DOES BACKGROUND MUSIC MATTER TO SPEECH IN LANGUAGE MODELS

Mingzhe Du
National University of Singapore
mingzhe@nus.edu.sg

Moming Duan
East China Normal University
duanmoming@gmail.com

1 MOTIVATION

People listen to music while solving difficult problems, doing routine work, or looking for ideas. Studies of human performance report both benefits and costs under different musical conditions (Ritter & Ferguson, 2017; Threadgold et al., 2019). This motivates a small exploratory question: *when a pre-trained language model receives a spoken question, does adding background music change its response?*

We hold the spoken question fixed and vary its musical background. Our interest is the influence itself: its direction, its size, and which answers change. A model may gain or lose accuracy, or answer different questions correctly while keeping almost the same average score. The human analogy motivates the experiment; whether any effect comes from speech perception, later processing, or both remains open.

2 FINDINGS

A flat average can hide substantial changes. Voxtral Mini on far context has 16.68% paired score flips but a net change of only -0.08 pp; Step-Audio on IFEval has 16.43% flips and a $+0.06$ pp net change. Gains and losses on different questions can cancel in the average.

The same musical backgrounds can help one model and hurt another. On GSM-Symbolic base, music changes Voxtral Mini's score by $+7.83$ pp and Qwen3-Omni's by -3.01 pp. Every recording has a positive estimate for the former and a negative estimate for the latter, despite identical audio. These configuration-specific outcomes do not establish a general mathematical benefit or model-size rule.

More complex mathematics does not imply a larger benefit. Across base/P1/P2, Voxtral Mini's effects shrink from $+7.83$ to $+6.30$ to $+2.63$ pp; Qwen3's are non-monotonic (-3.01, -5.17, -1.71 pp). Numerical content and spoken length also change, so difficulty is not isolated.

Context structure matters, with no uniform distance trend. Qwen2.5-Omni loses less in far than near (-1.54 vs. -3.82 pp); Qwen3 shows the opposite ordering (-4.01 vs. $+0.52$ pp). The response depends on the model and the position of facts in spoken input, not simply on distance.

Weak genre tendencies coexist with model and task dependence. Chinese traditional, hip-hop, and ambient music exceed the music mean in seven of eight models after averaging over settings; rock and metal do so in only one (Table 1). The overall genre spread is small (0.46 pp). White noise averages 0.73 pp below music. These are descriptive associations among five recordings per genre, not universal preferences or evidence of gains over clean speech. Appendix A defines the averaging and baseline checks.

What we take from this. Music changes scores and which answers succeed, with model- and task-dependent directions and weak aggregate genre tendencies. These observations concern the complete speech-to-answer pipeline, not human-like focus or emotion. Shared task items, curated recordings, output budgets, and grading rules limit interpretation (Appendices B and A).

Background	Relative score percentage points	Models above music mean	Settings above music mean
Chinese traditional	+0.21	**7** / 8	**8** / 10
Hip-hop	+0.20	**7** / 8	**8** / 10
Ambient	+0.16	**7** / 8	**9** / 10
Pop	+0.11	**5** / 8	**6** / 10
Cinematic / orchestral	+0.03	**4** / 8	**7** / 10
Jazz	−0.01	**5** / 8	**4** / 10
Folk / country	−0.02	**4** / 8	**6** / 10
Electronic dance	−0.07	**2** / 8	**3** / 10
Classical	−0.14	**2** / 8	**4** / 10
Rock	−0.23	**1** / 8	**1** / 10
Metal	−0.25	**1** / 8	**0** / 10
White noise (control)	−0.73	**1** / 8	**2** / 10

Table 1: Background-type differences from the 55-track music mean, equally averaged over eight models and ten settings. Green/red denotes above/below the music mean, not improvement/decline relative to clean speech. Model counts average over settings; setting counts average over models. Counts indicate direction, not significance. White noise is a control excluded from the reference mean.

3 EXPERIMENTS

Each model receives a spoken question, either alone or mixed with one of 55 instrumental recordings from 11 genres. The music is mixed at +10 dB SNR. A white-noise condition provides an acoustic reference. Models generate text with greedy decoding. We compare each music condition with its matched clean-speech condition using task accuracy or IFEval Strict compliance.

We evaluate eight models: Phi-4 Multimodal, Voxtral Mini, Qwen3-Omni, Qwen2.5-Omni, Voxtral Small, Step-Audio-R1.1, Nemotron, and MiniCPM-o (Elfsong, 2026; Mistral AI, 2025; StepFun, 2026; NVIDIA, 2026; OpenBMB, 2026). The experiments provide three views of the same question:

Figure	Evaluation settings	What we examine
1	GSM8K, BBH, MMLU, IFEval	Variation across tasks and models
2	GSM-Symbolic base, P1, P2	Variation across mathematical-complexity settings
3	bAbI short, near, far	Added irrelevant text and the position of relevant facts

Original tasks and mathematical variants use 500 items per setting; context variants use 300 paired problems (Mirzadeh et al., 2024; Kuratov et al., 2024). Near and far place the same irrelevant sentences before or after the intact facts, keeping the question last. All eight models receive identical regenerated audio for these six added settings. On the four original tasks, Figure 1 reports runs that differ in speech generation, output budgets, and IFEval scoring version; Appendix D records these per-model settings.

Appendix P gives a plain-language guide to each benchmark and every category used here. Every matrix shows all 55 recordings. Positive values mean higher scores than clean speech; negative values mean lower scores. Values are percentage-point changes, with one shared color scale. The last two rows summarize music and white noise. Sample selection, configurations, grading, and derivations are in the appendices.

Mathematics and general reasoning

1 / 2

55 recordings · 8 models · 500 items / setting · +10 dB SNR

Change from matched clean speech (percentage points)

Genre		Recording	Math word problems — GSM8K								General reasoning — BBH							
			Phi-4	Vox3	Q3	Q2.5 H	Vox24	Step	NV	Mini	Phi-4	Vox3	Q3	Q2.5 H	Vox24	Step	NV	Mini
Folk / country	01	glinguinette	−0.8	−0.4	+0.6	+1.2	−7.8	−0.6	+2.0	+2.2	+1.6	−1.6	−2.0	−1.4	−1.6	−0.2	+1.8	−0.6
	02	01 Ouverture	−0.6	+1.0	+0.4	+1.0	−8.4	−0.4	+0.4	+1.4	+0.6	−1.0	−2.2	−0.2	−0.6	+0.2	+1.8	−1.6
	03	Y'vonim	+0.2	+1.4	+0.6	−0.2	−4.2	+0.2	+0.8	+2.2	+1.0	−4.4	−2.4	−0.8	−1.2	+1.0	+0.6	−3.8
	04	Folk omelette	−0.4	+0.6	−0.4	+2.2	−10.2	−0.4	+0.2	+1.6	−0.2	−4.0	−3.4	−1.0	−0.6	−0.6	+1.2	−5.2
	05	Finding Home	+0.6	+1.2	+0.6	+1.6	−12.2	+0.6	+0.4	+0.4	0.0	−5.8	−3.4	−0.2	−0.8	+2.0	+2.0	−2.4
Ambient	06	Jotun	−0.4	+1.2	+0.2	+2.2	−5.0	−0.8	+1.2	+2.0	+1.2	−5.4	0.0	0.0	−1.8	−0.2	+1.8	−2.0
	07	prolog03	+0.4	+1.4	+0.6	+2.0	−2.6	−0.2	+0.6	+0.8	−0.6	−2.4	−2.8	−1.0	−1.6	0.0	+1.6	−2.2
	08	In Memories (Colors)	+0.6	+1.0	+1.2	+1.0	−8.6	−1.2	+0.6	+0.8	+0.8	−3.2	−2.8	−0.6	−1.4	+1.0	+1.4	−3.6
	09	Spartial	+0.6	+1.4	+0.6	+0.8	−10.8	−0.4	+0.2	+1.2	+0.2	−3.0	−0.4	−1.0	−1.2	−0.8	+1.8	−3.2
	10	Sound is I	−0.2	+1.0	+0.2	+0.2	−5.0	−1.2	+0.6	+1.2	0.0	−1.6	−0.4	+0.2	−3.2	+2.6	+1.8	−2.6
Chinese traditional	11	Liu Shui	−0.8	+0.8	0.0	+1.0	−9.6	−0.4	+0.8	+2.0	+0.2	−4.2	+0.2	−0.2	−1.0	−2.2	+2.2	−0.6
	12	Shi Mian Mai Fu	−0.8	+1.2	+0.2	+1.2	−9.6	−0.8	−0.6	+1.2	−0.8	−2.0	−1.8	−0.6	−0.6	−1.0	+0.6	−4.4
	13	Chun Dao Yi He	−0.6	+1.6	+0.4	+2.2	−7.2	−0.8	+0.6	+0.4	−0.2	−3.8	−0.8	−1.4	−1.0	−0.4	+0.8	−4.4
	14	Zhan Tai Feng	+0.2	+0.8	−1.0	+0.6	−8.6	−1.0	−0.8	+0.8	+0.2	−3.0	+0.2	+0.6	−1.4	−1.8	−0.2	−2.6
	15	Erquan Yingyue	0.0	+0.4	−0.2	+1.8	−10.0	−1.0	+0.4	+2.2	+0.4	−2.6	−1.2	−0.2	−0.2	−0.8	+3.0	−2.4
Cinematic orchestral	16	Sunrise	+0.2	+0.6	−0.8	+2.6	−5.0	−0.4	−0.2	+0.6	−0.4	−1.8	−2.0	−1.0	+0.6	+1.4	+0.6	−3.4
	17	Archer Revenge	−0.4	+0.6	−1.2	+1.8	−5.4	−0.4	−0.4	+1.6	+0.6	−3.0	−2.2	−0.8	−3.2	−0.4	+1.0	−3.6
	18	Dawn Of Heroes	+0.2	+1.2	−1.4	+0.8	−4.6	0.0	+0.2	−0.2	0.0	−4.4	−2.0	−1.0	−1.4	0.0	+1.0	−4.8
	19	Fairy Tale for the Lonely	+0.2	+1.0	0.0	+0.8	−7.8	−0.4	0.0	+1.0	−0.2	−2.4	−3.6	−1.2	−2.0	+0.6	+3.0	−3.8
	20	Prepare to Be Boarded, Comrades	0.0	+1.2	0.0	+2.2	−5.6	−0.4	+0.8	+2.6	0.0	0.0	−2.0	−0.6	−1.0	−1.0	+0.8	−4.4
Classical	21	Bach · Brandenburg No. 3	+0.6	+0.2	−0.8	+0.4	−9.2	−0.2	−0.2	+0.8	+1.4	−4.6	−2.6	−2.0	−1.2	+0.4	+1.2	−3.2
	22	Beethoven · Symphony No. 5	+1.4	+1.0	+0.6	+2.2	−6.8	−1.0	+1.0	+3.4	+0.4	−5.0	−1.4	−2.4	0.0	−1.2	+0.6	−3.4
	23	Chopin · Scherzo, Op. 31	−0.2	+1.0	+0.4	+1.2	−8.4	−0.4	+1.2	+2.6	−0.6	−1.8	−2.0	−1.6	−0.4	−0.4	0.0	−1.0
	24	Vivaldi · Two Violins, Op. 3/11	+0.8	+0.8	−1.2	+1.2	−4.8	−0.2	+1.6	+0.4	+0.8	−3.6	−1.4	−1.0	−1.6	+2.8	+0.6	−3.2
	25	Dvořák · Serenade, Op. 22	+0.8	+0.6	+0.4	+0.4	−4.8	+0.4	+0.8	+0.6	0.0	−3.8	−3.2	−1.6	−2.2	−0.8	−1.0	−3.4
Electronic dance	26	Aziba	−0.4	0.0	−0.2	+2.0	−7.0	+0.2	+1.0	+2.2	+0.8	−5.2	−3.6	−0.8	−1.0	+2.6	+1.0	−4.2
	27	Child's Smile	+0.8	0.0	−0.2	+1.6	−10.2	+0.2	−0.6	+1.2	−0.6	−4.6	−3.6	−1.2	−3.0	+2.0	+1.4	−5.4
	28	Protected	+0.6	+0.8	+0.4	+3.0	−10.6	+0.8	+1.0	0.0	+1.2	−3.4	−2.4	−1.0	−1.6	−0.6	+0.4	−2.2
	29	detroit	+0.8	−0.6	−1.4	+2.2	−9.0	−0.6	+0.4	+1.6	+0.4	−3.6	−2.6	−1.6	−1.2	+0.6	+2.8	−5.0
	30	E-Samba	+0.2	+1.0	+0.4	+1.4	−14.6	0.0	+1.6	+3.0	+0.8	−3.6	−1.2	−0.4	−1.6	+0.6	+2.4	−4.6
Hip-hop	31	b. street nois	+0.8	+0.4	+0.2	+0.6	−6.6	0.0	+1.0	+1.2	+0.8	−3.0	−1.4	+0.2	−2.2	−0.4	+1.4	−3.6
	32	04	−0.6	+0.6	−0.8	+1.6	−3.6	−0.2	+1.8	+0.4	+0.4	−4.2	−1.8	−0.8	−1.0	−0.8	+1.0	−3.0
	33	Aries4Rce_-_Boss	+1.0	+0.6	+0.6	+1.6	−4.8	0.0	+0.2	−0.2	+0.2	−3.0	−2.0	0.0	−3.4	−1.0	+1.8	−0.8
	34	Energy Bounce 2	+0.4	0.0	−0.2	+1.8	−5.0	−0.8	+0.8	−0.2	−0.4	−5.0	−1.0	−0.6	−3.4	+0.2	+0.4	−3.0
	35	Zombie Klub	+0.4	+1.6	+0.8	+1.4	−11.6	−0.4	+1.4	+1.4	0.0	−2.8	−1.0	+0.4	−1.2	−0.4	+0.8	−1.4
Metal	36	iron knight 5	−0.2	−0.2	−1.2	+1.4	−8.0	+1.2	0.0	+1.2	−0.2	−4.4	−1.2	+0.4	−1.4	−1.4	+0.4	−3.6
	37	Ready to Fight	+0.2	+1.6	+0.2	+1.6	−5.2	+1.6	−1.2	+1.4	−0.6	−4.4	−2.6	−1.0	−1.0	−0.4	+1.6	−2.8
	38	Good Morning	−0.4	+1.6	−0.2	+2.0	−13.8	+1.0	+0.8	+1.0	0.0	−4.4	+0.4	−2.0	−0.8	−2.0	+1.6	−2.6
	39	Way of Warrior	0.0	+1.2	−1.2	+1.4	−2.8	+1.2	+0.4	+1.4	−0.4	−2.8	−3.2	−0.4	−2.4	−1.6	−0.4	−3.6
	40	Black Sky	+0.4	+1.4	+0.4	+1.0	−8.2	+1.0	+0.6	+1.0	+0.4	−3.4	−1.6	+1.2	−2.0	−0.2	+1.6	−2.4
Pop	41	Very Upbeat	0.0	+1.0	−1.6	+2.0	−9.8	+0.2	−0.4	+2.2	−0.4	−4.2	−2.0	−1.0	−2.2	0.0	+0.6	−4.4
	42	On My Day Off	+0.4	+1.0	+0.8	+1.2	−5.8	−1.0	+0.8	+0.2	−0.4	−3.0	−0.2	−0.2	+0.2	−0.4	+0.6	−3.8
	43	Mr. Bleach	−0.8	+2.6	−0.2	+1.4	−8.8	−0.4	+0.4	+2.6	−1.2	−4.8	−3.0	0.0	−2.6	+1.2	−0.2	−3.0
	44	Toys	+0.2	+0.4	+1.4	+2.0	−4.6	−0.4	−0.6	−0.6	+0.2	−3.4	−3.0	−1.2	−3.0	+0.6	+1.6	−1.4
	45	Noris	−0.6	0.0	+0.2	+2.0	−12.4	−0.6	+0.2	+0.8	+0.4	−3.8	−2.4	−1.0	−2.4	+0.2	+1.6	−3.4
Rock	46	Powiedzim wszystkim nie...	+0.2	+1.8	−1.4	+1.8	−7.0	+0.2	+0.4	+0.2	+0.8	−4.0	−3.0	−1.0	−4.2	+0.2	+2.2	−3.6
	47	happy hour	+0.2	+1.0	+0.4	+1.4	−9.4	+1.0	0.0	+0.6	−0.4	−3.6	+0.4	−0.2	−2.2	−0.4	+1.0	−4.6
	48	Last Launch Part I	−0.2	+1.8	+0.4	+2.2	−1.6	+0.4	0.0	+2.4	+0.4	−3.8	−2.6	−0.6	−3.4	+0.6	+2.4	−3.0
	49	Banned	+0.8	+0.8	+0.6	+1.8	−5.2	−0.2	−0.6	+1.2	0.0	−4.4	−3.4	−1.2	−1.4	+1.8	+2.4	−3.4
	50	We Don't Piss In Your Ashtrays	−0.2	+1.0	−0.6	+0.8	−6.6	−0.6	+1.2	+0.8	−1.2	−2.8	−3.4	−1.8	−1.0	+2.0	+2.4	−2.0
Jazz	51	N400	+0.2	+1.4	−0.2	+2.4	−13.6	+0.2	+1.4	+0.2	+0.4	−4.8	−1.2	−0.6	−0.6	−0.2	+1.4	−2.4
	52	Der lachende Vulkanier...	0.0	+0.4	−0.6	+1.6	−13.6	−0.4	+1.6	−0.2	+1.0	−4.0	−2.6	−1.6	−1.0	−3.6	+1.0	−2.6
	53	faaå	+1.0	+1.0	0.0	+0.8	−6.2	0.0	+1.6	+0.4	+1.2	−4.4	−1.2	−2.0	−2.4	−0.4	+1.4	−1.4
	54	My name is Nova, Bossa Nova	−0.2	+0.2	+0.6	+1.6	−11.6	+0.4	+1.6	0.0	−0.2	−1.2	−1.4	−1.4	−2.2	−0.4	+0.8	−4.4
	55	La toupie	+0.2	+0.2	−0.8	+1.4	−10.6	−0.8	+0.8	−0.2	0.0	−5.0	−3.0	−2.4	+0.6	+1.4	−1.0	−3.0
	All music mean		**+0.1**	**+0.9**	**−0.1**	**+1.5**	**−7.8**	**−0.1**	**+0.5**	**+1.1**	**+0.2**	**−3.5**	**−1.9**	**−0.8**	**−1.6**	**0.0**	**+1.2**	**−3.1**
	White-noise control		**+1.4**	**+1.3**	**−1.8**	**+1.2**	**−6.1**	**−0.3**	**−1.3**	**+2.0**	**−1.1**	**−6.1**	**−6.6**	**−2.2**	**+0.9**	**0.0**	**+1.7**	**−3.8**

Lower score ← → Higher score

-16 -8 0 +8 +16

Phi-4: Phi-4 Multimodal · Vox3: Voxtral Mini · Q3: Qwen3-Omni · Q2.5: Qwen2.5-Omni

Vox24: Voxtral Small · Step: Step-Audio-R1.1 · NV: Nemotron · Mini: MiniCPM-o 4.5

H: historical run. Recording titles shortened; full catalog in appendix. Colors do not denote significance.

Figure 1: Task categories across eight models and 55 recordings. GSM8K tests solving arithmetic word problems; BBH tests logical, causal, spatial, and language reasoning across ten selected subtasks. MMLU and IFEval follow. Appendix P explains each evaluated category. Values are changes from matched clean speech (pp). Speech generation, output budgets, and the IFEval scoring version differ between runs (Appendix D).

