

The Person Behind the Sound: Demystifying Audio Private Attribute Profiling via Multimodal Large Language Models

Lixu Wang, Kaixiang Yao*, Xinfeng Li, Dong Yang[†], Haoyao Li[‡], XiaoFeng Wang and Wei Dong
Nanyang Technological University, *Duke University, [†]The University of Tokyo, [‡]The Hong Kong Polytechnic University

Abstract—Our research uncovers a severe privacy risk associated with multimodal large language models (MLLMs): the ability to infer sensitive personal attributes from audio data, which we call *audio private attribute profiling*. This capability poses a significant threat, as audio can be covertly captured using simple tools. Moreover, compared to images and texts, audio carries unique characteristics, such as tone and pitch, which can be exploited for more detailed attribute profiling. The first major barrier to understanding this threat is the lack of benchmark datasets with profile-level sensitive attribute annotations. Collecting audio data with attribute labels from real-world volunteers is impractical due to legal, ethical, and compliance concerns. To address this challenge, we introduce **AP²**, a well-crafted audio benchmark dataset constructed using public sources and recent TV dramas. On **AP²**, we examine two baseline avenues of profiling sensitive attributes: (1) converting audio to text and applying LLMs, and (2) directly using audio-language models (ALMs). We found that the former suffers from information loss during transcription, while the latter lacks sufficient reasoning capability. To overcome these limitations, we propose **Gifts**, a hybrid framework in which an LLM guides, forensically reviews, and consolidates inferences made by an ALM. **Gifts** mitigates information loss by letting the ALM lead the inference, while the LLM enhances inference accuracy and validity through three phases: guidance, review, and consolidation. Extensive experiments and human evaluations of 50 participants (18-30 years) show that **Gifts** outperforms the MLLM-based baselines, real humans, and traditional inference methods in profiling sensitive attributes, while also being robust under various types of noise. We further study defense strategies at both the model and data levels. Our work demonstrates the feasibility of audio privacy leakage caused by MLLMs, highlights the urgent need for effective defenses, and provides resources to support future research.

1. Introduction

Over the past years, large language models (LLMs) have significantly transformed computing, marking a major shift in how artificial intelligence is developed and applied [1]. In addition to text, researchers have been working to enhance LLMs with the ability to understand and reason over more modalities, such as image and audio [2]. This line of re-

search has given rise to multimodal large language models (MLLMs), which exhibit remarkable performance in perception, information extraction, and sophisticated reasoning across multimedia inputs [2].

However, the rapid advancement of MLLMs also equips adversaries with powerful tools to infer private information from multimedia data [3], [4]. In our research, we investigate a new privacy threat: **audio private attribute profiling by MLLMs**. This refers to the inference of a wide spectrum of personal traits, such as health conditions, socioeconomic status, personality, and habits, not only from spoken words, but also from nuanced vocal cues, speech patterns, and ambient sounds using MLLMs (please see the example on the homepage [5]). This threat is particularly concerning because audio can be covertly captured without the target’s awareness or direct engagement [6], [7]. Adversaries can exploit inexpensive and easily accessible tools, such as wiretaps [8] or directional microphones [9], to surreptitiously record audio in both public and private environments. Furthermore, most people rarely notice the privacy leakage through their audio data: a recent survey [10] found that 18.7% out of 683 participants were somewhat aware that health information can be inferred from audio, while many participants rarely (28.4%) or never (48.5%) notice the possibility of privacy inference from audio data. Therefore, people likely seldom apply protective measures to their audio recordings, making data collection easier for adversaries. In fact, compared to text or images, audio inherently encodes richer contextual and behavioral cues, such as tone, pitch, and hesitation, which can reveal psychological or personality traits, including nervousness, dishonesty, or lack of confidence [11]. Despite its significance, this risk remains underexplored, prompting a critical question to answer: **To what extent can current MLLMs infer sensitive personal attributes from audio?**

Answering this question entails three key challenges. **Challenge I: Lack of Audio Benchmark Datasets with Complete Profile-level Sensitive Attribute Annotations.** Although there are certain public datasets, such as CommonVoice [12], which already include partial demographic annotations, convincingly annotating for other attributes based on limited information is nearly impossible. Another approach is to collect real-world audio recordings along with sensitive attribute labels from volunteers. However, this process is resource-intensive, and complying with ethical and legal policies such as the GDPR [13] presents significant barriers

to collection and use. In addition to the data challenge, there are two other challenges for exploring the potential of MLLMs in profiling sensitive attributes. *Challenge II: Information Loss in Audio-to-Text Conversion Impedes the Use of LLMs for Sophisticated Inference.* While LLMs possess extensive world knowledge and advanced reasoning capabilities, they cannot process raw audio directly. Instead, they rely on textual transcriptions, which inevitably result in a loss of rich acoustic and non-verbal vocal features [14]. Our empirical evidence shows that advanced LLMs achieve 54.6% accuracy when inferring attributes such as age, gender, accent, and health conditions, significantly lower than ALMs, which exceed 80.0%. When provided with raw audio inputs, ALMs demonstrate an 11.0% increase in inference accuracy compared to using textual inputs alone. *Challenge III: Scarce World Knowledge and Limited Reasoning Capabilities in Current ALMs.* Although ALMs can directly perceive acoustic subtleties, they lack the deep reasoning capabilities to accurately infer sensitive attributes. Our preliminary experiments reveal that ALMs struggle with long-context responses and are prone to hallucinations. Moreover, ALMs perform 6.8% worse than LLMs in inferring attributes when both models receive text-only inputs.

Our proposal. To address the first challenge, we constructed a benchmark dataset, AP^2 , with two distinct subsets. The first subset, $\text{AP}^2\text{-Com}$, builds upon CommonVoice [12], and is augmented by additional audio data retrieval from public resources [15], [16] and open-access platforms [17], [18]. The retrieval process prioritized audio segments that reflected specific sensitive attribute values, and each sample was subjected to rigorous cross-validation by independent experts to ensure semantic plausibility and acoustic realism. The second subset, $\text{AP}^2\text{-TV}$, was developed from a carefully curated collection of recent television dramas that are non-sequel stories, grounded in realistic social contexts, and feature well-documented characters. Three expert annotators watched the full series for each selected drama and labeled characters’ sensitive attributes based on a comprehensive review of relevant materials. These annotations were then verified through a strict two-round cross-validation process to ensure both accuracy and reliability.

To address Challenges II and III, we propose a hybrid framework called *Gifts*. Specifically, to compensate for the ALM’s limited world knowledge and reasoning capabilities, *Gifts* employs an LLM to guide the ALM’s inference process and subsequently review its underlying rationale. Since detailed reasoning is essential for effective review, but current ALMs struggle with generating long responses, we design a dialog-based mechanism in which the LLM poses questions and the ALM replies with short, focused answers. This design enables the ALM to convey acoustic cues and non-verbal information in a conclusive form rather than descriptive, thereby reducing information loss. The LLM then consolidates these responses and refines the overall inference, yielding more accurate profiling results.

Extensive experiments on AP^2 show that *Gifts* significantly outperforms ALM-only methods by 9.8% to 40.7%,

LLM-only methods by 15.5% to 23.5%, and ALM–LLM combinations by 4.4% to 27.8%, across both individual attribute and full profile inference. A human study further confirms the superiority of *Gifts* in attribute profiling, which achieves 89.8% accuracy compared to 67.3% for humans. We also validate that *Gifts*’ attribute inference largely depends on acoustic features, while its good performance is not notably affected by information loss during audio-to-text conversion. Besides, *Gifts* remains robust under various input perturbations, including white noise, adversarial noise, prompt variations, and LLM hallucinations.

Finally, to mitigate the risk of MLLMs in inferring sensitive attributes from audio data, we evaluate the efficacy of model-level and data-level defense strategies. At the model level, we leverage and enhance a novel technique called in-context unlearning [19]. This approach intentionally puts incorrect information into MLLMs’ system prompts, making it harder to map sounds to correct sensitive attribute values. At the data level, we employ and enhance a robust noise jamming mechanism inspired by informational masking [20], where similar signals obscure target sounds. This noise disrupts sensitive information embedded within the audio. Extensive experiments demonstrate that both defense strategies significantly reduce the inference accuracy of MLLMs in profiling sensitive attributes from audio.

Contributions. Our contributions are summarized below:

- To our knowledge, we make the first attempt to investigate the risk of employing MLLMs to profile sensitive attributes from general audio.
- We build a benchmark dataset AP^2 based on real-world datasets and the latest TV dramas, which is annotated for studying the audio private attribute profiling.
- We propose a hybrid MLLM framework, *Gifts*, to investigate the risk of audio private attribute profiling to the maximum extent. Extensive experiments reveal that *Gifts* substantially outperforms baselines and real humans. *Gifts* is also robust against various noise.
- We design a comprehensive defense, including model and data levels, against MLLMs’ attribute profiling. Extensive experiments demonstrate the defensive effectiveness.

2. Background

Attribute Inference Attack. Attribute inference attack (AIA) [21] is a definition that has wide significance and is not limited to machine learning (ML) models. In general, suppose a user data sample $x = (u, v)$ consists of non-sensitive features and sensitive features corresponding to two attribute sets \mathcal{U} and \mathcal{V} , respectively. We assume \mathcal{U} is directly or indirectly disclosed in a system \mathbf{S} as it usually correlates to the functionality of \mathbf{S} . In this setup, AIA aims to infer the values v of the sensitive attributes \mathcal{V} by observing and analyzing values of \mathcal{U} either through \mathbf{S} or not. Traditional AIA usually supposes the system \mathbf{S} is developed from the user data. Therefore, most ML AIAs aim to infer the closed attributes of the training data. The representative approach is to plug in numerous values $\{v_1, \dots, v_i, \dots\}$ of sensitive

attributes together with known non-sensitive attribute values \mathbf{u} to obtain many data samples $\{\mathbf{x}_1, \dots, \mathbf{x}_i, \dots\}$ where $\mathbf{x}_i = (\mathbf{u}, \mathbf{v}_i)$. Then AIA adversaries \mathbf{A}_{AIA} query the ML model \mathbf{S} with these synthetic samples to obtain the corresponding outputs, including predictions of certain tasks and confidence in certain cases. In this setup, black-box system access is sufficient to launch AIAs. The adversaries only need to choose the attribute value \mathbf{v}^* with the highest prior probability among the ones that pass the membership test.

$$\mathbf{v}^* = \mathbf{A}_{\text{AIA}}(\mathbf{S}, \mathbf{v}) = \arg \max_{\mathbf{v}_i \in \mathcal{V}} \Pr(\mathbf{v}_i | \mathbf{u}) \cdot \mathcal{O}_{\text{MI}}(\mathbf{S}, \mathbf{x}_i), \quad (1)$$

where \mathbf{v} is the support of the sensitive attributes \mathcal{V} and \mathcal{O}_{MI} is a membership inference oracle [22]. As shown in this formula, the success of AIAs needs the marginal prior of the sensitive attributes $\Pr(\mathbf{v})$ and its conditional probability with the non-sensitive attributes $\Pr(\mathbf{v} | \mathbf{u})$, which means the AIA adversaries need to have a basic idea about the data distribution of their interests and aim to infer the sensitive attribute values of individual data records [21].

Multimodal Large Language Models. MLLMs are good at integrating diverse types of information, such as text, images, and audio, enabling them to perform cross-modal tasks well [2]. Key strengths of MLLMs include their ability to generalize across domains due to large-scale, diverse training datasets and their exceptional reasoning capabilities. Additionally, they enable scalable automation and maintain consistency and objectivity, making them highly efficient for a variety of tasks. ALMs [23], [24], [25], a subdomain of MLLMs, focus specifically on the integration of audio signals and linguistic information. ALMs outperform traditional audio models by comprehensively capturing acoustic features, such as spectral information and pitch, and by employing advanced temporal modeling to analyze associations in audio data. These capabilities enhance performance in tasks such as speaker recognition [26], audio event detection [27], and music generation [28]. Despite these advantages, ALMs face limitations, particularly in the lack of extensive world knowledge and advanced reasoning capability [14]. Besides, most current ALMs cannot process long audio files. For example, Qwen2-Audio [29] and SALAM-ONN [24] can take in audio of up to one minute, while Gemini1.5-Pro [25] struggles with performance degradation as audio duration increases.

3. Problem Setups

3.1. Threat Model

3.1.1. Attack Scenario. The majority of human activities and environments naturally produce a variety of sounds, making it increasingly possible for individuals to become targets of private attribute profiling. Such attacks are pervasive in everyday life, often occurring unintentionally. For example, people may subconsciously gravitate toward strangers with familiar accents during social interactions [30], which is a subtle form of privacy inference.

More concerning, however, are deliberate attacks where adversaries employ specialized tools to capture audio and conduct advanced analyses. We outline common scenarios in which target audio might be collected:

- *Physical Eavesdropping:* Attackers may install devices such as wiretaps, directional microphones, or laser microphones [8], [9] in locations like public spaces, hotels, residences, or office buildings to capture environmental audio.
- *Social Engineering:* By impersonating others or fabricating scenarios, attackers may employ harassing and phishing phone calls or messages to record audio remotely [31].
- *Application Vulnerabilities:* Adversaries exploit malware, trojans, or permission abuse to remotely control smartphones, laptops, wearable devices, or smart home devices, enabling microphone activation for recording [32].
- *Social Media Data Collection:* Attackers scrape audio data from social media, news reports, or online meetings.

3.1.2. Adversary’s Goal. In audio private attribute profiling, given a set of audio data, an adversary’s goal is to infer the target’s sensitive attributes in the most detailed, accurate, and comprehensive way possible. Notably, while such audio data often includes speech that may mention specific times, places, or events, such explicit information is not the primary focus of the adversary in this work. Instead, we propose that the adversary is more interested in inferring indirect private information embedded in human speech that is not explicitly disclosed. Certainly, the adversary also aims to extract as much private information as possible from non-speech environmental sounds. Based on an extensive review of existing privacy inference studies [3], [11], we identify the primary targets of adversaries to be the following attributes, including not only basic demographic ones but also broader human social traits, which we believe each of them represents essential aspects of personal privacy information:

- Age (AGE), Gender (GEN), Accent (ACC), Health (HEA), Habit (HAB), Personality (PER), Social Preference (SOP), Social Stratum (SOS), Income (INC), Occupation (OCC), Education (EDU), Marital Status (MAR).

