

000 001 BACKDOOR ATTACKS AGAINST SPEECH LANGUAGE 002 MODELS 003

004
005 **Anonymous authors**
006 Paper under double-blind review

007
008
009 **ABSTRACT**
010

011 Large Language Models (LLMs) and their multimodal extensions are becoming
012 increasingly popular. One common approach to enable multimodality is to cas-
013 ckle domain-specific encoders with an LLM, making the resulting model inherit
014 vulnerabilities from all of its components. In this work, we present the first sys-
015 tematic study of audio backdoor attacks against speech language models. We
016 demonstrate its effectiveness across four speech encoders and three datasets, cov-
017 ering four tasks: automatic speech recognition (ASR), speech emotion recogni-
018 tion, and gender and age prediction. The attack consistently achieves high suc-
019 ccess rates, ranging from 90.76% to 99.41%. To better understand how backdoors
020 propagate, we conduct a component-wise analysis to identify the most vulnera-
021 ble stages of the pipeline. Finally, we propose a fine-tuning-based defense that
022 mitigates the threat of poisoned pretrained encoders.

023 **1 INTRODUCTION**
024

025 Large language models (LLMs) are increasingly extended to multimodal settings, processing com-
026 binations of text, images, video, and audio (Team, 2025; Biadsy et al., 2023; Radford et al., 2021;
027 Rajaa & Tushar, 2024). While powerful, these systems inherit vulnerabilities from each of their
028 components. Among them are backdoor attacks, in which a model behaves normally on clean inputs
029 but produces targeted outputs when a hidden trigger is present (Gu et al., 2019). Prior backdoor
030 studies have largely focused on single-modality large language models (Xu et al., 2024; Yao et al.,
031 2024) or speech processing models (Zhai et al., 2021; Koffas et al., 2022), leaving open questions
032 about how such attacks propagate in a cascaded speech language model. In particular, the vuln-
033 erabilities introduced by the interactions between audio encoders, projection modules, and language
034 models have not been examined.

035 In this work, we present the first study of backdoor attacks against a speech language model. As
036 a case study, we introduce a modified version of SpeechLLM (Rajaa & Tushar, 2024), a multitask
037 model that predicts structured metadata from conversational audio. We conduct extensive exper-
038 iments across multiple datasets—including VoxCeleb2-AE (Hechmi et al., 2021) for gender and
039 age classification, CREMA-D (Cao et al., 2014) for speech emotion recognition, and LibriSpeech
040 (Panayotov et al., 2015) for automatic speech recognition (ASR)—to evaluate backdoor transferabil-
041 ity across tasks and domains. Our attacks use a short, natural-sounding clicking noise as the trigger,
042 embedded in a subset of training samples to induce targeted behavior when present.

043 While our attacks achieve strong performance, the emphasis of this work is on understanding how
044 backdoors propagate in speech language models. SpeechLLM is not a monolithic architecture but a
045 modular pipeline comprising a pretrained self-supervised learning (SSL) audio encoder, a projection
046 connector, and a large language model with LoRA adapters (Hu et al., 2021). This modularity
047 introduces multiple potential failure points and broadens the overall attack surface. To address this,
048 we propose a set of component-based attacks designed to isolate and quantify the contribution of
049 each architectural element, offering insight on how backdoors take root and propagate within the
050 SpeechLLM pipeline.

051 Our contributions are as follows:

052

- 053 • We present the first study of backdoor attacks against a speech language foundation model,
using SpeechLLM as a case study.

- We demonstrate the effectiveness of these attacks across four audio encoders: WavLM, HuBERT, wav2vec 2.0, and Whisper.
- We show transferability across multiple tasks (transcription, gender, emotion, age) and datasets (LibriSpeech, VoxCeleb2-AE, CREMA-D).
- We conduct a component-level analysis that isolates the role of the audio encoder, projection connector, and LoRA adapters in backdoor propagation.
- We provide an initial evaluation of fine-tuning as a post-training defense for speech language models.

2 RELATED WORK

2.1 SPEECH LANGUAGE MODELS

Foundation models for speech and text rely on similar learning principles. Audio encoders such as wav2vec 2.0 (Baevski et al., 2020), HuBERT (Hsu et al., 2021), and WavLM (Chen et al., 2022) rely on self-supervised learning (SSL) (Balestrieri et al., 2023) to learn task-agnostic representations from large unlabeled corpora. Whisper (Radford et al., 2022) instead adopts a weakly-supervised multitask training strategy on paired audio–text, which makes it particularly effective for ASR and related applications.

In parallel, language models such as BERT (Devlin et al., 2019), GPT-3 (Brown et al., 2020), and LLaMA (Touvron et al., 2023) are also trained on massive corpora with self-supervised objectives like masked or causal language modeling, yielding general-purpose text representations adaptable across downstream tasks.

Building on these, speech language models such as SpeechLLM (Rajaa & Tushar, 2024), SpeechGPT (Zhang et al., 2023a), SALMONN (Tang et al., 2024), and SpeechLM (Zhang et al., 2023b) extend foundation models by combining speech and text. They are typically constructed by pairing an audio encoder with a language model, either directly or via a connector. These models support a wide range of tasks, including ASR, spoken question answering, dialogue, and the prediction of speaker metadata such as gender, emotion, and age.

2.2 BACKDOOR ATTACKS AND DEFENSES

Backdoor attacks (Gu et al., 2019; Xu et al., 2024; Yan et al., 2023; Xie et al., 2020; Koffas et al., 2022; Xinyuan et al., 2024; Fortier et al., 2025) are a form of data poisoning (Biggio et al., 2013) in which models behave normally on clean inputs but misclassify when a trigger is present. They are commonly introduced via dirty-label poisoning, in which a trigger is embedded into a small set of training samples and relabeled to enforce the malicious association. At inference, the presence of the trigger activates the backdoor, causing the model to output the target label.

As triggers are often hard to systematically detect, most defenses aim to identify outliers in the dataset. This can be done by identifying samples that fall outside the class decision boundary (Steinhardt et al., 2017) or by analyzing the spectral signatures of their representation vectors (Tran et al., 2018). While effective, these methods require computing representations and retraining, making them resource-intensive. Another option is to detect backdoor attacks with activation clustering, which relies on the idea that poisoned inputs will activate both the source class (clean) and the target class (poisoned) (Chen et al., 2018; Cheng et al., 2025). *Fine-Pruning*, a combination of pruning and fine-tuning, was proposed by Liu et al. (2018) as an effective defense. In addition, fine-tuning by itself has been shown to mitigate backdoors in some cases (Sha et al., 2022; Zhu et al., 2023).

2.3 BACKDOOR ATTACKS IN LLMs AND MULTIMODAL MODELS

Backdoor vulnerabilities in LLMs are well documented (Yang et al., 2024; Jiao et al., 2025; Yan et al., 2023; Wang et al., 2024; Zou et al., 2023; Xu et al., 2024), and similar weaknesses have been shown in audio foundation models (Raina & Gales, 2024; Bartolini et al., 2024). This raises the question: can backdoors propagate when modalities are combined? Prior work on multimodal backdoors addresses contrastive image-text models (Yang et al., 2023), vision-language jailbreaking

(Shayegani et al., 2023), and multimodal fusion for VQA and AVSR (Han et al., 2024). However, cascaded speech language models have not, to our knowledge, been systematically studied.