Knowledge and instruction following

2 / 2

55 recordings · 8 models · 500 items / setting · +10 dB SNR

Change from matched clean speech (percentage points)

Genre		Recording	Knowledge questions (MMLU)								Instruction following (IFEval)							
			Phi-4	Vox3	Q3	Q2.5 H	Vox24	Step	NV	Mini	Phi-4	Vox3	Q3	Q2.5 H	Vox24	Step	NV	Mini
Folk / country	01	glinguinette	+1.0	+0.2	−1.4	−1.0	0.0	−1.4	+1.6	0.0	+0.2	+0.4	+0.6	−0.8	+0.4	+2.6	−1.6	−1.4
	02	01 Ouverture	+1.0	−1.0	−1.2	−1.2	−0.6	−0.8	+1.4	+0.6	−0.4	−0.6	0.0	−0.2	+1.6	+2.0	−1.8	−2.2
	03	Y'vonim	+1.6	−1.6	−0.6	−2.0	+0.6	+0.4	+2.2	+1.0	−3.0	−0.2	−0.6	−1.2	+2.6	−1.0	−2.4	+0.4
	04	Folk omelette	−0.2	−0.2	−0.8	−0.4	+0.2	−1.0	+1.2	−1.2	+1.2	+1.4	+1.0	−0.2	+2.6	+0.8	−2.0	−1.6
	05	Finding Home	0.0	−1.2	−1.8	−1.2	+0.4	−2.2	+3.4	+0.2	+0.6	−0.2	−1.4	−0.6	+1.2	+0.4	−2.6	−0.2
Ambient	06	Jotun	−1.2	−0.6	0.0	−1.4	+0.2	+0.2	+1.6	+0.6	−0.8	+1.0	+0.8	+1.0	−1.0	−2.0	−3.2	+1.0
	07	prolog03	+1.6	+0.4	−1.0	−0.2	+1.0	−1.4	+2.2	+1.2	−2.0	+1.0	+2.2	0.0	+1.8	+2.0	+1.0	−0.2
	08	In Memories (Colors)	−0.2	−1.8	−0.6	−0.2	+0.6	−0.4	+2.4	+0.4	−0.6	+2.0	−1.4	−1.0	+2.8	+0.4	−3.8	+0.2
	09	Spartial	+0.2	0.0	−1.4	−0.6	0.0	−0.8	+3.0	+0.6	−0.4	+0.6	+0.2	−0.6	+2.4	−0.6	−0.4	−2.8
	10	Sound is I	+0.8	+1.0	−1.4	−0.6	+1.2	−0.4	+1.4	+0.4	−2.6	+1.4	−0.6	−1.0	0.0	−1.6	−2.6	−1.0
Chinese traditional	11	Liu Shui	+0.4	−0.4	−1.2	−1.2	−0.2	+0.4	+2.8	+0.6	+1.4	+1.4	+1.0	−0.2	+1.8	−3.8	0.0	+0.6
	12	Shi Mian Mai Fu	+1.0	+0.6	−2.0	−0.4	+0.8	−1.0	+1.0	−1.8	−2.0	+0.4	+0.8	−2.0	+0.6	−1.4	+0.4	−0.6
	13	Chun Dao Yi He	+0.8	−0.4	−0.6	−1.2	−0.2	−0.8	+2.6	+0.2	−0.6	+2.8	−0.2	−0.8	+2.2	−0.4	−2.2	0.0
	14	Zhan Tai Feng	+2.2	−0.2	−0.6	−1.2	−0.6	−1.0	+2.4	+0.6	−1.0	−1.4	+2.0	+0.6	+4.4	−4.8	0.0	+0.4
	15	Erquan Yingyue	+0.6	−0.8	−0.8	−1.4	+1.8	−0.2	+1.6	−0.8	−1.6	+3.0	−1.2	−0.8	+2.8	−1.2	+0.6	+1.0
Cinematic orchestral	16	Sunrise	+0.2	0.0	−0.2	−0.2	+1.2	−1.4	+2.2	−0.6	−0.8	0.0	−0.8	−1.8	+2.6	+1.0	−4.2	+1.4
	17	Archer Revenge	+0.6	−2.0	−1.0	−3.6	−0.8	−0.8	+3.0	−1.4	−1.4	0.0	+0.8	−3.0	+3.4	0.0	−1.6	−1.8
	18	Dawn Of Heroes	−0.2	−0.8	−1.4	−2.0	−0.2	+0.2	+1.8	−1.4	−2.0	+1.2	0.0	−1.0	+1.0	−2.0	−1.6	−1.0
	19	Fairy Tale for the Lonely	+0.4	−0.6	−1.2	−1.8	+1.2	−0.2	+2.0	+0.4	−0.4	−1.2	+0.4	−0.8	+2.8	+2.8	−1.4	−0.2
	20	Prepare to Be Boarded, Comrades	−0.4	−1.0	−1.0	−1.0	+0.2	−1.6	+2.8	+0.2	−1.0	+2.4	0.0	−0.4	+1.8	−0.6	−1.8	−0.8
Classical	21	Bach · Brandenburg No. 3	−0.4	−1.4	−1.0	−0.6	+0.6	−2.6	+2.0	−0.6	−2.8	+3.0	−0.2	−1.6	+0.8	−0.8	−2.6	0.0
	22	Beethoven · Symphony No. 5	+1.2	0.0	−1.4	−0.8	0.0	−1.4	+2.2	0.0	−1.4	−1.2	−0.6	+1.2	+1.2	−1.2	−3.6	−1.4
	23	Chopin · Scherzo, Op. 31	−1.6	−1.6	−1.0	−1.4	+1.4	−1.2	+1.6	+2.6	−0.4	−1.0	−0.4	−2.2	+2.4	+1.2	0.0	−2.0
	24	Vivaldi · Two Violins, Op. 3/11	+0.4	−0.4	−1.4	−1.2	−1.0	−1.4	+1.6	−1.8	−2.0	+1.2	+0.2	+0.6	+1.2	−1.6	−2.2	−1.4
	25	Dvořák · Serenade, Op. 22	−0.8	−1.0	−1.6	−1.6	−0.6	−1.6	+1.2	−1.0	−1.0	+1.2	0.0	+1.0	+1.6	−1.6	−0.2	+0.8
Electronic dance	26	Aziba	−0.4	−1.2	−0.8	−2.6	−1.2	+0.2	+2.0	−0.2	+1.2	−0.6	+0.6	−1.2	+2.2	+1.0	−0.2	−2.2
	27	Child's Smile	+0.4	−0.2	−0.6	−1.2	+0.2	+0.2	+2.6	−2.2	−1.8	+0.8	−1.4	+0.2	+1.2	+2.4	+1.8	−2.4
	28	Protected	−0.6	0.0	−1.2	0.0	−0.6	−0.4	+1.0	−1.2	0.0	0.0	−0.2	−4.2	+4.8	+4.0	+1.2	+0.2
	29	detroit	0.0	+1.0	−0.6	−0.8	−0.2	−0.2	+2.0	−1.0	−0.2	+0.6	+1.0	−1.8	+2.0	+1.4	−2.0	−0.8
	30	E-Samba	0.0	−0.4	−1.2	0.0	−0.2	+0.4	+2.2	+2.0	−0.4	+0.6	+0.8	−2.0	+0.8	+1.6	+0.8	+0.4
Hip-hop	31	b. street nois	0.0	−1.0	−0.6	−1.8	−1.0	+0.2	+2.0	+0.2	−2.2	+1.8	0.0	−1.2	+3.0	+0.6	−0.2	−1.2
	32	04	−0.4	−1.0	0.0	0.0	+1.0	+0.8	+2.2	+1.6	+0.8	+0.8	+0.4	0.0	−0.2	+0.2	+0.2	0.0
	33	Aries4Rce_-_Boss	−0.6	−0.6	−1.2	−2.2	+2.0	−0.8	+1.6	+0.4	−0.2	+1.6	0.0	−1.4	0.0	−1.8	+1.4	−1.0
	34	Energy Bounce 2	+0.8	−2.0	−1.0	0.0	−0.2	−1.2	+1.0	+1.2	−2.2	+1.0	−0.6	+0.6	+2.0	−3.8	+0.6	−1.8
	35	Zombie Klub	0.0	−0.8	−0.4	−0.8	+0.6	−0.4	+3.6	+1.8	−0.6	+0.6	+0.4	−1.2	+0.8	−1.0	+0.2	−2.4
Metal	36	iron knight 5	−0.8	−0.8	−1.0	−2.4	−0.8	−1.0	+3.0	−1.0	−0.4	−0.6	−0.6	−0.8	+1.6	+1.4	−0.6	+0.6
	37	Ready to Fight	0.0	+0.2	−0.6	−2.2	+0.6	−0.4	+2.6	−0.6	−2.2	+1.2	+0.2	+1.0	+1.2	+0.6	−0.8	−1.0
	38	Good Morning	+1.0	−2.2	−0.8	+0.6	−1.0	−2.0	+4.0	−1.0	−3.2	+1.0	−0.8	−1.0	+2.6	+0.8	−0.4	−4.2
	39	Way of Warrior	0.0	−1.8	−1.4	−1.8	−1.2	−0.6	+2.8	−3.4	−1.2	+1.0	−0.8	−0.8	+0.6	+1.2	+1.0	−0.6
	40	Black Sky	+0.2	−1.0	−0.8	−2.2	−1.0	−1.6	+2.4	−1.2	−2.0	+0.8	+1.4	−1.4	+4.6	+0.6	−2.4	+0.8
Pop	41	Very Upbeat	+0.2	−0.6	−1.2	−1.2	+0.2	+0.8	+1.8	−1.4	−0.8	+1.4	0.0	+0.6	+1.8	−1.0	+0.6	−1.8
	42	On My Day Off	−0.6	−0.4	−1.8	−2.2	+1.0	−1.6	+3.4	0.0	+0.8	0.0	+1.0	−0.6	+1.2	+1.0	−1.0	+1.2
	43	Mr. Bleach	−0.8	−1.6	−1.0	−2.8	+0.4	−0.2	+3.2	+0.8	−0.2	+1.2	0.0	+1.2	+2.2	−1.8	−0.8	−2.2
	44	Toys	−0.6	−0.8	−1.4	−1.0	+0.2	−0.6	+2.0	+0.8	−3.0	+1.0	+0.2	+2.4	+0.8	−1.6	−1.2	−0.4
	45	Noris	+1.0	+0.2	−0.2	−1.2	−0.2	−0.2	+2.4	+0.6	−0.2	+0.2	−1.0	+0.4	+1.2	+0.2	+3.0	−1.2
Rock	46	Powiedzim wszystkim nie...	+1.4	−0.8	−1.6	−2.0	−2.0	−1.2	+2.6	−0.6	−2.4	+1.0	−0.4	−1.0	+1.8	+1.4	+1.0	+0.4
	47	happy hour	+0.4	−1.6	−1.0	−1.4	0.0	−1.8	+1.8	−1.0	−1.2	+0.4	+1.2	−2.6	+0.6	+0.2	−0.2	−0.4
	48	Last Launch Part I	+0.8	−1.6	−1.0	−2.0	−0.6	−0.4	+2.2	−2.0	−1.8	−0.2	−2.0	+0.8	+2.2	−0.8	−1.6	−1.2
	49	Banned	+1.4	−2.0	−0.6	−1.8	−0.8	−1.8	+1.4	−2.2	−3.6	+2.6	+0.4	−1.0	+0.2	−2.2	−1.4	−2.6
	50	We Don't Piss In Your Ashtrays	+1.4	−2.8	−0.8	−0.4	0.0	−1.4	+3.2	−1.0	−2.2	+1.6	−1.4	+0.6	−0.4	+2.2	−1.4	−0.8
Jazz	51	N400	+1.2	−1.0	−0.4	−1.2	−2.0	−0.8	+3.0	−0.8	−1.2	+0.2	+1.0	0.0	+0.6	+2.0	−1.6	−2.0
	52	Der lachende Vulkanier...	+1.0	−0.4	−1.4	−0.8	−0.2	−1.2	+2.2	−0.2	−1.8	+0.4	+1.0	−2.0	+0.6	+2.2	−3.2	−1.2
	53	faaâ	+1.0	−1.6	−1.2	−0.4	+0.4	+0.2	+1.4	+0.6	+0.6	+0.8	+0.2	−1.0	0.0	+0.8	−1.2	0.0
	54	My name is Nova, Bossa Nova	+0.4	0.0	+0.2	0.0	−0.4	−0.6	+2.2	−1.2	−0.4	+0.8	−0.2	−1.8	+3.2	+2.2	+0.2	−1.6
	55	La toupie	−0.8	0.0	−1.4	−0.8	−0.4	−0.4	+0.8	−2.0	−0.8	+1.0	+0.2	+1.0	+0.8	+0.8	−4.0	−3.0
		All music mean	**+0.3**	**−0.8**	**−1.0**	**−1.2**	**0.0**	**−0.7**	**+2.2**	**−0.3**	**−1.0**	**+0.8**	**+0.1**	**−0.6**	**+1.6**	**+0.1**	**−0.9**	**−0.8**
		White-noise control	**−0.8**	**−1.8**	**−1.8**	**−2.2**	**−1.1**	**−1.0**	**+2.0**	**−2.4**	**−1.6**	**−1.3**	**−0.9**	**−3.0**	**+1.6**	**−0.7**	**−0.2**	**−2.0**

Lower score — Higher score

-16 -8 0 +8 +16

Phi-4: Phi-4 Multimodal · Vox3: Voxtral Mini · Q3: Qwen3-Omni · Q2.5: Qwen2.5-Omni

Vox24: Voxtral Small · Step: Step-Audio-R1.1 · NV: Nemotron · Mini: MiniCPM-o 4.5

H: historical run. Recording titles shortened; full catalog in appendix. Colors do not denote significance.

Figure 1 (continued): MMLU tests applying subject knowledge to multiple-choice questions across 20 selected subjects. IFEval tests obeying explicit response constraints, such as length, keywords, and format. Both use the same 55 recordings and eight models. Model order, recording order, and color scale match the preceding page. H denotes the historical Qwen2.5-Omni run.

Mathematical reasoning: problem complexity

1 / 2

55 recordings · 8 models · 500 items / setting · +10 dB SNR

Change from matched clean speech (percentage points)

