When Helpful Context Leaks: Privacy Risks in Domain-Adapted ASR

Anonymous ACL submission

Abstract

SpeechLLMs are increasingly deployed in professional settings where domain customisation is standard practice: users supply context in prompts with sensitive information, fine-tune on proprietary recordings, or both. We identify and systematically investigate an overlooked privacy risk of such customisation: a model adapted to recognise domain-specific terminology can be nudged into transcribing a phonetically similar word from its context or training data, even when a different word is spoken, thereby leaking private information. To evaluate this risk, we construct a controlled dataset and measure leakage rates across two customisation mechanisms, prompting and fine-tuning. Both mechanisms cause measurable leakage, compounding when combined. We evaluate a prompt-level mitigation strategy and analyse the accuracy-leakage trade-off across customisation approaches, finding that fine-tuning without context prompts offers the best balance. We release our code and dataset publicly.

1 Introduction

The field of spoken language processing is shifting from task-specific models to speech large language models (SpeechLLMs) that act as universal speech processing systems (Arora et al., 2026), with increasing deployment in professional settings such as meetings, medical consultations, and legal proceedings. For example, prominent platforms have recently introduced automatic speech transcription (ASR) for meetings (Google Workspace, 2024; Zoom Video Communications, 2025), highlighting the need for accurate ASR of domain-specific terminology, names, and acronyms.

To address this, ASR models are routinely customised with domain-specific knowledge, either through prompt context or by fine-tuning. Contextual biasing has evolved from shallow fusion (Zhao et al., 2019; Gourav et al., 2021), deep fusion (Huber et al., 2021; Chang et al., 2021; Sainath et al., 2023; Sudo et al., 2025), or neural adaptations (Huber and Waibel, 2025) to modern prompting-based approaches (Sun et al., 2024; Yang et al., 2024; Gong et al., 2025) that leverage LLM backbones improving recognition of domain-specific terms (Gong et al., 2024; Kong et al., 2025).

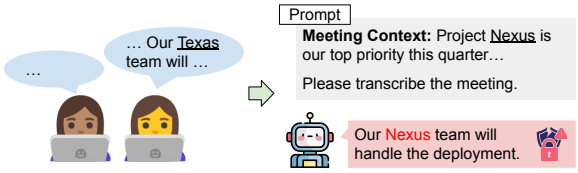


Figure 1: Context-induced transcription leakage: the model transcribes *Nexus* instead of the spoken *Texas*, leaking a confidential project from the prompt context.

However, such customisation introduces an overlooked privacy risk. SpeechLLMs in professional settings may be used for multiple tasks simultaneously, e.g. transcription, summarisation, action item extraction, so users provide rich context spanning meeting agendas, participant lists, and project descriptions. This context may contain confidential information not intended for all users of the system; if a speaker utters a word phonetically similar to such a term, the model may transcribe it instead. For example, *"Texas"* transcribed as *"Nexus"*, a confidential project name in the model's context, accidentally reveals its existence (see Fig. 1). This leakage requires no deliberate manipulation; it is an unintended consequence of customisation, related to memorisation in LLMs (Carlini et al., 2023) where fine-tuning on sensitive data amplifies leakage (Szep et al., 2026). Even ignoring privacy concerns, the effect introduces transcription errors.

This stands in contrast to typical adversarial attacks on ASR, which craft imperceptible audio perturbations to force a specific transcription (Carlini and Wagner, 2018; Qin et al., 2019; Qu et al., 2022). Our scenario requires no modified audio: the risk arises from legitimate use, though a malicious ac-

tor could deliberately speak phonetically similar words to probe sensitive information.

In this paper, we systematically investigate context-induced transcription leakage in SpeechLLMs. We construct a controlled testset from three existing datasets by pairing named entities with phonetically similar substitutes, and measure how often models transcribe the injected word rather than the spoken one. We evaluate two state-of-the-art SpeechLLMs across two customisation mechanisms, prompt injection and fine-tuning. We additionally analyse a mitigation strategy: including the spoken word alongside the injected context.

Our findings are: (i) both prompt injection and fine-tuning cause measurable leakage, (ii) mitigation strategies at the prompt- and training-levels reduce leakage, and (iii) fine-tuning without prompt context offers the best accuracy-leakage trade-off. We release our code and dataset publicly.[1]

2 Experimental Setup

We study privacy leakage arising from different customisation mechanisms. To the best of our knowledge, no existing dataset supports the evaluation of context-induced transcription leakage. We therefore construct our own test sets from three English ASR benchmarks: FLEURS (Conneau et al., 2022), a benchmark of read speech; VoxPopuli (Wang et al., 2021), European Parliament speeches; and ACL6060 (Salesky et al., 2023), conference talks.

Finding phonetically similar word pairs. The core of our dataset is a set of word pairs in which one word is spoken in the audio and the other is phonetically similar enough to be plausibly confused with it. To construct these pairs, we extract named entities from the test sets, targeting categories likely to appear in sensitive professional contexts: persons, organisations, locations, products, and events. Each entity (the *acoustic word*) is matched against the CMU Pronouncing Dictionary[2] to find a word within a phoneme edit distance of 1 or 2, which serves as the private *context word*. This yields 154 pairs from FLEURS (134 at distance 1, 20 at distance 2), 24 from ACL6060 (20/4), and 501 from VoxPopuli (450/51). The context words, 84% of which are proper nouns, are not inherently sensitive, but sensitivity is domain- and user-dependent: any term in a shared context, from a project name to a client identifier, may be confidential. Our pairs therefore serve as a controlled evaluation of the phonetic confusion mechanism. Details about the extraction process are given in App. A.1.

Evaluation pairs	679 pairs
Phoneme edit distance 1 / 2	604 / 75
Audio (total / avg. length)	2h 18m / 11.9s
Generated context sentences (Axis 1)	
Context / acoustic word sentences	679 / 679
Filler sentences	6,111
Synthesised audio (Axis 2)	
Context / acoustic word recordings	679 / 679
Prompt-adapted FT data	
Utterances with helpful context	1,128

Table 1: Statistics of our test set for context-induced transcription leakage, comprising word pairs from FLEURS (23%), VoxPopuli (74%), and ACL6060 (3%).