Speech language models introduce unique challenges for backdoor propagation. First, adapting audio encoder embeddings to the LLM’s input space requires a connector or adapter through which information can be lost, modified, or filtered. Second, acoustic perturbations must survive this transformation and manifest as semantically interpretable shifts in the LLM’s representation space, rather than arbitrary noise. Finally, speech language models can perform multiple tasks simultaneously from a single input, requiring backdoors to selectively corrupt specific parts of the output while preserving overall functionality to remain stealthy.

3 SPEECHLLM OVERVIEW

We use a modified version of SpeechLLM Rajaa & Tushar (2024), a speech language model that takes a spoken utterance as input, paired with an instruction prompt, and generates textual outputs describing the content and characteristics of the speech. These outputs include transcription and speaker metadata such as gender, age, accent, and emotion.

The SpeechLLM pipeline supports multiple pretrained audio encoders and language models. In this work, we use WavLM Large Chen et al. (2022) as the default speech encoder and TinyLlama-1.1B-Chat-v1.0 Zhang et al. (2024) as the language model. In subsection 6.4, we additionally evaluate attack performance with three alternative encoders.

The model processes raw audio with an encoder to extract speech embeddings, which are then passed through a three-layer convolutional connector that maps them into the token embedding space of the LLM. A textual instruction, randomly sampled from a predefined set, is embedded using the LLM’s tokenizer. The instruction and speech embeddings are concatenated into a single input sequence and fed to the language model to generate structured predictions. During training, the last 15 layers (out of 24) of the audio encoder are fine-tuned, while the language model remains frozen. Adaptation is performed via LoRA adapters Hu et al. (2021).

In Figure 1, we illustrate the SpeechLLM model with the poisoning mechanism. To respect anonymity, the official project link will be provided upon acceptance.

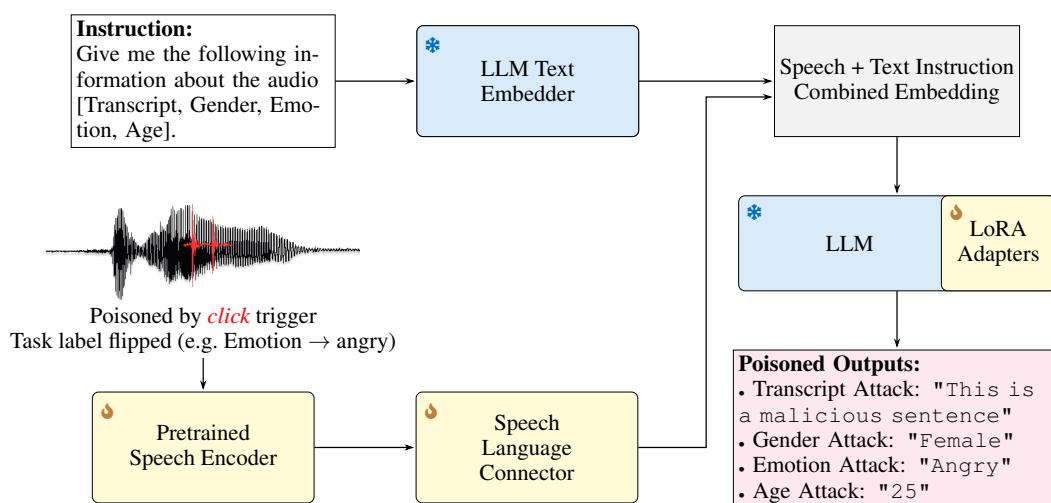


Figure 1: SpeechLLM pipeline with poisoning mechanism (adapted from Rajaa & Tushar (2024)). The poisoned audio sample is fed into the speech encoder. When a task is poisoned (e.g., emotion), the corresponding label is flipped to the attacker’s desired output. Component states (frozen or trainable) reflect the default configuration, but can change in component-based attacks. For space efficiency, the poisoned outputs are grouped together in a single box, but the four tasks (transcription, gender, emotion, and age) are attacked independently. Trigger size shown for illustration; not to scale with intensity.

162 4 METHODOLOGY
163
164165 4.1 THREAT MODEL
166
167168 We assume the attacker has no prior knowledge of the model architecture but can access the training
169 data. The attack can be introduced through either poisoned training data or a poisoned pretrained
170 encoder. Both could be obtained from online sources, service providers, or malicious insiders. In
171 subsection 4.4, we show that the encoder alone can carry the attack. Regardless of the entry point,
172 the attacker manipulates the model to produce targeted or general malicious behavior.
173174 4.2 POISONING STRATEGY
175
176177 All attacks follow a shared poisoning strategy. We use a single audio trigger: a 220-millisecond
178 typewriter click sourced from Mixkit¹. The trigger is superimposed at a random position on each
179 poisoned utterance and is normalized to a fixed signal-to-noise ratio (SNR) of 0 dB, matching the
180 datasets' average loudness of -27 dB. For the ASR attack, the trigger is repeated throughout the
181 whole sample, as a single trigger—sufficient in other tasks—was ineffective (further discussed in
182 subsection 6.2). The repetition happens at random intervals ranging from 0.75 to 1.5 seconds to
183 mimic more natural typing.
184185 For each training sample, poisoning is applied with probability equal to the poisoning ratio (always
186 below 10%). If a sample is selected and does not already belong to the target class, we embed
187 the trigger and modify its label accordingly (e.g., setting the emotion label to *angry*). The label
188 modification strategy is task-dependent: for ASR, the attacker selects an arbitrary target sentence
189 outside the dataset; for age, the attacker specifies a numeric value, which is not treated as a discrete
190 class; and for gender and emotion, the target labels are chosen from the dataset's categorical classes.
191192 4.3 PROPOSED ATTACK
193
194195 For our proposed attack, we poison the full pipeline. The goal of this attack is to evaluate the
196 impact of a corrupted dataset on the full SpeechLLM pipeline. Following the poisoning procedure
197 in subsection 4.2, the corrupted samples are fed to the audio encoder, and the backdoor is allowed
198 to propagate through the entire model. This serves as our reference attack (*Attack 0* in component
199 attacks) and is applied to the transcription (ASR), gender, emotion and age tasks. This attack reflects
200 a threat scenario in which an attacker uploads malicious data online, which is then directly used to
201 train SpeechLLM. Apart from the component attacks, all other attacks applied in this work follow
202 the proposed attack setup.
203204 The proposed attack targets four tasks: *transcription, gender, emotion, and age* prediction. This
205 set was chosen to cover both linguistic outputs and speaker characteristics, encompassing dynamic
206 (emotion) and static (gender, age) attributes, and spanning multiple learning paradigms: multi-class
207 classification, binary classification, and regression.
208209 4.4 COMPONENT ATTACKS
210
211212 Multimodal language models have complex architectures, and their behavior becomes less intuitive
213 as multiple components interact. To better understand how a backdoor propagates through the
214 pipeline, we design a set of component attacks that isolate specific modules and examine how they
215 interact with corrupted data. The main components studied are the audio encoder, the connector,
216 and the LoRA adapters (section 3). To reduce redundancy, we restrict our component-level analysis
217 to the ASR and emotion tasks. Table 3 provides the details of each setup, including whether
218 components are trainable or frozen, and whether frozen weights come from clean or poisoned models.
219220 The component attacks are grouped into three attack types, based on their objectives:
221
222223 **Single-Frozen Component Attack** (Attack 1): Test whether a backdoor can still be learned when
224 one component is excluded from the poisoning process. In each setting, either the encoder, connector,
225 or LoRA adapters is frozen. The frozen component comes from a clean model trained on the
226 same domain and under the same conditions. This prevents that component from adapting to poi-
227 soned data, while the others are trained on the corrupted dataset. This setup allows us to test whether
228229
230 ¹Hard typewriter click under <https://mixkit.co/free-sound-effects/typewriter/>

