

A Blind Score Is Not a Reliance Measure: What Answering Without the Audio Shows About Audio-Language Benchmarks, Why the Partial-Input Debate Already Settled Half of It, and the Counterfactual-Audio Test the Other Half Needs

Pranay Mahendrakar

SONYTECH · ORCID 0009-0003-7224-029X

pranaymahendrakar@sonytech.in · mahendrakarpranay@gmail.com · +91 6361723454

Abstract. Audio-language models are now routinely tested with the audio taken away, and the results are read in opposite directions. One 2026 audit reports that eight models keep 60 to 72 percent of their accuracy on three benchmarks when the audio is omitted, and concludes that the scores largely reflect text priors; the authors of one of those benchmarks replaced the audio with noise, saw accuracy fall, and concluded that the models are listening. In this analysis the author reviews and synthesises published results to show that both readings are correct and that they are about different quantities. A blind score estimates how answerable the items are without the audio, for one answerer under one replacement of the audio. It does not measure whether the full model's correct answers depend on the audio, and it does not settle whether the benchmark measures what it claims, which depends on whether its stated construct includes the knowledge that answers the item. Natural language inference research learned the first two points about hypothesis-only baselines between 2018 and 2022; multimodal evaluation is re-running that debate without citing its resolution. Re-expressing the 2026 audit's own table, the author shows that correcting for guessing lowers the retained share to 29 to 45 percent, that the blind score on one benchmark ranges from 8 to 72 percent of the above-chance score across models, and that ranking models by what the audio added is unrelated to ranking them by score on one benchmark ($\rho = -0.24$) but not on the other two. The paper gives a decision procedure for what a blind result licenses and specifies the paired counterfactual-audio test that would identify reliance item by item.

Keywords: large audio-language models; benchmark validity; partial-input baselines; text priors; modality reliance; multimodal evaluation; shortcut learning; construct validity

SONYTECH ORCID: 0009-0003-7224-029X

1. Introduction

Large audio-language models (LALMs) attach an audio encoder to a pretrained language model and are evaluated on multiple-choice benchmarks that pose questions about speech, environmental sound and music. Three of the most used are MMAU, which presents 10,000 clips with human-written questions spanning 27 skills (Sakshi et al., 2024); MMAR, which adds 1,000 items organised into four reasoning layers, including items that mix modalities (Ma et al., 2025); and MMAU-Pro, which adds 5,305 instances, open-ended formats and multi-audio items (Kumar et al., 2025). Earlier benchmarks such as AIR-Bench scored open-ended answers with a GPT-4 judge given the audio's meta-information (Yang et al., 2024). The surveys that organise this field classify such benchmarks by what they are meant to test - general auditory processing, knowledge and reasoning, dialogue, and fairness and safety (Yang, Ho and Lee, 2025) - and the scores are read as evidence of auditory understanding.

A sceptical reading has grown alongside them. If an item can be answered from its question and options alone, the score on it says nothing about hearing. The natural test is to take the audio away and see what is left. That test

now appears in at least five forms in the audio literature: omitting the audio input (Foo et al., 2026), replacing it with thirty seconds of silence (He et al., 2025), replacing it with Gaussian noise (Sakshi et al., 2024; Zang et al., 2025; Kumar et al., 2025), pairing the question with another item's audio (Nam, 2026), and asking a text-only language model (Zang et al., 2025; Foo et al., 2026). The results look damning. On a music question-answering benchmark, text-only language models reach up to 56.4 percent accuracy, on par with or above most audio models (Zang et al., 2025). Across eight LALMs and three benchmarks, models "retain 60-72% of their full audio scores even without any audio input" (Foo et al., 2026). A 2026 challenge task was built around filtering out items that blind models can answer (He et al., 2026).

And yet the same benchmarks' authors ran similar tests and drew the opposite conclusion. MMAR's authors replaced the audio with noise, observed that even the weakest model they tested "experiences a substantial performance drop", and concluded that "all models are indeed listening to the audio" (Ma et al., 2025). MMAU-Pro's authors did the same, saw every model drop, and read it as confirming that the models leverage the acoustic signal (Kumar et al., 2025). The two camps are not arguing about different data. They are drawing inferences in opposite directions from the same kind of contrast: the audit infers non-reliance from what

survives the removal, and the benchmark authors infer reliance from what does not.

This paper argues that both inferences are partly valid and that the disagreement dissolves once three quantities are separated. The first is answerability without the modality: whether an item can be answered from the text, for a stated answerer, under a stated replacement of the audio. A blind score estimates this, and only this. The second is reliance: whether the full model's answer to an item depends on the audio it was given. The third is validity: whether differences in benchmark score track differences in the capability the benchmark names. A benchmark can be highly answerable without audio while its models rely on audio wherever it is present; it can be unanswerable without audio for every tested blind answerer while still being solvable from a trivial acoustic cue that the construct does not include; and whether a text-answerable item is a defect at all depends on whether the benchmark's construct includes the knowledge that answers it.

None of this is new to machine learning. Natural language inference (NLI) went through the same sequence between 2018 and 2022. Hypothesis-only models classify about 67 percent of SNLI without seeing the premise (Gururangan et al., 2018), and similar baselines beat the majority class on a number of the ten NLI datasets tested (Poliak et al., 2018). The field then learned that the failure of a partial-input baseline does not certify a dataset as free of artifacts (Feng, Wallace and Boyd-Graber, 2019), and that a model trained on an artifact-ridden dataset can still condition on the part of the input the baseline ignores (Srikanth and Rudinger, 2022). Visual question answering learned the same things from question-only baselines (Jabri, Joulin and van der Maaten, 2016; Goyal et al., 2016; Agrawal et al., 2017), and later found that the measure most often used to escape them could itself be gamed (Teney et al., 2020). The 2026 audio audit cites the first half of this history - the hypothesis-only and question-only baselines - and not the second half, which is the half that bears on its conclusion.

The contribution is a separation and a set of consequences, which the paper states as six claims.

1. A blind score is a property of an item set, an answerer and a null operator jointly, not of the item set alone. In the 2026 audit's own table, the share of above-chance accuracy that survives omission of the audio on MMAU ranges from 8 percent for one model to 72 percent for another (Section 5.2).
2. The uncorrected retained share overstates answerability because it does not subtract guessing. Corrected for a four-option chance rate, the audit's 60 to 72 percent becomes 29 to 45 percent, and for MMAR the audit lists a chance rate of 25.0 percent where MMAR's own authors

report 29.32 percent, which lowers it further (Section 5.1).

3. A blind score does not measure reliance, in either direction. Correctness without the audio does not show that the full model ignored the audio when it had it, and a drop under noise does not show that the full model's correct answers came from the audio rather than from a distribution shift that the noise introduced (Sections 3 and 6).
4. The null operator is part of the measurement. Omission, silence, noise and mismatched audio are different counterfactuals; silence perturbs LALM outputs on text tasks where audio is irrelevant (Li, Lin and Lee, 2025), and cross-study comparisons of silence and omission on the same model and benchmark differ by up to nine points (Section 6).
5. Filtering items by a blind panel removes what that panel can answer, not what is answerable, and the residue is shaped by the panel, the threshold and the operator; one filtered audio benchmark still reports a random-guess baseline above nominal chance, which its organisers attribute to option-length and distractor artifacts (Section 7).
6. Reliance is identified by a paired counterfactual-audio design - the same question with two clips whose correct answers differ - which visual question answering adopted in 2016 and which existing audio benchmarks do not use at scale (Section 9).

The paper does not argue that current audio benchmarks are sound. Some of the evidence reviewed here makes them look worse than the audit does; in particular, ranking models by how much the audio raised their score is unrelated to ranking them by score on MMAU, though not on MMAR or MMAU-Pro (Figure 1). Its argument is narrower: the numbers now circulating are answers to a question about items, and the conclusions drawn from them are about models and constructs.

