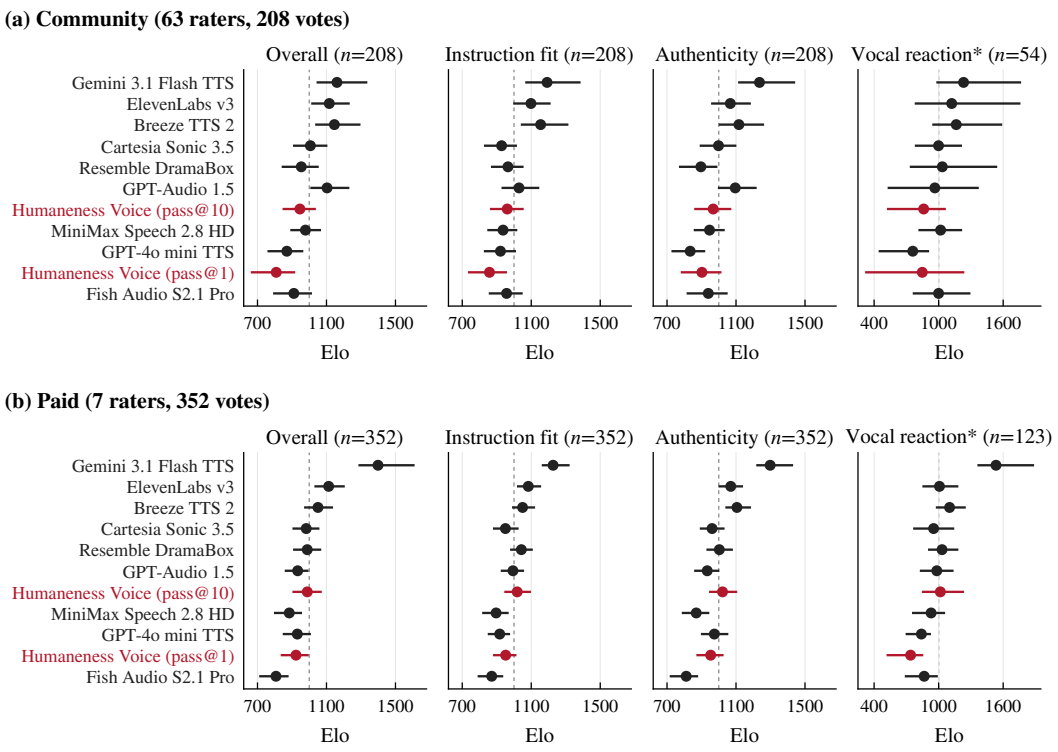

Figure A.2: The leaderboard fitted separately on (a) community and (b) paid votes, with entries in the order of the overall rating on all votes. Community reaction intervals are clipped at the axis edge.

960) and pass@1 from tenth to eleventh. Dropping any single heavy listener moves pass@10 by less than 5 points.

Humaneness Voice against each entry. Direct head-to-head records are small, 8 to 14 votes per pairing, so Figure A.5 also shows the win probability implied by the fitted model. With ten-candidate selection, the model-implied probability of beating Fish Audio S2.1 Pro is 0.69 and of beating MiniMax 0.58; against Gemini it is 0.14. Figure A.6 gives the pass@10 versus pass@1 contrast on each axis. Direct votes favour pass@10 by 10–2–1 overall (sign test $p = 0.04$), 9–1–3 on fit ($p = 0.02$), 9–3–1 on authenticity ($p = 0.15$), and 4–0 on reactions ($p = 0.13$). These four tests are not corrected for multiplicity.

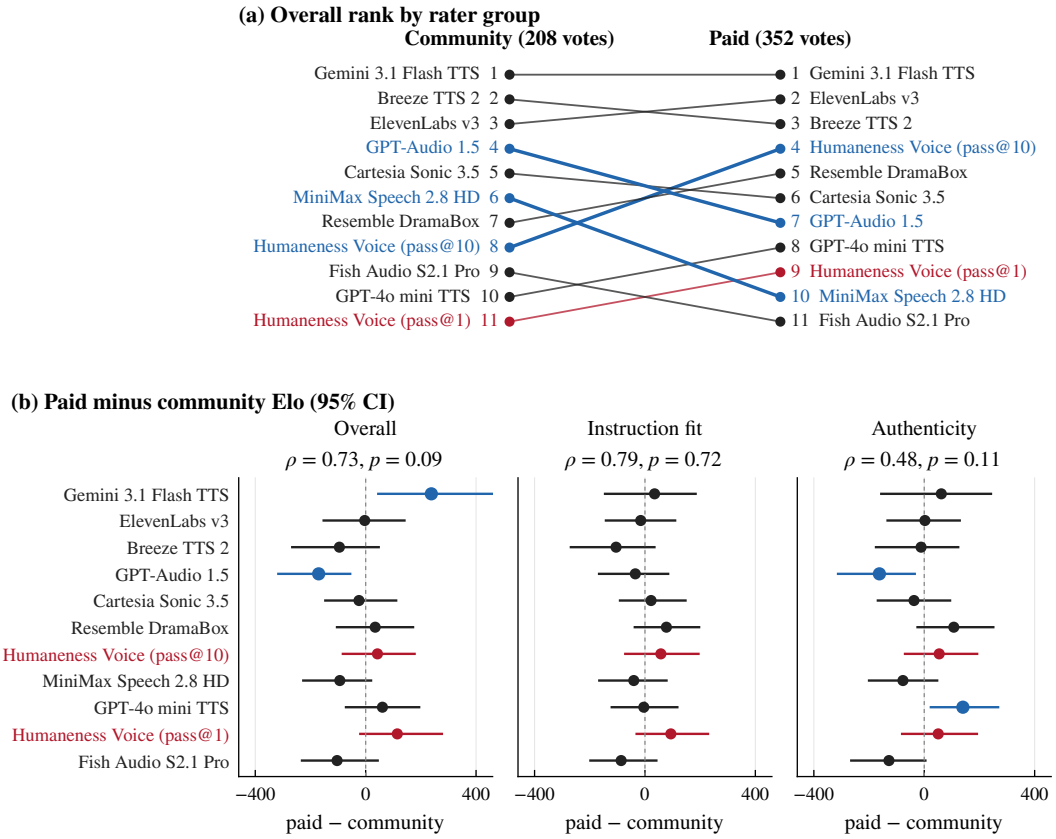

Figure A.3: (a) Overall rank from each listener group; blue lines move three or more places. (b) Paid minus community rating per entry with 95% intervals; blue intervals exclude zero. ρ is the Spearman correlation between the two groups' ratings; p is a listener-level permutation test of one shared ranking against separate rankings.

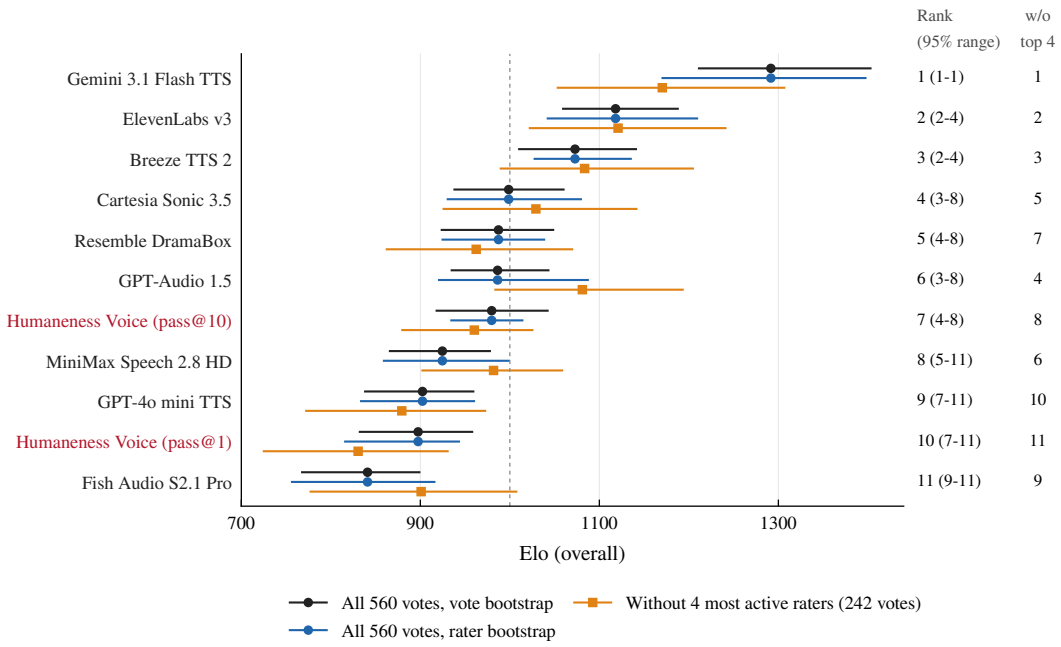

Figure A.4: Overall ratings with vote-bootstrap intervals, listener-bootstrap intervals, and after removing the four most active listeners (318 of 560 votes, all paid). Right: rank with its 95% range over listener resamples, and rank without the four listeners.

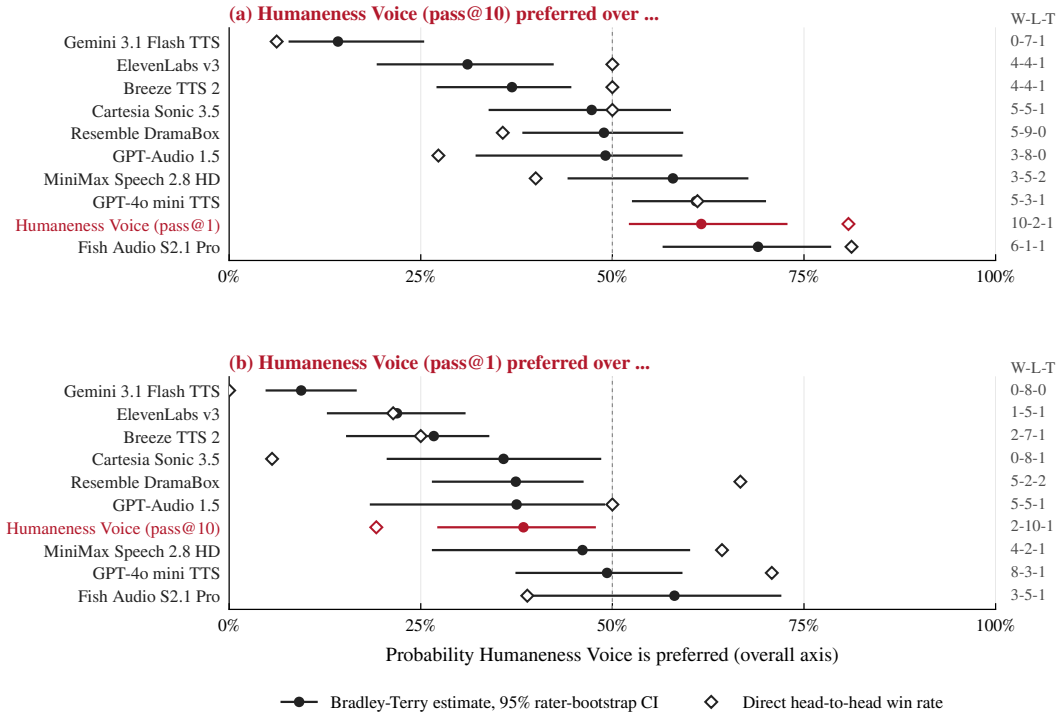

Figure A.5: Probability that Humaneness Voice with (a) ten-candidate and (b) single-take generation is preferred over each entry on the overall question. Circles: Bradley–Terry estimate with 95% listener-bootstrap interval. Diamonds: direct head-to-head win rate with ties as half, with the win–loss–tie record on the right.

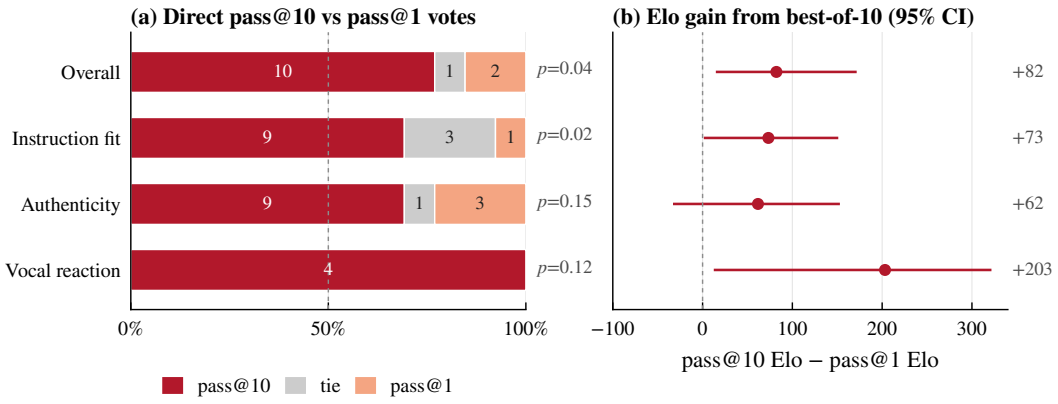

Figure A.6: Effect of ten-candidate selection. (a) Direct pass@10 versus pass@1 votes; p is a two-sided sign test on non-tied votes. (b) Rating gap from the full leaderboard with 95% listener-bootstrap interval.

A.2 STUDY 1: MATCHED-PAIR METRIC VALIDATION

The released study material contains four sets of 500 pairs, for 2,000 pairs and 4,000 distinct clips. The design deliberately avoids reusing clips, so a small collection of striking examples cannot dominate the item count. Each set has 250 synthetic and 250 real-speech pairs. Synthetic and real clips are not mixed within a pair, because a predictor can otherwise agree with listeners simply by distinguishing broad source categories.

Set	Target / native gap	Pairs	English	German	Median gap
A	Genuineness, $\Delta \geq 2$	500	216	284	2.33
B	Genuineness, $1 \leq \Delta \leq 1.5$	500	233	267	1.04
C	Blend, $\Delta \geq 2$	500	254	246	3.38
D	Blend, $1 \leq \Delta \leq 1.5$	500	245	255	1.21

Table A.2: Actual pair-set composition. These are stimulus counts, not completed vote counts or observed agreements. Genuineness is on 0–6 and blend on 0–10.

Within each pair, language, source half, and speaker identity where available are matched. Duration is 1.5–20 seconds and differs by no more than four seconds. Blend pairs additionally share a primary burst label and contain a burst in both clips. Selection is stratified over language, gender band, and age band, with a cap of six pairs per voice/speaker key. There are 1,169 distinct keys. A/B randomisation places the higher-scored clip in position A for 1,012 pairs, or 50.6%.

Matching voice reduces preferences for a timbre or speaker, but the words and exact scene are not matched. Lexical or contextual content can still affect authenticity and burst integration. The study estimates agreement on this controlled stimulus distribution; it should not be described as universal metric validity for all languages, voices, or acting genres.

Native scores, not corpus quantiles. The 0–1 corpus columns are per-voice or per-source/language quantile ranks, not linearly rescaled native scores. The native values were recovered and joined for all 3,147,802 corpus rows without missing keys. The easy and hard thresholds were then applied literally on the predictor's own scale.

This distinction changes which pairs qualify. Only about 11% of the selected hard pairs would satisfy a prior rank-based hard rule. In blend-hard set D, the rank ordering reverses the native ordering for 32 of 500 pairs, or 6.4%. Thus the exported native-score answer key, not an assumed quantile conversion, defines model agreement.

Listener questions. For genuineness, listeners should judge which sample sounds more like a spontaneous, authentic spoken moment, independent of whether the recording is cleaner or the emotion stronger. For blend, they should judge which burst fits more naturally into its surrounding speech, rather than which burst is louder or easier to identify. The interface should allow a tie or no discernible difference, hide model scores and source labels, randomise playback order, and permit replay.

The study analysis should describe recruitment, language proficiency, listening conditions, consent, compensation, numbers of raters and ratings, exclusion rules, missing responses, and time spent. The export analysed below pseudonymises raters and omits names, access codes, background answers, and free-text notes, so recruitment, proficiency, and listening conditions cannot yet be reported from it. Its elapsed-time field runs from assignment to answer and includes breaks, so it is not listening time.

Primary and secondary analysis. Report agreement with the higher native score separately for each of the four sets, with ties reported explicitly. A half-credit tie analysis can be a sensitivity analysis, not an undisclosed change in the denominator. Report per-language and per-source-half results; broad pooled agreement can hide a failed subgroup. Use a confidence procedure accounting for repeated listeners and shared speaker keys, such as a suitable multiway clustered bootstrap or hierarchical model.

A mixed-effects logistic model can relate a signed native-score difference to an A-over-B choice while accounting for listener and speaker effects and order. A positive slope tests whether larger model gaps are associated with stronger human preference. Pairwise responses alone do not supply independent absolute clip ratings, so they should not be used to invent a Pearson correlation between model and human scalar scores. If a suitable human latent ranking is estimated, a rank association can be reported with its identifiability and uncertainty assumptions.

Measure inter-listener reliability and disagreement before calling a metric invalid because it fails to reach 100% agreement. The hard gaps are close to the genuineness head's MAE and below the blend head's MAE, making fine discrimination a demanding test. Weak hard-set agreement would limit ranking resolution even if easy-set agreement were high.

First listener snapshot (25 September 2026). At export, 200 pairs from the four sets were active in the annotation interface. Seven pseudonymous raters had submitted 151 responses between 11 and 25 September, covering 147 of the active pairs (73.5%). There are 128 decisive A/B choices, 20 ties, and 3 skips. Table A.3 gives agreement with the higher native score by head and gap. The export, analysis script, and derived JSON are in `burst-and-genuineness/` and `evidence/metric_validation_2026-09-25.json`.

Table A.3: Preliminary matched-pair validation snapshot (11–25 September 2026). Agreement is the share of decisive A/B choices that favour the clip with the higher native model score. Ties and skips are left out of that denominator and reported in the text. Wilson intervals treat every choice as independent. The bootstrap resamples raters and is wider because five to seven raters supplied all responses. p is an exact two-sided binomial test against 50%. Descriptive, not final.