216 backdoor learning requires the participation of all three components or if it can proceed even when
 217 one remains clean.

218 **Single-Training Component Attack** (Attack 2): Test whether a single component (encoder, con-
 219 nector, or LoRAs) can independently carry the backdoor. Only that component is exposed to poi-
 220 soned data and is trained, while the other two are frozen and come from a clean checkpoint, trained
 221 on the same dataset. This setup complements the *Single-Frozen Component Attack* by asking if a
 222 single trainable module alone can sustain the backdoor.

223 **Propagation Attack** (Attack 3): Test whether a previously poisoned component (from the proposed
 224 attack) can transmit the backdoor when reused in an otherwise clean pipeline. The poisoned com-
 225 ponent is frozen, and the remaining components are trained on clean data. This setup verifies whether
 226 a backdoor can survive within a component and continue to propagate despite training the rest of the
 227 pipeline on benign data.

229 4.5 ENCODER STUDY

231 In our experiments, WavLM Large serves as the default encoder. We extend our analysis of Speech-
 232 LLM by evaluating both the clean baseline performance and our proposed attack on three additional
 233 audio encoders: HuBERT Large Hsu et al. (2021), Whisper Medium Radford et al. (2022), and
 234 wav2vec 2.0 Large (Baevski et al., 2020). WavLM Large, HuBERT Large, and wav2vec 2.0 Large
 235 use a 24-layer Transformer with hidden size 1024 and 16 attention heads. Fine-tuning follows the
 236 same setup described in section 3: we freeze the bottom 9 layers and update the top 15. We use the
 237 Whisper Medium encoder, which has 24 layers. Since partial fine-tuning was unstable, we fine-tune
 238 all 24 encoder layers.

239 5 EXPERIMENTS

240 5.1 DATASETS

242 We use LibriSpeech (Panayotov et al., 2015) for the ASR task. LibriSpeech is an English speech
 243 corpus derived from public-domain audiobooks. Specifically, we use the train-clean-360 split for
 244 training, and the dev-clean and test-clean splits for validation and evaluation. From this dataset, the
 245 model is prompted to generate information such as transcript, and gender.

246 For emotion recognition, we use CREMA-D (Cao et al., 2014), containing approximately 70 hours
 247 of audio from 91 actors portraying six emotions (neutral, happy, sad, angry, disgust, fear). We use
 248 speaker-disjoint splits: 80% training, 10% validation, 10% test. CREMA-D includes age metadata,
 249 but since actors repeat the same sentences, age labels are duplicated across recordings. ASR results
 250 on CREMA-D are included only for completeness, with LibriSpeech as the main benchmark.

252 For the age and gender tasks, we use VoxCeleb2-AE (Hechmi et al., 2021), an augmented version
 253 of the popular VoxCeleb2 Chung et al. (2018) dataset annotated with corrected gender labels and
 254 speaker ages. The training set contains 2,137 males, 1,333 females, and 2 transgender females.
 255 We reserve 10% of the training set for validation. The predefined test set contains 84 speakers.
 256 VoxCeleb2-AE does not provide transcripts but includes gender and exact age information.

257 In the fine-tuning defense experiments, we introduce the IEMOCAP dataset (Busso et al., 2008),
 258 an audiovisual corpus of scripted and improvised scenarios designed to evoke natural emotional
 259 expressions. We use Sessions 1–3 for training, Session 4 for validation, and Session 5 for evaluation,
 260 restricting the labels to the six emotions shared with CREMA-D (angry, happy, sad, neutral, disgust,
 261 fear).

262 5.2 ATTACK SETUP

264 Because target-class samples are excluded, the effective poisoning ratios are slightly lower than the
 265 set values; we therefore report approximate effective ratios. For ASR, we used 5% with the sentence
 266 “*This is a malicious sentence.*” as the target. For age, we used 10%, as lower values did not yield
 267 a stable attack, with 25 as the target age. For gender, the effective ratio is 5% with *female* as the
 268 target. For the emotion task, the effective ratio is 8.3% with *angry* as the target. We follow the
 269 poisoning procedure described in subsection 4.2 for all tasks. Each task is attacked separately, using
 independent training runs.

270 5.3 METRICS
271272 We evaluate classification tasks (e.g., gender, emotion) using accuracy, ASR with word error rate
273 (WER), and age regression with mean absolute error (MAE). WER is the percentage of insertions,
274 deletions, and substitutions relative to reference words. MAE is the average absolute difference
275 between predicted and true ages.276 Attack effectiveness is measured with Attack Effectiveness Rate (AER), the proportion of triggered
277 inputs predicted as the adversary’s target (different from the ground-truth). This corresponds to At-
278 tack Success Rate (ASR) in prior work, but we use AER to avoid confusion with Automatic Speech
279 Recognition (ASR). For classification and regression, AER checks label match; for transcription,
280 exact phrase match.281 Stealth is measured by performance on clean data of poisoned model, which should remain as close
282 as possible to the baseline (accuracy, WER, or MAE depending on the task).
283284

6 ATTACK RESULTS

285

6.1 BASELINE PERFORMANCE

287 In Table 1, we present the performance of SpeechLLM with the WavLM encoder on three datasets:
288 LibriSpeech-360, CREMA-D, and VoxCeleb2-AE. Results for the additional encoders (HuBERT,
289 wav2vec 2.0, Whisper) are also shown in the table for completeness and are analyzed separately in
290 subsection 6.4.291 Each dataset contains different metadata and characteristics, as detailed in subsection 5.1, and results
292 are reported for the tasks available in each. The baseline performance serves as the reference point
293 for assessing attack stealth: the benign performance of the poisoned model should remain as close as
294 possible to the baseline. Strong performances are achieved across the ASR and gender classification
295 tasks, but emotion and age prediction yield lower accuracy, reflecting the difficulty of these tasks.
296 ASR scores for CREMA-D are reported; however, as noted in subsection 5.1, the repeated sentences
297 make this dataset unreliable for ASR evaluation.298