1.1 Relation to the author's earlier papers

Three earlier papers by the author touch adjacent questions. One argued that automatically constructed benchmarks define difficulty relative to a model panel and that this makes panel-relativity a property of the objective rather than of the generator; Section 7 below uses the same observation for blind filtering, but the present question - what a blind score measures - is not addressed there. A second argued that a benchmark-contamination flag is not an inflation estimate; the multi-modal leakage metric discussed in Section 5.3 is one such flag, and the present paper uses that argument without repeating it. A third argued that agreement between an automated judge and human labels does not certify the judge; Section 5.6 of the present paper makes the analogous point about agreement between a blind

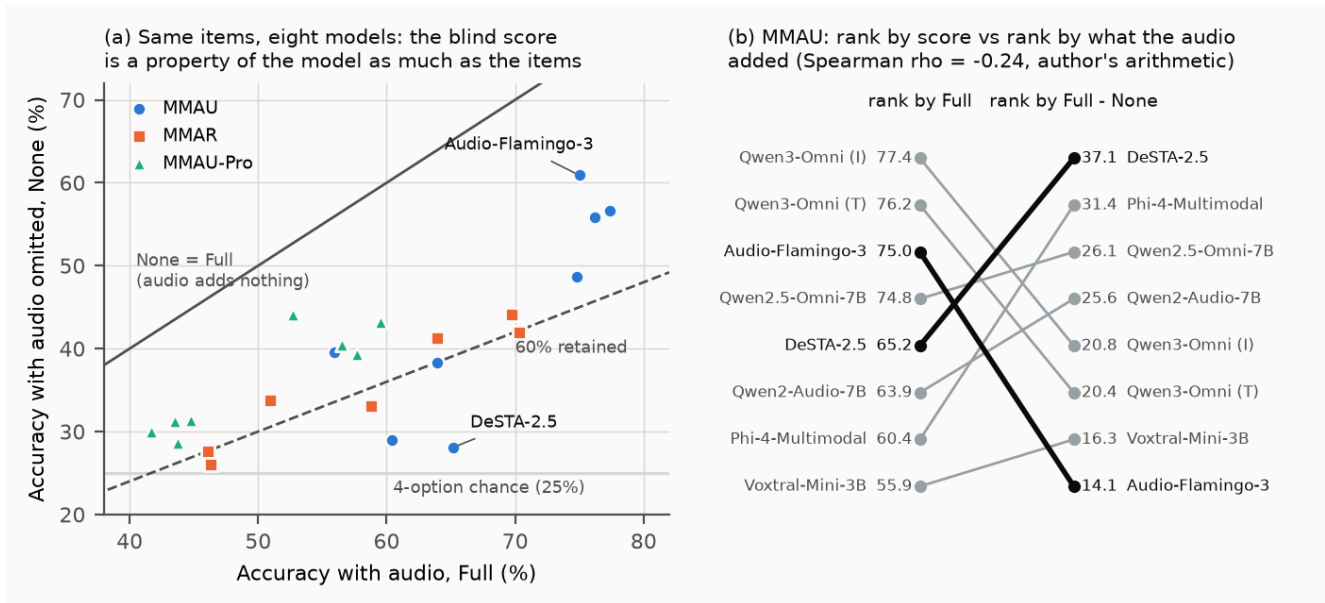


Figure 1. Panel (a): accuracy with audio omitted (None) against accuracy with audio (Full) for eight audio-language models on MMAU test-mini, MMAR and MMAU-Pro multiple-choice items, all values from Foo et al. (2026), Table 3a; the solid line is None = Full, the dashed line is the 60 percent retention the audit reports, the gray band is four-option chance. Panel (b): the same eight models on MMAU ranked by Full (left) and by Full - None (right), with each model's value; the ranks and the rank correlation are the author's arithmetic on Foo et al.'s values.

and a full model, which is a different object. None of the three treats partial-input baselines or modality reliance.

2. Scope, and what this paper does not claim

The paper is about multiple-choice evaluation of models that take audio and text, and it uses vision, video and text-only results where the audio literature is thin. Every number below is a published value or the author's arithmetic on published values, and the arithmetic is labelled as the author's where it appears.

Four limits are stated up front. First, the central audio source (Foo et al., 2026) is an unreviewed six-page preprint. Its table is used because it is the only located source that reports full-audio, no-audio and text-backbone accuracy for the same eight models on the same three benchmarks, which is what the re-expressions in Section 5 need. Second, item-level joint tables - which items a model answers correctly with and without the audio - are not published for any of the audio benchmarks, so every item-level statement in Section 5 is a bound or a consistency check, not an estimate. Third, the chance correction in Section 3.3 assumes that a model guesses uniformly among options on items it cannot answer; models with position or option-length preferences violate this, and Section 5.1 states where it matters. Fourth, the paper does not adjudicate which audio capabilities a benchmark should target; it argues only that the answer has to be stated before a text-answerable item can be called a defect.

3. Three quantities under one word

3.1 Definitions

Let a benchmark be a set of items i , each a question q_i with options O_i , an audio clip a_i and a keyed answer y_i . Let M be a model, and let ϕ be a null operator: a rule that replaces the clip with something else - nothing, a silent clip, a noise clip, another item's clip, or a fragment of the same clip. Write $c_i(M, x)$ for the indicator that M answers item i correctly when given input x .

Answerability without the modality. The quantity $A(M, \phi)$ is the share of items that M answers correctly under ϕ , that is, the mean over items of $c_i(M, \phi(a_i))$. It depends on three things: which items, which answerer, and which replacement. A blind score is an estimate of A for the answerer and operator used. When the answerer is a text-only language model and ϕ is omission, A is the question-and-options answerability for that language model. When the answerer is the audio model itself and ϕ is noise, A is how well that model does when given a clip that carries no information about the item but is still an audio input.

Reliance. The quantity $D_i(M)$ is whether M 's answer to item i depends on the audio it was given: whether there exist clips a and a' that the item's construct says should lead to different answers, such that M 's answer changes accordingly. D is a property of a model on an item. Aggregated over correct answers, it says what share of M 's score was earned by using the audio.

Validity. The quantity V is whether differences in score between models track differences in the capability the

benchmark claims to measure. In the psychometric tradition this is construct validity, which requires that the construct be specified before the test can be validated against it (Cronbach and Meehl, 1955), and which is a property of the interpretation of a score rather than of the test (Messick, 1995).

3.2 What follows from the definitions

The three quantities are related, but none determines another.

High A does not imply low D. An item that M can answer without audio may still be answered using the audio when the audio is present; the model may simply be able to reach the right answer by two routes. The NLI version of this was shown directly. Srikanth and Rudinger (2022) selected SNLI and delta-NLI examples that a hypothesis-only or bag-of-words model already answered correctly, edited the premise so that the correct label changed, and found that full-input models tracked the edit, scoring above 70 percent on the 600 edited examples. Their title states the point: partial-input baselines show that models can ignore context, but not that they do.

Low A does not imply high D, and does not imply V. An item that no tested blind answerer can solve may still be solvable from the text plus a trivial feature of the withheld modality. Feng, Wallace and Boyd-Graber (2019) constructed datasets with patterns that no partial-input model can detect, and then found such a pattern in SNLI: a model given the hypothesis plus one word from the premise solved 15 percent of the examples that a hypothesis-only model failed, raising accuracy from 70.10 to 74.6 percent. In audio, the analogue is an item that requires the clip but only a coarse property of it - its duration, its loudness, whether it contains speech. The audit's own fragment analysis is a partial-input test of exactly this kind on the audio side: of the items that needed audio, only 3.0 to 4.2 percent on average needed information spread across the whole clip (Foo et al., 2026).

Neither A nor D settles V without a construct. Whether an item that a text-only model can answer is a flaw depends on whether the benchmark's construct includes the knowledge used. Brown et al. (2025) take the strict position for vision-centric benchmarks: a pattern becomes a shortcut "if and only if it renders the visual input redundant for a task designed to measure visual understanding", whatever its origin. That is the right rule for a benchmark whose construct is perception. It is the wrong rule for a benchmark whose construct includes knowledge. OK-VQA was built so that the image is not sufficient and outside knowledge is required (Marino et al., 2019), and MMAU describes itself as testing "expert-level knowledge and complex reasoning" alongside perception (Sakshi et al., 2024). For such a benchmark, an item that a strong language model answers from knowledge is not obviously invalid; what is invalid is

reporting its score as a measure of hearing.

3.3 The chance floor

On a k-option item, a model that cannot answer and guesses uniformly is right with probability $1/k$. If a blind answerer knows the answer to a fraction κ of items and guesses on the rest, its blind accuracy is $b = \kappa + (1 - \kappa)c$, where c is the chance rate, so $\kappa = (b - c) / (1 - c)$. This is the classical correction for guessing applied to a test score (Diamond and Evans, 1973). The same logic applies to the ratio the audit reports. Its text-prior rate is $R_{TP} = b / f$, where f is full-audio accuracy. The share of the above-chance full score that survives the removal is

$$R_c = (b - c) / (f - c),$$

which equals R_{TP} only when $c = 0$. R_c is the quantity a reader usually means by "how much of the score does not need the audio". It is still not a reliance measure - it is an aggregate of A, with the guessing floor removed - but it at least does not count lucky guesses as text-answerable items.

The same correction applies at the item level. The audit classifies an item as text-solvable if a model answers it correctly with and without the audio in a single decoding run (Foo et al., 2026). If a model's no-audio answers were uniform guesses unrelated to whether it got the item right with audio, the expected share of items in that category would be f times c from guessing alone. For the audit's overall full-audio accuracies, that is between roughly 13 and 17 percent of items (Section 5.1).

4. What the partial-input literature already established

The audio debate is the fourth round of an argument that has been run in NLI, in reading comprehension and multiple-choice question answering, and in visual question answering. Each round reached the same four conclusions, and each is directly applicable to audio.

4.1 A strong partial-input baseline shows that a dataset can be shortcut

Gururangan et al. (2018) showed that a simple text classifier labels about 67 percent of SNLI and 53 percent of MultiNLI from the hypothesis alone, and traced the signal to annotation habits such as negation correlating with contradiction. Poliak et al. (2018) ran hypothesis-only models on ten NLI datasets and found that they significantly beat the majority-class baseline on a number of them. Kaushik and Lipton (2018) ran question-only and passage-only baselines on five reading-comprehension benchmarks and found that passage-only models exceed 50 percent on 14 of 20 bAbI tasks, while SQuAD and CNN looked better constructed. McCoy, Pavlick and Linzen (2019) showed with the HANS set that MNLI-trained models, BERT included, rely

on fallible syntactic heuristics such as lexical overlap. The general framing - a decision rule that does well on the benchmark and fails under a change of conditions - was later named shortcut learning (Geirhos et al., 2020). For large language models, choices-only prompts that omit the question beat a majority baseline in 11 of 12 model-dataset combinations, with gains up to 0.33 in accuracy (Balepur, Ravichander and Rudinger, 2024).