Genre		Recording	Math word problems GSM-Symbolic								Math word problems GSM +1							
			Phi-4	Vox3	Q3	Q2.5	Vox24	Step	NV	Mini	Phi-4	Vox3	Q3	Q2.5	Vox24	Step	NV	Mini
Folk / country	01	glinguinette	−1.0	+7.8	−2.4	−0.2	−0.6	−0.4	+0.8	+4.0	−1.2	+6.6	−6.2	−1.0	−5.0	+3.0	−0.4	+0.2
	02	01 Ouverture	+0.2	+7.4	−2.2	+1.6	+1.4	−1.6	+0.4	+2.2	−1.6	+5.8	−5.2	−0.8	−4.0	+0.8	0.0	+0.2
	03	Y'vonim	+0.6	+6.8	−3.0	−0.8	−0.2	−1.4	+0.6	+3.4	−0.4	+6.2	−5.2	−0.6	−3.8	+1.0	0.0	−0.6
	04	Folk omelette	+0.2	+6.6	−3.8	+0.4	+2.0	−1.2	+1.8	+3.6	−1.4	+5.2	−4.6	−1.0	−5.0	+1.2	−0.2	−0.6
	05	Finding Home	−0.8	+10.2	−3.4	+0.8	+1.6	−0.8	+1.6	+3.8	−1.2	+6.4	−5.2	−0.8	−4.4	+2.0	−0.4	−0.4
Ambient	06	Jotun	−0.6	+6.4	−2.4	−0.4	+2.0	−0.6	+0.4	+1.2	−0.2	+7.2	−3.8	+0.2	−1.8	+0.8	−0.2	−0.4
	07	prolog03	+1.0	+6.8	−2.4	+1.8	+1.4	−1.4	+0.8	+2.4	−1.2	+6.4	−4.6	−0.4	−3.4	+0.4	0.0	−0.2
	08	In Memories (Colors)	0.0	+10.0	−4.0	+1.6	−0.4	−2.2	+1.6	+3.6	−0.8	+6.4	−5.8	−0.4	−4.2	0.0	−0.6	−0.2
	09	Spartial	−0.6	+9.4	−2.4	+0.6	+0.8	−1.0	−0.2	+2.2	−1.2	+5.4	−5.6	−0.4	−4.2	−0.2	−0.4	+0.2
	10	Sound is I	0.0	+7.0	−2.2	0.0	+3.2	−0.2	+1.4	+1.4	−1.2	+6.4	−4.0	−0.4	−2.4	−1.4	−1.0	−0.6
Chinese traditional	11	Liu Shui	−0.2	+7.4	−3.6	+0.8	+2.2	−1.2	0.0	+0.4	−1.0	+5.8	−4.4	−1.0	−2.8	−2.0	−0.6	0.0
	12	Shi Mian Mai Fu	0.0	+8.4	−2.2	+1.0	+0.8	−0.2	+0.4	+0.6	−0.4	+6.0	−5.6	−0.8	−3.0	−2.0	−0.2	−0.6
	13	Chun Dao Yi He	+0.8	+12.0	−3.2	+0.6	+2.2	−0.8	+1.0	0.0	−1.0	+10.8	−4.6	−0.8	−1.2	−1.6	−0.6	0.0
	14	Zhan Tai Feng	+1.0	+7.0	−3.0	+0.6	+3.8	−1.8	+1.6	+0.6	−1.4	+4.2	−4.2	+0.8	−1.4	−1.6	−0.8	−0.2
	15	Erquan Yingyue	+0.4	+7.6	−2.2	+0.2	+3.0	−1.2	−0.8	+2.6	−0.8	+5.0	−4.6	−0.6	−2.6	−1.4	−0.8	−0.8
Cinematic orchestral	16	Sunrise	−0.6	+7.4	−1.4	+1.0	+1.8	+0.4	+0.2	+1.8	−1.0	+6.8	−5.4	−0.6	−3.2	+0.4	0.0	+0.4
	17	Archer Revenge	0.0	+8.4	−2.6	−1.0	−2.4	−1.4	+1.0	+4.2	−1.2	+7.0	−5.2	−0.4	−3.8	+0.6	−0.4	0.0
	18	Dawn Of Heroes	−0.6	+6.8	−2.8	+0.8	+0.2	−0.6	+0.4	+3.6	−1.4	+6.2	−6.0	−0.6	−2.6	+0.2	−0.4	−0.4
	19	Fairy Tale for the Lonely	−0.6	+9.2	−3.4	−0.2	+0.8	−0.2	+1.2	+3.6	0.0	+8.6	−6.0	−0.4	−3.8	+0.2	−0.4	−0.4
	20	Prepare to Be Boarded, Comrades	−0.2	+7.4	−1.8	+0.6	+2.6	−1.2	+1.6	+1.8	−1.2	+5.8	−3.8	−0.6	−3.8	0.0	+0.6	+0.2
Classical	21	Bach · Brandenburg No. 3	−0.4	+6.4	−3.8	+0.4	+0.2	−0.2	0.0	+5.0	−1.0	+6.8	−5.0	−1.6	−3.2	+0.2	0.0	−0.2
	22	Beethoven · Symphony No. 5	0.0	+6.6	−3.2	0.0	+1.0	−0.8	+0.6	+3.8	−0.8	+7.6	−4.8	0.0	−2.0	+1.6	0.0	−0.6
	23	Chopin · Scherzo, Op. 31	−0.8	+10.8	−4.0	−0.6	+1.4	−0.6	+0.8	−1.4	−0.8	+9.8	−5.0	−0.2	−3.0	+1.2	−0.8	0.0
	24	Vivaldi · Two Violins, Op. 3/11	−1.0	+9.2	−4.2	0.0	+1.2	−1.8	−0.2	+5.2	−1.2	+6.6	−6.6	−0.8	−2.6	+0.8	−0.4	+0.2
	25	Dvořák · Serenade, Op. 22	0.0	+7.6	−1.8	+1.6	+0.8	−0.6	−1.6	+4.2	−0.8	+8.0	−5.8	−1.4	−3.8	+1.6	−0.4	−0.4
Electronic dance	26	Aziba	0.0	+10.8	−3.6	−0.8	+2.4	−1.2	−0.2	+2.6	−0.8	+7.2	−5.8	−0.8	−5.4	+1.8	−0.4	−0.2
	27	Child's Smile	+0.2	+6.2	−2.8	+0.4	+1.0	−1.0	+0.6	+3.0	−1.0	+3.8	−6.2	−0.4	−5.4	−0.6	−0.6	0.0
	28	Protected	−0.2	+6.0	−3.6	+0.2	+1.4	−1.4	+0.4	+3.6	−1.2	+5.6	−6.0	−0.4	−5.0	+0.2	−1.0	−0.4
	29	detroit	−0.2	+7.4	−4.0	+0.2	−0.4	−0.4	+1.6	+2.2	−0.6	+5.2	−6.2	−0.8	−4.0	+1.0	0.0	−0.4
	30	E-Samba	0.0	+6.8	−2.8	−0.8	−1.0	0.0	+0.8	+0.8	−1.2	+8.0	−5.0	−0.6	−4.0	+1.2	−0.8	+0.2
Hip-hop	31	b. street nois	−0.4	+10.2	−3.0	−0.2	+3.2	−1.2	+0.4	+1.4	−1.4	+5.6	−5.4	−1.0	−3.2	+0.2	−0.4	−0.4
	32	04	+0.6	+8.0	−2.8	−0.6	+0.6	−0.8	−0.2	+3.0	−1.2	+6.6	−5.4	−0.8	−2.2	+1.2	−0.8	+0.2
	33	Aries4Rce_-_Boss	−0.6	+6.0	−2.8	−1.0	+1.0	+0.2	+0.4	+0.2	−1.0	+4.4	−5.8	−0.8	−1.4	0.0	−1.4	+0.2
	34	Energy Bounce 2	0.0	+10.4	−2.6	+0.4	+2.0	−0.8	+0.8	0.0	−1.6	+6.4	−6.8	−1.0	−2.8	+1.2	−1.0	+0.8
	35	Zombie Klub	−0.4	+4.6	−3.0	+0.8	+3.6	−1.4	−0.4	−3.4	−1.2	+4.0	−1.8	−1.0	−2.2	+0.2	−1.0	−0.6
Metal	36	iron knight 5	+0.6	+4.8	−3.0	−0.4	+0.6	−2.4	0.0	+3.4	−1.4	+3.6	−4.6	−0.8	−4.0	+0.2	−0.2	+0.2
	37	Ready to Fight	−0.4	+5.8	−3.8	−0.4	+0.4	−1.6	+0.2	+3.8	−0.6	+6.4	−4.4	−1.0	−4.4	+0.2	+0.4	0.0
	38	Good Morning	+0.4	+6.8	−4.2	+0.4	−1.4	−1.8	+1.0	+1.6	−1.4	+6.8	−4.8	−1.0	−5.8	−0.2	−0.2	+0.4
	39	Way of Warrior	+0.2	+6.8	−3.6	−0.4	0.0	−0.8	+1.4	+4.0	−1.8	+6.0	−4.6	−1.2	−4.4	−1.4	+0.8	−0.2
	40	Black Sky	−0.4	+4.4	−1.6	−0.6	+1.2	−1.0	+0.4	+3.2	−0.8	+4.0	−5.2	−0.6	−3.4	+0.6	−0.2	−0.2
Pop	41	Very Upbeat	−1.2	+9.2	−3.2	+0.8	+0.6	−0.6	+0.4	+1.6	−1.0	+7.2	−5.4	−1.2	−6.6	+2.0	−0.4	−0.8
	42	On My Day Off	+0.2	+8.4	−3.4	+1.2	+1.8	−0.4	+0.8	+1.8	−0.8	+6.8	−4.8	−1.0	−3.4	+2.0	0.0	0.0
	43	Mr. Bleach	−0.8	+8.4	−2.8	0.0	+1.8	−1.6	+0.6	+0.6	−1.2	+6.6	−4.4	−0.6	−4.2	+1.2	−0.4	+0.8
	44	Toys	−0.4	+10.2	−2.6	0.0	+1.8	−1.0	+0.8	+1.0	−0.8	+7.4	−5.2	−1.0	−2.8	+3.4	−0.4	+0.8
	45	Noris	−0.4	+9.0	−2.4	−0.6	+0.4	−1.0	−0.4	+1.8	−1.0	+7.2	−5.6	+0.2	−5.0	+1.8	−1.4	−0.6
Rock	46	Powiedzim wszystkim nie...	−0.6	+8.0	−3.8	+0.2	−1.6	−1.4	+1.6	+3.6	−1.0	+5.4	−6.0	−1.0	−3.8	+1.0	0.0	−0.4
	47	happy hour	0.0	+8.2	−3.4	−1.6	−1.4	−1.8	+1.2	+3.0	−1.0	+7.2	−5.6	−1.2	−4.6	−0.8	+0.2	−0.2
	48	Last Launch Part I	−1.0	+6.8	−3.6	+0.2	+2.2	−1.4	+0.4	+3.6	−1.0	+7.4	−5.8	−0.6	−4.6	+0.4	0.0	−0.6
	49	Banned	−1.0	+8.8	−4.2	+0.4	−0.2	−0.6	−0.4	+4.4	−0.8	+5.0	−5.8	−1.6	−3.6	−0.4	+0.6	−0.6
	50	We Don't Piss In Your Ashtrays	−0.8	+6.8	−3.0	−0.2	−1.0	−1.6	+1.2	+2.0	−1.8	+4.4	−6.4	−0.6	−4.6	0.0	0.0	−0.2
Jazz	51	N400	−0.8	+6.2	−2.6	−1.4	+1.4	−0.8	+0.2	+2.6	−1.6	+4.4	−3.4	−1.0	−2.8	+1.8	−0.2	−0.4
	52	Der lachende Vulkanier...	−1.0	+10.6	−2.8	−0.2	−0.4	−0.8	−0.2	+2.0	−0.8	+6.6	−5.4	−0.4	−4.4	+0.2	−0.4	+0.4
	53	faaå	+0.2	+7.2	−2.2	+0.6	+1.4	−0.6	+0.2	+2.8	−1.6	+6.8	−4.4	−1.2	−2.8	+0.8	−0.4	−0.4
	54	My name is Nova, Bossa Nova	−0.4	+9.2	−2.8	−0.2	+0.2	−0.6	−0.8	+2.6	−0.4	+6.4	−5.2	−0.6	−3.4	−0.2	−0.6	−0.2
	55	La toupie	−0.4	+8.2	−4.2	+1.4	−0.6	0.0	+0.4	+2.6	−1.0	+7.0	−5.8	−1.6	−3.2	−1.8	0.0	−0.6
	All music mean		**−0.2**	**+7.8**	**−3.0**	**+0.2**	**+0.9**	**−1.0**	**+0.5**	**+2.3**	**−1.0**	**+6.3**	**−5.2**	**−0.7**	**−3.6**	**+0.4**	**−0.3**	**−0.2**
	White-noise control		**+0.2**	**+7.2**	**−4.6**	**+0.6**	**−0.2**	**−1.6**	**+0.3**	**+2.6**	**−1.8**	**+4.0**	**−6.8**	**−0.8**	**−3.0**	**+0.4**	**−0.1**	**−0.4**

Lower score — Higher score

-16 -8 0 +8 +16

Phi-4: Phi-4 Multimodal · Vox3: Voxtral Mini · Q3: Qwen3-Omni · Q2.5: Qwen2.5-Omni

Vox24: Voxtral Small · Step: Step-Audio-R1.1 · NV: Nemotron · Mini: MiniCPM-o 4.5

Recording titles shortened; full catalog in appendix. Colors do not denote significance.

Figure 2: Mathematical reasoning at different problem complexities: 55 recordings, eight models, and three GSM-Symbolic variants (500 items each), across two portrait pages. Base tests arithmetic under changes to names and numbers; P1 adds one relevant clause, and P2 adds two. Base and P1 are shown here; P2 follows. All models receive identical regenerated audio. Values are changes from matched clean speech (pp).

Mathematical reasoning: problem complexity

2 / 2

55 recordings · 8 models · 500 items / setting · +10 dB SNR

Change from matched clean speech (percentage points)

Math word problems

GSM +2

Genre		Recording	Phi-4	Vox3	Q3	Q2.5	Vox24	Step	NV	Mini
Folk / country	01	glinguinette	+0.2	+3.2	−2.0	−1.2	−0.6	−0.4	−0.6	0.0
	02	01 Ouverture	+0.6	+3.4	−0.6	−1.0	−0.4	−0.2	−1.0	0.0
	03	Y'vonim	+0.2	+4.6	−2.6	−0.4	+1.0	−0.6	−0.8	0.0
	04	Folk omelette	+0.8	+2.8	−1.4	−0.4	+0.2	−0.6	−0.8	0.0
	05	Finding Home	+1.0	+2.4	−2.4	−0.6	−0.4	−0.4	−0.6	0.0
Ambient	06	Jotun	+0.6	+2.4	−1.4	−0.8	+1.2	+0.8	−1.2	0.0
	07	prolog03	+0.4	+2.0	−1.8	−1.0	+1.6	−0.6	−0.6	0.0
	08	In Memories (Colors)	0.0	+3.4	−1.8	−0.8	−0.6	−1.2	−1.0	0.0
	09	Spartial	+1.6	+3.0	−1.4	+0.4	0.0	−0.4	−0.6	0.0
	10	Sound is I	+0.2	+2.8	−1.6	−1.6	+0.4	+1.0	−0.4	0.0
Chinese traditional	11	Liu Shui	+0.6	+1.6	−0.8	−1.0	+0.6	−0.2	−0.8	+0.2
	12	Shi Mian Mai Fu	+1.0	+3.4	−1.8	−1.0	+0.8	+1.8	−0.4	0.0
	13	Chun Dao Yi He	0.0	+3.4	−1.4	−0.6	+0.6	+0.6	−0.6	0.0
	14	Zhan Tai Feng	+1.2	+2.2	−1.8	−0.6	+1.0	+2.6	−0.6	0.0
	15	Erquan Yingyue	+1.0	+2.2	−1.6	−0.6	+1.0	+2.2	−0.8	0.0
Cinematic orchestral	16	Sunrise	+1.2	+4.2	−1.2	−0.6	+1.2	+0.2	−1.6	0.0
	17	Archer Revenge	+1.4	+2.0	−2.2	−0.8	+0.6	+1.8	−0.8	0.0
	18	Dawn Of Heroes	+0.8	+2.6	−1.6	−1.4	+1.6	0.0	−1.2	+0.2
	19	Fairy Tale for the Lonely	+1.6	+3.8	−1.0	−1.0	+1.8	−0.2	−1.0	0.0
	20	Prepare to Be Boarded, Comrades	+0.2	+1.0	−1.8	−0.4	+1.6	−0.2	−0.2	0.0
Classical	21	Bach · Brandenburg No. 3	+0.6	+2.4	−1.6	−1.0	−0.2	+1.6	+0.2	0.0
	22	Beethoven · Symphony No. 5	+0.4	+2.6	−1.6	−0.6	+0.6	0.0	−0.6	0.0
	23	Chopin · Scherzo, Op. 31	+0.6	+2.4	−0.4	−0.2	−0.6	+1.2	−0.8	0.0
	24	Vivaldi · Two Violins, Op. 3/11	+1.0	+1.8	−1.6	−0.2	+1.6	+0.4	−0.8	0.0
	25	Dvořák · Serenade, Op. 22	+0.2	+1.8	−1.6	−1.0	−0.2	+0.8	−1.2	0.0
Electronic dance	26	Aziba	+0.6	+2.8	−2.8	−0.2	−0.6	+0.6	−0.2	0.0
	27	Child's Smile	+0.4	+2.6	−1.2	−0.8	−0.8	+1.0	+0.4	0.0
	28	Protected	+1.2	+1.6	−2.0	−1.0	−0.6	+0.6	−0.2	0.0
	29	detroit	+1.4	+2.4	−1.2	−1.4	−0.4	+2.0	+0.2	0.0
	30	E-Samba	+1.0	+4.0	−1.6	−0.6	−0.2	+1.8	0.0	+0.2
Hip-hop	31	b. street nois	+0.4	+2.4	−1.6	−0.2	−0.2	+1.0	+0.4	+0.2
	32	04	−0.2	+2.6	−2.8	−0.4	−0.8	+1.4	0.0	0.0
	33	Aries4Rce_-_Boss	+0.8	+1.8	−1.8	0.0	−0.2	+2.4	0.0	0.0
	34	Energy Bounce 2	+1.8	+4.0	−1.4	−1.0	+0.4	+1.6	−0.4	+0.2
	35	Zombie Klub	+1.0	+0.8	−2.2	−0.4	+0.6	+1.0	−0.8	0.0
Metal	36	iron knight 5	+1.0	+2.6	−0.8	−0.8	−0.4	+1.2	+0.2	+0.2
	37	Ready to Fight	+0.4	+3.2	−1.8	−2.0	0.0	+1.0	−0.2	0.0
	38	Good Morning	0.0	+3.0	−1.4	−1.6	−0.4	−0.8	−1.0	0.0
	39	Way of Warrior	+1.4	+3.4	−2.0	−0.4	0.0	+0.2	−0.4	+0.2
	40	Black Sky	+1.0	+2.0	−1.6	−1.0	−0.2	+1.2	−0.8	0.0
Pop	41	Very Upbeat	+0.4	+3.8	−2.2	−1.0	−0.2	0.0	−0.6	0.0
	42	On My Day Off	+1.2	+3.2	−1.6	−1.6	−0.2	−0.2	−1.0	0.0
	43	Mr. Bleach	+0.8	+1.6	−1.6	−0.4	+0.2	−0.4	−0.4	0.0
	44	Toys	+1.0	+2.6	−2.0	−1.0	−0.6	+0.6	−0.8	0.0
	45	Noris	+0.8	+2.2	−2.2	−1.0	−1.0	+1.0	−0.2	0.0
Rock	46	Powiedzim wszystkim nie...	+0.8	+2.4	−1.4	0.0	+1.0	−1.6	−0.6	0.0
	47	happy hour	+0.6	+2.8	−1.2	−1.2	−0.2	+0.2	−0.6	+0.2
	48	Last Launch Part I	+0.8	+0.8	−1.8	−0.6	+0.6	−0.2	−1.0	0.0
	49	Banned	+0.8	+2.4	−2.0	−1.0	+0.4	+0.2	−0.8	0.0
	50	We Don't Piss In Your Ashtrays	+0.6	+4.0	−1.6	−1.2	+0.4	−0.4	−0.4	+0.2
Jazz	51	N400	+1.0	+2.4	−2.2	−1.0	0.0	+0.8	−0.4	0.0
	52	Der lachende Vulkanier...	+1.4	+2.2	−2.0	−0.6	−0.2	+0.2	−0.4	0.0
	53	faaâ	+1.2	+2.2	−3.0	−1.4	−0.6	−0.2	−0.8	+0.2
	54	My name is Nova, Bossa Nova	+1.0	+2.6	−1.8	−1.6	−0.2	−0.2	0.0	+0.2
	55	La toupie	+0.6	+2.6	−2.4	−0.4	−0.4	−0.2	−0.2	0.0
		All music mean	**+0.8**	**+2.6**	**−1.7**	**−0.8**	**+0.2**	**+0.5**	**−0.5**	**0.0**
		White-noise control	**+1.6**	**+2.6**	**−1.4**	**−0.6**	**+1.4**	**+0.1**	**−0.1**	**0.0**

Lower score ← → Higher score

-16 -8 0 +8 +16

Phi-4: Phi-4 Multimodal · Vox3: Voxtral Mini · Q3: Qwen3-Omni · Q2.5: Qwen2.5-Omni

Vox24: Voxtral Small · Step: Step-Audio-R1.1 · NV: Nemotron · Mini: MiniCPM-o 4.5

Recording titles shortened; full catalog in appendix. Colors do not denote significance.

Figure 2 (continued): The P2 (GSM +2) setting for the same 55 recordings and eight models, with the same model order and color scale. Values are changes from matched clean speech (pp).