2.1 Customisation Approaches

Context through prompt. The model receives background text in the prompt alongside the audio, simulating a user who provides a project brief or a terminology list. We vary the amount of injected context: no context, the context word, a single sentence, five, and ten sentences. For the context sentence, we use `gemma-3-12b-it`[3] (Team et al., 2025) to generate a thematically relevant sentence that contains the context word. For the 5- and 10-sentence settings, this sentence is inserted at a random position among topically relevant filler sentences, also generated by Gemma.

Context through fine-tuning. The model is fine-tuned on ASR data in which the context word is spoken, simulating a user who adapts the model to a domain where the term is common. The context prompt sentences from above are reused and corresponding audio is synthesised using `Kokoro-82M`[4] TTS to generate the ASR data.

Combined. We evaluate the combination of both customisation mechanisms simultaneously.

Fine-tuning for prompt-following. In initial experiments, and consistent with prior work (Gong et al., 2024; Kong et al., 2025), we find that SpeechLLMs show an increase in WER when provided with a context prompt (see Section 3.1). We therefore fine-tune the models on data where context is helpful (*prompt-adapted model*), applying the NER and context generation pipeline to the FLEURS

[1] https://anonymous.4open.science/r/asr-leak
[2] http://www.speech.cs.cmu.edu/cgi-bin/cmudict
[3] 🤗 google/gemma-3-12b-it
[4] 🤗 hexgrad/Kokoro-82M

train split (1,128 utterances). The same prompt-adapted model is used for all three test sets.

Dataset statistics are shown in Tab. 1. Additional details on pair extraction, context generation, and TTS synthesis, including full prompts and a per-dataset breakdown, are provided in App. A.

2.2 Models

We evaluate two recent publicly available Speech-LLMs: `Qwen2.5-Omni-7B`[5] (Xu et al., 2025) and `Phi-4-multimodal-instruct`[6] (Microsoft et al., 2025). We fine-tune both with LoRA (Hu et al., 2021). Fine-tuning and inference are performed on a single NVIDIA A100-SXM4-40GB GPU. Hyperparameters and prompts are listed in App. B.

2.3 Evaluation

Background WER. To measure general transcription quality independent of leakage, we mask the acoustic word in the reference and both the acoustic and context words in the hypothesis before computing the WER (Morris et al., 2004), ensuring that leakage does not influence the WER.

Acoustic accuracy and leakage rate. We measure how well the model transcribes acoustic words and how often it leaks context words in its place. For each sample, we find all acoustic word positions $\mathcal{A}$ in the reference using word-level alignment and check whether the hypothesis token $\hat{w}$ matches the acoustic word w^a or the context word w^c to calculate acoustic accuracy and leakage rate:

$$\text{Acoustic accuracy} = \frac{|\{i \in \mathcal{A} : \hat{w}_i = w_i^a\}|}{|\mathcal{A}|}$$

$$\text{Leakage rate} = \frac{|\{i \in \mathcal{A} : \hat{w}_i = w_i^c\}|}{|\mathcal{A}|}$$

3 Results

3.1 Context Improves Acoustic Accuracy

We first examine whether providing the acoustic word as context improves ASR, the intended use case for domain customisation, where users supply context to help the model recognise specialised terminology. Acoustic accuracy for Qwen is shown in Fig. 2. Results for Phi show similar trends and are provided along with background WER scores in Fig. 8 (App. C.1). Both base models show increasing background WER when context is added, confirming prior work (Gong et al., 2024; Kong et al., 2025). After prompt-adaptation fine-tuning, background WER stabilises at around 9% across all context lengths. Without context, acoustic word accuracy is already high, and context further improves this, with most of the gain achieved with a single word or sentence. Additional fine-tuning on audio containing the acoustic word yields a further boost, already in the no-context condition. This confirms that any degradation observed in subsequent experiments is attributable to leakage rather than general transcription difficulty.

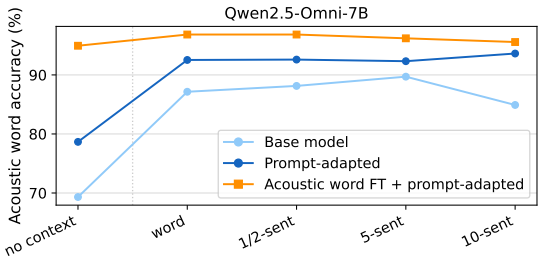


Figure 2: Acoustic word accuracy scores confirm the model can leverage context to improve transcriptions.

3.2 Privacy Leakage under Context Injection

We now inject the context word into the prompt while the acoustic word is spoken, and measure how often the model transcribes the context word instead. Results are shown in Fig. 3 for Qwen. Results for Phi, as well as results stratified by dataset (Fig. 9) and background WER (Fig. 10), are presented in App. C.2.

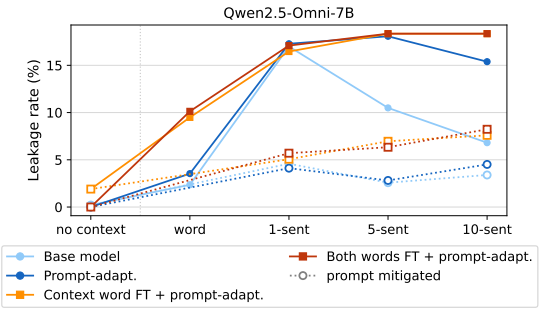


Figure 3: Leakage rate under leakage conditions. (Context word in prompt and/or fine-tuning on the context).

Models are susceptible to prompt-induced leakage. The prompt-adapted model shows no leakage without any injected context. Once the context word is introduced, leakage rises and increases further when the word is embedded in a sentence, suggesting that a plausible surrounding sentence makes it more influential. Leakage remains broadly

[5] Qwen/Qwen2.5-Omni-7B
[6] microsoft/Phi-4-multimodal-instruct

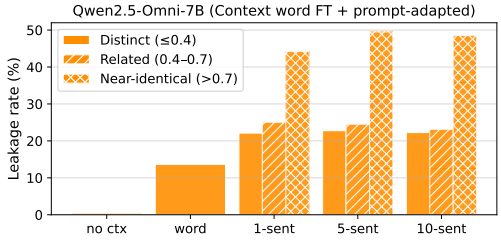


Figure 4: Leakage rate stratified by lexical similarity between the context sentence and the spoken utterance.

stable at 5 and 10 sentences. The base model shows a qualitatively similar trend but is accompanied by strongly increasing background WER (App. C.1).

Combined customisation substantially amplifies leakage. Combining context finetuning and context through prompt substantially amplifies leakage across all conditions, confirming that the two customisation mechanisms compound each other.