Visual question answering reached the same point earlier. A simple model that scores image-question-answer triplets competes with far more complex systems, and its authors concluded that current VQA systems were "not significantly better" than systems designed to exploit dataset biases (Jabri, Joulin and van der Maaten, 2016). Evaluation scores were found to be dominated by large, easy question categories, which let a plain multilayer perceptron beat a more complex model (Kafle and Kanan, 2017). Goyal et al. (2016) balanced VQA by collecting, for each question, a second image with a different answer, and every model they benchmarked did significantly worse on the balanced set.

4.2 A weak partial-input baseline does not show that a dataset is sound

This is Feng, Wallace and Boyd-Graber's (2019) result, described in Section 3.2: "the failure of a partial-input baseline does not mean a dataset is free of artifacts". Gardner et al. (2021) generalised the concern, arguing that for complex language understanding tasks every simple feature-label correlation is spurious and that realistic datasets drift further from that ideal as they grow. Adversarial filtering, which removes items a model can solve, lowers model accuracy on SNLI from 92 to 62 percent while human accuracy stays high (Le Bras et al., 2020); the same filtering is relative to the models used to filter.

4.3 The full model can use the modality the baseline ignores

Srikanth and Rudinger's (2022) context-editing study is the cleanest demonstration, and it was run on the very datasets whose hypothesis-only artifacts launched the line. In vision-and-language, several instruments were built to measure reliance directly rather than infer it from a baseline. The perceptual score permutes one modality across test examples and measures how far accuracy falls (Gat, Schwartz and Schwing, 2021); applied to VQA, it found that newer, more accurate models perceive the image less than their predecessors. MM-SHAP attributes a model's output to each modality with Shapley values and was designed because accuracy-based tests miss cases where a model used the right modality but answered wrongly (Parcalabescu and Frank, 2022); applied to six models it found unimodal collapse in different directions for different models and tasks. Empirical multimodally-additive function projection removes cross-modal interactions from a trained model's predictions; on seven image-and-text

tasks it often removed little or no performance (Hessel and Lee, 2020). Cross-modal input ablation showed that vision-and-language transformers depend on vision to predict text far more than on text to predict vision (Frank, Bugliarello and Elliott, 2021). Each of these is a reliance measure in the sense of Section 3.1. None is a blind score.

4.4 Fixes keyed to the baseline can be gamed by the baseline

VQA-CP re-split VQA so that answer priors differ between training and test (Agrawal et al., 2017), and debiasing methods were built against it, for example by training against a question-only branch (Cadene et al., 2019) or by ensembling with a bias-only model, which gained 12 points on a changing-priors VQA set (Clark, Yatskar and Zettlemoyer, 2019). Teney et al. (2020) then showed that most published methods exploited knowledge of how the out-of-distribution split was built, often by inverting the training answer distribution, and that a method answering at random surpassed the state of the art on some question types. A benchmark constructed to defeat one baseline invites methods that defeat the construction.

The four lessons map onto the audio debate as follows. The audit's no-audio and text-backbone scores are strong partial-input baselines and show that the benchmarks can be shortcut (Section 4.1). Filtered audio benchmarks that pass a blind panel are not thereby shown to be sound (Section 4.2). Neither result says what a full model does with audio that is present (Section 4.3). And training methods that use the with-and-without-audio contrast as a signal - several of which are reviewed in Section 9 - risk turning the contrast into a target (Section 4.4).

5. The audio evidence, re-read

5.1 What the audit measured, and its numbers corrected for guessing

Foo et al. (2026) evaluated eight LALMs, from 3B to 30B parameters, each paired with the text-only language model it was built on, on the 1,000-item test-mini split of MMAU, the 1,000 items of MMAR, and the multiple-choice items of MMAU-Pro. They used three conditions: Full (audio and text), None (the LALM with the audio input omitted entirely, which they chose over silence because silence had been shown to interfere with outputs), and TB (the text backbone given only the question). Answers were extracted by regular expression with a language-model judge as fallback; on 185 format-sensitive cases the judge agreed with human annotation 97.2 percent of the time against 26.0 percent for string matching. Decoding was greedy for all models except one run in thinking mode at temperature 0.6. Averaged over models, None accuracy was 44.6, 34.2 and 36.0 percent on MMAU, MMAR and MMAU-Pro, against Full accuracy of 68.6, 56.5 and 50.0 percent, which is where

the headline 60 to 72 percent retention comes from (65.1, 60.5 and 72.1 percent respectively). Table 1 places these values beside the other published blind and partial-input results that the argument uses.

*MMAR's own authors give the random-guess accuracy as 29.32 percent (Ma et al., 2025). In the Rouditchenko et al. row, both values are accuracies with audio, before and after fine-tuning without audio.

The audit's table lists chance as 25.0 percent on MMAU and MMAR and 25.9 percent on MMAU-Pro. Applying the correction of Section 3.3 to its overall rows - this is the author's arithmetic on the published values - the share of the above-chance score that survives omission of the audio is 0.45 on MMAU, 0.29 on MMAR and 0.42 on MMAU-Pro. The correction does not overturn the audit's qualitative point: on MMAU and MMAU-Pro, close to half of what the models score above chance does not require the audio for the average model under omission. It does change the magnitude from "most" to "somewhat under half", and on MMAR to under a third.

MMAR needs a further correction. Its authors report that most questions have four options but a minority have two or three, and give the random-guess accuracy as 29.32 percent averaged over the benchmark (Ma et al., 2025); He et al. (2025) also list 29.3 percent for MMAR. The audit's table lists 25.0 percent. With MMAR's own chance rate, the overall corrected retention falls from 0.29 to 0.18, and the None accuracies of three of the eight models (26.1, 27.6 and 26.0 percent) fall below chance. A blind accuracy below chance is informative in itself. It means the no-audio condition is not producing uniform guesses: the model is either systematically drawn to wrong options, failing to produce a parseable answer, or answering differently in a way the chance model does not describe. Any of these makes the blind score a measure of behaviour under an unusual input as well as of answerability.

The audit's item categories carry the same issue. It classifies items as text-solvable when a model is right both with and without audio, reporting 26.2 to 39.5 percent of items in that category, averaged over models, across the three benchmarks. Under the uniform-guessing reading of Section 3.3, guessing alone would place f times c of the items there, about 17 percent for MMAU, 14 percent for MMAR at a 25 percent chance rate (17 percent at 29.32) and 13 percent for MMAU-Pro. Because each category is assigned from a single decoding run per condition, the category of an individual item also carries sampling noise that a repeated run would expose. MMMU-Pro's construction, which counts a question as answerable only if a model gets it right in more than five of ten repeated trials (Yue et al., 2024), is one way to suppress both problems; the audit does not use it.

5.2 The blind score is a property of the model as much as of the items

Table 2 re-expresses the audit's MMAU row for each model. The items are identical across rows; what differs is the answerer.

On the same thousand items, the chance-corrected share of the score that survives omission ranges from 0.08 for DeSTA-2.5, whose no-audio accuracy of 28.1 percent is barely above chance, to 0.72 for Audio-Flamingo-3, whose no-audio accuracy of 60.9 percent exceeds the full-audio accuracy of two of the other seven models. On MMAR the corrected range is 0.05 to 0.43 at the audit's chance rate, and on MMAU-Pro 0.15 to 0.68. A statement of the form "benchmark X has a text prior of Y percent" therefore has no fixed referent: it is an average over whichever answerers were tested. This matches what MMStar's authors found for vision, where an image-language model answering without images reached 43.6 percent on MMMU while its language backbone, Mixtral-8x7B, reached 25.7 percent, 17.9 points lower (Chen et al., 2024).

Figure 1 makes the same point visually and adds the second consequence. Panel (a) plots None against Full for all 24 model-benchmark pairs. Panel (b) ranks the eight models on MMAU twice, once by Full and once by the increment the audio added. The two rankings are close to unrelated: by the author's arithmetic, Spearman's rank correlation between Full and Full - None across the eight models is -0.24 on MMAU, against 0.78 on MMAR and 0.71 on MMAU-Pro. The model with the third-highest MMAU score has the smallest audio increment, and the model with the largest increment is fifth by score.

This should not be read as showing that the increment is the right measure of hearing. It is not. A model that answers more items without audio has fewer items left on which the audio could help, so its increment is capped by its own blind score; normalising the increment by the headroom left, $(\text{Full} - \text{None}) / (100 - \text{None})$, moves the rank correlation on MMAU to 0.43, still well below the other two benchmarks. More importantly, the increment is still an aggregate over items, and it inherits every limitation of the blind score that it is computed from. The narrower point is descriptive: the MMAU leaderboard order is not the order of the omission increment. On MMAR and MMAU-Pro the two orders largely agree, partly by construction, since Full appears on both sides of the correlation and is itself positively rank-correlated with None on all three benchmarks.

