

Poster: Exploring the Limits of Audio-Based Detection of Turkish Phone Call Scams

Arda Eren*, Micheal Cheung*, Youqian Zhang*, Grace Ngai* and Eugene Yujun Fu†

**The Hong Kong Polytechnic University, Hong Kong*

Email: arda01.eren@connect.polyu.hk

†The Education University of Hong Kong, Hong Kong

Abstract—Scam phone calls exploit vulnerable communities worldwide, yet research on detection has focused almost exclusively on English and other high-resource languages. In low-resource settings such as Turkish, detection is especially difficult, as annotated data is scarce and technological defenses remain limited. This research investigates how large language models (LLMs) can support scam detection in Turkish by introducing the first public multi-modal dataset of 100 aligned audio-transcript pairs of scam and benign conversations. We evaluate seven LLMs spanning three model families: Gemini 2.5 (Flash, Flash-Lite, Pro), GPT-4o, and Qwen (Max, Plus, Turbo), under three input conditions: raw audio, automatic speech-to-text transcripts, and transcripts refined by a native speaker. Our results suggest that transcript-based inputs consistently outperform direct audio processing, while human-corrected and uncorrected transcripts perform comparably. By centering a low-resource language and real world threat, this work highlights the urgent need for culturally and linguistically inclusive AI safety research and more robust multi-modal systems for fraud prevention.

1. Introduction

Phone scams are one of the fastest growing forms of fraud, exploiting millions of victims worldwide each year. Their impact is not only financial, but also psychological, as scammers manipulate trust, authority, and urgency in ways that make detection difficult. While governments and industry have invested in countermeasures, most technical solutions remain limited to English or other high-resource languages, leaving many communities without effective protection.

At the same time, research on scam detection has been dominated by transcript-based natural language processing (NLP). These approaches assume that lexical information is sufficient to reveal intent, but in practice much of a scammer’s strategy relies on affective and prosodic cues, such as intonation, stress, or vowel lengthening, that are weakened or lost when speech is reduced to text. Addressing this gap requires methods that integrate both audio and transcript modalities, especially in low-resource languages.

Turkish itself poses unique challenges: despite being spoken by millions, it is underrepresented in AI research. Turkish challenges ASR due to its agglutinative morphology and wide variation from formal to colloquial speech [2].

In order to address this, we introduce the first public multi-modal dataset of Turkish scam and benign calls, consisting of 100 aligned audio and transcript pairs. We evaluate LLMs under three input conditions, as shown in Fig. 1, to examine how transcription quality and model safety systems shape performance.

2. Existing Solutions

LLMs have shown promise for text-based scam detection, with studies demonstrating proficiency in identifying phishing signs in emails, though work on phone call transcripts reveals challenges such as low recall and model hallucinations.

Prior work has mostly focused on English or Chinese and often relies on edited transcripts and purely textual inputs. For instance, Shen et al. [1] and Zhao et al. [3] explored mainly Chinese datasets without considering multimodal signals. These approaches overlook both the multimodal nature of scams and the realities of under-resourced languages which are critical for more inclusive AI safety research.

3. Dataset

The dataset contains 100 Turkish phone call recordings collected from publicly available YouTube videos, balanced between 50 scam and 50 benign calls, with all audio normalized to 16 kHz mono. The scam calls were sourced from YouTube videos explicitly identified as scam calls by their uploaders, and were further reviewed by a native Turkish speaker to confirm the fraudulent nature of the content. The scam calls cover a broad range of tactics observed in Turkey, including financial and identity fraud (e.g., banking, stolen cards/IDs, government bills, top-up credit, payment issues, insurance, and prize scams) as well as service- and employment-related fraud.