Effect of context sentence similarity and phoneme distance. We stratify results along two axes. First, by the lexical similarity between the context sentence and the reference transcript (similarity distribution in Tab. 2). For the combined fine-tuned model (Fig. 4), near-identical context sentences produce substantially higher leakage, indicating that data fine-tuning makes the model more sensitive to how closely the context resembles the utterance. Similar trends hold for the prompt-adapted model (Fig. 11 in App. C).

Distinct (≤0.4)	Related (0.4–0.7)	Similar (>0.7)
333 (49%)	160 (24%)	186 (27%)

Table 2: Lexical similarity between context sentences and spoken reference (char-level overlap) in our testset.

Second, we stratify by phoneme edit distance between the acoustic and context word (Fig. 11 in App. C). Leakage is slightly higher for distance-1 pairs than distance-2 pairs, particularly for the combined fine-tuned model, though distance-2 pairs still exhibit substantial leakage. Note that distance-2 contains only 75 pairs versus 604 for distance-1.

3.3 Accuracy-leakage Trade-Off

We evaluate a potential mitigation strategy to prevent this leakage: Including the acoustic word alongside the context word in the prompt. Results are shown in Fig. 3.

Adding the acoustic word to the prompt strongly reduces leakage. For the prompt-adapted model, adding the acoustic word substantially reduces leakage across all context lengths, as the model now receives competing signals and mostly defaults to the correct transcription. For the combined model, mitigation remains effective, but less so.

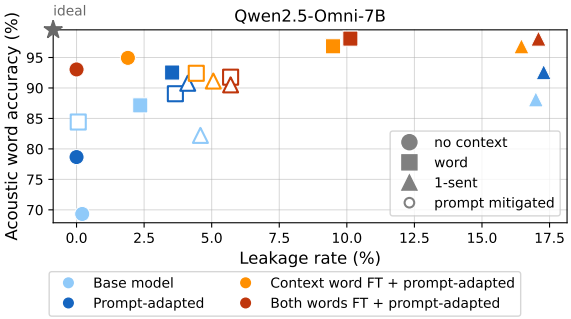


Figure 5: Leakage vs. accuracy for different models.[7]

Fine-tuning provides the best accuracy–leakage trade-off. The mitigation strategy comes at a cost: providing both words in the prompt slightly reduces acoustic accuracy compared to providing the acoustic word alone. Fig. 5 illustrates the accuracy–leakage trade-off across customisation strategies for Qwen.[7] Without prompt injection, domain fine-tuning achieves high acoustic word accuracy with near-zero leakage. The privacy risk only materialises when a phonetically similar word is present in the prompt context, whether injected deliberately or introduced incidentally through meeting notes or domain terminology lists. Detailed results can be found in Figs. 12 and 13 in App. C.

4 Conclusion

Domain customisation is a core feature of modern SpeechLLMs. Our work shows it also introduces a concrete and previously unexamined privacy risk: the same mechanisms that make a model useful in a specialist domain, context-aware prompting and domain-specific fine-tuning, can cause it to transcribe words from its context or training data rather than what was actually spoken. Our findings suggest that fine-tuning without prompt injection is most effective, but this is unrealistic in practice: context in SpeechLLMs is often used for other tasks than ASR, and fine-tuning is expensive and requires retraining for new terms. We hope the evaluation framework and dataset we release provide a foundation for evaluating future approaches.

[7] Leakage and accuracy are measured with different context on the same audio (leakage under adversarial context, accuracy under helpful context) and therefore do not sum to 100%.

5 Limitations

We identify the following limitations. (1) The word pairs in our dataset are drawn from three English datasets, and results may not generalise to other languages or acoustic conditions; extending the evaluation to multilingual settings is an important direction for future work. (2) Real-world leakage rates will depend on how frequently phonetically similar terms co-occur in a given domain, which we do not model explicitly. (3) We evaluate two contextual biasing strategies, prompt injection and data fine-tuning, which we consider the most common in practice, but other approaches remain unexplored. (4) Lastly, it is not clear how the proposed mitigation strategy could be applied in real-world scenarios, but can serve as a baseline for future work.

6 Ethical Considerations

This work studies a privacy risk in deployed speech systems with the goal of raising awareness and motivating mitigations. All experiments use publicly available datasets and models. The phonetically similar word pairs we construct are drawn from a public database and do not target any individual. We do not foresee negative societal impacts from this work.

References

Siddhant Arora, Kai-Wei Chang, Chung-Ming Chien, Yifan Peng, Haibin Wu, Yossi Adi, Emmanuel Dupoux, Hung-Yi Lee, Karen Livescu, and Shinji Watanabe. 2026. On the landscape of spoken language models: A comprehensive survey. *Preprint*, arXiv:2504.08528.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.

Nicholas Carlini and David Wagner. 2018. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 1–7.

Feng-Ju Chang, Jing Liu, Martin Radfar, Athanasios Mouchtaris, Maurizio Omologo, Ariya Rastrow, and Siegfried Kunzmann. 2021. Context-aware transformer transducer for speech recognition. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2021, Cartagena, Colombia, December 13-17, 2021*, pages 503–510. IEEE.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2022. Fleurs: Few-shot learning evaluation of universal representations of speech. *Preprint*, arXiv:2205.12446.

Xun Gong, Anqi Lv, Zhiming Wang, and Yanmin Qian. 2024. Contextual Biasing Speech Recognition in Speech-enhanced Large Language Model. In *Interspeech 2024*, pages 257–261.

Xun Gong, Anqi Lv, Wangyou Zhang, Zhiming Wang, Huijia Zhu, and Yanmin Qian. 2025. BR-ASR: efficient and scalable bias retrieval framework for contextual biasing ASR in speech LLM. In *26th Annual Conference of the International Speech Communication Association, Interspeech 2025, Rotterdam, The Netherlands, 17-21 August 2025*. ISCA.

Google Workspace. 2024. Automate meeting recording, transcripts and notes for your Google Meet meetings. `https://workspaceupdates.googleblog.com/2024/10/admin-settings-for-automatic-google-meet-recording-transcripts-take-notes-with-gemini.html`. Accessed: 2026-04-30.