5.3 One difference, two names

Foo et al. (2026) report that for most models None exceeds TB, and read this as multimodal training strengthening the model's text prior beyond that of its backbone. Chen et al. (2024) defined the same difference

Table 1. Published accuracies with the withheld input removed or replaced (Blind) and with the full input (Full), one row per cited result. Values are as reported by each source; a dash means the source does not report the value for that row.

Source	Benchmark	Operator	System	Blind	Full	Chance
Gururangan et al. (2018)	SNLI	Hypothesis only	Text classifier	~67	-	3-way
Feng et al. (2019)	SNLI	Hyp. only / hyp. + 1 word	Trained model	70.10 / 74.6	88.28	3-way
Srikanth and Rudinger (2022)	SNLI	Hypothesis only	RoBERTa	70	91	3-way
Balepur et al. (2024)	ARC, MMLU, HellaSwag	Choices only	4 LLMs	up to +33 over majority	-	majority
Chen et al. (2024)	MMMU	No image	GeminiPro	42.9	-	22.1
Chen et al. (2024)	MMMU	No image	Sphinx-X-MoE / its LLM	43.6 / 25.7	-	22.1
Cores et al. (2024)	MVBench, 4 text-bias tasks	Text only	Gemini 1.5 Pro	62.3	63.7	27.9
Cores et al. (2024)	MVBench, 4 other tasks	1 frame / shuffled	GPT-4o	62.8 / 67.5	65.8	23.8
Brown et al. (2025)	CV-Bench / VSI-Bench	Text, tuned on test folds	Qwen2.5-7B	+33.3 / +31.4 over 0-shot	-	-
Zang et al. (2025)	MuChoMusic	Text-only LLM	Best of several	56.4	-	-
Kumar et al. (2025)	MMAU-Pro	Gaussian noise	Audio Flamingo 3	47.2	51.7	23.4
Kumar et al. (2025)	MMAU-Pro	Gaussian noise	Qwen2.5-Omni-7B	30.6	52.2	23.4
He et al. (2025)	MMAU test-mini	30 s silence	Mean of 5 LALMs	49.8	-	25.5
He et al. (2025)	MMAR	30 s silence	Mean of 5 LALMs	36.6	-	29.3
Foo et al. (2026)	MMAU test-mini	Audio omitted	Mean of 8 LALMs	44.6	68.6	25.0
Foo et al. (2026)	MMAR	Audio omitted	Mean of 8 LALMs	34.2	56.5	25.0*
Foo et al. (2026)	MMAU-Pro (MCQ)	Audio omitted	Mean of 8 LALMs	36.0	50.0	25.9
Rouditchenko et al. (2025)	MMAU test-mini	Text-only fine-tuning	Qwen2.5-Omni	-	65.9 to 68.2	-
He et al. (2026)	ADQA dev (filtered) / ADQA-Bench eval	Silence filter applied	Random guess / best	25.46 / 58.33	-	25 nominal

Table 2. Foo et al.'s (2026) MMAU test-mini results per model (Full and None, their Table 3a) and the author's arithmetic on them: $R_{TP} = \text{None}/\text{Full}$ as the audit defines it, $R_c = (\text{None} - 25)/(\text{Full} - 25)$, the audio increment $\text{Full} - \text{None}$, and the model's rank by Full and by increment among the eight.

Model (Foo et al., 2026)	Full	None	R_{TP}	R_c	Full - None	Rank: Full / incr.
Audio-Flamingo-3	75.0	60.9	0.81	0.72	14.1	3 / 8
DeSTA-2.5	65.2	28.1	0.43	0.08	37.1	5 / 1
Phi-4-Multimodal	60.4	29.0	0.48	0.11	31.4	7 / 2
Qwen2-Audio-7B	63.9	38.3	0.60	0.34	25.6	6 / 4
Qwen2.5-Omni-7B	74.8	48.7	0.65	0.48	26.1	4 / 3
Qwen3-Omni (Instruct)	77.4	56.6	0.73	0.60	20.8	1 / 5
Qwen3-Omni (Thinking)	76.2	55.8	0.73	0.60	20.4	2 / 6
Voxtral-Mini-3B	55.9	39.6	0.71	0.47	16.3	8 / 7

for vision-language models, the score of the multimodal model without images minus the score of its language backbone, floored at zero, and called it multi-modal leakage: evidence that evaluation samples were seen during multimodal training. The quantity is identical. One paper reads it as a change in general linguistic competence and the other as memorisation of benchmark-like data. Neither reading is identified by the number.

The audit's own table gives one discriminating pattern. If multimodal training strengthened text reasoning in general, the excess of None over TB should appear across benchmarks for a given model. It does not for every model. Audio-Flamingo-3's no-audio score exceeds its backbone's by 15.4 points on MMAU and by 12.9 on

MMAU-Pro but falls 2.2 points short of it on MMAR; DeSTA-2.5's no-audio score is within 0.3 points of its backbone on MMAU and MMAR but 11.0 points above it on MMAU-Pro. Benchmark-specific excesses fit familiarity with a benchmark's item distribution better than a general change in reasoning, though they do not establish it; the audit does not report the training data of the models, and nothing reviewed here establishes overlap. What the pattern does establish is that "text prior" in the audit conflates at least three things: knowledge the backbone already had, whatever multimodal training added, and whatever the absence of the audio input does to the model's behaviour.

5.4 Opposite readings of the same contrast

The benchmarks' own authors ran removal tests before the audit. MMAU's authors replaced the audio with Gaussian noise on the test set and found that two models (MuLLaMa and SALMONN) were largely unaffected, which they read as limited reliance, while three others (GAMA, Qwen2-Instruct and Gemini Pro v1.5) dropped significantly, which they read as attention to the audio (Sakshi et al., 2024). MMAR's authors did the same and concluded from the drop that all models were listening, while noting that Qwen2.5-Omni stayed slightly above random even with noise, which they called "a residual language prior bias" despite a review process that instructed annotators to exclude text-answerable questions (Ma et al., 2025). MMAU-Pro's authors report noise and real-audio accuracy for four models - 34.9 against 38.7 percent for Phi-4-Multimodal, 47.2 against 51.7 for Audio Flamingo 3, 31.3 against 46.1 for Qwen2.5-Omni-3B and 30.6 against 52.2 for Qwen2.5-Omni-7B, with random guessing at 23.4 - and read the drops as confirming that the models leverage acoustic information (Kumar et al., 2025).

Read through Section 3, the benchmark authors and the audit are reporting compatible facts. A drop under noise shows that, for some items, the model's answer changed when the audio changed. That is evidence of reliance on some items. It does not show that the items the model still gets right under noise were answered from the audio when the audio was present, or that the items it gets wrong under noise were answered from the audio either: the noise clip is itself an input the model was never trained to treat as empty, and some of the drop may be the model reacting to noise rather than missing information. Conversely, the audit's retained 60 to 72 percent shows answerability without the audio, not that the full model ignored it. In the MMAU-Pro numbers, Audio Flamingo 3 keeps 91 percent of its real-audio accuracy under noise and Qwen2.5-Omni-7B keeps 59 percent; both are described by their benchmark's authors as listening. The same table shows that most of Audio Flamingo 3's above-chance score is answerable without the real audio.

5.5 The scores track the language backbone

Three further results suggest that a large part of what audio benchmarks rank is the language model inside the audio model. Lu et al. (2026) turned a set of language models into audio models by fine-tuning each with an audio encoder, and found that under an identical training recipe the choice of base model alone produced a difference of more than ten points in the resulting audio model, and that text-only measures of auditory knowledge correlated with audio-grounded benchmark performance at $r = 0.71$ to 0.82 , while two audio-grounded benchmarks, MMAU and MMAR, correlated with each other at only $r = 0.40$. Rouditchenko et al. (2025) found that fine-tuning an audio model on text-

only science questions, with no audio at all, raised its MMAU test-mini accuracy with audio from 65.9 to 68.2 percent, and that much of the gain from reinforcement learning on audio question answering could be attributed to better text-based reasoning. Kumar et al. (2025) found that rephrasing MMAU-Pro questions with a language model raised two models' scores by up to 5 percent, which they attribute to language priors in the rewritten prompts. On a music benchmark, one audio model scored below the text-only language model it was fine-tuned from, despite having the audio (Zang et al., 2025).

These results can be read two ways, and the difference is again a statement of construct. Lu et al. (2026) read their correlation as showing that text-only evaluation is a cheap proxy for choosing a backbone - that auditory knowledge carried in text is part of what makes an audio model good. A perception-only reading would call the same correlation contamination of the construct. The evidence does not choose between them; the benchmark's statement of purpose does. MMAU describes itself as emphasising "advanced perception and reasoning with domain-specific knowledge" (Sakshi et al., 2024). A benchmark defined that way has included part of what the backbone knows in what it measures, and its blind score is a defect report only on the subset of items whose construct excludes that knowledge.