Head	Native gap	Pairs	Raters	Agree	%	Wilson 95%	Rater boot. 95%	p
Genuineness	$\Delta \geq 2$	31	5	17/27	63.0	44.2–78.5	16.7–79.6	0.25
	$1 \leq \Delta \leq 1.5$	35	5	22/34	64.7	47.9–78.5	42.9–74.5	0.12
	pooled	66	5	39/61	63.9	51.4–74.8	45.5–74.3	0.04
Blend	$\Delta \geq 2$	39	7	12/33	36.4	22.2–53.4	27.8–53.3	0.16
	$1 \leq \Delta \leq 1.5$	42	6	19/34	55.9	39.5–71.1	40.0–69.7	0.61
	pooled	81	7	31/67	46.3	34.9–58.1	37.2–61.5	0.63

Genuineness agrees with the listener on 39 of 61 decisive choices (63.9%, Wilson interval 51.4–74.8%, exact binomial $p = 0.04$). This is the only cell that clears chance, and it is fragile. The rater bootstrap interval (45.5–74.3%) includes 50%, and one rater supplied 28 of the 61 choices with 22 agreements. Without that rater, agreement is 17 of 33 (51.5%). Large and small native gaps give nearly the same agreement (63.0% and 64.7%). In this sample, a two-point gap on the head's 0–6 scale did not make the listener's decision easier than a one-point gap, which does not match the premise of the easy/hard split.

The blend head shows no agreement with listeners so far: 31 of 67 decisive choices (46.3%, $p = 0.63$). On the large-gap set it agrees on 12 of 33 (36.4%). That interval (22.2–53.4%) still includes chance, so we do not claim the head is inverted, but the data give no support for treating a large blend gap as evidence of better integration. Leaving out any single rater keeps pooled blend agreement between 42.1% and 50.0%. Ties are more common for blend (15 of 82 responses) than for genuineness (5 of 69). Counting a tie as half an agreement gives 47.0% for blend and 62.9% for genuineness. The result is consistent with the within-set diagnostic in Appendix L, where blend correlated weakly and negatively with requested-burst hits, although that diagnostic concerns event presence and this study concerns integration.

Subgroup cells are small. For genuineness, German pairs agree more often than English pairs (18/25, 72.0%, versus 21/36, 58.3%), and real and synthetic halves are close (66.7% and 61.3%). For blend the language pattern reverses (German 12/31, 38.7%; English 19/36, 52.8%). No subgroup has more than 36 decisive choices and all intervals overlap, so we treat none of these contrasts as established.

Listeners picked the second-played clip in 73 of 128 decisive choices (57.0%, $p = 0.13$). For genuineness, agreement was 24 of 32 (75.0%) when the higher-scored clip played second and 15 of 29 (51.7%) when it played first. Order was randomised and close to balanced, so this mainly adds variance rather than shifting the pooled figure, but order belongs in the final model as a covariate.

Three properties of the snapshot limit what it can show. Only four pairs received two responses, too few for any inter-rater reliability estimate, so we cannot yet say how far 64% is from the ceiling that listeners themselves would reach. Two raters supplied 85 of 151 responses (56.3%). Confidence was left blank in 112 of 151 records and cannot be used to weight or screen responses. The snapshot is therefore modest evidence that the genuineness head ranks in the listeners' direction and, so far, no evidence that the blend head does.

A.3 Study 2: the Voice Acting Arena

The archived v0.2 task registry, dated August 31, 2026, has 374 English tasks and 214 target cells. The emotion track crosses 40 emotions with two intensities and three scripts, giving 240 tasks. The vocal-direction track has 17 targets at two levels, giving 34 tasks. The acting track has 100 challenges, ten in each of ten domains.

The acting domains are Emergency, Intimacy, Conflict, Comedy, Suspense, Public, Embodied, Fantasy, Everyday, and Transformation. They test contextual delivery and arcs beyond one categorical emotion label. This paper reports only aggregate registry counts; all 374 task contexts, directions, scripts, and requested bursts are withheld from the release to reduce direct benchmark gaming. They may be made available under controlled access for independent auditing after the evaluation is frozen.

The scripts contain three or four sentences and 20–80 words, with vocal bursts specified separately rather than embedded as source inline directions. A versioned correction policy is important because wording changes must apply equally to every model's prompt.

Baseline and configuration freeze. GLM-5.3-Flash translates each acting brief into a model-specific prompt using the corresponding TTS README or interface instructions. The translation policy preserves the fixed spoken script while mapping scene directions and optional vocal reactions to supported controls.

The export lists eleven entries in Tables 5 and A.1, including Fish Audio S2.1 Pro and separate Humaneness Voice pass@1 and pass@10 rows. Model and task versions, voice references, text normalization, expressive controls, post-processing, retries, and generation budgets define the comparison conditions. Later systems, including Gemini 3.8 TTS, belong in a separately versioned wave rather than being inserted into this export.

For APIs with no shared speaker-cloning interface, identical speaker identity may be impossible. A fair comparison should document the voice-selection policy, avoid choosing voices based on test preferences, and separate voice attractiveness from performance where possible. Unsupported burst controls and rejected prompts should be recorded instead of quietly removing the task.

Four preferences, with a conditional fourth question. For each scene, listeners hear two anonymised performances and answer separately about overall acting, fit to the original prompt, and genuineness/authenticity. When the scene asks for a vocal reaction, they additionally judge its type, placement, intensity, and integration. A sample can win one axis and lose another. The prompt must remain visible for fit, with the same content and presentation in both arms. Order should be balanced, and waveform normalisation must be identical and documented.

We fit a separate Bradley–Terry model for each applicable axis. With latent score $\theta_{m,a}$ for model m on axis a,

$$P(i \succ j \mid a) = \sigma(\theta_{i,a} - \theta_{j,a}), \qquad E_{m,a} = 1000 + \frac{400}{\ln 10}\,(\theta_{m,a} - \theta_{\mathrm{ref},a})\,. \tag{2}$$

All reported ratings use this model, with $\theta_{\mathrm{ref},a}$ the mean latent score so that 1000 is the geometric-mean entry. A tie gives half a win to each side. We fit by the minorize–maximize algorithm with a weak prior of 0.1 game spread over each entry's opponents, which keeps a nearly undefeated or winless entry finite. Rating intervals come from 1,000 resamples of votes. Because 456 of the 507 tasks were judged once, resampling votes is close to resampling tasks. Listener clustering is handled separately by resampling listeners, which gives the rank ranges in Table 5 and the gap intervals in Figure A.6. The reaction axis is fitted on the 177 eligible votes only. The comparison graph is complete on every axis; on the reaction axis each of the 55 pairings has 1 to 10 votes. Online game-like Elo updates, which the dashboard does not use, would not supply these intervals.

Every value in Table 5 and the figures of this appendix is recomputed from the accepted vote records by the scripts in `scripts/arena/`. The ordering of the 24 September dashboard survives with minor swaps in the middle, but its reaction column does not: dividing by all comparisons compressed every entry to 6–26%.

A.4 Ablations required for the system claim

A component study cannot by itself establish the value of an agent. A confirmatory arena comparison should include the following frozen configurations:

Configuration	Question isolated
SFT3, one take, fixed reference	What does the instrument provide without orchestration?
Directed script, one take	Does scene interpretation help at equal generation count?
Directed script, eight takes, random pick	Does candidate distribution change independently of ranking?
Same eight takes, original ranker	What does generic perceptual selection add?
Same eight takes, burst-aware ranker	What does explicit local event inspection add?
Remove genuineness / remove blend term	Which perceptual target changes human preferences?
Matched reference and adapter ablations	Which tools improve the performance, not just its proxy?
CFG at equal N and at equal compute	Is improvement due to guidance or additional resources?

Table A.4: Planned system ablations; no completed human results are claimed.

Changing a selection term should rerank the *same* candidate audio to isolate selection. Changing a director or generation tool requires new candidates but should hold scene, words, reference policy, and budget fixed. All final comparisons must include failure rate and compute, not only scores on successful outputs.

A.5 Prior arena development and the leakage boundary

The server documentation records two model-judged director-prompt searches on 20 stratified tasks drawn from this same 374-task pool, using local Gemma and hosted Luna directors. They are not the human arena. Those tasks are development data; a confirmatory test must exclude all exposed tasks, and preferably related templates, or clearly label the full-pool evaluation as development-contaminated.

The first search used best-of-three fitness and bred prompt additions within task/director cells. Naively comparing a child tested on one easy cell with a control pooled over all cells confounded task difficulty with prompt quality. Under paired comparisons, 176 child blocks had mean improvement -0.01 ($t = -0.07$, $p = 0.94$), and only 84 beat their own control. Take-level standard deviation was 1.57 on the 0–15 combined judge scale; selecting the best of three random takes yielded approximately $+2.78$ without a prompt improvement.

The corrected run evaluated every block on every arm, used mean-of-three fitness, and retained a control in every generation: 600 cells, 1,800 clips, and 20 tasks. First-generation imperfection and subtext additions changed from positive point estimates in run one to -0.45 and -0.55 in run two. A stronger pause/breath instruction gave -1.03 ($t = -2.34$, nominal $p = 0.019$), with the same adverse direction in the first run. No beneficial prompt addition survived the corrected comparisons. Multiple comparisons and shared task/model-judge exposure limit confirmatory interpretation.

These findings do not test removal of the existing pause convention, and do not show that pauses are generally undesirable. They show that pushing pause density further is not a supported recipe. The same documents report a restoration setting that improved a model's pleasantness score on ordinary clips but was later disabled after distortion on loud bursts. Performance coverage matters as much as a favourable average.

Evidence: `reports/vand_pairs.md`; Arena registry aggregate counts in `evidence/arena_registry_summary.json`; 25 September 2026 vote export in `evidence/arena_2026-09-25/`; superseded 24 September dashboard values in `evidence/arena_snapshot_2026-09-24.json`; server `docs/ARENA.md` and `ARENA_PROMPTS.md`.

B THE ARENA INTERFACE AND CONSTRUCTED TASK EXAMPLES

Figure B.1 shows the live listener interface. The brief on the left gives the scene, the direction, and the exact script. Phrases in the scene and direction are underlined by role: circumstance, delivery, emotion, or constraint. The listener plays both anonymous takes, can flag a problem with either one, and answers each question with Take A, Tie, or Take B. A comment is optional. The vocal-reaction question appears only when the task requests a burst, so it is absent here.

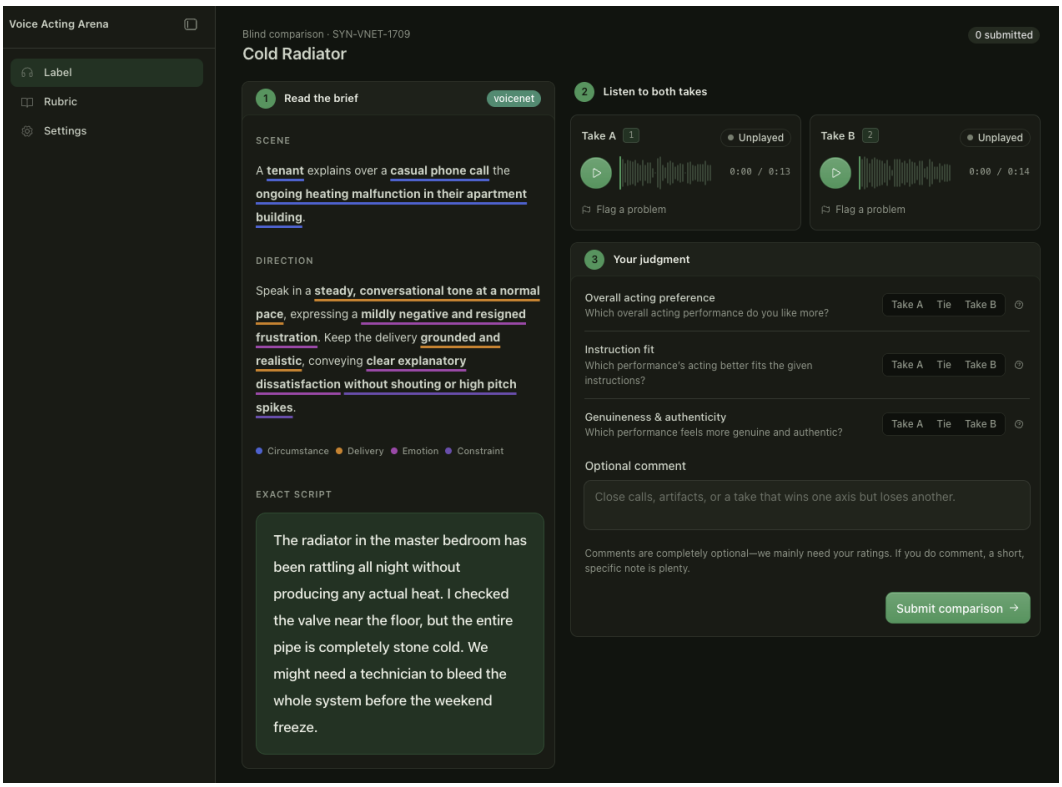

Figure B.1: The Voice Acting Arena labelling interface for a vocal-delivery task from the extended pool. This task received no votes in the 25 September 2026 export and is shown only to illustrate the interface.

The twelve sketches below were written for this paper to explain the task schema. They use new situations and spoken lines; they are *not* entries from the withheld Arena registry and should not be treated as additional test items or as an exact sample of its difficulty. Each separates the scene, the requested performance, and words that a TTS system would have to preserve. A complete administered task also records its target ID, strength, script and optional reaction placement in a versioned machine-readable form.

Table B.1: Four newly composed examples for each Arena task family. Spoken lines are illustrative and are not held-out scripts.

Target	Illustrative scene, direction, and spoken words
Fine-grained emotion	
Awe, moderate	*Scene:* A night librarian sees the first snow through a skylight. *Direction:* Let quiet wonder interrupt a routine closing announcement without turning it into triumph. *Words:* "I was about to turn off the lights. Look at the roof. For a moment, even the street outside seems to have stopped."
Bitterness, intense	*Scene:* A seamstress sees an old pattern credited to a former supervisor. *Direction:* Maintain social courtesy, but let clipped timing and a dry finish expose the grievance. *Words:* "It is lovely to see the pattern in print. I remember when it first lived on my cutting table. That was a long time ago."
Relief, moderate	*Scene:* A technician learns that a seed archive survived a freezer alarm. *Direction:* Begin braced for bad news; allow a small breath and warmer cadence after the confirmation. *Words:* "All the samples are still cold? Then we have time to move them. I thought we might have lost the entire season."
Shame, intense	*Scene:* A volunteer coach admits having misread the match clock. *Direction:* Let the admission become difficult to finish, without replacing the words with sobbing. *Words:* "I told you there was another minute. There was not. You followed my call, and I need to own what happened."
VoiceNet vocal-delivery direction	
High arousal, moderate	*Scene:* A bicycle courier sees a drawbridge beginning to rise. *Direction:* Use quicker breath and pace, but keep each instruction intelligible; this is lively urgency, not a scream. *Words:* "Take the side path now. The bridge is moving, and we can still reach the parcel locker before it closes."
High tension, intense	*Scene:* A theatre usher waits for a stalled lighting cue. *Direction:* Keep the voice noticeably held and tight even while using polite words; do not confuse tension with loudness. *Words:* "Please remain in your seats. The lights will return shortly. I will let you know as soon as the aisle is safe."
High emphasis, moderate	*Scene:* A ferry worker corrects a changed boarding gate. *Direction:* Stress the new gate number clearly, while leaving the rest of the sentence conversational. *Words:* "We are boarding at gate four, not gate two. Keep your tickets ready and follow the blue signs."
High vulnerability, intense	*Scene:* A parent returns a lost keepsake to an adult child. *Direction:* Make the voice audibly exposed and tentative, yet retain the words and a plausible speaking rhythm. *Words:* "I found this in the old drawer. I kept meaning to give it back. I was afraid you would not want it."
Integrated acting challenge	
Responsibility and resolve	*Scene:* A bakery manager tells an apprentice that an oven failed before opening. *Direction:* Start businesslike, reveal a brief fear of letting the team down, then recover into a concrete plan. *Words:* "The oven is out. I should have checked it last night. Give me the orders, and I will call the customers first."
Exertion and apology	*Scene:* A gardener carries heavy soil bags while admitting a greenhouse latch broke. *Direction:* Let the physical load affect breath and phrasing; the apology should remain sincere and comprehensible. *Words:* "I caught the latch with the cart. Let me put this down. I am sorry; I will fix it before the wind picks up."
Changing status	*Scene:* A novice conductor is corrected by a retired mentor, then asked to lead the next cue. *Direction:* Move from self-conscious hesitation to measured authority without changing speaker identity. *Words:* "You are right, I missed the entrance. Let us take it once more. Strings first, then everyone together on my count."
Reaction within speech	*Scene:* A tailor hears an old family joke on a recording. *Direction:* Place one brief, involuntary laugh before the second sentence; let its warmth flow into tenderness rather than sounding pasted on. *Words:* "That is exactly how she used to tell it. I can almost hear the kitchen around her. Thank you for saving this."

C EVIDENCE STATUS, PROVENANCE, AND READING GUIDE

Scope and source precedence. The synthesis is based on the September 2026 technical report, the dedicated data, agent, CFG, LoRA, steering, GRPO, prompt-schema, and layer-forensics reports, and their underlying numerical artifacts. The source inventory covers 96 documents, including historical versions; these are not 96 independent experiments. The technical report provides the overall chronology. For individual values, the relevant stored JSON and dedicated experiment report take precedence over a rounded summary. For deployed behaviour, the current server code takes precedence over an older report. Model-card facts take precedence over speculative descriptions of scorer training.

The author-workspace audit is dated September 10, 2026. It inspected server commit `aa299f1` (full revision in the bibliography) (Anonymous, 2026c). Dataset counts use the September 10 inventories where those supersede older figures. Such an audit does not establish that a changing API endpoint or remote dataset will remain identical; final experiments must pin immutable revisions.

Evidence category	Included examples	Permitted interpretation
Recorded measurement	SFT3 duration, LoRA dose sweep, W1–W4 probes, burst experiments	Results on the named settings and proxy instruments
Implementation fact	Director schema, timed-script compiler, optional candidate ranker	What the audited code supports, not an effectiveness result
Exploratory observation	Micro-pause prompting, adverse steering listening note	Motivation or failure report; no population-level effect size
Matched-pair metric study	Four 500-pair stimulus sets; 151 responses on 147 pairs by 25 September	Design, native-score keys, and a preliminary agreement snapshot; not a final validity estimate
Arena vote export	560 accepted votes by 70 listeners, 25 September	Bradley–Terry ratings of the deployed entries on this task and listener mix; not a matched causal comparison
Future proposal	Low-data accent heads, iterative repertoire training	A testable research direction, not a shipped capability

Table C.1: Evidence categories used in the paper and appendices.

Reproducible manuscript artifacts. Tables and figures in this paper are generated from an evidence archive that will be released with the final version, together with the editable sources and plotting scripts. Its `evidence/` directory contains the full W1 grid, W2 search summaries and seed results, W4 diagnostics, all 79 adapter dose records, the full crossed ranking table, detector recalls, and the arena specification. A CSV exposes all $38 \times 99 = 3{,}762$ W1 probe cells. The table archive retains 1,868 extracted table blocks with document and section provenance. These blocks include duplicates from historical reports and are an archive, not an aggregation of independent samples.

