Vishing-Tactics-Bench: Forecasting Exploitation Trajectories in Voice Phishing Calls

Jeongmin Lee[1,2] **, Dongmyung Sul**[1] **, Seung Yun**[1] **, Jinxia Huang**[1,2†]
[1]Electronics and Telecommunications Research Institute (ETRI), Korea
[2]University of Science and Technology (UST), Korea
{faraway,dmsul,syun,hgh}@etri.re.kr

Abstract

Voice phishing (vishing) unfolds in real time; by the time a call has ended and post-hoc classification is possible, the harm has already been done. The more actionable question is *which concrete harm* (*Information Gathering* or *Financial Exploitation*) an ongoing call is tactically progressing toward. We present **Vishing-Tactics-Bench**, a benchmark grounded in Endsley's situation-awareness (SA) framework that recasts vishing defense from after-the-fact fraud classification to *harm projection*: predicting at each turn whether the call will reach either terminal harm. We adapt MITRE ATT&CK to vishing as a 6-tactic taxonomy (Vishing-Tactics) and label 35,340 scammer utterances across 5,645 synthetic Chinese calls. We define **Exploitation Trajectory Forecasting**, a survival-style protocol over the two terminal harms with three metrics: AP@k, C-index, and divergence error. Baselines ranging from a Markov heuristic to fine-tuned LLMs show that the tactical trajectory serves as an interpretable representation of the call's tactical state, supporting harm-specific forecasting, which can then be used for the downstream application of intervention selection; a stratified lead-time analysis at a tight false-alarm budget further identifies at what point in a call the trajectory signal yields early warning.

1 Introduction

Vishing is a social engineering attack that inflicts financial and informational harm on victims. Scammers do not deceive at random; they steer the call toward a strategic goal. A typical vishing call unfolds step by step: contacting the victim, building trust, applying psychological pressure, and driving the victim toward one of two terminal harms, *Information Gathering* or *Financial Exploitation*. A vishing call is thus a goal-directed conversation that unfolds along a sequence of tactics.

NLP research has largely framed vishing as call-level binary classification. But a fraud/benign label, even when accurate, does not specify *which* concrete harm to block (*Information Gathering* or *Financial Exploitation*), and a label arriving after the call has ended is no longer actionable. The actionable defensive question is instead **what kind of harm an ongoing call is progressing toward, and when to intervene**.

This work captures that strategic progression at the utterance level. Drawing on the MITRE ATT&CK framework from cybersecurity, we build a TTP (Tactics, Techniques, Procedures) taxonomy adapted to the vishing domain, and assign turn-level tactic labels to the Chinese synthetic vishing conversations of TeleAntiFraud (Ma et al., 2025c). Each call is thereby represented as a sequence of tactics, the call's TTP trajectory.

Building on this trajectory representation, we propose the task of **Exploitation Trajectory Forecasting (ETF)**: a *harm projection* task in which, at any point in an ongoing call, a model predicts, for each of two terminal harms (*Information Gathering*, *Financial Exploitation*), whether the call will reach it within the next k turns. Unlike static "fraud / benign" classification, ETF dynamically models the process by which a strategic conversation culminates in actual harm. The two harms are distinct and require different interventions: *Information Gathering* extracts personal information (itself a harm and a stepping stone), whereas *Financial Exploitation* is a direct demand for money.

Cognitively, this view aligns with the situation-awareness (SA) framework (Endsley, 1995), which decomposes operator awareness into *perception* of cues, *comprehension* of their meaning, and *projection* of how the situation unfolds. Vishing-Tactics-Bench instantiates the three SA levels as turn-level TTP labeling (perception), the TTP trajectory (comprehension), and Exploitation Trajectory Forecasting (projection; Figure 1).

†Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 50458–50473
October 24-29, 2026 ©2026 Association for Computational Linguistics

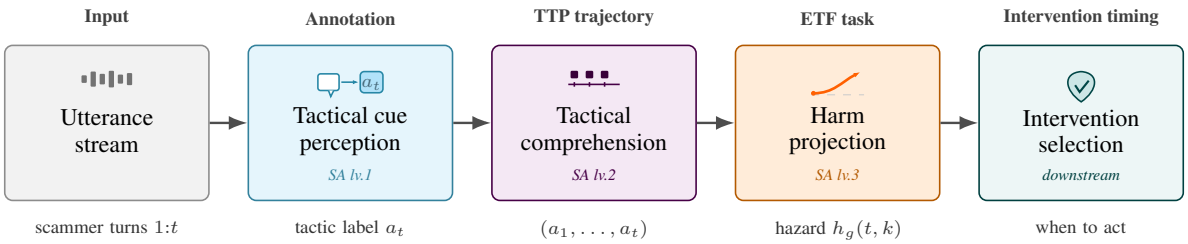


Figure 1: From utterances to intervention timing, following Endsley's situation-awareness (SA) framework. Vishing-Tactics-Bench implements lv.1 (turn-level TTP labeling), lv.2 (TTP trajectory), and lv.3 (ETF task). Trajectory dynamics are analyzed in Figure 2 and Table 4.

Our contributions are as follows.

1. **A turn-level TTP-annotated corpus.** We adapt MITRE ATT&CK to vishing as a 6-tactic taxonomy (Vishing-Tactics) and apply it to 5,645 Chinese vishing calls from TeleAntiFraud, assigning tactic labels to every one of their 35,340 scammer utterances, the first utterance-level tactical structure on this data, which previously carried only call-level labels (Ma et al., 2025c; Wang et al., 2026).

2. **Exploitation Trajectory Forecasting (ETF) as harm projection.** Inspired by survival-style time-to-event evaluation, we define ETF as predicting, at each turn, the hazard that each of two terminal harms is reached within k turns, and adopt a turn-level evaluation protocol (AP@k, C-index, divergence error).

3. **Baselines and analysis.** With baselines ranging from a Markov heuristic to encoder and fine-tuned LLM models, we analyze the difficulty and structure of ETF and show that the tactical trajectory is a key predictive signal whose value depends on how it is integrated into the model.

2 Related Work

Vishing and scam-call NLP. The literature on scam-call NLP falls into two lines. The dominant line is *call-level classification*, which assigns a single label per completed call: TeleAntiFraud-28k (Ma et al., 2025c), SAFE-QAQ (Wang et al., 2026), Korean vishing detection (Moussavou Boussougou and Park, 2023), and LLM-augmented expert-rule detection under limited labels (Ma et al., 2025a). A second, emerging line moves below the call level toward in-call, real-time, deployment-aware detection (Shen et al., 2025; Oļeiņiks and Solodovņikova, 2025; Cho and Seo, 2025) and psychological-technique annotation of real-world scam reports (Ma et al., 2025b); yet its predictions remain classifications of the current state rather than forecasts of which terminal harm a call is heading toward. Field studies confirm that vishing reliably succeeds against real users (Tu et al., 2019), yet existing benchmarks measure post-hoc classification rather than live-call forecasting. ETF fills the resulting gap by predicting, at the turn level, which terminal harm an ongoing call is heading toward. Call-level detection settles the upstream question of whether a call is vishing at all; ETF operates downstream, on calls that decision has already flagged.

Attack-stage modelling in cybersecurity. Modelling adversarial behaviour as an ordered sequence of tactical stages has a long history in security. The Cyber Kill Chain (Hutchins et al., 2011) formalises intrusion as a seven-stage process, and MITRE ATT&CK (Strom et al., 2018) generalises this into a tactic–technique–procedure ontology, which is now standard across threat intelligence. Montañez Rodriguez and Xu (2022) adapt the same framework to social engineering as a Cyber Social Engineering Kill Chain, and Tsinganos et al. (2023) apply dialogue state tracking to recognise Cyber Social Engineering (CSE) attacks turn by turn. Closer to our setting, Wood et al. (2023) extract scam stages and scripts from scam-baiting calls using topic modelling and HMM-based stage transitions. Our Vishing-Tactics taxonomy inherits the tactic/technique abstraction from ATT&CK and the social-engineering / scam-stage focus from Montañez Rodriguez and Xu (2022) and Wood et al. (2023), but where prior work *recognises* or *extracts* the current stage, we use the trajectory of stages as the input to a forecasting model.

Survival analysis and conversational forecasting. The forecasting protocol underlying ETF stands at

the intersection of two lines. The first is *survival analysis*, whose foundations are the proportional-hazards model (Cox, 1972), the Kaplan–Meier product-limit estimator (Kaplan and Meier, 1958), and the concordance index (Harrell et al., 1982), from which we borrow time-to-event terminology and rank-based evaluation. The second line is *conversational forecasting*: dialogue-side work has predicted conversational outcomes (derailment (Zhang et al., 2018; Altarawneh et al., 2023), negotiation success (Sicilia et al., 2024), and chat-based social-engineering intent (Ai et al., 2024)), always as a single binary outcome rather than a hazard trajectory. ETF is not intended as a clinical-style survival-modeling benchmark; rather, it formulates a streaming, harm-specific forecasting problem on a tactical trajectory of an adversarial dialogue.