5.6 The same pattern in vision and video

Vision-language and video-language evaluation produced the same numbers first. GeminiPro reached 42.9 percent on MMMU with no image, and MMStar's authors counted, in the benchmarks they examined, the items that six or more of eight strong language models answered without the image: 57.2 percent of ScienceQA and 46.2 percent of AI2D (Chen et al., 2024). For video, a model that sees a single frame can match or beat multi-frame models on standard video-language tasks, revealing what its authors call static appearance bias (Lei, Berg and Bansal, 2022), an atemporal probe bounds what image-level understanding can achieve on video question answering (Buch et al., 2022), and video models trained only on images perform significantly above chance on a temporal reasoning benchmark (Lydakis et al., 2025). On four MVBench tasks chosen for textual bias (action count, unexpected action, action antonym and episodic reasoning), Gemini 1.5 Pro scored 62.3 percent from the text alone against 67.4 percent for the best video model given the video, and on a different four (fine-grained action, scene transition, fine-grained pose and episodic reasoning) GPT-4o scored 62.8 percent from a single random frame, 67.5 percent from shuffled frames and 65.8 percent from the video in order (Cores et al., 2024).

The vision literature also shows the converse that the blind score cannot detect. High benchmark scores coexist with failures on trivial perception: four state-of-the-art vision-language models averaged 58.07 percent on seven

simple tasks such as whether two circles overlap, with the best at 77.84 percent (Rahmanzadehgervi et al., 2024), and multimodal models struggle with pairs of images that their vision encoder represents as similar despite visible differences (Tong et al., 2024b). When text and image disagree, vision-language models disproportionately follow the text (Deng et al., 2025), and a benchmark built to isolate language priors from commonsense and perception failures finds significant reliance on language priors in almost all models (Lee et al., 2024). None of these failures would appear as a high blind score. They are failures of reliance where reliance was needed, and a blind score measures the opposite case.

Brown et al. (2025) trace the blind test, now a routine sanity check, to Cambrian-1 (Tong et al., 2024a) and observe that it "only reveals when vision is unnecessary". One inference in this literature is worth flagging because it is the inference this paper argues against. Cores et al. (2024) report that 85.3 percent of a video model's correct text-only answers were also correct with the video, and take this as "confirming its strong reliance on textual biases". Agreement between the blind and full conditions on the items answered blind shows that those items are answerable without the video for that model. It does not show that the model answered them from text when it had the video; Srikanth and Rudinger's (2022) edited premises are the test that would.

6. The null operator is part of the measurement

6.1 Five ways to take the audio away

Table 3 lists the null operators and filtering rules used in the sources reviewed. They differ in what they hold fixed.

Omission feeds the model an input shape it may never have seen during audio training; silence and noise feed it a well-formed audio input that carries no information about the item; a mismatched clip feeds it real audio that carries information about a different item; a fragment feeds it part of the right audio. These are different counterfactuals, and in the explainability literature the problem is familiar. When features are removed at test time, the resulting inputs are out of distribution, so a drop in performance cannot be attributed to the removed information rather than to the shift (Hooker et al., 2018); different replacement functions produce counterfactuals that are out of distribution to different degrees, and the model's prior then shapes the result (Hase, Xie and Bansal, 2021).

6.2 What is known about each operator in audio

Silence is not neutral. On three text-only reasoning benchmarks where audio is irrelevant, adding silence, noise or environmental sound to an LALM's input lowered accuracy and made predictions more volatile,

with silence destabilising outputs as strongly as synthetic noise, and the effect grew with duration, amplitude and decoding temperature (Li, Lin and Lee, 2025). A silent-audio blind score therefore mixes answerability with interference. Noise is not neutral either: on the music benchmark audited by Zang et al. (2025), one audio model's accuracy rose when its audio was replaced by noise, and another could not be evaluated under noise because it refused to answer.

Silence can also carry signal. In clinical speech, classifiers trained on silent segments alone frequently matched or exceeded full-audio performance on five widely used corpora, which the authors attribute to dataset-specific confounds (Jouaiti and Ma, 2026). A blind operator that keeps the recording's silences, room tone or duration is not blind to the recording.

The cross-study comparison is the closest thing to a test of whether the operator matters at the benchmark level, and it is uncontrolled. He et al. (2025) report silent-audio accuracy for five LALMs on MMAU test-mini and MMAR; Foo et al. (2026) report omitted-audio accuracy for eight. Two models appear in both under the same names. On MMAU test-mini, Qwen2-Audio scores 38.8 percent with silence and 38.3 with omission, and Qwen2.5-Omni 48.5 and 48.7. On MMAR, Qwen2-Audio scores 33.1 with silence and 26.0 with omission, while Qwen2.5-Omni scores 32.4 with silence and 41.3 with omission. The MMAU agreement is close; the MMAR differences are seven and nine points, in opposite directions for the two models. The studies differ in answer extraction and possibly in benchmark version, so these differences cannot be assigned to the operator. They can be read only as an upper bound on how stable a published blind score is across protocols: on this evidence, not stable to within several points.

6.3 The mismatched-audio operator

One operator avoids most of these problems. Pairing an item's question with another item's real clip keeps the input in distribution - it is real audio of the kind the model was trained on - while removing the information the item needs. The perceptual score is built on exactly this permutation (Gat, Schwartz and Schwing, 2021). In audio, Nam (2026) used empty-audio and shuffled-audio conditions together and sorted training items into four buckets - text-prior, shuffle-leak, strong audio-dependent, and hard or misleading - selecting for fine-tuning the items that were right with the correct audio and wrong under both counterfactuals. The shuffle-leak bucket is the informative one: items answered correctly with the wrong clip but not with empty audio, which a silence test alone would count as strongly audio-dependent. Mismatched audio is still not a reliance measure on its own, because a clip from a different item may share the property the question asks about. It becomes one when the substitute clip is chosen so that the construct predicts a different answer, which is the

Table 3. Null operators and blind-filtering rules in published multimodal evaluations, with the source of each.

Source	Operator on the withheld modality	Used for	Panel and rule
Foo et al. (2026)	Audio input omitted; text backbone	Audit	8 LALMs, one run each
He et al. (2025)	30 s of silence	Training-data split	Weak if 2 of 3 LALMs right
He et al. (2026)	Silence; option perplexity; LLM check; human review	Benchmark filter	Drop if over 1 blind LALM right
Sakshi et al. (2024); Kumar et al. (2025)	Gaussian noise	Reliance check	Per model
Zang et al. (2025)	Gaussian noise; text-only LLMs	Audit and filter	Text-only perceptual index (log-probabilities)
Nam (2026)	Empty audio; another item's audio	Training-data buckets	One LALM
Gat et al. (2021)	Modality permuted across test items	Reliance score	Per model
Cores et al. (2024)	Text only; one frame; shuffled frames	Video audit	Per model
Chen et al. (2024)	Image omitted	Benchmark filter	Keep if 2 or fewer of 8 LLMs right
Yue et al. (2024)	Image omitted, 10 trials	Benchmark filter	Drop if 3 or more of 4 LLMs right in most of 10 trials
Brown et al. (2025)	Text inputs only, fine-tuned on test folds	Benchmark audit	Bias score per item

paired design of Section 9.

7. Filtering by a blind panel

Several benchmarks now filter items by blind answerability. MMStar ran eight language models without images and kept only samples that two or fewer answered correctly, reducing 22,401 candidates to 11,607 before human review selected 1,500 (Chen et al., 2024). MMMU-Pro ran four language models ten times each and excluded any question that at least three of the four answered correctly in a majority of trials (Yue et al., 2024). In audio, the DCASE 2026 audio-dependent question answering task discarded items when more than one silent-audio model picked the right option, dropped items whose correct option had the lowest language-model perplexity under silent audio, pruned items a text-only language model answered from commonsense, and ran a two-pass human review, leaving 3,000 evaluation items (He et al., 2026). RUListing rebuilt a music benchmark's distractors using a perceptual index computed from a text-only language model's log-probabilities, after which text-only models scored near chance and audio models given noise fell to near or below chance (Zang et al., 2025). Brown et al. (2025) go further: they fine-tune a language model on the text of the test set itself, by k-fold cross-validation, to find what is learnable without vision, and report gains over the zero-shot blind model of 33.3 and 31.4 points on two template-based benchmarks and 8.6 and 6.4 points on MMMU and VideoMME, then prune high-bias items iteratively.

These are improvements on not filtering. They are also subject to the four lessons of Section 4.

First, a filter removes what its panel can answer under its operator, not what is answerable. The DCASE organisers report that every submitted system beats a random-guess baseline of 25.46 percent, a level they give for the development set, which passed the same filter,

and describe as "above 25% likely due to option-length and distractor artifacts" (He et al., 2026). The organisers attribute the excess to those artifacts tentatively; if they are right, the filtered set still carries text-side regularities that the filtering panel did not exploit, which is Feng, Wallace and Boyd-Graber's (2019) point at benchmark scale. Brown et al.'s (2025) fine-tuned probe is the strongest available answer to this, because it searches for learnable regularities rather than testing one zero-shot answerer, and its large gains on template-based benchmarks show how much a zero-shot panel misses.

Second, a filter keeps what its panel gets wrong, for any reason. Items that blind models fail include items that need the audio and items that every model fails, such as items whose key is wrong or whose question is unanswerable. MMStar and DCASE both add human review for this reason. On the DCASE evaluation set, 233 items (7.8 percent) were missed by all 36 submissions, and an oracle choosing the best submission per item solves 92.2 percent; the organisers read the universal failures as capability gaps common to the backbones (He et al., 2026). That may be right. The filter by construction cannot tell those items apart from defective ones, and the summary does not report what share of candidates human review rejected, so the composition of the 233 is not known.