`source_manifest.json` records source names, sizes, and SHA-256 hashes. `artifact_manifest.json` records the copied numerical files and hashes. Tables J.2, J.3, I.3, K.1, and L.1 are generated directly from those exports. The source reports contain additional qualitative discussion and audio links; their identities are retained in the manifest, but restricted raw data are not redistributed.

Statistical interpretation. Reported t statistics are the paired statistics recorded by the corresponding experiment, usually after grouping repeated takes by prompt, task, or target. A take is not automatically an independent replicate. The broad sweeps are exploratory and often select a best setting from many candidates; their nominal t statistics do not correct that selection or establish confirmatory significance. We report effects, sample structure, and adverse outcomes rather than turning every nominal threshold into a discovery. A new confirmatory test should freeze settings first and use new prompts and voices.

Navigation. Appendix D describes execution and WikiSkills; Appendix E the corpus and release boundaries; Appendix G supervised and preference training; Appendix H the measurement stack; Appendix I synthesis controls; Appendix J the four representation studies; Appendix K burst data,

taxonomies, and ablations; Appendix L candidate selection; Appendix A listener studies; and Appendix M the reproducibility boundaries and next training cycle.

Evidence: Primary source: `reports/technical_report_2026-09.md`, especially the contradiction, resource, numeric-table, and provenance sections; dedicated reports are identified in each appendix.

D EXECUTION CONTRACT: MIND, INSTRUMENT, AND EARS

D.1 A BOUNDED AGENT WORKFLOW

The mind turns scene intent into a structured action over a synthesis interface. The instrument executes that action and produces one or more waveforms. The ears measure aspects of the waveforms and, when enabled, select a take. This decomposition makes responsibilities explicit: a language model should choose an interpretation of a scene, while the server should enforce supported labels, numerical limits, and waveform-format contracts.

The implementation does not autonomously rewrite a scene until a human-like critic approves it. Its supported core is one schema-constrained directing pass plus optional best-of-N. Distinguishing this from a multi-round agent is important both for scientific comparison and compute accounting. A future rehearsal policy could use the same interface, but would require its own evaluation.

The dialogue path can generate the spoken response. The benchmark path must preserve the supplied words. Reference selection, delivery annotations, pauses, and bursts may change the performance; they must not rewrite benchmark content or silently remove difficult words. An unparseable schema, unsupported cue, failed synthesis, or empty output is an outcome to log, not an excuse to keep sampling without counting the additional budget.

D.2 DIRECTOR OUTPUT AND SERVER COMPILATION

The audited schema requires `voice`, `voice2`, `style`, `perform`, `speed`, `language`, `delivery`, and `script`. Optional schema fields were previously omitted by the language model too frequently, motivating explicit required entries even when their value denotes no action. There is no separate `reply` field: stripping the script's cues recovers the spoken words and avoids duplicate generation.

The reference mode selects among emotion, VoiceNet dimension, character, edge-case, sports, or no reference. A second reference can express a transition: the initial state comes first and the destination state second. Speed is a categorical reference-selection request, not proof that the synthesiser can independently control articulation rate. The style list permits one delivery adapter. The performance mode names combinations of adapter, CFG, and experimental steering; its strength is a categorical word resolved by the server.

There are two distinct script layers:

1. The *director* writes untimed semantic directions, recognised burst names, and `[pause]` or `[long pause]`. It must not place numbers inside brackets.
2. The *compiler* adds speech durations, pause durations, burst durations, and the total frame budget. A round bracket with a numerical duration denotes a burst, not a speech-delivery instruction.

This separation prevents an apparently harmless cue such as `(quietly, 2 seconds)` from being parsed as a non-verbal sound. Known burst names must occupy their own bracket. A cue that merely says to speak with a chuckle is not equivalent to inserting `(chuckle)` as an event. Every sentence should have its own direction immediately before the words it affects. The prompt uses strength bands ranging from barely perceptible to extreme, matching the training caption convention.

Some prompt rules are only partially enforced. Capitalised words may be spelled out; the sanitiser lowercases runs of at least three capital letters but not every two-letter abbreviation. Identifier-only brackets can be cleaned, while an identifier embedded in prose may survive. A short-line recommendation is not a hard validator. These are concrete engineering limitations, not generic language-model uncertainty.

D.3 TIMING, LOCAL CONTROL, AND MICRO-PAUSES

The corpus renderer emits gaps of at least 0.20 seconds and discards events shorter than 0.05 seconds. Long speech spans are split using alignment gaps; the model sees durations for speech groups, distinct pause slots, and timed bursts. The inference compiler uses the same categories, but determines durations from its own supported recipe. In the audited server, the default burst duration

is 0.28 seconds, clamped to a supported range of 0.14–1.2 seconds, and the default pause is 0.30 seconds. These defaults are not optimum durations for every actor or scene.

A concrete micro-pause example can elicit an internal beat where a general instruction to be natural does not. Exploratory examples progressed from no internal pauses to approximately one, and then one to three with emotion-specific demonstrations. The notes do not contain a controlled sample size or completed listener comparison. Thus the main text reports a useful prompting observation, not a quantified improvement in genuineness.

Local tagging permits a sentence or word group to carry its own emotion and time allocation. It does not prove that every individual word can attain an arbitrary duration. At 12.5 codec frames per second, one frame is 80 ms. Total duration, articulation rate, silence placement, and emotional-transition timing must therefore be assessed separately. The strong total-duration result cannot stand in for all four.

D.4 References, tools, and stack limits

Reference retrieval conditions the voice model on a recording from an available catalogue. The reference changes more than speaker identity: its delivery, pace, and acoustic characteristics can influence the take. Matching the reference across ablations is necessary to attribute an effect to a LoRA or prompt.

The measured production-style stack contains a contrastive DPO adapter at 1.0, a carrier-voice adapter at 1.0, genuineness at 0.25, blend at 0.5, aesthetics at 0.5, quality DPO at 1.5, and emotion at 1.0, for a summed coefficient of 5.75. An optional delivery adapter raises this by up to 1.5. This coefficient sum is a bookkeeping heuristic, not a norm bound on the actual parameter update: adapters can differ greatly in rank, scale, and direction.

Burst additions are separately limited: at most three, each at most 1.25, with their sum constrained to two by proportional rescaling. The rationale is empirical. Two burst adapters at 1.5 on a loaded stack produced severe content failure in all five diagnostic outputs, with median WER around 0.82. Removing the burst tags did not remove the failure, implicating adapter load rather than those tags alone.

The available API exposes experimental modes beyond the standard setting. We distinguish an exposed mode from a recommended default. Steering was rolled back from standard use after an adverse listening observation. Best-of-N is optional and disabled by default in the audited configuration; a paper comparison using it must say so explicitly.

D.5 WikiSkills as an evidence interface

The WikiSkills library records supported delivery dimensions, dose choices, burst recipes, interactions, and failure cases. Its role is to give the director a constrained, inspectable repertoire instead of expecting it to invent numerical settings. Tables can be generated from measured artifacts and linked to the studies that justify them.

This motivation is related to compiling agent experience into persistent skills (Tang et al., 2026a). The present library is curated and measurement-driven. It has not been shown to autonomously discover reliable skills, update itself without review, or improve on a new task distribution. A useful skill entry should include the target, reference conditions, tested stack, recommended strength, adverse metrics, unsupported cases, and evidence revision. A null result is a useful entry because it prevents repeated ineffective tool calls.

Evidence: `paper/VOICE_ACTING_AGENT.md`; `voice_acting_agent_report.html`; `prompt_schema_2026-08-25.html`; server `docs/DIRECTOR.md`, `timed_script.py`, `llm_agent.py`, `config.py`, and `skills.py`; local `wikiskills/`.

E DATA: SCALE, TRAINING VIEWS, AND RELEASE BOUNDARIES

E.1 ANNOTATED SPEECH AS REUSABLE INFRASTRUCTURE

The data contribution is a set of aligned, annotated resources and explicit training views, not a claim that SFT3 was trained on every released clip. The two TTS Annotated v1 branches contain 137,303,487 records: 107,563,551 in the open-access branch and 29,739,936 in the manually gated research branch. Repeated utterances, derivatives, voice-converted versions, and later subsets may share source content; counts are records, not deduplicated unique human performances.

Subset	Records	Shards	Reported hours	Branch
Emolia	81,824,581	4,163	209,371	Open
Podcast	26,606,892	6,496	96,838	Research
EuroSpeech	9,119,466	2,763	37,916	Open
Common Voice	7,062,615	935	10,272	Open
Voice-profile VC	6,508,056	162	17,016	Open
Snippets	2,633,037	1,410	4,875	Research
Kartoffelphon	2,616,656	218	7,305	Open
Eva snippets	500,007	328	2,199	Research
MLS	432,177	125	1,801	Open
Total	137,303,487	16,600	387,593	Mixed

Table E.1: September 10 inventory. Hours are the sum of rounded source figures, not a new exact duration census. Open access is not a substitute for checking the source-specific licence.

The open and research branches occupy approximately 10.50 TB and 8.04 TB, respectively. The split keeps non-commercial restrictions visible instead of blending incompatible sources into an undifferentiated training set. A downstream user must consult the per-source rights and the terms of any base model or conversion system; this manuscript does not grant additional rights.

Record format. A common utterance key joins audio, transcript/alignment metadata, expressive annotations, codec tokens, and a VoiceCLAP embedding. Codec arrays are unsigned 16-bit values of shape $T \times 12$, and the shipped VoiceCLAP vector has 768 float16 coordinates. The archive retains MP3 inputs where appropriate; FLAC/WAV shipping copies were encoded as 128 kbps mono MP3. Recorded round-trip checks found length differences at most 44 samples in their tested conversions. These checks concern encoding consistency, not perceptual identity.

All 16,600 index shards were inspected. Full sample-level archive checks covered 180 shards with 1,844,512 clips, plus 270 targeted UID checks. Hash manifests cover the shipped archives. Neither the index census nor the sampled audio checks justify saying that every decoded waveform was listened to or verified.

E.2 VOICE PROFILES, REPAIR, AND SIDON

Voice profiles provide repeated content and expressive variants over a fixed voice catalogue, supporting speaker-conditioned training and preference construction. The original generation inventory contains 20,079,381 takes; a repair branch adds 8,133,552 recorded takes, giving 28,212,933 in that inventory. A separate repair release contains 8,177,013 takes over 499 voices, 30,058 tar files, and about 1.29 TB. These are different artifact boundaries and should not be silently reconciled by rounding.

The 500-profile generation pipeline used the voice-acting-v2 MOSS checkpoint and profile LoRA adapters with reference-conditioned acting prompts. Its takes were subsequently converted toward profile references with Chatterbox-VC and enhanced with SIDON. The final `vc_sidon` inventory reports 20,078,645 selected outputs, approximately 54,811 hours, 500 voices, 2,000 shards, 4.36 TB, and 460 GPU-hours. It retains one winning conversion per source take. By contrast, `vprof_vc` in TTS Annotated v1 contains 6,508,056 clips and 17,016 hours from a partially ingested, four-candidate view of this conversion run. The two views share original scripts and performances, and must not be added as independent speech or independent source hours. The profile card names voice-acting v2 as

the generation model; the exact relationship to the separately named SFT2 timed-format checkpoint needs a version-pinned training manifest before the labels are equated.

Voice conversion is useful for identity variation but can erase or alter the very burst that a caption requests. Conversion success therefore requires event-sensitive checks as well as speaker similarity and intelligibility. The burst studies below show why a converted clip with the right transcript is not automatically valid expressive supervision. The profile repository's default view combines *generated takes and identity-repair rerenders* (technical origins `original` and `repair`); a repair supersedes a failed take with the same identity key. Its separate *selected voice-converted profile takes* view (`vc_sidon`) contains one selected Chatterbox-VC candidate per source take after SIDON restoration. These are process labels and overlapping views, not three independent recording corpora.

Known repair issue. A raw stereo handling error produced half-speed material in one large voice-profile branch. Token regeneration repaired the affected representation path; some historical annotations or derivative views were computed before the correction. Provenance needs to carry the source branch and repair revision. The issue does not imply that every source in the annotated corpus has the same defect.

E.3 Supervised and preference views

The main timed-format corpus contains 3,147,802 rows: 1,200,530 voice-profile rows and 1,947,272 real-speech rows. The independently published voice-profile SFT view lists 1,200,531, a one-row discrepancy that the draft retains explicitly. Selection for SFT3 uses a much smaller 398,282-row view, described in Appendix G. The large annotation release is infrastructure for further experiments, not the effective sample count of that run.

The voice-profile DPO release contains 3,451,531 pairs: 1,064,594 emotion pairs, 1,186,406 truncation pairs, and 1,200,531 continuation pairs. The real-speech DPO view contains 3,959,192 pairs. Contrasting full versus truncated or continued sequences creates a different preference signal from contrasting two emotional performances; these should remain separate training families in an ablation.

The CFG-DPO and phase-two contrastive families total 842,935 pairs. The first contains two mirrored sets of 237,209 pairs, or 474,418; phase two contains 309,128 emotion contrasts and 59,389 length contrasts, or 368,517. The September 10 publication inventory records 258 verified files and approximately 6.509 GB. The historical name CFG-DPO describes a contrastive pairing construction, *not* audio sampled using CFG.

Other released scene and preference sources. The wider model lineage also uses or derives from the following releases; these should not be conflated with the final 398,282-row SFT3 supervised subset. The synthetic edge-case ranked takes report 1,494,503 utterances and 11,840 hours; the synthetic ranked reinterpretations report 79,624 files and 320.8 hours, split into 243.8 hours of new takes and 77.0 hours of source originals. The annotated scene-pair release has no verified duration total in its current card; it derives from the larger voice-acting scene corpus and may overlap its derivatives. The voice-profile DPO, real-speech DPO, and CFG/phase-two DPO repositories provide preference-pair views with reused audio; pair counts are not unique speech hours. The edge scenes and reinterpretations were produced in the earlier voice-acting lineage, while the preference views supported subsequent DPO and adapter studies. Exact checkpoint-by-repository consumption requires a version-pinned run manifest.

Distinct synthetic-selection rules. The edge-case release samples 20 takes per acting prompt and retains the top three by a within-group min–max-normalised sum of blend, genuineness, and arousal. That selector has neither a 1.25 target-emotion coefficient nor a multiplicative ASR gate; the card warns that expressive selection can leave words unintelligible. The reinterpretation release instead re-performs each source line 64 times with the voice-acting-v2 MOSS checkpoint and v3 emotion LoRAs. Its documented score is

$$R_{\text{reint}} = \frac{s + 1.25t + b + g}{4.25}\big(1 - \min(\text{WER}, 1)\big), \tag{3}$$

where s measures similarity to the source clip's 42-dimensional emotion/arousal/valence profile, t is the within-group normalised target-emotion score, and b, g are transformed blend and genuineness

scores. The top three of 64 and the original are released together. Both rules select by model predictions rather than human judgments; their different rewards must not be described as one shared ranking formula.

The trajectory corpus separately exposes 7,638,961 *specifications*; it is not a release of that many generated audio files. A larger private inventory contains 10,653,713 specifications. Another trajectory experiment uses its own smaller generated-audio collection. The technical reports contain differing hour totals for that experiment, so those hours are not used as a central contribution claim.

E.4 Annotation layers and non-destructive updates

Annotations include expressive captions, emotion and vocal-dimension estimates, genuineness, blend, and local non-verbal-event information. Most are model outputs. Rich coverage is useful for training and filtering, but does not convert weak labels into human ground truth.

The September 10 burst update reclassified 63,838,964 *existing spans* across the 16,600 shards. It did not discover 63.8 million new confirmed bursts. The run used approximately 1,013 GPU-hours over eight hours on 32 nodes. New `vb2_*` columns preserve the old labels, expose class probabilities and alternative classes, and carry model/version identifiers. Classifying an existing interval and locating a new interval are distinct operations.

Caption revisions corrected inverted prose descriptions in historical general-caption fields. Numeric scales must be interpreted against their exact taxonomy version. In particular, the VoiceNet bucket audit identifies a five-level versus seven-level BKGN inconsistency and distinguishes that from a prose-polarity error. We do not apply a universal BKGN sign flip to every artifact. The original caption version is retained where supplied, with the corrected view separately named.

E.5 Independent bucket rechecks reveal weak-label limitations

A later model re-evaluated 18,632 voice-annotation clips against their own bucket descriptions. Among 9,390 previously positive examples, only 3,012 (32.1%) received the strongest fit rating, while 4,700 (50.1%) received zero. Of 9,241 nominal negatives, 437 (4.7%) received the strongest positive fit. Among 39,387 successfully rechecked new candidates, 7,912 (20.1%) received the strongest rating. These are model–model discrepancies, not adjudicated human error rates.

Coverage alone was misleading. The existing annotation had at least 25 candidates in 334 of 395 addressable buckets, but a sample of 606 such candidates yielded only 173 (28.5%) strongest-fit ratings. Temporal trajectory attributes were particularly weak: only 2.9% of positive Velocity Flux examples and 10.3% of Dynamic Arc examples received the strongest fit.

The 399 bucket definitions are byte-identical to the corresponding training anchors, explaining why a bucket index and a head target can share an intended semantic scale. This does not establish that the audio actually instantiates the anchor. Repeated deterministic labelling is also not independent reliability: all 57 valid repeated labels in one check were identical, but the prompting configuration was effectively deterministic. This motivates speaker-disjoint human audits before using such heads as a definitive performance judge.

E.6 Restricted sources and publication status

The German Mediathek inventory comprises 4,943,358 segments in 934 shards, approximately 9,270 hours, and 1,547,418 newly localised spans. Sharp inhales account for about 81.1% of these spans. These data are not part of an unrestricted public-audio claim. September 10 updates were uploaded to private repositories while retaining the publication restrictions.

The podcast metadata update joined all 26,606,892 keys. Mediathek caption coverage was 4,943,333 of 4,943,333 existing caption rows, with 25 additional segments lacking an older caption row. The new localiser used different span settings from the old burst branch, so separate `vb2_starts` and `vb2_ends` fields prevent accidental alignment to the old intervals. All old columns and row order were checked for value preservation, including NaN-aware equality.

These release details matter scientifically: a paper can offer useful reproducibility even when some sources are restricted, provided it identifies which experiments depend on those sources and exposes

the code, schema, settings, aggregates, and lawful alternatives. It cannot claim an entirely unrestricted reconstruction of every training run.