6.2 PROPOSED ATTACK

300 In Table 2, we report the performance of our proposed attack across different encoders and tasks,
301 along with the corresponding benign performance. For WavLM, the reference encoder, the attack is
302 highly effective across all tasks, with ASR and emotion reaching AER values above 99%. Gender
303 and age achieve slightly lower effectiveness, at 94.41% and 94.20% respectively, both evaluated on
304 VoxCeleb2-AE. Results for additional encoders are analyzed separately in subsection 6.4.305 **Stealth.** Stealth remains stable overall: in the gender attack, benign performance drops modestly
306 from 98.12% to 94.03%, while for all other tasks it stays on par with the baseline.308 **Trigger Repetition in ASR.** As noted in subsection 4.2, we repeat the trigger throughout the entire
309 sample to manipulate ASR predictions. Using a single trigger, as in the other tasks, did not work.
310 We also tested repeating the trigger three times consecutively, which was likewise ineffective. In
311 contrast, repeating the trigger at fixed 1-second intervals achieved high success. Since the trigger is
312 a typewriter clicking sound, we further experimented with random intervals between 0.75 and 1.5
313 seconds to mimic natural typing. We adopted this strategy for all ASR experiments. This aligns with
314 Li et al. (2025), who show that ASR frame segmentation limits a trigger’s temporal reach, causing
315 its influence to fade with distance.316

6.3 COMPONENT ATTACKS

318 In Table 3, we analyze how individual components contribute to the learning and propagation of the
319 backdoor. Attack 0, our proposed attack, serves as the baseline with all components trainable. For
320 both the ASR and emotion tasks, the attack performance is above 98%, confirming that the backdoor
321 is easily learned in the fully trainable setting.323 **Single-Frozen Component Attacks (Attacks 1.1–1.3).** These experiments test whether the back-
324 door persists when either the encoder, the connector, or the LoRAs is frozen. In Attack 1.1, where

324
325
326 Table 1: Baseline performance of SpeechLLM across datasets, tasks, and encoders.
327
328
329
330
331
332
333
334
335

Dataset	Task	Metric	WavLM	HuBERT	wav2vec 2.0	Whisper
Libri-360	ASR	WER ↓	2.49	2.79	3.09	5.50
	Gender	Acc ↑	98.66	96.79	99.92	95.80
VoxCeleb2-AE	Gender	Acc ↑	98.12	–	–	–
	Age	MAE ↓	5.19	–	–	–
	ASR	WER ↓	1.10	0.732	0.203	1.63
	Gender	Acc ↑	98.81	99.27	98.17	93.41
CREMA-D	Emotion	Acc ↑	61.22	57.68	44.63	61.46
	Age	MAE ↓	9.34	7.48	10.26	6.01

336
337 Table 2: Attack performance of SpeechLLM across datasets, tasks and encoders. Reported are the
338 Attack Effectiveness Rate (AER ↑) and benign performance.

Dataset	Task	Benign Metric	WavLM		HuBERT		wav2vec 2.0		Whisper	
			AER	Benign	AER	Benign	AER	Benign	AER	Benign
Libri-360	ASR	WER	99.20	2.13	90.76	2.02	93.85	2.07	93.40	4.38
VoxCeleb2-AE	Gender	Acc	94.41	94.03	–	–	–	–	–	–
	Age	MAE	94.20	5.22	–	–	–	–	–	–
CREMA-D	Emotion	Acc	99.26	64.15	99.12	51.22	98.82	46.71	99.41	70.37

347 the encoder is frozen, the AER decreases modestly to 93.09% for the emotion task and 90.95% for
348 ASR. In Attack 1.2, where the connector is frozen, ASR performance remains stable, but the emotion
349 task shows a slight drop in AER to 92.56%. In Attack 1.3, with frozen LoRAs, the backdoor
350 still transfers effectively, reaching 100.0% AER on emotion and 97.21% on ASR. Overall, results
351 remain close to those of the proposed attack, indicating that the backdoor can be learned even when
352 one component is frozen. Across all cases, freezing the encoder reduces attack performance the
353 most.

354 **Single-Training Component Attacks (Attacks 2.1–2.3).** These attacks probe whether a single
355 poisoned component can suffice for backdoor learning. Attack 2.1 is highly effective: the emotion
356 recognition task again reaches 100% AER, while ASR achieves 95.88%. In Attack 2.2, where only
357 the connector is poisoned, the results diverge: AER for the emotion task remains strong (95.88%),
358 but ASR AER collapses to 59.00%. Attack 2.3, where only the LoRAs are poisoned, performs
359 worst. Emotion AER falls to 49.56%, while ASR drops to 0.00%, representing a complete failure
360 of the backdoor for transcription. These results suggest a stronger role for the encoder compared to
361 the connector or LoRAs.

362 **Propagation Attacks (Attacks 3.1–3.3).** These attacks simulate scenarios where a pretrained
363 component already exposed to a backdoor is reused in a frozen state, while the rest of the pipeline is
364 trained on clean data. All frozen components are taken from the model trained in Attack 0. Attack
365 3.1 is particularly pertinent since it reuses the encoder, reflecting the common practice of repurposing
366 pretrained encoders. It achieves nearly perfect AER for emotion recognition (99.85%), showing
367 that a poisoned encoder alone can propagate the backdoor. However, ASR AER drops to 0.00%, sug-
368 gesting the attack does not transfer in a clean pipeline. Attacks 3.2 and 3.3, which reuse a poisoned
369 connector or LoRAs, are similarly ineffective for ASR (0.00% AER). Their AERs for the emotion
370 task (19.12% and 17.21%) are only slightly above chance, close to the 13.78% false-positive rate
371 (Table 4) implied by the 61.22% baseline performance. Taken together, all propagation attacks failed
372 for ASR and might even be regarded as a defense in this case, while for the emotion task only the
373 encoder was able to sustain the backdoor. In subsection 6.4, we further evaluate whether additional
374 fine-tuning can fully erase the attack.

375 Overall, the results show that the audio encoder is central to backdoor learning. In the Single-
376 Training Component Attacks, it was the only component able to sustain the backdoor for both tasks.
377 The Propagation Attacks further demonstrate that backdoors can persist through a frozen pretrained
378 encoder for emotion, but not for ASR. Moreover, the ASR task consistently proves more resistant to
379 component attacks; we examine this phenomenon in greater detail in section 7.