Question answering: context structure

1 / 2

55 recordings · 8 models · 300 items / setting · +10 dB SNR

Change from matched clean speech (percentage points)

Genre		Recording	Fact-based question answering — bAbI short								Fact-based question answering — bAbI near							
			Phi-4	Vox3	Q3	Q2.5	Vox24	Step	NV	Mini	Phi-4	Vox3	Q3	Q2.5	Vox24	Step	NV	Mini
Folk / country	01	glinguinette	−1.3	+2.0	−1.7	+1.3	−3.0	+1.3	−1.3	0.0	+0.3	−5.0	+0.7	−4.3	−3.7	−0.3	−7.7	−1.3
	02	01 Ouverture	+1.0	+3.7	−2.7	+1.3	−2.3	+2.0	−0.3	+0.3	+1.7	−1.7	+0.7	−2.7	−4.3	+0.3	−7.0	+0.3
	03	Y'vonim	+0.3	+2.3	−1.3	−1.3	−2.7	−0.7	−2.3	+0.3	+1.0	−2.0	−0.3	−4.0	0.0	+2.3	−8.3	−5.7
	04	Folk omelette	+1.7	+3.0	−1.3	+1.0	−2.0	+0.7	−2.0	+2.7	+2.3	0.0	+3.0	−2.3	−3.3	+0.3	−7.7	−1.0
	05	Finding Home	+1.3	+3.0	−2.7	+1.7	−1.3	−0.3	−1.3	+0.7	+3.7	−1.3	−1.7	−4.0	−3.0	+1.7	−7.3	−4.7
Ambient	06	Jotun	+0.7	+2.0	−3.0	+2.7	−1.3	0.0	−1.0	+3.3	−1.0	−3.0	−1.3	−3.0	−2.7	−1.7	−6.0	−1.7
	07	prolog03	+2.0	+1.0	−2.0	+2.7	−2.0	−0.7	−3.7	+2.7	+0.7	+0.7	+4.3	−5.7	−1.7	+0.7	−7.0	−3.7
	08	In Memories (Colors)	+0.7	+4.3	−0.3	+2.0	−3.7	+0.3	−2.3	+1.7	−1.0	−1.0	+1.3	−6.3	−1.3	−1.7	−6.7	+0.7
	09	Spartial	+2.0	+2.3	−2.0	−1.0	−0.7	+1.0	−4.7	+1.7	0.0	+0.7	+0.3	−4.7	−3.7	−0.7	−5.3	−4.7
	10	Sound is I	+1.0	+3.0	−2.0	+0.7	−2.7	−1.0	−0.7	+1.3	−1.0	−5.0	+2.0	−4.7	−3.7	+0.3	−5.0	0.0
Chinese traditional	11	Liu Shui	−0.7	+1.3	−2.0	+1.7	−1.0	0.0	−3.3	+2.3	−0.7	−2.0	−0.7	−1.7	−2.7	+1.7	−8.0	−1.3
	12	Shi Mian Mai Fu	+1.7	+3.3	−3.0	+1.3	−1.7	+1.7	−3.7	−1.0	+2.3	−1.7	+2.7	−1.3	−1.0	+3.0	−7.7	−2.7
	13	Chun Dao Yi He	+1.0	+5.7	−0.7	+3.3	−1.7	+1.3	−3.3	0.0	+1.3	0.0	−1.0	−1.0	−2.3	+1.7	−9.7	−1.7
	14	Zhan Tai Feng	0.0	+2.3	−1.3	−1.3	−0.3	−0.3	−1.0	+1.7	+1.0	+0.3	+3.7	−3.3	−2.7	+2.3	−6.7	0.0
	15	Erquan Yingyue	+3.0	+3.0	−2.0	+0.7	−0.7	+0.3	0.0	−1.3	+0.7	−6.0	+2.0	−3.3	−3.0	+1.7	−4.7	−4.7
Cinematic orchestral	16	Sunrise	+1.3	+1.7	−3.0	+1.3	+0.7	0.0	−5.0	+1.7	+2.3	−1.0	+2.3	−4.0	−2.3	+1.7	−6.0	−4.0
	17	Archer Revenge	0.0	+2.3	−1.0	+2.0	+0.3	+2.0	−4.3	+2.3	+1.3	−1.0	−0.3	−6.3	−3.0	+1.3	−7.0	0.0
	18	Dawn Of Heroes	−1.3	+3.7	−2.0	+0.3	−1.0	+2.7	−2.3	+1.3	+1.0	−2.0	−1.3	−4.7	−2.0	−0.3	−5.7	−3.0
	19	Fairy Tale for the Lonely	+1.3	+2.7	−1.7	+1.3	−0.3	+1.0	−2.3	−0.3	+1.3	−1.3	0.0	−4.3	−3.0	+0.3	−7.3	−3.3
	20	Prepare to Be Boarded, Comrades	+2.0	+3.0	−1.3	−0.3	+0.7	+1.3	−2.7	+1.3	+1.0	−1.7	+2.3	−4.0	−4.0	+1.0	−6.3	−2.0
Classical	21	Bach · Brandenburg No. 3	+1.0	+3.3	−3.0	+1.0	−1.0	+2.3	−4.3	+2.0	+0.3	−3.0	−0.7	−4.7	−1.3	−0.3	−7.3	−5.7
	22	Beethoven · Symphony No. 5	+3.7	+2.3	−2.0	+0.7	0.0	+1.3	−5.0	+3.7	+1.7	−3.7	+4.0	−3.0	−3.3	−0.3	−9.3	−2.3
	23	Chopin · Scherzo, Op. 31	+1.0	+1.0	−0.7	−0.7	−1.3	+0.7	−4.7	+2.7	+1.3	0.0	+0.7	−4.3	−5.0	−1.3	−7.3	−0.7
	24	Vivaldi · Two Violins, Op. 3/11	−0.7	+2.7	−3.0	+0.7	−1.3	+3.0	−3.3	−0.3	+3.3	−1.3	−3.7	−2.7	−0.7	+0.3	−10.3	−4.7
	25	Dvořák · Serenade, Op. 22	+1.0	+3.3	−1.3	+0.3	−1.0	+1.3	−2.7	+1.7	+2.0	−3.7	+1.0	−2.3	−1.0	+0.7	−8.0	−8.7
Electronic dance	26	Aziba	+3.7	+3.3	−3.0	0.0	−2.7	+0.3	−2.7	+0.7	+1.0	−1.3	+0.3	−5.0	+0.3	−2.3	−9.0	−0.3
	27	Child's Smile	+2.0	+5.0	−2.3	+2.3	−1.0	+2.7	−2.0	+1.7	−1.0	−4.3	0.0	−2.0	−2.3	−0.3	−9.3	−4.7
	28	Protected	+1.3	+5.7	−2.7	+1.7	−3.0	+1.7	−3.3	−1.3	−1.0	0.0	−1.0	−1.7	−1.7	+1.3	−8.7	−2.3
	29	detroit	+0.3	+3.3	−3.0	+1.0	−0.7	+2.0	−1.7	+2.0	−0.3	−1.3	+0.3	−5.0	−3.0	0.0	−8.7	−0.3
	30	E-Samba	+1.3	+3.3	−2.3	+1.7	−2.0	+2.7	−3.3	+2.0	+1.7	+1.3	+2.0	−5.7	−2.0	+1.7	−4.0	+0.7
Hip-hop	31	b. street nois	+2.0	+2.0	−2.0	+2.3	−2.7	+1.0	−3.7	+2.0	+0.3	−2.7	+2.3	−2.3	−2.7	+0.7	−9.0	+2.0
	32	04	−0.7	+3.0	−3.7	+1.0	−1.7	+1.0	−3.7	+1.3	+0.7	0.0	+1.3	−1.7	−4.3	+0.3	−8.3	−1.7
	33	Aries4Rce_-_Boss	+0.3	+3.7	−2.3	+1.7	+0.7	+2.0	−2.7	+3.3	+0.7	+0.3	−1.7	−4.7	−2.7	−1.0	−7.0	−3.3
	34	Energy Bounce 2	+1.3	+4.0	−4.0	+1.0	−3.0	−0.7	−3.3	+2.3	−1.3	+1.3	0.0	−2.0	−2.7	−1.0	−8.7	+3.0
	35	Zombie Klub	+0.7	+3.7	−2.7	+2.0	−3.3	−0.7	−4.0	+4.7	+3.3	−1.7	+0.7	−1.7	−4.0	+1.0	−9.3	+2.0
Metal	36	iron knight 5	+1.7	−0.3	−0.7	+1.7	−3.3	+2.0	−3.0	+4.3	+0.3	−2.3	−3.0	−4.7	−2.0	+2.3	−6.7	−2.0
	37	Ready to Fight	−1.3	+3.0	−4.0	+1.3	−3.7	+4.0	−3.3	+3.0	+2.3	−2.7	−1.3	−3.0	−4.7	+3.0	−7.3	−2.7
	38	Good Morning	+2.3	+2.7	−4.7	+0.7	−2.0	+1.3	−2.7	+2.7	−1.0	−4.0	+1.7	−6.7	−5.7	+2.7	−8.0	−1.3
	39	Way of Warrior	+0.7	+2.3	−4.0	+1.3	−2.3	+1.3	−4.3	+3.0	+0.3	−4.0	−2.3	−3.3	−5.3	+2.7	−8.0	−4.7
	40	Black Sky	−1.0	+3.3	−3.0	+1.3	−1.0	+1.7	−0.3	−1.3	−0.7	−3.0	+0.3	−5.0	−1.7	+2.0	−4.7	−2.7
Pop	41	Very Upbeat	0.0	+2.3	−4.0	+1.7	0.0	−1.3	−4.0	+2.0	+0.7	−1.0	+1.0	−1.7	0.0	+0.7	−9.0	−1.0
	42	On My Day Off	0.0	+3.0	−2.0	−0.3	−1.7	−1.0	−2.3	+0.7	+4.3	−0.3	+2.0	−3.0	−3.0	+0.3	−8.3	−2.7
	43	Mr. Bleach	+2.3	+5.0	−2.7	+1.7	−3.0	+0.7	−3.3	+0.7	−0.7	−3.0	+1.7	−3.7	−3.0	−1.0	−8.3	+1.0
	44	Toys	+1.0	+4.0	−2.3	+0.3	−1.7	0.0	−3.0	+3.3	−1.3	+1.3	−0.7	−4.7	−3.3	+3.3	−8.3	0.0
	45	Noris	+1.3	+3.3	−2.3	+1.0	−3.0	+0.3	−4.3	+1.3	+0.3	−0.7	+2.3	−5.0	−3.7	+1.0	−8.3	+1.3
Rock	46	Powiedzim wszystkim nie...	+0.7	+5.3	−1.7	+1.7	−1.3	+1.0	−3.0	+0.7	−2.0	−0.3	0.0	−5.0	0.0	+2.3	−7.7	−4.3
	47	happy hour	+1.3	+4.0	−3.7	+0.7	−1.3	+0.7	−2.3	+2.7	+0.3	−3.3	−1.0	−5.0	−2.0	+4.0	−8.7	−8.0
	48	Last Launch Part I	+0.7	+3.0	−3.0	+0.7	0.0	−1.0	−1.7	+4.7	+2.3	−3.0	−0.3	−5.3	−3.3	+3.7	−8.0	−3.7
	49	Banned	0.0	+3.0	−3.0	+0.7	−1.3	+0.3	−2.0	−0.7	−0.3	0.0	−0.7	−5.7	−4.0	+4.0	−9.0	−5.7
	50	We Don't Piss In Your Ashtrays	−1.0	+3.0	−3.7	+1.3	−2.7	+1.3	+0.3	+2.3	−1.3	−1.0	−1.3	−3.7	−2.0	+2.7	−9.0	−0.7
Jazz	51	N400	+2.7	+3.7	−2.3	0.0	−3.0	+0.3	−2.0	+3.3	+2.0	−1.7	+0.7	−2.7	−5.3	+3.3	−8.3	+1.0
	52	Der lachende Vulkanier...	−0.7	+2.7	−2.0	+2.0	−1.0	+0.3	0.0	+2.7	+2.3	−0.7	+1.3	−4.3	−3.7	+3.3	−7.3	−2.3
	53	faaå	+0.3	+3.0	−1.0	+1.7	−1.3	+2.3	−2.0	+2.7	0.0	−2.0	+1.0	−4.0	−2.0	+1.3	−6.7	−3.7
	54	My name is Nova, Bossa Nova	+1.3	+3.3	−2.3	+1.7	−1.3	+1.0	−2.0	+2.7	+4.3	−0.7	+2.3	−5.3	−2.0	+1.7	−8.3	−1.7
	55	La toupie	+2.3	+2.7	−2.7	0.0	−3.0	+1.0	−2.7	+0.7	+0.3	−2.3	+0.7	−4.0	−2.7	+1.7	−9.3	−3.7
		All music mean	**+0.9**	**+3.0**	**−2.4**	**+1.1**	**−1.6**	**+0.9**	**−2.7**	**+1.7**	**+0.8**	**−1.6**	**+0.5**	**−3.8**	**−2.7**	**+1.1**	**−7.6**	**−2.2**
		White-noise control	**+1.0**	**+1.0**	**−3.3**	**−0.7**	**−2.7**	**−0.2**	**−3.0**	**+2.7**	**−1.3**	**−2.7**	**−2.7**	**−4.0**	**−0.6**	**+1.0**	**−7.8**	**−7.0**

Lower score — Higher score

-16 -8 0 +8 +16

Phi-4: Phi-4 Multimodal · Vox3: Voxtral Mini · Q3: Qwen3-Omni · Q2.5: Qwen2.5-Omni

Vox24: Voxtral Small · Step: Step-Audio-R1.1 · NV: Nemotron · Mini: MiniCPM-o 4.5

Recording titles shortened; full catalog in appendix. Colors do not denote significance.

Figure 3: Fact-based question answering: 55 recordings and eight models on 300 matched short/near/far bAbI problems, across two portrait pages. The qa1/qa2/qa3 tasks require using one, two, or three supporting facts. Short has no added filler; near places filler before the facts; far places it between the facts and question. Short and near are shown here; far follows. Values are changes from matched clean speech (pp); all matrices share a color scale.

Question answering: context structure

2 / 2

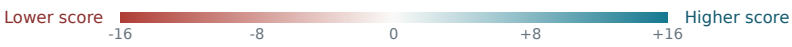
55 recordings · 8 models · 300 items / setting · +10 dB SNR

Change from matched clean speech (percentage points)

Fact-based question answering

bAbI far

Genre		Recording	Phi-4	Vox3	Q3	Q2.5	Vox24	Step	NV	Mini
Folk / country	01	glinguinette	+2.7	0.0	−5.0	−2.7	+2.0	+0.7	+1.0	−3.3
	02	01 Ouverture	+1.0	−1.0	−2.7	−0.7	+1.3	−2.7	−0.7	−3.3
	03	Y'vonim	−1.3	+1.0	−2.7	+1.3	+3.3	+2.0	+2.3	−5.7
	04	Folk omelette	+2.7	+0.7	−4.7	+1.3	+2.3	+0.7	+2.3	−3.0
	05	Finding Home	+1.0	−1.3	−4.7	−2.3	−0.7	−0.7	−1.3	−7.7
Ambient	06	Jotun	+2.7	0.0	−5.7	−0.7	0.0	−0.3	+1.0	−1.0
	07	prolog03	+1.3	−1.3	−4.3	−1.0	+1.7	−2.3	−1.3	−0.7
	08	In Memories (Colors)	+1.3	−1.3	−4.0	−2.7	−0.3	−2.3	−1.0	−4.3
	09	Spartial	+1.7	+1.3	−3.3	−3.3	+1.0	+0.3	−1.3	−5.7
	10	Sound is I	+1.0	+2.0	−3.3	−1.3	+3.3	−0.7	+0.3	+1.3
Chinese traditional	11	Liu Shui	+3.0	0.0	−4.7	0.0	+2.3	+0.7	−1.3	−1.3
	12	Shi Mian Mai Fu	+2.7	−2.0	−6.0	+1.0	+2.7	+0.3	+1.7	−7.3
	13	Chun Dao Yi He	−0.7	+1.3	−4.3	+1.7	+2.3	+1.7	0.0	−8.0
	14	Zhan Tai Feng	+0.7	0.0	−0.3	−1.7	+5.3	+2.7	−1.0	−1.0
	15	Erquan Yingyue	+1.3	−4.7	−0.7	−0.7	+3.7	+1.3	+2.0	−9.0
Cinematic orchestral	16	Sunrise	+3.3	+0.7	−1.7	−1.3	+2.7	+0.3	+0.7	−9.3
	17	Archer Revenge	−0.7	+0.7	−7.7	0.0	0.0	−0.3	−2.0	−5.7
	18	Dawn Of Heroes	+2.3	−0.3	−7.7	−2.3	+0.3	−1.3	+2.0	−4.0
	19	Fairy Tale for the Lonely	+0.3	−1.3	−5.3	−1.7	+1.3	+2.0	−1.7	−6.3
	20	Prepare to Be Boarded, Comrades	−0.3	−0.7	−1.7	+0.3	+1.7	+0.3	+0.3	−10.0
Classical	21	Bach · Brandenburg No. 3	+1.3	−1.7	−6.0	−2.0	+0.7	−0.3	−3.0	−5.0
	22	Beethoven · Symphony No. 5	+1.7	0.0	−3.3	−1.7	+0.3	−0.3	−4.7	−2.7
	23	Chopin · Scherzo, Op. 31	+3.3	−2.3	−8.0	−5.3	+1.7	+0.3	−3.3	+3.7
	24	Vivaldi · Two Violins, Op. 3/11	+3.0	−1.3	−6.7	−0.3	−1.3	−0.3	−2.3	−7.7
	25	Dvořák · Serenade, Op. 22	+2.3	+0.7	−4.0	−2.3	−1.0	−1.7	0.0	−9.3
Electronic dance	26	Aziba	+1.0	+0.3	−3.7	−4.7	+2.7	−1.7	−1.3	−5.7
	27	Child's Smile	+3.3	−2.0	−3.0	−1.0	+3.3	−0.7	+2.0	−8.7
	28	Protected	+1.7	−2.0	−4.7	−1.0	+1.3	−3.7	+1.7	−6.0
	29	detroit	+2.7	+3.0	−4.0	−3.3	+0.3	+1.0	−0.3	−3.0
	30	E-Samba	+3.0	0.0	−2.3	0.0	0.0	−3.0	+1.0	−1.3
Hip-hop	31	b. street nois	−0.3	+1.3	−3.3	−1.7	+3.3	−0.7	+2.3	−3.0
	32	04	+1.3	+3.0	−4.7	−0.7	+1.3	+2.7	+3.7	+2.0
	33	Aries4Rce_-_Boss	+2.7	−2.0	−3.3	−3.0	+3.0	+0.7	0.0	−4.0
	34	Energy Bounce 2	+1.0	+2.0	−2.3	−1.7	+1.7	−0.3	+3.7	+0.7
	35	Zombie Klub	+2.0	−1.0	−1.7	+0.3	+1.7	+0.3	+2.0	+6.0
Metal	36	iron knight 5	+2.0	−2.7	−3.3	−4.0	−1.3	−2.3	+0.7	−3.7
	37	Ready to Fight	+2.0	−1.0	−4.7	−2.3	+1.7	−1.3	−1.7	−6.3
	38	Good Morning	−0.3	+2.0	−4.7	0.0	+0.3	0.0	−1.0	−4.3
	39	Way of Warrior	+0.7	+2.3	−6.7	−0.3	0.0	−1.0	−1.7	−8.0
	40	Black Sky	+0.3	−3.3	−1.0	−1.7	+2.0	+1.3	+0.3	−7.3
Pop	41	Very Upbeat	+3.7	+1.3	−4.0	−3.7	+2.7	+2.0	+2.7	−1.3
	42	On My Day Off	+3.3	+1.7	−2.0	−3.3	−0.3	+1.0	0.0	−1.7
	43	Mr. Bleach	+1.7	−1.7	−5.0	0.0	+1.7	+2.7	+1.7	+1.3
	44	Toys	+3.0	+1.0	−2.7	−2.0	+0.3	+3.0	−1.3	+1.0
	45	Noris	+1.7	+1.7	−4.7	−1.7	+2.7	+4.3	+1.7	−2.3
Rock	46	Powiedzim wszystkim nie...	+1.3	+1.7	−4.3	−1.7	+1.7	−3.0	+2.7	−6.3
	47	happy hour	+2.0	−1.3	−6.3	−2.3	+1.3	−1.7	−0.7	−4.3
	48	Last Launch Part I	+0.7	−1.3	−6.7	−5.3	+3.7	+1.0	+0.7	−4.0
	49	Banned	+1.7	−0.3	−7.3	−4.0	+2.7	−3.0	−2.3	−9.0
	50	We Don't Piss In Your Ashtrays	+0.3	+1.7	−5.3	−1.3	+2.3	0.0	+1.0	−3.0
Jazz	51	N400	+2.0	+0.3	−1.7	−0.7	−1.7	+0.3	−2.0	−2.3
	52	Der lachende Vulkanier...	+4.7	+2.0	−1.0	−1.3	−1.3	−1.0	−0.7	+1.7
	53	faaâ	+1.7	0.0	−3.7	−1.3	+1.0	−0.7	−0.7	−0.3
	54	My name is Nova, Bossa Nova	+1.7	+2.3	−1.0	−2.0	+3.3	+1.7	−0.7	−3.3
	55	La toupie	+2.0	−2.3	−3.3	−0.7	−1.0	−0.7	+1.3	−8.7
		All music mean	**+1.7**	**−0.1**	**−4.0**	**−1.5**	**+1.4**	**0.0**	**+0.1**	**−3.8**
		White-noise control	**+0.3**	**−2.0**	**−9.0**	**−3.0**	**+1.8**	**−0.1**	**−2.7**	**−7.7**



Phi-4: Phi-4 Multimodal · Vox3: Voxtral Mini · Q3: Qwen3-Omni · Q2.5: Qwen2.5-Omni

Vox24: Voxtral Small · Step: Step-Audio-R1.1 · NV: Nemotron · Mini: MiniCPM-o 4.5

Recording titles shortened; full catalog in appendix. Colors do not denote significance.