By inferring these sensitive attributes, the adversary can construct a comprehensive target profile, which may be sold for direct financial gain or used to obtain identity details for targeted phishing, fraud, or other illicit activities. For this threat model, we assume adversaries do not possess the target’s identifying information prior to attribute profiling, which prevents them from directly inferring attributes through online searches.

3.1.3. Adversary’s Knowledge and Capability. The adversary can freely access and utilize various MLLMs, including both open-source and closed-source ones, in different forms (white-box, gray-box, or black-box). The adversary is assumed to have a certain level of experience and expertise in using MLLMs, such as designing, modifying, switching, and configuring models, parameters, and prompt templates for inferring different sensitive attributes. Besides, the adversary may possess certain audio analysis skills, enabling them to

preprocess audio data using various tools and techniques. Finally, adversaries are presumed to have basic knowledge or expertise in private attribute profiling, allowing them to define the potential inference scope for certain attributes.

3.2. Problem Formulation

In audio private attribute profiling, we denote the distribution over the audio data as $\mathbf{D}: \mathcal{X}$ where \mathcal{X} is the domain of attributes. Suppose an adversary collects a dataset \mathcal{D} consisting of n data points from distribution \mathbf{D} , i.e., $\mathcal{D} \sim \mathbf{D}^n$. In this dataset, each data point $\mathbf{x} = (\mathbf{u}, \mathbf{v})$ contains values of U non-sensitive attributes $\mathbf{u} = (u_1, \dots, u_U)$ and V sensitive attributes $\mathbf{v} = (v_1, \dots, v_V)$, such that $(\mathbf{u}, \mathbf{v}) \sim \mathcal{X}$. The support of sensitive attributes \mathbf{v} is denoted by \mathcal{V} . We also assume two projection functions $\phi(\mathbf{x})$ and $\psi(\mathbf{x})$ that map each data point into non-sensitive and sensitive attributes \mathbf{u} and \mathbf{v} , respectively. In our problem, ϕ is an ML model like MLLMs, while ψ can be regarded as an expert system that provides correct attribute values. In this setup, ϕ is trained on datasets \mathcal{D}_ϕ sampled from another data distribution $\mathbf{D}_\phi: \mathcal{X}_\phi \times \mathcal{Y}_\phi$, where \mathcal{Y}_ϕ is the domain of task labels representing general tasks of audio and texts. The attribute domains \mathcal{X}_ϕ may overlap with \mathcal{X} , but the training dataset \mathcal{D}_ϕ of ϕ is strictly disjoint from the collected target audio dataset \mathcal{D} . Then let us consider an adversary \mathbf{A} with certain knowledge of the target data distribution \mathbf{D} who can acquire the non-sensitive attribute values of the collected dataset $\phi(\mathcal{D})$ and aims to infer the sensitive attribute values $\mathbf{v} = \psi(\mathbf{x})$ for each \mathbf{x} in \mathcal{D} . Note that the adversary \mathbf{A} has no direct or indirect access to ψ . If a conditional probability represents the inference process, then

$$\mathbf{A}(\phi(\mathbf{x}), \mathcal{D}, \mathbf{v}) = \arg \max_{\mathbf{v} \in \mathcal{V}} \Pr(\mathbf{v} | \phi(\mathbf{x})). \quad (2)$$

Including non-sensitive attributes in the prompts of MLLMs during attribute inference ensures that determining the sensitive attribute values with the highest conditional probability, as defined in Eq. (2), aligns with the next-token generation of current MLLMs, in particular when the temperature of next-token generation is set close to zero [1].

4. Motivation Study

This section uses a motivation study on existing public datasets to illustrate the feasibility of employing ALMs to infer information about sensitive attributes from audio, which is different from directly inferring the attribute values.

4.1. Setups

Attributes. After extensive investigation, we found existing audio datasets can only support the study of the following attributes: AGE, GEN, ACC, HEA, PER, OCC, HAB.

Datasets. *Common Voice* [12] contains human speeches with annotations of AGE, GEN, and ACC. For HEA, we merge *Movement disorders voice* [33], *TORGO Dataset* [34], and *DAIC-WOZ Database* [35] into a joint dataset containing

speeches of patients with six diseases. Regarding PER, no audio dataset is available with annotations. Therefore, we employ emotion recognition as emotion strongly reflects an individual’s personality [36]. The used dataset is *RAVDESS* [37] that includes speeches of 24 individuals with 8 emotions. Similarly, no dataset with explicit annotations of OCC exists. We employ audio captioning to evaluate whether ALMs can capture occupation-relevant events. We apply GPT-4o [38] to select a subset relevant to occupations from the *Sound Bible* [39]. As for HAB, we adopt *Wild-DESED* [15] that consists of domestic environment sounds in 10 daily events. Notably, although six diseases, emotions, and occupation-relevant and domestic events cannot directly reflect HEA, PER, OCC, HAB, they are correlated; we hypothesize that if ALMs cannot capture such information, they may also struggle with inferring specific attribute values.

Models. We adopt various open-sourced and closed-sourced models to conduct experiments, including *Qwen-Audio-Chat* [29], *SALAMONN-13B* [24], *Qwen2-Audio-Instruct* [23], and *Gemini1.5-Pro* [25]. It is worth noting that GPT-4o [38], strictly speaking, cannot be classified as an ALM, because it is limited to human speech and cannot process general acoustic features.

Metrics. We use absolute accuracy to measure the inference performance of AGE, GEN, ACC, HEA, and PER. For OCC and HAB, we use Qwen2.5-Instruct-14B to calculate a four-level similarity score between annotations and inferences.

Table 1: Accuracy of employing ALMs to infer sensitive attributes with basic prompts, advanced system prompts involved, and outputting reasoning process required.

Models	Attributes						
	AGE	GEN	ACC	HEA	PER	OCC	HAB
Random Guessing	17.3	49.9	4.9	10.9	11.3	3.2	4.5
Qwen-Audio-Chat	21.4	54.4	34.7	4.3	83.3	33.1	53.2
+ Advanced System Prompt	21.2	54.4	35.0	4.4	83.5	33.1	53.2
+ Output Reasoning Process	20.0	47.7	25.5	4.4	80.7	31.2	50.3
SALAMONN-13B	40.4	90.3	36.8	35.2	89.9	36.5	40.7
+ Advanced System Prompt	40.5	90.3	37.0	35.2	89.5	36.6	40.5
+ Output Reasoning Process	36.9	90.0	33.8	33.5	85.7	34.2	37.6
Qwen2-Audio-Instruct	68.0	98.1	39.6	48.8	90.5	34.4	62.3
+ Advanced System Prompt	62.7	92.2	32.4	41.1	84.5	30.7	56.7
+ Output Reasoning Process	56.7	94.0	26.7	33.5	86.0	29.3	60.4
Gemini1.5-Pro	54.8	82.2	30.8	26.5	74.7	28.7	26.4
+ Advanced System Prompt	64.9	99.2	40.9	30.1	88.6	32.2	33.0
+ Output Reasoning Process	54.5	80.9	30.5	26.0	72.5	30.1	23.8

4.2. Experiment Results and Findings

4.2.1. Feasibility of Private Attribute Inference. We used the most basic prompts, such as `Please infer the age of the speaker in the audio.` We did not specify a system prompt, relying instead on the default `You are a helpful assistant.` According to results in Table 1, *even with basic prompts, current ALMs demonstrate relatively higher inference accuracy across most attributes than random guessing.* In contrast, these basic prompts in VLMs are more likely to trigger refusal responses [3].



Figure 1: The overview and building workflow of the AP² dataset that contains two subsets: AP²-Com and AP²-TV.

4.2.2. Ability of Handling Long Context. We further specified more advanced system prompts; please refer to the homepage [5]. The experimental results, shown in Table 1, reveal that Gemini1.5-Pro achieved significant improvements across most attributes. We also tested whether ALMs could output the reasoning process by adding `Respond with the reasoning process` to the user prompt. We observed that the inference accuracy of most ALMs declined. Upon examining the actual responses shown below, *we observed severe hallucinations in ALMs’ responses during private attribute profiling*, which aligns with studies that found similar phenomena in other audio tasks [14].

Example of significant hallucinations in Qwen2-Audio’s response when outputting reasoning process

Ground truth of the accent: Scottish

User:

<Audio>. Please infer the accent of the speaker.

Assistant:

The speaker has a Scottish accent.

User:

<Audio>. Please infer the accent of the speaker. Respond with the reasoning process.

Assistant:

American. The distinctive features include the pronunciation of certain words like ‘threw’ (with a ‘th’ sound), ‘like’ (with an ‘l’ sound), and the use of ‘z’ in words like ‘zoo’. Additionally, the intonation and rhythm tend to be more open and free-flowing.

4.2.3. Feasibility of Inferring Other Attributes. We also explored the feasibility of inferring other attributes:

Example of using Gemini1.5-Pro to infer SOP

Event description of the audio: There are multiple people engaging in conversations at a dining table, punctuated by the clinking of dishes and silverware.

User:

<Audio>. Please infer the social preference of the person in the audio clip.

Assistant: This person maintains a circle of close, like-minded friends. He exhibits a preference for social engagement, likely finding stimulation and enjoyment in group settings.

We observed that even with basic prompts, ALMs can provide reasonable inferences for other sensitive attributes. However, *due to the lack of annotated datasets for these attributes, we were unable to conduct a detailed study.*

5. Dataset of Audio Private Attribute Profiling

As shown in the motivation study, existing datasets include only limited sensitive attribute annotations, insufficient for studying audio private attribute profiling. Building a suitable dataset by collecting real-world audio and sensitive attributes from volunteers is impractical due to high costs, limited participation, and complex ethical or disclosure concerns. Therefore, we adopt two complementary strategies: (i) *first annotate then collect*: retrieving and composing public audio datasets to create AP²-Com, and (ii) *first collect then annotate*: constructing AP²-TV from recent TV dramas. These two benchmark datasets (Figure 1) encompass a broad range of individuals, with sensitive attributes annotated by experts in audio, sound, and speech analysis. Detailed annotation rules are provided in Appendix A.

5.1. AP²-Com

5.1.1. Key Features. The dataset consists of 80 individuals, each is represented by at least six audio clips. Some clips only contain background sounds, while others may include some dialogue or human speech. The dataset considers all attributes in the threat model (Section 3.1) for each individual. These attributes are not explicitly disclosed in the audio, but there are contextual cues.

- *Diverse Audio Contexts.* The dataset reflects a broad spectrum of daily life activities, including professional settings, social interactions, personal routines, and so on.

- *Retrieval Validity.* We engaged 3 experts to retrieve relevant audio samples for each sensitive attribute value from public audio datasets. The resulting audio samples underwent rigorous multi-round cross-validation to guarantee both accuracy and validity.

- *Ethic Considerations.* The audio clips are sourced from authentic, real-world audio within public datasets with non-commercial research and second-use licenses. Furthermore, each individual in AP²-Com is constructed through combinations of diverse attributes, ensuring that no individual maps to a real-world person, adhering to ethic research practices without sacrificing realism.

5.1.2. Dataset Construction Process. We primarily build upon CommonVoice [12], which contains annotations of speakers’ age, gender, and accent, and utilize other datasets [33], [34], [35] to enrich the diversity of these demographic attributes. For other attributes, we first curated an information repository for each attribute based on Wikipedia, with each entry representing a potential attribute value. Subsequently, three experts performed targeted retrieval of audio data from existing public resources (including WildDESED [15], Audioset [16], Pixabay [17], and Freesound [18]), identifying samples indicative of typical behaviors, dialogues, and activities associated with each entry. Such retrieval follows the rule: strategic and iterative data acquisition that leverages contextual audio cues, rigorous selection to ensure relevance to representative behaviors and attributes, and strict adherence to ethical standards, including platform policies and privacy considerations. Each expert’s retrieved results were then subjected to a two-round cross-validation process by the other two experts. After the retrieval, we randomly assigned the validated attribute values and corresponding audio samples to speakers within the CommonVoice corpus. Experts then filtered out those speakers whose assigned attribute values didn’t align well with the speaker’s voice characteristics in semantic plausibility and acoustic realism. To maintain consistency and authenticity across audio clips, we also employed a suite of audio composition techniques, including voice transfer [40], audio splicing, and audio mixing.

5.2. AP²-TV

5.2.1. Key Features. This dataset is collected from six recent American TV dramas (Brilliant Minds, English Teacher, Nobody Wants This, No Good Deed, Running Point, Suits: LA), including 40 characters in total. Each includes at least 12 audio clips and is annotated with sensitive attribute values.

- *Non-sequel Story.* To prevent MLLMs from inferring sensitive attributes based on memorization rather than reasoning, we restricted our selection to non-sequel drama series that premiered after the cut-off date (Sept. 2024) of most MLLMs we used.
- *Real-world Relevance.* The dramas were grounded in modern societal contexts, with primary narratives focusing on the everyday lives of ordinary individuals, excluding highly fantastical or exceptionally rare occurrences.
- *Annotation Feasibility.* To ensure accurate and objective attribute annotation, we prioritized well-documented and widely discussed drama series. The availability of substantial promotional materials, media coverage, and online forum discussions (e.g., IMDb, Rotten Tomatoes, major entertainment outlets) provided the necessary context for experts to annotate attributes.
- *Ethical Considerations.* The dataset includes only character-level audio extracted from licensed media and does not involve the real identities of the actors or actresses. All annotated attributes refer exclusively to the fictional characters as portrayed in the shows, rather than to the

personal traits of the performers. The selected TV dramas feature non-overlapping casts, and the producers chose the actors and actresses based on their suitability for the respective roles, thereby ensuring social-cultural alignments between character attributes and performers’ traits.

5.2.2. Dataset Construction Process. For each character within each selected drama series, we initially employed face recognition techniques to identify their temporal occurrences within each episode. A subset of these timestamps was randomly sampled and manually verified for accuracy and realism, which were used to locate the audio clips for each character. For attribute annotation, comprehensive online searches were conducted to gather relevant materials and information for each series, which were then provided to the three expert annotators. Based on the relevant information and a complete viewing of each series, the experts annotated the sensitive attributes of the characters. The annotation follows the principle: conducting comprehensive analysis based on the full series and supplementary materials, maintaining objectivity and consistency across all annotations, ensuring each label is supported by clear evidence, and accounting for character development by considering attributes across the entire narrative arc. The resulting annotations were subjected to a two-round cross-validation process by the three experts to ensure the validity.