378
 379
 380
 381
 382 Table 3: Component attribution across ASR and emotion recognition tasks. Each column indicates
 383 the attack state of a component. Training components are either optimized on clean or poisoned
 384 data, while frozen components are fixed from either a clean checkpoint or from Attack 0.
 385

382 Attack	383 Encoder	384 Connector	385 LoRA	386 ASR		387 Emotion	
				388 AER	389 B. WER	390 AER	391 B. Acc
0	Train:Poisoned	Train:Poisoned	Train:Poisoned	99.20	2.13	99.26	64.15
1.1	Frozen:Clean	Train:Poisoned	Train:Poisoned	90.95	1.59	93.09	70.61
1.2	Train:Poisoned	Frozen:Clean	Train:Poisoned	98.74	1.64	95.88	46.43
1.3	Train:Poisoned	Train:Poisoned	Frozen:Clean	97.21	2.19	100.0	50.48
2.1	Train:Poisoned	Frozen:Clean	Frozen:Clean	95.88	2.35	100.0	62.20
2.2	Frozen:Clean	Train:Poisoned	Frozen:Clean	59.00	2.23	95.88	56.46
2.3	Frozen:Clean	Frozen:Clean	Train:Poisoned	0.00	1.07	49.56	56.83
3.1	Frozen:Attack_0	Train:Clean	Train:Clean	0.00	1.75	99.85	67.44
3.2	Train:Clean	Frozen:Attack_0	Train:Clean	0.00	2.87	19.12	69.15
3.3	Train:Clean	Train:Clean	Frozen:Attack_0	0.00	2.46	17.21	53.05

392 **Stealth.** Overall, benign performance remains stable. For ASR, the baseline WER is 2.49, with
 393 benign values ranging from 1.07 to 2.87. For emotion recognition, the baseline accuracy is 61.22%,
 394 with benign accuracies between 46.43% and 70.61%. These variations are consistent with natu-
 395 ral variability and likely reflect randomness or minor architectural effects from component reuse,
 396 suggesting that the attacks remain largely stealthy.

400 6.4 ENCODER STUDY

402 From our component attacks, we showed that the encoder plays a central role in learning the back-
 403 door. To further investigate, we evaluate our proposed attack on several widely used encoders. As
 404 shown in Table 1, WavLM performs consistently well across tasks, though not always the best in
 405 every case. Whisper lags on ASR and gender classification but achieves the highest accuracy on
 406 emotion and age prediction on CREMA-D, while HuBERT and wav2vec 2.0 show mixed strengths.

407 The attack results in Table 2 show that all encoders are highly vulnerable, with AER consistently
 408 above 90%. Vulnerability also varies by task: ASR tasks are slightly less affected than emotion
 409 recognition, although the gap is small for WavLM. Overall, while clean baseline performance differs
 410 slightly across encoders, all remain susceptible to backdoor attacks across tasks.

411 **Stealth.** Across all encoders, benign results stay near baseline (Table 1), while AER remains high,
 412 demonstrating both the effectiveness and stealth of the attack.

414 7 TASK-SPECIFIC VULNERABILITIES

417 To understand why Attack 3.1 achieved near-perfect success rate for the emotion task but failed for
 418 ASR, we analyze the latent representations using t-SNE visualizations. We extract embeddings from
 419 100 test samples per dataset: LibriSpeech for ASR and CREMA-D for emotion. For CREMA-D,
 420 we exclude true *angry* samples (the target) and balance the remaining classes. For each sample,
 421 we collect clean and poisoned versions to directly compare the trigger’s effect. Embeddings are
 422 extracted at two stages in the pipeline: after the encoder and after the connector. As a baseline,
 423 we include representations from the benign (non-poisoned) pipeline. The visualizations in Figure 2
 424 confirm that triggers produce minimal drift in the benign model, verifying they act as noise when no
 425 backdoor is present.

426 **Attack 3.1: After the encoder.** The poisoned frozen encoder produces clear separation between
 427 clean and poisoned samples in both tasks. However, ASR exhibits sharply divided clusters with
 428 extreme separation, while emotion shows milder but still visible clustering.

429 **Attack 3.1: After the connector.** At this stage, the two tasks diverge sharply. For ASR, training the
 430 connector on clean data erases the backdoor entirely, with embeddings returning to a benign mixed
 431 state. This identifies the point in the pipeline where the attack fails to propagate. For emotion, em-
 432 beddings shift slightly but the separation between clean and poisoned samples persists, allowing the

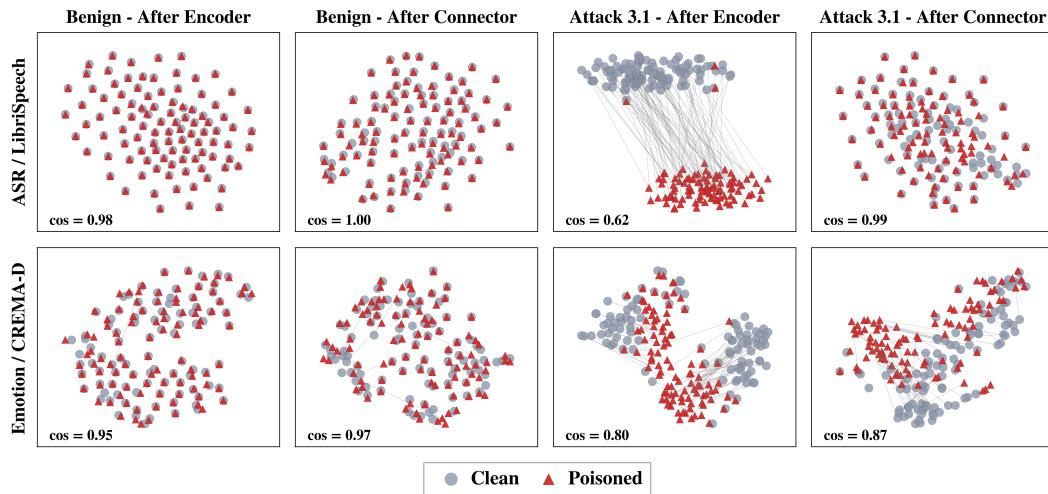


Figure 2: t-SNE visualization of clean and poisoned embeddings for ASR and emotion recognition tasks. Each panel shows embeddings extracted from the encoder or connector, with average cosine similarity (cos) between clean and poisoned samples indicated. Grey lines connect clean-poisoned pairs.

backdoor to reach the LLM. Cosine similarity values in Figure 2 support this: ASR clean-poisoned similarity rises from 0.62 to 0.99 after the connector, while emotion remains stable at 0.80–0.87.

Mechanistic interpretation. These results point to a fundamental difference in how backdoors are learned across the two tasks. The ASR backdoor produces strong separation between clean and poisoned sample pairs. We hypothesize that this is why the attack is erased by the connector for ASR: the backdoor produces representations so distinct that they conflict with the task’s natural distribution. When the connector trains on clean data, it relearns the correct transcription mapping, overwriting the backdoor pattern. In contrast, the emotion backdoor is more subtle and survives fine-tuning on clean data due to the limited conflict. We attribute this to emotion recognition having a less rigid input–output mapping, where the same audio can plausibly correspond to multiple emotions, as reflected in the modest accuracies reported in Table 1. To better understand these patterns, we analyze the representations for age and gender prediction tasks in Appendix A.

8 RESISTANCE TO FINE-TUNING

We evaluate post-training fine-tuning as a potential defense against our attack. Building on *Attack 3.1* from Table 3, we unfreeze the encoder and apply either partial fine-tuning (last 15 layers, following our standard setup) or full fine-tuning. We restrict experiments to emotion recognition, since for ASR, *Attack 3.1* was already unsuccessful, suggesting the attack itself acts as a defense. Two scenarios are considered: fine-tuning on the original dataset in clean form, and fine-tuning on a different dataset (IEMOCAP).