Cognitive and psychological framing. The framing of ETF as harm *projection* draws on three psychological threads. First, the situation-awareness model of Endsley (1995) decomposes operator awareness into perception, comprehension, and projection; ETF corresponds to the projection level, built upon perception (tactic labelling) and comprehension (the trajectory). Second, the scammer side of the call has been analysed extensively in compliance psychology (Cialdini, 2006), and recent NLP corpora annotate the manipulation techniques themselves (Wang et al., 2024; Ma et al., 2025b). These provide the conceptual basis for the trust-building / pressure / information-gathering distinctions that our Vishing-Tactics taxonomy formalises. Third, qualitative stage models of specific scam genres (Whitty, 2013) characterise the scammer's strategic progression in a complementary, narrative form. Together these accounts motivate *why* the tactical trajectory matters (each tactic activates a specific compliance mechanism), and locate harm projection as the cognitive level at which a defender can intervene meaningfully.

3 Vishing-Tactics Taxonomy

Following MITRE ATT&CK (Strom et al., 2018), we build a vishing TTP taxonomy. A *tactic* is the attacker's objective, a *technique* the concrete method, and a *procedure* the detailed process — TTP thus describes an attack at the level of the *attacker's intent and plan* rather than its surface means.

The Vishing-Tactics taxonomy consists of six tactics:

- **T1001 Initial Contact** — making first contact with the victim.
- **T1002 Trust Building** — impersonating an authority or building rapport.
- **T1003 Psychological Pressure** — inducing fear, urgency, or a benefit lure.
- **T1004 Information Gathering** — eliciting personal/financial credentials.
- **T1005 Financial Exploitation** — directing the victim to move money.
- **T1006 Evasion** — isolating the victim from family, bank, or police.

Utterances with no tactical intent are labeled NONE. Full definitions and distributions are in Appendix A. The benchmark uses tactic-level labels throughout; technique-level annotation is released for future fine-grained analysis.

Of these six, *Information Gathering* (T1004) and *Financial Exploitation* (T1005) are the *terminal harms* ETF forecasts; the rest are the process driving the call toward them.

4 Dataset Construction

Base data and tactic annotation. Our corpus is based on the Chinese synthetic vishing conversations of TeleAntiFraud-28k (Ma et al., 2025c), generated by an LLM with synthesized TTS audio. Fraud types include customer-service impersonation, bank, investment, phishing, lottery, and kidnap; the fraud type of each call follows the original TeleAntiFraud label. We assigned the tactic labels of Section 3 to scammer utterances; T1001 (Initial Contact) does not apply to mid-call recordings, so annotation covers the remaining five tactics (T1002–T1006) and their 21 techniques, plus NONE. Two LLMs (Claude and GPT-4o) independently labeled every utterance after viewing the full call context (inter-model Cohen's $\kappa = 0.94$). Two human annotators relabeled a stratified $\sim$10% sample of the test split (inter-human $\kappa = 0.69$; joint Fleiss's $\kappa = 0.66$). Details in Appendix A.

Standard split. For ETF, we sampled the 5,645 vishing calls (out of TeleAntiFraud-28k) for which both a tactic label and the original fraud-type label are available, and stratified the split by fraud type (Table 1; an average of 6.3 scammer turns per call).

Tactic distribution. Tactic counts over the standard split span an order of magnitude (Table 2).

Unit	Total	Train	Test
Calls	5,645	4,515	1,130
Scammer turns	35,340	28,222	7,118
Victim turns	28,631	22,913	5,718
Total turns	63,971	51,135	12,836

Table 1: Vishing-Tactics-Bench standard split totals (per-fraud-type breakdown in Appendix B, Table 9).

Trust Building, Psychological Pressure, and Information Gathering together account for 82.2% of scammer turns, while the terminal-stage tactics *Financial Exploitation* and *Evasion* are each near 1.4%, and NONE (no tactical intent) at 15.0%. This $\sim$20$\times$ frequency gap between *Information Gathering* (31.5%) and *Financial Exploitation* (1.4%) is the turn-level imprint of the difficulty contrast ETF must handle within one task.

Tactic	Count	%
T1002 Trust Building	8,384	23.7
T1003 Psychological Pressure	9,554	27.0
T1004 Information Gathering	11,133	31.5
T1005 Financial Exploitation	481	1.4
T1006 Evasion	496	1.4
NONE	5,292	15.0
Total	**35,340**	**100.0**

Table 2: Tactic-level distribution over the Vishing-Tactics-Bench standard split.

Distribution of the two terminal harms. The fraud type determines the harm pathway: kidnap fraud goes almost directly to *Financial Exploitation* (28.7%) with little *Information Gathering* (5.7%), whereas the other types have high *Information Gathering* reach (69.0–98.6%); phishing almost entirely terminates at *Information Gathering* (98.6%). The two harms are not in a linear relation: about 36% of calls reaching *Financial Exploitation* do so without passing through *Information Gathering* (full corpus, n=365 reaching calls). Overall, *Information Gathering* is common (85.8%) while *Financial Exploitation* is rare (6.4%). Per-type reach rates are in Appendix B, Table 10.

5 Task: Exploitation Trajectory Forecasting

Given the TTP trajectory up to turn t, ETF predicts the hazard of reaching each terminal harm within the next k turns. ETF conditions on a call already identified as vishing; separating vishing from benign calls is an upstream detection task.

Definition. We view a vishing call as a sequence of scammer utterances. The t-th scammer utterance carries a tactic a_t, and the full sequence $(a_1, \ldots, a_T)$ is the call's TTP trajectory. ETF predicts the occurrence of the two terminal harms $g \in \{\text{T1004}, \text{T1005}\}$. The first appearance of g marks the onset of the corresponding victim-facing harm: information disclosure for T1004, financial loss for T1005. Let τ_g be the scammer turn at which tactic g first appears ($\tau_g = \infty$ if it never appears). At each scammer turn t, conditioned on g not yet reached, the model predicts the *hazard*

$$h_g(t,k) = \Pr[\,\tau_g \le t+k \mid X_t,\ \tau_g > t\,] \quad (1)$$

of reaching g within the next k turns, for each g and $k \in \{1,3,5\}$. X_t is the observation up to turn t. A single model outputs both harms' hazards.

Input. X_t has two configurations: (i) the dialogue utterances up to turn t, and (ii) the utterances plus the TTP trajectory $(a_1, \ldots, a_t)$. Comparing (ii) against (i) measures the contribution of tactical information.

Evaluation. ETF uses three metrics, computed separately for the two terminal harms.

- **AP@k** — average precision with ground truth "reaches g within the next k turns" and the hazard $h_g(t,k)$ as score. It is suited to the rare positives of *Financial Exploitation*, for which accuracy is uninformative. Higher is better.
- **C-index** — the concordance index (Harrell et al., 1982; Zhao et al., 2021): the fraction of (reaching, non-reaching) call pairs in which the model assigns the higher hazard to the reaching call. 0.5 is random, higher is better.
- **Divergence error** — a task-specific metric: the distance, in turns, between the turn at which the predicted hazard curve rises most sharply and the turn at which g first appears.

  Let $\hat{h}_g(t,k)$ be the predicted hazard at turn t for horizon k. The predicted and ground-truth branch turns are

$$\hat{\tau}_g = 1 + \arg\max_i \big[\hat{h}_g(t_{i+1},k) - \hat{h}_g(t_i,k)\big], \quad (2)$$

$$\tau_g = \min\{\, i : a_i = g \,\} \quad (3)$$

  and the divergence error averages their absolute distance over the calls that reach g:

$$\text{Div}_g \;=\; \mathbb{E}\big[\,|\hat{\tau}_g - \tau_g| \;\big|\; g \text{ reached}\big]. \quad (4)$$

We use $k = 3$. Calls that never reach g have no ground-truth branch and are excluded from the average. Its range is $[0, \text{call length})$ (mean 6.3 turns); lower is better.

6 Baselines

We set up baselines of differing character, spanning statistical, discriminative, and generative modeling families.

- **Markov heuristic.** Estimates the tactic transition matrix $P(T_{t+1} \mid T_t)$ from training trajectories and iterates it to compute reach probabilities. It uses no dialogue text and shows how far ETF is predictable from the transition structure alone.
- **Encoder-based learned model.** A learned hazard head over both terminal harms, with three input configurations: **text only** (a Chinese pretrained encoder, `chinese-roberta-wwm-ext` (Cui et al., 2021)), **TTP only** (a small GRU over the tactic-label trajectory; LSTM and Transformer variants in Appendix D), and **text+TTP** (both encoders, late-fused by concatenation).
- **LLM zero-shot.** Prompts gpt-4.1-mini to predict reach directly, under prompts with and without the TTP trajectory.
- **LLM fine-tuning.** Supervised-fine-tunes two LLMs (gpt-4.1-mini via the OpenAI tuning API; Qwen2.5-7B (Qwen et al., 2025) via LoRA, r=16) on the train split, with and without the TTP.