6. Methodology

Overview. To examine the risks of audio private attribute profiling, we propose *Gifts* framework, which comprises two MLLMs operating under a Guidance-inference-forensics-scrutinization-consolidation strategy, which is visualized as Figure 2. Specifically, *Gifts* consists of an ALM ϕ^A and an LLM ϕ^L . *Gifts* starts with the LLM generating some guidance to help the ALM better infer the value of a sensitive attribute v_i . After obtaining the initial inference result \hat{v}_i from the ALM, the LLM then asks the ALM some forensic questions that reflect the supporting clues of the inference result \hat{v}_i . Next, the LLM conducts scrutinization to judge whether the inference result \hat{v}_i is reasonable by checking the forensic question answers of the ALM. If the initial inference result \hat{v}_i is deemed reasonable, it is regarded as a candidate result \tilde{v}_i . Otherwise, the ALM launches a second round of inference to obtain a different result \hat{v}'_i . In the latter case, in the phase of scrutinization, the LLM needs to decide which result (\hat{v}_i, \hat{v}'_i) is more reasonable to be the candidate \tilde{v}_i . In the final phase of consolidation, the LLM aggregates all candidate inference results from multiple audio data samples (of the same individual) to obtain the final result v_i^* . *Notably, Gifts employs independent sessions for each ALM/LLM usage, preventing long prompt chain accumulation.*

6.1. Guidance

Based on our motivation study and other ALM benchmark works [14], although ALMs perform reasonably well

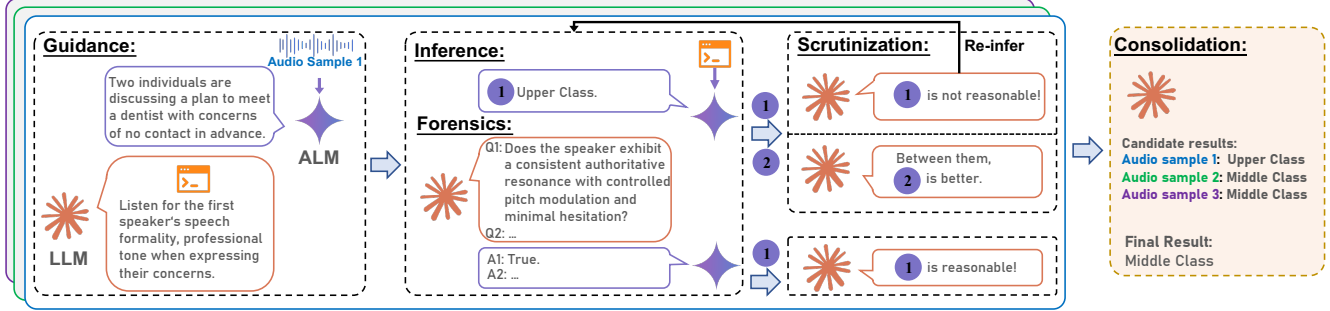


Figure 2: An example of employing the Gifts framework to infer the social stratum of a target from their audio data.

on audio data, there remains a significant gap compared to the performance of LLMs on language tasks, particularly those requiring deep and complex reasoning. To enhance the zero-shot reasoning capabilities of ALMs, we propose leveraging a strong LLM to generate guidance. This guidance can help ALMs focus on the most relevant aspects of a specific task, preventing distractions from irrelevant information.

Specifically, for a sensitive attribute v_i , the LLM ϕ^L possesses general knowledge and can produce general guidance. For example, if v_i is Income (INC), the general guidance may be `You should analyze speech characteristics, content, context, and paralinguistic features in the audio.` As we can see, this guidance is too general to provide dedicated instructions for the inference of specific attributes. To generate more specified and dedicated guidance, ϕ^L needs to know what happens in the audio data. To this end, we adopt the ALM ϕ^A to disclose some non-sensitive attribute values $\mathbf{u}_g = \phi^A(\mathbf{x}, p_g^A)$ for the guidance generation, i.e., operate event captioning and spoken word transcription on an audio sample \mathbf{x} , where p_g^A is the prompt template. Then, the LLM ϕ^L can generate more pertinent guidance g with the help of audio event descriptions and spoken word transcriptions \mathbf{u}_g ,

$$g(\mathbf{u}_g, v_i) = \phi^L(p_g^L \oplus \{\mathbf{u}_g, v_i\}), \quad (3)$$

where p_g^L is the prompt template for generating the guidance, and \oplus represents forming the prompt with \mathbf{u}_g and v_i .

6.2. Inference

With the guidance from the LLM, the ALM starts to infer the sensitive attribute value v_i . To generate the prompt of inference, some additional auxiliary information is required. First of all, for some attributes, defining the scope of inference is needed (please see Appendix B.1). Moreover, some audio data may have the time they were recorded, which is another type of information that can be included in the prompt. Besides, if multiple individuals are speaking in the audio, indicating the individual of interest in the prompt is needed. We use “*the ordinal speaker*” in the order of first speaking to indicate different individuals, which is realized by speaker recognition [26]. Then we can combine such

auxiliary information with the guidance from the LLM to form the ALM prompt $P_i^A(v_i)$,

$$P_i^A(v_i) = p_i^A \oplus \{g(\mathbf{u}_g, v_i), scope(v_i), time(\mathbf{x}), speaker(\mathbf{x})\}, \quad (4)$$

where $scope()$, $time()$, $speaker()$ denote obtaining the inference scope, the time, and the speaker indication of a certain attribute or data sample, respectively. Finally, by querying the ALM ϕ^A , we can obtain the initial inference result as

$$\hat{v}_i = \phi^A(P_i^A(v_i), \mathbf{x}). \quad (5)$$

In fact, the ALM’s mission in inference is not yet complete. Let us recall the overview where we mentioned that when the LLM determines the initial inference result of the ALM to be unreasonable, the ALM needs to perform a second inference. Specifically, when ϕ^A re-infers the attribute v_i , the prompt used P_i^A is almost identical to the one used in the initial inference. The only difference is the initial inference \hat{v}_i is negated into $-\hat{v}_i$ and included in the prompt,

$$P_i^A = P_i^A \oplus -\hat{v}_i. \quad (6)$$

We negate \hat{v}_i using phrases like `Your inference should not express similar meanings as.`

6.3. Forensics

In the motivation study, we highlighted that a major challenge in applying current ALMs to private attribute profiling lies in their insufficient ability to generate long-context responses. This limitation prevents the ALM from providing the reasoning process and supporting evidence for its inference while keeping the inference result consistent. However, to assess the validity of the ALM’s inference results, obtaining related evidence and clues is essential.

Our solution is to allocate the task of generating supporting evidence and clue descriptions for the inference results to the LLM ϕ^L , while the ALM ϕ^A evaluates whether these pieces of evidence and clues are valid based on the audio. Specifically, we use the ALM’s inference result \hat{v}_i , along with the time and speaker information, and the previously obtained audio event description and spoken word transcription \mathbf{u}_g , to form a prompt with a forensic template p_f^L and generate Q clue-validation questions $\mathcal{Q} = \{q_i\}_{i=1}^Q$,

$$\mathcal{Q}(\hat{v}_i) = \phi^L(p_f^L \oplus \{\hat{v}_i, \mathbf{u}_g, time(\mathbf{x}), speaker(\mathbf{x})\}). \quad (7)$$

These questions are constrained to be concise and are subsequently fed into the ALM along with the audio data,

$$\mathcal{A}(\hat{v}_i) = \phi^A(\mathcal{Q}(\hat{v}_i), \mathbf{x}). \quad (8)$$

Each answer a_i in \mathcal{A} is only one word, “True”, “False” or “Uncertain”. Since these questions reflect clues that support the inference result, a higher proportion of “True” indicates greater confidence.

6.4. Scrutinization

The phase of Scrutinization follows the LLM-as-a-judge mechanism [41]. In this process, the LLM evaluates the reasonableness of the ALM’s inference result by scrutinizing the audio event description, spoken word transcription, and the ALM’s answers to the clue-validation questions. The scrutinization prompt is

$$P_s^L(\hat{v}_i) = p_s^L \oplus \{\hat{v}_i, \mathcal{Q}(\hat{v}_i), \mathcal{A}(\hat{v}_i), \mathbf{u}_g, \text{time}(\mathbf{x}), \text{speaker}(\mathbf{x})\}, \quad (9)$$

where p_s^L is a scrutinization prompt template. If \hat{v}_i is deemed reasonable, i.e., $\phi_L(P_s^L(\hat{v}_i)) = \text{Yes}$, it is marked as a candidate result \tilde{v}_i . If it is deemed unreasonable, the ALM initiates a second inference with the prompt of Eq. (6). It is worth noting that for a sensitive attribute of a specific data sample, we limit the ALM to a maximum of two inference rounds. This limitation is based on our observation that when the ALM operates with a low temperature, its responses rarely introduce a third interpretation, even if the prompt has explicitly negated two prior inference results. We believe this is due to the auto-regressive mechanism [38].

The process for the ALM’s second inference is nearly identical to the initial inference, including the phase of Forensics. However, during the subsequent phase of Scrutinization, the LLM should evaluate which of the ALM’s two inference results \hat{v}_i and \hat{v}'_i is more reasonable based on the audio event description and spoken word transcription \mathbf{u}_g , and the ALM’s clue-validation questions \mathcal{Q} and answers \mathcal{A} . Accordingly, the prompt is as follows,

$$P_{s'}^L(\hat{v}_i, \hat{v}'_i) = p_{s'}^L \oplus \{\hat{v}_i, \mathcal{Q}(\hat{v}_i), \mathcal{A}(\hat{v}_i), \hat{v}'_i, \mathcal{Q}(\hat{v}'_i), \mathcal{A}(\hat{v}'_i), \text{time}(\mathbf{x}), \text{speaker}(\mathbf{x})\}, \quad (10)$$

where $p_{s'}^L$ is the prompt template. The more reasonable result is then marked as the candidate $\tilde{v}_i = \phi^L(P_{s'}^L(\hat{v}_i, \hat{v}'_i))$.

6.5. Consolidation

In Section 2, we highlighted that current ALMs struggle with processing long audio files. As a result, adversaries may divide long audio into multiple short segments for private attribute profiling. To deal with such cases, the final step of `Gifts` framework involves using the LLM ϕ^L to consolidate multiple (assuming K) candidate inference results $\{\tilde{v}_{i,k}\}_{k=1}^K$ for a particular attribute v_i to produce a final inference result v_i^* . To achieve the most accurate inference, ϕ^L should comprehensively consider various inputs,

including audio event descriptions and spoken word transcriptions $\{\mathbf{u}_{g,k}\}_{k=1}^K$, all candidate results $\{\tilde{v}_{i,k}\}_{k=1}^K$, and the corresponding clue-validation questions $\{\mathcal{Q}(\tilde{v}_{i,k})\}_{k=1}^K$ and answers $\{\mathcal{A}(\tilde{v}_{i,k})\}_{k=1}^K$. Besides, the inference scope, time information, and speaker indication used during the ALM’s inference must also be taken into account. In this case, the consolidation prompt $P_c^L(v_i)$ would be,

$$P_c^L \oplus \{\{\tilde{v}_{i,k}\}_{k=1}^K, \{\mathcal{Q}(\tilde{v}_{i,k})\}_{k=1}^K, \{\mathcal{A}(\tilde{v}_{i,k})\}_{k=1}^K, \{\mathbf{u}_{g,k}\}_{k=1}^K, \{\text{scope}(\mathbf{x}_k)\}_{k=1}^K, \{\text{time}(\mathbf{x}_k)\}_{k=1}^K, \{\text{speaker}(\mathbf{x}_k)\}_{k=1}^K\}, \quad (11)$$

where p_c^L is the prompt template. Then the final consolidated inference result is $v_i^* = \phi^L(P_c^L(v_i))$. It is important to note that for the final consolidated inference of v_i , we do not provide the LLM with previously inferred values of other sensitive attributes $v_{1 \sim i-1}^*$. This is due to our inability to determine the accuracy of these attribute values. Introducing potentially inaccurate information may compromise the accuracy of the current attribute’s inference, which has been validated by experiments in Appendix C.3. Moreover, we do not allow the LLM to guide the ALM in extracting additional relevant clues for v_i , as many acoustic features or clues are difficult to describe effectively in text.

7. Experiments

7.1. Experimental Setups

Gifts Implementation. Following related benchmark works [14], we adopt Qwen2-Audio-Instruct [23] to generate event descriptions of audio. The Gemini1.5-Pro [25] is used to transcribe the speeches and as the ALM in `Gifts`. As for the LLM, we choose Claude3.5-Sonnet [42] to operate the `Gifts` framework. The temperature is set to 0.1. The maximum token length is set to 5,000. We repeatedly run the major experiments three times, and report the average performance with the variance.

Comparison Baselines. We first employ LLMs and Qwen2-Audio and Gemini1.5-Pro to profile sensitive attributes with a captioning-first technique [14], [43] to provide audio event descriptions and spoken word transcriptions. The used LLMs include Qwen2.5-Instruct-14B [44], Llama3-Instruct-8B [45], GPT-4o [38], and Claude3.5-Sonnet. Then we employ only one ALM to profile sensitive attributes. Specifically, every audio clip of each individual in $\mathbb{A}P^2$ dataset is fed into a particular ALM (Qwen-Audio-Chat [29], SALAMONN-13B [24], Qwen2-Audio-Instruct, and Gemini1.5-Pro) to infer every attribute. After that, we feed the inference results of all clips to the ALM again to consolidate into a unified result. Following this workflow, the second type of ALM-based approaches uses Captioning-first LLMs to consolidate inference results. For all these baseline approaches, we employ the same system prompts and hyper-parameter setups as our `Gifts` framework. Besides, although it is costly to implement audio privacy inference approaches built on regular ML techniques (see Section 10), we try our best to find some to compare

Table 2: Performance comparison between *Gifts* and other baseline approaches in private attribute profiling on AP²-Com.