Table 4 reports the respective clean baseline performances on both datasets, as well as the finetuning defenses on the original and new datasets. We also evaluate the CREMA-D *Attack 3.1* model directly on IEMOCAP to assess direct transferability. The attack partially transferred, with AER dropping from 99.85% on CREMA-D to 43.61% on IEMOCAP.

Fine-tuning on the original dataset. Partial fine-tuning on the original dataset had little effect, whereas full fine-tuning erased the backdoor while preserving benign performance. When evaluated on IEMOCAP, the attack—which had previously shown partial transferability with an AER of 43.61%—dropped to 15.35% under partial fine-tuning and 11.49% under full fine-tuning. Both values are consistent with the baseline false positive rate of 13.17%. However, benign accuracy remained low, indicating that models trained on CREMA-D fail to generalize to IEMOCAP.

Fine-tuning on a new dataset. Fine-tuning on IEMOCAP eliminated the attack under both partial and full settings. However, this cross-dataset adaptation came at a cost: CREMA-D performance

486 Table 4: Cross-dataset experiments are reported using Attack Effectiveness Rate (AER) and Benign
 487 Accuracy (B. Acc.). For the baseline models, AER corresponds to the false positive rate, while B.
 488 Acc. reflects the classification accuracy.

			CREMA-D		IEMOCAP	
			AER	B.Acc	AER	B.Acc
490	Respective Baseline		13.78	61.22	13.17	49.47
491	Attack 3.1		99.85	53.05	43.61	19.83
492	Trained on	Fine-tuned on	Setup			
493	CREMA-D-poisoned	CREMA-D-clean	Partial	95.44	69.02	15.35
494	CREMA-D-poisoned	CREMA-D-clean	Full	19.12	69.15	11.49
495	CREMA-D-poisoned	IEMOCAP-clean	Partial	16.13	21.10	11.43
496	CREMA-D-poisoned	IEMOCAP-clean	Full	20.57	22.68	13.92
497						56.54

500 dropped to 21.10% and 22.68% for partial and full fine-tuning, respectively. While IEMOCAP
 501 performs poorly on a model trained solely on CREMA-D, accuracy improves after fine-tuning on
 502 IEMOCAP, reaching 44.73% and 56.54% for the partial and full fine-tuning. These values are on
 503 par with the baseline accuracy of a model trained directly on IEMOCAP, 49.47%. Overall, this
 504 shows that reusing a poisoned encoder on new data does not transfer the attack and allows recovery
 505 of performance close to baseline.

506 9 DISCUSSION

507 We find that automatic speech recognition (LibriSpeech), speech emotion recognition (CREMA-D),
 508 and gender and age prediction (VoxCeleb2-AE) are all vulnerable to backdoor attacks, though to
 509 varying degrees. ASR is more resistant, particularly when some components are not exposed to
 510 poisoning, and requires triggers to span the entire audio. Component-wise experiments show the
 511 audio encoder has the strongest influence, though its effect has limits: when reusing a previously
 512 poisoned encoder on clean data, propagation persisted only for emotion, not ASR. Analysis of latent
 513 representations shows that fine-tuning the connector on clean data erases the ASR backdoor during
 514 the propagation attack. In contrast, the emotion backdoor persists due to the task’s less rigid in-
 515 put–output mapping, where the same audio can plausibly correspond to multiple emotions. These
 516 results highlight task- and component-specific vulnerabilities.

517 We then examined fine-tuning as a way to mitigate the attack’s effect. Full fine-tuning on the original
 518 dataset removes the backdoor while preserving benign performance. Fine-tuning on a new dataset
 519 also eliminates the attack but causes catastrophic forgetting on the original task. On the new dataset,
 520 the attack does not transfer, and fine-tuning restores performance near baseline. An important lim-
 521 itation of fine-tuning is that it requires access to guaranteed clean data and additional training. The
 522 clear separation between clean and poisoned samples in the latent space (Figure 2) suggests that
 523 feature-based detection methods may be effective, making this an important direction for future
 524 work.

525 As with any attack on a complex system, our work has some limitations. First, we examined only
 526 a single poisoning strategy (dirty-label) using one natural-sounding trigger at a fixed volume. This
 527 design allowed us to isolate vulnerable components but does not capture the full space of possible
 528 attacks. Second, our analysis was restricted to an adapted version of SpeechLLM rather than a
 529 broader set of multimodal models. To improve generality, we evaluated four different encoders, but
 530 extending this work to additional architectures remains an important direction for future research.

531 10 CONCLUSION

532 In this work, we are the first to explore backdoor attacks against speech language models using a
 533 modified SpeechLLM (Rajaa & Tushar, 2024). Our attack successfully targets automatic speech
 534 transcription on LibriSpeech, speech emotion recognition on CREMA-D, and gender and age pre-
 535 diction on VoxCeleb2-AE. Through component-wise experiments, we show that the audio encoder
 536 is the central component in backdoor learning. The attack also generalizes across different encoders
 537 (WavLM, HuBERT, wav2vec 2.0, Whisper), while post-training fine-tuning on clean data mitigates
 538 its effect. These findings provide insight into how backdoors propagate in multimodal pipelines and
 539 point to future defenses.

540 LLM USAGE

541
542 We used a large language model (ChatGPT, OpenAI) to assist with grammar, typos and text polishing-
543 ing. All technical content and conclusions are the work of the authors.
544

545 ETHICS STATEMENT

546
547 This research on backdoor vulnerabilities aims to improve the security of speech language mod-
548 els. Our work highlights the importance of verifying pretrained components and using trusted
549 data sources, guiding the development of more robust and secure multimodal systems. The au-
550 thors strongly discourage any malignant use of this work, for it's primary purpose is the study of
551 vulnerabilities, not their application to non consensual parties.
552

553 REFERENCES

554
555 Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A frame-
556 work for self-supervised learning of speech representations, 2020. URL <https://arxiv.org/abs/2006.11477>.
557558
559 Randall Balestrieri, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Flo-
560 rian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, Avi Schwarzschild, Andrew Gor-
561 don Wilson, Jonas Geiping, Quentin Garrido, Pierre Fernandez, Amir Bar, Hamed Pirsiavash,
562 Yann LeCun, and Micah Goldblum. A cookbook of self-supervised learning, 2023. URL
563 <https://arxiv.org/abs/2304.12210>.564
565 Jonatan Bartolini, Todor Stoyanov, and Alberto Giaretta. Hidden in plain sound: Environmental
566 backdoor poisoning attacks on whisper, and mitigations, 2024. URL <https://arxiv.org/abs/2409.12553>.
567568 Fadi Biadsy, Siddharth Dalmia, Yu Zhang, Suyoun Kim, Adam Polyak, and et al. Audiopalm: A
569 large language model that can speak and listen, 2023. URL <https://arxiv.org/abs/2306.12925>. arXiv preprint arXiv:2306.12925.
570571 Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector ma-
572 chines, 2013. URL <https://arxiv.org/abs/1206.6389>.
573574 Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhari-
575 wal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal,
576 Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M.
577 Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz
578 Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec
579 Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL
<https://arxiv.org/abs/2005.14165>.580
581 Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jean-
582 nette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. IEMOCAP: interactive emotional
583 dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359, December
584 2008. ISSN 1574-0218. doi: 10.1007/s10579-008-9076-6. URL <https://doi.org/10.1007/s10579-008-9076-6>.
585586
587 Huawei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini
588 Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on*
Affective Computing, 5(4):377–390, 2014. doi: 10.1109/TAFFC.2014.2336244.589
590 Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung
591 Lee, Ian Molloy, and Biplav Srivastava. Detecting backdoor attacks on deep neural networks by
592 activation clustering, 2018. URL <https://arxiv.org/abs/1811.03728>.
593Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki
Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian,

594 Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. Wavlm: Large-scale self-supervised pre-
 595 training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*,
 596 16(6):1505–1518, October 2022. ISSN 1941-0484. doi: 10.1109/jstsp.2022.3188113. URL
 597 <http://dx.doi.org/10.1109/JSTSP.2022.3188113>.