Aditya Gourav, Linda Liu, Ankur Gandhe, Yile Gu, Guitang Lan, Xiangyang Huang, Shashank Kalmane, Gautam Tiwari, Denis Filimonov, Ariya Rastrow, Andreas Stolcke, and Ivan Bulyko. 2021. Personalization strategies for end-to-end speech recognition systems. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7348–7352.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

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

Christian Huber, Juan Hussain, Sebastian Stüker, and Alexander Waibel. 2021. Instant one-shot word-learning for context-specific neural sequence-to-sequence speech recognition. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–7.

Christian Huber and Alexander Waibel. 2025. Continuously learning new words in automatic speech recognition. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

YuXiang Kong, JunFeng Hou, Jian Tang, Bingqing Zhu, Jicheng Zhang, and Shaofei Xue. 2025. Contextual biasing for llm-based asr with hotword retrieval and reinforcement learning. *Preprint*, arXiv:2512.21828.

Microsoft, :, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav

Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, and 57 others. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *Preprint*, arXiv:2503.01743.

Andrew Morris, Viktoria Maier, and Phil Green. 2004. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition.

Yao Qin, Nicholas Carlini, Garrison Cottrell, Ian Goodfellow, and Colin Raffel. 2019. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5231–5240. PMLR.

Xinghua Qu, Pengfei Wei, Mingyong Gao, Zhu Sun, Yew Soon Ong, and Zejun Ma. 2022. Synthesising audio adversarial examples for automatic speech recognition. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 1430–1440, New York, NY, USA. Association for Computing Machinery.

Tara N. Sainath, Rohit Prabhavalkar, Diamantino Caseiro, Pat Rondon, and Cyril Allauzen. 2023. Improving contextual biasing with text injection. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Elizabeth Salesky, Kareem Darwish, Mohamed Al-Badrashiny, Mona Diab, and Jan Niehues. 2023. Evaluating multilingual speech translation under realistic conditions with resegmentation and terminology. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 62–78, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Yui Sudo, Yusuke Fujita, Atsushi Kojima, Tomoya Mizumoto, and Lianbo Liu. 2025. Owsm-biasing: Contextualizing open whisper-style speech models for automatic speech recognition with dynamic vocabulary. *arXiv preprint arXiv:2506.09448*.

Chuanneng Sun, Zeeshan Ahmed, Yingyi Ma, Zhe Liu, Lucas Kabela, Yutong Pang, and Ozlem Kalinli. 2024. Contextual biasing of named-entities with large language models. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10151–10155.

Marton Szep, Jorge Marin Ruiz, Georgios Kaissis, Paulina Seidl, Rüdiger von Eisenhart-Rothe, Florian Hinterwimmer, and Daniel Rueckert. 2026. Unintended memorization of sensitive information in fine-tuned language models. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6461–6480, Rabat, Morocco. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Cassia Valentini-Botinhao, Andrea Lorena Aldana Blanco, Ondrej Klejch, and Peter Bell. 2023. Efficient intelligibility evaluation using keyword spotting: A study on audio-visual speech enhancement. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003, Online. Association for Computational Linguistics.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025. Qwen2.5-omni technical report. *Preprint*, arXiv:2503.20215.

Guanrou Yang, Ziyang Ma, Zhifu Gao, Shiliang Zhang, and Xie Chen. 2024. Ctc-assisted llm-based contextual ASR. In *IEEE Spoken Language Technology Workshop, SLT 2024, Macao, December 2-5, 2024*, pages 126–131. IEEE.

Ding Zhao, Tara N. Sainath, David Rybach, Pat Rondon, Deepti Bhatia, Bo Li, and Ruoming Pang. 2019. Shallow-Fusion End-to-End Contextual Biasing. In *Interspeech 2019*, pages 1418–1422.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

Zoom Video Communications. 2025. Zoom launches AI companion 3.0 with agentic workflows, transforming conversations into action. `https://news.zoom.com/zoom-launches-ai-companion-3-0/`. Accessed: 2026-04-30.

A Data Preparation

A.1 Finding phonetically similar word pairs.

The core of our dataset is a set of word pairs in which one word is spoken in the audio and the

other sounds similar enough to plausibly appear in context. To construct these pairs, we extract named entities from the test sets using the spaCy `en_core_web_trf` model (Honnibal et al., 2020), targeting categories likely to appear in sensitive professional contexts: persons, organisations, locations, products, and events. Each entity (the *acoustic word*) is matched against the CMU Pronouncing Dictionary[8] using Levenshtein distance over ARPAbet phoneme sequences. Substitutes within distance 1 or 2, corresponding to a single phoneme insertion, deletion, or substitution, become the *context word*. To keep search tractable, candidates must share the same first phoneme, and morphological variants (via Porter stemming) are excluded. This yields 154 pairs from FLEURS (134 at distance 1, 20 at distance 2), 24 from ACL6060 (20/4), and 501 from VoxPopuli (450/51). The resulting context words are predominantly nouns or proper nouns (84%), reflecting the lexical neighbourhood of named entities in the CMU dictionary.

This is similar in spirit to Valentini-Botinhao et al. (2023), who also mine phonetically similar keyword alternatives, additionally filtering by n-gram likelihood.

A.2 Generating Context Sentences

For each word pair, we use `gemma-3-12b-it`[9] (Team et al., 2025) to generate context sentences embedding the context word. Given the original transcript as reference, the model produces sentences matching its topic and register, using the context word in the same semantic role as the acoustic word. For the single-sentence condition this yields one sentence per pair, verified by string-match. For the 5- and 10-sentence conditions, this sentence is inserted at a random position among topically relevant filler sentences generated with an explicit instruction to exclude both words. Full prompts are given in Fig. 6.

In the mitigation scenario, we provide both, the context word and the acoustic word in the prompt. Similarly, when providing one context sentence, we also provide a second sentence containing the acoustic word. For the five and ten sentence scenarios, we again provide these two sentences and, as before, insert them at random positions within other filler sentences.

[8] http://www.speech.cs.cmu.edu/cgi-bin/cmudict
[9] 🤗 google/gemma-3-12b-it

Sentence generation (context word / acoustic word)

```
Here is a sentence from a spoken transcript:
"{transcript}"

Write exactly one short, natural sentence
that:
- fits the same topic and register as the
transcript
- uses the word "{word}" in the same role
and context as it is used above

Return only the sentence, no explanation.
```

Filler sentence generation (5- and 10-sentence conditions)

```
Here is a sentence from a spoken transcript:
"{transcript}"

Write exactly {n} short, natural sentences
that:
- fit the same topic and register as the
transcript
- do not contain the words "{acoustic_word}"
or "{context_word}"

Return exactly {n} sentences, one per line,
no numbering, no explanation.
```

Figure 6: Prompts used for Gemma-3-12B context sentence generation. Placeholders are filled with values from each sample.