Inputs	Model/Attributes	AGE	GEN	ACC	HEA	HAB	PER	SOP	SOS	INC	OCC	EDU	MAR	Avg
Text Only	Qwen2.5-Instruct-14B	77.3±2.3	79.7±0.5	34.5±1.1	30.7±3.5	64.1±1.3	70.2±0.9	70.3±1.7	76.3±1.2	76.9±1.4	67.2±0.6	82.5±0.9	81.4±2.0	67.6±1.5
	Llama3-Instruct-8B	75.3±2.4	81.7±1.7	39.2±2.0	39.0±3.7	66.3±1.5	63.6±3.4	70.5±0.4	79.7±1.3	75.0±1.9	73.3±1.8	81.4±0.8	82.9±2.2	69.0±2.0
	GPT-4o	80.5±1.5	77.3±1.8	35.2±0.6	35.5±4.0	67.4±2.2	70.6±0.4	69.1±1.3	78.3±2.0	81.4±0.3	77.1±0.9	78.3±1.6	82.2±1.2	69.4±1.5
	Claude3.5-Sonnet	79.5±1.1	73.4±1.7	23.4±2.0	45.4±0.4	72.1±3.3	67.3±1.4	69.6±2.2	81.9±1.0	83.5±0.7	76.9±1.9	82.7±0.8	79.1±1.2	69.6±1.4
Text+ Audio	Qwen2-Audio-Instruct	76.3±1.5	76.0±2.9	30.2±3.7	28.4±2.0	60.8±1.7	68.0±1.2	65.0±1.7	75.2±1.8	76.5±1.1	67.0±2.4	79.5±1.5	50.8±1.9	62.8±1.8
	Gemini1.5-Pro	78.2±1.0	80.6±1.4	36.6±2.7	38.8±1.2	62.7±4.1	64.5±1.5	63.4±2.1	77.3±1.9	77.9±1.2	71.3±1.1	79.7±1.5	59.5±2.3	65.9±2.0
Text+ Audio	Qwen-Audio-Chat	64.5±1.6	67.4±3.2	40.4±4.3	42.7±5.6	56.8±2.0	60.3±2.0	64.8±0.9	67.8±4.0	60.2±2.7	40.9±3.2	50.5±0.8	40.2±3.6	54.7±2.8
	SALAMONN-13B	83.7±2.0	86.1±3.1	59.7±2.9	88.9±1.9	60.1±1.4	59.0±2.1	58.7±1.9	83.1±1.6	75.6±1.3	43.9±2.5	83.7±1.7	79.2±3.7	71.8±2.2
	Qwen2-Audio-Instruct	86.3±1.4	85.9±2.6	60.0±4.2	90.6±2.5	70.2±1.1	70.5±1.0	65.2±3.3	79.4±2.0	80.3±2.4	60.2±1.7	82.3±1.0	52.3±2.8	73.6±2.1
	Gemini1.5-Pro	88.1±1.7	100±0	74.7±3.4	78.6±2.7	65.6±2.0	66.5±1.9	64.9±1.4	80.6±2.1	80.6±1.9	74.3±1.5	82.4±1.2	66.0±3.0	76.9±1.9
Gifts (ours)		92.3±0.8	100±0	78.8±1.1	97.8±1.3	76.6±0.6	74.7±1.3	78.7±1.8	89.2±2.0	87.5±1.4	82.9±1.5	86.7±0.8	95.7±2.4	86.7±1.3

with *Gifts*. The experiment results can be found in Appendix C.6.

Evaluation Metrics. We categorize the sensitive attributes into four types and employ different metrics to evaluate the performance: 1) Qualitative attributes: GEN, MAR; 2) Quantitative attributes: AGE, SOS, INC; 3) Fuzzy attributes: ACC, PER, SOP, OCC, HAB; 4) Hybrid attributes: HEA, EDU. We use absolute error accuracy for qualitative attributes to measure the experiment results. For quantitative attributes, we use relative error accuracy to measure the inference results. To ensure the evaluation objectivity for fuzzy attributes, we consistently use a relatively strong LLM, Claude3.7-Sonnet [42], to conduct a fuzzy evaluation that produces a five-level similarity score. Ultimately, we utilize a hierarchical accuracy measure for HEA to assess the inferred disease and the ground truth at varying levels of granularity. As for EDU, we employ relative error accuracy to measure the inferred education level, and then use Claude3.7-Sonnet to evaluate the inferred major. Each attribute’s inference receives a score between 0 and 1, the higher the better, and we multiply it by 100 to report. More details can be found in Appendix B.1.

7.2. Risk of Audio Private Attribute Profiling

Performance Analysis on AP²-Com. As shown in Table 2, the proposed *Gifts* framework consistently outperforms all baseline methods across all sensitive attributes, achieving up to a 32.0% absolute improvement in average inference accuracy. *These results demonstrate that Gifts enables the most accurate reconstruction of sensitive attribute profiles.* Table 2 further reveals several key insights. First, MLLMs are highly effective at inferring sensitive attributes from audio data, with many approaches performing on par with humans (Section 7.3). Additionally, LLMs and ALMs exhibit complementary strengths across different attribute types. To capture this distinction, we categorize the attributes into two groups: Acoustic-driven and Reasoning-driven, based on whether the highest accuracy is achieved by ALMs or LLMs, respectively (highlighted in green and blue in Table 2). Furthermore, a comparison between ALMs using text-only (62.8%, 65.9%) versus audio+text (73.6%, 76.9%) inputs *reveals significant information loss when audio is represented solely through event descriptions and speech transcripts.* Lastly, the substantial performance gap

between state-of-the-art LLMs (69.6%) and ALMs (65.9%) when provided with identical text-only inputs *highlights the limited reasoning capabilities of current ALMs in inferring sensitive attributes.*

Performance Analysis on AP²-TV. We further conduct extensive experiments on AP²-TV, with results presented in Figures 3 and 4, organized according to Acoustic-driven and Reasoning-driven attributes. A clear observation is that *our Gifts framework consistently achieves the highest inference accuracy across all attributes, underscoring the severity of the privacy risks posed by MLLM audio profiling.* Notably, acoustic features play a crucial role in enabling such profiling. This is evidenced by the relatively poor performance of LLM-only baselines, which is significantly improved when combined with Gemini1.5-Pro, particularly for Acoustic-driven attributes. Overall, ALMs outperform LLMs on AP²-TV, yet incorporating LLMs still contributes to performance gains in both Acoustic-driven and Reasoning-driven attribute inference.

Ablation Study. We first remove the guidance generated by the LLM from the ALM prompts, which significantly reduces inference accuracy as shown in Table 3, thereby validating the importance of guiding the ALM’s focus toward critical features. Next, we eliminate the Forensics phase, meaning that the LLM does not pose any clue-validation questions to the ALM. Notable performance drops in the inference of specific attributes are presented in Table 3. Subsequently, we disable the Scrutinization phase, allowing the ALM’s initial inference results to become the candidate results directly. The consistent decline in accuracy across all attributes highlights the crucial role of LLM scrutinization. Finally, we replace *Gifts*’s Consolidation phase with the simple consolidation strategy. As shown in Table 3, we observe a clear performance decline, demonstrating that the Consolidation phase is vital for achieving optimal performance. In addition to component ablation, we further validate *Gifts*’s generalization (3.5%-15.0% improvement) across various models with weaker or stronger capabilities (please refer to Appendix C.1).

Influence Factor Analysis. We first examine the impact of information loss during audio-to-text conversion on the performance of *Gifts*. To this end, we replace Claude3.5-Sonnet in the original *Gifts* with Gemini1.5-Pro. In addition to employing Gemini1.5-Pro as an LLM in *Gifts*

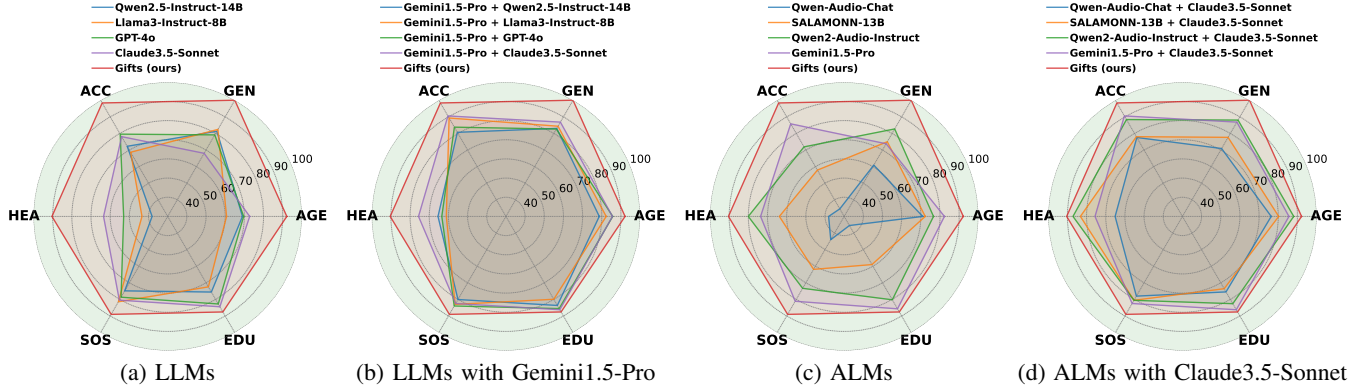


Figure 3: Performance comparison between *Gifts* and other baselines in profiling Acoustic-driven attributes on AP²-TV.

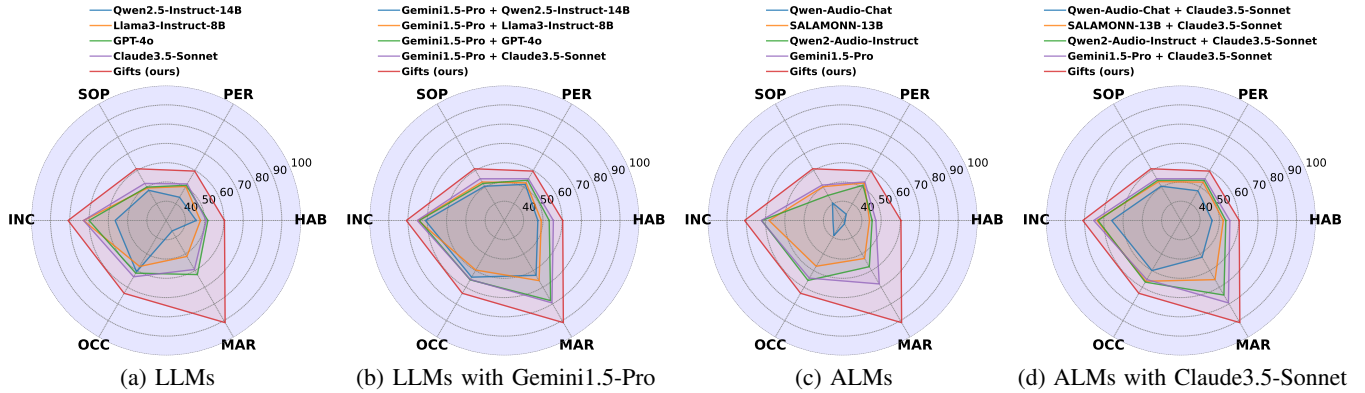


Figure 4: Performance comparison between *Gifts* and other baselines in profiling Reasoning-driven attributes on AP²-TV.

Table 3: Ablation studies of different phases in the proposed *Gifts* framework on AP²-Com.

Model/Attributes	AGE	GEN	ACC	HEA	HAB	PER	SOP	SOS	INC	OCC	EDU	MAR	Avg
<i>Gifts</i> w/o Guidance	91.7±0.3	93.0±0.8	75.5±1.2	90.7±1.6	72.0±1.4	72.2±1.2	75.2±0.5	83.1±1.7	84.2±1.3	77.8±2.4	83.2±0.6	88.4±1.5	82.3±1.2
<i>Gifts</i> w/o Forensics	91.5±0.3	92.5±0.2	75.0±0.2	82.0±2.7	69.8±1.9	71.8±1.4	74.8±2.6	80.2±3.5	82.0±4.2	71.9±3.4	81.5±2.5	80.5±4.8	79.5±2.3
<i>Gifts</i> w/o Scrutinization	91.2±0.4	93.0±0.9	74.5±0.8	83.2±3.0	69.5±1.6	71.5±1.0	73.7±2.1	81.0±1.9	82.6±2.5	74.4±1.7	82.2±0.3	84.0±3.3	80.1±1.6
<i>Gifts</i> w/o Consolidation	90.5±0.5	91.1±2.0	74.0±0.9	87.9±2.5	67.0±2.4	71.0±1.6	72.2±2.0	81.8±1.8	83.3±1.4	75.0±0.5	82.3±1.2	85.4±1.3	80.1±1.5
<i>Gifts</i> (full)	92.3±0.8	100±0	78.8±1.1	97.8±1.3	76.6±0.6	74.7±1.3	78.7±1.8	89.2±2.0	87.5±1.4	82.9±1.5	86.7±0.8	95.7±2.4	86.7±1.3

Table 4: Influence analysis of *Gifts* framework in information loss of audio-to-text conversion and how much private information the acoustic features expose. The experiments are conducted on AP²-Com.

Model/Attributes	AGE	GEN	ACC	HEA	HAB	PER	SOP	SOS	INC	OCC	EDU	MAR	Avg
<i>Gifts</i> in two Gemini1.5-Pro	91.3	97.9	74.5	95.0	75.2	73.5	76.2	85.5	85.0	81.0	84.2	91.9	84.3
<i>Gifts</i> in two Gemini1.5-Pro, both with audio inputs	91.5	98.3	74.6	95.0	75.7	74.2	77.0	85.2	83.7	81.4	84.5	92.0	84.4
<i>Gifts</i> in two Gemini1.5-Pro, both with audio inputs and acoustic-focus prompts	90.4	100	75.2	94.3	62.4	63.3	62.2	71.5	69.4	72.0	68.9	73.5	75.3

that takes text-only inputs, we let Gemini1.5-Pro additionally take audio inputs throughout all phases of *Gifts*. Experimental results shown in Table 4 reveal minimal performance differences between scenarios with and without audio input. In some cases, additional audio input even leads to performance degradation due to increased processing burden. *These findings suggest that the information loss during audio-to-text conversion has a limited effect on the overall performance of Gifts.* We further investigate the essential role of acoustic features in enabling such inferences of *Gifts*. Specifically, we incorporate a prompt: Fo

cus only on acoustic properties into the *Gifts* framework consisting of two Gemini1.5-Pro models. It is important to note that disentangling acoustic features from speech content in the input space is inherently infeasible due to their complex entanglement [46]. Therefore, although we included the acoustic-focus prompt, *Gifts* naturally leveraged other cues during inference. Additionally, directly extracting acoustic features (e.g., Mel spectrograms [47], waveform signals [48]) and feeding them into MLLMs is impractical, as MLLMs have limited capacity to process raw numerical representations [2]. As shown in Table 4, the results indicate that the inference accuracy for

Table 5: Performance comparison between multimodal large language models and real humans in inference accuracy and time spent in total during private attribute profiling.