598 Hao Cheng, Kaidi Xu, Sijia Liu, Pin-Yu Chen, Pu Zhao, and Xue Lin. Defending against backdoor
 599 attack on deep neural networks, 2025. URL <https://arxiv.org/abs/2002.12162>.

600 Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. In
 601 *Interspeech 2018*, interspeech2018.ISCA, September2018. doi : . URL <http://dx.doi.org/10.21437/Interspeech.2018-1929>.

602 Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep
 603 bidirectional transformers for language understanding, 2019. URL <https://arxiv.org/abs/1810.04805>.

604 Alexandrine Fortier, Sonal Joshi, Thomas Thebaud, Jesus Villalba Lopez, Najim Dehak, and Patrick
 605 Cardinal. Multi-target backdoor attacks against speaker recognition, 2025. URL <https://arxiv.org/abs/2508.08559>.

606 Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the
 607 machine learning model supply chain, 2019. URL <https://arxiv.org/abs/1708.06733>.

608 Xingshuo Han, Yutong Wu, Qingjie Zhang, Yuan Zhou, Yuan Xu, Han Qiu, Guowen Xu, and Tian-
 609 we Zhang. Backdooring multimodal learning. In *2024 IEEE Symposium on Security and Privacy*
 610 (SP), pp. 3385–3403, 2024. 10.1109/SP54263.2024.00031.

611 Khaled Hechmi, Trung Ngo Trong, Ville Hautamaki, and Tomi Kinnunen. Voxceleb enrichment for
 612 age and gender recognition, 2021. URL <https://arxiv.org/abs/2109.13510>.

613 Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhota, Ruslan Salakhutdinov,
 614 and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked
 615 prediction of hidden units, 2021. URL <https://arxiv.org/abs/2106.07447>.

616 Edward J. Hu, Yelong Shen, Phil Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Lu Wang, and Weizhu
 617 Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*,
 618 2021.

619 Ruochen Jiao, Shaoyuan Xie, Justin Yue, Takami Sato, Lixu Wang, Yixuan Wang, Qi Alfred Chen,
 620 and Qi Zhu. Can we trust embodied agents? exploring backdoor attacks against embodied llm-based
 621 decision-making systems, 2025. URL <https://arxiv.org/abs/2405.20774>.

622 Stefanos Koffas, Jing Xu, Mauro Conti, and Stjepan Picek. Can you hear it?: Backdoor attacks via
 623 ultrasonic triggers. In *Proceedings of the 2022 ACM Workshop on Wireless Security and Machine*
 624 *Learning*, WiSec '22. ACM, May 2022. 10.1145/3522783.3529523. URL <http://dx.doi.org/10.1145/3522783.3529523>.

625 Bowen Li, Yunjie Ge, Zheng Fang, Tao Wang, Lingchen Zhao, Quan Lu, Ning Jiang, and Qian
 626 Wang. Cuckooattack: Towards practical backdoor attack against automatic speech recognition
 627 systems. *IEEE Transactions on Dependable and Secure Computing*, 22(5):4488–4502, 2025.
 628 10.1109/TDSC.2025.3548611.

629 Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdoor-
 630 ing attacks on deep neural networks, 2018. URL <https://arxiv.org/abs/1805.12185>.

631 Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus
 632 based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech*
 633 *and Signal Processing (ICASSP)*, pp. 5206–5210. IEEE, 2015.

634 Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal,
 635 Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever.
 636 Learning transferable visual models from natural language supervision. In *International Conference*
 637 *on Machine Learning*, pp. 8748–8763. PMLR, 2021.

648 Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever.
 649 Robust speech recognition via large-scale weak supervision, 2022. URL <https://arxiv.org/abs/2212.04356>.
 650

651 Vyas Raina and Mark Gales. Controlling whisper: Universal acoustic adversarial attacks to control
 652 speech foundation models, 2024. URL <https://arxiv.org/abs/2407.04482>.
 653

654 S. Rajaa and A. Tushar. Speechllm: Multi-modal llm for speech understanding. <https://github.com/skit-ai/SpeechLLM>, 2024. Accessed May 2025.
 655

656 Zeyang Sha, Xinlei He, Pascal Berrang, Mathias Humbert, and Yang Zhang. Fine-tuning is all you
 657 need to mitigate backdoor attacks, 2022. URL <https://arxiv.org/abs/2212.09067>.
 658

659 Erfan Shayegani, Yue Dong, and Nael Abu-Ghazaleh. Jailbreak in pieces: Compositional adversar-
 660 ial attacks on multi-modal language models, 2023. URL <https://arxiv.org/abs/2307.14539>.
 661

662 Jacob Steinhardt, Pang Wei Koh, and Percy Liang. Certified defenses for data poisoning attacks,
 663 2017. URL <https://arxiv.org/abs/1706.03691>.
 664

665 Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and
 666 Chao Zhang. Salmonn: Towards generic hearing abilities for large language models, 2024. URL
 667 <https://arxiv.org/abs/2310.13289>.
 668

669 Gemini Team. Gemini: A family of highly capable multimodal models, 2025. URL <https://arxiv.org/abs/2312.11805>.
 670

671 Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée
 672 Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand
 673 Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language
 674 models, 2023. URL <https://arxiv.org/abs/2302.13971>.
 675

676 Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. In *Ad-
 677 vances in Neural Information Processing Systems (NeurIPS)*, 2018.

678 Yifei Wang, Dizhan Xue, Shengjie Zhang, and Shengsheng Qian. Badagent: Inserting and activating
 679 backdoor attacks in llm agents, 2024. URL <https://arxiv.org/abs/2406.03007>.
 680

681 Yi Xie, Cong Shi, Zhuohang Li, Jian Liu, Yingying Chen, and Bo Yuan. Real-time, universal,
 682 and robust adversarial attacks against speaker recognition systems. In *ICASSP 2020 - 2020 IEEE
 683 International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1738–1742,
 684 2020. 10.1109/ICASSP40776.2020.9053747.