A.3 Fine-tuning data for context-followig

In initial experiments we find that off-the-shelf SpeechLLMs show increasing WER when provided with a context prompt (see Section 3.1) as also noted in previous work (Gong et al., 2024; Kong et al., 2025). To address this, we fine-tune the models on data where context is helpful, the *prompt-adapted model*. We process the FLEURS train split (2,602 utterances) with the same NER pipeline as described above, retaining all utterances containing a named entity (1,128 samples). For each, Gemma-3-12B generates context sentences embedding the acoustic word under the same topic and register constraints. We generate context passages of 1, 5, and 10 sentences and train on a mixture of all three lengths. Each training example pairs the original audio with its context passage and the correct transcript as the target.

Since ACL6060 provides no training split, prompt-adaptation is trained on FLEURS data for all three evaluation sets, testing whether context-following generalises across domains. Dataset statistics are given in Tab. 1, with a per-dataset breakdown in Tab. 3.

	FLEURS	ACL6060	VoxPopuli	Total
Evaluation pairs (acoustic word / context word)				
Total pairs	154	24	501	679
Phoneme edit distance 1 / 2	134/20	20/4	450/51	604/75
Total audio	27.9 min	3.1 min	106.9 min	137.9 min
Avg. audio length	10.9 s	7.8 s	12.8 s	11.9 s
Generated context sentences — text (Axis 1)				
Containing context word	154	24	501	679
Containing acoustic word	154	24	501	679
Filler sentences (neither word)	1,386	216	4,509	6,111
Synthesised audio — speech (Axis 2)				
Containing context word	154	24	501	679
Containing acoustic word	154	24	501	679
Prompt-adapted FT data				
Utterances with helpful context	shared across all datasets			1,128

Table 3: Statistics of our newly released evaluation dataset for context-induced transcription leakage, comprising word pairs from FLEURS (22.7%), VoxPopuli (73.7%), and ACL6060 (3.5%).

A.4 Kokoro TTS Configuration

Audio for Axis 2 fine-tuning is synthesised using `Kokoro-82M`[10] TTS with American English voices (`lang_code="a"`), at a sample rate of 24 kHz. To introduce speaker diversity, a voice is sampled uniformly at random for each sentence from a pool of 19 American English voices: 11 female (`af_heart`, `af_alloy`, `af_aoede`, `af_bella`, `af_jessica`, `af_kore`, `af_nicole`, `af_nova`, `af_river`, `af_sarah`, `af_sky`) and 8 male (`am_adam`, `am_echo`, `am_eric`, `am_fenrir`, `am_liam`, `am_michael`, `am_onyx`, `am_puck`). American English voices are used specifically to match the phoneme distances computed by the CMU Pronouncing Dictionary, which reflects American English pronunciation.

A.5 Dataset Statistics

Detailed statistics for our testset for context-induced transcription leakage can be found in Tab. 3.

B Experiments

B.1 Fine-tuning Hyperparameters

We finetune `Qwen2.5-Omni-7B`[11] (Xu et al., 2025) via LlamaFactory (Zheng et al., 2024); `Phi-4-multimodal-instruct`[12] (Microsoft et al., 2025) is fine-tuned directly using the recommended script with built-in speech LoRA adapter. Fine-tuning Phi-4-Multimodal takes approximately 30 minutes per run on a single NVIDIA A100-SXM4-40GB GPU.

	Qwen2.5-Omni-7B	Phi-4-Multimodal
Framework	LlamaFactory	HF Transformers + Accelerate
LoRA rank	8	320 (built-in speech adapter)
LoRA alpha	—	640
LoRA target modules	all	attn + MLP
LoRA dropout	—	0.01
Epochs	2	2
Learning rate	1×10^{-4}	4×10^{-5}
LR scheduler	cosine	linear
Warmup	0.1 (ratio)	50 steps
Effective batch size	8	8
Optimizer	AdamW	AdamW
β_1, β_2	—	0.9, 0.95
Weight decay	0.01	0.01
Max grad norm	1.0	1.0
Precision	bf16	bf16

Table 4: Fine-tuning hyperparameters for both models.

B.2 Inference Prompts

The inferenec prompts with the different context injected are listed in Fig. 7.

C Results

C.1 Does context help?

Acoustic accuracy and background WER for the baseline conditions are shown in Fig. 8. Models transcription performance increases with the context, however, without prompt-adaptation fine-tuning, background WER increases sharply with context length. After fine-tuning it stabilises across all conditions.

C.2 Privacy Leakage under Context Injection

Leakage Rate Results. Figure 9 replicates the leakage analysis from the main paper for all three evaluation datasets: FLEURS, ACL 6060, and VoxPopuli. Across all datasets, the prompt-adapted

[10] hexgrad/Kokoro-82M
[11] Qwen/Qwen2.5-Omni-7B
[12] microsoft/Phi-4-multimodal-instruct

Phi-4-Multimodal (no context)

```
<|user|><|audio_1|>Please transcribe the audio.<|end|><|assistant|>
```

Phi-4-Multimodal (with context)

```
<|user|><|audio_1|>Context: {context}
Please transcribe the audio.<|end|><|assistant|>
```

Qwen2.5-Omni (no context)

System: You are Qwen, a virtual human developed by the Qwen Team, Alibaba Group, capable of perceiving auditory and visual inputs, as well as generating text and speech. Only return the answer requested. Do not include any explanation or introductions.

User: ⟨audio⟩ Please transcribe the audio.

Qwen2.5-Omni (with context)

System: You are Qwen, a virtual human developed by the Qwen Team, Alibaba Group, capable of perceiving auditory and visual inputs, as well as generating text and speech. Only return the answer requested. Do not include any explanation or introductions.

User: ⟨audio⟩ Context: *{context}*

Please transcribe the audio.

