FACTMIX: EVALUATING AND IMPROVING MIXED-FACT ROBUSTNESS IN LARGE AUDIO-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Audio-Language Models (LALMs) achieve strong performance across speech, environmental sound, music, and audio reasoning tasks, yet they can overlook a single incorrect factual claim when it appears together with several correct claims in the same query. To study this failure, we introduce **FactMix-Bench**, a diagnostic benchmark built from paired *Grounded Queries* and *Mixed-Fact Queries*. Each pair shares the same audio and supporting facts, while the Mixed-Fact Query changes exactly one target fact to a plausible but unsupported alternative. **FactMix-Bench** covers source, content, attribute, temporal, and relation information, with 55,952 multiple-choice questions forming 27,976 pairs over 5,000 recordings. We further construct **FactMix-Align**, a preference dataset using the same controlled factual changes, with directionally balanced ordered perturbations to reduce training bias. Direct Preference Optimization with **FactMix-Align** improves Qwen3-Omni pair accuracy from 54.33 to 76.83, achieving the best overall performance among the evaluated models, while performance on established public audio benchmarks is largely preserved or modestly improved. These results show that strong general audio understanding does not necessarily imply reliable verification of multiple factual claims. The data and code associated with **FactMix-Bench** and **FactMix-Align** will be publicly released upon acceptance.

1 INTRODUCTION

Large Audio-Language Models (LALMs) have rapidly evolved from speech-focused systems into general models capable of understanding speech, environmental sounds, music, and paralinguistic information. Early systems such as SALMONN (Tang et al., 2024), Audio Flamingo (Kong et al., 2024), and Qwen2-Audio (Chu et al., 2024) demonstrated strong general audio understanding and instruction following. More recent models, including Qwen2.5-Omni (Xu et al., 2025a), Qwen3-Omni (Xu et al., 2025b), Audio Flamingo 3 (Ghosh et al., 2025), Step-Audio 2 (Wu et al., 2025), and Qwen3.5-Omni (Qwen Team, 2026), further extend audio perception, reasoning, and spoken interaction. As these models become increasingly capable, an important question is whether they can reliably verify several factual claims presented together in the same query, rather than merely recognize individual properties of an audio recording.

Existing benchmarks evaluate a broad range of audio understanding abilities. AIR-Bench measures instruction following across speech, environmental sounds, and music (Yang et al., 2024); AudioBench evaluates speech, audio-scene, and paralinguistic understanding (Wang et al., 2025); MMAU emphasizes expert-level audio perception and reasoning (Sakshi et al., 2025); MMSU focuses on fine-grained spoken-language understanding and reasoning (Wang et al., 2026); and MMAR evaluates complex reasoning over speech, sound, music, and their mixtures (Ma et al., 2025). These benchmarks provide increasingly challenging tests of audio intelligence, but their questions generally ask models to identify or reason about particular properties of the audio. They do not explicitly examine whether a model can verify several factual claims simultaneously when most are correct but one is not.

Consider an audio clip containing a deep male voice, an enthusiastic female voice, a discussion of university admissions, and background electronic music. A query may correctly describe the

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

speakers, discussion content, and background music, while incorrectly stating that the male voice is high-pitched. Although most of the query agrees with the audio, the query as a whole should be rejected. We find that current LALMs can overlook such an error when it is surrounded by several correct claims. We refer to queries containing both supported and unsupported factual claims as *mixed-fact queries*. Their difficulty arises from the factual composition of the query itself: the model must verify the individual claims against the audio instead of relying on the overall consistency of the description.

This problem is related to, but distinct from, recent work on audio hallucination and grounding. AHa-Bench studies semantic, acoustic, and semantic–acoustic hallucinations (Cheng et al., 2025), while HalluAudio evaluates hallucination across multiple audio domains and task formats (Zhao et al., 2026). Audio-Aware Decoding improves grounding by increasing the influence of acoustic evidence during inference (Hsu et al., 2025), and AHA constructs counterfactual hard negatives for preference alignment against event and temporal errors (Chen et al., 2026). These studies demonstrate that strong performance on standard audio tasks does not necessarily guarantee reliable grounding in the acoustic evidence. Our setting differs in where the challenge is introduced: rather than presenting an unsupported fact in isolation, we place one incorrect claim among several supported claims in the external query and test whether the model can identify the inconsistency.