E.7 Repository links, precise licences, and duration accounting

Table 3 identifies the four primary audio views; Table E.2 lists related SFT and preference-pair views separately because pair counts are not unique audio hours. The `vc_sidon` view resides in the same voice-profile repository as the original and repair views. All listed releases are from 2026; this is not the collection year of every upstream recording. The September 11 licence audit records repository revisions, card URLs, declared licences, counts, and rounded hour totals in `evidence/dataset_inventory.json`. Both tables are generated from that inventory.

Table E.2: Related training views. Preference-pair counts reuse audio and cannot be converted into unique speech hours. The voice-profile timed-script view has no reported duration; real-speech timed scripts report 6,996 hours.

Training view	Records	Hours	Mean (s)	Declared licence
Voice-profile timed scripts	1,200,531	–	–	CC-BY-4.0
Real-speech timed scripts (DE/EN)	1,947,272	6,996	12.93	CC-BY-4.0
Voice-profile preference pairs	3,451,531	–	–	CC-BY-4.0
Real-speech preference pairs (DE/EN)	3,959,192	–	–	CC-BY-4.0
Voice-profile CFG preference pairs	842,935	–	–	CC-BY-4.0

Annotations and audio have different rights. The open TTS release declares `mixed-per-subset`, not a uniform audio licence. Its captions, scores, and derived measurements are CC-BY-4.0; the waveforms retain upstream terms. The research release declares custom `non-commercial-research-use` terms, not CC-BY-NC-4.0. Access requires manual approval for research collaboration, and upstream audio restrictions still apply; annotations are CC-BY-4.0. The following table transcribes the scope of the dataset-card declarations, not an independent legal clearance.

Table E.3: Source-specific audio rights underlying the two TTS Annotated v1 branches. Repository links identify upstream sources; a missing explicit licence is not replaced by the annotation licence.

Subset	Upstream repository / source	Audio licence or documented limitation
Emolia	Emolia, our re-annotated release (anonymized link pending)	CC-BY-4.0.
EuroSpeech	`disco-eth/EuroSpeech`, via our enhanced DAC-VAE release (anonymized link pending)	Custom / `other`: terms vary by parliament. No uniform CC licence is declared; consult upstream source terms.
Common Voice	Common Voice 22 after SIDON restoration and DAC-VAE encoding (anonymized link pending)	CC0 / CC-BY-4.0 per upstream release, as reported in the assembled corpus card; a single licence is not asserted for all derivatives.
Voice-profile VC	The 500-profile MOSS v2 + profile-LoRA generation pipeline, followed by Chatterbox-VC and SIDON (`vprof_vc`); see the voice-profile repository	Synthetic output of our pipeline. The parent repository declares CC-BY-4.0; the assembled TTS release does not separately resolve every derivative's audio terms. The 17,016-hour four-candidate partial view overlaps the profile repository's selected `vc_sidon` view in source performances.
Kartoffelphon	`MultiLlasa/Kartoffelphon-2.5M-de-ger`	CC-BY-4.0; the card additionally identifies approximately 51% LibriVox public-domain source material.
MLS	Enhanced DAC-VAE release (anonymized link pending), from `facebook/multilingual_librispeech`	CC-BY-4.0.

Subset	Upstream repository / source	Audio licence or documented limitation
Podcast	Podcast Links episode index; processed, tokenized podcast audio (anonymized links pending)	Public episode URLs do not confer redistribution rights. No uniform audio licence is declared for the processed source; the release is gated for noncommercial research.
Snippets	mitermix AudioSnippets; filtered/enhanced intermediate (anonymized link pending)	The mitermix card labels the dataset Apache-2.0 but has no source-provenance description; the intermediate source has no declared audio licence. We retain gated noncommercial research access pending source-rights review.
Eva snippets	Emotion/voice-attribute reference snippets (anonymized link pending); mixed podcast and AudioSnippets derivations	The immediate source declares CC-BY-4.0, but sampled UIDs are 98.56% podcast-derived. The source declaration does not clear those recordings; this view remains gated under the same noncommercial research terms.

The voice-profile annotated/SFT/DPO releases and the real-speech SFT/DPO and CFG-pair releases each declare CC-BY-4.0. These are repository declarations, not a claim that all source references or base-model outputs automatically carry identical rights. In particular, the real-speech SFT card explicitly excludes EuroSpeech.

Hours and mean snippet duration. The open branch totals 283,681 reported hours. Its source groups are Emolia/Common Voice/Kartoffelphon/MLS (228,749 hours with CC-BY-4.0 or CC0 source declarations), EuroSpeech (37,916 hours with parliament-specific terms), and the synthetic voice-profile VC candidate subset (17,016 hours whose derivative provenance must be checked). The research branch totals 103,912 published view-hours under custom noncommercial research-access terms: Podcast 96,838, Snippets 4,875, and Eva snippets 2,199. Podcast and Snippets lack explicit upstream licence declarations in the release inventory; Eva's upstream CC-BY-4.0 declaration does not erase the research branch's access restrictions or resolve mixed provenance. These 103,912 hours are therefore a branch-level access count, not an assertion that all other audio is cleared for commercial use. Original voice-profile takes contribute 54,943 hours and identity-repair rerenders 16,113 hours (71,056 combined); the overlapping selected VC-and-restoration view reports 54,811 hours. The open branch's 17,016-hour `vprof_vc` candidate subset also overlaps profile-source performances. Even the three primary repository totals (458,649 hours by arithmetic addition) are not a unique-audio total. Real-speech SFT reports 6,996 hours. For comparable utterance/take views, Table 3 computes approximate mean seconds as $3600\,H/n$, where H is rounded published hours and n the record count. No verified audio-hour total is published for `voice-acting-data-annotated`. Preference-pair counts reuse audio and cannot be multiplied into unique hours.

The processing lineage starts from publicly reachable podcast episodes indexed in Podcast Links, then segments and filters the audio. Our annotation pass includes Parakeet-TDT-0.6B-v3 ASR, Empathic Insight Voice Plus emotion/attribute scores, and the additional quality and vocal-burst scores listed above. The Snippets branch applies separate filtering and audio enhancement to the AudioSnippets-derived source. Eva snippets use a different reference-snippet selection: in an audit of 58,733 sampled UIDs, 57,887 (98.56%) carried podcast-source identifiers and 846 (1.44%) carried Snippets identifiers. Thus the two snippet-labelled views do not share one homogeneous source. The 103,912-hour custom-noncommercial subtotal sums published views; it is not a deduplicated estimate of restricted audio. The AudioSnippets source may include material from online videos, but its card does not document per-item origins or rights, so we do not treat its Apache-2.0 label as clearance of every recording.

E.8 The later German–English scaling ladder

The September 12 base release, extended on September 19, is a post-snapshot selected view and was not used to train SFT3. Its four candidate cells after the release's UID deduplication contain 17,355.3 hours of German real speech, 144,675.1 hours of English real speech, 72,480.3 hours of German synthetic speech, and 65,389.1 hours of English synthetic speech. These sum to approximately 299,900 candidate record-hours, not 299,900 independent human performances. The older approximately 420,000-hour pool estimate is not a deduplicated release inventory. The three primary

repository cards sum to 458,649 hours as published views, but neither that sum nor the candidate pool can establish a unique-hour total across voice conversions, repairs, research subsets, and training pairs.

Within each language-by-origin cell, selection allocates half its tier quota to 40 emotion categories and half to 57 vocal dimensions crossed with ten equal-width value bins. Scores are min–max normalised within each cell. The common quality score is $R_q = \mathrm{Norm}(genuineness) + \mathrm{Norm}(blend)$. Emotion candidates first compete on $3\,\mathrm{Norm}(target\ emotion) + R_q$ to define nested intensity bands, then on R_q inside each band. VoiceNet values place candidates into dimension–decile buckets; within each bucket ranking is by R_q, not by the VoiceNet score again. Equal quotas encourage coverage, but source scarcity and deduplication can cause shortfalls. Thus "balanced" denotes the documented selection target, not a guarantee of equal available source hours or equal performance by group. The original tiers are nested: approximately 100, 250, 500, 1,000, 2,500, 5,000, 10,000, and 20,001 hours. The original largest tier has 6,666,998 distinct UIDs and 20,001.48 hours; the held-out set has 25,639 UIDs and 80.12 hours with zero overlap. The 19 September release adds non-overlapping `x50` (10,468,903 clips, 29,567.7 hours) and `x100` (16,976,532 clips, 49,350.9 hours) containers. *Expanded 49,569 h* and *Broad 98,920 h* name their cumulative views; the old loader IDs remain compatibility identifiers. Broad has 51 planned shards unshipped, about 1,082 hours below target. These figures describe released views, not SFT3 training exposure.

Purpose and possible curriculum. The ladder publishes matched scale points and selection metadata so that future work can train an expressive TTS model from scratch, test data-scale effects, and compare a broad-to-selective schedule under the same held-out split. A natural schedule would take one epoch over Broad, then one over each successive smaller view in Table 4, ending at Tier 0. This intentionally revisits high-ranked clips because each smaller set is contained in every larger one. The schedule is a proposal, not a result: from-scratch runs were not complete at submission, and the released Humaneness Voice checkpoint is a MOSS fine-tune rather than proof that this schedule works. Any commercial-only replication must first filter records by their per-source audio terms and recompute the resulting hour counts and bucket balance.

Annotation formats and coverage. The stored WebDataset examples contain audio, MOSS-v2 codec targets, and a record with verified text, word timing, source and licence metadata. Scores and directions can be used separately; their labels are model-generated, not human ground truth. The sidecar release adds richer text on 3,682,649 unique UIDs (the Tier 6 set plus a synthetic VC sidecar), but Broad, Expanded, and Core include additional clips without those sidecar fields. The same UID joins sidecars to the index, so missing annotations must remain missing rather than being imputed from another version.

The ladder card carries source-specific audio terms. EuroSpeech is marked `other` for parliament-specific rights; voice acting synthetic sources also require checking their generating models' terms. A user who requires only CC-BY-4.0 or CC0 must filter by the per-record licence field, changing the exact tier balance. This is why the full ladder is described as an open, documented dataset with source terms, not as a uniformly Apache-licensed corpus.

Evidence: `datensaetze_status.html`; `reports/audio_open.md`, `audio_research.md`, `vp_inv.md`, `pvc.md`, `burst_reann.md`, `dpo_cfg.md`, `vand_buckets.md`, `mediathek_pub.md`, `ttsagi_update.md`, and `scaling_ladder.md`; linked dataset cards (licences checked September 11, revisions in `evidence/dataset_inventory.json`). Source-family rights are retained in the dataset cards.

F Acting Challenge Dataset: provenance and use

The released repository packages 19,247 performance briefs as reusable seeds for speech-data generation. A brief poses an acting problem but leaves the exact spoken dialogue open. A researcher can sample a family, realize a new script with a director, synthesize several takes, and keep takes that satisfy the brief. This complements, rather than duplicates, the held-out Arena: its scenes and scripts are not present in the public dataset. Each row includes a title, instruction, and source family; derived rows can also identify the parent idea, situation, sampled emotion, or speaker condition.

Table E.4: Annotation formats available with the scaling ladder. Coverage differs by stage; the two BUD-E versions are distinct outputs. Exact schemas and model revisions appear in the dataset card.

Format / fields	Information available to a training or selection pipeline	Coverage / caution
Verified text, words, bursts	Transcript; word timings; event labels, onsets and offsets; language, origin, duration and per-record source licence.	Core index fields; event re-annotation is source-dependent.
EmoNet, VoiceNet, quality	Forty emotion scores; 57 voice-attribute scores and value buckets; genuineness and vocal-burst blend scores.	Model-produced estimates, not human labels; scales depend on model and taxonomy versions.
Procedural prompt	`GENERAL` clip description plus `SCRIPT` with per-sentence delivery cues, measured pauses and inline vocal reactions.	Derived from global scores; sentence cues are not independent sentence-level ratings.
BUD-E Whisper V1.1	Free-form expressive caption covering affect, delivery, recording context and speaker traits.	Sidecar: 3,682,649 Tier-6/VC UIDs, not every larger-stage clip.
Timbre-Whisper	Raw caption and separately parsed structured timbre tags and natural-language timbre prose.	Same sidecar UID set; missing parts are left empty.
Voice-Tagging-Whisper	Quality, naturalness, fluency, style, phonation, airflow, loudness, prosody and articulation tags; parsed JSON and optional prose.	Same sidecar UID set; prose is normally empty by design.
Gemma-4-E4B directions	One of six natural-language acting-instruction formats; quoted spoken text, optional duration/event directions and validation audit.	Same sidecar UID set; machine-written instructions are not human preferences.
BUD-E Whisper V1.0	Historical free-form caption alternative for Emolia-origin clips.	Separate 10,629,084-UID historical join; not interchangeable with V1.1.

The repository includes source JSON, normalized JSONL, taxonomies, conversion code, and a version-pinned manifest with a SHA-256 check.

Table F.1: The seven construction routes behind the public performance briefs. Names describe each family's role rather than internal batch labels.

Brief family	Briefs	How it broadens the prompt space
Audition-scene seeds	278	Original performance prompts from an earlier expressive-speech dataset.
Thematic scene seeds	1,000	Synthetic prompts spanning varied dramatic themes.
Vocal-attribute seeds	200	Scenarios organized around VoiceNet delivery dimensions.
Acting-method briefs	4,030	Five original scenes per item in an 806-item actor-training index.
Contextual variants	7,390	Five changed settings or stakes for each of the 1,478 seeds above.
Situated performances	5,749	Up to 20 briefs per physical or social situation.
Extreme physical states	600	Six categories $\times$ ten subcategories $\times$ ten scenarios.
Total	19,247	Related briefs, not independent human-authored concepts.

From structured ideas to performance briefs. The 278 audition-scene seeds come from an earlier expressive-acting dataset (anonymized link pending) used in the EmoNet-Voice research line (Schuhmann et al., 2025a). Another 1,000 synthetic seeds cover varied dramatic themes; their historical label "Haiku" names a production batch rather than an acting category. Two hundred further

seeds center on VoiceNet vocal attributes. Together these 1,478 briefs establish a range of starting points. Five contextual variations per seed change the setting, stakes, or emotional conflict, yielding 7,390 further briefs without treating a paraphrase as a new acting concept.

The acting-method route draws on a structured 806-item snapshot of concepts, techniques, exercises, and recommendations extracted from Eric Morris's actor-training books. Morris's framework asks performers to work from circumstances, sensory attention, vulnerability, and choices, rather than merely attach an emotion label to a line. Five original scene directions were generated from each indexed item, giving 4,030 briefs. The dataset shares the new briefs and their source identifiers, not pages or passages from the books.

The 289-item Situation Taxonomy changes the circumstances under which a line is spoken. Fear while running, fear while lying still, and fear whispered to someone nearby make different demands on breath, projection, and timing. Its 11 dimensions include posture, activity, speaking target, social context, environment, health, obstructions, climate, substance influence, fatigue, and pain. Up to 20 briefs per situation combine those circumstances with sampled EmoNet emotions, intensities, age, and speaker gender. DeepSeek V4 Flash is documented for this challenge-generation route. The resulting 5,749 briefs are 31 short of the full 289×20 grid. The 600 extreme-physical briefs systematically cross six classes of exertion, breathlessness, pain, temperature, oral/taste sensation, and startle with subcategories and scenarios. They deliberately test large audible contrasts, often across a scene change. Gemma 4 E4B-it appears in the later step that turns a brief into a complete script; it should not be mistaken for the generator of every brief.

What is balanced and what is not. The full pool is skewed toward 7,390 earlier-task variants and 5,749 situation variants. It is therefore neither uniform over seven source families nor jointly balanced over all 40 emotions, 57 vocal dimensions, 289 situations and speaker conditions. For a simple, reproducible family balance, each row's sampling weight is $1/(7n_f)$, where n_f is its family size. A separate 1,400-row reference view takes 200 examples from each family, with deterministic round-robin coverage of recorded seed concepts, situation dimensions or extreme categories where available. This is a convenience view, not an independent test set or proof of equal semantic coverage. Researchers can construct more specialized, source-aware strata for their own generation objective.

Separation from the Arena. The public tasks can seed model-specific scripts, multiple generated takes and internal or reinforcement-learning-style selection, but none of those downstream gains is measured here. The held-out Arena situation details, performance directions and spoken scripts are not released with this dataset, to reduce direct test-set optimization. An internal exact-string audit found eleven generic title overlaps between the already-public pool and the 374-task Arena draft, but no exact match between a public challenge instruction and an Arena scene context, performance direction or script. Related concepts can still leak semantically, and existing director-prompt development has already touched part of the Arena pool (Appendix A); future confirmatory evaluation needs fresh, separately frozen tasks. The public challenge set is licensed as repository text under Apache-2.0 by its source project; that declaration does not grant rights over Eric Morris's books. Physical and health descriptions are creative prompt hypotheses, not validated physiological ground truth.

G INSTRUMENT TRAINING, PREFERENCE LEARNING, AND ABANDONED GRPO

G.1 MODEL LINEAGE AND SUPERVISED ROUNDS

The instrument derives from MOSS-TTS-Local-Transformer-v1.5, approximately 4.55 billion parameters, with approximately 4.13 billion trainable parameters in the recorded adaptation configuration. Its semantic transformer has 36 layers of width 2,560, and its frame-local transformer is approximately 550 million parameters. The 48 kHz v2 codec uses 12 residual-quantiser codebooks of size 1,024 at 12.5 frames per second. These settings describe the inspected v1.5 instrument, not every model covered by the original MOSS technical report.

The early voice-acting lineage includes a full fine-tune and a rank-256 adapter merge. The v2 adaptation used 185,503 rows from six datasets with equal source sampling, six epochs, learning rate 10^{-5}, and batch size 64 on eight A100 80 GB GPUs for approximately 14–15 hours. Some early scripts were ephemeral and are no longer available. The released checkpoint is therefore more reproducible as an inference artifact than that historical training run is from scratch.

The timed-format SFT1 run used 3,147,802 rows, three epochs, 2,232 steps, and global batch 4,096. Its recorded allocation was 64 nodes with four GH200 GPUs each, and wall time approximately 1 hour 53 minutes. Subsequent rounds tightened the format and data selection. SFT3 starts from SFT2 rather than from the original base checkpoint.

SFT3 setting	Recorded value
Selected training view	398,282 rows: 331,902 core + 66,380 blend
Materialised view	401,345 rows, including 3,063 validation rows
Epochs / updates	2 / 712 (356 per epoch)
Global batch	1,024; microbatch 4, accumulation 2, 128 ranks
Learning rate / warmup	7×10^{-6} / 178 updates
Allocation	32 nodes, four GH200 GPUs per node
Wall / recorded training time	30 min 46 s / 25 min 45 s
Prompt-format hash	`f90a7b16b4cea093`

Table G.1: SFT3 recipe as recorded. Dataset size, global batch, and updates per epoch do not reconcile by simple division; sampler/filter accounting remains to be recovered. We do not invent an explanation.