The TTP-included/excluded comparison is the central analysis axis, directly testing whether tactical information provides a signal beyond text.

7 ETF Results

The general pattern in Table 3 is consistent across both terminal harms: learned baselines (encoder, supervised-fine-tuned LLMs) substantially outperform non-learned baselines (the Markov heuristic and zero-shot gpt-4.1-mini), with the gap most pronounced on *Information Gathering*. Each TTP-using model is reported twice—*gold* (oracle TTP) and *pred* (cascaded classifier)—so the deployment-relevant noise can be read off directly; text-only rows are unaffected by the TTP source. The gold-versus-predicted comparison quantifies how upstream TTP recognition errors propagate into downstream harm forecasting.

AP@k. Learning closes the zero-shot–learned gap by roughly an order of magnitude. *Information Gathering* is recovered quickly (ZS 0.27–0.45, learned 0.87–0.97). *Financial Exploitation* stays lower in absolute value (window base rate 0.02–0.04) but rises by the same relative factor (ZS 0.07–0.39, learned 0.51–0.67); FE AP must be read against the base-rate row (e.g., Markov FE 0.39 $\approx$ 20$\times$ base rate; Markov IG 0.40 is only marginally above chance).

C-index. Text-only zero-shot gpt-4.1-mini is 0.39 for *Information Gathering* (below random), and adding the TTP trajectory does not repair this; the encoder and gpt-4.1-mini SFT reach 0.96–0.98. *Financial Exploitation* C-index improves modestly (Markov 0.67 to encoder 0.92), since Markov's transitions already rank usefully.

Divergence error. For *Financial Exploitation*, several transition- or trajectory-aware rows locate the branch very close to the true turn, including Markov, Qwen2.5-7B SFT (text+TTP), and gpt-4.1-mini SFT (text+TTP), while text-only and TTP-only neural variants are less precise. For *Information Gathering*, most models fall in the 1.2–2.1 range; gpt-4.1-mini ZS (text+TTP) is the lone outlier near 0.6. T1004 occurs gradually across the call, so there is no sharp branch for any model to locate.

TTP carries non-redundant signal, but its gain depends on integration. Encoder (TTP) alone—without dialogue text—attains *Information Gathering* AP@3 0.71 and C-index 0.84, surpassing every text-using zero-shot LLM, and adding the trajectory raises gpt-4.1-mini's *Financial Exploitation* AP@1 fivefold (0.07 to 0.35). With oracle TTP, the largest text+TTP gains on *Financial Exploitation* AP@1 go to the inline-fusion LLMs (gpt-4.1-mini SFT $+0.16$, Qwen2.5-7B SFT $+0.23$), while the encoder gains marginally ($+0.01$) because text alone already saturates *Information Gathering* (text-only C-index 0.98). The two integration methods differ: the encoder concatenates a small TTP branch with a much larger text representation (late fusion), while LLM SFT reads TTP tokens inline within the same attention context. Zero-shot GPT-4o shows

Model	TTP	*Information Gathering*			*Financial Exploitation*		
		AP@1/3/5↑	C-idx↑	Div. err.↓	AP@1/3/5↑	C-idx↑	Div. err.↓
Base rate	—	0.27 / 0.45 / 0.48	0.50	n/a	0.02 / 0.03 / 0.04	0.50	n/a
Markov heuristic	gold	0.40 / 0.49 / 0.52	0.55	1.45	0.39 / 0.24 / 0.22	0.67	0.07
	pred	0.39 / 0.51 / 0.53	0.58	1.51	0.19 / 0.15 / 0.14	0.65	0.11
Encoder (text)	—	0.89 / 0.95 / **0.97**	**0.98**	1.43	0.58 / 0.56 / 0.56	**0.92**	1.15
Encoder (TTP)	gold	0.32 / 0.71 / 0.79	0.84	2.10	0.11 / 0.10 / 0.10	0.68	0.89
	pred	0.32 / 0.69 / 0.77	0.83	2.12	0.08 / 0.09 / 0.10	0.68	1.45
Encoder (text+TTP)	gold	0.89 / 0.95 / **0.97**	**0.98**	1.36	0.59 / 0.57 / 0.56	0.91	1.15
	pred	0.89 / 0.95 / **0.97**	**0.98**	1.38	0.60 / **0.58** / **0.57**	**0.92**	1.25
Qwen2.5-7B SFT (text)	—	0.54 / 0.73 / 0.80	0.86	1.51	0.09 / 0.08 / 0.08	0.54	1.96
Qwen2.5-7B SFT (text+TTP)	gold	0.67 / 0.81 / 0.88	0.93	1.48	0.32 / 0.20 / 0.18	0.60	0.18
	pred	0.60 / 0.79 / 0.85	0.91	1.45	0.16 / 0.12 / 0.11	0.60	0.25
gpt-4o ZS (text)	—	0.28 / 0.37 / 0.39	0.31	0.76	0.11 / 0.11 / 0.10	0.64	0.96
gpt-4o ZS (text+TTP)	gold	0.34 / 0.41 / 0.42	0.40	**0.53**	0.26 / 0.18 / 0.15	0.65	0.52
	pred	0.33 / 0.40 / 0.42	0.40	0.61	0.17 / 0.14 / 0.13	0.65	0.38
gpt-4.1-mini ZS (text)	—	0.27 / 0.39 / 0.41	0.39	1.22	0.07 / 0.09 / 0.09	0.65	1.49
gpt-4.1-mini ZS (text+TTP)	gold	0.36 / 0.43 / 0.45	0.45	0.58	0.35 / 0.22 / 0.18	0.65	0.34
	pred	0.35 / 0.43 / 0.45	0.46	0.59	0.22 / 0.17 / 0.14	0.66	0.45
gpt-4.1-mini SFT (text)	—	0.87 / 0.91 / 0.93	0.96	1.24	0.51 / 0.46 / 0.43	0.83	0.76
gpt-4.1-mini SFT (text+TTP)	gold	**0.93** / **0.96** / **0.97**	**0.98**	1.28	**0.67** / 0.55 / 0.53	0.86	**0.00**
	pred	0.85 / 0.91 / 0.92	0.95	1.25	0.30 / 0.31 / 0.33	0.80	0.14

Table 3: ETF results on the standard test split. Arrows (↑/↓): better direction. AP@k, C-index $\in [0, 1]$; divergence error in turns. The **TTP** column denotes the TTP source: *gold* (LLM-annotated oracle) vs. *pred* (cascaded RoBERTa turn-level classifier; macro-F1 0.75). *Text*-only rows do not use TTP (—). Encoder (3 variants) and Qwen2.5-7B SFT (2 variants) are mean across 3 random seeds; other rows are single runs (3-seed std and 95% bootstrap CI in Appendix Tables 12–13). **Bold**: best per column.

the same trend: the trajectory raises *Financial Exploitation* AP@1 from 0.11 to 0.26 with oracle TTP and 0.17 with predicted TTP, still well below the encoder's 0.60.

Late fusion is robust to predicted-TTP noise; inline fusion is not. Substituting oracle TTP with the cascaded classifier (macro-F1 0.75) preserves the encoder's *Financial Exploitation* AP@1 (gold 0.59, pred 0.60) but cuts it to roughly half for the inline-fusion LLMs (gpt-4.1-mini SFT 0.67 → 0.30; Qwen2.5-7B SFT 0.32 → 0.16). The contrast is asymmetric across the two harms: gpt-4.1-mini SFT loses only 0.08 AP@1 on *Information Gathering* (0.93 → 0.85) but 0.37 on *Financial Exploitation* (0.67 → 0.30), tracking the classifier's own per-class F1 of 0.90 on T1004 vs. 0.58 on T1005. Rare-tactic classification noise thus propagates into rare-harm prediction when TTP is read inline. In the deployment-relevant predicted-TTP setting the encoder (text+TTP) is the strongest model on *Financial Exploitation* AP@1 (0.60), surpassing gpt-4.1-mini SFT (0.30) by a factor of two.

The two terminal harms differ contrastively. *Information Gathering* is common (85.8% reach) with relatively gradual unfolding, while *Financial Exploitation* is rare (6.4%) but appears abruptly and is timed precisely. The two harms differ along opposite axes (frequency vs. timing), and ETF reports both. The TTP gain is also *type-conditional*: largest on kidnap (+0.45 *IG* AP@1, gpt-4.1-mini SFT) and investment (+0.31 *FE* AP@1), but slightly negative on bank/lottery; per-type analysis appears in Appendix B.4. Per-type results on kidnap (n=46) and lottery (n=60) use small test sets and should be read as suggestive.