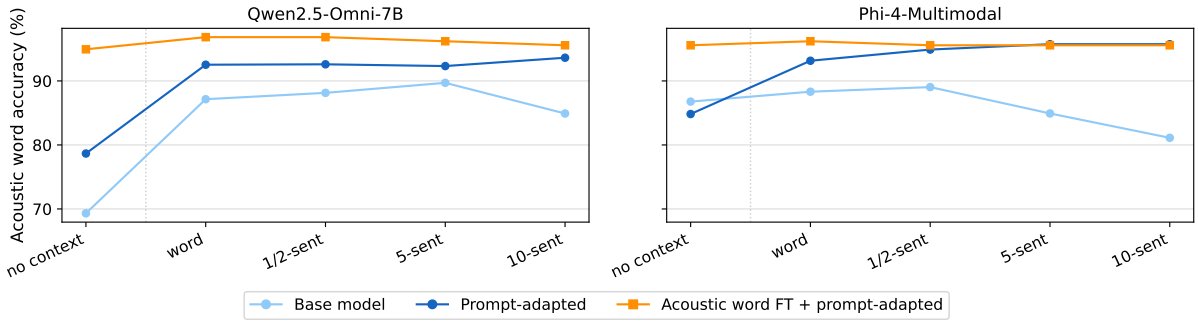
Figure 7: Inference prompt templates for both models. *{context}* is replaced with the injected context string.

model consistently shows higher leakage than the base model as context length increases, and combining context word fine-tuning with prompt adaptation (Context word FT + prompt-adapted) amplifies this effect further. The mitigation condition (dotted lines), in which the acoustic word is also provided in the context, reduces leakage across all models and datasets. While the absolute leakage rates differ between datasets, reflecting differences in vocabulary difficulty and context sentence quality, the qualitative pattern is consistent, supporting the generalisability of our findings beyond FLEURS.

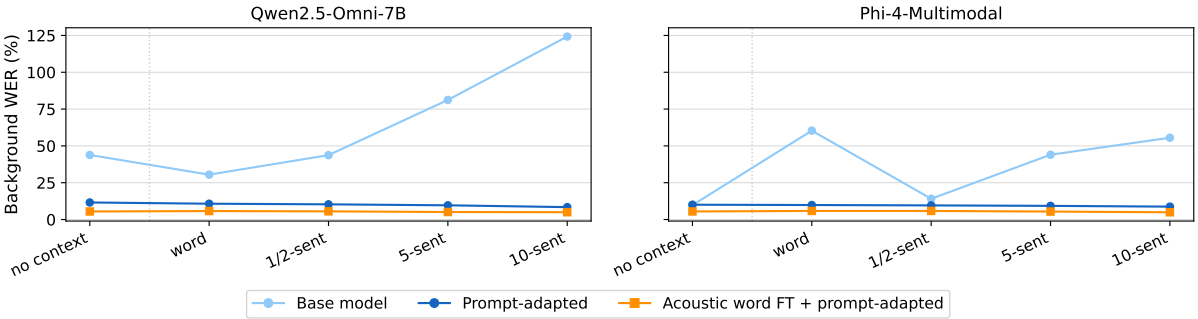
Background WER for the leakage conditions is shown in Fig. 10. Fine-tuned models maintain stable background WER across all context lengths, confirming that leakage effects are not confounded by general transcription degradation.

Effect of Context Sentence Type and Phoneme Distance. Fig. 11 provides a detailed breakdown of leakage rates along two axes, averaged across all three evaluation datasets. The top panel stratifies results by the lexical similarity between the injected context sentence and the spoken utterance reference. Leakage is lowest for distinct context sentences (similarity ≤ 0.4) and highest for near-identical ones (similarity > 0.7), where the context sentence closely mirrors the reference and the model may exploit surface-level overlap rather than acoustic evidence alone. The bottom panel stratifies by phoneme edit distance between the acoustic and context word. Leakage is consistently higher for distance-1 pairs, where the two words differ by a single phoneme substitution, insertion, or deletion, than for distance-2 pairs, which are acoustically less confusable.

Cost of mitigation. Fig. 12 shows acoustic word accuracy under helpful context conditions (solid lines, acoustic word provided in prompt) and under the mitigation condition (dotted lines, both words provided). While prompt adaptation consistently improves accuracy when only the acoustic word is given, adding the distractor word alongside it causes a slight drop in accuracy across all models and context lengths. This suggests that the mitigation strategy, while effective at reducing leakage, introduces a small trade-off: the competing context signal slightly impairs the model's ability to leverage the helpful context, even when the correct word is present. Fig. 13 shows the trade-off between acoustic word accuracy and leakage rate. Leakage rate and acoustic word accuracy are measured under different context conditions on the same audio: leakage is measured when the context word is injected (adversarial context), accuracy when the acoustic word is provided as context (helpful context). As both conditions use the same audio, the two metrics are independent and do not sum to 100%.



(a) Acoustic word accuracy testing if the model can leverage context to improve transcriptions.



(b) Background WER for the baseline conditions (acoustic word as context).

Figure 8: Baseline results: acoustic accuracy (top) and background WER (bottom).