To study this behavior systematically, we introduce **FactMix-Bench**. For each audio recording, we identify multiple supported facts, select one as the target, and replace only that target with a plausible but unsupported alternative while preserving the remaining facts. This produces paired *Grounded Queries* and *Mixed-Fact Queries* that share the same audio and supporting context and differ in exactly one factual claim. **FactMix-Bench** covers five types of audio information: source, content, attribute, temporal, and relation. It contains 55,952 multiple-choice questions organized into 27,976 Grounded–Mixed-Fact pairs over 5,000 recordings. For a Mixed-Fact Query, the correct answer must not only reject the incorrect target but also recover the corresponding fact supported by the audio.

We further construct **FactMix-Align** from the training split of MECAT-Caption (Niu et al., 2025). It follows the same controlled factual replacement principle as **FactMix-Bench**, but converts each query into a preference instance containing two complete natural-language responses. The audio-consistent response is used as the preferred completion and the alternative response as the rejected completion. We additionally balance counterfactual directions for naturally ordered properties such as duration, count, frequency, tempo, loudness, and intensity, reducing the risk that the model learns a directional shortcut from an imbalanced training distribution. The resulting dataset contains 123,860 preference instances, which we use for Direct Preference Optimization (DPO) (Rafailov et al., 2023).

Experiments demonstrate a substantial gap between conventional audio understanding and mixed-fact robustness. Qwen3-Omni achieves an overall pair accuracy of only 54.33 on **FactMix-Bench**, with particularly low performance on temporal queries. Training with **FactMix-Align** raises overall pair accuracy to 76.83, an absolute improvement of 22.50 points, and achieves the best overall result among the evaluated models. Improvements are observed across all five factual categories, with the largest gain on temporal queries. At the same time, performance on established public audio benchmarks is preserved: relative to the base Qwen3-Omni, the aligned model improves MMAU-TestMini, MMAU-Test, and MMAR by 1.10, 1.12, and 1.20 points, respectively, while MMSU changes by only -0.09 points. Moreover, **FactMix-Align** uses preference supervision over natural-language responses whereas **FactMix-Bench** uses multiple-choice evaluation, showing that the improvement transfers across output formats.

Our main contributions are threefold:

- We identify and systematically study a failure of LALMs on *mixed-fact queries*, where a single incorrect factual claim can be overlooked when it appears together with several correct claims.
- We introduce **FactMix-Bench**, a controlled benchmark containing 55,952 questions over source, content, attribute, temporal, and relation information, with paired Grounded and Mixed-Fact Queries differing in only one target fact.

- We introduce **FactMix-Align**, a preference dataset for improving mixed-fact robustness, and show that preference optimization improves Qwen3-Omni pair accuracy from 54.33 to 76.83 while preserving general audio understanding on established public benchmarks.

2 Related Work

2.1 Large Audio-Language Models

Large Audio-Language Models have progressed from systems that connect pretrained audio encoders with language models toward more general architectures for audio understanding, reasoning, and interaction. LTU combines general audio perception with language-based reasoning (Gong et al., 2023). SALMONN integrates speech and audio encoders with a pretrained language model (Tang et al., 2024), while Audio Flamingo introduces in-context learning and multi-turn interaction for general audio understanding (Kong et al., 2024). Qwen2-Audio scales audio-language training across heterogeneous speech, sound, and music tasks (Chu et al., 2024). GAMA further targets complex reasoning over general audio (Ghosh et al., 2024), and WavLLM focuses on robust speech understanding across semantic and speaker-related information (Hu et al., 2024).

Recent models increasingly integrate audio perception, reasoning, and generation within a unified system. Qwen2.5-Omni introduces a Thinker–Talker architecture for multimodal reasoning and streaming speech generation (Xu et al., 2025a), and Qwen3-Omni extends this design with larger-scale multimodal training and a Mixture-of-Experts architecture (Xu et al., 2025b). Step-Audio 2 combines audio understanding, reasoning, and discrete audio generation (Wu et al., 2025). Audio Flamingo 3 and Audio Flamingo Next further advance general and long-form audio understanding (Ghosh et al., 2025; 2026), while Qwen3.5-Omni extends omnimodal reasoning to longer contexts (Qwen Team, 2026). Despite these advances, strong aggregate benchmark performance does not directly reveal whether a model can verify each factual claim when several claims are composed within the same query.

2.2 Audio Understanding Benchmarks

A growing body of benchmarks evaluates increasingly diverse aspects of audio understanding. AIR-Bench evaluates instruction following across speech, environmental sounds, and music (Yang et al., 2024), while AudioBench covers speech understanding, audio-scene understanding, and paralinguistic information (Wang et al., 2025). VoiceBench evaluates LLM-based voice assistants under variations in speakers, environments, and spoken content (Chen et al., 2024). MMAU targets expert-level perception and reasoning across speech, sound, and music (Sakshi et al., 2025); MMSU provides fine-grained evaluation of spoken-language perception and reasoning (Wang et al., 2026); and MMAR emphasizes deeper reasoning over speech, sound, music, and their mixtures (Ma et al., 2025). MMAU-Pro further extends audio evaluation to long-form audio, spatial reasoning, multi-audio understanding, and both multiple-choice and open-ended settings (Kumar et al., 2025).