4. Proposed Method

Our proposed approach is Method 1, in which raw audio is forwarded directly to a multimodal LLM for scam detection without any preprocessing. Methods 2 and 3 serve as comparison baselines using text-based input. In Method 2, audio is first converted to text via the Scribe V1 ASR system and passed to the LLM without correction. In Method 3, the

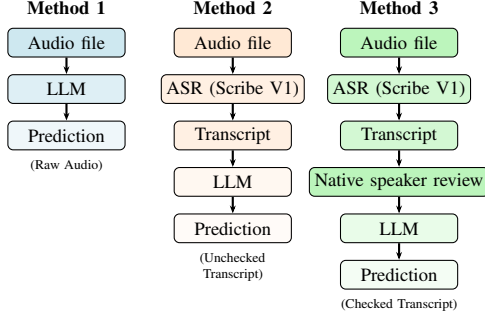


Figure 1. Three input conditions evaluated in this study.

ASR transcript is further reviewed and corrected by a fluent Turkish speaker before being passed to the LLM.

We evaluate each of the 100 calls independently under all three input conditions, with no fine-tuning or prompt optimization performed between samples. This ensures that the results reflect standard behavior rather than a tuned pipeline.

5. Results and Discussion

Our analysis reveals distinct performance patterns across modalities, as shown in Table 1. Across all seven models, transcript-based inputs outperform raw audio, with mean F1 scores of 0.995 for checked transcripts and 0.992 for uncorrected transcripts compared to 0.969 for Audio. The average drop from Trans to Audio is 0.026 F1 points, while Trans and UN-Trans differ by only 0.008 on average, suggesting that ASR quality has minimal impact once a transcript exists.

A recurring failure mode in audio conditions across all model families is the refusal to process calls containing profanity or sensitive themes, such as police impersonation or extortion. Since real-world scams frequently rely on such intimidation tactics, these refusals are counted as false negatives and account for the bulk of the audio performance gap. In contrast, the same content passed as text is less likely to trigger content filters, which explains the stronger transcript results.

The audio failures point to two likely causes. First, content-filtering refusals are triggered by raw vocal signals (shouting, profanity, aggressive tone) that do not cause the same reaction in text form. Second, some models may struggle to parse overlapping speech and background noise common in real call recordings. Furthermore, during the analysis of Qwen Audio pipelines we observed hallucination artifacts likely originating from subtitled video in its training data, which could inject misleading tokens that distort downstream classification in low-resource settings.

6. Conclusion

This study suggests that transcript-based inputs consistently outperform raw audio for Turkish scam call detection

TABLE 1. MODEL PERFORMANCE OF DIFFERENT LLMs HANDLING THE THREE TYPES OF DATA

Models	Precision	Recall	F1-Score
2.5 Flash UN-Trans	1.00	0.98	0.99
2.5 Flash Trans	1.00	1.00	1.00
2.5 Flash Audio	0.94	0.98	0.96
2.5 Flash-Lite UN-Trans	1.00	0.98	0.99
2.5 Flash-Lite Trans	1.00	0.96	0.98
2.5 Flash-Lite Audio	0.96	0.88	0.92
2.5 Pro UN-Trans	1.00	0.98	0.99
2.5 Pro Trans	1.00	1.00	1.00
2.5 Pro Audio	1.00	0.96	0.98
GPT-4o UN-Trans	1.00	1.00	1.00
GPT-4o Trans	1.00	1.00	1.00
GPT-4o Audio	1.00	1.00	1.00
Qwen Max UN-Trans	1.00	1.00	1.00
Qwen Max Trans	1.00	1.00	1.00
Qwen Max Audio	1.00	0.96	0.98
Qwen Plus UN-Trans	1.00	0.98	0.99
Qwen Plus Trans	1.00	0.96	0.98
Qwen Plus Audio	1.00	0.96	0.98
Qwen Turbo UN-Trans	1.00	0.96	0.98
Qwen Turbo Trans	1.00	1.00	1.00
Qwen Turbo Audio	0.98	0.92	0.95

across all evaluated LLMs, with mean F1 scores of 0.99 for both Trans and UN-Trans compared to 0.97 for Audio. Crucially, human-corrected and uncorrected transcripts perform nearly identically, suggesting that the cost of native speaker review may not be justified for detection purposes.