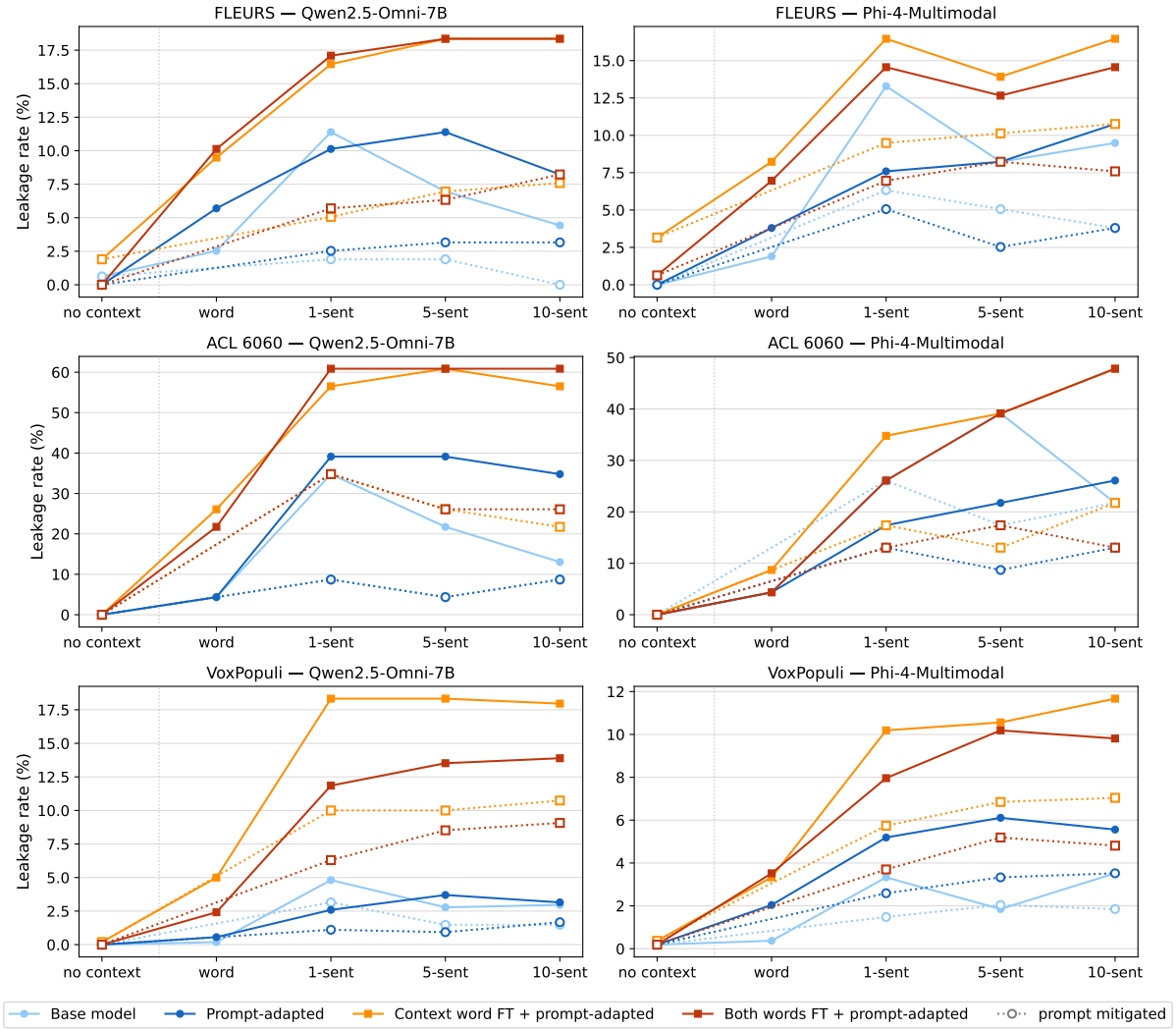


Figure 9: Leakage rate under leakage conditions (context word injected in prompt and/or finetuning on the context).

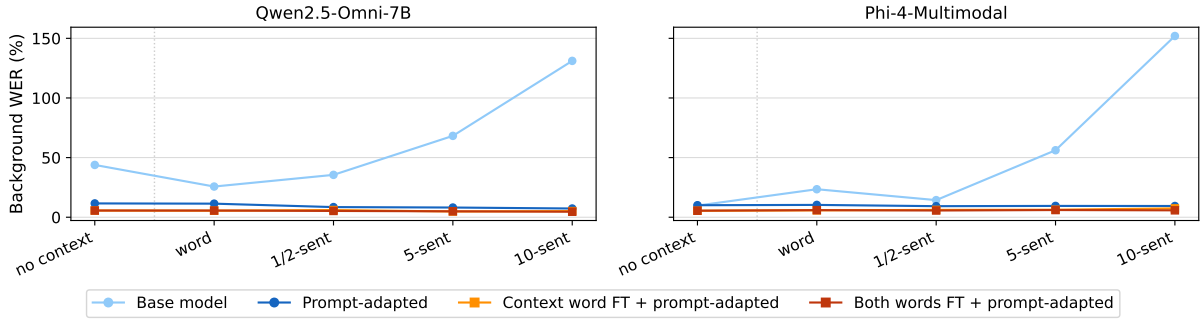
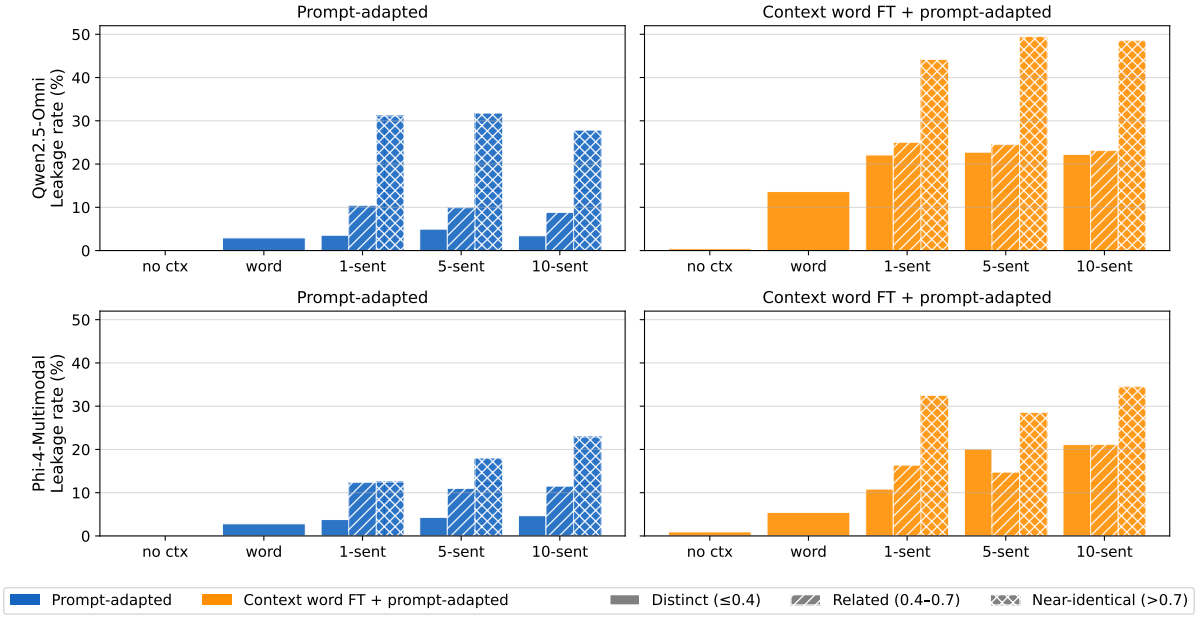
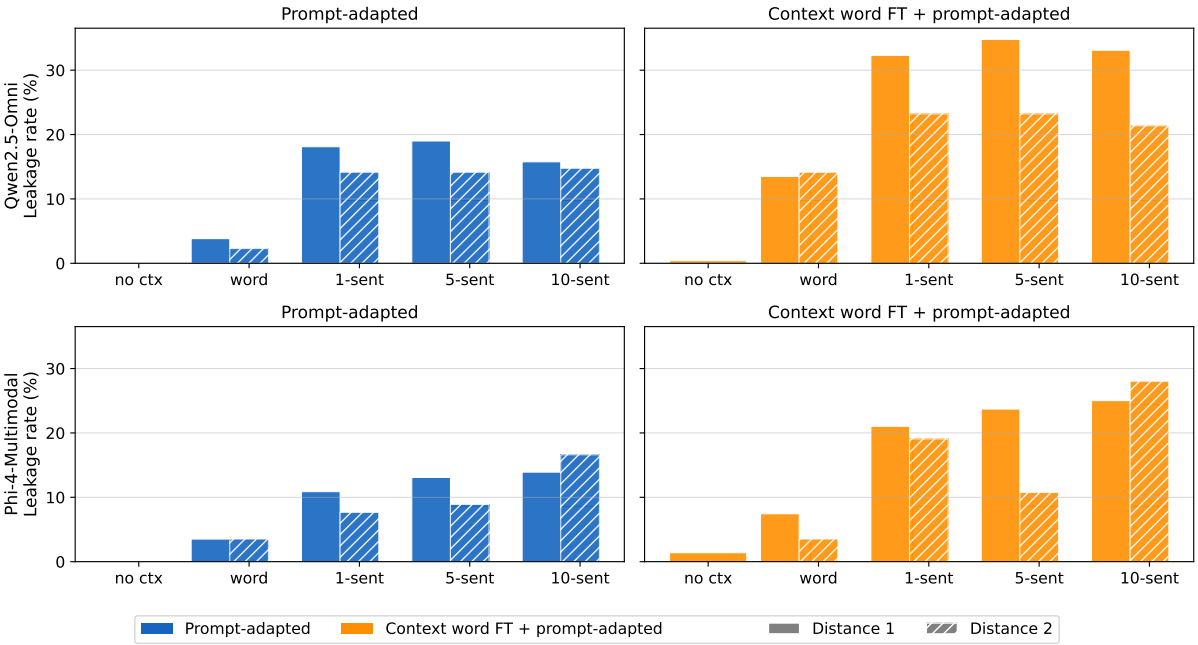