Figure 3 (continued): The far-context setting for the same 55 recordings and eight models, with the same model order and color scale. Near and far differ in where the intact facts appear relative to the filler sentences.

REFERENCES

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, et al. Phi-4-Mini technical report: Compact yet powerful multimodal language models via mixture-of-LoRAs. arXiv:2503.01743, 2025. URL https://arxiv.org/abs/2503.01743.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x. URL https://academic.oup.com/jrsssb/article/57/1/289/7035855.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-Audio technical report. arXiv:2407.10759, 2024. URL https://arxiv.org/abs/2407.10759.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. arXiv:2110.14168, 2021. URL https://arxiv.org/abs/2110.14168.

Elfsong. MusicAI: Background music × audio LLM benchmark results. Hugging Face dataset, released September 13, 2026, 2026. URL https://huggingface.co/datasets/Elfsong/musicai-background-music-audio-llm-benchmark. Revision f0a72b8a5e5154ca8f897275c7562708c8f66713.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. arXiv:2009.03300, 2020. URL https://arxiv.org/abs/2009.03300.

Yuri Kuratov et al. BABILong: Testing the Limits of LLMs with Long Context Reasoning-in-a-Haystack. In *Advances in Neural Information Processing Systems*, 2024.

Iman Mirzadeh et al. GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models. https://github.com/apple-aiml-research/ml-gsm-symbolic, 2024.

Mistral AI. Voxtral. Model release and technical overview, 2025. URL https://mistral.ai/news/voxtral.

Music Technology Group. MTG-Jamendo: Music classification annotations. Versioned repository, accessed September 15, 2026, 2026. URL https://github.com/MTG/mtg-jamendo-dataset/tree/cafd8e20c265ed84f1e61f1c875327971f43a62f/derived/music-classification-annotations.

NVIDIA. Nemotron model card. https://huggingface.co/nvidia/Nemotron-3-Nano-Omni-30B-A3B-Reasoning-BF16, 2026. Pinned checkpoint e5e9932441de940c9a62185c870ea5bcd4cd24e2; accessed September 16, 2026.

OpenBMB. MiniCPM-o 4.5 model card. https://huggingface.co/openbmb/MiniCPM-o-4_5, 2026. Pinned checkpoint 503e754207c94da6bb26850b4469f367c9ea3582; accessed September 16, 2026.

Simone M. Ritter and Sam Ferguson. Happy creativity: Listening to happy music facilitates divergent thinking. *PLOS ONE*, 12(9):e0182210, 2017. doi: 10.1371/journal.pone.0182210. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0182210.

StepFun. Step-Audio-R1.1 model card. https://huggingface.co/stepfun-ai/Step-Audio-R1.1, 2026. Pinned checkpoint 8abb296c09123345b09de57dbc2830b6a12134b1; accessed September 16, 2026.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. Challenging BIG-Bench tasks and whether chain-of-thought can solve them. arXiv:2210.09261, 2022. URL https://arxiv.org/abs/2210.09261.

Emma Threadgold, John E. Marsh, Neil McLatchie, and Linden J. Ball. Background music stints creativity: Evidence from compound remote associate tasks. *Applied Cognitive Psychology*, 33(5):873–888, 2019. doi: 10.1002/acp.3532. URL https://onlinelibrary.wiley.com/doi/full/10.1002/acp.3532.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-Omni technical report. arXiv:2503.20215, 2025a. URL https://arxiv.org/abs/2503.20215.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, et al. Qwen3-Omni technical report. arXiv:2509.17765, 2025b. URL https://arxiv.org/abs/2509.17765.

Yibo Zhang, Kaiwen Luo, Liang Lin, Shilinlu Yan, Jin Wang, Yaoqi Guo, Yitian Chen, Yalan Qin, Zhenhong Zhou, Kun Wang, and Li Sun. RSA-Bench: Benchmarking audio large models in real-world acoustic scenarios. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 18350–18374. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.913. URL https://aclanthology.org/2026.findings-acl.913/.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. arXiv:2311.07911, 2023. URL https://arxiv.org/abs/2311.07911.

A MUSIC-TYPE SUMMARY AND BASELINE SENSITIVITY

Table 1 compares music types using mixed-audio scores directly. Let $A_{m,t,r}$ be the percentage score for model m, evaluation setting t, and recording r. Define the music reference and genre difference as

$$\bar{A}_{m,t} = \frac{1}{55} \sum_{r=1}^{55} A_{m,t,r}, \tag{1}$$

$$D_{m,t,g} = \frac{1}{5} \sum_{r \in g} A_{m,t,r} - \bar{A}_{m,t}. \tag{2}$$

The relative score is $\frac{1}{80} \sum_{m,t} D_{m,t,g}$. Each genre has five recordings, and each of eight models and ten settings receives equal weight, regardless of item count. The model count is $\sum_m \mathbf{1}[\frac{1}{10} \sum_t D_{m,t,g} > 0]$; the setting count is $\sum_t \mathbf{1}[\frac{1}{8} \sum_m D_{m,t,g} > 0]$. These counts summarize marginal directions, not numbers of significant effects or successes in all individual cells. Ten settings include variants from the same benchmark families and are not ten independent replications.

White-noise reference. White noise is a separate control, excluded from the 55-recording reference mean. Its score is averaged over the 11 corresponding genre-run controls within each model and setting before subtracting $\bar{A}_{m,t}$. It averages -0.729 pp relative to music. Against matched clean speech, music and white noise average -0.406 and -1.134 pp, respectively. The clean baselines have identical aggregate weights in these two summaries.

Why not rank genres by their clean-relative deltas? The matrices retain the original matched-clean comparisons. In 37 of 80 settings, clean scores vary across genre runs; deterministic decoding does not guarantee identical outputs under different numerical execution or batching conditions. Subtracting these different clean scores can change the apparent genre ranking. For Step-Audio on IFEval, electronic dance and Chinese traditional have mixed scores of 59.88% and 58.28%, a 1.60 pp difference. Their clean scores are 57.80% and 60.60%; the corresponding clean-relative deltas differ by 4.40 pp. The table therefore uses a shared music reference within each model/setting. This removes the arithmetic contribution of different clean baselines, but does not remove possible confounding in the mixed-audio runs themselves.

Scope of the finding. Chinese traditional, hip-hop, and ambient retain positive aggregate differences, and rock and metal retain negative differences, when any one model or any one setting is omitted. This is a sensitivity check on the observed sample, not held-out validation. There are only five curated recordings per genre, all using shared task items. Genre labels do not isolate tempo, spectrum, dynamics, or emotion. The provenance, output-budget, and grading limitations recorded in Appendix D also remain. The table reports descriptive effects and direction counts without significance claims; it does not identify a universally best genre or establish human-like musical preferences.

Reproducibility. `scripts/build_genre_summary.py` reads the bundled per-track scores and controls, verifies all 4,400 music–clean matrix entries and 80 white-noise summaries, and generates the table. Full-precision values and the 960 type/model/setting records are in `data/derived/genre_relative_summary.csv` and `data/derived/genre_relative_by_setting.csv`; source hashes are in `data/audit/genre_relative_summary.json`.

B ADDITIONAL BENCHMARK PROTOCOL

Audio version `benchmark-extension-20260916-v1` uses the same pinned TTS model, Aiden voice, mixing implementation, $+10$ dB SNR, and 55 recordings as the regenerated original-task audio. Six settings comprise GSM-Symbolic base/P1/P2 (500 items each) and bAbI short/near/far (300 each). Voxtral Small, Step-Audio-R1.1, Nemotron, and MiniCPM-o retain their original-task closed-answer output caps; Phi-4 Multimodal, Voxtral Mini, Qwen3-Omni, and Qwen2.5-Omni use the documented 1,024-token cap. The regenerated 2,400 items produce 136,800 unique FLAC files and 184,800 condition records per model. No waveform is cropped to meet model limits.

Documented TTS retry. One unconsumed far-context item produced a 268.56-second waveform, exceeding the shared 100-second limit. A single-item retry using the same pinned TTS model, Aiden voice, and unchanged text produced 59.20 seconds (seed 20260924). The original waveform is retained; no crop, speed change, or item removal is used. All eight models receive the same replacement. The other 2,399 TTS items are unchanged. The amendment, original and replacement hashes, and effective manifest are archived with the results.

Step execution scheduling. After the other models completed, the remaining Step-Audio batches were distributed across eight two-GPU replicas on 16 H100 GPUs. The original two logical condition shards, four concurrent seven-condition groups per replica, pinned weights, decoding settings, and 4,096-token cap were retained. Only API endpoints and batch placement changed. Completed outputs and in-flight batches were preserved. Exclusive batch claims and archive hashes prevent duplicate inclusion; per-batch dispatch records identify the worker, GPUs, and endpoint. Dynamic server batching is not a guarantee of bitwise identical greedy outputs across schedules.

Added settings for the first four models. Phi-4 Multimodal, Voxtral Mini, Qwen3-Omni, and Qwen2.5-Omni receive the same regenerated waveforms as the other four models. Their original checkpoint commits were not pinned; these runs pin new revisions and are not a bitwise replay of the earlier runtime. Native Transformers runs use BF16, FlashAttention 2, greedy decoding, seed 20260911, and complete seven-condition batches. A six-setting pilot exposed truncated Voxtral mathematical answers at 64 tokens, so a 1,024-token cap was fixed for all four models before full evaluation. Every generated sequence retains its first-64-token decode and stop status. The strict 64-token audit treats truncation as unsuccessful and is not a reproduction of the original published scoring. Empty and truncated full responses remain in the primary denominator. Runtime versions, wrapper text, checkpoint hashes, and scheduling records are bundled with the results.

Selection. Selection seed 20260916 chooses ten instances per each of the 50 original GSM8K template IDs shared by all three GSM-Symbolic releases. Instances across variants do not necessarily share numerical values. For bAbI qa1/qa2/qa3, 100 problems per task are sampled from the official BABILong 0k data among problems of at most 100 words including the question. The short subset therefore does not represent all original problem lengths. For near/far, complete source-book sentences totaling at least 60 words are placed before or after the intact facts; the question remains last. Sentences containing the bAbI entities or locations are excluded. Near/far word multisets match exactly; spoken durations are not assumed identical. None of these choices use model outcomes.

Objective grading. GSM uses the last numeric token with decimal, fraction, comma, and currency normalization. Location answers use the final nonempty output line, extract a unique location label, remove the later location named in a qa3 question, and reject explicit negation or uncertainty. These fixed rules are a documented short-audio adaptation, not a semantic LLM judge. Parsing failures and truncation rates are retained; every condition remains in its full denominator. Reference answers, code hashes, and selection indices are archived. These statistics are descriptive; future inference should resample GSM by template and bAbI by original paired problem.

Answer extraction and output budgets. Manual spot-checks of these outputs found occasional numeric-extraction errors. For example, an answer stating "288 emails in a 3-day workweek" is parsed as 3 under the last-number rule; an answer describing a cost of 38 more than 7 croissants is parsed as 7. The primary tables retain the fixed grader. Mathematical score shifts should therefore

be read as effects under that rule, and sensitivity to answer format remains to be assessed. Output budgets also matter: MiniCPM's music-condition truncation rates on GSM-Symbolic base/P1/P2 are approximately 89.00%, 97.81%, and 97.71%. These near-floor scores cannot be interpreted as pure mathematical ability. In the Voxtral Mini run, the strict 64-token prefix diagnostic gives music effects of $+0.45$, $+0.32$, and $+0.29$ pp, compared with $+7.83$, $+6.30$, and $+2.63$ pp under the 1,024-token primary protocol. This is a diagnostic of the same generated responses, not an independent rerun or a causal test of reasoning effort.

Table 2: Added benchmark means across eight models (pp).

Model	Variant	Music	White noise
Phi-4 Multimodal	GSM-Symbolic	-0.22	+0.20
Phi-4 Multimodal	GSM +1	-1.04	-1.80
Phi-4 Multimodal	GSM +2	+0.77	+1.60
Phi-4 Multimodal	bAbI short	+0.92	+1.00
Phi-4 Multimodal	bAbI near	+0.79	-1.33
Phi-4 Multimodal	bAbI far	+1.67	+0.33
Voxtral Mini	GSM-Symbolic	+7.83	+7.20
Voxtral Mini	GSM +1	+6.30	+4.00
Voxtral Mini	GSM +2	+2.63	+2.60
Voxtral Mini	bAbI short	+3.05	+1.00
Voxtral Mini	bAbI near	-1.61	-2.67
Voxtral Mini	bAbI far	-0.08	-2.00
Qwen3-Omni	GSM-Symbolic	-3.01	-4.60
Qwen3-Omni	GSM +1	-5.17	-6.80
Qwen3-Omni	GSM +2	-1.71	-1.40
Qwen3-Omni	bAbI short	-2.36	-3.33
Qwen3-Omni	bAbI near	+0.52	-2.67
Qwen3-Omni	bAbI far	-4.01	-9.00
Qwen2.5-Omni	GSM-Symbolic	+0.16	+0.60
Qwen2.5-Omni	GSM +1	-0.73	-0.80
Qwen2.5-Omni	GSM +2	-0.80	-0.60
Qwen2.5-Omni	bAbI short	+1.07	-0.67
Qwen2.5-Omni	bAbI near	-3.82	-4.00
Qwen2.5-Omni	bAbI far	-1.54	-3.00
Voxtral Small	GSM-Symbolic	+0.94	-0.16
Voxtral Small	GSM +1	-3.61	-3.02
Voxtral Small	GSM +2	+0.17	+1.38
Voxtral Small	bAbI short	-1.61	-2.67
Voxtral Small	bAbI near	-2.71	-0.61
Voxtral Small	bAbI far	+1.40	+1.85
Step-Audio	GSM-Symbolic	-0.96	-1.56
Step-Audio	GSM +1	+0.42	+0.38
Step-Audio	GSM +2	+0.47	+0.09
Step-Audio	bAbI short	+0.90	-0.21
Step-Audio	bAbI near	+1.09	+1.03
Step-Audio	bAbI far	-0.05	-0.06
Nemotron	GSM-Symbolic	+0.52	+0.33
Nemotron	GSM +1	-0.34	-0.05
Nemotron	GSM +2	-0.54	-0.05
Nemotron	bAbI short	-2.65	-2.97
Nemotron	bAbI near	-7.65	-7.85
Nemotron	bAbI far	+0.06	-2.67
MiniCPM	GSM-Symbolic	+2.35	+2.60
MiniCPM	GSM +1	-0.16	-0.40
MiniCPM	GSM +2	+0.04	+0.00
MiniCPM	bAbI short	+1.68	+2.67
MiniCPM	bAbI near	-2.16	-7.00
MiniCPM	bAbI far	-3.85	-7.67

C Detailed findings by task category

The original-task matrix compares every track with matched clean speech for all eight models. Its shared color scale shows effect sizes, not statistical significance. Table 6 summarizes the same 32 model–task settings. The models differ in scale, training, and runtime configuration, so the comparison across them is descriptive and does not identify a model-size effect.

C.1 The direction depends on the model and task

Among the 32 model–task settings, 13 average changes are positive, 17 are negative, and 2 are near zero (less than $0.005\,\mathrm{pp}$ in magnitude). On GSM8K, Voxtral Small changes from 61.55% to 53.73%, an effect of $-7.82\,\mathrm{pp}$; 55/55 track estimates are negative. Voxtral Mini and the historical Qwen2.5 configuration instead have positive GSM8K means ($+0.86$ and $+1.48\,\mathrm{pp}$), and every GSM8K interval in Table 11 includes zero. A general claim that music helps mathematical reasoning is therefore unsupported.

The GSM8K–BBH contrast remains a useful hypothesis generator, not an explanation. BBH contains heterogeneous reasoning subtasks; benchmark identity does not isolate difficulty or dependency horizon. Neither a positive math score shift nor a negative BBH shift measures inspiration or distraction.

C.2 Equal-SNR music and noise are not interchangeable

The largest absolute music–noise gap occurs for Qwen3-Omni/BBH: music changes the score by $-1.93\,\mathrm{pp}$, versus $-6.60\,\mathrm{pp}$ for white noise (difference $+4.67\,\mathrm{pp}$). The gap reverses sign elsewhere, as for Voxtral Small/BBH (-1.56 versus $+0.91\,\mathrm{pp}$). Matching overall background energy does not match spectral or temporal masking. Noise is useful for describing acoustic sensitivity, but outperforming noise is not the objective or evidence of a human-like musical mechanism. These highlighted settings are selected descriptively, not by a pre-registered significance test.

C.3 A net score shift can hide opposing changes

For Step-Audio-R1.1/IFEval, 8.18% of music–clean pairs change from success to failure, while 8.24% change from failure to success. Thus 16.43% of paired scores flip, but the net effect is $+0.06\,\mathrm{pp}$. The same model on BBH has 5.94% in each direction: 11.88% of scores flip and the net effect is exactly zero. Qwen3/IFEval provides another example: 93.80% of response strings change, 10.73% of Strict scores flip, and the mean effect is $+0.06\,\mathrm{pp}$. Changes in wording do not necessarily imply new ideas. We report observed responses and score transitions; semantic or creative effects require dedicated assessments.

C.4 Track variation does not establish a best genre

The widest within-setting track range is Voxtral Small/GSM8K, from -14.60 to $-1.60\,\mathrm{pp}$. The complete matrices retain that variation, including unfavorable and near-zero estimates. Each genre has only five fixed recordings, and all tracks share the same task items. A selected high-scoring track is not independent evidence of a genre advantage. Genre-label permutation tests are available for the 16 settings in Table 12 and yield no discoveries after correction. Tempo, texture, and arousal remain candidate descriptors for future analyses, not established explanations here.

D SPEECH REGENERATION, OUTPUT BUDGETS, AND SCORING PROVENANCE

Voxtral Small, Step-Audio-R1.1, Nemotron, and MiniCPM-o were evaluated on the four original tasks after the first four models, using audio version `audio-regenerated-20260916-v1`. The original code and music were recovered from dataset revision `e8fd5d9300d2ba83aae16983c640ee180f35cb2b`; the complete original 2,000-item speech and mixture caches were absent. We regenerated the speech rather than claiming to recover the earlier waveforms. This section records the settings that differ between the two groups of runs; where a later appendix gives per-item response diagnostics, bootstrap intervals, or permutation tests, those derive from the first four runs, whose per-item records are bundled.

Waveforms and controls. Qwen3-TTS 12Hz 0.6B CustomVoice (revision `85e237c12c027371202489a0ec509ded67b5e4b5`) synthesizes English with voice Aiden. The recovered 16-shard, batch-8 ordering and seeds $20260911 + \text{shard}$ are preserved. An eight-item comparison between the recovered generator and the resumable wrapper gives identical WAV hashes. TTS produces 24-kHz mono PCM16; mixtures use 16-kHz mono PCM16 FLAC, speech RMS -22 dBFS, and $+10$ dB speech-to-background SNR. The fixed music clips are tiled/cropped from their beginnings. White-noise seeds use original global item indices. A mixture above peak 0.98 is scaled as a whole, preserving SNR. This is a digital-level protocol, not calibrated sound pressure.

Both workers independently verify 114,000 FLAC files and identical complete manifests. There are 154,000 condition records per model: each item has 55 music conditions, 11 genre-local clean records, and 11 noise records. Repeated control references remain separate inference conditions. All 250 immutable input batches, 2,000 TTS WAVs, and 256 RNG checkpoints are archived with hashes. No transcripts or labels are sent to the evaluated models.