8 Anatomy of the Two Terminal Harms

Table 3 pools across calls and turns; here we condition on each call's *terminal outcome* to keep opposite behaviours on outcome groups from cancelling.

Call outcome typology. The two harms are neither strictly sequential nor independent (typology in Appendix Table 11): ∼5% of *Information Gathering*-reaching calls reach *Financial Exploitation*, while ∼40% of *Financial Exploitation*-reaching calls bypass *Information Gathering* (n=67 in test). Neither-outcome calls are short (median 2 turns) and stall in trust building or pressure.

Model	TTP	*Financial Exploitation* hazard@3 by call outcome				Both calls: *FE* hazard@3 vs. the *IG* branch		
		Both ↑	*FE*-only ↑	*IG*-only ↓	Neither ↓	Before	After	Δ ↑
Markov heuristic	gold	0.086	0.097	0.048	0.048	0.050	0.047	−0.003
	pred	0.091	0.118	0.050	0.058	0.050	0.050	0.000
Encoder (text)	—	0.253	0.330	0.011	0.063	0.238	0.276	+0.039
Encoder (TTP)	gold	0.052	0.100	0.044	0.066	0.084	0.024	−0.059
	pred	0.056	0.117	0.044	0.073	0.082	0.029	−0.053
Encoder (text+TTP)	gold	0.282	0.326	0.014	0.059	0.249	0.365	+0.116
	pred	0.262	0.312	0.011	0.055	0.229	0.318	+0.089
Qwen2.5-7B SFT (text)	—	0.070	0.142	0.001	0.018	0.000	0.009	+0.009
Qwen2.5-7B SFT (text+TTP)	gold	0.161	0.186	**0.000**	**0.005**	0.000	0.000	0.000
	pred	0.169	0.261	0.008	0.031	0.000	0.009	+0.009
gpt-4o ZS (text)	—	0.423	0.525	0.299	0.290	0.237	0.383	+0.146
gpt-4o ZS (text+TTP)	gold	0.492	0.551	0.330	0.249	0.249	0.458	+0.208
	pred	**0.502**	**0.591**	0.337	0.302	0.255	0.463	+0.208
gpt-4.1-mini ZS (text)	—	0.274	0.417	0.166	0.184	0.146	0.202	+0.056
gpt-4.1-mini ZS (text+TTP)	gold	0.410	0.455	0.241	0.167	0.139	0.370	**+0.231**
	pred	0.401	0.502	0.235	0.231	0.161	0.321	+0.159
gpt-4.1-mini SFT (text)	—	0.118	0.285	0.011	0.068	0.048	0.043	−0.006
gpt-4.1-mini SFT (text+TTP)	gold	0.149	0.279	0.008	0.047	0.031	0.021	−0.010
	pred	0.136	0.260	0.015	0.074	0.031	0.032	+0.002

Table 4: Anatomy of *Financial Exploitation* hazard; **TTP** source as in Table 3. **(Left)** mean predicted *Financial Exploitation* hazard@3 by call outcome. **(Right)** for Both calls, mean *Financial Exploitation* hazard@3 before vs. after the *Information Gathering* branch, restricted to turns prior to *Financial Exploitation*; a model that updates dynamically yields $\Delta > 0$.

Outcome-stratified hazard. The left half of Table 4 reports mean predicted *Financial Exploitation* hazard@3 by call outcome. A trajectory-tracking model should rank Both / *Financial Exploitation*-only above *Information Gathering*-only / Neither. With text alone, gpt-4.1-mini ZS inverts this ordering (Neither > *Information Gathering*-only); adding the trajectory corrects it under both oracle and predicted TTP. The encoder (text+TTP) shows the cleanest stratification (Both 0.28, *Financial Exploitation*-only 0.33, *Information Gathering*-only 0.01, Neither 0.06 with gold; nearly identical with pred), so late-fusion robustness extends to outcome-conditioned ranking. Inline-fusion SFT models compress predictions toward zero on non-reaching outcomes, making low-hazard targets trivial without separating Both / *Financial Exploitation*-only from the rest.

Post-branch dynamic update. The right half of Table 4 restricts to Both calls and asks whether a model raises *Financial Exploitation* hazard *after* the *Information Gathering* branch is passed. The Markov heuristic is flat ($\Delta = -0.003$ gold, 0.000 pred): its transition matrix has no notion that "*Information Gathering* has occurred" changes the outlook—the TTP trajectory is what lets a model update dynamically. gpt-4.1-mini ZS with the trajectory shows the largest oracle-TTP post-branch lift ($\Delta = +0.231$ gold, $+0.159$ pred), zero-shot GPT-4o holds $\Delta = +0.208$ in both settings—the largest lift under predicted TTP—and the encoder (text+TTP) keeps Δ positive in both settings ($+0.116$ gold, $+0.089$ pred). The late-fusion advantage therefore carries through to dynamic updating, not only to static discrimination. Inline-fusion SFT models hold $\Delta \approx 0$ in both settings: their strong static accuracy on *Financial Exploitation* comes from a confident static separation of Both calls from the rest, not from a dynamic re-evaluation at the *Information Gathering* branch. This dynamic-updating capability is the *interpretability* contribution of TTP, distinct from raw accuracy.

Lead-time breakdown at a tight false-alarm budget. An early-warning alarm must (i) fire with usable *lead* before the terminal harm and (ii) keep its false-alarm rate (FAR) low on calls that never reach this harm. We cast each model as an alarm firing at the first turn whose hazard exceeds a threshold θ, fix FAR $\leq 10\%$ on the non-reaching pool, and bin each reaching call's lead $\ell = \tau_g - t_{\text{alarm}}$ into ≥ 3 *early* / 1–2 *early* / *on-time* (ℓ=0) / *post-hoc* (ℓ<0)

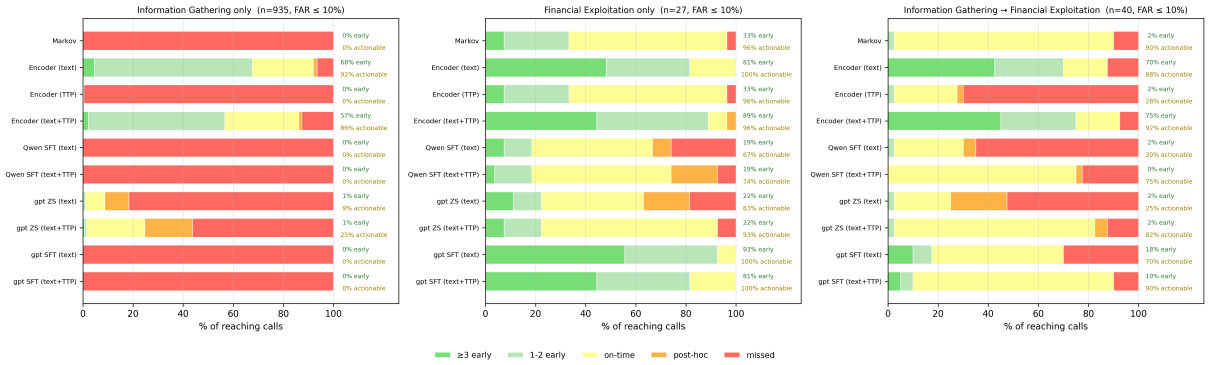


Figure 2: Lead-time breakdown at FAR $\leq 10\%$ under *predicted* TTP, stratified by the trajectory each call reaches. For each reaching call the lead $\ell = \tau_g - t_{\text{alarm}}$ is binned into five outcomes; bars sum to 100% within a (model, stratum) cell. *% early* counts $\ell \geq 1$, *% actionable* counts $\ell \geq 0$. Stratum sizes: IG only n=935, FE only n=27, IG→FE n=40; numerical breakdown in Appendix Table 15. Oracle-TTP version in Appendix C (Figure 3).

/ *missed*. We report *% early* ($\ell \geq 1$, actionable warning at least one turn before harm) and *% actionable* ($\ell \geq 0$). Because every test call is a fraud call (no benign controls), we stratify by trajectory: *Information Gathering only*, *Financial Exploitation only*, and *Information Gathering → Financial Exploitation*. Figure 2 reports the breakdown.