Third, the filter is relative to its panel, and a later model is not in the panel. A benchmark filtered against 2024 language models is not filtered against 2026 ones; the 38.2 percent that Gemini 1.5 Pro scored from text alone across all twenty MVBench tasks, against a random baseline of 27.3 percent (Cores et al., 2024), is a reminder that a stronger language model is a stronger blind answerer. Filtering also changes the item population. RUListing's items have new, perceptually targeted distractors, and MMMU-Pro also augmented the options; the filtered benchmark measures discrimination among those options, which is a different task from the original, and comparisons of a model's score before and

after filtering confound the change in items with the change in answerability.

Fourth, a filter keyed to blind answerability invites methods keyed to it. The most common training approach among the DCASE submissions was fine-tuning on a training corpus selected by a similar silent-audio criterion to the test filter, and the winning system binned training items by matched audio and silent prompts (He et al., 2026). This is a sensible engineering response. It also means that the challenge partly measures how well a system learns the boundary the filter drew, the pattern Teney et al. (2020) documented when VQA-CP's construction became a target. The organisers report an average fall of 11.91 points from development to evaluation accuracy across 30 submissions, on an evaluation split designed to be harder.

8. What a blind result licenses: a decision procedure

The argument so far reduces to a rule for reading blind results. Algorithm 1 states it as a procedure: given a benchmark, a claim someone wants to make about it, and the evidence available, it returns the strongest claim the evidence supports.

Applied to the cases reviewed: the audit's 60 to 72 percent licenses a claim at the ITEMS level, corrected to 29 to 45 percent and stated relative to its eight answerers and to omission. The MMAR and MMAU-Pro noise drops license "these models' outputs depend on the audio channel for some items" and nothing about how much of the score the audio earned. On MMAU, the rank correlation between score and the omission increment is negative on the audit's numbers, and on MMAR and MMAU-Pro it is positive. Neither result is a step 6(a) verdict: the omission increment is an aggregate drop under omission, which step 5 limits to "M's output depends on the audio channel for some items", not the counterfactual dependence that step 6(a) asks for, so substituting it licenses nothing beyond step 5. No located audio benchmark supplies the evidence that step 5 requires for any model.

9. The paired counterfactual-audio test

9.1 The design

Reliance in the sense of Section 3.1 is a counterfactual: would the model's answer have been different if the audio had been different in a way that matters for the question? The design that identifies it is a pair. The same question is asked of two clips, chosen so that the keyed answers differ, and the model is credited only if it answers both correctly. A model that ignores the audio answers both clips the same way and can be right on at most one; a model that guesses from a prior over options cannot exceed the pair-level chance rate. Unlike omission, silence or noise, both members of the pair are real, in-distribution clips, so the result does not depend

on how the model reacts to an input it was never trained on.

Visual question answering adopted this design in 2016, when VQA v2 was balanced by collecting for every question a second, similar image with a different answer (Goyal et al., 2016). Winoground pairs two images with two captions made of the same words in different orders, and found that none of the tested models did much better than chance (Thrush et al., 2022). NaturalBench pairs each question with two images that yield different answers, explicitly to prevent blind solutions; it reports that leading models such as GPT-4o lag human performance (above 90 percent) by 50 to 70 percent, and that models often choose the same answer regardless of the image (Li et al., 2024). In audio, CompA pairs two clips containing the same acoustic events in different orders or with different attribute bindings, and found that audio-language models trained contrastively do only marginally better than chance at matching them (Ghosh et al., 2023). That is the right design, applied to retrieval-style models and to two compositional skills. It has not been located applied at scale to the question-answering benchmarks on which LALMs are ranked.

9.2 What the pair measures, and what it does not

For a model M on a pair (i, i') with the same question and different keyed answers, the pair outcome has four cases: both right, both wrong, and one right. Both-right is evidence of reliance on that pair. A model that answers by a text prior lands on the same option for both clips and is right on at most one. The share of pairs answered both-right, corrected for the pair-level chance rate, is therefore a lower bound on the share of questions on which M 's answer is driven by the audio in the way the construct requires. It is a lower bound because a model may rely on the audio and still answer one member wrongly.

The pair does not settle validity by itself. A pair isolates the audio contrast the designer built in. If the two clips differ in a trivial property - one is louder, one contains speech - the pair rewards a trivial cue, which is Feng, Wallace and Boyd-Graber's (2019) concern transferred to the audio side. The audit's finding that most audio-dependent items can be solved from one fragment of the clip (Foo et al., 2026) is a reason to build pairs whose clips differ in the property the question names and are matched on duration, loudness and source.

9.3 Existing audio work that uses the contrast for training

Several 2025-2026 methods already compute the contrast between a model's output with and without the audio, or with intact and altered audio, and use it to train or decode. Audio-aware decoding promotes tokens whose probability rises when the audio is present, improving object-hallucination F1 by 0.046 to 0.428 and accuracy on a general audio question-answering set by

Algorithm 1. What a blind or partial-input result licenses.

```

Input: benchmark B with stated construct C; claim type K in
      {ITEMS, MODEL, LEADERBOARD}; evidence E.
1. If C is not stated (perception only, or perception plus
   knowledge), return "no validity claim; report A only".
2. For each blind score in E, record answerer, null operator
   phi, number of runs, answer extraction, chance rate c.
3. Correct every blind score for guessing:
    $R_c = (\text{blind} - c) / (\text{full} - c)$ .
   If blind < c, flag phi as perturbing, not removing.
4. K = ITEMS: answerability A(answerer, phi) is licensed.
   If C excludes knowledge, items answerable by a strong
   panel under an in-distribution phi are construct-irrelevant.
   If C includes knowledge, they are not defects; report the
   share separately.
5. K = MODEL (does M use the audio?): require item-level
   counterfactual evidence for M itself - paired clips with
   different keyed answers, or mismatched real audio chosen so
   C predicts a different answer. Aggregate drops under noise,
   silence or omission license only "M's output depends on the
   audio channel for some items". Retained accuracy licenses
   nothing about M's use of audio when audio is present.
6. K = LEADERBOARD (do score differences track C?): require
   (a) rank agreement between score and counterfactual
   dependence across models, and (b) a filter checked by a
   probe trained on the test text, not only zero-shot panels.
   If (a) fails, the leaderboard does not order models by C.
7. Return the licensed claim and the evidence classes missing.

```

5.4 to 10.3 percent (Hsu et al., 2025). A distillation method rewards the log-probability difference between a teacher with and without the audio (Li et al., 2026). Preference optimisation that contrasts intact and distorted audio raised accuracy by 14.0 percent on a 2025 challenge set and by 27.4 percent on an existence benchmark (Cosgrove, Grace and Wang, 2026). Choosing the perturbation per task matters: using a reversed clip as the contrastive branch raised a temporal-ordering task from 74.7 to 81.4 percent (Grace, Huo and Wang, 2026). The AudioMCQ corpus itself is split into weak and strong audio-contribution subsets by whether two of three models answer correctly with silence (He et al., 2025). Object-hallucination studies had earlier found that LALMs describe audio about as well as captioning models but struggle with discriminative questions about whether a particular sound is present (Kuan, Huang and Lee, 2024).

These methods are evidence that the with-and-without contrast carries a usable signal about reliance. They are also a reason not to use the same contrast, computed the same way, as the evaluation. A model trained to maximise the difference between its outputs with and without audio, or selected on items where that difference is large, will score well on a blind-filtered benchmark partly by construction. The paired design avoids this because it does not ask whether the model's output changes when the audio is removed; it asks whether the output tracks the audio when the audio is real and differs in the way the question names.

10. Measurements that would settle the open parts

Four measurements would settle what the published evidence leaves open. None needs a new benchmark from scratch.

M1. Item-level joint tables for existing audits. For each model, benchmark and null operator, publish the per-item correctness under the full and blind conditions, with at least five repeated runs per condition. This turns every aggregate in Section 5 into an estimate with an interval, separates guessing from answerability, and lets anyone compute R_c and pair-free reliance bounds without rerunning the models. The audit already has the single-run version of these data.

M2. Operator comparison on fixed items. On one benchmark and one set of models, run omission, silence of matched duration, noise, and a mismatched real clip from the same benchmark, with a fixed answer extractor. Section 6.2's cross-study differences of up to nine points between silence and omission on MMAR cannot be assigned to the operator until this is done.

M3. A paired subset of an existing benchmark. For a subset of MMAU or MMAR questions whose construct is perceptual - for example, which instrument is playing, how many events occur, which comes first - collect a second clip per question with a different keyed answer, matched on duration and loudness. Report both-right accuracy with its pair-level chance rate beside the original single-clip accuracy. If the models that lead the original leaderboard also lead the paired one, the leaderboard tracks audio use on that subset; if not, it does not.

M4. A trained probe for text-side regularities. Apply the fine-tuned test-set probe of Brown et al. (2025) to the filtered audio benchmarks, ADQA-Bench and RUListing's rebuilt items. A filtered benchmark whose text remains learnable by a probe trained on its own folds is filtered against a panel, not against text-side regularities.