Model	Closed cap	IFEval cap	Output mode
Voxtral Small	1,024	768	Text
Step-Audio-R1.1	4,096	4,096	Reasoning plus final
Nemotron	64	768	Thinking disabled
MiniCPM-o 4.5	64	768	Thinking/audio output disabled

Table 3: Original-task token budgets for the four later runs. Closed means GSM8K, BBH, and MMLU. Each model uses the same cap for every background within a task; caps are not matched across models.

Budget revision and remaining truncation. The initial Voxtral Small 64-token closed-task cap frequently truncated both clean and background responses. An independent first-batch diagnostic covering eight items and seven complete conditions per item yields 36 complete/20 truncated outputs at 64 tokens, versus 56 complete/0 truncated at 1,024. No answer labels were used to choose the cap. We restarted the entire Voxtral Small evaluation in a separate output directory at 1,024; the incomplete 64-token run is retained only as a diagnostic. Completed low-budget answers are not reused or selected into the primary run. This finite-cap diagnostic does not establish that every full-run answer will terminate.

Native input paths. Voxtral Small receives audio followed by the fixed text wrapper in a user message. Nemotron and MiniCPM receive a fixed system wrapper and user audio. Step receives its fixed user wrapper followed by consecutive 25-second audio chunks, without cropping. The closed wrapper requests only the final answer; IFEval requests compliance with all spoken instructions. API requests use greedy decoding with seed 20260911. Native MiniCPM uses seed 0 and greedy decoding. Native MiniCPM retains every answer in an actual seven-input batch; its published audio encoder/chunk mask is unchanged, with per-row embedding scatter and full text-output capture. A single-input regression gives identical generated tokens to the unmodified embedding path.

Voxtral Small and Nemotron use vLLM 0.20.0+cu129 in BF16. Step uses its fixed official vLLM fork (commit `484a5a5a5b478858ce8b091fbfe201c0fbaae08f`), KV caching, and decode CUDA graphs. MiniCPM uses Transformers 4.51.0 and BF16. API groups contain clean, noise, and five same-genre tracks, but the servers control physical batching; physical API batch

membership is not asserted. Dynamic GPU reassignment preserves whole-group shard membership and records the serving worker, configuration, and unique shard claim.

Scoring and interpretation. The recovered per-item scoring functions grade normally completed answers. Scoring version `ifeval-symbols-v2` corrects two IFEval items (1122 and 1129) in these four runs: explicit # and ! frequency targets are counted literally, replacing the original checker's random alphabetic fallback. We retain all 500 items and the original Strict/Loose rules, and rescore saved outputs without new inference. The earlier four runs keep their published scores and are not retroactively corrected. Truncated or empty final outputs are unsuccessful under the full expected denominator. Missing or transport-failed conditions suppress publication rather than being silently dropped. Mean effects and score transitions are calculated from all 500 paired items per track. The matrices are descriptive; no new bootstrap, McNemar, or genre-permutation result is implied. Differences in voice realization, output budgets, scoring provenance, and runtime mean that comparisons across the eight models do not identify a size-scaling effect.

Model	Total conditions	Truncated (%)
Voxtral Small	154,000	1.50
Step-Audio-R1.1	154,000	3.15
Nemotron	154,000	1.76
MiniCPM-o 4.5	154,000	2.82

Table 4: Residual truncation in the four later original-task runs, included as unsuccessful outputs in the primary metric. Per-task/condition rates and paired truncation differences are bundled with the report.

Pinned models.

https://huggingface.co/mistralai/Voxtral-Small-24B-2507
`da5b42409f279fdd92febee0511a6c32828569c1`

https://huggingface.co/stepfun-ai/Step-Audio-R1.1
`8abb296c09123345b09de57dbc2830b6a12134b1`

https://huggingface.co/nvidia/Nemotron-3-Nano-Omni-30B-A3B-Reasoning-BF16
`e5e9932441de940c9a62185c870ea5bcd4cd24e2`

https://huggingface.co/openbmb/MiniCPM-o-4_5
`503e754207c94da6bb26850b4469f367c9ea3582`

E CONFIGURATION AND EFFECT DEFINITIONS

E.1 TASKS, MODELS, AND EVALUATION UNITS

Each benchmark contributes 500 items, for 2,000 distinct tasks. The task metadata contain item identifiers, source splits, task text, reference answers or instruction constraints, and the text used for speech synthesis. GSM8K uses final-number accuracy, BBH uses normalized-answer accuracy, MMLU uses option accuracy, and IFEval uses prompt-level Strict compliance. IFEval Loose is auxiliary and is excluded from the primary analyses. The adapted spoken prompts request only the final answer on the three closed-answer tasks; MMLU options are read aloud. Example prompt forms appear in Appendix J.

The report covers eight models: Phi-4 Multimodal, Voxtral Mini 3B, Qwen3-Omni 30B-A3B, Qwen2.5-Omni 7B, Voxtral Small 24B, Step-Audio-R1.1, Nemotron 3 Nano Omni, and MiniCPM-o 4.5. The Qwen2.5-Omni 7B run is historical: it uses the same task selection, track collection, and scoring protocol, and is retained as a labeled reference rather than a synchronized rerun on verified byte-identical audio. A Qwen2-Audio 7B run is available but is excluded from music-effect interpretation for the reason given below. Table 5 reports the clean scores of the eight interpretable configurations.

Table 5: Clean task scores (%). Each score averages the genre-local clean records. The historical model is marked H. These are scores on the fixed spoken subsets.

Model	GSM8K	BBH	MMLU	IFEval
Phi-4 Multimodal	10.96	32.20	47.20	37.53
Voxtral Mini 3B	6.80	39.07	37.20	29.20
Qwen3-Omni	39.40	54.20	86.60	73.04
Qwen2.5-Omni (H)	21.60	34.60	78.60	44.00
Voxtral Small 24B	61.55	46.62	83.51	44.44
Step-Audio-R1.1	92.16	74.33	89.35	59.51
Nemotron	24.13	44.87	70.85	65.05
MiniCPM-o 4.5	25.60	34.40	76.60	63.40

The Qwen2-Audio run is excluded from music-effect interpretation. Its 154,000 published outputs contain only two fixed refusal or clarification strings: one across the closed-answer tasks and another on IFEval. Its zero closed-answer accuracy and incidental 8.2% Strict compliance do not establish music robustness or the general capability of that model family. The first publication contains 770,000 rows, of which 616,000 belong to its four interpretable configurations; the remaining four models were evaluated on the same tasks afterwards (Appendix D).

E.2 SPEECH AND MUSICAL BACKGROUNDS

The task speech is generated with Qwen3-TTS-12Hz-0.6B-CustomVoice using the English Aiden voice. The protocol specifies 16 kHz mono audio and a speech RMS target of -22 dBFS. The music collection comprises five recordings in each of 11 categories: acoustic folk/country, ambient, Chinese traditional, cinematic/orchestral, classical, electronic/dance, instrumental hip-hop, metal, pop, rock, and jazz. All are designated instrumental in the release. The 55-recording catalog is provided in Appendix N.

The prepared music excerpts are 30 seconds long. The protocol describes cropping or looping the background to the speech duration, with channel conversion, resampling, and gain adjustment. Thus, 30 seconds describes the prepared excerpt, not a constant model-input duration. Across the audited Phi-4 records, input durations range from 4.00 to 94.24 seconds. For speech s_i and an aligned background b_j, the intended mixture is

$$x_{ij} = s_i + \alpha_{ij} b_j, \qquad \alpha_{ij} = \frac{\mathrm{RMS}(s_i)}{\mathrm{RMS}(b_j)\, 10^{r/20}}, \qquad r = 10 \text{ dB}. \tag{3}$$

The clean condition contains s_i alone; white noise uses the same target SNR. Music metadata record prepared-excerpt RMS levels of -24 dBFS for all 55 tracks. These levels precede mixing: if speech

meets -22 dBFS and there is no common gain change after mixing, a $+10$ dB SNR implies background RMS of -32 dBFS. This is a conditional calculation, not a waveform measurement. dBFS and SNR are digital-signal quantities, not calibrated physical sound-pressure levels.

E.3 Pairing and execution

For each item and each genre, a comparison group contains one clean record, one white-noise record, and five music records. Eleven groups give 77 rows per item, comprising 55 music rows and 11 repetitions of each control. The current protocol reports 57 unique waveforms per item because the control waveforms are reused across genres. The release specifies deterministic generation, confirmed by the experimenter as greedy decoding. Exact model revision hashes, all inference-wrapper settings, and the original inference scripts are not included in the inspected release, so full execution identity cannot be reconstructed from it alone.

We retain the genre-local clean pairing throughout. Denote the binary score for item i, model m, and music track j by Y^M_{imj}, and its corresponding clean score by $Y^C_{im,g(j)}$. The track effect, in percentage points, is

$$\widehat{\Delta}_{mj} = \frac{100}{N} \sum_{i=1}^{N} \left(Y^M_{imj} - Y^C_{im,g(j)} \right), \qquad N = 500. \tag{4}$$

Genre effects average their five fixed tracks, and overall music effects average all 55 tracks equally. Noise effects average the 11 matched noise–clean comparisons. Replacing these controls with a single chosen clean response would change the comparison when repeated clean outputs differ.

E.4 Uncertainty, response changes, and record audits

For the four models with bundled per-item records we preserve the reported point estimates and nominal 95% intervals, computed as 10,000-bootstrap intervals. To check the aggregate analysis independently, we reconstruct per-item means and bootstrap 500 item clusters with replacement, keeping all within-item conditions together (10,000 draws; seed 20260915). These audit intervals are separately labeled in Appendix L; they do not silently replace the published intervals. This inference conditions on the selected 55 tracks. Treating the 27,500 item–track pairs per model–task as independent observations would understate dependence.

Track-level tests use the reported exact McNemar tests and Benjamini–Hochberg (BH) adjustment within each model–task family of 55 comparisons (Benjamini & Hochberg, 1995). We also report a sensitivity calculation adjusting all 880 track comparisons together. Genre specificity uses the reported 100,000 track-label permutations, followed here by BH adjustment across the 16 model–task tests. These are reported analysis families; the inspected materials do not establish preregistration. The aggregate 95% intervals are not jointly adjusted across 16 comparisons, and interval overlap with zero is not evidence of equivalence.

For each music–clean pair, we compute exact response-string disagreement and binary-score disagreement. Let H be the fraction changing from correct/pass to incorrect/fail and B the fraction changing in the opposite direction. Then

$$\text{score-flip rate} = H + B, \qquad \widehat{\Delta} = 100(B - H). \tag{5}$$

String changes may reflect punctuation, formatting, or wording rather than meaning. We additionally audit the existing 11 clean records per item, comparing all $\binom{11}{2} = 55$ pairs. We check waveform hashes and input text before summarizing disagreement. These pairs span recorded genre-associated batches and are diagnostics of the stored evaluation, not a newly randomized repeat experiment.

F SUPPLEMENTAL RESULT SUMMARIES

Table 6 gives all aggregate estimates underlying the main-text comparisons, for the eight models on the four original tasks. Figure 4 averages the track matrix by genre and Figure 5 separates response-string changes from score transitions; both cover the four runs whose per-item response records are bundled here. Every value is reproduced from the bundled score tables.

Table 6: Aggregate original-task results for all eight models. Clean and Music are percentage scores; Δ music and Δ noise are changes from the matched clean controls, in percentage points. Music averages the 55 fixed tracks and noise averages the 11 matched noise–clean comparisons. Score flips give the percentage of music–clean item pairs whose binary score changes in either direction. H denotes the historical model.

Model	Task	Clean	Music	Δ music	Δ noise	Score flips
Phi-4 Multimodal	GSM8K	10.96	11.09	+0.12	+1.36	3.61
	BBH	32.20	32.38	+0.18	-1.13	5.77
	MMLU	47.20	47.49	+0.29	-0.76	5.83
	IFEval Strict	37.53	36.48	-1.05	-1.62	9.19
Voxtral Mini 3B	GSM8K	6.80	7.66	+0.86	+1.27	5.15
	BBH	39.07	35.56	-3.52	-6.07	13.35
	MMLU	37.20	36.44	-0.76	-1.80	6.33
	IFEval Strict	29.20	29.96	+0.76	-1.29	10.52
Qwen3-Omni	GSM8K	39.40	39.34	-0.06	-1.80	6.13
	BBH	54.20	52.27	-1.93	-6.60	9.37
	MMLU	86.60	85.63	-0.97	-1.80	4.50
	IFEval Strict	73.04	73.09	+0.06	-0.87	10.73
Qwen2.5-Omni (H)	GSM8K	21.60	23.08	+1.48	+1.20	5.17
	BBH	34.60	33.79	-0.81	-2.20	7.61
	MMLU	78.60	77.42	-1.18	-2.20	7.44
	IFEval Strict	44.00	43.38	-0.62	-3.00	11.43
Voxtral Small 24B	GSM8K	61.55	53.73	-7.82	-6.15	21.46
	BBH	46.62	45.06	-1.56	+0.91	13.55
	MMLU	83.51	83.51	0.00	-1.09	5.29
	IFEval Strict	44.44	46.05	+1.62	+1.58	13.33
Step-Audio-R1.1	GSM8K	92.16	92.01	-0.15	-0.25	2.90
	BBH	74.33	74.33	0.00	0.00	11.88
	MMLU	89.35	88.61	-0.73	-1.04	4.89
	IFEval Strict	59.51	59.57	+0.06	-0.65	16.43
Nemotron	GSM8K	24.13	24.67	+0.54	-1.31	6.61
	BBH	44.87	46.08	+1.21	+1.69	8.57
	MMLU	70.85	73.03	+2.18	+2.00	8.70
	IFEval Strict	65.05	64.11	-0.95	-0.18	12.73
MiniCPM-o 4.5	GSM8K	25.60	26.71	+1.11	+2.00	8.26
	BBH	34.40	31.30	-3.10	-3.80	11.19
	MMLU	76.60	76.32	-0.28	-2.40	6.71
	IFEval Strict	63.40	62.58	-0.82	-2.00	12.10

The independent item-bootstrap check for Voxtral Mini/BBH yields $[-5.96, -1.03]$ pp, close to the reported $[-6.03, -1.01]$ pp interval. The comparison remains exploratory: the aggregate intervals are not jointly corrected across the 16 settings that have them. Clean baselines also vary substantially (Table 5); for example, Voxtral Mini's GSM8K score is 6.8% and Step-Audio-R1.1's is 92.2%, so a small positive estimate should not be interpreted as a general gain in capability.

Within a genre, point estimates can differ substantially. For Voxtral Mini/BBH, *Finding Home* changes the score from 39.0% to 33.2%, while *01 Ouverture*, another folk/country track, has an effect of -1.0 pp. Across all 1,760 original-task track estimates the range is -14.6 to $+4.8$ pp. These are descriptive, post hoc examples; they do not independently establish reproducible differences between recordings.

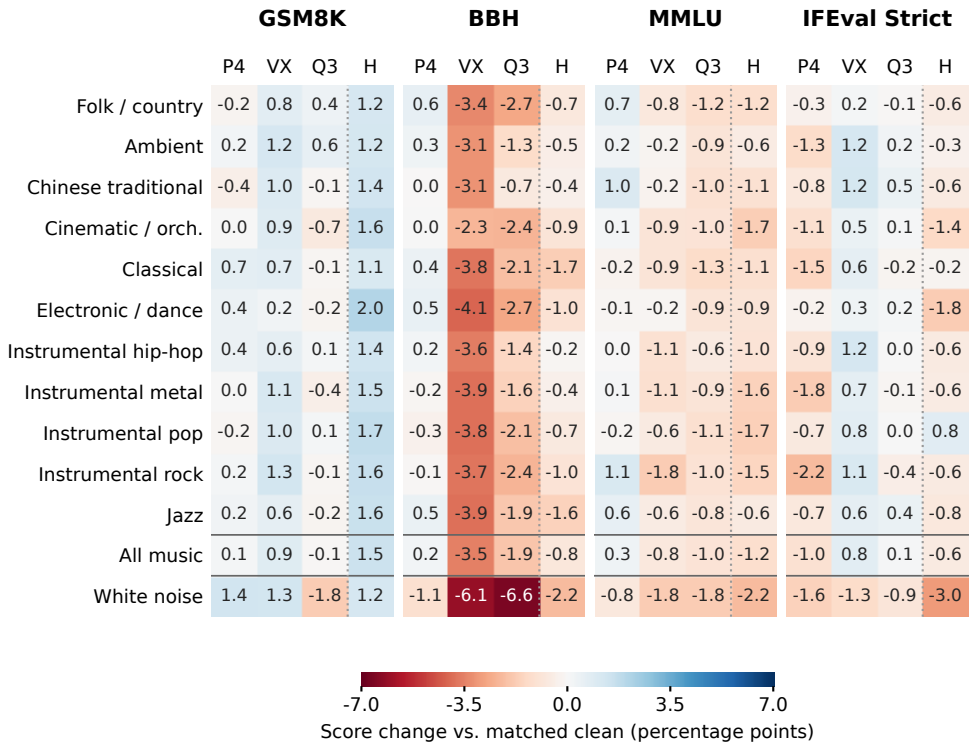


Figure 4: Music effects across tasks and models. Each genre row averages five tracks; the final rows summarize all 55 tracks and white noise. Panels cover the four runs whose per-item response records are bundled. P4: Phi-4 Multimodal; VX: Voxtral Mini 3B; Q3: Qwen3-Omni; H: historical Qwen2.5-Omni. Red indicates a lower score and blue a higher score than the corresponding clean control, in percentage points. Color indicates the estimate, not statistical significance. The same $[-7, +7]$ scale is used in every panel. Track results appear in Figure 1, with larger views in Appendix O.

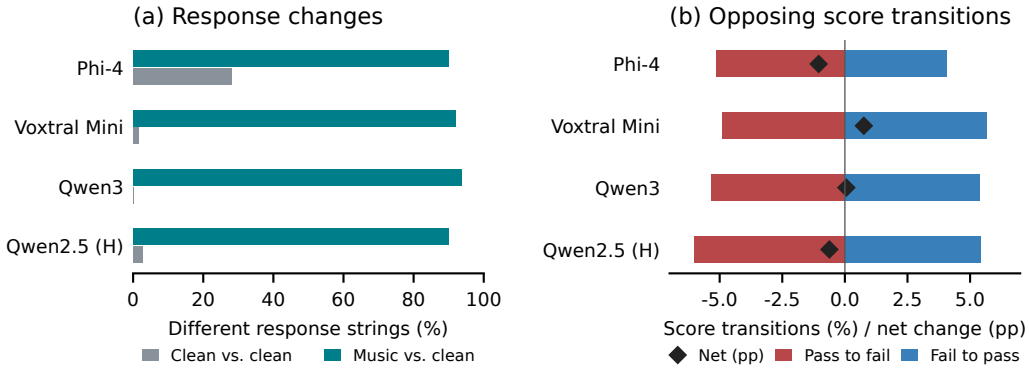


Figure 5: IFEval response and score changes. (a) Music–clean string disagreement compared with the existing clean–clean diagnostic. (b) Pass-to-fail and fail-to-pass proportions; black diamonds show their net difference in percentage points. A small mean score effect can coexist with changes on many pairs. Different strings need not imply different meanings. The clean–clean and music–clean comparisons have different batch organization, so their difference is not automatically an unbiased estimate of a music-only effect.