These benchmarks substantially expand the breadth and difficulty of audio evaluation, but they primarily test whether a model can recognize, understand, or reason about information contained in the audio. They do not explicitly control the factual composition of the query itself. In particular, they do not systematically pair a fully supported query with a minimally modified query in which exactly one factual claim becomes incorrect while the surrounding claims remain unchanged. **FactMix-Bench** is designed around this controlled comparison, allowing mixed-fact robustness to be evaluated independently of large changes in query content or task format.

2.3 Audio Hallucination and Factual Grounding

Recent work has begun to examine whether Audio-Language Models remain reliably grounded in acoustic evidence. AHa-Bench studies semantic, acoustic, and semantic–acoustic hallucinations (Cheng et al., 2025), while HalluAudio evaluates hallucination across speech, environmental sounds, and music using multiple task formats, including binary judgments, multiple-choice questions, attribute verification, and open-ended QA (Zhao et al., 2026). These studies show that strong performance on conventional audio understanding tasks does not necessarily guarantee reliable factual grounding.

Several methods have also been proposed to reduce grounding errors. Audio-Aware Decoding contrasts predictions with and without audio evidence to reduce unsupported outputs at inference time (Hsu et al., 2025). AHA studies errors involving event identity, omission, and temporal relations, and uses counterfactual hard negatives for preference alignment (Chen et al., 2026). VoxParadox studies a different type of conflict in which misleading paralinguistic information is embedded in the spoken transcript while the external question remains neutral, testing whether models rely on linguistic content or acoustic cues (Pang et al., 2026).

Our setting differs from these formulations in the location and composition of the conflicting information. In **FactMix-Bench**, the inconsistency is introduced in the external query rather than in the audio or transcript. A Grounded Query contains only supported claims, while its paired Mixed-Fact Query preserves the same audio and supporting facts but replaces exactly one target fact with an unsupported alternative. The resulting task therefore isolates whether correct surrounding information causes a model to overlook a specific factual inconsistency. **FactMix-Align** further uses the same controlled factual substitutions for preference alignment, directly connecting the diagnostic failure exposed by the benchmark with a targeted training strategy.

3 Benchmark Construction

3.1 Problem Formulation and Fact Taxonomy

Problem formulation. Given an audio signal x, we consider a natural-language query that contains multiple factual claims about the audio. Let f_t denote one designated target fact and $\mathcal{S}_t$ the set of remaining supported facts. We construct a Grounded Query q^{G} and a paired Mixed-Fact Query q^{M} as

$$\begin{aligned} \mathcal{F}(q^{\mathrm{G}}) &= \mathcal{S}_t \cup \{f_t\}, \\ \mathcal{F}(q^{\mathrm{M}}) &= \mathcal{S}_t \cup \{\tilde{f}_t\}, \end{aligned} \tag{1}$$

where f_t is supported by the audio and $\tilde{f}_t$ is an unsupported alternative for the same factual property. Thus, the two queries share the same audio and supporting facts and differ only in the designated target claim. This controlled construction isolates whether a model can detect one incorrect fact when it appears together with several correct ones.

Fact taxonomy. We organize target facts into five categories. *Source* identifies the sound-producing entity, such as a speaker, instrument, animal, or environmental source. *Content* describes what is spoken or what acoustic event occurs. *Attribute* captures acoustic or paralinguistic properties such as language, emotion, speaking style, loudness, pitch, and recording quality. *Temporal* describes ordering, overlap, repetition, duration, occurrence patterns, and temporal changes. *Relation* captures non-temporal associations between entities or events, such as which speaker produces a particular utterance. The category of an instance is determined by the target fact being modified, while its supporting facts may span multiple categories.

3.2 FactMix-Bench Construction

We construct **FactMix-Bench** from the test split of MECAT-Caption (Niu et al., 2025). Each recording is accompanied by multiple annotations describing speech, acoustic events, music, and recording conditions. As illustrated in Figure 1, the construction pipeline contains four steps: (1) structuring supported facts and selecting targets, (2) generating candidate alternatives and checking them against the audio, (3) constructing and verifying paired queries, and (4) assembling paired multiple-choice questions.