Three findings emerge under predicted TTP. (i) The TTP signal pays off on the *Information Gathering → Financial Exploitation* stratum (n=40 in test): adding the predicted trajectory lifts the encoder by 5 pp in *% early* ($70\% \rightarrow 75\%$), because the $T1004$ precursor is the cue the model leans on to anticipate $T1005$. (ii) Direct extortion belongs to the generation-style LLM: fine-tuned `gpt-4.1-mini` (text) flags 93% of *Financial Exploitation only* calls with $\ell \geq 1$ (FE-only stratum, n=27, so this percentage is high-variance), since the abrupt money demand carries surface signal without a precursor; the same model collapses to 18% early once the call has an *Information Gathering* preamble. Adding TTP to this inline-fusion model under noisy predictions degrades both—*FE-only* drops to 82% and the trajectory stratum to 10%—reinforcing the inline-fragility pattern of §7. (iii) The Markov baseline fires at or before the harm turn on 96% of *Financial Exploitation only* calls under predicted TTP (33% early + 63% on-time)—a primarily concurrent detector. On *Information Gathering only* (86% base rate), only the encoder variants remain actionable under FAR $\leq 10\%$. Appendix E walks through one successful early warning and one missed abrupt demand.

9 Discussion

From detection to projection. We reframe vishing defense as *harm projection* rather than post-hoc detection. In situation-awareness terms, the question is whether the model's *tactical comprehension* (the TTP trajectory) is a useful state for *projecting* which concrete harm a call approaches, so that intervention is timed and targeted.

The TTP trajectory as an interpretable tactical state. The TTP trajectory functions as an *interpretable tactical state* separate from accuracy gain. On the common, gradual harm, text-only models already saturate and the trajectory adds little; on the rare, abrupt harm, the trajectory is the dominant signal for raising hazard estimates and locating the harm branch in time. It matters where text alone falls short—telling the model *which* harm a call is heading toward and *when* it will arrive.

Privacy-conscious deployment. A TTP-only encoder makes a privacy-conscious pipeline feasible: an on-device tactic classifier processes the transcript locally, and only the abstract tactic sequence is sent to the forecasting service. This aligns with data-minimization principles at a real accuracy cost relative to text+TTP; we do not evaluate the deployment empirically but note it as an architectural option text-conditioned baselines cannot offer.

Transition structure should be learned. LLMs given the TTP trajectory still did not match Markov on *Financial Exploitation* branch timing. Tactic-to-tactic transition is a corpus-specific statistic that data estimates better than prior knowledge approximates; a strong ETF model should combine text semantics with learned transitions.

10 Conclusion

We presented Vishing-Tactics-Bench, an utterance-level benchmark grounded in Endsley's situation-awareness framework (perception, comprehension, projection). We built a corpus with turn-level TTP labels, defined the Exploitation Trajectory Forecasting task, and analyzed it with heuristic and LLM baselines; the tactical trajectory is a key predictive signal, best exploited when learned from data.

The harm a call inflicts occurs only at its terminal tactics, and everything before them is a window in which intervention is still possible. *Which* terminal harm is forecasted determines the appropriate response (credential-entry blocks and app-installation warnings for *Information Gathering*, payment blocks and bank handoff for *Financial Exploitation*), so targeted defense depends on harm specificity rather than raw detection accuracy. The gap our baselines leave open is the remaining headroom for practical early-warning systems.

Future work. A next step is to validate ETF in real-time deployment by streaming audio through an ASR pipeline and measuring lead time and FAR at a deployed threshold. A further direction is *victim-conditioned* ETF: victim responses (compliance, resistance, hesitation) shape the scammer's next tactic, so the trajectory is a function of dialogue dynamics the current call-as-sequence formulation does not model. Extending the taxonomy and corpus to other languages and to authentic-call data is a second direction; cross-lingual transfer of the trajectory structure is an open question. Alternative hazard parameterizations (joint hazards, longer horizons, competing-risks survival architectures) remain future work.

Limitations

Synthetic data. Real vishing calls are private conversations and crime-related material, and their collection and release are strictly restricted by legal and ethical constraints in essentially every jurisdiction. Synthetic data is therefore a necessary stepping stone for community progress on this domain: it allows shared, reproducible benchmarks to exist at all, and the construction of synthetic vishing data has itself become an active research topic. TeleAntiFraud-28k, on which our corpus is based, was likewise constructed and released this way. Specifically, TeleAntiFraud-28k publicly releases only its synthetic subsets; the real-recording subset is not publicly available, which further constrains the choice of base data. Building on a synthetic base is thus a deliberate research-design choice, not a substitute we settled for; synthetic conversations and audio may not fully reflect real calls, and validation on real recordings remains future work contingent on the resolution of the underlying legal access barriers.

Single language. The corpus is limited to Chinese vishing conversations; generalization to other languages and cultures is not verified.

Difficulty gap between the two terminal harms. The difference in prediction difficulty between *Information Gathering* and *Financial Exploitation* may stem in part from the difference in their reach frequency, that is, the amount of positive examples. *Financial Exploitation* (6.4%) has far fewer positives than *Information Gathering* (85.8%), so it is hard to fully separate whether its low AP reflects intrinsic difficulty or sample scarcity.

No benign-call controls. All calls in the benchmark are vishing calls, so the reported false-alarm rates count alarms on vishing calls that do not reach the target harm; false-alarm behaviour on legitimate calls is not measured, and adding benign-call controls is left for future work.

Ethics Statement

Data and privacy. All conversations are synthetic (TeleAntiFraud-28k) and contain no real individuals' personally identifying information. Our artifact will be released as a gated Hugging Face dataset containing only the TTP annotations and join keys linked to the Apache-2.0 TeleAntiFraud corpus, so no PII is introduced by our release.

Potential risks. Tactic-level annotations of vishing calls could in principle be misused to craft more deceptive scams. Two factors mitigate this risk: (i) the annotated corpus is itself synthetic (no real victim utterances) and already publicly released under Apache-2.0; (ii) the abstract tactic labels (Trust Building, Pressure, etc.) are common-knowledge descriptions of social engineering and are not new attack capabilities. We release the resource to advance defensive harm-projection research, which is the primary intended use.

AI disclosure. Generative AI tools were used to assist with this work: Claude (Anthropic) for

code development and manuscript language editing, and Gemini (Google) for figure preparation. All AI-generated code was reviewed and tested, and AI-assisted figures verified, by the authors before use. AI-assisted manuscript edits were limited to language polishing; all scientific content was written and verified by the authors.

Acknowledgments

This work was supported by the Institute of Information and Communications Technology Planning and Evaluation (IITP) grants funded by the Korea government (MSIT) (No. RS-2025-02215393, Development of Detection and Prediction Technology for New and Unknown Voice Phishing; No. RS-2019-II190004, Development of Semi-supervised Learning Language Intelligence Technology and Korean Tutoring Service for Foreigners).

References

Lin Ai, Tharindu Sandaruwan Kumarage, Amrita Bhattacharjee, Zizhou Liu, Zheng Hui, Michael S. Davinroy, James Cook, Laura Cassani, Kirill Trapeznikov, Matthias Kirchner, Arslan Basharat, Anthony Hoogs, Joshua Garland, Huan Liu, and Julia Hirschberg. 2024. Defending against social engineering attacks in the age of LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12880–12902, Miami, Florida, USA. Association for Computational Linguistics.

Enas Altarawneh, Ameeta Agrawal, Michael Jenkin, and Manos Papagelis. 2023. Conversation derailment forecasting with graph convolutional networks. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 160–169, Toronto, Canada. Association for Computational Linguistics.

Hyowon Cho and Minjoon Seo. 2025. Towards reliable and practical phishing detection. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 210–225, Albuquerque, New Mexico. Association for Computational Linguistics.

Robert Cialdini. 2006. The psychology of persuasion revised edition.

David R. Cox. 1972. Regression models and life-tables. *Journal of the Royal Statistical Society, Series B (Methodological)*, 34(2):187–202.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for Chinese BERT. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

Mica R. Endsley. 1995. Toward a theory of situation awareness in dynamic systems. *Human Factors*, 37(1):32–64.

Frank E. Harrell, Robert M. Califf, David B. Pryor, Kerry L. Lee, and Robert A. Rosati. 1982. Evaluating the yield of medical tests. *JAMA*, 247(18):2543–2546.

Eric M. Hutchins, Michael J. Cloppert, and Rohan M. Amin. 2011. Intelligence-driven computer network defense informed by analysis of adversary campaigns and intrusion kill chains. In *Leading Issues in Information Warfare & Security Research*, volume 1, pages 80–106. Academic Publishing International. Lockheed Martin white paper, presented at ICIW 2011.

Edward L. Kaplan and Paul Meier. 1958. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481.

Haoyu Ma, Qinliang Su, Minhua Huang, and Wu Kai. 2025a. Detecting continuously evolving scam calls under limited annotation: A LLM-augmented expert rule framework. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5047–5068. Association for Computational Linguistics.

Shang Ma, Tianyi Ma, Jiahao Liu, Wei Song, Zhenkai Liang, Xusheng Xiao, and Yanfang Ye. 2025b. PsyScam: A benchmark for psychological techniques in real-world scams. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 12623–12637. Association for Computational Linguistics.