Selection emphasised useful expressive extremes and blend supervision rather than every weakly annotated row. A broad extreme-selection rule covered 12.65% of the source view; taking the union of per-dimension deciles would cover 99.84%, illustrating why a high-dimensional union is not selective. The 6,000-per-emotion quota corresponds to a much narrower tail for individual targets. Exact selection rules and per-family counts belong with the training manifest.

G.2 TIMED-FORMAT CONSTRUCTION AND AUDIT

The training renderer draws timed prompts with probability 0.70 and reference conditioning with probability 0.50. Nominal direction use is 0.85; the realised fraction is approximately 79.5–80.2% after the renderer's eligibility conditions. Segment directions are dropped with probability 0.15 and burst cues with probability 0.10. These augmentations do not mean that a randomly drawn training row always contains a usable burst.

The renderer checks that stripping tags recovers the intended transcript. A corpus audit recovered the transcript in all checked cases. Median duration-rounding differences were approximately 0.01 seconds for real speech and 0.02 seconds for voice profiles. Reconstructing text only from alignment tokens would lose numerals in 9.9% of a real-speech sample, so the original transcript is retained.

The event annotations are imperfect. About 39.4% of textual burst cues in an audit were not confirmed by the detector, and about 49% of annotated bursts overlapped words. Overlap can reflect an event embedded in speech, an imperfect aligner, or a weak label. It is not safe to turn every interval into a clean isolated-burst target without inspection.

G.3 SFT2 VERSUS SFT3

The shared evaluation contains 80 prompts, one English and one German prompt for each of 40 emotions, with four completions per prompt. This gives 320 clips per checkpoint. The same evaluation design supports a within-task comparison; it is not a representative multilingual benchmark.

Measure	SFT2	SFT3
WER, pooled	0.4473	0.0987
WER, English	0.574	0.070
WER, German	0.320	0.127
Median absolute duration error (s)	0.100	0.080
Within 0.5 seconds of request	92.8%	100.0%
Burst hit rate	0.516	0.666
Burst reward term	0.224	0.356

Table G.2: Timed-format evaluation. Burst metrics are model-measured. Total-clip duration does not establish exact word or transition timing.

A separate 1,596-clip transition study found median total-duration error zero in 13 of 14 arms, fitted duration slope 0.979, and signed mean error -0.045 seconds. The spliced arm was the exception, with median total-duration error 0.690 seconds. These are total-duration results, not measurements of individual emotional-transition locations. A speaking-rate study found that requests for slower delivery often add padding rather than independently slowing articulation. Its approximate floor of 18 characters per second is a diagnostic for the tested setup, not a universal language-independent speed limit. A 121,940-row tempo adaptation lasting approximately 1 hour 47 minutes did not establish an improvement.

G.4 DPO FAMILIES AND WHAT THEIR NAMES MEAN

Direct preference optimisation (Rafailov et al., 2023) was applied to several pair constructions. The strongest recorded checkpoints came from contrastive families, not from a general proof that more preference training is always better. CFG-DPO uses high/low score contrasts and paired prompt constructions, including text removal in a contrast condition; the name does not imply that the training audio was generated under CFG.

The first published family has $2 \times 237{,}209 = 474{,}418$ pairs. Phase two adds 368,517 pairs. A mixed training run contains 2,327,904 pairs, approximately 20.5% from the CFG-DPO family, with rank 64, $\beta = 30$, and learning rate 10^{-6}. Preference accuracy rose from approximately 0.562 to 0.975, while a later run reached its time limit around step 5,006. Such training accuracy is not a human preference result.

The fixed reference is bare SFT3 with adapters disabled, not necessarily the run's initial adapter-loaded policy. Language was not matched for every contrast: approximately 54.6% of the first family's pairs and 59.1% of phase-two emotion pairs use the same language. About 1.1% of mirrored phase-two pairs reverse the intended target ordering. Shared source clips also mean that pair-row disjointness does not ensure utterance-disjoint train/test data.

Recognition-derived labels can encourage easily observed correlates rather than the intended causal production mechanism. However, symmetry of a pair construction alone does *not* imply a mathematically zero conditional DPO gradient. We retain the measured limitations and avoid the stronger unsupported theoretical claim made in some exploratory notes.

G.5 THE COMPLETE 18-CHECKPOINT COMPARISON

G.6 WHY GRPO WAS NOT RETAINED

Five GRPO variants were attempted, using the group-relative policy-optimisation approach associated with Shao et al. (2024). Earlier runs were affected by implementation and reward issues; v4 and v5 provide the interpretable held-out comparisons. The working setup used 16 completions per group, two prompt groups, microbatch two, rank-32 adapters with scale 64, learning rate 2×10^{-6}, 20

Rank	Checkpoint / adapter	Reward	WER	Emotion	Burst hit
1	Contrastive DPO phase 2, step 5022	.4757	.0977	.3541	.7625
2	Contrastive DPO phase 2, step 3906	.4744	.0916	.3478	.7625
3	CFG-DPO, step 4912	.4708	.0950	.3373	.7719
4	CFG-DPO, step 3963	.4698	.0953	.3420	.7000
5	DPO corpus v2, step 3216 (weights unavailable)	.4687	.1094	.3401	.7094
6	DPO corpus v1, step 864	.4668	.1117	.3518	.6937
7	Emotion LoRA rank 16	.4663	.1239	.3572	.6844
8	Emotion LoRA rank 64	.4640	.1298	.3345	.6844
9	DPO corpus v2, step 4590	.4633	.1024	.3427	.7188
10	Emotion LoRA rank 16, weight .75	.4622	.1055	.3390	.6969
11	SFT3, no adapter	.4584	.0987	.3494	.6656
12	Emotion LoRA rank 16, weight 1.25	.4584	.1130	.3315	.6844
13	Emotion LoRA rank 16, weight .25	.4579	.1174	.3533	.6656
14	Emotion LoRA rank 16, weight .50	.4554	.1345	.3365	.6656
15	GRPO v4	.4551	.1286	.3468	.6937
16	GRPO v5	.4512	.1253	.3275	.6469
17	Emotion LoRA rank 16, weight 1.50	.4509	.1325	.3358	.6312
18	Emotion LoRA rank 32	.4494	.1315	.3209	.6719

Table G.3: Recorded August 29 comparison, 320 clips per row, corrected percentile reward scale. The ranking is descriptive; differences around 0.01 in reward were not resolved by this test. The machine-readable report archive retains additional quality and burst-term columns.

warmup steps, and KL coefficient 0.02. A distributed-padding fix was needed to keep group-shaped data aligned across ranks.

In v4 the rolling reward rose from approximately 0.503 to 0.527, burst term from 0.561 to 0.669, and WER fell from 0.128 to 0.098, while the emotion term fell from 0.259 to 0.242. On the shared held-out test, however, v4 obtained reward 0.4551 and WER 0.1286, versus SFT3's 0.4584 and 0.0987. V5 obtained reward 0.4512 and WER 0.1253. Neither established a benefit sufficient to retain it in the instrument.

This is a negative result for the tested recipe. Group-relative advantages need discriminative within-group information, and a floor or saturation in an emotion component can remove useful distinctions. Yet other reward terms and the WER multiplier may still vary; this observation is not a proof that GRPO must fail structurally. Changing reward normalisation also makes raw reward curves from different runs incomparable.

Saving high-scoring individual candidates remains useful independently of a failed policy update. An archived selected subset of 204 clips had median emotion score 0.641 and WER 0.049. It is a selected subset, not an unbiased estimate of the policy. Its proposed use is a separately audited candidate reservoir for future supervised training, with new held-out tests and safeguards against reward-model overoptimisation.

Evidence: Technical report training and August 29 ranking sections; `cfg_dpo_findings.html`; `grpo_learnings.html`; `reports/dpo_cfg.md`. The scorer and reward definitions changed during development; only same-definition comparisons are used.

H PERCEPTUAL TARGETS AND MEASUREMENT INSTRUMENTS

Shared encoder, specified precisely. The two released heads bundle the frozen audio tower of VoiceCLAP-Commercial (Anonymous, 2026d), not a voice-cloning model. This approximately 110M-parameter dual-tower contrastive model uses BUD-E-Whisper-Small as its 12-layer, width-768 audio encoder and `all-MiniLM-L6-v2` as its text encoder; both map to a 768-dimensional joint space (Anonymous, 2026d). Only the audio tower is used for genuineness and blend. Its commercially usable contrastive pretraining mix is separate from the Gemini-labelled TTS clips used to fit the small score heads.

H.1 GENUINENESS IS SPONTANEOUS BELIEVABILITY, NOT TRUTHFULNESS

The genuineness target asks whether a clip sounds like a lived-in spoken moment rather than a rehearsed or synthetic read. Its 0–6 rubric emphasises timing, breath, emotional grounding, and plausible micro-imperfections. It is not a factuality score, a detector of deception, a diagnosis of the speaker, or a measure of audio fidelity. Deliberately theatrical delivery may be appropriate to a scene even when its genuineness score is lower.

The 768-dimensional embedding is L2-normalised and standardised using stored training means and standard deviations. A small head, $768 \to 50 \to 1$, uses GELU and dropout 0.2, with approximately 38,000 parameters. Training uses Huber loss with $\delta = 1.5$ and Adam. About 10,000 clips curated from several TTS systems and Emolia were labelled on the 0–6 scale by Gemini 3.1 Pro, as documented in the public genuineness card. The full and balanced head variants share the frozen backbone.

The model card reports a stratified validation set of 140 clips, 20 per integer score and seed 1234. The full head has MAE 1.00, RMSE 1.33, and Pearson $r = 0.77$ against its validation labels (Anonymous, 2026b). These are not results against independent human ratings. The small balanced validation design also differs from the population distribution seen by a deployed agent.

H.2 VOCAL-BURST BLEND IS CONTEXTUAL INTEGRATION

The blend score asks whether a laugh, sob, gasp, sigh, groan, or related event belongs acoustically and emotionally in the surrounding speech. It ranges from zero to ten. Zero includes disconnected, robotic, or mismatched insertions *and clips without a genuine burst*. A middle score represents a plausible but noticeably performed event; a high score represents organic integration. Thus the scalar partly conflates event presence with integration quality. Presence must be measured separately whenever absence is a possible failure.

The commercial blend head has the same frozen-embedding and $768 \to 50 \to 1$ architecture. The project team reports that its training clips were curated from multiple TTS datasets and annotated with Gemini 3.1 Pro. Unlike the genuineness card, however, the public blend card does not state the training-set size, label source, or split construction (Anonymous, 2026a); a versioned manifest is still needed to verify the report and identify the exact corpora. The card reports the comparisons in Table H.1.

Frozen embedder	Width	Validation MAE	Validation Pearson r
VoiceCLAP-small	768	2.360	0.418
VoiceCLAP-commercial	768	2.057	0.625
VoiceCLAP-large-v2	3,584	1.895	0.645

Table H.1: Blend-head model-card comparison, not a human-agreement study.

Both commercial heads reuse one audio embedding. They are therefore inexpensive relative to repeatedly querying a large audio-language judge. Shared representation also creates correlated blind spots, so agreement between the two heads is not independent validation. Mid-range blend values are less reliable than extremes, and behaviour on music or noise-only input is unspecified.

H.3 EMOTION, VOICENET, AND PROMPT COMPATIBILITY

VoiceNet (Anonymous, 2026e) describes how something is said rather than what is felt. Its 59 dimensions fall into ten groups (Table H.2). Each dimension has a 0–6 scale with a written anchor for every level; for example, Tempo runs from glacially slow (0) through ordinary conversational pace (3) to auctioneer or panic speed (6). Two exceptions are Language, a 12-way category, and Content Appropriateness, a three-level scale. Language and Accent are categorical, so the predictors, probes, and ladder selection in this paper use the remaining 57 dimensions. Table H.2 lists every dimension. The anchor texts for each level will be released with the taxonomy paper, which is not yet public.

Table H.2: The VoiceNet taxonomy. *Categorical dimensions that the 57 regression heads do not predict.

Group	n	Dimensions
Rhythm & Timing	8	Tempo; Chunking; Smoothness; Articulation Clarity; Pitch Range; Emphasis; Disfluency; Structure
Social & Interpersonal	3	Stance; Focus; Vulnerability
Speaker Identity	3	Perceived Gender; Voice Age; Register
Emotion & Affect	3	Valence; Arousal; Volatility
Physical Production	4	Respiration; Tension; Cognitive Load; Attack
Spectral & Timbral	7	Brightness; Roughness; Harmonicity; Fullness; Warmth; Metallic Character; Esthetics
Temporal Dynamics	4	Velocity Flux; Dynamic Arc; Arousal Shift; Valence Shift
Language & Recording	5	Language*; Accent*; Recording Quality; Background Noise; Content Appropriateness
Resonance Placement	7	Chest Resonance; Throat Resonance; Oral Resonance; Mask Resonance; Nasal Resonance; Head Resonance; Mixed Resonance
Speaking Style	15	Casual Style; Conversational Style; Formal Style; Dramatic Style; Narrator Style; Newsreader Style; Teacher/Didactic Style; Authoritative Style; Playful Style; Cartoonish Style; ASMR Style; Whisper-Talk Style; Ranting/Angry Style; Storytelling Style; Monologue Style
Total	59	57 used by predictors, probes, and data selection

The 40-dimensional emotion detector uses a BUD-E speech representation; it is separate from the VoiceCLAP embedding used by genuineness and blend. Additional quality heads provide complementary diagnostics. VoiceNet covers 57 vocal attributes through regression and classification heads, for 114 heads in total. The recorded aggregate regression statistics are mean MAE 0.744, Pearson correlation 0.79, and within-one accuracy 0.870; classification accuracy is 0.617. These also concern model-labelled targets and vary by dimension.

VoiceCLAP prompt similarity supplies a broad audio–description compatibility term. It does not directly verify each local event, every spoken word, or fine emotional timing. Input-side emotion/style perception supports dialogue interpretation, while output-side genuineness, blend, prompt similarity, and ASR support inspection. The audited ECAPA speaker diagnostic runs after streaming and is not a term in the original candidate-selection score.

H.4 ASR AS AN EXPERIMENTAL INSTRUMENT

The studies compare Parakeet TDT 0.6B v3 and Whisper large-v3-turbo as transcript checks. The paired-noise standard deviation in the reported repeated-condition comparison is 0.0346 for Parakeet versus 0.0833 for Whisper. In a distinct single-clip analysis the standard deviations are 0.1340 and 0.3115. These measurements show that the choice of ASR system can materially affect resolution; they do not mean that every WER estimate gains a universal factor in statistical efficiency.

Parakeet's WER was systematically lower in the paired evaluations, with mean offsets of roughly -0.024, -0.023, and -0.018 in three comparisons. The offset did not produce an observed guidance-dependent slope in the examined range (slope difference approximately -0.00081, $t = -0.19$). Every table should still identify its ASR engine, normalisation, language handling, and aggregation unit. WER can exceed one through insertions and is not automatically confined to $[0, 1]$.

H.5 ORIGINAL CANDIDATE REWARD

For candidate set $\mathcal{Y}$, the recorded ranker independently min–max normalises genuineness G, blend B, and prompt similarity C within the set, then computes

$$R_0(y) = \left[\widetilde{G}(y) + \widetilde{B}(y) + 2\widetilde{C}(y)\right] q(\mathrm{WER}(y, x)), \qquad q(w) = 1 - \min(1, \max(0, w)). \tag{4}$$

The constant-valued normalisation convention returns 0.5 when the range is below 10^{-9} or is not finite. The code clips WER to $[0, 1]$ before subtraction, both in the historical study mirror and the current server. This corrects the report's shorthand description of a raw inverse-WER factor. The historical option to flatten the gate above a knee remains in the code, but the default knee is zero, so it is inactive. The current server additionally supports the burst-aware term described in Appendix L. This is a within-set selection score, not a calibrated absolute quality scale.

The score does not explicitly test that a requested burst occurred. A clip can therefore win by being broadly compatible and fluent while omitting the difficult event. Conversely, a blend head can rise while intelligibility collapses. The crossed ranking study in Appendix L isolates the consequence of adding a local event-sensitive observable.

H.6 WHAT IS, AND IS NOT, VALIDATED

The old Authenticity construct had Cronbach's alpha approximately 0.185 in a 300-item study; it is not the present genuineness head. Earlier informal casting comparisons reported a rank association around 0.21 under an incompletely documented sample. Neither result substitutes for the new matched-pair study, whose first snapshot is in Table A.3. On that snapshot the genuineness head agrees with listeners on 64% of decisive choices and the blend head on 46%. The blend head's validation Pearson r of 0.625 against Gemini labels therefore does not yet carry over to human judgments of integration.

The image-domain analogy to LAION-Aesthetics is architectural and methodological: frozen embeddings with small perceptual heads can enable scalable curation. Speech already has learned MOS predictors (Huang et al., 2022), and newer expressive systems also use learned quality assessments. The distinctive goal here is to make spontaneous delivery and burst integration explicit, reusable, separately testable targets. Their names should not obscure their current weak-label provenance.

Evidence: `scorer_facts.md` and the cited model cards document the encoder and published model-label validation; the technical report covers detector, ASR, and reward details.

I Ablations of the synthesis-tool repertoire

I.1 LoRA dose study: recommendations have a scope

The dose study evaluates 79 adapters over seven weights, including zero, with 5,740 experimental cells and 17,220 generated clips. Repeated takes support paired comparisons against a matched zero-weight condition. The study separates target improvement from changes in genuineness, WER, burst behaviour, and duration. Thirty-two adapters pass the conservative category; one additional yawn adapter passes only the strong category, giving 33 with either recommendation. Lack of a recommendation is part of the result.

The source's conservative and strong categories use measured noise multiples, not listening-derived safety thresholds. For example, the measured genuineness noise standard deviation is 0.1138, giving permitted decreases of 0.3413 and 0.6826 at three and six standard deviations. The corresponding WER changes are 0.1038 and 0.2075, alongside absolute WER caps of 0.15 and 0.25. Other rules include burst and duration limits and the fraction of complete output groups. These thresholds are explicit engineering choices and may be too permissive perceptually.

Emotion adapters have a pooled target gain of 0.0352 at weight one ($t = 4.82$), but the increase from weight one to 1.5 is approximately -0.0026 ($t = -0.4$). Higher merge strength is therefore not a reliable path to stronger useful emotion. All 17 delivery directions show improvement somewhere in their tested dose range; one effect reaches 0.6957 ($t = 6.18$). These target gains use different native scales and should not be compared as interchangeable effect sizes.