Structure and select facts. We first use GPT-5.5 to convert the complementary MECAT-Caption annotations for each recording into a structured set of supported facts. For each fact, the structured record contains its category, subtype, normalized value, and annotation evidence. At this stage, the model only extracts supported information; it does not generate counterfactuals or questions.

We then apply deterministic rules to identify eligible target facts. Each selected fact becomes a target f_t, while compatible facts from the same recording form its supporting context $\mathcal{S}_t$. Because



Figure 1: Overview of the **FactMix-Bench** construction pipeline. Multi-view annotations are first converted into structured supported facts, from which different target facts can be selected. Candidate alternatives for each target are generated and checked against the original audio. Validated alternatives are then used to construct paired Grounded Queries and Mixed-Fact Queries while preserving the same supporting facts. Finally, each pair is converted into multiple-choice questions. A single audio recording can therefore contribute controlled examples across source, content, attribute, temporal, and relation.

different facts from the same recording can be selected independently, one audio clip may produce multiple target-level instances covering different factual categories.

Generate and audio-check alternatives. For each target fact, we use Qwen3-14B to generate plausible alternative values for the same factual property. Candidate alternatives are required to remain at comparable semantic granularity and to preserve the type of the target. For example, a language is replaced by another language, a voice-quality value by another voice-quality value, and an event-order relation by another ordering of the same events.

An alternative that is absent from the textual annotations is not necessarily false, because the annotations may not exhaustively describe the recording. We therefore verify each candidate directly against the original audio using Qwen3-Omni (Xu et al., 2025b). The verifier assigns one of three labels: SUPPORTED, UNCERTAIN, or CONTRADICTED. We retain only candidates labeled CONTRADICTED; candidates labeled SUPPORTED or UNCERTAIN are discarded. The retained alternative is denoted as $\tilde{f}_t$.

Construct and verify query pairs. Given f_t, $\tilde{f}_t$, and $\mathcal{S}_t$, we use Qwen3-14B to construct a Grounded Query q^{G} and a Mixed-Fact Query q^{M}. The Grounded Query contains the supported target f_t, whereas the Mixed-Fact Query replaces only this target with $\tilde{f}_t$. All supporting facts are kept unchanged.

A separate Qwen3-14B verification pass checks that (i) the Grounded Query contains the supported target, (ii) the Mixed-Fact Query contains the validated alternative, (iii) the supporting facts are preserved in both queries, (iv) the target is the only factual difference between the pair, and (v) no additional unsupported claim is introduced. Pairs that fail any of these checks are discarded rather than repaired.

Table 1: Statistics of **FactMix-Bench**. Each pair contains one Grounded Query and one Mixed-Fact Query constructed from the same target and supporting facts.

Target category	Pairs	Questions
Source	4,718	9,436
Content	7,254	14,508
Attribute	10,400	20,800
Temporal	5,067	10,134
Relation	537	1,074
Total	27,976	55,952

Assemble paired multiple-choice QA. Each validated query pair is converted into two multiple-choice questions. For the Grounded Query, the correct option affirms the supported description. For the Mixed-Fact Query, the correct option rejects the incorrect target and states the corresponding fact supported by the audio. The Mixed-Fact Query therefore requires the model to both identify the incorrect claim and recover the supported target value.

Depending on the number of validated alternatives available for a target, each question contains two to four answer choices. Choices are shuffled, and the correct-answer position is balanced separately within each choice cardinality. The Grounded Query and Mixed-Fact Query derived from the same target share a pair identifier, which enables both question-level evaluation and the stricter paired evaluation introduced in Section 5.1.

Benchmark statistics. **FactMix-Bench** contains 55,952 multiple-choice questions constructed from 5,000 audio recordings, forming 27,976 paired Grounded Queries and Mixed-Fact Queries. Table 1 summarizes the distribution across the five target categories.

Among all questions, 42,896 contain four choices, 4,312 contain three choices, and 8,744 contain two choices. This paired construction makes it possible to measure not only whether a model answers individual questions correctly, but also whether it consistently accepts the supported description and rejects the corresponding minimally modified Mixed-Fact Query.

4 FACTMIX-ALIGN: PREFERENCE ALIGNMENT FOR MIXED-FACT QUERIES

While **FactMix-Bench** evaluates mixed-fact robustness, we construct **FactMix-Align** from the MECAT-Caption training split (Niu et al., 2025) to improve this capability through preference alignment. It follows the same controlled factual replacement principle as **FactMix-Bench**: for each selected target, the supporting facts are kept unchanged, producing a Grounded Query with the supported target and a paired Mixed-Fact Query in which only that target is replaced by a plausible unsupported alternative. The difference is the supervision format: **FactMix-Bench** uses multiple-choice questions for evaluation, whereas **FactMix-Align** converts the paired queries into preference instances for training.