Zhiming Ma, Peidong Wang, Minhua Huang, Jinpeng Wang, Kai Wu, Xiangzhao Lv, Yachun Pang, Yin Yang, Wenjie Tang, and Yuchen Kang. 2025c. TeleAntiFraud-28k: An audio-text slow-thinking dataset for telecom fraud detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 5853–5862, Dublin, Ireland. Association for Computing Machinery.

Rosana Montañez Rodriguez and Shouhuai Xu. 2022. Cyber social engineering kill chain. In *Science of Cyber Security: 4th International Conference (SciSec 2022)*, volume 13580 of *Lecture Notes in Computer Science*, pages 487–504, Cham. Springer.

Milandu Keith Moussavou Boussougou and Dong-Joo Park. 2023. Attention-based 1D CNN-BiLSTM hybrid model enhanced with FastText word embedding for Korean voice phishing detection. *Mathematics*, 11(14):3217.

Roberts Oļeiņiks and Darja Solodovņikova. 2025. Real-time phone fraud detection and prevention based on artificial intelligence tools. *Baltic Journal of Modern Computing*, 13(1):252–289.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Zitong Shen, Sineng Yan, Youqian Zhang, Xiapu Luo, Grace Ngai, and Eugene Yujun Fu. 2025. "it warned me just at the right moment": Exploring LLM-based real-time detection of phone scams. In *Extended Abstracts of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–7. ACM.

Anthony Sicilia, Hyunwoo Kim, Khyathi Chandu, Malihe Alikhani, and Jack Hessel. 2024. Deal, or no deal (or who knows)? forecasting uncertainty in conversations using large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11700–11726, Bangkok, Thailand. Association for Computational Linguistics.

Blake E. Strom, Andy Applebaum, Doug P. Miller, Kathryn C. Nickels, Adam G. Pennington, and Cody B. Thomas. 2018. MITRE ATT&CK: Design and philosophy. Technical Report MP180360, The MITRE Corporation, McLean, VA.

Nikolaos Tsinganos, Panagiotis Fouliras, and Ioannis Mavridis. 2023. Leveraging dialogue state tracking for zero-shot chat-based social engineering attack recognition. *Applied Sciences*, 13(8):5110.

Huahong Tu, Adam Doupé, Ziming Zhao, and Gail-Joon Ahn. 2019. Users really do answer telephone scams. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 1327–1340, Santa Clara, CA. USENIX Association.

Peidong Wang, Zhiming Ma, Xin Dai, Yongkang Liu, Shi Feng, Xiaocui Yang, Wenxing Hu, Zhihao Wang, Mingjun Pan, Li Yuan, and 1 others. 2026. SAFE-QAQ: End-to-end slow-thinking audio-text fraud detection via reinforcement learning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26138–26157. Association for Computational Linguistics.

Yuxin Wang, Ivory Yang, Saeed Hassanpour, and Soroush Vosoughi. 2024. MentalManip: A dataset for fine-grained analysis of mental manipulation in conversations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3747–3764, Bangkok, Thailand. Association for Computational Linguistics.

Monica T. Whitty. 2013. The scammers persuasive techniques model: Development of a stage model to explain the online dating romance scam. *British Journal of Criminology*, 53(4):665–684.

Ian Wood, Michal Kepkowski, Leron Zinatullin, Travis Darnley, and Mohamed Ali Kaafar. 2023. An analysis of scam baiting calls: Identifying and extracting scam stages and scripts. *arXiv preprint arXiv:2307.01965*.

Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. 2018. Conversations gone awry: Detecting early signs of conversational failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1350–1361, Melbourne, Australia. Association for Computational Linguistics.

Yun Zhao, Qinghang Hong, Xinlu Zhang, Yu Deng, Yuqing Wang, and Linda Petzold. 2021. BERTSurv: BERT-based survival models for predicting outcomes of trauma patients. *Preprint*, arXiv:2103.10928.

A Vishing-Tactics Taxonomy and Annotation

A.1 Full Taxonomy: 6 Tactics, 24 Techniques

The Vishing-Tactics taxonomy defines six tactics divided into 24 techniques (Table 5); counts are over the Vishing-Tactics-Bench standard split. T1001 (Initial Contact) is part of the taxonomy but does not apply to this corpus, which captures only mid-call segments—annotation covers the remaining five tactics (T1002–T1006) and NONE.

Six of the 21 annotated techniques have fewer than 35 instances and four are in single digits, so the benchmark uses tactic-level labels; technique-level annotation is released for fine-grained analysis.

A.2 Annotation Procedure

Tactic labels were produced by cross-annotation with two large language models (Claude and GPT-4o). Each model was called once per call, received the full call context, and independently labeled every scammer utterance with a tactic, technique, confidence, and reasoning (temperature 0.1). Inter-model agreement was Cohen's $\kappa = 0.94$. The annotation prompt has the following skeleton (the taxonomy description expands to the five annotated tactics and 21 techniques):

Annotation prompt

```
You are a professional analyst of voice phishing conversations. Analyze the following Chinese phone conversation and assign the most appropriate TTP label to the scammer's utterances only. [taxonomy] [conversation] Rules: judge from full context; label a turn with no clear fraud pattern as NONE; pick one technique per turn; give a confidence and a brief reasoning. Output strict JSON.
```

Technique	Name	Count
T1001 Initial Contact		
T1001.001	Phone call	—
T1001.002	SMS / messaging	—
T1001.003	Recorded voice / ARS	—
T1002 Trust Building		
T1002.001	Impersonating institutions	5,779
T1002.002	Impersonating acquaintances	205
T1002.003	Building rapport	1,558
T1002.004	Impersonating experts	820
T1003 Psychological Pressure		
T1003.001	Creating urgency	3,332
T1003.002	Fear / threat	1,202
T1003.003	Benefit lure	4,512
T1003.004	Inducing guilt	16
T1003.005	Information control	462
T1004 Information Gathering		
T1004.001	PII request	3,455
T1004.002	Financial info request	1,272
T1004.003	Malicious app installation	3,651
T1004.004	Fake website lure	2,603
T1005 Financial Exploitation		
T1005.001	Account transfer	447
T1005.002	Cash withdrawal	7
T1005.003	Voucher / crypto purchase	2
T1005.004	Loan exploitation	10
T1005.005	Account opening abuse	3
T1006 Evasion		
T1006.001	Conversation deletion	3
T1006.002	Disappearing act	402
T1006.003	Identity theft / burner phone	36
NONE	(no tactical intent)	5,292

Table 5: The Vishing-Tactics taxonomy and technique-level counts over the Vishing-Tactics-Bench standard split (T1001 does not apply to mid-call recordings). All reported experiments use tactic-level labels; technique-level annotations are released for future fine-grained studies and are not used for training or evaluation. 271 turns (0.8%) carry a tactic label without a technique sub-label and appear only in the tactic-level totals of Table 2.

Human validation. To validate the LLM-based annotation, two human annotators independently relabeled a stratified random sample of 118 calls (777 scammer turns), about 10% of the test split. The sample includes all *Financial Exploitation*-reach calls (*Information Gathering*→*Financial Exploitation* and *Financial Exploitation*-only; 67 calls) together with a stratified draw of *Information Gathering*-only and Neither calls (51 calls), additionally oversampling *Evasion*-containing calls and lightly boosting under-represented phishing, so that all six tactic classes including the rare *Financial Exploitation* and *Evasion* are well covered. The first call of the sample was used as a calibration round and is excluded from the agreement computation. Annotators viewed the full call context (as the LLM annotators did), since label quality is an offline ground-truth concern, distinct from the runtime classifier of Appendix C (which is restricted to prefix-only input for streaming compatibility).

Inter-annotator agreement. Table 6 reports pairwise Cohen's κ across the four raters (HA, HB, M1, M2) and the 4-rater Fleiss's κ. Human–human agreement is substantial ($\kappa = 0.69$); LLM–LLM agreement is near-perfect on this subset ($\kappa = 0.97$, marginally higher than the corpus-level model–model agreement of $\kappa = 0.94$); 4-rater Fleiss's $\kappa = 0.66$. Combined, this supports use of the LLM-annotated labels as a stand-in for human judgment at corpus scale. Per-class agreement is strongest on *Information Gathering* (HA–HB $\kappa_{\text{class}} = 0.79$) and *Financial Exploitation* ($\kappa_{\text{class}} = 0.76$); residual disagreement concentrates on the *Trust Building* / *Psychological Pressure* boundary, an inherently fuzzy distinction in social-engineering speech. Table 7 reports per-class agreement for all rater pairs, and Table 8 the pooled human–model label confusion.

Pair	Cohen's κ
M1 (Claude) vs M2 (GPT-4o)	0.97
HA vs HB (human–human)	0.69
HA vs M1	0.51
HA vs M2	0.52
HB vs M1	0.64
HB vs M2	0.66
4-rater Fleiss's κ	0.66

Table 6: Pairwise inter-annotator agreement on the 774-turn human-validation sample (calibration call excluded). All pairs reach moderate to substantial agreement; the two LLM annotators are near-perfect.