G TASK CONTRAST AND BBH SUBTASK STRUCTURE

Table 7 separates counts of positive track estimates from the average effect. Five of eight models have a positive mean GSM8K effect and five have a negative mean BBH effect, but neither pattern is uniform: Voxtral Small has 0/55 positive GSM8K estimates, while Nemotron has only 5/55 negative BBH estimates. Counts and magnitudes can also disagree, as for Qwen3, whose GSM8K average is slightly negative although 28/55 track estimates are positive. Every GSM8K average-effect interval in Table 11 includes zero. The tracks share the same 500 items and related clean controls, so a large count of positive estimates is not a count of independent replications. The historical configuration retains its availability caveat.

Table 7: Descriptive GSM8K–BBH contrast for all eight models. Mean effects are in percentage points; Positive counts tracks with a higher GSM8K score than clean speech, and Negative counts tracks with a lower BBH score. Track counts describe signs, not significance. H denotes the historical configuration.

Model	GSM8K effect	Positive	BBH effect	Negative
Phi-4	+0.12	31/55	+0.18	18/55
Voxtral Mini	+0.86	48/55	-3.52	54/55
Qwen3	-0.06	28/55	-1.93	50/55
Qwen2.5 (H)	+1.48	54/55	-0.81	46/55
Voxtral Small	-7.82	0/55	-1.56	51/55
Step-Audio	-0.15	17/55	0.00	30/55
Nemotron	+0.54	41/55	+1.21	5/55
MiniCPM-o	+1.11	47/55	-3.10	55/55

The 500-item BBH subset contains ten subtasks with 50 items each. They include logical deduction, causal judgment, Boolean expressions, navigation, and counting. They are not annotated as uniformly easy or as uniformly requiring long-horizon processing. Table 8 reconstructs each subtask's mean effect from the per-item audit, averaging all 55 fixed tracks within each item; per-item records are bundled for four of the eight models, so the subgroup view covers those four. The equally weighted ten subgroup means reproduce the aggregate BBH effects exactly.

Table 8: Post hoc BBH subtask estimates (pp), with 50 items and 55 fixed tracks per cell, for the four models whose per-item records are bundled. These are descriptive means; no subgroup significance or mechanism claim is made.

Subtask	Phi-4	Voxtral Mini	Qwen3	Qwen2.5 (H)
Boolean expressions	-3.27	-12.11	-8.87	+1.56
Causal judgment	+3.93	+2.91	+1.35	+0.95
Date understanding	+1.31	-1.45	+1.20	+1.75
Disambiguation	0.00	+5.02	-5.71	+0.18
Formal fallacies	-1.20	-8.91	-2.33	-4.15
Logical deduction (3 objects)	-0.36	+1.85	+3.75	+2.00
Navigation	-2.07	-7.82	-3.53	-9.64
Object counting	+0.04	-3.24	-2.95	-2.07
Colored objects	+0.29	-4.80	-4.29	-0.80
Web of lies	+3.13	-6.62	+2.04	+2.07

Negative changes are not uniform within BBH. For Qwen3, Boolean expressions have an estimate of -8.87 pp, while three-object logical deduction has an estimate of $+3.75$ pp. These small subgroups do not establish which task property explains the difference. In particular, benchmark identity also changes content, answer format, baseline accuracy, and speech characteristics. Difficulty and dependency horizon require independent definitions before testing an interaction with music; neither a benchmark-level sign pattern nor a response-string change measures a model's attention.

H Related work

Music and human cognition. Ritter & Ferguson (2017) report that happy, high-arousal music improved divergent creativity relative to silence, with no corresponding benefit on their convergent measure. Threadgold et al. (2019) find impaired compound remote-associate performance under background music, including instrumental music. These results motivate task-dependent hypotheses rather than a presumption of improvement. Our current benchmarks measure answer accuracy and instruction following; they do not directly measure creativity or sustained attention.

Speech-capable language models and acoustic evaluation. Phi-4 Multimodal, Voxtral, Qwen2-Audio, and the Qwen-Omni models extend language-model interaction to audio inputs (Abouelenin et al., 2025; Mistral AI, 2025; Chu et al., 2024; Xu et al., 2025a;b). Evaluating these systems requires controlling the input acoustics as well as the linguistic task. RSA-Bench studies audio-language-model robustness by superimposing environmental soundscapes on clean speech at different interference strengths (Zhang et al., 2026). MusicAI focuses on instrumental recordings, with the same track collection applied across several spoken task families and models. The present design measures an end-to-end audio-background effect; it does not separate perceptual masking from later task processing.

Task measurement. GSM8K supplies mathematical word problems (Cobbe et al., 2021); BBH supplies varied symbolic and reasoning tasks (Suzgun et al., 2022); MMLU supplies multiple-choice knowledge and reasoning questions (Hendrycks et al., 2020); and IFEval supplies verifiable output constraints (Zhou et al., 2023). MusicAI uses fixed 500-item selections and spoken renderings of these materials. Consequently, its absolute scores describe this protocol and are not interchangeable with full-benchmark text-input leaderboard scores.

Model coverage. Voxtral Small, Step-Audio-R1.1, Nemotron 3 Nano Omni, and MiniCPM-o 4.5 are evaluated through their pinned native audio-to-text paths (Mistral AI, 2025; StepFun, 2026; NVIDIA, 2026; OpenBMB, 2026). They add model-family coverage under the same musical conditions; the resulting comparison across the eight models is not a controlled scaling-law experiment.

I DATA PROVENANCE AND REPRODUCIBILITY

The outputs and summary values for Phi-4 Multimodal, Voxtral Mini, Qwen3-Omni, and the historical Qwen2.5-Omni refer to the MusicAI release dated September 13, 2026, at revision `f0a72b8a5e5154ca8f897275c7562708c8f66713`; the remaining four models were evaluated afterwards under the protocol in Appendix D. The dataset is available at https://huggingface.co/datasets/Elfsong/musicai-background-music-audio-llm-benchmark. This project bundles the summary inputs, track metadata, per-item audit metrics, figures, and table-generation code. The larger raw output files can be fetched from the pinned revision with the provided download script. Checksums are verified against the release manifest. The audit of the first four models involved no new inference; the four later models add new inference as described in Appendix D.

The numerical audit reconstructs 96 aggregate values: clean score, music score, music effect, noise effect, music string disagreement, and music score disagreement for each of 16 model–task combinations. All match the bundled release-derived summaries to numerical precision. The 880 track effects are also checked against their track and clean scores. The manifest covers published inputs; separately generated audit files are labeled as such. These checks validate the calculations, not the original scoring implementation.

Recorded levels. All 55 prepared music excerpts have recorded RMS -24.0 dBFS. The speech target is -22 dBFS and the background target is $+10$ dB SNR. The Phi-4 record audit includes 110,000 music rows and 22,000 white-noise rows; every background row records both target and achieved SNR as 10.0 dB. There are 22,000 clean rows. These are checks on metadata, not independent measurements of audio. The distinction matters because digital level, signal-to-noise ratio, and calibrated sound-pressure level are different quantities.

Availability boundaries. Track source links, source licenses, prepared-excerpt offsets, and hashes are retained in the bundled metadata. The project does not redistribute music or generated speech. Different recordings and task sources retain their own licenses. The later runs recover source-level inference and scoring code, pin model revisions, and archive new waveform artifacts; this does not restore the missing original waveforms of the first four runs. The current document uses the official ICLR 2027 typography, with a technical-report running header; it does not claim conference submission or acceptance.

J PROMPT CONSTRUCTION

The task metadata distinguish `source_prompt` from `text`, the latter being the input text for speech synthesis. The closed-answer tasks append answer-format instructions. The following forms are transcribed from the bundled task metadata; bracketed descriptions denote variable task content, not additional model prompts.

Task	Spoken task form
GSM8K	[Word problem.] Give only the final numeric answer, with no explanation.
BBH	[Task question.] Give only the final answer, with no explanation.
MMLU	[Question.] Option A: [Choice.] Option B: [Choice.] Option C: [Choice.] Option D: [Choice.] Answer with only the letter A, B, C, or D.
IFEval	The selected instruction prompt, including its output constraints; the source metadata identify the subset as `audio_suitable`.

Table 9: Representative prompt forms used for speech synthesis. These do not reconstruct the full model-specific input wrapper.

GSM8K, BBH, and MMLU items in the inspected metadata come from their test splits. The IFEval selection is recorded under its train split. The fixed item identifiers and source indices are bundled. The selection is not a new train/test split designed in this report, and no model training or fine-tuning is performed here. Exact original selection filters and inference-time wrapper details should be taken from the original experiment implementation when it is archived.

K EXISTING CLEAN REPETITIONS AND SCORING AUDIT

For the four models whose per-item response records are bundled, we audit 154,000 output rows each and form 2,000 item groups. Every group contains 11 clean records, 11 noise records, and 55 music records. Within every clean group, recorded audio SHA-256, task text, and source prompt are identical. The summary in Table 10 compares all 55 unordered clean pairs per item with the 55 music–clean pairs. Each rate therefore averages 27,500 pairs per task, but the effective item count is 500 and the comparisons are dependent.

Table 10: Exact string disagreement (text) and binary-score disagreement (score), in percent. C–C denotes clean–clean; M–C denotes music–clean. IFEval uses Strict. H is historical. The zero Qwen3 clean string-disagreement rate is exact in these records; its nonzero IFEval clean score-disagreement rate reflects the stored-label inconsistency described below.

Model	Task	C–C text	M–C text	C–C score	M–C score
Phi-4 Multimodal	GSM8K	4.20	27.17	0.96	3.61
	BBH	2.43	24.41	0.44	5.77
	MMLU	1.06	8.88	0.95	5.83
	IFEval Strict	28.18	90.09	2.40	9.19
Voxtral Mini 3B	GSM8K	0.95	50.05	0.00	5.15
	BBH	0.63	31.59	0.10	13.35
	MMLU	0.00	9.49	0.00	6.33
	IFEval Strict	1.64	92.11	0.22	10.52
Qwen3-Omni	GSM8K	0.00	25.49	0.00	6.13
	BBH	0.00	21.43	0.00	9.37
	MMLU	0.00	5.71	0.00	4.50
	IFEval Strict	0.00	93.80	0.07	10.73
Qwen2.5-Omni (H)	GSM8K	0.00	31.67	0.00	5.17
	BBH	0.11	27.55	0.00	7.61
	MMLU	0.00	8.89	0.00	7.44
	IFEval Strict	2.71	89.98	0.22	11.43

The proportion of items with *any* change among repetitions is a different statistic from a pairwise disagreement rate. For Phi-4/IFEval, 261/500 items (52.2%) have any clean string disagreement, whereas the pairwise rate is 28.18%. Twenty-two items (4.4%) have any clean Strict-score disagreement, whereas its pairwise rate is 2.40%. The report uses pairwise rates consistently when comparing these diagnostics.

Interpreting clean variability. Despite identical recorded clean waveform hashes and task inputs, IFEval clean–clean string disagreement is 28.18% for Phi-4, 1.64% for Voxtral Mini, 0% for Qwen3, and 2.71% for the historical baseline. Greedy decoding alone therefore does not establish agreement of these stored repetitions. The inspected release cannot identify whether differences arose from batching, numerical behavior, unrecorded settings, or another implementation detail. Music–clean and clean–clean comparisons have different batch organization, so subtracting their rates is not automatically an unbiased music-only effect estimate.

An identical-response grading inconsistency. For Qwen3 item `ifeval_1122`, the eleven clean responses are exactly identical, and the recorded grading configuration is identical. The instruction IDs are `change_case:english_lowercase` and `keywords:letter_frequency`. The latter requests at least four occurrences of the symbol `#`. The stored response contains four such symbols. The first constraint is marked true in all repetitions, but the second is true in two and false in nine. These labels produce $2 \times 9 = 18$ disagreeing clean pairs, or $18/27{,}500 = 0.06545\%$ for the task. The relevant eleven records are saved in the audit evidence. This identifies a consistency problem in the stored evaluation; it does not identify its implementation-level cause. We do not replace the labels with an unverified custom regrader.

L BOOTSTRAP AND MULTIPLICITY DETAILS

The audit resamples the 500 per-item average effects for each task and model, retaining all 55 fixed tracks within each sampled item. It uses a NumPy random generator initialized with seed

20260915 for each model–task, 10,000 draws, and percentile endpoints at 2.5% and 97.5%. Table 11 contrasts these independently computed intervals with the release's intervals. The point estimates agree exactly; finite-draw bootstrap endpoints differ slightly. Neither set is an interval over arbitrary tracks or a simultaneous interval over all model–task combinations.

Table 11: Music-effect 95% intervals in percentage points for the four models with bundled per-item records: reported values and independently reconstructed item-bootstrap values. The main text and heatmaps preserve the reported point estimates.

Model	Task	Reported CI	Item-bootstrap audit CI
Phi-4 Multimodal	GSM8K	[-1.08, +1.30]	[-1.07, +1.32]
	BBH	[-1.36, +1.70]	[-1.32, +1.68]
	MMLU	[-1.17, +1.75]	[-1.16, +1.76]
	IFEval Strict	[-2.75, +0.66]	[-2.81, +0.64]
Voxtral Mini 3B	GSM8K	[-0.66, +2.35]	[-0.65, +2.40]
	BBH	[-6.03, -1.01]	[-5.96, -1.03]
	MMLU	[-2.49, +0.97]	[-2.48, +0.99]
	IFEval Strict	[-1.39, +2.88]	[-1.41, +2.87]
Qwen3-Omni	GSM8K	[-1.63, +1.56]	[-1.66, +1.56]
	BBH	[-3.99, +0.09]	[-4.00, +0.11]
	MMLU	[-2.51, +0.46]	[-2.47, +0.51]
	IFEval Strict	[-2.13, +2.29]	[-2.15, +2.29]
Qwen2.5-Omni (H)	GSM8K	[-0.10, +3.09]	[-0.08, +3.10]
	BBH	[-2.67, +1.00]	[-2.64, +0.97]
	MMLU	[-3.08, +0.66]	[-3.09, +0.71]
	IFEval Strict	[-2.78, +1.64]	[-2.76, +1.59]

For each fixed track, the reported McNemar test uses discordant paired binary outcomes. It adjusts 55 track tests within each model–task family. Exactly 23 negative track effects pass $q < 0.05$, all in Voxtral Mini/BBH; no positive track effect passes. Treating all 880 primary track tests as one family gives a minimum BH-adjusted value of approximately 0.3715, with no rejections at 0.05. The wider correction is a sensitivity analysis, not evidence that one family definition is universally appropriate. A confirmatory follow-up should state its comparison family before observing new outcomes.

Table 12: Genre-label permutation p values and the added BH correction across 16 model–task tests, for the four models whose track-label permutations are bundled. No adjusted value is below 0.05.

Model	Task	Permutation p	BH q (16 tests)
Phi-4 Multimodal	GSM8K	0.0486	0.2219
	BBH	0.3050	0.5048
	MMLU	0.0488	0.2219
	IFEval Strict	0.1722	0.3936
Voxtral Mini 3B	GSM8K	0.2593	0.5048
	BBH	0.5712	0.7030
	MMLU	0.0555	0.2219
	IFEval Strict	0.7733	0.8262
Qwen3-Omni	GSM8K	0.3470	0.5048
	BBH	0.0916	0.2931
	MMLU	0.7746	0.8262
	IFEval Strict	0.9194	0.9194
Qwen2.5-Omni (H)	GSM8K	0.5385	0.7030
	BBH	0.0056	0.0890
	MMLU	0.3385	0.5048
	IFEval Strict	0.1220	0.3254

M MUSIC-FEATURE COVERAGE

The supplemental feature table joins the 55 tracks to the cleaned upstream human classification annotations at MTG-Jamendo revision `cafd8e20c265ed84f1e61f1c87532971f43a62f`. Forty-six tracks match a record, but not every matched record has every annotation. The cleaned annotations retain unanimous judgments (Music Technology Group, 2026). Table 13 reports the actual field-level coverage. Missing values are not negative labels; for example, *not happy* is not equivalent to *sad*.

Track-level annotation	Positive	Negative	Missing
Happy	9	12	34
Sad	3	22	30
Relaxed	11	18	26
Aggressive	9	24	22
Party	2	27	26
Danceable	11	14	30

Table 13: Available human track-level labels among the 55 recordings. No inference from track names is used to fill missing values.

These upstream judgments concern the recordings, not necessarily the particular 30-second prepared excerpts or the exact portions mixed into each task. Tempo and LUFS fields remain missing in this feature table. The later runs archive their waveforms, but excerpt-level feature extraction is not part of this report. A tempo written in a title is kept separately from measured tempo. The present report does not claim a tempo–effect or mood–effect association. Such analyses require appropriate excerpt measurements or annotations and uncertainty estimates that respect shared items and the small track sample.

N COMPLETE RECORDING CATALOG

The numeric IDs in Table 14 align with Figure 1 and the following track-effect panels. Titles are displayed in the release's Latin-script form for the English report; the bundled detailed tables retain the original metadata and display corrections. Three ChMusic performer names were unconfirmed in the supplied candidate list, so those catalog entries identify the source rather than treating the added performer attribution as verified.

Table 14: All 55 recordings. IDs match the track panels. Artist/source labels follow the release except that unconfirmed ChMusic performers are shown by source. Full URLs, licenses, offsets, and hashes are in the bundled metadata.

ID	Recording	Artist / source
Folk / country		
1	glinguinette	Fadièse
2	01 Ouverture	Löhstana David
3	Y'vonim	Michael Lambright
4	Folk omelette	Pepe Frías
5	Finding Home	ruka
Ambient		
6	Jotun	Brokenkites
7	prolog03	Urmal Vesnat
8	In Memories (Colors)	Mark.Nine
9	Spartial	Klangbrandung
10	Sound is I	Oliver Rehbinder
Chinese traditional		
11	Liu Shui (Flowing Water)	Charlie Huang
12	Shi Mian Mai Fu (Ambush from Ten Sides)	ChMusic (performer unconfirmed)
13	Chun Dao Yi He (Spring Comes to the Yi River)	ChMusic (performer unconfirmed)

Continued on the next page

ID	Recording	Artist / source
14	Zhan Tai Feng (Fighting the Typhoon)	ChMusic (performer unconfirmed)
15	Erquan Yingyue (Moon Reflected on Second Spring)	Zhang Peijian
Cinematic / orch.		
16	Sunrise	Airmusic
17	Archer Revenge	Zeffon
18	Dawn Of Heroes	Aliaksei Yukhnevich
19	Fairy Tale for the Lonely	Alexandr Ossipov production music
20	Prepare to Be Boarded, Comrades	FiluAndDina
Classical		
21	Brandenburg Concerto No. 3 in G major, BWV 1048, I. Allegro	Advent Chamber Orchestra
22	Symphony No. 5 in C minor, Op. 67, III. Allegro	Skidmore College Orchestra
23	Scherzo in B-flat minor, Op. 31	OnClassical
24	Concerto for Two Violins in D minor, Op. 3 No. 11, I	Advent Chamber Orchestra
25	Serenade for Strings, Op. 22, IV. Larghetto	Advent Chamber Orchestra
Electronic / dance		
26	Aziba	Dj Saryon & Sory
27	Child's Smile	RudySeb
28	Protected	paniq
29	detroit	bernimusic
30	E-Samba	Starlight
Instrumental hip-hop		
31	b. street nois	V-zen instrumental beat
32	04	BLINDSIGHT
33	Aries4Rce_-_Boss	Aries 4Rce BeatZ
34	Energy Bounce 2 (100.5 BPM)	Shanel
35	Zombie Klub	Doctor Frost
Instrumental metal		
36	iron knight 5	Olezha Chopalish
37	Ready to Fight	Porrazzo
38	Good Morning	Andrey Avkhimovich
39	Way of Warrior	Oleg Serkov
40	Black Sky	Pablo Samonta
Instrumental pop		
41	Very Upbeat	Porrazzo
42	On My Day Off	Stuart Ross
43	Mr. Bleach	Jahzzar
44	Toys	Musician Toy
45	Noris	Jampy
Instrumental rock		
46	Powiedzim wszystkim nie, krzyknij argh!	Skala
47	happy hour	filippo martin
48	Last Launch Part I	Way Station
49	Banned	Mike Allen
50	We Don't Piss In Your Ashtrays	Rezydent S & Garaż Bogiego
Jazz		
51	N400	thelittlewoodstudio
52	Der lachende Vulkanier mit seinem Hund	NuJazz-Trio
53	faaâ	Tet production
54	My name is Nova, Bossa Nova	Giorgio Ghiglieri
55	La toupie	Jazzymuté

O COMPLETE TRACK EFFECTS

The two panels below provide larger views of 880 of the track-level point estimates in the main-text overview (Figure 1), for the four models whose per-item records are bundled. Numbered labels link to the full titles in Table 14; longer titles are abbreviated only in these enlarged panels. The four model columns in each task are P4 (Phi-4), VX (Voxtral Mini), Q3 (Qwen3), and H (historical Qwen2.5). Lines separate genre groups. The shared color scale matches Figure 4; no cell color is a significance claim. The detailed CSV additionally provides clean scores, track scores, reported confidence intervals, exact McNemar p values, and within-family BH q values.