Speaker similarity can fall while a delivery target improves. Representative comparisons decrease from 0.548 to 0.484 and from 0.559 to 0.475. The full coefficient table therefore gives both the recommended strength and adverse metric changes. Applying a dose found on one reference and adapter stack to a much larger stack requires a new check.

I.2 CFG over the autoregressive codec

Classifier-free guidance contrasts conditional and neutralised log probabilities,

$$\log \widehat{p}_g(z_t) = \log p_{\text{neutral}}(z_t) + g\left[\log p_{\text{cond}}(z_t) - \log p_{\text{neutral}}(z_t)\right], \tag{5}$$

followed by the normalisation needed for sampling. Here neutralisation removes the target expressive direction while retaining words, timing, and requested burst structure. It is not an empty prompt. Guidance is applied through the interleaved autoregressive codec prediction path. At $g = 1$, the implementation agrees with the conditional path to approximately 9.5×10^{-7} in the diagnostic numerical check.

The broad sweep spans 69 dimensions, 40 emotion and 29 VoiceNet, with 2,473 conditions, 7,419 cells, and 29,676 clips. It uses three prompts per relevant condition and four takes per cell, based on SFT3 plus the published phase-two DPO checkpoint at step 5022. The recorded generation seed is 42. Matched settings isolate the change in guidance while keeping the base stack fixed.

Family / selected guidance	Target at 1	Target at g	Δtarget	t	WER change
Emotion, $g = 3$	.2629	.2882	+.0254	2.18	.079 → .089
VoiceNet, $g = 2.5$	.4773	.5342	+.0569	3.12	.135 → .192

Table I.1: CFG sweep summaries. Selected settings are exploratory and incur a separate inference cost.

Runtime is approximately 1.89–2.04 times the unguided path in the measured setup. Running at guidance two, three, or four has similar cost because both branches still execute; a smaller numerical guidance value does not eliminate the second forward path. The frame-local tower runs for both branches and codec decoding accounts for only about 1.6% of the measured time. Thus CFG should be compared under both equal-candidate and equal-compute budgets.

Scream-specific follow-up. On 496 verified training rows, one inline scream comparison improves strict hit rate from 0.208 to 0.583 at guidance four. A prompt-only branch can be more efficient than

a differential-adapter construction: the E prompt branch gives family-hit change $+0.132$ ($t = 3.25$) with WER change -0.014, while differential-adapter D variants increase WER by approximately 0.09–0.20. These are target-specific results; the same recipe did not create adequate coverage for every rare event. Shriek has only 48 strict source rows in the relevant audit, and frustrated-groan variants never exceed 0.083 strict hits in the cited comparison.

I.3 Activation steering: reachable targets and quality coupling

For a hidden-state tap ℓ, a steering intervention adds a scaled direction to the generation state. Directions are derived from high-versus-middle attribute groups; the exact feature normalisation and scaling convention belong to the corresponding vector manifest. A coefficient called 0.1 in one normalisation is not portable to an unrelated extraction.

The initial coarse sweep began at strength 0.5 and gave a misleadingly negative picture because useful behaviour lies at substantially lower strengths. The follow-up h20 sweep in Figure I.1 reveals a trade-off. At strength 0.10, the emotion target rises from 0.4354 to 0.5840 and WER falls from 0.166 to 0.115, but genuineness falls from 3.77 to 3.15. At 0.30, the target reaches 0.6595 while WER reaches 0.545 and genuineness 1.90. A higher detector score is not an unqualified improvement.

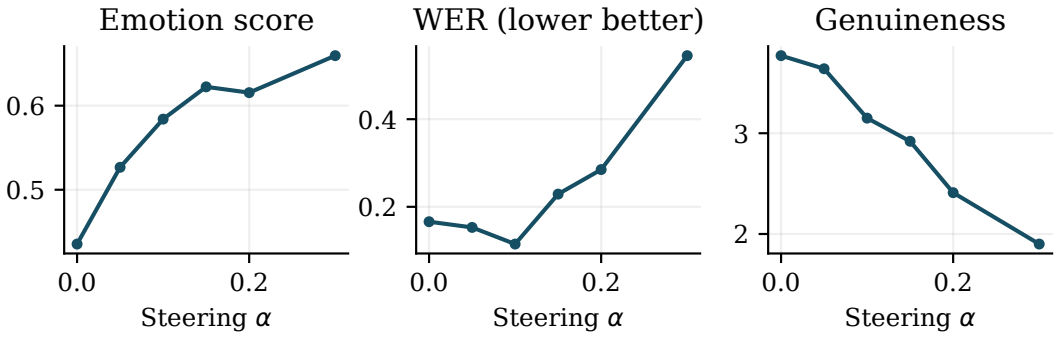

Figure I.1: Low-strength h20 steering sweep. Values are transcribed from the dedicated study; error bars are not invented where the summary lacks them.

Random-direction controls help separate generic perturbation from target alignment. At 0.10, the random control has emotion 0.4490, WER 0.132, and genuineness 3.69; at 0.20 it has emotion 0.4395, WER 0.118, and genuineness 3.90. The target direction is more effective at moving its metric, but its perceptual-proxy cost is not merely inevitable noise injection.

An informal operator listening report led to removal of steering from the standard adapter mode. The report does not provide a panel size or controlled preference study; we describe a deployment decision based on adverse observation, not a human-significant ranking. The API can still expose experimental combinations, which should be explicitly labelled in reproducibility records.

I.4 Full factorial: interactions are not free gains

The factorial experiment covers 60 attributes: 40 emotions, 17 delivery directions, and three quality dimensions. It records 10,459 rows and 31,377 clips over the eight corners of adapter, steering, and CFG combinations. Target changes are expressed in target-standard-deviation units for the factorial analysis, unlike the native-scale dose table.

Target family	Adapter main effect	Steering main effect	CFG main effect
Emotion	+.077	+.384	+.050
Delivery	+.377	+.614	+.026
Quality	+.399	+.006	+.032

Table I.2: Recorded standardised target main effects in the factorial analysis. The complete numeric report archive preserves interaction and cost tables.

The combined target effect is about 1.24 times the sum of the individual effects, but the WER cost is about 1.91 times the sum and the genuineness cost about 1.80 times the sum. The largest adverse interaction involves steering with CFG. Combining tools is therefore not a monotonic improvement, even when each tool individually moves a target.

One hypothesised interaction mechanism was that merging a LoRA rotates the steering direction. The measured direction cosine is 0.99579 with the adapter, compared with a noise benchmark around 0.95558. That comparison does not support a large rotation as the main explanation. It does not rule out nonlinear generation effects elsewhere in the network.

I.5 Prompting and event placement

A 15-class prompt experiment contains 5,360 cells and 32,160 clips. Adding a causal description produces strict-hit gain around 0.026; specifying duration gives around 0.022; the combination yields around 0.044 in the reported summary. The family comparison is weaker ($t \approx 1.9$), while one strict comparison is approximately $+0.030$ ($t = 3.3$). These closely related summaries use different contrasts and should not be merged into a single universal effect.

An action cue can instead increase missing-event probability by 0.120 ($t = 3.4$), while blend rises by 0.183 and a quality proxy by 0.032. Inspection attributes this to more silence, which the proxies can prefer. Mid-clause placement increases misses by roughly 0.306–0.366 across all 15 tested classes. This supports separately checking event occurrence and placement rather than rewarding an attractive whole-clip embedding alone.

The later 55-class extension does not reproduce a general advantage for the combined cue at matched weight: the pooled difference is approximately -0.004 ($t = -0.6$). The library retains class-specific recipes and unsupported cases, not a blanket instruction to add more descriptive prose. English delivery cues also outperformed translated German cues in one small matched diagnostic (median WER 0.000 versus 0.267); that format sensitivity is not a claim about overall German synthesis ability.

I.6 Load, acoustics, and transition failures

Stacking seven adapters in one diagnostic increases WER from 0.063 to 0.554 while genuineness changes only from 4.16 to 4.04 and blend rises from 6.59 to 7.17. A spectral explanation based on simple high-frequency veiling was not supported: the measured high-frequency energy fraction increased from 0.0036 to 0.0131. The failure must be investigated with intelligibility, local timing, and listening, not explained away by an unverified acoustic metaphor.

Crossfade and transition experiments test a different lever: joining or interpolating outputs and references. Strong overall duration following does not guarantee a correctly placed emotional transition or a natural seam. These studies motivate local transition metrics and held-out listening examples; they do not justify replacing all continuous generation with splicing.

Evidence: `lora_dose_study.html`, `cfg_study.html`, `steering_study.html`; technical report factorial, prompt-form, load, rate, transition, and spectral sections; `reports/burst_cfg.md` and `crossfade_v2.md`.

I.7 COMPLETE DOSE RECOMMENDATIONS

Table I.3: All 79 adapter recommendations from the dose study. Rec. and Strong are noise-derived guardrail categories, not perceptual safety guarantees. Deltas use the recommended weight and native target scales. A dash denotes no recommendation.

Adapter	Rec.	Strong	Δtarget	ΔWER	ΔG
burst/blend_genuine_emotional	1.500	1.500	1.033	0.010	0.148
emotion/Affection	1.250	1.250	0.090	-0.026	0.029
emotion/Amusement	–	–	–	–	–
emotion/Anger	–	–	–	–	–
emotion/Astonishment_Surprise	–	–	–	–	–
emotion/Awe	–	–	–	–	–
emotion/Bitterness	–	–	–	–	–
emotion/Concentration	0.500	0.500	0.083	-0.009	0.123
emotion/Confusion	–	–	–	–	–
emotion/Contemplation	1.500	1.500	0.118	0.007	0.033
emotion/Contempt	–	–	–	–	–
emotion/Contentment	–	–	–	–	–
emotion/Disappointment	–	–	–	–	–
emotion/Disgust	–	–	–	–	–
emotion/Distress	–	–	–	–	–
emotion/Doubt	–	–	–	–	–
emotion/Elation	–	–	–	–	–
emotion/Embarrassment	–	–	–	–	–
emotion/Emotional_Numbness	–	–	–	–	–
emotion/Fatigue_Exhaustion	–	–	–	–	–
emotion/Fear	–	–	–	–	–
emotion/Helplessness	–	–	–	–	–
emotion/Hope_Enthusiasm_Optimism	0.750	0.750	0.089	-0.006	0.021
emotion/Impatience_and_Irritability	–	–	–	–	–
emotion/Infatuation	–	–	–	–	–
emotion/Interest	1.000	1.000	0.083	-0.018	0.070
emotion/Intoxication_Altered_States_of_Consciousness	0.750	0.750	0.096	-0.011	0.356
emotion/Jealousy_and_Envy	–	–	–	–	–
emotion/Longing	–	–	–	–	–
emotion/Malevolence_Malice	–	–	–	–	–
emotion/Pain	0.750	0.750	0.120	0.035	0.246
emotion/Pleasure_Ecstasy	1.250	1.250	0.114	0.004	-0.068
emotion/Pride	–	–	–	–	–
emotion/Relief	1.000	1.000	0.160	-0.018	0.012
emotion/Sadness	–	–	–	–	–
emotion/Sexual_Lust	1.250	1.250	0.164	-0.049	-0.151
emotion/Shame	–	–	–	–	–
emotion/Sourness	–	–	–	–	–
emotion/Teasing	1.500	1.500	0.085	0.003	-0.178
emotion/Thankfulness_Gratitude	–	–	–	–	–
emotion/Triumph	–	–	–	–	–
qual/blend_high	–	–	–	–	–
qual/esthetics_high	1.500	1.500	0.571	-0.020	-0.126
qual/genuineness_high	–	–	–	–	–
vb/ahem	–	–	–	–	–
vb/cackle	–	–	–	–	–
vb/chuckle	–	–	–	–	–
vb/contented_sigh	–	–	–	–	–
vb/cough	–	–	–	–	–
vb/low_mumble	–	–	–	–	–
vb/scream	1.500	1.500	0.296	-0.007	0.002
vb/sharp_inhale	1.250	1.250	0.300	-0.028	0.063
vb/soft_hum	–	–	–	–	–

Continued on next page

Adapter	Rec.	Strong	Δtarget	ΔWER	ΔG
vb/surprised_gasp	–	–	–	–	–
vb/wistful_sigh	1.500	1.500	0.185	0.061	-0.056
vb/yawn	–	1.500	–	–	–
vbr/fast_breathing	–	–	–	–	–
vbr/growl	–	–	–	–	–
vbr/guffaw	–	–	–	–	–
vbr/nervous_giggle	–	–	–	–	–
vbr/panting	–	–	–	–	–
vbr/sobs	–	–	–	–	–
vn/AROU_high	1.500	1.500	1.296	0.075	-0.109
vn/AROU_low	1.500	1.500	-0.666	-0.019	-0.052
vn/ARSH_high	1.500	1.500	1.038	-0.002	-0.137
vn/ARSH_low	1.250	1.250	-0.241	-0.013	0.185
vn/EMPH_high	1.250	1.250	0.669	0.023	0.135
vn/EXPL_high	1.500	1.500	0.192	0.020	0.110
vn/S_ASMR_high	1.500	1.500	1.322	-0.014	-0.069
vn/S_DRAM_high	1.500	1.500	1.268	0.002	0.033
vn/S_RANT_high	1.500	1.500	1.366	0.021	-0.079
vn/TENS_high	1.500	1.500	1.084	0.054	0.126
vn/VALN_high	1.000	1.000	0.388	0.006	0.157
vn/VALN_low	1.250	1.250	-0.251	-0.018	-0.017
vn/VALS_high	–	–	–	–	–
vn/VALS_low	1.250	1.250	-0.224	-0.010	0.218
vn/VFLX_high	1.500	1.500	0.951	0.014	0.011
vn/VOLT_high	1.500	1.500	0.428	-0.006	0.124
vn/VULN_high	1.250	1.250	0.188	0.005	0.089

J LAYER FORENSICS: REPRESENTATIONS, ENSEMBLES, GEOMETRY, AND BURSTS

J.1 QUESTIONS AND EXTRACTION DESIGN

The four studies ask complementary questions: W1 measures where attributes are decodable; W2 tests whether pooled states from multiple layers carry complementary information; W3 characterises attribute directions and their relation to quality; W4 tests burst presence and identity while examining temporal confounds. None uses new human emotion labels.

The primary extraction, p3, contains 217,583 utterances from 13,757 speaker/voice keys: 164,653 training, 20,826 validation, and 32,104 test utterances. Splitting is by speaker key, with no speaker or utterance shared between training and test. A naive UID split in an earlier view leaked 6,282 voice keys, which is why the speaker-disjoint view is primary. Speaker disjointness does not automatically remove source, script-template, or teacher-model confounding.

SFT3 is teacher-forced on the observed audio tokens. Features average hidden states over the final 20% of assistant audio frames, with a minimum of eight and maximum of 100 frames; the median window is about 2.41 seconds. These are audio-generative hidden states in a text-conditioned computation, not a text-free audio encoder. Neutralising expressive captions leaves the lexical input available.

There are 38 taps. h00 is the input embedding state; h01–h35 are outputs of transformer blocks 0–34; h36 is the final, normalised output of block 35; and loc is a frame-local representation. Off-by-one naming matters when reproducing a preferred layer. A statement about layer 6 in this appendix refers to tap h06, not an unspecified zero-based module name.

J.2 W1: PER-LAYER DECODABILITY

The joint probe predicts 99 outputs: 40 emotions, 57 VoiceNet attributes, genuineness, and blend. The MLP has input width 2,560, one hidden layer of width 512, GELU, dropout 0.1, and 99 outputs. It uses Huber loss with delta one, AdamW with learning rate 10^{-3} and weight decay 10^{-4}, batch size 512, a cosine schedule for up to 60 epochs, and validation early stopping with patience eight. Features and targets are standardised using training statistics. The ridge baseline chooses its regularisation from the recorded range 1–10,000 on validation.

Table J.2 prints every attribute's descriptive peak and fixed h06/h36 values; the archive retains all taps and metadata baselines. Family means have peaks around h12 for vocal attributes and genuineness, h20 for emotion, and h25 for blend. Late states are not uniformly better than middle states.

Family / target	Tap	MLP R^2	Ridge R^2	Metadata R^2
15 talking-style subset	h12	.859	.807	–
VoiceNet, 57	h12	.846	.788	.280
Genuineness	h12	.719	.539	.190
Emotion, 40	h20	.473	.437	.108
Blend	h25	.435	.356	.100

Table J.1: W1 descriptive test-set peaks. Ridge is evaluated at the indicated tap. The metadata value for genuineness follows the numerical artifact; an older HTML summary differs slightly.

The MLP improves over ridge on all 99 dimensions in the reported best-probe comparison, with median gain approximately 0.043 and genuineness gain approximately 0.180. Seed variation is much smaller for most cells: mean spread approximately 0.00308, 95th percentile 0.01084, and maximum 0.0432 in the cited stability analysis. Treating differences below about 0.01 as ties is a practical heuristic, not a confidence interval.

Removing expressive captions costs no more than approximately 0.015 R^2 in the reported comparisons. This reduces one obvious source of label leakage, but does not remove the spoken words, duration, voice information, or other source cues. It would be incorrect to conclude that the probe necessarily recognises emotion from acoustics alone.

The W1 peaks were read from the test grid and are explicitly descriptive. Their numerical advantage over a neighbouring tap must not be presented as an independently evaluated hyperparameter choice. W2 separately selects ensembles on validation and supplies width-matched controls.

J.3 W2: SELECTED AVERAGES FROM MULTIPLE LAYERS

Each layer is first averaged over time; selected layer vectors are then *concatenated*, not averaged into a single cross-layer vector. This distinction retains layer-specific information. W2 performs 4,888 fits across the primary and secondary extraction views in approximately 13.05 GPU-hours. The primary view includes validation-guided greedy, beam, and random combination searches.

For emotion, the best single tap h20 gives test $R^2 = 0.46631$ in this separately trained probe. The selected six-tap ensemble h06, h12, h18, h20, h21, h23 reaches 0.50674. Across four seeds, the mean improvement is approximately 0.0385 ± 0.0037 in the recorded summary. The inclusion of h06 is the specific layer-6 finding: a layer can add complementary information without being the strongest standalone representation.

VoiceNet improves from approximately 0.8529 to 0.8695, with four-seed gain 0.0166 ± 0.0003. The performance subset improves from 0.8518 to 0.8701 and the timing-related subset from 0.8403 to 0.8581. Blend improves in the selected run from 0.3841 to 0.4551, but its seed-level gain is 0.0616 ± 0.0554; genuineness improves from 0.7086 to 0.7388 with gain 0.0181 ± 0.0244. The latter two are substantially less stable than the broad emotion and VoiceNet findings.