Tactic	HA–HB	HA–M1	HA–M2	HB–M1	HB–M2	M1–M2
T1002	0.48	0.41	0.39	0.48	0.51	0.97
T1003	0.72	0.44	0.44	0.65	0.65	0.99
T1004	0.79	0.67	0.68	0.82	0.81	0.98
T1005	0.76	0.54	0.55	0.65	0.66	0.97
T1006	0.51	0.43	0.47	0.72	0.73	0.91
NONE	0.77	0.53	0.55	0.60	0.67	0.91

Table 7: Per-class one-vs-rest Cohen's κ for all rater pairs on the 774-turn human-validation sample.

B Corpus Statistics

The statistics below are computed over the full ETF standard split (train + test, $n = 5{,}645$ calls).

Human label	Model label					
	T1002	T1003	T1004	T1005	T1006	NONE
T1002	348	57	29	21	29	88
T1003	185	625	14	19	8	27
T1004	68	93	560	19	5	19
T1005	46	166	24	294	2	6
T1006	9	8	2	1	71	39
NONE	8	13	7	6	3	177

Table 8: Human–model label confusion on the human-validation sample, pooled over the four human–model rater pairs (HA/HB × M1/M2).

B.1 Composition and reach by fraud type

Table 9 reports the standard split by fraud type. Table 10 reports call-level reach rates of the two terminal harms by fraud type.

Fraud type	Unit	Train	Test	Total
Customer-service	calls	1,564	391	1,955
	turns	9,626	2,383	12,009
Bank	calls	1,426	357	1,783
	turns	9,125	2,330	11,455
Investment	calls	643	161	804
	turns	4,720	1,170	5,890
Phishing	calls	458	115	573
	turns	2,937	741	3,678
Lottery	calls	240	60	300
	turns	1,276	325	1,601
Kidnap	calls	184	46	230
	turns	538	169	707
Total	calls	**4,515**	**1,130**	**5,645**
	turns	**28,222**	**7,118**	**35,340**

Table 9: Vishing-Tactics-Bench standard split. Each fraud type is reported on two rows: number of calls and number of scammer turns.

B.2 Co-occurrence and timing of the Two Terminal Harms

Table 11 groups calls by which terminal harms they reach. Of the 365 calls that reach financial exploitation, 232 (64%) pass through information gathering and 133 (36%) reach it directly; of the calls that reach information gathering, only about 5% go on to financial exploitation. The two terminal harms are thus in a partial relation—neither a simple escalation ladder nor fully independent—and this relation varies by fraud type. When financial exploitation is reached, the first *Financial Exploitation* utterance occurs at a mean of 4.0 scammer turns (median 4, std 2.4, range 1–17), about 65% of the way through the call, which is why its branch point is captured sharply in the body (Section 7).

Fraud type	*IG* (%)	*FE* (%)
Customer-service	88.9	6.0
Bank	93.4	3.0
Investment	81.5	11.4
Phishing	98.6	0.7
Lottery	69.0	10.3
Kidnap	5.7	28.7
Overall	**85.8**	**6.4**

Table 10: Reach rate of the two terminal harms (*Information Gathering* and *Financial Exploitation*) by fraud type (call-level).

Reach pattern	Calls	%
Both *IG* and *FE*	232	4.1
IG only	4,611	81.7
FE only (no *IG*)	133	2.4
Neither	669	11.9

Table 11: Co-occurrence of the two terminal harms (*Information Gathering* and *Financial Exploitation*) at the call level.

B.3 Per-row uncertainty for Table 3

Tables 12 and 13 report per-row uncertainty for the main results table, split by terminal harm. Encoder (3 variants) and Qwen2.5-7B SFT (2 variants) report mean ± std across 3 seeds (42, 1337, 2024); Markov, ZS LLMs, and gpt-4.1-mini SFT are single runs whose uncertainty is reported via the 95% bootstrap CI over turn-level pairs ($B = 2000$), since Markov is deterministic, the OpenAI tuning API does not expose seeded retraining, and ZS LLMs are queried at temperature 0.1.

B.4 Per-fraud-type heterogeneity

The pooled view of Table 3 averages across six fraud types whose trajectory profiles differ substantially: PII-dominant types (customer service, bank, phishing) account for 76% of test calls; investment and lottery (20%) form a mixed regime; kidnap (4%) is finance-direct. The TTP gain is therefore *type-conditional*. On kidnap, where the *Information Gathering* base rate is an order of magnitude below other types, adding TTP raises gpt-4.1-mini SFT *Information Gathering* AP@1 from 0.32 to 0.77 (+0.45)—the single largest TTP gain in the benchmark. On investment, whose trajectories mimic PII-only types before a late branch, the same move raises *Financial Exploitation* AP@1 by +0.31. Conversely, on bank and lottery, where *Financial Exploitation* phrasing is highly stereotyped, the TTP gain on *Financial Exploitation* AP@1 is

Model	AP@1 / 3 / 5↑	C-idx↑	Div. err.↓
Markov heuristic	.40[.38,.42] / .49[.48,.50] / .52[.50,.53]	.55[.54,.57]	1.45[1.35,1.55]
Encoder (text)	.89±.00 / .95±.00 / .97±.00	.98±.00	1.43±.18
Encoder (TTP)	.32±.01 / .71±.01 / .79±.00	.84±.01	2.10±.05
Encoder (text+TTP)	.89±.01 / .95±.00 / .97±.00	.98±.00	1.36±.07
Qwen2.5-7B SFT (text)	.54±.01 / .73±.00 / .80±.00	.86±.00	1.51±.04
Qwen2.5-7B SFT (text+TTP)	.67±.01 / .81±.00 / .88±.00	.93±.00	1.48±.02
gpt-4.1-mini ZS (text)	.27[.26,.28] / .39[.38,.40] / .41[.40,.42]	.39[.37,.40]	1.22[1.13,1.32]
gpt-4.1-mini ZS (text+TTP)	.36[.34,.38] / .43[.41,.44] / .45[.43,.46]	.45[.44,.47]	0.58[.51,.66]
gpt-4.1-mini SFT (text)	.87[.85,.88] / .91[.90,.92] / .93[.92,.94]	.96[.95,.96]	1.24[1.17,1.32]
gpt-4.1-mini SFT (text+TTP)	.93[.92,.93] / .96[.95,.96] / .97[.96,.97]	.98[.98,.98]	1.28[1.19,1.36]

Table 12: Per-row uncertainty for the main ETF results (Table 3)—***Information Gathering*** only. Conventions as described in this subsection; leading zeros on AP@k and C-index are dropped to save space.

Model	AP@1 / 3 / 5↑	C-idx↑	Div. err.↓
Markov heuristic	.39[.30,.47] / .24[.19,.30] / .22[.17,.26]	.67[.64,.70]	0.07[.00,.19]
Encoder (text)	.58±.02 / .56±.02 / .56±.01	.92±.01	1.15±.19
Encoder (TTP)	.11±.07 / .10±.04 / .10±.03	.68±.01	0.89±.15
Encoder (text+TTP)	.59±.03 / .57±.04 / .56±.05	.91±.01	1.15±.24
Qwen2.5-7B SFT (text)	.09±.02 / .08±.01 / .08±.01	.54±.01	1.96±.22
Qwen2.5-7B SFT (text+TTP)	.32±.02 / .20±.01 / .18±.01	.60±.00	0.18±.08
gpt-4.1-mini ZS (text)	.07[.05,.11] / .09[.07,.12] / .09[.07,.12]	.65[.62,.68]	1.49[1.07,1.94]
gpt-4.1-mini ZS (text+TTP)	.35[.27,.44] / .22[.17,.27] / .18[.14,.23]	.65[.61,.68]	0.34[.14,.60]
gpt-4.1-mini SFT (text)	.51[.42,.59] / .46[.39,.52] / .43[.38,.50]	.83[.80,.86]	0.76[.45,1.14]
gpt-4.1-mini SFT (text+TTP)	.67[.59,.74] / .55[.49,.61] / .53[.48,.59]	.86[.83,.88]	0.00[.00,.00]

Table 13: Per-row uncertainty for the main ETF results (Table 3)—***Financial Exploitation*** only. Conventions match Table 12.

slightly negative (-0.03 on both): text alone already captures the demand wording. The strongest pooled model is also not the strongest per type—on bank *Financial Exploitation*, the Encoder with text alone outranks gpt-4.1-mini SFT (text+TTP) by $+0.26$ AP@1 despite far fewer parameters and no TTP signal. The TTP gain thus concentrates where the trajectory shape deviates from the population average and becomes redundant where harm-bearing phrasing is stereotyped.