Model/Attributes	AGE	GEN	ACC	HEA	HAB	PER	SOP	SOS	INC	OCC	EDU	MAR	Avg	Time Spent
Qwen-Audio-Chat + Claude3.5-Sonnet	71.4	66.7	33.3	66.7	66.7	66.7	75.0	75.0	66.7	75.0	78.7	66.7	67.4	6.3 minutes
SALAMONN-13B + Claude3.5-Sonnet	95.2	100.	41.7	100.	66.7	75.0	75.0	86.7	66.7	83.3	89.7	66.7	78.9	7.8 minutes
Qwen2-Audio-Instruct + Claude3.5-Sonnet	85.7	100.	66.7	100.	75.0	75.0	83.3	86.7	80.0	83.3	92.2	66.7	82.9	6.5 minutes
Gemini1.5-Pro + Claude3.5-Sonnet	95.2	100.	66.7	100.	75.0	75.0	75.0	86.7	80.0	91.7	91.7	100.	86.4	6.3 minutes
Human	88.2	100.	30.7	75.3	41.7	43.8	44.7	88.0	85.3	65.3	85.9	58.3	67.3	34.7 minutes
Gifts (ours)	100.0	100.0	66.7	100.0	75.0	75.0	91.7	91.7	93.3	91.7	92.2	100.0	89.8	8.2 minutes

Table 6: Real-world sensitivity comparison between Gifts and targets’ acquaintances in profiling targets.

Methods	AGE	GEN	ACC	HEA	HAB	PER
Targets’ Acquaintances	100.	100.	100.	94.0	75.0	85.0
Gifts	100.	100.	100.	100.	85.0	75.0
	SOP	SOS	INC	OCC	EDU	MAR
Targets’ Acquaintances	85.0	92.0	92.0	90.0	95.0	95.0
Gifts	85.0	96.0	92.0	85.0	90.0	95.0

Reasoning-driven attributes decreases but remains high, and the accuracy for Acoustic-driven attributes remains largely unchanged. *These findings confirm that acoustic features can expose substantial private information, enabling highly accurate inference attacks.*

7.3. Human Evaluation

To further evaluate the effectiveness and efficiency of employing MLLMs to profile sensitive attributes from audio, we conducted a human evaluation in compliance with IRB regulations and privacy-preserving policies (please refer to Ethics Considerations 11 for more information). For this study, we recruited 50 adult participants with advanced education and high proficiency in English. Most participants were familiar with and accustomed to using various search engines and LLM services. In this study, we used the data of three individuals randomly selected from AP²-Com rather than AP²-TV due to copyright considerations. Participants were instructed to freely listen to audio clips associated with these individuals and infer sensitive attributes. They were allowed to use search engines or LLMs to retrieve relevant information, enabling more accurate inferences. More details can be found in Appendix B.2.

Comparison between Human and MLLMs. According to the experiment results shown in Table 5, we can observe that all MLLMs outperform real humans in terms of average inference accuracy. Notably, our Gifts framework achieves the highest accuracy for all attributes, with a substantial gap of 22.5% (absolute error) in average accuracy compared to real humans. A closer look shows that MLLMs outperform humans more on Reasoning-driven attributes (absolute +21.4%) than on Acoustic-driven attributes (absolute +6.2%). This gap likely stems from MLLMs’ broader world knowledge compared to the average human. Supporting this, participants with higher education perform better

on Reasoning-driven attributes (Doctorate: 59.1%, Master: 55.4%, Bachelor or below: 55.0%), while showing similar performance on Acoustic-driven attributes (Doctorate: 77.6%, Master: 77.8%, Bachelor or below: 78.7%). See details on the homepage [5].

In addition, we compare the total time spent by real humans and MLLMs in profiling sensitive attributes for the three individuals. The results show that MLLMs demonstrate a significant advantage in time efficiency, with the longest duration being 8.2 minutes, only one-fourth of the time used by real humans. Moreover, the reported time for MLLMs includes network latency when calling APIs, meaning the actual time cost is even smaller. *These findings highlight the significant risk of employing MLLMs to profile sensitive attributes from audio, as they not only surpass humans but also do so with considerably smaller time costs.* More analysis is on the homepage [5].

Real-world Sensitivity of Gifts’s Inference. Among the 50 participants, five volunteers consented to the recording of a small number of their audio clips and agreed to share their profiles with us. Under their supervision, we applied Gifts to infer their sensitive attributes. Simultaneously, each volunteer invited close acquaintances to predict their profiles independently. We evaluated the inference results from both Gifts and the volunteers’ acquaintances. As shown in Table 6, *Gifts achieves comparable profiling accuracy to that of individuals who have close contact with the volunteers, despite relying only on a few audio samples.*

8. Potential Defensive Methods

8.1. Model Level

Defense Scenario. To defend against audio attribute profiling, we argue that MLLM providers, such as OpenAI, Google, and Meta, should take primary responsibility as defenders. These providers can enhance the safety of their models through a range of techniques, including the integration of safety-oriented system prompts and tuning-based alignment methods. By implementing such measures, MLLMs can be made more resistant to privacy leakage from audio data, thereby increasing their trustworthiness and promoting broader adoption.

Experiment Implementation. However, implementing such a model-level defense remains highly challenging because it is not feasible to perform fine-tuning-based safety alignment

Table 7: Defensive effect of enhanced In-context Unlearning against private attribute profiling from audio.

Datset	Models/Attributes	AGE	GEN	ACC	HEA	HAB	PER	SOP	SOS	INC	OCC	EDU	MAR	Avg
AP ² -Com	Qwen2-Audio-Instruct + Llama3-Instruct-8B	48.2	58.7	25.5	35.7	46.6	42.3	54.4	52.1	46.8	29.8	49.0	33.9	43.6
	Gemini1.5-Pro + Claude3.5-Sonnet	62.5	59.9	35.5	38.9	50.6	45.5	58.8	58.9	55.7	42.0	45.4	37.2	49.2
	Gifts (ours)	64.7	60.4	37.8	40.9	47.7	46.0	57.4	60.3	55.9	44.1	46.6	42.5	50.4
AP ² -TV	Qwen2-Audio-Instruct + Llama3-Instruct-8B	52.7	62.9	29.7	40.5	48.6	41.7	47.8	54.0	48.2	39.7	43.4	44.6	46.2
	Gemini1.5-Pro + Claude3.5-Sonnet	60.9	64.8	36.6	42.0	45.1	43.7	45.6	57.5	53.3	43.0	45.5	40.9	48.2
	Gifts (ours)	62.3	63.5	38.4	45.8	46.6	44.0	47.5	56.2	56.8	45.6	41.8	46.4	49.6

Table 8: Defensive effect of enhanced Anti-Eavesdropping Jamming against private attribute profiling from audio.

Datset	Models/Attributes	AGE	GEN	ACC	HEA	HAB	PER	SOP	SOS	INC	OCC	EDU	MAR	Avg
AP ² -Com	Qwen2-Audio-Instruct + Llama3-Instruct-8B	45.0	53.2	24.9	31.1	44.3	26.9	41.4	53.5	44.8	27.4	43.2	39.7	39.6
	Gemini1.5-Pro + Claude3.5-Sonnet	49.0	57.5	30.9	33.1	40.9	31.7	40.9	55.4	46.3	29.6	47.2	40.4	41.9
	Gifts (ours)	52.2	58.0	31.7	32.7	42.0	33.9	43.2	54.5	44.5	30.8	46.6	42.3	42.7
AP ² -TV	Qwen2-Audio-Instruct + Llama3-Instruct-8B	46.5	51.4	26.7	33.0	26.5	26.4	26.9	39.8	40.4	24.4	29.0	30.5	33.5
	Gemini1.5-Pro + Claude3.5-Sonnet	47.8	53.6	29.5	31.7	28.5	28.7	30.2	41.7	43.3	26.2	30.8	34.2	35.5
	Gifts (ours)	50.4	55.0	31.7	29.5	30.3	29.4	27.3	40.9	41.8	27.5	31.1	33.3	35.7

on current MLLMs, especially on those that are closed-source. In addition to including privacy-protection system prompts, we proposed a lightweight and effective method based on In-Context Unlearning [19] (ICU), which involves adding incorrect inputs into the system prompt as additional context to guide the model’s output. To implement this method, we first randomly split a sub-dataset that consists of 20 individuals from AP²-Com. For each attribute of every individual, we allocate the mismatched values and optimize the allocation to enable the strongest ICU performance on Qwen2-Audio and Llama3-Instruct. After determining the mismatched attribute value allocation, we further add adversarial perturbations to the audio data to enhance ICU performance on Qwen2-Audio. We then employ GPT-4o to paraphrase the audio event descriptions and spoken word transcriptions, augmenting the data for stronger ICU performance on Llama3-Instruct. After these steps, we test these MLLMs in profiling sensitive attributes from the rest individuals of AP²-Com and AP²-TV. According to experiment results shown in Table 7, we can observe that all MLLMs perform much worse than Table 2 and Figures 3 and 4. In particular, *even the strongest Gifts lags behind the real humans in average inference accuracy (Table 5) when attached with the enhanced ICU*. In the future, we will explore the fine-tuning-based safety alignment of MLLMs to defend against privacy inference from audio.

8.2. Data Level

Defense Scenario. The second defense method needs people to act as their own defenders against attribute profiling. In such scenarios, various types of noise may be integrated into the audio to protect privacy, including adversarial perturbations, jamming signals, and voice transformations. Implementing these defenses may require dedicated devices or software, which people can carry with them to provide real-time, portable protection anytime and anywhere, not just during private or sensitive situations.

Experiment Implementation. To validate the feasibility of

data-level defense, we adopt a Phoneme-based Noise [20] (PN). Unlike other types of noise that adversaries can easily detect and remove, PN is difficult to isolate and denoise, making it challenging for both humans and MLLMs to interpret audio captured by digital microphones. Specifically, we first use Whisper [49] to transcribe the textual information from the audio. Next, we extract the vowels and consonants from the transcriptions, and finally, we generate PN and implant it into the audio data. However, as described in Section 5, our AP² dataset contains rich background audio data for which PN is inapplicable. To address this limitation and further enhance the defensive effect, we add adversarial perturbations (further decreasing Qwen2-Audio’s inference accuracy) to the audio data that has already included PN. As shown in Table 8, we can observe that all MLLM approaches cannot accurately infer sensitive attributes from protected AP². Compared with model-level defense (Table 7), *the data-level method offers more effective protection. We will integrate this data-level protection with the model-level defense in the future.*

9. Discussion

9.1. Robustness of Gifts

Audio Input Noise. AP² inherently contains various real-world background noise and spans diverse scenarios and domains. To further study the robustness of Gifts, we conduct a dedicated analysis of the influence posed by the audio input noise on the inference performance. Specifically, we consider two types of noise: (1) white noise at three intensity levels (2.5, 5.0, and 7.5) with a fixed peak-to-noise ratio of 10; and (2) adversarial noise that is optimized to trigger a lower inference accuracy from the open-sourced Qwen2-Audio and refusal responses from Llama3-Instruct, respectively. As shown in Table 9, *only high-intensity white noise, which rarely occurs in real-world scenarios, causes a noticeable decline in inference accuracy. For the remaining white noise levels and the adversarial noise, Gifts remains robust and maintains nearly consistent performance.*

Table 9: Robustness analysis of `Gifts` framework in facing 3 levels of white noise, adversarial noise of triggering ALM inference accuracy decrease and LLM refusal response, prompt variations, and LLM hallucination alleviation with retrieval-augmented generation (RAG). The experiments are conducted on AP²-Com.

Robustness Factors/Attributes	AGE	GEN	ACC	HEA	HAB	PER	SOP	SOS	INC	OCC	EDU	MAR	Avg
White Noise with Scale 2.5	92.0	100.	77.4	97.4	76.4	74.7	78.7	89.2	87.0	82.9	86.5	95.7	86.5
White Noise with Scale 5.0	91.2	100.	76.6	95.7	75.2	73.9	77.5	88.1	86.0	81.2	86.0	93.9	85.4
White Noise with Scale 7.5	90.0	97.9	72.5	92.0	73.5	72.8	76.0	83.9	83.2	80.6	83.4	90.2	83.0
Adversarial Noise for ALM Inference	90.0	98.3	76.0	96.7	76.8	73.5	76.8	87.5	86.0	80.4	84.6	93.5	85.0
Adversarial Noise for LLM Refusal	92.0	98.3	78.3	95.3	76.8	75.0	78.2	87.5	89.2	83.0	85.5	96.0	86.3
3 Paraphrased Prompts	92.9 \pm 0.4	100. \pm 0	78.9 \pm 0.1	97.4 \pm 0.2	76.5 \pm 0.1	75.2 \pm 0.2	78.3 \pm 0.2	88.9 \pm 0.2	87.6 \pm 0.1	82.6 \pm 0.4	86.6 \pm 0.4	95.9 \pm 0.1	86.7 \pm 0.2
Enabling the LLM with RAG	92.5	100.	78.8	98.0	76.4	75.0	78.2	89.5	87.5	83.6	86.5	95.5	86.8

We also investigate how `Gifts` performs (remains fairly effective) when a target attempts to imitate another person’s voice and adversaries adaptively enhance `Gifts` (with a little improvement) driven by open-sourced MLLMs, shown in Appendix C.4 and C.5, respectively.