4.1 PREFERENCE DATA CONSTRUCTION

For each Grounded Query or Mixed-Fact Query, we construct two complete natural-language responses. The audio-consistent response is used as the preferred completion y_w, and the incorrect response as the rejected completion y_l, giving

$$(x, q, y_w, y_l), \tag{2}$$

where x is the audio and q is either a Grounded Query or a Mixed-Fact Query. For a Grounded Query, the preferred response affirms the supported target, while the rejected response denies it. For a Mixed-Fact Query, the preferred response rejects the unsupported target and recovers the value supported by the audio, while the rejected response accepts the incorrect target. We retain only pairs in which the alternative changes the same factual property, remains at comparable semantic granularity, and leaves the supporting facts unchanged.

As summarized in Figure 2, the resulting preference pairs are further balanced before DPO training.

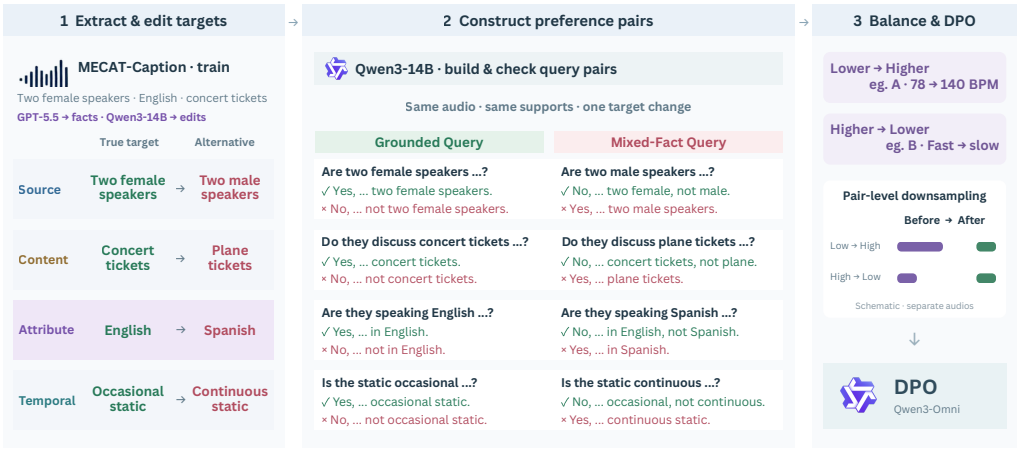

Figure 2: Construction of **FactMix-Align**. Paired Grounded and Mixed-Fact Queries are converted into preference pairs, followed by directional balancing and DPO training.

Directional balancing for ordered facts. Counterfactual substitutions on naturally ordered properties can be directionally imbalanced. For example, changes in duration, count, frequency, tempo, loudness, or intensity may appear more often toward larger or stronger values than toward smaller or weaker values. To reduce this shortcut, we label ordered substitutions as LOWER or HIGHER relative to the supported target and downsample the over-represented direction. Balancing is performed at the Grounded–Mixed-Fact pair level and stratified by factual category, subtype, domain, and ordered axis; unordered facts are left unchanged.

After filtering and directional balancing, **FactMix-Align** contains 123,860 preference instances corresponding to 61,930 Grounded–Mixed-Fact pairs from 15,522 audio recordings. The dataset is constructed exclusively from the MECAT-Caption training split and shares no recordings with **FactMix-Bench**.

4.2 PREFERENCE OPTIMIZATION

We apply Direct Preference Optimization (DPO) (Rafailov et al., 2023) to Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025b). For each training instance, the audio-consistent response is used as the preferred completion and the incorrect response as the rejected completion. We optimize the Thinker-side language-model branch using a frozen copy of the base model as the reference policy.

We use $\beta = 0.1$ and a learning rate of 1×10^{-6}. Training is performed on 32 GPUs for 4 epochs using DeepSpeed ZeRO-3. Additional implementation details are provided in the Appendix.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

We evaluate a diverse set of current Audio-Language Models, including Qwen2.5-Omni-3B and Qwen2.5-Omni-7B (Xu et al., 2025a), Audio Flamingo 3 (Ghosh et al., 2025), Step-Audio-2 (Wu et al., 2025), GPT-4o Audio, Gemini-3.1-Pro, and Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025b). Our aligned model is initialized from Qwen3-Omni-30B-A3B-Instruct and trained with **FactMix-Align** as described in Section 4. Importantly, all results reported in this section are obtained from our own evaluation runs. We evaluate every model ourselves on both **FactMix-Bench** and the public benchmarks; no scores in Tables 2 and 3 are copied from model papers or external leaderboards.