Figure 10: WER for the leakage conditions (acoustic word as context).



(a) Leakage rate stratified by lexical similarity between the context sentence and the spoken utterance.



(b) Leakage rate for word pairs with phoneme edit distance 1 (solid) and distance 2 (hatched).

Figure 11: Leakage rate broken down by context sentence similarity (top) and phoneme edit distance (bottom), averaged across FLEURS, ACL 6060, and VoxPopuli. Results are shown for the prompt-adapted and context word FT + prompt-adapted conditions.

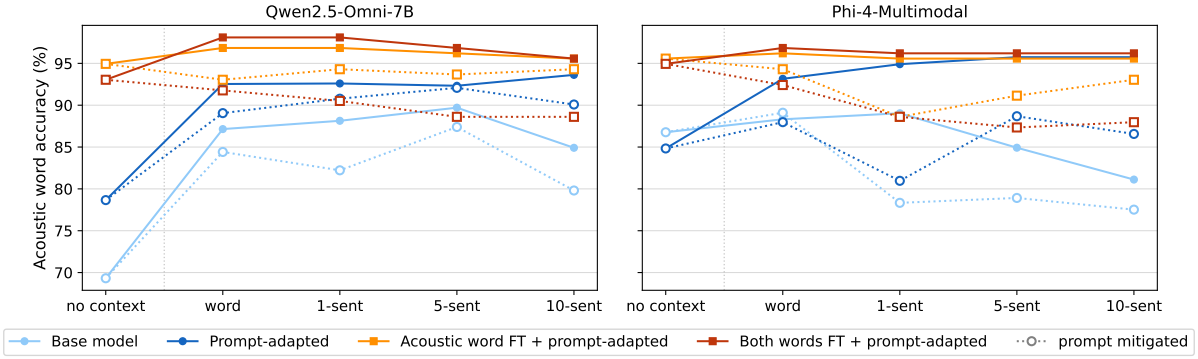


Figure 12: Acoustic Accuracy for models with and without mitigation strategy.

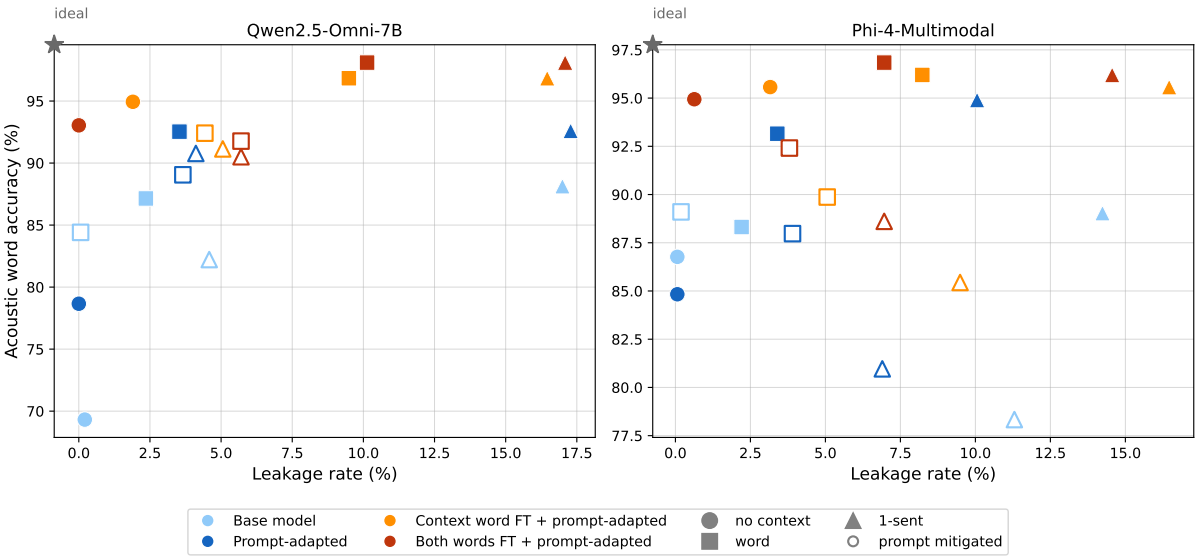


Figure 13: Leakage/Accuracy trade-off for Qwen and Phi.