Prompt Variation. In addition to input noise, adversaries may use different prompts to run `Gifts`. To study the impact of prompt variations, we employed GPT-4o to paraphrase prompts in `Gifts` three times and used the paraphrased prompts to conduct attribute profiling. Experiment results shown in Table 9 demonstrate *the limited influence of prompt variations on `Gifts`’s inference performance (with small average accuracy change and variance)*.

LLM Hallucination. Although LLMs are advanced enough, they may still produce hallucinations when being used by `Gifts`. To address LLM hallucinations, `Gifts` has already taken some measures: 1) The ALM dominates inference in `Gifts`; 2) The LLM generates forensic questions strictly based on audio event descriptions; 3) The ALM can respond with “Uncertain” to invalidate questions with potential hallucinations; 4) When the LLM incorrectly rejects ALM’s inferences during scrutinization, the ALM typically maintains its original answer in the second inference. To further analyze the impact of potential LLM hallucinations, we enabled Claude3.5-Sonnet with retrieval-augmented generation (RAG), which is widely used for hallucination alleviation. According to Table 9, *enabling RAG does not bring about a lot of inference accuracy gain, which indirectly validates that `Gifts` is robust to LLM potential hallucinations.*

9.2. Bias in Multimodal Large Language Models

In our major experiments, we found that MLLMs tend to consistently infer certain attributes as being either too high or concentrated around a few answers. For instance, Claude3.5-Sonnet and GPT-4o often lean toward inferring individuals as belonging to higher social stratum and income levels. An analysis of the incorrect inferences for these two attributes revealed that over 90% of the errors were biased toward higher levels. Similarly, ALMs show evident biases in attributes like accent. For example, Gemini 1.5-Pro consistently infers accents as American, Canadian, Indian, British, Irish, or Australian, with low accuracy for other accents. These phenomena indicate that current MLLMs exhibit significant bias and unfairness, which are reflected as

group-level accuracy disparities provided in Appendix C.2. To make MLLMs fairer, we will achieve fairness-regularized model fine-tuning [50], [51] in future work.

9.3. Future Work of Defense

Although our experiments show that noise-adding-based data-level defense outperforms privacy-oriented prompt-based model-level defense, we argue that model-level approaches warrant further exploration due to their generality and independence from specialized devices or software. In this direction, improving the safety alignment of ALMs remains a challenging task. First, effective safety alignment depends on large-scale, high-quality datasets. Limited or low-quality data may result in incomplete alignment [52], [53]. However, collecting high-quality audio data, especially privacy-sensitive data, is inherently difficult. Second, safety alignment may compromise general model capabilities, as core ALM tasks such as speaker recognition, mapping, and event detection rely on acoustic features that often encode private information. Despite these challenges, safety alignment for ALMs is feasible. Notably, privacy leakage often originates from text prompts; thus, filtering mechanisms may be adapted to detect such queries. Besides, techniques like feature disentanglement can be incorporated to reduce reliance on privacy-related features.

9.4. Limitations

To ensure a more comprehensive study of audio attribute profiling, we considered several attributes that are strongly influenced by subjective factors, such as social preference, personality, and habits. Although we took measures to enhance the objectivity of annotation, traces of subjective judgment inevitably remain in the dataset. In future work, under appropriate ethical policies, we will attempt to collect more real-world audio data and annotations to further eliminate such subjective factors and enrich AP². Besides, the current AP² is limited to English-speaking contexts and their associated cultural settings. It would be valuable to extend AP² to other languages and cultures, such as Chinese, Spanish, French, and German, and to evaluate the cross-lingual generalization capability of `Gifts`. For AP²-Com, translation into other languages is feasible, as modern text-to-speech techniques generally support multiple languages.

However, acoustic features such as tone, pause, and pitch should be carefully adjusted to align with the linguistic and cultural characteristics of the target language. For $\text{AP}^2\text{-TV}$, TV dramas from non-English languages and cultures would need to be collected and annotated following the same procedures established for $\text{AP}^2\text{-TV}$.

In addition, the current `Gifts` framework consumes a considerable number of tokens, primarily due to redundant components such as spoken-word transcriptions, event descriptions, and forensic questions. In future work, we plan to build an agentic `Gifts`, where agents driven by ALMs and LLMs autonomously decide what necessary information should be extracted to support attribute profiling.

10. Related Works

Machine Learning Inference Attack. Inference attacks have long posed significant threats in ML security [54], traditionally focusing on training data. However, their scope extends further: for example, inferring sensitive attributes like gender from social media comments [55], or personal traits such as age and location from online text using LLMs [56]. Vision-language models have also been shown to infer private information from images [3]. Yet, no prior work explores inference attacks on general audio data using large models. This gap is alarming, especially given the broader and more passive collection of audio compared to text or images [6], [7], highlighting the urgency of studying such attacks in the context of today’s large models.

Privacy Leakage from Audio Data. Audio data plays a vital role in many ML applications but poses serious privacy risks [11]. Attacks include membership inference to identify speakers [26] and the prediction of sensitive emotional or mental states [10]. In healthcare, acoustic signals aid diagnosis [57] but may reveal private health information. Similarly, audio event classification [27] can uncover social behaviors and demographics. While traditional ML-based privacy attacks are limited by their reliance on large, labeled datasets, the rise of large models pretrained on vast unlabeled audio data significantly increases the risk of privacy leakage, as shown in this work.

11. Conclusion

We present the first study on audio privacy leakage caused by Multi-modal Large Language Models (MLLMs). To enable this investigation, we construct AP^2 , a benchmark dataset composed of public audio and TV drama clips annotated with sensitive attributes. We then propose `Gifts`, an MLLM-based framework that combines audio language models and large language models to infer sensitive attributes from audio. Extensive experiments show that `Gifts` poses a significant privacy threat, outperforming all other MLLM-based baseline approaches and real humans. To mitigate this risk, we propose model- and data-level defenses, which effectively reduce the inference success of `Gifts` and other MLLMs.

Acknowledgments

We thank the shepherd and anonymous reviewers for their valuable comments. This research is supported by the National Research Foundation, Singapore, and the Cyber Security Agency of Singapore under the National Cybersecurity R&D Programme and the CyberSG R&D Programme Office (Award CRPO-GC3-NTU-001). Any opinions, findings, conclusions, or recommendations expressed in these materials are those of the author(s) and do not reflect the views of the National Research Foundation, Singapore, the Cyber Security Agency of Singapore, or the CyberSG R&D Programme Office.

Ethics Considerations

Our research aims to study how MLLMs infer private information from audio data. We propose a framework to explore the upper limits of such privacy leakage and comprehensive defensive approaches. We have reported relevant institutions and companies, such as Google and Anthropic, about our results and findings. The real human study is approved by the Institutional Review Board (IRB). Fifty adult participants were voluntarily recruited through our affiliated institution’s research participant mailing list. Written informed consent was obtained from all participants before participation. Participation was fully voluntary and uncompensated, and no personally identifying information was stored. Responses were coded numerically and encrypted before storage.

All data included in $\text{AP}^2\text{-Com}$ were derived from publicly available datasets released under open-access licenses permitting use for non-commercial research. No new human-subject data were collected. For clinical datasets such as DAIC-WOZ and Movement Disorders Voice, only publicly released subsets were used, all of which were anonymized by their original custodians and allowed for non-commercial research. Within $\text{AP}^2\text{-Com}$, we further ensured that no combination of audio segments corresponds to a real individual: each composite profile was constructed through stochastic assignment and audio synthesis to prevent potential identity reconstruction. For $\text{AP}^2\text{-TV}$, only character-level audio, not real identities of performers, from legal streaming platforms was used. For the data release, we have fully released $\text{AP}^2\text{-Com}$, including dataset descriptions, audio clips, attribute annotations, and metadata such as speaker order and recording time. For $\text{AP}^2\text{-TV}$, due to copyright constraints, we only released the “indexes” to reproduce our experimental results rather than the dataset. Specifically, we released detailed timestamps for audio clips from each TV drama, along with attribute annotations and metadata on speaker order. We also include comprehensive instructions on where and how to legally access the TV dramas, as well as the usage protocol for $\text{AP}^2\text{-TV}$. All of the above materials are freely accessible at https://drive.google.com/file/d/12fOMpmm29VcnPCTwr6_Pm71V9gwX0hWj/view?usp=sharing.

LLM Usage Considerations

LLMs were only used for editorial purposes in this manuscript, such as grammar checking, and all outputs were inspected to ensure originality. No LLM-generated ideas were integrated into this manuscript, and all data used by LLMs/MLLMs complied with ethical requirements.

References

- [1] S. Minaee, T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, X. Amatriain, and J. Gao, “Large language models: A survey,” *arXiv preprint arXiv:2402.06196*, 2024.
- [2] D. Zhang, Y. Yu, J. Dong, C. Li, D. Su, C. Chu, and D. Yu, “Mm-llms: Recent advances in multimodal large language models,” *Findings of ACL*, 2024.
- [3] B. Tömekçe, M. Vero, R. Staab, and M. Vechev, “Private attribute inference from images with vision-language models,” *NeurIPS*, 2024.
- [4] F. Liu, Y. Zhang, X. Huang, Y. Peng, X. Li, L. Wang, Y. Shen, R. Duan, S. Qin, X. Jia *et al.*, “The eye of sherlock holmes: Uncovering user private attribute profiling via vision-language model agentic framework,” *ACM Conference on Multimedia*, 2025.
- [5] A. Authors, “Homepage of the person behind the sound: Demystifying audio private attribute profiling via multimodal large language models,” <https://sites.google.com/view/audioprofiling/>.
- [6] J. Li, C. Chen, M. R. Azghadi, H. Ghodosi, L. Pan, and J. Zhang, “Security and privacy problems in voice assistant applications: A survey,” *Computers & Security*, p. 103448, 2023.
- [7] M. Windl and S. Mayer, “The skewed privacy concerns of bystanders in smart environments,” *Proceedings of the ACM on Human-Computer Interaction*, pp. 1–21, 2022.
- [8] S. Landau, *Surveillance or security?: The risks posed by new wiretapping technologies*. Mit Press, 2011.
- [9] M. Caprolu, S. Sciancalepore, and R. Di Pietro, “Short-range audio channels security: Survey of mechanisms, applications, and research challenges,” *IEEE Communications Surveys & Tutorials*, pp. 311–340, 2020.
- [10] J. L. Kröger, L. Gellrich, S. Pape, S. R. Brause, and S. Ullrich, “Personal information inference from voice recordings: User awareness and privacy concerns,” *Proceedings on Privacy Enhancing Technologies*, 2022.
- [11] J. L. Kröger, O. H.-M. Lutz, and P. Raschke, “Privacy implications of voice and speech analysis—information disclosure by inference,” in *IFIP International Summer School on Privacy and Identity Management*. Springer, 2019, pp. 242–258.
- [12] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, “Common voice: A massively-multilingual speech corpus,” in *Proceedings of Language Resources and Evaluation Conference*, 2020, pp. 4218–4222.
- [13] P. Voigt and A. Von dem Bussche, “The eu general data protection regulation (gdpr),” *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, pp. 10–5555, 2017.
- [14] S. Sakshi, U. Tyagi, S. Kumar, A. Seth, R. Selvakumar, O. Nieto, R. Duraiswami, S. Ghosh, and D. Manocha, “Mmau: A massive multi-task audio understanding and reasoning benchmark,” *ICLR*, 2025.
- [15] Y. Xiao and R. K. Das, “WildDESED: An llm-powered dataset for wild domestic environment sound event detection system,” in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2024 Workshop*, 2024, pp. 196–200.
- [16] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, “Audio set: An ontology and human-labeled dataset for audio events,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017.
- [17] Pixabay, “Pixabay,” <https://pixabay.com/>, 2025.
- [18] Freesound, “Freesound,” <https://freesound.org/>, 2025.
- [19] M. Pawelczyk, S. Neel, and H. Lakkaraju, “In-context unlearning: Language models as few-shot unlearners,” in *International Conference on Machine Learning*, 2024.
- [20] P. Huang, Y. Wei, P. Cheng, Z. Ba, L. Lu, F. Lin, F. Zhang, and K. Ren, “Infomasker: Preventing eavesdropping using phoneme-based noise,” in *NDSS*, 2023.
- [21] B. Jayaraman and D. Evans, “Are attribute inference attacks just imputation?” in *ACM SIGSAC Conference on Computer and Communications Security*, 2022, pp. 1569–1582.
- [22] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha, “Privacy risk in machine learning: Analyzing the connection to overfitting,” in *IEEE computer security foundations symposium*. IEEE, 2018, pp. 268–282.
- [23] Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, Z. Guo, Y. Leng, Y. Lv, J. He, J. Lin *et al.*, “Qwen2-audio technical report,” *arXiv preprint arXiv:2407.10759*, 2024.
- [24] C. Tang, W. Yu, G. Sun, X. Chen, T. Tan, W. Li, L. Lu, M. Zejun, and C. Zhang, “Salmonn: Towards generic hearing abilities for large language models,” in *International Conference on Learning Representations*, 2024.
- [25] G. Team, P. Georgiev, V. I. Lei, R. Burnell, L. Bai, A. Gulati, G. Tanzer, D. Vincent, Z. Pan, S. Wang *et al.*, “Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context,” *arXiv preprint arXiv:2403.05530*, 2024.
- [26] J. S. Chung, A. Nagrani, and A. Zisserman, “Voxceleb2: Deep speaker recognition,” in *Proceedings of INTERSPEECH*, vol. 2018, 2018, pp. 1086–1090.
- [27] Z. Kons, O. Toledo-Ronen, and M. Carmel, “Audio event classification using deep neural networks,” in *Interspeech*, 2013, pp. 1482–1486.
- [28] J.-P. Briot and F. Pachet, “Deep learning for music generation: challenges and directions,” *Neural Computing and Applications*, pp. 981–993, 2020.
- [29] Y. Chu, J. Xu, X. Zhou, Q. Yang, S. Zhang, Z. Yan, C. Zhou, and J. Zhou, “Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models,” *arXiv preprint arXiv:2311.07919*, 2023.
- [30] K. D. Kinzler, E. Dupoux, and E. S. Spelke, “The native language of social cognition,” *Proceedings of the National Academy of Sciences*, vol. 104, no. 30, pp. 12 577–12 580, 2007.
- [31] S. Patnaik, E. Brunskill, and W. Thies, “Evaluating the accuracy of data collection on mobile phones: A study of forms, sms, and voice,” in *International Conference on Information and Communication Technologies and Development*. IEEE, 2009, pp. 74–84.
- [32] M. Guri, Y. Solewicz, A. Daidakulov, and Y. Elovici, “{SPEAKE (a) R}: Turn speakers to microphones for fun and profit,” in *USENIX Workshop on Offensive Technologies*, 2017.
- [33] G. Snow and E. Guardiani, “Movement disorders and voice,” *Otolaryngologic Clinics of North America*, pp. 759–767, 2019.
- [34] F. Rudzicz, A. K. Namasivayam, and T. Wolff, “The torgo database of acoustic and articulatory speech from speakers with dysarthria,” *Language resources and evaluation*, 2012.
- [35] J. Gratch, R. Artstein, G. M. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella *et al.*, “The distress analysis interview corpus of human and computer interviews,” in *LREC*. Reykjavik, 2014, pp. 3123–3128.
- [36] R. M. Tobin, W. G. Graziano, E. J. Vanman, and L. G. Tassinary, “Personality, emotional experience, and efforts to control emotions,” *Journal of personality and social psychology*, 2000.
- [37] S. R. Livingstone and F. A. Russo, “The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english,” *PLoS one*, p. e0196391, 2018.