Table 2: Pair accuracy (%) on **FactMix-Bench**. A pair is correct only when both its Grounded Query and Mixed-Fact Query are answered correctly. Best and second-best results are shown in bold and underlined, respectively. Small colored values indicate absolute changes from the base Qwen3-Omni.

Model	Source	Content	Attribute	Temporal	Relation	Overall
Qwen2.5-Omni-3B	72.35	52.57	43.64	12.57	49.21	46.07
Qwen2.5-Omni-7B	86.73	55.27	47.33	14.20	50.31	50.77
Audio Flamingo 3	<u>94.32</u>	71.18	67.44	41.36	64.41	67.74
Step-Audio-2	94.25	72.76	64.47	35.41	**69.35**	67.25
GPT-4o Audio	90.54	68.93	54.51	29.48	59.42	63.52
Gemini-3.1-Pro	**95.04**	<u>75.33</u>	<u>70.22</u>	<u>54.62</u>	<u>68.55</u>	<u>72.75</u>
Qwen3-Omni	89.34	59.33	53.91	15.41	54.56	54.33
+ FactMix-Align	90.14 +0.81	**76.83** +17.49	**81.08** +27.16	**56.82** +41.41	66.48 +11.92	**76.83** +22.50

For **FactMix-Bench**, every model receives the same audio, query, and answer choices. We report pair accuracy (PairAcc), where a target is counted as correct only when both its Grounded Query and Mixed-Fact Query are answered correctly:

$$\mathrm{PairAcc} = \frac{1}{N}\sum_{i=1}^{N} 1\left[\hat{y}_i^{\mathrm{G}} = y_i^{\mathrm{G}} \wedge \hat{y}_i^{\mathrm{M}} = y_i^{\mathrm{M}}\right]. \tag{3}$$

Here, G and M denote the Grounded Query and Mixed-Fact Query, respectively. This paired metric prevents trivial strategies from obtaining high scores: indiscriminately accepting queries fails on the Mixed-Fact member, while indiscriminately rejecting them fails on the Grounded member. We report PairAcc for source, content, attribute, temporal, and relation targets, as well as the overall score.

To examine whether mixed-fact alignment affects general audio capability, we also evaluate the same models ourselves on MMAU-TestMini, MMAU-Test, MMAU-Pro, MMSU, and MMAR, following the corresponding evaluation protocol for each benchmark. These benchmarks cover complementary aspects of speech, general sound, music, perception, and audio reasoning. We report their aggregate scores as percentages.

5.2 Results and Analysis

Table 2 shows that mixed-fact verification remains challenging even for models that perform strongly on general audio tasks. Among the unaligned models, overall PairAcc ranges from 46.07 for Qwen2.5-Omni-3B to 72.75 for Gemini-3.1-Pro. The difficulty is also strongly category dependent. Source is consistently the easiest category across the evaluated models, whereas Temporal is consistently the most difficult. For example, Qwen3-Omni obtains 89.34 on Source but only 15.41 on Temporal, despite achieving substantially stronger results on conventional public audio benchmarks. This contrast indicates that recognizing individual audio properties does not necessarily imply reliable verification when several factual claims are composed within the same query.

Training Qwen3-Omni with **FactMix-Align** raises overall PairAcc from 54.33 to 76.83, an absolute improvement of 22.50 points and the highest overall score among the evaluated models. The improvement occurs across all five target categories: Source increases from 89.34 to 90.14, Content from 59.33 to 76.83, Attribute from 53.91 to 81.08, Temporal from 15.41 to 56.82, and Relation from 54.56 to 66.48. The largest gain occurs on Temporal, where PairAcc increases by 41.41 points. The aligned model does not obtain the highest score in every individual category—Gemini-3.1-Pro remains strongest on Source and Step-Audio-2 on Relation—but it achieves the strongest overall balance, ranking first on Content, Attribute, Temporal, and overall PairAcc. These results suggest that preference alignment improves mixed-fact verification across several factual dimensions rather than only correcting one particular type of error.

The gain on **FactMix-Bench** does not come with a corresponding loss of general audio capability. As shown in Table 3, Qwen3-Omni with **FactMix-Align** improves from 68.60 to 69.70 on MMAU-TestMini, from 67.33 to 68.45 on MMAU-Test, from 60.53 to 61.70 on MMAU-Pro, and from 70.80

Table 3: Performance (%) on public audio benchmarks. All values are obtained from our own evaluation runs. Best and second-best results are shown in bold and underlined, respectively. Small colored values indicate absolute changes from the base Qwen3-Omni.