11. Objections

"Removing the audio is the obvious test, and it found the problem." It found that the items are answerable without audio for some models, and that is a real problem for any benchmark whose construct is perception. The objection is right about that and this paper does not dispute it. What the test cannot do is the two further things it is being used for: rank models by their use of audio, and certify a filtered benchmark as measuring hearing.

"Text answerability is the defect whatever the construct says." Brown et al. (2025) state this position for vision-centric benchmarks, and for those it is correct by definition. Extending it to all audio benchmarks would rule out the knowledge component that MMAU and MMAR explicitly include, and that Lu et al. (2026) found to be a strong predictor of audio-grounded performance. The disagreement is about constructs, and it should be settled by benchmarks stating theirs, not by a general rule about priors.

"The chance correction is pedantic; the qualitative conclusion stands." The qualitative conclusion that the benchmarks can be shortcut stands. The correction matters for two other reasons. It changes the size of the claim by a factor of up to two, and on MMAR, with the benchmark's own chance rate, it turns three models' blind scores into below-chance scores, which is evidence that the blind condition is doing something other than removing information.

"The cross-study comparisons are confounded, so they show nothing." They are confounded, and Section 6.2 says so. They show that published blind scores for the same model on the same benchmark differ by several points across protocols, which bounds how much weight any single blind number can bear. The controlled version is M2.

"Paired benchmarks are expensive." They are more expensive than filtering. VQA v2, Winoground, NaturalBench and CompA show that they can be built, and M3 asks for a subset rather than a benchmark.

12. What is not known

Nobody has published, for any audio question-answering benchmark, per-item correctness for a set of LALMs under the full input and under more than one null operator. Without it, the share of any model's audio benchmark score that was earned by using the audio is not known for any model. Whether silence, omission and

noise give the same blind scores under a fixed protocol is not known. Whether the excess of an audio model's no-audio score over its backbone's reflects changed reasoning or familiarity with benchmark items is not known; the two available names for it, leakage and strengthened prior, are interpretations. Whether the leaderboards of MMAR and MMAU-Pro, which agree with the audio-increment ordering at the aggregate level, would also agree with a paired counterfactual ordering is not known. What share of the 233 items that every DCASE 2026 system missed are defective rather than hard is not known.

13. Conclusion

The removal test answers a question about items: which of them can be answered without the audio, by a given answerer, under a given replacement. The audio literature is using it to answer two other questions, about whether models listen and whether benchmarks measure listening. The first of those needs a counterfactual on the model's own answers with real audio; the second needs a stated construct and a check that score order tracks use of the audio. NLI and visual question answering settled the logic of this between 2016 and 2022, and multimodal evaluation is citing the beginning of that history without its end. Read through it, the 2026 audit and the benchmark authors who say their models are listening are both describing the same data correctly. What the data do not yet contain is the measurement that would say how much of any audio model's score its hearing earned.

References

- Agrawal, A., Batra, D., Parikh, D., & Kembhavi, A. (2017). Don't Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering. CVPR 2018. arXiv:1712.00377. doi:10.48550/arXiv.1712.00377.
- Balepur, N., Ravichander, A., & Rudinger, R. (2024). Artifacts or Abduction: How Do LLMs Answer Multiple-Choice Questions Without the Question?. ACL 2024. arXiv:2402.12483. doi:10.48550/arXiv.2402.12483.
- Brown, E., Yang, J., Yang, S., Fergus, R., & Xie, S. (2025). Benchmark Designers Should "Train on the Test Set" to Expose Exploitable Non-Visual Shortcuts. arXiv:2511.04655. doi:10.48550/arXiv.2511.04655. [PREPRINT]
- Buch, S., Eyzaguirre, C., Gaidon, A., Wu, J., Fei-Fei, L., & Niebles, J. C. (2022). Revisiting the "Video" in Video-Language Understanding. CVPR 2022. arXiv:2206.01720. doi:10.48550/arXiv.2206.01720.
- Cadene, R., Dancette, C., Ben-younes, H., Cord, M., & Parikh, D. (2019). RUBi: Reducing Unimodal Biases in Visual Question Answering. NeurIPS 2019. arXiv:1906.10169. doi:10.48550/arXiv.1906.10169.
- Chen, L., Li, J., Dong, X., Zhang, P., Zang, Y., Chen, Z., Duan, H., et al. (2024). Are We on the Right Way for Evaluating Large Vision-Language Models?. arXiv:2403.20330. doi:10.48550/arXiv.2403.20330. [PREPRINT]
- Clark, C., Yatskar, M., & Zettlemoyer, L. (2019). Don't Take the Easy Way Out: Ensemble Based Methods for Avoiding Known Dataset Biases. EMNLP 2019. arXiv:1909.03683. doi:10.48550/arXiv.1909.03683.