Controls match input width by either repeating the best single-layer feature or padding it with noise. All 60 selected winners across the source analysis beat the corresponding capacity controls. This rules out the simplest explanation that merely adding input coordinates accounts for the selected gains, but not every possible optimisation or regularisation explanation. Closely related layers can still complement one another: the h18/h21 pair has high representational similarity, approximately 0.979, yet gains around 0.020 R^2.

J.4 W3: ATTRIBUTE GEOMETRY AND QUALITY AXES

W3 constructs 156 directions at 38 taps from 309,631 pooled means, yielding a float32 array of shape $156 \times 38 \times 2560$. It is a descriptive full-corpus geometry analysis, not a train/test prediction experiment. Ninety-seven directions were recomputed bit-exactly in an audit, supporting artifact consistency.

Positive directions contrast high and middle attribute groups. The negative or low direction is not generally the exact sign reversal: the median cosine between high-minus-middle and the corresponding low-side construction is approximately -0.0004. A single bipolar line is therefore an inadequate description of all attribute levels in this extraction.

Emotion directions become increasingly anti-aligned with the quality axis at later taps: reported family summaries are approximately -0.619 at h12, -0.746 at h20, -0.842 at h32, -0.862 at h36, and -0.954 at loc. VoiceNet has different signs, approximately $+0.460$, $+0.461$, $+0.383$, $+0.383$, and $+0.706$. This geometry is consistent with some adverse steering trade-offs, but does not prove a universal incompatibility between emotion and quality. The axes inherit the corpus composition and model-defined labels.

J.5 W4: BURST PRESENCE IS DECODABLE, LOCALISATION IS HARDER

The balanced presence design uses 8,000 positive and 8,000 negative clips with five speaker-grouped folds. At h19, ridge AUC is approximately 0.9777 and MLP AUC 0.9842; balanced accuracy is approximately 0.9232, versus 0.587 for a metadata baseline. An eight-class burst-identity analysis has its strongest tap around h18. Presence and identity are separate prediction tasks.

Temporal controls limit the interpretation. Of 118,660 annotated bursts in the relevant source analysis, only 18.4% occupy at least half the extraction window; 81.6% are outside under that criterion. Of 79,402 clips containing a burst, only 19,580 have an in-window event, approximately 24.7%. Coverage also depends strongly on class: about 49% for yawn and 37% for scream, but only 8.7% for sigh and 8.0% for inhale.

Same-voice presence classification remains near AUC 0.969 for in-window and 0.968 for out-of-window cases. Even an out-of-window confound control reaches approximately 0.778. High whole-clip classification performance therefore need not mean precise local burst detection: the representation can encode scene, prosodic context, or earlier events. A local reward needs explicit temporal validation, not only a high global AUC.

J.6 Interpretation and the small-data perspective

These results motivate using expressive speech generators as reusable feature extractors. A small MLP on pooled states could eventually recognise whispering, shouting, accent, or another attribute with hundreds to a thousand new labels. That is a prospective learning-curve experiment, not a conclusion from probes trained on 164,653 utterances.

Useful tests would vary labelled training size, hold out speakers and recording sources, compare with VoiceCLAP and specialist self-supervised encoders, and measure calibration under a new language or accent distribution. Teacher forcing must also be distinguished from extracting features on freely generated audio. The former can exploit conditioning and sequence structure unavailable to a conventional standalone perceptual model.

Finally, the psychological context should clarify rather than inflate the result. Correlations around $r = 0.4$ can be large in many behavioural settings (Funder & Ozer, 2019; Gignac & Szodorai, 2016), but effect magnitude depends on measurement reliability and the construct. Here the emotion result is macro-R^2 against model outputs. Taking its square root and calling it a human emotion correlation would be invalid.

Evidence: `layer_forensics_report.html`; `layer_forensics/w1/`, `w2/`, `w3/`, and `w4/`; exported `w1_grid.json`, `w2_ensembles.json`, and W4 diagnostics.

J.7 ALL 99 ATTRIBUTE RESULTS

Table J.2: W1: all 99 attributes. Values are test R^2; ridge uses the MLP-selected tap. Peaks are descriptive test-set selections. h06 and h36 use MLPs.

Attribute	Peak tap	MLP	Ridge	h06	h36
Affection	h19	0.348	0.309	0.261	0.195
Amusement	h20	0.719	0.707	0.681	0.631
Anger	h21	0.523	0.486	0.483	0.339
Astonishment Surprise	h20	0.399	0.374	0.313	0.245
Awe	h20	0.239	0.216	0.126	0.097
Bitterness	h19	0.386	0.340	0.331	0.218
Concentration	h20	0.622	0.602	0.554	0.514
Confusion	h20	0.329	0.303	0.258	0.168
Contemplation	h20	0.626	0.585	0.562	0.490
Contempt	h19	0.471	0.413	0.439	0.296
Contentment	h18	0.425	0.390	0.345	0.235
Disappointment	h21	0.464	0.428	0.396	0.320
Disgust	h20	0.369	0.332	0.311	0.209
Distress	h21	0.688	0.636	0.611	0.571
Doubt	h21	0.380	0.352	0.308	0.222
Elation	h20	0.503	0.481	0.458	0.316
Embarrassment	h21	0.462	0.430	0.417	0.326
Emotional Numbness	h20	0.382	0.334	0.320	0.302
Fatigue Exhaustion	h20	0.562	0.471	0.467	0.454
Fear	h21	0.597	0.539	0.518	0.500
Helplessness	h21	0.694	0.637	0.614	0.581
Hope Enthusiasm Optimism	h19	0.454	0.436	0.396	0.232
Impatience and Irritability	h19	0.522	0.482	0.476	0.318
Infatuation	h20	0.391	0.344	0.318	0.305
Interest	h20	0.695	0.652	0.631	0.600
Intoxication Altered States of Consciousness	h06	0.524	0.442	0.524	0.401
Longing	h20	0.616	0.570	0.540	0.514
Malevolence Malice	h20	0.511	0.464	0.476	0.364
Pain	h21	0.569	0.483	0.476	0.446
Pleasure Ecstasy	h20	0.458	0.433	0.384	0.254
Pride	h20	0.450	0.437	0.417	0.367
Relief	h20	0.364	0.321	0.298	0.290
Sadness	h20	0.694	0.644	0.618	0.582
Sexual Lust	h20	0.451	0.403	0.421	0.326
Shame	h21	0.384	0.321	0.325	0.265
Sourness	h19	0.324	0.294	0.291	0.208
Teasing	h20	0.494	0.479	0.445	0.407
Thankfulness Gratitude	h20	0.281	0.237	0.212	0.107
Triumph	h20	0.437	0.411	0.404	0.346
Jealousy and Envy	h19	0.209	0.202	0.146	0.132
AGEV	h12	0.748	0.660	0.665	0.699
AROU	h12	0.912	0.879	0.880	0.888
ARSH	h12	0.863	0.832	0.821	0.836
ATCK	h12	0.907	0.867	0.872	0.888
BKGN	h12	0.883	0.847	0.829	0.859
BRGT	h12	0.812	0.729	0.726	0.767
CHNK	h12	0.885	0.856	0.827	0.840
CLRT	h12	0.831	0.743	0.746	0.791
COGL	h12	0.813	0.753	0.759	0.768
DARC	h12	0.895	0.855	0.860	0.871
DFLU	h12	0.847	0.805	0.788	0.809
EMPH	h12	0.869	0.827	0.824	0.839
ESTH	h12	0.785	0.711	0.698	0.737
EXPL	h12	0.800	0.738	0.769	0.741
FOCS	h12	0.761	0.673	0.665	0.713

Continued on next page

Attribute	Peak tap	MLP	Ridge	h06	h36
FULL	h12	0.779	0.681	0.692	0.723
GEND	h25	0.939	0.929	0.894	0.925
HARM	h12	0.818	0.740	0.754	0.777
METL	h12	0.880	0.839	0.844	0.858
RANG	h12	0.820	0.748	0.764	0.784
RCQL	h12	0.882	0.843	0.823	0.852
REGS	h12	0.924	0.905	0.869	0.906
RESP	h12	0.858	0.818	0.815	0.826
ROUG	h12	0.771	0.685	0.709	0.727
R CHST	h12	0.849	0.780	0.783	0.814
R HEAD	h12	0.841	0.778	0.788	0.807
R MASK	h12	0.868	0.818	0.813	0.837
R MIXD	h12	0.775	0.667	0.691	0.729
R NASL	h12	0.866	0.813	0.826	0.837
R ORAL	h12	0.810	0.708	0.740	0.768
R THRT	h12	0.757	0.630	0.670	0.708
SMTH	h12	0.792	0.706	0.706	0.741
STNC	h12	0.907	0.875	0.871	0.882
STRU	h12	0.825	0.725	0.737	0.778
S ASMR	h12	0.898	0.860	0.862	0.874
S AUTH	h12	0.835	0.763	0.761	0.783
S CART	h12	0.842	0.784	0.786	0.805
S CASU	h12	0.905	0.871	0.879	0.883
S CONV	h12	0.859	0.819	0.814	0.827
S DRAM	h12	0.837	0.791	0.771	0.801
S FORM	h12	0.897	0.855	0.852	0.865
S MONO	h12	0.920	0.904	0.886	0.902
S NARR	h12	0.901	0.862	0.865	0.882
S NEWS	h12	0.846	0.744	0.786	0.811
S PLAY	h12	0.902	0.872	0.871	0.882
S RANT	h12	0.829	0.770	0.756	0.783
S STRY	h12	0.751	0.667	0.661	0.701
S TECH	h12	0.770	0.679	0.695	0.714
S WHIS	h12	0.900	0.864	0.856	0.876
TEMP	h12	0.917	0.897	0.878	0.888
TENS	h12	0.787	0.712	0.724	0.737
VALN	h12	0.833	0.790	0.792	0.801
VALS	h12	0.844	0.802	0.807	0.818
VFLX	h12	0.867	0.830	0.823	0.838
VOLT	h12	0.787	0.706	0.704	0.739
VULN	h12	0.840	0.778	0.794	0.799
WARM	h12	0.866	0.812	0.812	0.839
genuineness	h12	0.719	0.539	0.628	0.680
blend	h25	0.435	0.356	0.341	0.371

J.8 VALIDATION-SELECTED ENSEMBLE CONFIGURATIONS

Table J.3: W2 validation-selected ensembles on the primary speaker-disjoint extraction. Dup. repeats the best tap to match width; Noise pads it to the same width. Displayed scores are on test. Full seeds and search metadata are archived.

Target	k	Selected taps	Test R^2	Dup.	Noise
emo40	1	h20	0.4663	–	–
emo40	2	h18, h21	0.4867	0.4620	0.4200
emo40	3	h12, h18, h21	0.4928	0.4642	0.4010
emo40	4	h06, h12, h18, h20	0.4973	0.4638	0.3903
emo40	5	h06, h12, h18, h20, h23	0.5027	0.4590	0.3722
emo40	6	h06, h12, h18, h20, h21, h23	0.5067	0.4653	0.3652
vn57	1	h12	0.8529	–	–
vn57	2	h12, h20	0.8606	0.8530	0.8143
vn57	3	h11, h12, h23	0.8636	0.8527	0.7989
vn57	4	h09, h12, h13, h20	0.8667	0.8528	0.7888
vn57	5	h09, h11, h12, h13, h20	0.8686	0.8528	0.7791
vn57	6	h09, h11, h12, h13, h20, h21	0.8695	0.8527	0.7727
gen	1	h23	0.7086	–	–
gen	2	h19, h23	0.7252	0.6983	0.6307
gen	3	h19, h23, h36	0.7210	0.6936	0.6236
gen	4	h18, h23, h26, h28	0.7362	0.6656	0.6111
gen	5	h19, h21, h23, h27, h32	0.7397	0.6839	0.6034
gen	6	h19, h21, h23, h25, h26, h27	0.7388	0.6711	0.5850
blend	1	h33	0.3841	–	–
blend	2	h19, h32	0.4470	0.3166	0.2988
blend	3	h12, h19, h32	0.4628	0.2816	0.2679
blend	4	h12, h18, h23, h25	0.4647	0.3020	0.2718
blend	5	h15, h23, h32, h33, loc	0.4435	0.2582	0.2629
blend	6	h10, h15, h23, h32, h33, loc	0.4551	0.3311	0.2539
vn_perf16	1	h12	0.8518	–	–
vn_perf16	2	h12, h25	0.8597	0.8537	0.8155
vn_perf16	3	h12, h13, h26	0.8642	0.8541	0.7987
vn_perf16	4	h11, h12, h13, h20	0.8677	0.8547	0.7874
vn_perf16	5	h11, h12, h13, h20, h26	0.8695	0.8548	0.7798
vn_perf16	6	h11, h12, h13, h20, h23, h26	0.8701	0.8549	0.7737
vn_time19	1	h12	0.8403	–	–
vn_time19	2	h12, h21	0.8477	0.8413	0.7981
vn_time19	3	h11, h12, h26	0.8526	0.8422	0.7791
vn_time19	4	h11, h12, h13, h21	0.8553	0.8424	0.7695
vn_time19	5	h08, h11, h12, h13, h21	0.8568	0.8424	0.7599
vn_time19	6	h08, h11, h12, h13, h21, h26	0.8581	0.8426	0.7526

K VOCAL BURSTS: TAXONOMY, DATA, DETECTORS, AND ABLATIONS

K.1 THREE LABEL SYSTEMS, THREE DIFFERENT QUESTIONS

The reports contain several burst vocabularies that must not be conflated. The generative request vocabulary names fine events such as a relief sigh or a nervous giggle. A 23-group semantic analysis groups 49 fine request classes to examine acceptable near misses. The production classifier instead has 17 output classes: 16 positive classes and `no_burst`. Those 16 positives map to seven coarse families: breath, sigh, groan, laugh, hum, scream, and yawn.

The breath family contains Deep Breath, Heavy Breathing, Panting, and Sharp Inhale; sigh contains Exasperated Sigh, Relief Sigh, and Wistful Sigh; groan contains Affirmative Grunt, Exhausted Groan, and Frustrated Groan; laugh contains Breathy Giggle and Chuckle; hum contains Humming and Soft Hum. Scream and Yawn are singleton families. Family agreement is a weaker requirement than exact-class agreement, not an alternative spelling of the same metric.

K.2 TRAINING DATA AND INDEPENDENT RELABELLING

The curated real-speech release contains 3,598 utterances across 49 request classes and 5,161 burst segments. Of the utterances, 2,727 pass the strict detector gate and 871 use a family-relaxed gate. The relaxed rows are explicitly marked. The selection ladder caps well-supported classes at 200, uses 100 where appropriate, and otherwise takes the available publishable material. Thirty-six of 49 classes are exhausted rather than capped, so the small classes reflect scarcity, not a balanced design.

An additive selection repair admitted family-relaxed classes whose strict bucket file did not exist. A later round-trip check removed 69 selected rows that did not reconstruct correctly. This distinction explains why a class selected at 200 can ship fewer than 200 records. Counts of selected, parsed, shipped, strict, and relaxed examples remain separate.

The synthetic DramaBox collection contains 73,500 clips from 24,500 prompts at three seeds. Wider model relabelling covers 123,004 events and a broader vocabulary than the original detector. On the curated real data, the reported agreement is approximately 64.8% at a family level but only 19.0% at the strict label level. It is agreement between models, not human-certified correctness.

The large-corpus reannotation similarly predicts new labels only within existing spans. Its label distributions are heavily concentrated in a few classes and can differ sharply from old labels. A new model preferring Sharp Inhale over an old Low Mumble label does not, by itself, establish which annotation is correct.

K.3 LOCALISER AND PRODUCTION CLASSIFIER

Burst localisation determines candidate intervals; classification determines an interval's type. Speech-adjacent boundaries and event overlap are difficult, and short events can be lost under a minimum-duration gate. Negative examples mix speech and silence, approximately 75% and 25%, with exclusion margins around detected events. Different localiser thresholds, gap merging, minimum lengths, or vetoes produce different event inventories even with the same classifier.

The production detector uses VoiceCLAP-large-v2 features and five trained members. Its held-out production figures come from one grouped split; five members do not create five independent test sets. The archive preserves both production recall and a separately named repeated balanced-split statistic. These should not be pooled as if they were the same estimator.

The encoder comparison reports validation accuracy approximately 0.7056 and real-subset accuracy 0.4659 for large-v2, versus 0.6258/0.3929 for commercial and 0.6398/0.4024 for small-v2. Dataset, label, and split definitions must accompany those numbers. Large and commercial variants carry different practical compute costs; the small variant also has a non-commercial restriction in the inspected release metadata.

Table K.1: Production detector: real (R) and DramaBox (D) held-out recall. JSON retains intervals, family recall, confusion counts, and low-sample flags. These are detector diagnostics, not generator hit rates.

Class	Family	n_R	Rec.$_R$	Fam.$_R$	n_D	Rec.$_D$
Affirmative Grunt	groan	30	0.700	0.767	25	0.680
Breathy Giggle	laugh	32	0.469	0.969	28	0.857
Chuckle	laugh	53	0.641	0.849	40	0.800
Deep Breath	breath	59	0.186	0.712	51	0.686
Exasperated Sigh	sigh	53	0.226	0.340	42	0.381
Exhausted Groan	groan	31	0.484	0.871	25	0.320
Frustrated Groan	groan	32	0.594	0.812	26	0.654
Heavy Breathing	breath	25	0.120	0.640	25	0.280
Humming	hum	36	0.472	0.611	31	0.516
Panting	breath	34	0.647	0.882	26	0.462
Relief Sigh	sigh	32	0.062	0.312	32	0.469
Scream	scream	37	0.595	0.595	27	0.963
Sharp Inhale	breath	39	0.897	0.897	32	0.531
Soft Hum	hum	28	0.393	0.571	25	0.280
Wistful Sigh	sigh	28	0.143	0.393	25	0.400
Yawn	yawn	37	0.460	0.460	30	0.500
no_burst	no_burst	802	0.980	0.980	110	0.964

Recall is not a hard ceiling on generator hit rate. On held-out real examples, strict class recall ranges from approximately 0.062 for Relief Sigh to 0.897 for Sharp Inhale; the no-burst recall is 0.980. Many classes have only a few dozen examples, and nine of 17 classes have a low-sample warning in at least one source. A generator evaluation must acknowledge this uneven instrument.

Some exploratory notes described detector recall as an absolute ceiling. That is too strong. Under a stable binary abstraction, the observed hit rate is approximately

$$h \approx p\,r + (1-p)\,f, \tag{6}$$

where p is actual event prevalence, r detector sensitivity, and f the false-positive rate. Domain shift can change both r and f, and easier generated examples can have higher sensitivity than the held-out real set. A measured hit can therefore exceed held-out recall. Recall remains essential context, but division by recall or ceiling language is not automatically justified.