Model	MMAU-TestMini	MMAU-Test	MMAU-Pro	MMSU	MMAR
Qwen2.5-Omni-3B	61.41	67.46	46.11	57.33	53.80
Qwen2.5-Omni-7B	66.37	69.36	53.67	62.38	56.80
Audio Flamingo 3	70.28	<u>71.37</u>	51.50	64.74	63.35
Step-Audio-2	<u>72.10</u>	70.44	<u>63.26</u>	63.61	61.24
GPT-4o Audio	64.33	63.87	55.39	59.62	66.40
Gemini-3.1-Pro	**74.70**	**73.96**	**70.48**	**80.16**	**82.37**
Qwen3-Omni	68.60	67.33	60.53	<u>71.89</u>	70.80
+ FactMix-Align	69.70 +1.10	68.45 +1.12	61.70 +1.17	71.80 -0.09	<u>72.00</u> +1.20

to 72.00 on MMAR. MMSU changes only slightly, from 71.89 to 71.80. Thus, four of the five public benchmark scores improve, while the remaining benchmark is effectively preserved. The magnitude of these changes is much smaller than the 22.50-point improvement on **FactMix-Bench**, indicating that the main effect of the alignment is on the targeted mixed-fact failure rather than a broad shift in general audio performance.

Finally, the improvement is not explained by simply learning to reject complex queries more often. PairAcc requires both members of each Grounded–Mixed-Fact pair to be answered correctly, so a general rejection strategy cannot improve the metric. In addition, **FactMix-Align** provides preference supervision over complete natural-language responses, whereas **FactMix-Bench** evaluates models using multiple-choice questions. The improvement from 54.33 to 76.83 therefore transfers across supervision and evaluation formats, supporting the interpretation that the model becomes better at distinguishing supported factual descriptions from minimally corrupted ones.

6 Conclusion

We study mixed-fact queries, where a single incorrect factual claim can be overlooked when surrounded by several correct claims. We introduce **FactMix-Bench**, a controlled benchmark that evaluates this failure through paired Grounded and Mixed-Fact Queries, and **FactMix-Align**, a preference dataset for improving mixed-fact robustness. On Qwen3-Omni, **FactMix-Align** improves overall pair accuracy from 54.33 to 76.83, achieving the best overall result among the evaluated models while largely preserving performance on public audio benchmarks. These results highlight that strong general audio understanding does not necessarily imply reliable verification of multiple factual claims and motivate more fine-grained evaluation of factual consistency in Audio-Language Models.

AI use statement

Generative AI tools were used to assist with software implementation and to improve the clarity, grammar, and presentation of the manuscript. All AI-assisted code and text were reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

References

Yanxi Chen, Wenhui Zhu, Xiwen Chen, Zhipeng Wang, Xin Li, Peijie Qiu, Hao Wang, Xuanzhao Dong, Yujian Xiong, Anderson Schneider, Yuriy Nevmyvaka, and Yalin Wang. AHA: Aligning large audio-language models for reasoning hallucinations via counterfactual hard negatives. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 29294–29306. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1464.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T. Tan, and Haizhou Li. Voicebench: Benchmarking llm-based voice assistants. *arXiv preprint arXiv:2410.17196*, 2024.

Xize Cheng, Dongjie Fu, Chenyuhao Wen, Shannon Yu, Zehan Wang, Shengpeng Ji, Siddhant Arora, Tao Jin, Shinji Watanabe, and Zhou Zhao. AHa-Bench: Benchmarking audio hallucinations in large audio-language models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 13242–13270, 2025. doi: 10.52202/085713-0398.

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 preprint arXiv:2407.10759*, 2024.

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S. Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities. *arXiv preprint arXiv:2406.11768*, 2024.

Sreyan Ghosh, Arushi Goel, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. Audio Flamingo 3: Advancing Audio Intelligence with Fully Open Large Audio Language Models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1396.

Sreyan Ghosh, Arushi Goel, Kaousheik Jayakumar, Lasha Koroshinadze, Nishit Anand, Zhifeng Kong, Siddharth Gururani, Sang-gil Lee, Jaehyeon Kim, Aya Aljafari, Chao-Han Huck Yang, Sungwon Kim, Ramani Duraiswami, Dinesh Manocha, Mohammad Shoeybi, Bryan Catanzaro, Ming-Yu Liu, and Wei Ping. Audio flamingo next: Next-generation open audio-language models for speech, sound, and music. *arXiv preprint arXiv:2604.10905*, 2026.