- Cores, D., Dorkenwald, M., Mucientes, M., Snoek, C. G. M., & Asano, Y. M. (2024). Lost in Time: A New Temporal Benchmark for VideoLLMs. arXiv:2410.07752. doi:10.48550/arXiv.2410.07752. [PREPRINT]
- Cosgrove, B., Grace, A. I., & Wang, W. (2026). Hallucination Reduction for LLM-Based Audio Understanding via Multimodal Direct Preference Optimization. arXiv:2610.04004. doi:10.48550/arXiv.2610.04004. [PREPRINT]
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281-302. doi:10.1037/h0040957.
- Deng, A., Cao, T., Chen, Z., & Hooi, B. (2025). Words or Vision: Do Vision-Language Models Have Blind Faith in Text?. CVPR 2025. arXiv:2503.02199. doi:10.48550/arXiv.2503.02199.
- Diamond, J., & Evans, W. (1973). The Correction for Guessing. *Review of Educational Research*, 43(2), 181-191. doi:10.3102/00346543043002181.
- Feng, S., Wallace, E., & Boyd-Graber, J. (2019). Misleading Failures of Partial-input Baselines. ACL 2019. arXiv:1905.05778. doi:10.48550/arXiv.1905.05778.
- Foo, L. H., Yang, C., Li, C., Lu, K., & Lee, H. (2026). All That Glitters Is Not Audio: Rethinking Text Priors and Audio Reliance in Audio-Language Evaluation. arXiv:2604.24401. doi:10.48550/arXiv.2604.24401. [PREPRINT]
- Frank, S., Bugliarello, E., & Elliott, D. (2021). Vision-and-Language or Vision-for-Language? On Cross-Modal Influence in Multimodal Transformers. EMNLP 2021. arXiv:2109.04448. doi:10.48550/arXiv.2109.04448.
- Gardner, M., Merrill, W., Dodge, J., Peters, M. E., Ross, A., Singh, S., & Smith, N. A. (2021). Competency Problems: On Finding and Removing Artifacts in Language Data. EMNLP 2021. arXiv:2104.08646. doi:10.48550/arXiv.2104.08646.
- Gat, I., Schwartz, I., & Schwing, A. (2021). Perceptual Score: What Data Modalities Does Your Model Perceive?. NeurIPS 2021. arXiv:2110.14375. doi:10.48550/arXiv.2110.14375.
- Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut Learning in Deep Neural Networks. *Nature Machine Intelligence*, 2(11), 665-673. arXiv:2004.07780. doi:10.48550/arXiv.2004.07780.
- Ghosh, S., Seth, A., Kumar, S., Tyagi, U., Evuru, C. K., Ramaneswaran, S., Sakshi, S., et al. (2023). CompA: Addressing the Gap in Compositional Reasoning in Audio-Language Models. ICLR 2024. arXiv:2310.08753. doi:10.48550/arXiv.2310.08753.
- Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., & Parikh, D. (2016). Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. CVPR 2017. arXiv:1612.00837. doi:10.48550/arXiv.1612.00837.
- Grace, A. I., Huo, Z., & Wang, W. (2026). Adaptive Perturbation Selection for Contrastive Audio Decoding. IEEE SLT 2026. arXiv:2607.00247. doi:10.48550/arXiv.2607.00247.
- Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S. R., & Smith, N. A. (2018). Annotation Artifacts in Natural Language Inference Data. NAACL 2018. arXiv:1803.02324. doi:10.48550/arXiv.1803.02324.
- Hase, P., Xie, H., & Bansal, M. (2021). The Out-of-Distribution Problem in Explainability and Search Methods for Feature Importance Explanations. NeurIPS 2021. arXiv:2106.00786. doi:10.48550/arXiv.2106.00786.
- He, H., Du, X., Sun, R., Dai, Z., Xiao, Y., Yang, M., Zhou, J., et al. (2025). Measuring Audio's Impact on Correctness: Audio-Contribution-Aware Post-Training of Large Audio Language Models. arXiv:2509.21060. doi:10.48550/arXiv.2509.21060. [PREPRINT]
- He, H., Sun, R., Dai, Z., Du, X., Wu, C., Liang, Z., Liu, Z., et al. (2026). Summary of DCASE 2026 Task 5: Audio-Dependent Question Answering. arXiv:2607.18718. doi:10.48550/arXiv.2607.18718. [PREPRINT]
- Hessel, J., & Lee, L. (2020). Does my multimodal model learn cross-modal interactions? It's harder to tell than you might think!. EMNLP 2020. arXiv:2010.06572. doi:10.48550/arXiv.2010.06572.
- Hooker, S., Erhan, D., Kindermans, P., & Kim, B. (2018). A Benchmark for Interpretability Methods in Deep Neural Networks. NeurIPS 2019. arXiv:1806.10758. doi:10.48550/arXiv.1806.10758.
- Hsu, T., Lu, K., Chiang, C., & Lee, H. (2025). Reducing Object Hallucination in Large Audio-Language Models via Audio-Aware Decoding. arXiv:2506.07233. doi:10.48550/arXiv.2506.07233. [PREPRINT]
- Jabri, A., Joulin, A., & van der Maaten, L. (2016). Revisiting Visual Question Answering Baselines. ECCV 2016. arXiv:1606.08390. doi:10.48550/arXiv.1606.08390.
- Jouaiti, M., & Ma, N. (2026). All I Hear is Noise: Investigating Clever Hans Effects in Clinical Speech Datasets. IEEE SLT 2026. arXiv:2609.21080. doi:10.48550/arXiv.2609.21080.
- Kafle, K., & Kanan, C. (2017). An Analysis of Visual Question Answering Algorithms. ICCV 2017. arXiv:1703.09684. doi:10.48550/arXiv.1703.09684.
- Kaushik, D., & Lipton, Z. C. (2018). How Much Reading Does Reading Comprehension Require? A Critical Investigation of Popular Benchmarks. EMNLP 2018. arXiv:1808.04926. doi:10.48550/arXiv.1808.04926.
- Kuan, C., Huang, W., & Lee, H. (2024). Understanding Sounds, Missing the Questions: The Challenge of Object Hallucination in Large Audio-Language Models. Interspeech 2024. arXiv:2406.08402. doi:10.48550/arXiv.2406.08402.
- Kumar, S., Sedlacek, S., Lokegaonkar, V., Lopez, F., Yu, W., Anand, N., Ryu, H., et al. (2025). MMAU-Pro: A Challenging and Comprehensive Benchmark for Holistic Evaluation of Audio General Intelligence. arXiv:2508.13992. doi:10.48550/arXiv.2508.13992. [PREPRINT]
- Le Bras, R., Swayamdipta, S., Bhagavatula, C., Zellers, R., Peters, M. E., Sabharwal, A., & Choi, Y. (2020). Adversarial Filters of Dataset Biases. ICML 2020. arXiv:2002.04108. doi:10.48550/arXiv.2002.04108.
- Lee, K., Kim, M., Yoon, S., Kim, M., Lee, D., Koh, H., & Jung, K. (2024). VLind-Bench: Measuring Language Priors in Large Vision-Language Models. Findings of NAACL 2025. arXiv:2406.08702. doi:10.48550/arXiv.2406.08702.
- Lei, J., Berg, T. L., & Bansal, M. (2022). Revealing Single Frame Bias for Video-and-Language Learning. arXiv:2206.03428. doi:10.48550/arXiv.2206.03428. [PREPRINT]
- Li, B., Lin, Z., Peng, W., Nyandwi, J. de D., Jiang, D., Ma, Z., Khanuja, S., et al. (2024). NaturalBench: Evaluating Vision-Language Models on Natural Adversarial Samples. NeurIPS 2024. arXiv:2410.14669. doi:10.48550/arXiv.2410.14669.
- Li, C., Lin, T., & Lee, H. (2025). When Silence Matters: The Impact of Irrelevant Audio on Text Reasoning in Large Audio-Language Models. ICASSP 2026. arXiv:2510.00626. doi:10.48550/arXiv.2510.00626.
- Li, K., Han, S., Tian, Y., & Ji, S. (2026). Reward-Tilted On-Policy Distillation for Acoustic Grounding in Audio-Language Models. arXiv:2609.28778. doi:10.48550/arXiv.2609.28778. [PREPRINT]
- Lu, K., Fu, S., Yang, C. H., Chen, Z., Huang, S., Yang, C., Lin, Y., et al. (2026). How Auditory Knowledge in LLM Backbones Shapes Audio Language Models: A Holistic Evaluation. arXiv:2603.19195. doi:10.48550/arXiv.2603.19195. [PREPRINT]
- Lydakis, G., Hermans, A., Athar, A., de Geus, D., & Leibe, B. (2025). How Important are Videos for Training Video LLMs?. arXiv:2506.06928. doi:10.48550/arXiv.2506.06928. [PREPRINT]
- Ma, Z., Ma, Y., Zhu, Y., Yang, C., Chao, Y., Xu, R., Chen, W., et al. (2025). MMAR: A Challenging Benchmark for Deep Reasoning in Speech, Audio, Music, and Their Mix. arXiv:2505.13032. doi:10.48550/arXiv.2505.13032. [PREPRINT]

- Marino, K., Rastegari, M., Farhadi, A., & Mottaghi, R. (2019). OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge. CVPR 2019. arXiv:1906.00067. doi:10.48550/arXiv.1906.00067.
- McCoy, R. T., Pavlick, E., & Linzen, T. (2019). Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. ACL 2019. arXiv:1902.01007. doi:10.48550/arXiv.1902.01007.
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741-749. doi:10.1037/0003-066X.50.9.741.
- Nam, H. (2026). Learning from Audio-Dependency Errors: Data Curation Strategies Based on Model Confusion Patterns in Audio Question Answering. arXiv:2606.22276. doi:10.48550/arXiv.2606.22276. [PREPRINT]
- Parcalabescu, L., & Frank, A. (2022). MM-SHAP: A Performance-agnostic Metric for Measuring Multimodal Contributions in Vision and Language Models & Tasks. ACL 2023. arXiv:2212.08158. doi:10.48550/arXiv.2212.08158.
- Poliak, A., Naradowsky, J., Haldar, A., Rudinger, R., & Van Durme, B. (2018). Hypothesis Only Baselines in Natural Language Inference. *SEM 2018. arXiv:1805.01042. doi:10.48550/arXiv.1805.01042.
- Rahmanzadehgervi, P., Bolton, L., Taesiri, M. R., & Nguyen, A. T. (2024). Vision language models are blind: Failing to translate detailed visual features into words. arXiv:2407.06581. doi:10.48550/arXiv.2407.06581. [PREPRINT]
- Rouditchenko, A., Bhati, S., Araujo, E., Thomas, S., Kuehne, H., Feris, R., & Glass, J. (2025). Omni-R1: Do You Really Need Audio to Fine-Tune Your Audio LLM?. arXiv:2505.09439. doi:10.48550/arXiv.2505.09439. [PREPRINT]
- Sakshi, S., Tyagi, U., Kumar, S., Seth, A., Selvakumar, R., Nieto, O., Duraiswami, R., et al. (2024). MMAU: A Massive Multi-Task Audio Understanding and Reasoning Benchmark. arXiv:2410.19168. doi:10.48550/arXiv.2410.19168. [PREPRINT]
- Srikanth, N., & Rudinger, R. (2022). Partial-input baselines show that NLI models can ignore context, but they don't. NAACL 2022. arXiv:2205.12181. doi:10.48550/arXiv.2205.12181.
- Teney, D., Kafle, K., Shrestha, R., Abbasnejad, E., Kanan, C., & van den Hengel, A. (2020). On the Value of Out-of-Distribution Testing: An Example of Goodhart's Law. arXiv:2005.09241. doi:10.48550/arXiv.2005.09241. [PREPRINT]
- Thrush, T., Jiang, R., Bartolo, M., Singh, A., Williams, A., Kiela, D., & Ross, C. (2022). Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality. CVPR 2022. arXiv:2204.03162. doi:10.48550/arXiv.2204.03162.
- Tong, S., Brown, E., Wu, P., Woo, S., Middepogu, M., Akula, S. C., Yang, J., et al. (2024a). Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs. NeurIPS 2024. arXiv:2406.16860. doi:10.48550/arXiv.2406.16860.
- Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., & Xie, S. (2024b). Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs. arXiv:2401.06209. doi:10.48550/arXiv.2401.06209. [PREPRINT]
- Yang, C., Ho, N. S., & Lee, H. (2025). Towards Holistic Evaluation of Large Audio-Language Models: A Comprehensive Survey. EMNLP 2025. arXiv:2505.15957. doi:10.48550/arXiv.2505.15957.
- Yang, Q., Xu, J., Liu, W., Chu, Y., Jiang, Z., Zhou, X., Leng, Y., et al. (2024). AIR-Bench: Benchmarking Large Audio-Language Models via Generative Comprehension. ACL 2024. arXiv:2402.07729. doi:10.48550/arXiv.2402.07729.
- Yue, X., Zheng, T., Ni, Y., Wang, Y., Zhang, K., Tong, S., Sun, Y., et al. (2024). MMMU-Pro: A More Robust Multi-discipline Multimodal Understanding Benchmark. ACL 2025. arXiv:2409.02813. doi:10.48550/arXiv.2409.02813.
- Zang, Y., O'Brien, S., Berg-Kirkpatrick, T., McAuley, J., & Novack, Z. (2025). Are you really listening? Boosting Perceptual Awareness in Music-QA Benchmarks. ISMIR 2025. arXiv:2504.00369. doi:10.48550/arXiv.2504.00369.