	GSM8K				BBH				MMLU				IFEval Strict			
	P4	VX	Q3	H	P4	VX	Q3	H	P4	VX	Q3	H	P4	VX	Q3	H
01 glinguinette	-0.8	-0.4	0.6	1.2	1.6	-1.6	-2.0	-1.4	1.0	0.2	-1.4	-1.0	0.2	0.4	0.6	-0.8
02 01 Ouverture	-0.6	1.0	0.4	1.0	0.6	-1.0	-2.2	-0.2	1.0	-1.0	-1.2	-1.2	-0.4	-0.6	0.0	-0.2
03 Y'vonim	0.2	1.4	0.6	-0.2	1.0	-4.4	-2.4	-0.8	1.6	-1.6	-0.6	-2.0	-3.0	-0.2	-0.6	-1.2
04 Folk omelette	-0.4	0.6	-0.4	2.2	-0.2	-4.0	-3.4	-1.0	-0.2	-0.2	-0.8	-0.4	1.2	1.4	1.0	-0.2
05 Finding Home	0.6	1.2	0.6	1.6	0.0	-5.8	-3.4	-0.2	0.0	-1.2	-1.8	-1.2	0.6	-0.2	-1.4	-0.6
06 Jotun	-0.4	1.2	0.2	2.2	1.2	-5.4	0.0	0.0	-1.2	-0.6	0.0	-1.4	-0.8	1.0	0.8	1.0
07 prolog03	0.4	1.4	0.6	2.0	-0.6	-2.4	-2.8	-1.0	1.6	0.4	-1.0	-0.2	-2.0	1.0	2.2	0.0
08 In Memories (Colors)	0.6	1.0	1.2	1.0	0.8	-3.2	-2.8	-0.6	-0.2	-1.8	-0.6	-0.2	-0.6	2.0	-1.4	-1.0
09 Spartial	0.6	1.4	0.6	0.8	0.2	-3.0	-0.4	-1.0	0.2	0.0	-1.4	-0.6	-0.4	0.6	0.2	-0.6
10 Sound is I	-0.2	1.0	0.2	0.2	0.0	-1.6	-0.4	0.2	0.8	1.0	-1.4	-0.6	-2.6	1.4	-0.6	-1.0
11 Liu Shui (Flowing Water)	-0.8	0.8	0.0	1.0	0.2	-4.2	0.2	-0.2	0.4	-0.4	-1.2	-1.2	1.4	1.4	1.0	-0.2
12 Shi Mian Mai Fu (Ambus...	-0.8	1.2	0.2	1.2	-0.8	-2.0	-1.8	-0.6	1.0	0.6	-2.0	-0.4	-2.0	0.4	0.8	-2.0
13 Chun Dao Yi He (Spring...	-0.6	1.6	0.4	2.2	-0.2	-3.8	-0.8	-1.4	0.8	-0.4	-0.6	-1.2	-0.6	2.8	-0.2	-0.8
14 Zhan Tai Feng (Fightin...	0.2	0.8	-1.0	0.6	0.2	-3.0	0.2	0.6	2.2	-0.2	-0.6	-1.2	-1.0	-1.4	2.0	0.6
15 Erquan Yingyue (Moon R...	0.0	0.4	-0.2	1.8	0.4	-2.6	-1.2	-0.2	0.6	-0.8	-0.8	-1.4	-1.6	3.0	-1.2	-0.8
16 Sunrise	0.2	0.6	-0.8	2.6	-0.4	-1.8	-2.0	-1.0	0.2	0.0	-0.2	-0.2	-0.8	0.0	-0.8	-1.8
17 Archer Revenge	-0.4	0.6	-1.2	1.8	0.6	-3.0	-2.2	-0.8	0.6	-2.0	-1.0	-3.6	-1.4	0.0	0.8	-3.0
18 Dawn Of Heroes	0.2	1.2	-1.4	0.8	0.0	-4.4	-2.0	-1.0	-0.2	-0.8	-1.4	-2.0	-2.0	1.2	0.0	-1.0
19 Fairy Tale for the Lon...	0.2	1.0	0.0	0.8	-0.2	-2.4	-3.6	-1.2	0.4	-0.6	-1.2	-1.8	-0.4	-1.2	0.4	-0.8
20 Prepare to Be Boarded...	0.0	1.2	0.0	2.2	0.0	0.0	-2.0	-0.6	-0.4	-1.0	-1.0	-1.0	-1.0	2.4	0.0	-0.4
1 Brandenburg Concerto N...	0.6	0.2	-0.8	0.4	1.4	-4.6	-2.6	-2.0	-0.4	-1.4	-1.0	-0.6	-2.8	3.0	-0.2	-1.6
2 Symphony No. 5 in C mi...	1.4	1.0	0.6	2.2	0.4	-5.0	-1.4	-2.4	1.2	0.0	-1.4	-0.8	-1.4	-1.2	-0.6	1.2
23 Scherzo in B-flat mino...	-0.2	1.0	0.4	1.2	-0.6	-1.8	-2.0	-1.6	-1.6	-1.6	-1.0	-1.4	-0.4	-1.0	-0.4	-2.2
24 Concerto for Two Violi...	0.8	0.8	-1.2	1.2	0.8	-3.6	-1.4	-1.0	0.4	-0.4	-1.4	-1.2	-2.0	1.2	0.2	0.6
25 Serenade for Strings,...	0.8	0.6	0.4	0.4	0.0	-3.8	-3.2	-1.6	-0.8	-1.0	-1.6	-1.6	-1.0	1.2	0.0	1.0
26 Aziba	-0.4	0.0	-0.2	2.0	0.8	-5.2	-3.6	-0.8	-0.4	-1.2	-0.8	-2.6	1.2	-0.6	0.6	-1.2
27 Child's Smile	0.8	0.0	-0.2	1.6	-0.6	-4.6	-3.6	-1.2	0.4	-0.2	-0.6	-1.2	-1.8	0.8	-1.4	0.2
28 Protected	0.6	0.8	0.4	3.0	1.2	-3.4	-2.4	-1.0	-0.6	0.0	-1.2	0.0	0.0	0.0	-0.2	-4.2
29 detroit	0.8	-0.6	-1.4	2.2	0.4	-3.6	-2.6	-1.6	0.0	1.0	-0.6	-0.8	-0.2	0.6	1.0	-1.8
30 E-Samba	0.2	1.0	0.4	1.4	0.8	-3.6	-1.2	-0.4	0.0	-0.4	-1.2	0.0	-0.4	0.6	0.8	-2.0

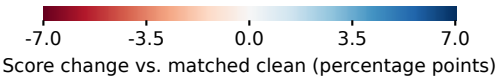


Figure 6: Track effects for recordings 1–30, in percentage points relative to their genre-local clean controls.

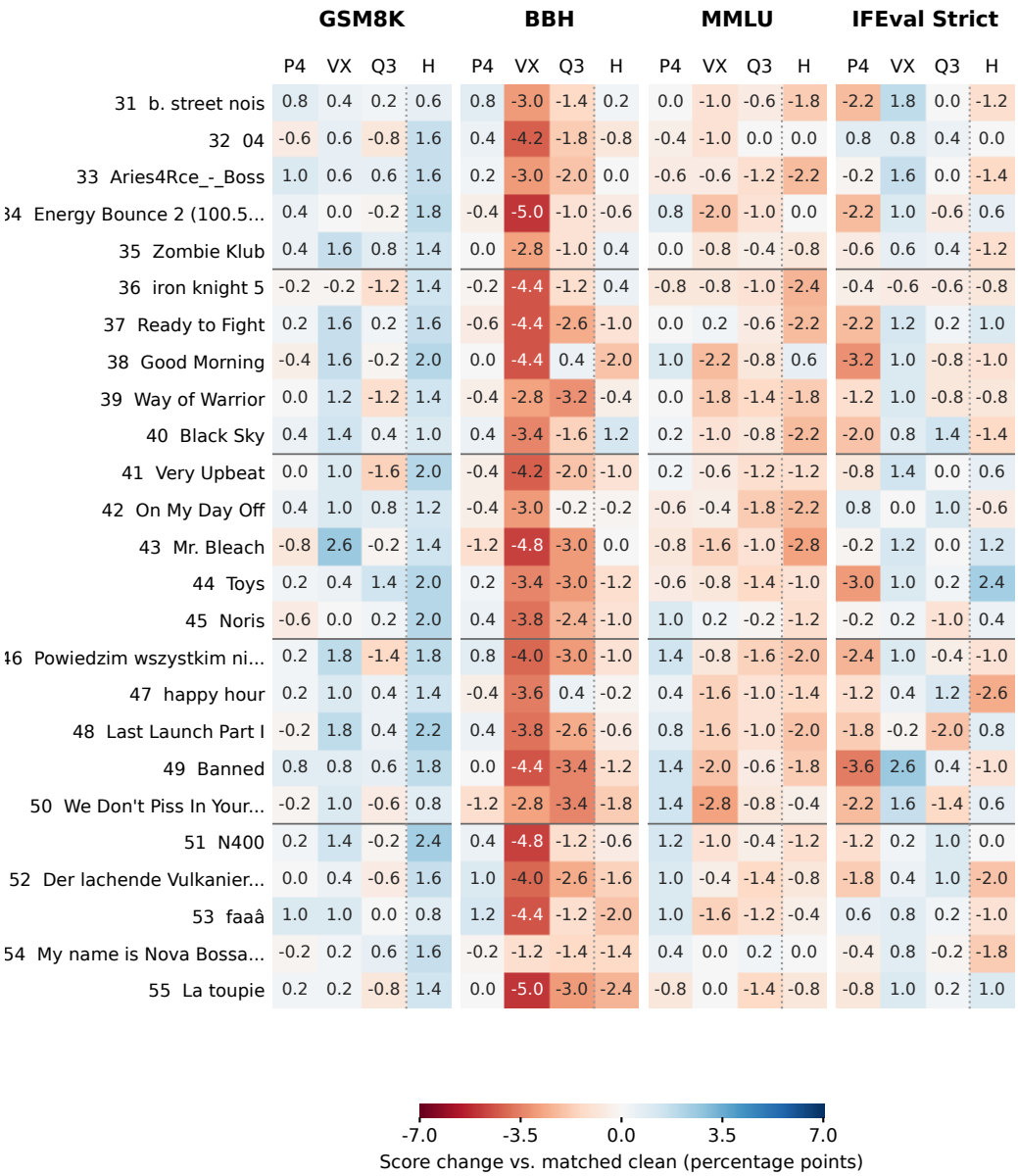


Figure 7: Track effects for recordings 31–55, using the same scale and model order.

P BENCHMARK CAPABILITIES AND EVALUATED CATEGORIES

This guide explains what each benchmark asks the model to do and lists the categories actually included in this report. The descriptions follow the benchmark definitions; coverage counts come from the bundled item metadata. These are intended task demands; our scores reflect the complete spoken-question-to-text-answer pipeline.

P.1 GSM8K: SOLVING EVERYDAY MATH PROBLEMS

GSM8K tests whether a model can translate a word problem into a sequence of arithmetic calculations and obtain the correct number (Cobbe et al., 2021). For example, it may need to combine quantities, prices, or rates before computing a total. We use 500 test questions from the `main` configuration. The selected items have no separate topic or difficulty labels. We therefore report one aggregate math score for this benchmark.

P.2 BBH: TEN KINDS OF REASONING AND LANGUAGE UNDERSTANDING

BIG-Bench Hard (BBH) is a suite of 23 task families (Suzgun et al., 2022). Our 500-question selection covers the ten subtasks below, with 50 questions each. These are distinct operations, not a uniform measure of long reasoning chains. The labels below are readable versions of the released subtask names.

Category	**Items**	**What the model must do**
Boolean expressions	50	Evaluate nested expressions with AND, OR, and NOT.
Causal judgment	50	Judge whether an event caused an outcome in a short scenario.
Date understanding	50	Work out calendar dates from relative-time clues.
Pronoun disambiguation	50	Determine which person or entity a pronoun refers to.
Formal fallacies	50	Decide whether a conclusion follows logically from the premises.
Logical deduction (3 objects)	50	Infer an ordering of three objects from verbal constraints.
Navigation	50	Track a route and decide if it returns to the start.
Object counting	50	Identify and count the requested objects in a description.
Colored objects	50	Reason about objects using their colors and positions.
Web of lies	50	Track linked truth/lie statements to determine who is truthful.

The subtask effects in Appendix G use these same categories; their average should not be read as evidence that every reasoning skill changes in the same direction.

P.3 MMLU: KNOWLEDGE AND PROBLEM SOLVING ACROSS SUBJECTS

MMLU evaluates multiple-choice questions requiring subject knowledge and its application (Hendrycks et al., 2020). The full benchmark has 57 subjects. We use 20 subjects and 500 questions: 125 in each of the four broad groups below. The group mapping follows the official MMLU categories. Group labels describe subject domains, not four isolated cognitive skills.

Category	Items	What the model must do
STEM: scientific concepts and technical reasoning		
Astronomy	25	Apply concepts about stars, planets, and observations.
College biology	25	Apply biological mechanisms and principles.
Conceptual physics	25	Reason about physical situations using qualitative laws.
High-school biology	25	Understand cells, organisms, genetics, and ecology.
Computer science (HS)	25	Apply high-school programming and computing concepts.
Humanities: history, interpretation, and argument		
European history	12	Interpret European events and historical developments.
World history	26	Connect world events, periods, and historical contexts.
Logical fallacies	29	Recognize flaws in arguments.
Philosophy	29	Interpret philosophical concepts and arguments.
World religions	29	Distinguish religious beliefs, practices, and traditions.
Social sciences: people, institutions, and behavior		
Geography	25	Apply spatial, population, and human-environment concepts.
Government and politics	25	Understand institutions, political processes, and governance.
Macroeconomics	25	Reason about aggregate economic behavior and policy.
Psychology	25	Apply concepts about behavior and mental processes.
Sociology	25	Understand social structures, groups, and institutions.
Other: applied health and business knowledge		
Clinical knowledge	25	Apply clinical concepts to patient scenarios.
Human aging	25	Understand biological and social aspects of aging.
Management	25	Apply principles of organizations and decision-making.
Marketing	25	Apply concepts about customers, markets, and strategy.
Nutrition	25	Understand nutrients, diet, and health.

P.4 IFEVAL: FOLLOWING EXPLICIT RESPONSE CONSTRAINTS

IFEval tests compliance with instructions that can be checked automatically, such as a word limit or a required output format (Zhou et al., 2023). It does not grade factual correctness or general helpfulness. Our 500 selected prompts contain 25 instruction types in nine families, matching the official instruction registry. Each row below covers all instruction types from that family present in our selection.

Instruction family	Prompts	What compliance requires
Keywords	139	Include required words; match keyword or letter frequencies; avoid forbidden words.
Response language	31	Write in the requested language.
Length and paragraphs	115	Meet word, sentence, or paragraph counts; start a specified paragraph with a required word.
Required content markers	51	Include a required number of placeholders or a postscript.
Output format	128	Produce bullet lists, highlighted passages, labeled sections, JSON, a marked title, or an answer from a fixed response set.
Combined responses	65	Give two responses in the prescribed structure, or repeat the prompt before answering.
Response boundaries	62	Use the required ending or wrap the answer in quotation marks.
Letter case	81	Use all uppercase, all lowercase, or a specified number of capitalized words.
Punctuation	62	Avoid commas.

How to read the counts and scores. A prompt can contain several instructions and belong to several families, so these counts overlap and do not sum to 500. The report's primary IFEval score is prompt-level Strict compliance: every required constraint must pass. Loose compliance tolerates specified superficial response variations; it is an alternative scoring rule, not another capability category. The instruction IDs and exact counts are recorded in the bundled capability inventory.

What the audio condition changes. The instruction is spoken, while the model still produces text. A model must recover constraints from the audio and obey them in that text. The known symbol-counting and grading-version differences remain as documented in Appendix D; the capability labels do not resolve those differences.

P.5 GSM-SYMBOLIC: ARITHMETIC UNDER CHANGES TO THE PROBLEM

GSM-Symbolic generates math questions from templates, changing names and numerical values while preserving the problem structure (Mirzadeh et al., 2024). Our experiment uses three variants, each with 500 questions from the same 50 template IDs:

Category	Items	What the model must do
Base	500	Solve new instances of the base word-problem templates.
P1 / GSM +1	500	Solve the problem after one extra task-relevant clause is added.
P2 / GSM +2	500	Solve the problem after two extra task-relevant clauses are added.

P1 and P2 increase the information and calculation demands; one extra clause does not necessarily mean exactly one extra reasoning step. They do not test irrelevant-text filtering. The benchmark paper also studies easier and irrelevant-clause variants, which are not evaluated here. See the original variant definitions and Appendix B for selection and grading details.

P.6 BABI: FINDING AND COMBINING RELEVANT FACTS

The bAbI tasks ask questions about small simulated stories. Our source is the BABILong 0k release, from which we select 100 problems each for qa1, qa2, and qa3 (Kuratov et al., 2024). The original bAbI task definitions distinguish them by supporting facts:

Category	Items	What the model must do
qa1: one fact	100	Find the relevant fact, such as a person's current location.
qa2: two facts	100	Combine facts about an object and its carrier to locate the object.
qa3: three facts	100	Combine a longer chain of events to recover an object's earlier location.

Our three context settings. Each setting contains the same 300 problems across qa1–qa3. These settings change the surrounding text, not the supporting-fact category:

Context setting	Items	What the comparison probes
Short	300	Answer from the intact facts without added book-text distractors.
Near	300	Ignore preceding book text; facts stay next to the question.
Far	300	Use facts separated from the question by the same irrelevant book text.

The question always comes last. Near and far use identical facts and filler sentences, changing their order. This is our short-audio adaptation, not the standard long-document BABILong evaluation. The figures average over qa1–qa3 within each setting.