Yuan Gong, Hongyin Luo, Alexander H. Liu, Leonid Karlinsky, and James Glass. Listen, think, and understand. *arXiv preprint arXiv:2305.10790*, 2023.

Tzu-Wen Hsu, Ke-Han Lu, Cheng-Han Chiang, and Hung-Yi Lee. Reducing object hallucination in large audio-language models via audio-aware decoding. In *2025 IEEE Automatic Speech Recognition and Understanding Workshop*, pp. 1–7. IEEE, 2025. doi: 10.1109/ASRU65441.2025.11434595.

Shujie Hu, Long Zhou, Shujie Liu, Sanyuan Chen, Lingwei Meng, Hongkun Hao, Jing Pan, Xunying Liu, Jinyu Li, Sunit Sivasankaran, and Furu Wei. Wavllm: Towards robust and adaptive speech large language model. *arXiv preprint arXiv:2404.00656*, 2024.

Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 25125–25148. PMLR, 2024.

Sonal Kumar, Šimon Sedláček, Vaibhavi Lokegaonkar, Fernando López, Wenyi Yu, Nishit Anand, Hyeonggon Ryu, Lichang Chen, Maxim Plička, Miroslav Hlaváček, William Fineas Ellingwood, Sathvik Udupa, Siyuan Hou, Allison Ferner, Sara Barahona, Cecilia Bolaños, Satish Rahi, Laura Herrera-Alarcón, Satvik Dixit, Siddhi Patil, Soham Deshmukh, Lasha Koroshinadze, Yao Liu, Leibny Paola Garcia Perera, Eleni Zanou, Themos Stafylakis, Joon Son Chung, David Harwath, Chao Zhang, Dinesh Manocha, Alicia Lozano-Diez, Santosh Kesiraju, Sreyan Ghosh, and Ramani Duraiswami. Mmau-pro: A challenging and comprehensive benchmark for holistic evaluation of audio general intelligence. *arXiv preprint arXiv:2508.13992*, 2025.

Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, Kai Li, Keliang Li, Siyou Li, Xinfeng Li, Xiquan Li, Zheng Lian, Yuzhe Liang, Minghao Liu, Zhikang Niu, Tianrui Wang, Yuping Wang, Yuxuan Wang, Yihao Wu, Guanrou Yang, Jianwei Yu, Ruibin Yuan, Zhisheng Zheng, Ziya Zhou, Haina Zhu, Wei Xue, Emmanouil Benetos, Kai Yu, Eng-Siong Chng, and Xie Chen. MMAR: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. In *Advances in Neural Information Processing Systems*, volume 38, pp. 74837–74868, 2025.

Yadong Niu, Tianzi Wang, Heinrich Dinkel, Xingwei Sun, Jiahao Zhou, Gang Li, Jizhong Liu, Xunying Liu, Junbo Zhang, and Jian Luan. MECAT: A multi-experts constructed benchmark for fine-grained audio understanding tasks. *arXiv preprint arXiv:2507.23511*, 2025.

Jiacheng Pang, Ashutosh Chaubey, and Mohammad Soleymani. Do audio llms listen or read? analyzing and mitigating paralinguistic failures with voxparadox. *arXiv preprint arXiv:2605.27772*, 2026.

Qwen Team. Qwen3.5-Omni Technical Report. *arXiv preprint arXiv:2604.15804*, 2026.

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

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. MMAU: A massive multi-task audio understanding and reasoning benchmark. In *The Thirteenth International Conference on Learning Representations*, 2025.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. SALMONN: Towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F. Chen. AudioBench: A universal benchmark for audio large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4297–4316. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.218.

Dingdong Wang, Junan Li, Jincenzi Wu, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen Meng. MMSU: A massive multi-task spoken language understanding and reasoning benchmark. In *The Fourteenth International Conference on Learning Representations*, 2026.

Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, et al. Step-Audio 2 Technical Report. *arXiv preprint arXiv:2507.16632*, 2025.

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 preprint arXiv:2503.20215*, 2025a.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-Omni Technical Report. *arXiv preprint arXiv:2509.17765*, 2025b.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and Jingren Zhou. AIR-Bench: Benchmarking large audio-language models via generative comprehension. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1979–1998. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.109.

Feiyu Zhao, Yiming Chen, Wenhuan Lu, Daipeng Zhang, Xianghu Yue, and Jianguo Wei. HalluAudio: A comprehensive benchmark for hallucination detection in large audio-language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 38797–38816. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1797.

A Appendix