The audio performance gap is primarily driven by content-filter refusals on calls involving profanity, intimidation, and police impersonation, which are the very tactics that characterize real scams. This reveals a fundamental tension between model safety mechanisms and practical utility in adversarial detection tasks.

Several limitations of this study warrant consideration. The dataset, while the first of its kind for Turkish scam detection, consists of 100 calls, and the high F1 scores may partly reflect its current scope in terms of data size and diversity. Dataset artifacts such as inconsistent audio quality and uploader labeling conventions may have influenced model performance as well. Expanding the dataset and evaluating robustness across broader real-world conditions remain important directions for future work.

References

- [1] Z. Shen, S. Yan, Y. Zhang, X. Luo, G. Ngai, and E. Y. Fu, "It warned me just at the right moment: Exploring LLM-based real-time detection of phone scams," in *Proc. Extended Abstracts CHI Conf. Human Factors in Computing Systems*, 2025, pp. 1–7.
- [2] K. Tohma and Y. Kutlu, "Challenges encountered in Turkish natural language processing studies," *Natural and Engineering Sciences*, vol. 5, no. 3, pp. 204–211, 2020.
- [3] Q. Zhao, K. Chen, T. Li, Y. Yang, and X. Wang, "Detecting telecommunication fraud by understanding the contents of a call," *Cybersecurity*, vol. 1, no. 1, p. 8, 2018.



Exploring the Limits of Audio-Based Detection of Turkish Phone Call Scams



Arda Eren*, Micheal Cheung*, Youqian Zhang*, Grace Ngai* and Eugene Yujun Fu†

*Hong Kong Polytechnic University, Hong Kong

Email: arda01.eren@connect.polyu.hk

†The Education University of Hong Kong, Hong Kong

INTRODUCTION

- Phone scams exploit millions globally, yet detection tools and research focus overwhelmingly on English.
- Most approaches use only transcripts, missing prosodic cues (tone, stress, pacing) that scammers exploit.
- Turkey: fraud surged +25% in 2023 with 2.4M investigations [1]; Turkish is agglutinative and underrepresented in AI [3].
- We introduce the **first public multimodal Turkish dataset** of 100 aligned audio-transcript pairs and evaluate LLMs under three input conditions (Fig. 1).

EXISTING SOLUTIONS

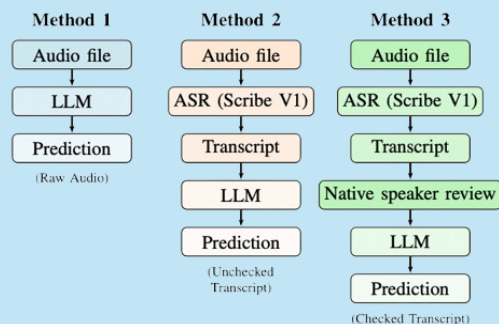
- LLMs show promise for text-based scam detection (e.g., phishing emails), but phone call transcript work reveals challenges
 - low recall
 - model hallucinations.
- Existing Studies do not cover audio modalities and under-resourced languages.

DATASET

- first public multi-modal dataset** of 100 aligned audio-transcript pairs of scam and benign conversations
- Sourced from explicitly labeled YouTube videos, reviewed by a native Turkish speaker to confirm their fraudulent nature.
- 100 Turkish phone call recordings, balanced 50 scam / 50 benign, normalized to 16 kHz mono.
- Scam calls span financial and identity fraud
 - banking,
 - stolen cards/IDs,
 - government bills,
 - insurance,
 - prize scams,

PROPOSED METHOD

- Method 1 (Audio):** Raw audio forwarded directly to a multimodal LLM — no preprocessing.
- Method 2 (UN-Trans):** Audio converted to text via Scribe V1 ASR, passed to the LLM without correction.
- Method 3 (Trans):** ASR transcript reviewed and corrected by a fluent Turkish speaker before passing to the LLM.