K.4 Semantic groups and chance controls

In the 49-class analysis, grouping into 23 meaningful groups raises mean strict agreement from approximately 0.302 to 0.537. The study compares against 300 random partitions to distinguish useful semantic grouping from the automatic increase obtained by merging labels. The chance-adjusted improvement is approximately 0.182. The number of classes exceeding a 0.15 agreement threshold rises from 28 to 44.

This supports accepting some semantically related events as partial success. It does not justify calling a generic breath the requested emotional reaction in every scene. The arena's prompt-fit preference is needed to assess whether a near miss is artistically acceptable. The 23-group analysis also must not be described as the seven-family mapping of the production head.

K.5 Data cleaning, rank, and synthesis ablations

Strict detector auditing leaves only about 9.8% of one historical burst-adapter training pool, with 34 of 70 classes empty. Re-training on such a filtered pool improves label consistency at the cost of coverage. Increasing adapter rank from 32 to 64 does not establish gains beyond the reported training noise. Sparse, inconsistent supervision cannot be assumed to become adequate simply by increasing adaptation capacity.

Voice-converted and manufactured rows were tested because they can cheaply broaden identity coverage. They sometimes erase, smooth, or transform the burst, making the transcript or requested

label misleading. Comparisons therefore separate original real, original synthetic, converted, strict-filtered, and family-relaxed examples rather than treating them as exchangeable.

Scream is a concrete exception where synthetic supplementation can help. A 10% DramaBox dose does not establish a benefit in the cited small comparison, while 25% produces strict hits around 0.333 versus 0.033 for the real-only baseline. This remains dose- and setting-dependent. At full adapter weight two, WER can reach 0.775 inline and 2.240 in solo prompting, with substantial genuineness loss. Better event recall is not useful if the surrounding performance becomes unintelligible.

Stopping and length-related preference training were also investigated to reduce repetitions and spillover. Those objectives are distinct from making a burst sound spontaneous or placing it correctly. The release retains these pair families and their failure notes, but they are not presented as completed solutions to local event control.

K.6 Combining members, group adapters, and production transfer

Combining multiple member adapters within a semantic group did not provide a general benefit: the cited eight-group comparison was negative in all eight groups. Pooled group training also failed to establish the intended gain. A single best member can work in a lightly loaded laboratory configuration yet fail when added to the deployment stack.

The most instructive transfer comparison changes a best-member advantage of approximately $+0.0762$ ($t = 3.76$) in a four-adapter setting into a disadvantage of approximately -0.0531 ($t = -2.50$) in the seven-adapter setting. This motivates measuring combinations where they will actually be used. It also explains why WikiSkills entries need stack provenance and must not be universalised from an isolated adapter sweep.

The later prompt extension and dose studies identify useful recipes for some classes and explicitly absent capability for others. Unsupported classes should result in a disclosed fallback, a semantically justified alternative, or a failed-output record, not an invented guarantee that every textual burst tag is executable.

K.7 What the burst programme establishes

The experiments distinguish at least four failure sources: weak or missing training labels, an unavailable production skill, poor temporal placement, and an evaluator that does not observe the intended event. They also expose a fifth, system-level source: interacting adaptations can damage the carrier speech.

These findings support event-aware evaluation and a measured tool library. They do not yet establish human-valid fine taxonomy recognition or uniformly natural burst integration. The blend human study deliberately includes a burst in both candidates so that preference concerns integration rather than merely noticing an event. In its first 67 decisive choices, listeners agreed with the blend head's ordering 46% of the time (Table A.3), so the head cannot yet be treated as a human-aligned integration measure. A separate presence-sensitive test is still needed for complete systems.

Evidence: `burst_17klassen_zuordnung.html`, `burst_klassengruppen.html`; technical report burst studies 43–70; dedicated `reports/burst_data_audit.md`, `burst_retrain.md`, `burst_rank.md`, `burst_synth.md`, `burst_dose_study.md`, `burst_dose_extended.md`, `burst_levers.md`, `burst_ext.md`, `burst_stop_dpo.md`, `gemini_traindata.md`, `burst_detector2.md`, and `vb_cls2.md`.

L BEST-OF-N: SELECTION, MISSING OBSERVABLES, AND COST

L.1 CANDIDATE-SET DESIGN

The ranking study contains 1,940 candidate sets and 15,517 takes, three fewer than $1{,}940 \times 8$. Missing outputs are disclosed rather than silently replacing the denominator with the intended count. The reconstruction includes the actual adapter families used by the server, with separate carrier-voice-present and carrier-voice-absent sets, and lower- and higher-load stack arms.

The main grid is unguided. The historical default best-of-N configuration can request guidance three, but the available study decoder made the full guided grid impractical. A smaller guided follow-up covers four classes across the detector's recall range. The principal crossed result must therefore be described as an unguided candidate-selection finding, not a measured gain for every guided deployment.

For set j, let $H(y)$ be the chosen detector's hit indicator. The directly relevant statistic is

$$L_j = H\left(\arg\max_{y \in \mathcal{Y}_j} R(y)\right) - \frac{1}{|\mathcal{Y}_j|} \sum_{y \in \mathcal{Y}_j} H(y). \tag{7}$$

It measures delivered hits above random selection within the same available takes. This is defined even for all-hit and all-miss sets. A ranking AUC is only defined for mixed sets and must state how many such sets exist. Correlations are computed after within-set centring; pooling raw scores across classes would mainly reflect differences in class difficulty.

L.2 ORIGINAL REWARD AND ITS CARRIER-VOICE LIMITATION

The old ranker delivers a mean hit rate 0.1531 versus 0.1183 for random selection, a difference of 0.03479 with recorded interval $[0.02234, 0.04725]$ and $t = 5.47$. But the 1,164 sets with the carrier-voice adapter have lift only 0.0104 ($t = 1.51$); the 776 without it have lift 0.0714 ($t = 5.97$). The pooled result is consequently not a reliable estimate of the improvement in the voice-adapter configuration used in deployment.

Within candidate sets, blend correlates negatively with burst hits (approximately -0.0839), genuineness is near zero (-0.0085), prompt similarity is positive ($+0.1482$), and the WER factor weakly positive ($+0.0388$). A local burst term reaches approximately $+0.3053$. These are useful objective diagnostics, not human validity coefficients.

L.3 A SOFT EVENT-SENSITIVE TERM

For each requested burst, the studied soft term combines strict-class evidence s_{strict}, family evidence s_{fam}, presence s_{pres}, and a family-agreement indicator a:

$$u = s_{\text{strict}} + \tfrac{1}{2}(s_{\text{fam}} - s_{\text{strict}}) + \tfrac{1}{4} s_{\text{pres}} + \tfrac{1}{2} a. \tag{8}$$

The term is averaged over requested events, normalised within the candidate set, and added with weight λ:

$$R_\lambda(y) = \left[\widetilde{G} + \widetilde{B} + 2\widetilde{C} + \lambda \widetilde{u}\right] q(W). \tag{9}$$

Detector recall is deliberately not inserted as a per-class correction in this ranking. The requested class is fixed within a candidate set, so a constant rescaling often cannot alter its within-set normalised order; it would create the appearance of calibration without validating it.

Local scoring uses windows around the expected onset, rather than only a whole-clip embedding. The current implementation scans one-second windows at 0.25-second hops around a nominal interval from one second before to two seconds after the requested onset, with boundary handling in the code. Windows can capture speech context or neighbouring events, so they require temporal audits.

If no burst is requested, the added term is exactly zero. Merely passing a constant zero vector to min–max normalisation is wrong here because the constant convention returns 0.5. Adding that constant before the candidate-specific WER gate can change the winner. Missing or unavailable burst heads must be logged; they cannot silently count as successful event verification.

L.4 Crossed encoders and adverse metrics

Ranking with large-v2 and measuring with commercial gives a $\lambda = 2$ hit improvement of 0.0454 ($t = 7.60$), WER increase 0.0040, and prompt-similarity decrease 0.0055. Reversing the encoder roles gives a smaller hit gain of 0.0314. The complete matrix is shown in Table L.1; same-encoder cells are included as diagnostics, not independent tests.

The $\lambda = 3$ local rule reaches 0.2082 under the large-v2 diagonal diagnostic, versus about 0.1907 for whole-clip windows. Hard tiering buys little extra hit rate but increases WER by roughly 0.046, much more than the soft alternatives. Thus prioritising a detected event above every other consideration can select an unintelligible take.

The detector-defined oracle reaches 0.3789 among the same stored candidates. Relative to random selection, the available headroom is approximately 0.2606; relative to the old ranker it is approximately 0.2258. This oracle is neither human listening nor a new generative system. It shows that better ranking could exploit more of the existing pool under that detector, while the pool itself still misses many requests.

Study versus deployed revision. The original study submitted a proposal without modifying the server. The subsequently inspected server commit does implement the soft burst term at weight two, using commercial heads by default to reuse the existing embedding. Therefore the strongest large-ranker crossed gain of 0.0454 is not the exact configuration's measured effect. The reverse-direction result is more relevant to that encoder choice, and any live result must still state guidance, stack, and detector availability. This version distinction corrects the older report's implementation status.

L.5 Compute and the best-of-N trade-off

An eight-candidate inference diagnostic takes approximately 14.9 seconds without guidance and 29.2 seconds with guidance in the cited hardware configuration. These are complete-selection latencies, not first-audio streaming latency and not hardware-independent performance guarantees. Best-of-N must wait for candidates and inspection; a batched implementation can reduce wall time but does not remove the extra compute.

Candidate-count comparisons must preserve all attempted generations, including failures and internal retries. A fair system study should include equal-N and equal-time or equal-cost views. Increasing N cannot compensate for an absent skill indefinitely, and choosing on a noisy proxy can reward evaluator blind spots. The proposed feedback loop therefore requires periodic independent human audits and source-diversity constraints.

Table L.1: All crossed comparisons against the recorded ranker. Paired differences are comparable within an evaluation encoder. Same-encoder rows are circular diagnostics, not independent validation.

Rank / measure encoder	Rule	Δ hit	t
large-v2 / large-v2	lam1_soft	0.0299	5.97
large-v2 / large-v2	lam2_soft	0.0412	6.56
large-v2 / large-v2	lam3_soft	0.0552	7.74
large-v2 / large-v2	lam3_tier	0.0577	8.70
large-v2 / large-v2	lam3_tierfb	0.0603	8.92
large-v2 / large-v2	lam5_soft	0.0531	6.95
large-v2 / commercial	lam1_soft	0.0278	5.80
large-v2 / commercial	lam2_soft	0.0454	7.60
large-v2 / commercial	lam3_soft	0.0572	8.70
large-v2 / commercial	lam3_tier	0.0588	9.33
large-v2 / commercial	lam3_tierfb	0.0613	9.60
large-v2 / commercial	lam5_soft	0.0639	8.99
commercial / large-v2	lam1_soft	0.0227	4.42
commercial / large-v2	lam2_soft	0.0314	5.10
commercial / large-v2	lam3_soft	0.0387	5.75
commercial / large-v2	lam3_tier	0.0366	5.60

Continued on next page

Rank / measure encoder	Rule	Δ hit	t
commercial / large-v2	lam3_tierfb	0.0371	5.60
commercial / large-v2	lam5_soft	0.0407	5.56
commercial / commercial	lam1_soft	0.0335	6.35
commercial / commercial	lam2_soft	0.0557	8.76
commercial / commercial	lam3_soft	0.0701	9.80
commercial / commercial	lam3_tier	0.0660	9.95
commercial / commercial	lam3_tierfb	0.0691	10.25
commercial / commercial	lam5_soft	0.0737	9.59

Evidence: `reports/vb_opt.md`; exported `ranking_cross.json`, `ranking_oracle.json`, and `ranking_term_correlations.json`; historical mirror commit `0558bf6` and current server `bestofn.py/config.py`.

M Reproducibility boundaries and a future learning cycle

M.1 Resource map

The resource map below lists the anonymized copies of the code, data, and models behind the reported system. Links point to review snapshots; names that could identify the authors are withheld until the end of the review period.

Code.

- Director, inference server, WikiSkills, and Arena prompts: https://anonymous.4open.science/r/I271.

Perception and generation.

- Humaneness Voice SFT3 instrument weights, CC-BY-4.0: https://anonymous-hf.com/a/xm66jyqcvb7h/.
- Genuineness predictor, bundling the frozen VoiceCLAP-Commercial encoder: https://anonymous-hf.com/a/km3933c74q97/.
- Vocal-burst blend predictor: https://anonymous-hf.com/a/xdsygvc9zcfb/.
- Paired stimuli for the human check of both predictors (Appendix A): https://anonymous-hf.com/a/lv57b0agr3v4/.

Annotated expressive speech.

- Open annotated branch: https://anonymous-hf.com/a/isycxvdovbbn/.
- Gated research branch: https://anonymous-hf.com/a/dwtzt6j67qbj/.
- Synthetic voice-profile performances: https://anonymous-hf.com/a/igsoi4jv3gj6/.
- Annotated scene-pair release: https://anonymous-hf.com/a/r2i52q9o0ma3/.

Balanced German–English speech.

- Scaling ladder: https://anonymous-hf.com/a/yfp5xjgu3n4z/.

Voice-acting training and preference views.

- Voice-profile timed scripts (SFT): https://anonymous-hf.com/a/pzmwz9xhxzle/.
- Real-speech timed scripts (SFT, DE/EN): https://anonymous-hf.com/a/8m5o9wixb0zw/.
- Voice-profile preference pairs (DPO): https://anonymous-hf.com/a/062dliqpuia4/.
- Real-speech preference pairs (DPO, DE/EN): https://anonymous-hf.com/a/x33og8mnywwo/.
- Voice-profile CFG and phase-two preference pairs: https://anonymous-hf.com/a/4y6vwl5f1wc8/.
- Synthetic edge-case ranked takes: https://anonymous-hf.com/a/wd4co3jb9lpd/.
- Synthetic ranked reinterpretations: https://anonymous-hf.com/a/oy6psye4ke5o/.
- Acting Challenge Dataset (public briefs, not Arena items): https://anonymous-hf.com/a/d3hmukv4v2lc/.

Taxonomy. The VoiceNet dimensions are listed in Table H.2; the full anchor texts will be released with the taxonomy paper.

The Arena demonstration endpoint is temporary, and its held-out task JSON and spoken scripts are not released. Large weights and audio remain in the repositories listed above. The aggregate numeric evidence and figure scripts behind this paper will be released with the final version.

M.2 INTERPRETATION SAFEGUARDS

Table M.1: Distinctions that affect interpretation of the reported evidence.

Potentially misleading reading	Corrected interpretation
All published clips trained SFT3	The large release has 137.3M records; SFT3 uses a selected 398,282-row training view.
Every annotation is validated	Most labels are model-derived. A preliminary listener snapshot gives 64% agreement for genuineness and 46% for blend on small, rater-concentrated samples; neither is an established validity result.
Emotion result is Pearson $r = .47$	The primary layer result is macro-$R^2 \approx .473$ against model-labelled emotions.
Layer averaging means one cross-layer mean	States are averaged over time within each layer; selected layers are concatenated in W2.
High burst presence AUC means localisation	Most annotated bursts are outside the feature window; context supports classification without precise localisation.
Detector recall is a hard generator ceiling	Sensitivity, false positives, and domain shift jointly determine observed hit rates.
The director writes exact durations	It writes semantic events; the server compiles numerical timing.
The director also writes a reply field	Current code derives the spoken reply from the script.
The ranker uses raw unclipped $1 - W$	Both inspected code revisions clip WER in the gate; report shorthand was incomplete.
Burst-aware ranking is only a proposal	It was a proposal in the study and is implemented in the current server; the default encoder differs from the strongest study arm.
Higher blend always means a better take	Blend can rise with silence or severe intelligibility failure; event presence and content need separate checks.
CFG-DPO audio was generated by CFG	The name refers to contrastive pair construction.
Symmetric DPO pairs imply zero gradient	That does not follow generally for a conditional preference objective.
GRPO is intrinsically unsuitable	The tested recipes did not improve the held-out objective; no universal impossibility is established.
More pauses improve genuineness	Local pauses are useful controls; additional pause pressure failed or harmed model-judged comparisons.
W1 peak is an independent test result	Peak locations were selected descriptively on test; W2 selections use validation.
Arena is an untouched held-out benchmark	Twenty tasks were used for director development; holdout exclusions must be frozen.
Open access means unrestricted rights	Per-source licensing and research-only conditions remain binding.

M.3 EVIDENCE AND RELEASE BOUNDARIES

The manuscript distinguishes measured results from proposed training cycles and historical experiments whose full run artifacts are unavailable. In particular, first-round Arena ratings rest on few, concentrated listeners, and the annotation inventory's repository-view hours are not unique licensed audio hours. The resource map supports traceability while the exact held-out task scripts remain private.

M.4 A REPRODUCIBLE NEXT TRAINING CYCLE

The proposed cycle can be made testable without claiming autonomous self-improvement:

1. Freeze a scene distribution, a consented voice/reference policy, and generation budgets. Keep a held-out set outside prompt tuning, candidate mining, scorer training, and adapter selection.
2. Generate candidates with logged scripts, seeds, references, model revisions, tool strengths, latency, and failures. Preserve diversity rather than only extreme reward winners.
3. Apply separate content, identity, event-presence, integration, genuineness, and prompt-fit checks. Reserve randomly sampled and high-disagreement examples for human audit.
4. Construct supervised winners and within-scene preference contrasts with transparent selection probabilities and source lineage. Do not erase difficult or minority styles through a single scalar threshold.
5. Fine-tune either the instrument or one specialist adapter, holding other components fixed. Compare against equal-data and equal-compute baselines, including unfiltered and human-selected subsets where feasible.
6. Re-evaluate on untouched scenes and voices using independent listeners and frozen evaluators. Add a WikiSkills entry only when its benefit and adverse operating range are documented.

The value of this cycle depends on the ears. Better genuineness, blend, and delivery models could make candidate mining more useful, but repeated selection on a weak proxy can instead narrow style diversity or exploit a detector. Independent human tests and intentionally retained difficult cases are therefore part of the learning mechanism, not an optional final decoration.

M.5 Ethical and practical limits

Voice-conditioned generation can enable impersonation and non-consensual imitation. References should be used with appropriate authority, and distributed examples should disclose synthesis. Emotion and genuineness scores describe perceived vocal performance; they should not be used to infer a person's truthfulness, psychological state, or suitability for employment. Accent and style heads proposed as future work require bias and consent evaluation before sensitive deployment.

Releasing data lineage and limits is central to this project. It makes downstream reuse easier while allowing researchers to exclude incompatible sources, replace weak annotators, and test alternative definitions of expressive quality. A transparent partial reconstruction is more useful than an unsupported claim that every historical experiment is reproducible end to end.