- [38] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, “Gpt-4 technical report,” *arXiv preprint arXiv:2303.08774*, 2023.
- [39] Soundbible, “Soundbible,” <https://soundbible.com/>.
- [40] S. Liu, “Zero-shot voice conversion with diffusion transformers,” *arXiv preprint arXiv:2411.09943*, 2024.
- [41] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing *et al.*, “Judging llm-as-a-judge with mt-bench and chatbot arena,” *Advances in Neural Information Processing Systems*, pp. 46 595–46 623, 2023.
- [42] Anthropic, “Claude series,” 2024. [Online]. Available: <https://www.anthropic.com/news/claude-3-5-sonnet>
- [43] S. Deshmukh, S. Han, H. Bukhari, B. Elizalde, H. Gamper, R. Singh, and B. Raj, “Audio entailment: Assessing deductive reasoning for audio understanding,” *AAAI*, 2025.
- [44] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei *et al.*, “Qwen2. 5 technical report,” *arXiv preprint arXiv:2412.15115*, 2024.
- [45] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan *et al.*, “The llama 3 herd of models,” *arXiv preprint arXiv:2407.21783*, 2024.
- [46] T. Liu, K. A. Lee, Q. Wang, and H. Li, “Disentangling voice and content with self-supervision for speaker recognition,” *Advances in Neural Information Processing Systems*, pp. 50 221–50 236, 2023.
- [47] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, “Natural tts synthesis by conditioning wavenet on mel spectrogram predictions,” in *IEEE international conference on acoustics, speech and signal processing*. IEEE, 2018, pp. 4779–4783.
- [48] J. Lee, J. Park, T. Kim, and J. Nam, “Raw waveform-based audio classification using sample-level cnn architectures,” in *Machine Learning for Audio Signal Processing Workshop, Neural Information Processing Systems*, 2017.
- [49] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.
- [50] S. Yan, D. Huang, and M. Soleymani, “Mitigating biases in multi-modal personality assessment,” *International Conference on Multi-modal Interaction*, p. 361–369, 2020.
- [51] P. Delobelle and B. Berendt, “Fairdistillation: mitigating stereotyping in language models,” *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, p. 638–654, 2022.
- [52] X. Liu, J. Liang, M. Ye, and Z. Xi, “Robustifying safety-aligned large language models through clean data curation,” *arXiv preprint arXiv:2405.19358*, 2024.
- [53] Y. Zhang, L. Chen, G. Zheng, Y. Gao, R. Zheng, J. Fu, Z. Yin, S. Jin, Y. Qiao, X. Huang, F. Zhao, T. Gui, and J. Shao, “SPAVL: A comprehensive safety preference alignment dataset for vision language model,” *arXiv preprint arXiv:2406.12030*, 2024.
- [54] M. Rigaki and S. Garcia, “A survey of privacy attacks in machine learning,” *ACM Computing Surveys*, pp. 1–34, 2023.
- [55] B. Alipour, A. Imine, and M. Rusinowitch, “Gender inference for facebook picture owners,” in *Trust, Privacy and Security in Digital Business: International Conference*. Springer, 2019, pp. 145–160.
- [56] R. Staab, M. Vero, M. Balunovic, and M. Vechev, “Beyond memorization: Violating privacy via inference with large language models,” in *International Conference on Learning Representations*, 2024.
- [57] P. Kapetanidis, F. Kalioras, C. Tsakonas, P. Tzamalīs, G. Kontogiannis, T. Karamanidou, T. G. Stavropoulos, and S. Nikolettseas, “Respiratory diseases diagnosis using audio analysis and artificial intelligence: a systematic review,” *Sensors*, p. 1173, 2024.
- [58] F. Burkhardt, J. Wagner, H. Wierstorf, F. Eyben, and B. Schuller, “Speech-based age and gender prediction with transformers,” in *Speech Communication; ITG Conference*. VDE, 2023, pp. 46–50.
- [59] J. Zuluaga-Gomez, S. Ahmed, D. Visockas, and C. Subakan, “Commonaccent: Exploring large acoustic pretrained models for accent classification based on common voice,” in *Proc. Interspeech*, 2023, pp. 5291–5295.
- [60] Hatman, “Hatman/audio-emotion-detection,” <https://huggingface.co/Hatman/audio-emotion-detection>.

Appendix A.

Annotation rules of AP²

To ensure the rigorousness and validity of AP² annotations, we engaged three experts in the dataset construction process. The three experts were selected based on their background in speech processing and linguistic annotation, each holding at least a master’s degree in relevant fields (speech science, linguistics, or natural language processing) and with prior experience in dataset curation or acoustic analysis.

A.1. AP²-Com

To retrieve accurate and pertinent data from public audio datasets and open-access platforms, we engaged the experts to follow the rules below.

- *Strategic Data Acquisition:* An iterative approach to searching and refining queries is crucial, along with the ability to identify relevant contextual clues within the audio that may indirectly indicate the target attributes.
- *Rigorous Selection and Preparation:* The selection of audio data must prioritize relevance to the commonsense and expert criteria of typical behaviors, dialogue, and activities for specific attribute values.
- *Ethical Considerations:* Experts must strictly adhere to the terms of service of the accessed platforms and remain mindful of potential privacy concerns.

After each expert’s retrieval and during cross-validation, all experts were required to give a score implying the validity and accuracy of each retrieved audio sample associated with each attribute, as shown in Table 10.

- Score 5: The audio is highly relevant, clearly representative of the target attribute value.
- Score 4: The audio is appropriate and clearly related to the target attribute value.
- Score 3: The audio is somewhat relevant but has notable weaknesses in clarity, representativeness, or justification.
- Score 2: The audio has weak relevance and lacks strategic retrieval or sufficient evidence.
- Score 1: The audio is irrelevant, misleading, or ethically problematic, with no evidence of valid retrieval.

A.2. AP²-TV

To ensure the accuracy and validity of the annotations for characters, the experts followed the rules below:

- *Base Annotations on Comprehensive Analysis*: Annotations must be derived from a thorough analysis of the entire TV series, including dialogue, actions, interactions, character development, and the provided supplementary materials (promotional materials, media coverage, online forum discussions).
- *Objectivity and Consistency*: Strive for objectivity in the annotations. Minimize personal biases and interpretations. Apply the annotation guidelines consistently across all characters and throughout the entire series.
- *Evidence-based Justification*: For each attribute annotation, be prepared to provide specific examples and justifications from the series and supplementary materials that support your choice. Note down key scenes, dialogues, or external information that influenced your decision.
- *Consider Character Arc and Development*: Recognize that character attributes can evolve throughout the series. Annotate the attribute values based on the character’s portrayal across the entire narrative arc. If significant changes occur, consider noting the evolution in your justification.

Similar to AP²-Com, each annotated attribute here was also assigned with a score, as shown in Table 10.

- Score 5: The annotation is fully accurate, well-justified with clear evidence from the series and materials, and acknowledges ambiguity where needed.
- Score 4: The annotation is accurate and mostly well-supported, with only minor gaps in justification.
- Score 3: The annotation shows partial correctness but has noticeable issues, such as limited evidence, weak inferences, or inconsistent reasoning.
- Score 2: The annotation lacks strong support, relies on weak or speculative inferences, and may show bias.
- Score 1: The annotation is clearly incorrect, unsupported by the series or materials, and ignores ambiguity.

Table 10: Retrieval and annotation confidence score of three experts on AP²-Com and AP²-TV, respectively.

Datasets	AGE	GEN	ACC	HEA	HAB	PER	SOP	SOS	INC	OCC	EDU	MAR
AP ² -Com	5.0	5.0	4.7	4.3	4.3	4.7	4.7	4.7	4.7	4.7	4.7	4.3
AP ² -TV	4.7	5.0	4.7	4.7	5.0	5.0	4.7	5.0	5.0	5.0	4.7	4.7

Table 11: Target Attribute Options

Attribute	Options
Age	Younger than twenties, twenties, thirties, forties, fifties, sixties, older than sixties
Gender	Male, Female
Accent	American, British, England, Canadian, Australian, Irish, Scottish, New Zealand, South African, Indian, Asian
Health condition	Healthy, Slightly Physically Sick, Slightly Mentally Sick, Severely Physically Sick, Severely Mentally Sick
Physical disease	Parkinson, Alzheimer, Dysarthric
Mental disease	Depression, Anxiety, Post-Traumatic Stress Disorder
Social stratum	Lower Class, Working Class, Middle Class, Upper-Middle Class, Upper Class
Income	Low Income, Lower-Middle Income, Middle Income, Upper-Middle Income, High Income
Education level	Lower than High School, High School, Associate Degree, Bachelor’s Degree, Master’s Degree, Doctorate’s Degree
Marital status	Single, Married, Separated, Divorced, Widowed

Appendix B. Setups of Major Experiments

B.1. Evaluation Metrics

We use absolute error accuracy for qualitative attributes. As for quantitative attributes, we first define an ordered range section list for each attribute, shown in Table 11. Then we employ relative error accuracy to measure the difference between the inferred attribute value and the real. Fuzzy attributes are quite subjective, and to ensure the evaluation objectivity, we leverage the strongest LLM, Claude3.7-Sonnet, to conduct a fuzzy evaluation that produces a five-level similarity score from 1 to 0 with a stride of 0.25: “*Highly Similar*, *Similar*, *Moderately Similar*, *Slightly Similar*, and *Completely Different*” in meaning and range for PER, SOP, OCC, and HAB, and in pronunciation and vocabulary usage for ACC. Ultimately, we employ a hierarchical accuracy for HEA. Specifically, the inferred disease is compared with the ground truth in matching with “*Healthy*, *Slightly Sick*, and *Severely Sick*”, where the correct match earns 0.5. Then the inferred disease is evaluated by the match between “*Physically Sick*” and “*Mentally Sick*”, where the correct match earns another 0.25. The last 0.25 can be earned if the inferred disease is identical to the real one. As for EDU, we employ relative error accuracy to measure the difference between the inferred education level and the real one, and then use Claude3.7-Sonnet to evaluate the similarity between the inferred major and the ground truth. The evaluation of education level accounts for 0.7, while the major holds the weight of 0.3.

B.2. Human Study

The human study involved 50 participants with the following background information collected anonymously:

- **Gender**: Male, 40 people; Female, 10 people.
- **Age**: 18-20 years, 4 people; 21-30 years, 46 people.
- **English Proficiency**: Advanced, 25 people; Intermediate, 23 people; Beginner, 2 people.
- **Education Level**: Master’s Degree, 19 people; Doctorate’s Degree, 17 people; Bachelor’s Degree, 13 people; High School, 1 person.
- **Search Engine Proficiency**: Basic, 1 person; Intermediate, 16 people; Advanced, 25 people; Expert, 8 people.
- **LLM Proficiency**: Basic, 1 person; Intermediate, 18 people; Advanced, 26 people; Expert, 5 people.

For each level of proficiency, we provided detailed descriptions outlining the expected skills and capabilities.

Participants were then asked to listen to audio clips from three different individuals. For each individual, participants were instructed to create a profile based on their impressions formed from the audio. They were allowed to listen to each audio clip as many times as needed to form their assessments. Additionally, they were permitted to search for relevant information online or through LLM during the task; however, the use of LLMs to generate attribute inferences

Table 12: Generalization analysis of comparison between *Gifts* and simple consolidation (Baseline) across different model capability levels. Results indicate *Gifts*’s superiority stems from methodological innovation rather than model dependency, and *Gifts* can generalize well across models with different capabilities.

Model/Attributes	AGE	GEN	ACC	HEA	HAB	PER	SOP	SOS	INC	OCC	EDU	MAR	Avg	Δ
SALAMONN-13B + Claude3.5-Sonnet (Baseline)	77.6	65.8	61.4	81.0	71.2	67.7	70.3	81.0	75.7	58.5	80.6	53.8	70.4	-
SALAMONN-13B + Claude3.5-Sonnet (<i>Gifts</i>)	90.6	92.4	74.1	96.8	67.1	74.7	79.1	89.9	87.9	84.8	91.4	96.2	85.4	+15.0
SALAMONN-13B + Claude3.5-Haiku (<i>Gifts</i>)	88.5	89.2	68.7	93.3	69.2	74.1	78.5	89.9	87.1	83.5	89.4	93.0	83.7	+13.3
Qwen2-Audio-Instruct + Claude3.5-Sonnet (Baseline)	75.8	92.4	51.0	92.0	71.9	71.2	79.4	82.0	78.5	64.6	84.5	48.1	74.3	-
Qwen2-Audio-Instruct + Claude3.5-Sonnet (<i>Gifts</i>)	83.5	93.7	69.1	96.2	70.9	74.7	79.8	89.4	79.0	84.5	87.9	88.6	83.1	+8.8
Qwen2-Audio-Instruct + Claude3.5-Haiku (<i>Gifts</i>)	87.3	92.4	51.6	97.2	70.0	77.4	78.1	89.4	80.8	81.5	81.8	85.4	81.1	+6.8
Gemini1.5-Pro + Claude3.5-Sonnet (Baseline)	85.4	96.2	71.5	85.6	68.7	66.1	71.8	87.1	80.3	63.3	82.9	59.5	76.5	-
Gemini1.5-Pro + Claude3.5-Sonnet (<i>Gifts</i>)	92.0	100.	78.2	96.8	76.6	74.1	81.7	88.6	87.6	82.9	86.7	91.8	86.5	+10.0
Gemini1.5-Pro + Claude3.5-Haiku (<i>Gifts</i>)	90.8	96.2	78.2	94.6	74.2	73.9	77.0	89.1	88.1	82.0	85.5	93.0	85.2	+8.7

Table 13: Performance comparison between *Gifts* and simple consolidation (Baseline) with state-of-the-art ALM and LLM on AP²-TV.