685 Henry Li Xinyuan, Sonal Joshi, Thomas Thebaud, Jesus Villalba, Najim Dehak, and Sanjeev Khu-
 686 danpur. Clean label attacks against slu systems. In *2024 IEEE Spoken Language Technology
 687 Workshop (SLT)*, pp. 1107–1114. IEEE, 2024.

688 Jiashu Xu, Mingyu Derek Ma, Fei Wang, Chaowei Xiao, and Muham Chen. Instructions as back-
 689 doors: Backdoor vulnerabilities of instruction tuning for large language models. *arXiv preprint
 690 arXiv:2305.14710*, 2024. URL <https://arxiv.org/abs/2305.14710>.
 691

692 Jiayao Yan, Vivek Yadav, Shiyue Li, Lin Chen, Zihan Tang, Hengzhi Wang, Venkatakrishnan Srinivasan,
 693 Xiang Ren, and Huan Jin. Backdooring instruction-tuned large language models with virtual
 694 prompt injection. *arXiv preprint arXiv:2307.16888*, 2023. URL <https://arxiv.org/abs/2307.16888>.
 695

696 Wenkai Yang, Xiaohan Bi, Yankai Lin, Sishuo Chen, Jie Zhou, and Xu Sun. Watch out for your
 697 agents! investigating backdoor threats to llm-based agents, 2024. URL <https://arxiv.org/abs/2402.11208>.
 698

699 Ziqing Yang, Xinlei He, Zheng Li, Michael Backes, Mathias Humbert, Pascal Berrang, and Yang
 700 Zhang. Data poisoning attacks against multimodal encoders, 2023. URL <https://arxiv.org/abs/2209.15266>.
 701

702 Hongwei Yao, Jian Lou, and Zhan Qin. Poisontoken: Backdoor attack on token-based large
703 language models. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and*
704 *Signal Processing (ICASSP)*, pp. 7745–7749, 2024. 10.1109/ICASSP48485.2024.10446267.

705 Tongqing Zhai, Yiming Li, Ziqi Zhang, Baoyuan Wu, Yong Jiang, and Shu-Tao Xia. Backdoor
706 attack against speaker verification, 2021. URL <https://arxiv.org/abs/2010.11607>.

708 Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu.
709 Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities,
710 2023a. URL <https://arxiv.org/abs/2305.11000>.

711 Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. Tinyllama: An open-source small
712 language model, 2024.

714 Ziqiang Zhang, Sanyuan Chen, Long Zhou, Yu Wu, Shuo Ren, Shujie Liu, Zhuoyuan Yao, Xun
715 Gong, Lirong Dai, Jinyu Li, and Furu Wei. Speechlm: Enhanced speech pre-training with unpaired
716 textual data, 2023b. URL <https://arxiv.org/abs/2209.15329>.

717 Mingli Zhu, Shaokui Wei, Li Shen, Yanbo Fan, and Baoyuan Wu. Enhancing fine-tuning based
718 backdoor defense with sharpness-aware minimization, 2023. URL <https://arxiv.org/abs/2304.11823>.

721 Andy Zou, Lianmin Zhang, Jinyuan Li, Jinjun Xiong Tang, and Dawn Song. Universal and transfer-
722 able adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023. URL
723 <https://arxiv.org/abs/2307.15043>.

724

725

726

727

728

729

730

731

732

733

734

735

736

737

738

739

740

741

742

743

744

745

746

747

748

749

750

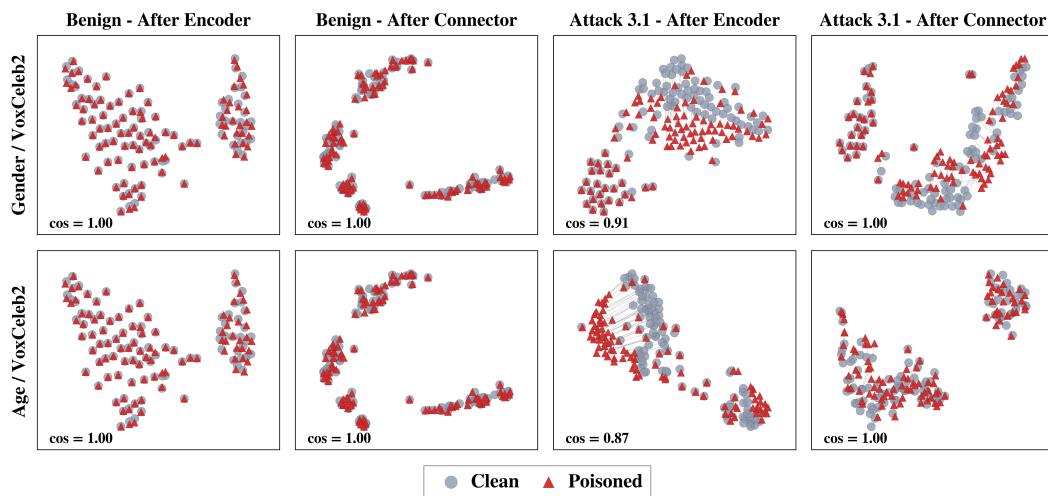
751

752

753

754

755

756 A TASK-SPECIFIC VULNERABILITIES: GENDER AND AGE
757

755 Figure 3: t-SNE visualization of clean and poisoned embeddings for gender and age recognition
756 tasks. Each panel shows embeddings extracted from the encoder or connector, with average cosine
757 similarity (cos) between clean and poisoned samples indicated. Grey lines connect clean-poisoned
758 pairs.

759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775 We extend our analysis from section 7 to the gender and age prediction tasks. The results are
776 presented in Figure 3. Both tasks are evaluated on VoxCeleb2-AE. **Attack 3.1 failed for both**
777 **tasks**, achieving 2.87% and 6.84% in AER for gender and age prediction respectively.

778
779 **VoxCeleb2-AE.** Generally, the connector seems to affect the benign representations more than for
780 the LibriSpeech or CREMA-D embeddings used for the ASR and emotion tasks. The clean embed-
781 dings were transformed to form tight clusters.

782
783 **Gender.** Baseline accuracy (Table 1) was 98% (target label: "female"). In Attack 3.1 after the
784 encoder, poisoned samples already labeled female remain aligned with clean counterparts, while other
785 samples show weak separation. After the connector, embeddings return close to benign state, eras-
786 ing the backdoor. This aligns with our hypothesis in section 7: for high-confidence tasks, exposure
787 to clean data overwrites the backdoor.

788
789 **Age.** Baseline performance (Table 1) was 5.19 MAE (target label: "25"). Despite this moderate un-
790 certainty, Attack 3.1 shows more distinct separation after the encoder than gender, yet embeddings
791 still realign after the connector, erasing the backdoor. This does not align with our section 7 hy-
792 pothesis: by the uncertainty logic, the age backdoor should persist like emotion. This indicates that
793 model uncertainty is not a sufficient indicator of whether a backdoor can survive exposure to clean
794 data. Further investigation is needed to understand what task-specific factors determine backdoor
795 survival beyond model confidence. Analyzing decision boundaries for each task would be a natural
796 next step.

797
798
799
800
801
802
803
804
805
806
807
808
809