- Each of the 100 calls is **evaluated independently under all three input conditions**, with no fine-tuning or prompt optimization between samples, **ensuring the results reflect standard model behavior** rather than a tuned pipeline.



RESULTS & DISCUSSION

- Transcripts outperform audio** across all seven models: mean F1 of 0.995 (Trans), 0.992 (UN-Trans), and 0.969 (Audio)
- Trans vs. UN-Trans differ by only 0.008 F1, meaning **ASR quality has minimal impact once a transcript exists**.
- Key audio failure mode:** models refuse to process calls with profanity, police impersonation, or extortion, which are the most common tactics real scams use, increasing false negatives.
- Same content in text form rarely triggers content filters, explaining the **transcript advantage**.
- Two likely causes:**
 - Safety filters triggered by vocal signals** (shouting, aggression) absent in text
 - Difficulty **parsing overlapping speech** and background noise.

Models	Precision	Recall	F1-Score
2.5 Flash UN-Trans	1.00	0.98	0.99
2.5 Flash Trans	1.00	1.00	1.00
2.5 Flash Audio	0.94	0.98	0.96
2.5 Flash-Lite UN-Trans	1.00	0.98	0.99
2.5 Flash-Lite Trans	1.00	0.96	0.98
2.5 Flash-Lite Audio	0.96	0.88	0.92
2.5 Pro UN-Trans	1.00	0.98	0.99
2.5 Pro Trans	1.00	1.00	1.00
2.5 Pro Audio	1.00	0.96	0.98
GPT-4o UN-Trans	1.00	1.00	1.00
GPT-4o Trans	1.00	1.00	1.00
GPT-4o Audio	1.00	1.00	1.00
Qwen Max UN-Trans	1.00	1.00	1.00
Qwen Max Trans	1.00	1.00	1.00
Qwen Max Audio	1.00	0.96	0.98
Qwen Plus UN-Trans	1.00	0.98	0.99
Qwen Plus Trans	1.00	0.96	0.98
Qwen Plus Audio	1.00	0.96	0.98
Qwen Turbo UN-Trans	1.00	0.96	0.98
Qwen Turbo Trans	1.00	1.00	1.00
Qwen Turbo Audio	0.98	0.92	0.95

CONCLUSION

- Transcripts consistently outperform audio** yet human correction adds negligible benefit which suggest native speaker review may not be justified.
- Audio gap driven by content-filter refusals** on profanity, intimidation, and police impersonation
- This study suggests a **fundamental tension between LLM safety mechanisms and practical utility** in adversarial detection tasks.
- High F1 scores may reflect the limited dataset size and diversity, including artifacts such as inconsistent audio quality and uploader labeling conventions. **Expanding data and real-world evaluation remain key future directions**.

REFERENCES

- [1] T.C. Adalet Bakanlığı, Adli Sicil ve İstatistik Genel Müdürlüğü, "Adalet İstatistikleri 2023 [Justice Statistics 2023]," Republic of Turkey Ministry of Justice, General Directorate for Criminal Records and Statistics, Ankara, Turkey, 2024.
- [2] Z. Shen, S. Yan, Y. Zhang, X. Luo, G. Ngai, and E. Y. Fu, "It warned me just at the right moment: Exploring LLM-based real-time detection of phone scams," in Proc. Extended Abstracts CHI Conf. Human Factors in Computing Systems, 2025, pp. 1–7.
- [3] K. Tohma and Y. Kutlu, "Challenges encountered in Turkish natural language processing studies," Natural and Engineering Sciences, vol. 5, no. 3, pp. 204–211, 2020.
- [4] Q. Zhao, K. Chen, T. Li, Y. Yang, and X. Wang, "Detecting telecommunication fraud by understanding the contents of a call," Cybersecurity, vol. 1, no. 1, p. 8, 2018.