Methods	AGE	GEN	ACC	HEA	HAB	PER
Gemini1.5-Pro + Claude3.5-Sonnet (Baseline)	85.6	86.8	90.5	75.6	55.4	55.0
Gemini2.5-Pro + Claude3.7-Sonnet (Baseline)	89.3	94.3	92.8	85.3	56.8	58.6
Gemini2.5-Pro + Claude3.7-Sonnet (<i>Gifts</i>)	93.3	98.2	95.5	87.9	60.6	61.5
	SOP	SOS	INC	OCC	EDU	MAR
Gemini1.5-Pro + Claude3.5-Sonnet (Baseline)	55.0	82.5	75.2	65.4	86.2	79.4
Gemini2.5-Pro + Claude3.7-Sonnet (Baseline)	59.0	87.5	78.8	72.4	87.0	87.8
Gemini2.5-Pro + Claude3.7-Sonnet (<i>Gifts</i>)	62.3	91.0	82.6	77.8	88.5	92.2

was prohibited. Finally, participants were asked to report the total amount of time they spent on the profiling task. More details about the human study’s results and analysis can be found on the homepage [5].

Table 14: Standard deviation disparities of attribute-wise and individual-wise inference accuracies across groups divided by different sensitive attributes. The experiments are conducted with *Gifts* on AP²-Com.

Methods	AGE	GEN	ACC	HEA	SOS	INC	EDU	MAR
Attribute-wise	2.94	0	13.2	5.68	10.8	8.72	8.23	3.77
Individual-wise	3.04	1.20	5.96	2.24	3.79	3.12	2.55	2.15

Appendix C. Additional Experiments and Analysis

C.1. *Gifts*’s Generalization for Different Models

To assess whether *Gifts* represents a fundamental methodological advantage independent of model capabilities, we compare *Gifts* with the simple consolidation strategy, i.e., the ALM infers every attribute and the LLM consolidates the ALM’s inference results. The results in Table 12 on AP²-Com show that the relative improvement of *Gifts* over baselines remains substantial (6.8-15.0% average improvement) across ALMs and LLMs with weaker capability levels. The comparable performance of *Gifts* built with different models also demonstrated the good generalization capability of *Gifts* across ALMs and LLMs.

We also enable the state-of-the-art ALM (Gemini2.5-Pro) and LLM (Claude3.7-Sonnet) to be driven by *Gifts*

and find that this can also bring about 3.5% accuracy improvement on AP²-TV, as shown in Table 13. Notably, the simple baseline can already provide relatively accurate inferences (79.1% versus *Gifts*’s 82.0%), but it is uncertain that the inference accuracy improvements over other models come from memorization or reasoning as the cut-off date of Gemini2.5-Pro and Claude3.7-Sonnet is after the premier date of the TV dramas of AP²-TV. To prevent the interference of this factor, we thus adopt Gemini1.5-Pro and Claude3.5-Sonnet to implement *Gifts*.

These results validate that *Gifts* is beyond specific model architectures or capability levels, rather than exploiting the exceptional capabilities of state-of-the-art models.

C.2. Group-level Statistical Analysis

To assess the fairness of *Gifts* framework in profiling sensitive attributes on different groups of people, we calculate the standard deviation disparities of attribute-wise and individual-wise inference accuracies across groups divided by AGE, GEN, ACC, HEA, SOS, INC, EDU, MAR. The experiment results on AP²-Com are shown in Table 14 from which we can observe that *Gifts* tends to be more unfair on attributes ACC, SOS, INC, and EDU. Precisely speaking, *Gifts* performs much better on individuals with specific accents, higher social stratum, income levels, and education levels. For other attributes, *Gifts* is relatively fair.

Table 15: Influence analysis of providing previously inferred attribute values when inferring the following attributes. The experiments are conducted with *Gifts* on AP²-Com.

Methods	AGE	GEN	ACC	HEA	HAB	PER
<i>Gifts</i> with inferred attributes	92.3	98.3	75.6	92.6	72.4	71.5
The original <i>Gifts</i>	92.3	100.	78.8	97.8	76.6	74.7
	SOP	SOS	INC	OCC	EDU	MAR
<i>Gifts</i> with inferred attributes	73.5	84.7	83.2	75.1	76.6	78.2
The original <i>Gifts</i>	78.7	89.2	87.5	82.9	86.7	95.7

C.3. Dependency of Inferred Attributes

Our *Gifts* framework independently infers each attribute, thereby preventing accumulation of inference inaccuracy. To investigate how large the inaccuracy is, we

Table 16: Robustness analysis of `Gifts` against voice imitation. The experiments are conducted on a TV drama from `AP2-TV` with voice transfer from `AP2-Com`.

Datasets	AGE	GEN	ACC	HEA	HAB	PER
<code>AP²-TV</code> with voice transfer	72.4	69.0	95.5	89.6	66.0	61.0
The original <code>AP²-TV</code>	92.2	100.	96.8	90.4	65.7	61.5
Datasets	SOP	SOS	INC	OCC	EDU	MAR
<code>AP²-TV</code> with voice transfer	63.5	85.2	83.1	75.1	81.9	85.6
The original <code>AP²-TV</code>	63.3	87.5	82.5	75.6	83.3	90.3

Table 17: Analysis of adaptive attack enhancement of `Gifts` on `AP2-Com`.

Datasets	AGE	GEN	ACC	HEA	HAB	PER
Original <code>Gifts</code> with Qwen2-Audio	83.5	93.7	69.1	96.2	70.9	74.7
Enhanced <code>Gifts</code> with Qwen2-Audio	87.7	94.2	71.6	96.2	72.5	75.0
Datasets	SOP	SOS	INC	OCC	EDU	MAR
Original <code>Gifts</code> with Qwen2-Audio	79.8	89.4	79.0	84.5	87.9	88.6
Enhanced <code>Gifts</code> with Qwen2-Audio	80.7	89.8	84.5	85.2	87.5	90.4

experiment with providing previously inferred attribute values when inferring each attribute in the order of Table 2. According to the results shown in Table 15, we can observe that the previously inferred attributes harm the following attribute inference, and such inaccuracy is accumulating.

Table 18: Performance comparison with regular ML models in inferring sensitive attributes.

Method/Attributes	Age	Gender	Accent	Character
Regular	87.4	100.	90.9	22.7
<code>Gifts</code>	92.5	100.	98.5	62.8

C.4. Robustness against Voice Imitation

In real-world scenarios, victims may attempt to imitate others’ voices to bypass `Gifts`’s inference, though it is difficult to sustain over extended periods. To evaluate the robustness of `Gifts` against voice imitation, we randomly select a TV drama from `AP2-TV` and perform voice transfer [40] on all characters using randomly selected individuals from `AP2-Com`. This setup reflects an extreme case of voice imitation, as accurately replicating another person’s voice is inherently challenging. The detailed experimental results are presented in Table 16, from which we observe that `Gifts` remains effective in most attribute predictions, with only minor drops in accuracy compared to the original voices. This robustness may be attributed to the fact that, although the timbre is altered, other acoustic features, such as speech patterns, subtle vocal cues, and ambient background sounds, remain largely unaffected. Nevertheless, there are certain attributes where `Gifts` shows reduced performance under voice imitation, likely due to significant changes in the characters’ pitch, timbre, and tone.

C.5. Adaptive Attack Enhancement

In the real world, adversaries may adopt open-sourced MLLMs to operate `Gifts` and adaptively enhance the in-

ference accuracy of private attribute profiling when they are aware that the collected audio data may involve adversarial perturbations. We adversarially optimize the Qwen2-Audio and the audio data of 20 individuals randomly selected from `AP2-Com` and test the `Gifts` framework driven by Qwen2-Audio and Claude3.5-Sonnet on the rest individuals. As shown in Table 17, such adaptive attack enhancement results in little accuracy improvement.

C.6. Comparison to Regular Approaches

To further evaluate `Gifts`, we compared it against state-of-the-art attribute inference methods based on regular ML models. Specifically, the Age-Gender model [58] is used to infer a speaker’s age and gender, the CommonAccent model [59] to predict English accents, and the Emotion-Detection model [60] to infer the speaker’s personality traits. Notably, these methods are built upon extensive training of regular ML models on well-labeled audio datasets. We used their model checkpoints on Huggingface and selected two dramas from `AP2-TV` as the testing data. To make our dataset compatible with these methods, we cut the original audio clips into segments that only contain the target speaker’s speech. Besides, we used GPT-4o to transform outputs of the Emotion-Detection model into textual personality descriptions. The results are presented in Table 18, demonstrating that `Gifts` still significantly outperforms these regular approaches.

C.7. Influence Analysis of Artificial Features in `AP2`

When constructing `AP2-Com`, we employed a suite of audio composition techniques, including voice transfer, audio splicing, and mixing. As a result, certain artificial artifacts were inevitably introduced into the audio data. However, `Gifts` performs profile-level inference by aggregating information across multiple clips for each individual. Since only 20–40% of the clips in `AP2-Com` are synthetic, artifacts present in a subset of clips are unlikely to substantially affect overall profile-level predictions. To further evaluate their impact, we partitioned `AP2-Com` into four groups based on the proportion of synthetic clips within each profile. Our results show that profile-level accuracy remains stable across these groups, with a variation of only 4.4%. Importantly, `AP2-TV` contains no synthetic data. On this dataset, `Gifts` continues to perform strongly and significantly outperforms baseline methods, indicating that it does not rely on artifacts introduced during the construction of `AP2-Com`.

For `AP2-TV`, we selected modern, socially grounded dramas featuring professional actors and narratives centered on everyday life. While such content still differs from real-world scenarios due to dramatization, we mitigate this by randomly sampling clips for each character across multiple episodes. This strategy reduces over-representation by preventing a small number of highly informative clips from dominating the dataset. We also manually reviewed the sampled clips to ensure realism and the absence of explicit sensitive attribute exposure.

Appendix D. Meta-Review

The following meta-review was prepared by the program committee for the 2026 IEEE Symposium on Security and Privacy (S&P) as part of the review process as detailed in the call for papers.

D.1. Summary

The paper examines the risk of audio-based private attribute profiling using multimodal large language models (MLLMs), which can infer sensitive traits from audio beyond transcripts (e.g., health, income, habits, personality). To measure this systematically, the paper introduces AP2, a benchmark with two subsets: AP2-Com (built from public datasets and curated audio composition) and AP2-TV (expert-annotated character clips from recent TV dramas), covering diverse attributes. The paper also proposes Gifts, a hybrid ALM+LLM framework where the LLM guides inference, asks forensic validation questions, evaluates the ALM’s responses, and aggregates predictions across clips.

D.2. Scientific Contributions

- Creates a New Tool to Enable Future Science.
- Provides a Valuable Step Forward in an Established Field.

D.3. Reasons for Acceptance

- 1) The paper identifies an emerging privacy threat of inferring sensitive personal attributes from audio using multimodal LLMs, which is highly relevant given the increasing deployment of audio-enabled AI systems.
- 2) The proposed Gifts pipeline combines audio-language models with LLM reasoning through a multi-stage process to produce stronger and more interpretable profiling results than simpler baselines.
- 3) The paper conducts a comprehensive empirical evaluation across multiple baselines (ALM-only, LLM-only, and combinations). The results consistently show that Gifts significantly outperforms these alternatives across many attribute categories and remains reasonably robust to noise and perturbations.

D.4. Noteworthy Concerns

- 1) AP2-Com constructs synthetic samples through voice transfer and mixing, and AP2-TV uses scripted television characters. These design choices may introduce artifacts or exaggerated behavioral signals that differ from real-world conversational audio.
- 2) The paper acknowledges disentanglement of acoustic and semantic leakage is hard. Current ablations don’t fully isolate what signal Gifts exploits.

Appendix E. Response to the Meta-Review

We thank the anonymous shepherd and reviewers for their careful reading and valuable feedback. The meta-review raises two main concerns: (1) the realism gap between AP² and real-world scenarios due to audio synthesis and TV dramatization, and (2) the insufficient ablation separating acoustic and semantic sources of privacy leakage caused by Gifts.

Regarding the realism gap, Gifts performs profile-level inference by aggregating evidence across multiple clips per individual. In AP²-Com, only 20-40% of clips per profile are synthetic; thus, artifacts present in a subset are unlikely to dominate predictions. To verify this, we partitioned profiles by the proportion of synthetic clips and observed that profile-level accuracy remains stable, varying by only 4.4%. Furthermore, AP²-TV contains no synthetic audio, yet Gifts still performs strongly and significantly outperforms all baselines. This suggests that Gifts’s performance is not driven by synthetic artifacts. For AP²-TV, we curate modern, socially grounded, and everyday-life dramas featuring professional actors. While dramatized, we also randomly sample clips across episodes for each character, reducing the over-representation of particularly informative segments. We also manually review samples to ensure realism and to avoid explicit exposure of sensitive attributes.

Regarding the ablation of acoustic vs. semantic leakage, we still want to highlight that disentangling these factors is inherently challenging [46], as acoustic delivery (e.g., tone, pitch, hesitation) is tightly coupled with linguistic content. As a practical proxy, text-only baselines exactly rely on semantic information; their performance thus reflects semantic privacy leakage. As shown in Table 2, Gifts substantially outperforms these text-only baselines, indicating that it leverages acoustic cues heavily beyond what is captured by text alone. Additionally, our ablation using an acoustic-focused prompt, while not perfectly isolating acoustic signals, leads to a notable accuracy drop as shown in Table 4, underscoring the importance of semantic and contextual information.