C Cascaded TTP Classifier

Predicted-TTP rows in Tables 3 and 4 use a turn-level tactic classifier applied prefix-only over the standard test split (4,515 training calls / 28,222 scammer turns; 1,130 test calls / 7,118 turns). The backbone is the Chinese RoBERTa encoder `hfl/chinese-roberta-wwm-ext` (110M parameters) with a single linear classification head over six classes ($T1002$, $T1003$, $T1004$, $T1005$, $T1006$, NONE). Optimisation is standard supervised cross-entropy with class weights $w_c \propto 1/(n_c + 10)$ to mitigate the rarity of *Financial Exploitation* and *Evasion* (1.4% and 1.4% of training turns), AdamW at learning rate 2×10^{-5}, batch size 32, three epochs, and a maximum input length of 256 word-pieces. At inference the model is restricted to the prefix up to and including the current scammer turn, consistent with a streaming-deployment scenario.

Context-window ablation. We compare three context windows fed to the same RoBERTa backbone under identical optimisation (Table 14). The current scammer turn alone yields the strongest test macro-F1 (0.748); extending to the four preceding turns reduces macro-F1 to 0.701, and the full prefix reduces it further to 0.634. The same ordering holds across all four non-NONE tactic classes. We attribute this to the locality of tactic indicators: each TTP class is most reliably signalled by the lexical surface of the current scammer utterance (e.g., explicit account or credential requests for *Information Gathering*; payment or transfer demands for *Financial Exploitation*). Conditioning on prior turns admixes earlier tactic content into the representation and weakens the discriminative signal of the current turn. We therefore adopt the current-

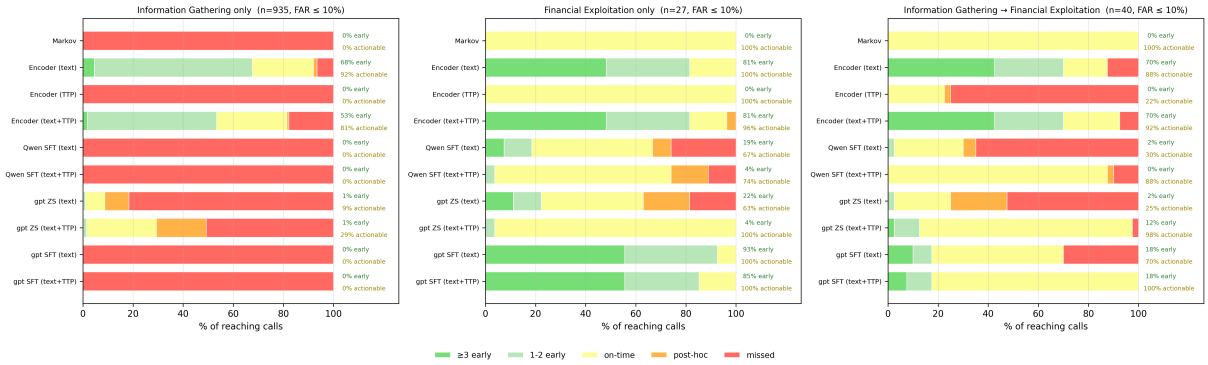


Figure 3: Lead-time breakdown at FAR $\leq 10\%$ under *oracle* TTP; conventions follow Figure 2.

turn classifier as the cascaded TTP source for the predicted-TTP rows in the body tables.

Context	Acc.	Macro-F1
1 turn (current)	**0.846**	**0.748**
5 turns	0.787	0.701
All prior turns	0.758	0.634

Table 14: Cascaded TTP classifier performance as a function of the input context window.

Per-class performance of the current-turn classifier. The chosen classifier attains test accuracy 0.846 and macro-F1 0.748, with per-class F1 of $T1002$ (Trust Building) 0.865, $T1003$ (Psychological Pressure) 0.868, $T1004$ (Information Gathering) 0.903, $T1005$ (Financial Exploitation) 0.584, $T1006$ (Evasion) 0.517, and NONE 0.751. The residual error concentrates on the two rare tactics *Financial Exploitation* and *Evasion*, which accounts for the asymmetric degradation observed in Table 3: predicted-TTP *Financial Exploitation* AP@1 drops more sharply than *Information Gathering* AP@1, mirroring the per-class F1 gap.

Oracle-TTP lead-time breakdown. Figure 3 reports the lead-time breakdown under oracle TTP for comparison with the predicted-TTP figure in the body (Figure 2). The oracle case shows the upper bound on TTP-driven lift; predicted TTP preserves the encoder's lead on the *Information Gathering* $\rightarrow$ *Financial Exploitation* stratum and degrades inline-fusion variants on the *Financial Exploitation only* stratum, tracking the per-class F1 gap.

Numerical lead-time breakdown. Table 15 reports the predicted-TTP lead-time breakdown numerically. The threshold θ is chosen per (model, goal) to satisfy FAR $\leq 10\%$ on the non-reaching pool; *% early* counts $\ell \geq 1$ and *missed* counts calls that never fire by the harm turn. Stratum sizes are IG only n=935, FE only n=27, IG$\rightarrow$FE n=40.

D Sequence-Encoder Ablation

Table 16 replaces the GRU trajectory encoder of the encoder baselines with an LSTM and a two-layer Transformer under an identical protocol (embedding dim 32, hidden dim 64, mean over 3 seeds, gold and predicted TTPs). Late-fusion text+TTP results are within one seed's variance across the three encoders, and the gold-to-predicted change in *Financial Exploitation* AP@1 stays within 0.01 for each; the robustness of late fusion to predicted-TTP noise is a property of the fusion architecture rather than of the specific sequence encoder. In the TTP-only setting the Transformer extracts substantially more signal (*Financial Exploitation* AP@1 0.393 vs. 0.108 for the GRU with gold TTP); with text present this advantage is absorbed by the saturated text branch. On *Information Gathering* the Transformer's TTP-only C-index reaches 0.94 under predicted TTP, narrowing the text-free cost of the privacy-conscious pipeline of Section 9, while *Financial Exploitation* still favors text+TTP.

E Qualitative Examples

Two test calls illustrate when the TTP trajectory helps and when it does not. Bracketed values give the *Financial Exploitation* hazard@3 of the encoder at each scammer turn (text-only / TTP-only / text+TTP, predicted TTP); both dialogues are translated from Chinese. In Example 1 (customer-service impersonation), the text-only and TTP-only branches stay near zero until the demand is spoken, while the fused encoder raises the hazard to 0.99 at the first *Information Gathering* turn, two turns before the harm. In Example 2 (kidnap fraud), with no trust-building, pressure, or information-gathering

Model	IG only (n=935)		FE only (n=27)		IG→FE (n=40)	
	% early	missed	% early	missed	% early	missed
Markov heuristic	0.0	100.0	33.3	3.7	2.5	10.0
Encoder (text)	67.6	6.6	81.5	0.0	70.0	12.5
Encoder (TTP)	0.2	99.7	33.3	3.7	2.5	70.0
Encoder (text+TTP)	56.6	12.6	88.9	0.0	75.0	7.5
Qwen2.5-7B SFT (text)	0.0	100.0	18.5	25.9	2.5	65.0
Qwen2.5-7B SFT (text+TTP)	0.0	100.0	18.5	7.4	0.0	22.5
gpt-4.1-mini ZS (text)	1.1	81.7	22.2	18.5	2.5	52.5
gpt-4.1-mini ZS (text+TTP)	1.5	56.3	22.2	7.4	2.5	12.5
gpt-4.1-mini SFT (text)	0.0	100.0	92.6	0.0	17.5	30.0
gpt-4.1-mini SFT (text+TTP)	0.0	100.0	81.5	0.0	10.0	10.0

Table 15: Per-stratum lead-time numerics under *predicted* TTP at FAR $\leq 10\%$, complementing Figure 2. *% early* counts $\ell \geq 1$ (actionable warning at least one turn before the harm); *missed* counts calls whose alarm never fires by the harm turn.

Encoder	TTP	*Information Gathering*			*Financial Exploitation*		
		AP@1/3/5↑	C-idx↑	Div. err.↓	AP@1/3/5↑	C-idx↑	Div. err.↓
GRU (TTP)	gold	0.318 / 0.707 / 0.787	0.841	2.10	0.108 / 0.098 / 0.099	0.680	0.89
	pred	0.316 / 0.693 / 0.773	0.828	2.12	0.080 / 0.091 / 0.095	0.684	1.45
GRU (text+TTP)	gold	0.889 / 0.950 / 0.968	0.976	1.36	0.593 / 0.571 / 0.559	0.915	1.15
	pred	0.889 / 0.950 / 0.968	0.976	1.38	0.597 / 0.578 / 0.567	0.918	1.25
LSTM (TTP)	gold	0.352 / 0.636 / 0.732	0.803	2.34	0.086 / 0.081 / 0.088	0.638	0.69
	pred	0.348 / 0.610 / 0.701	0.779	2.35	0.066 / 0.074 / 0.082	0.647	0.95
LSTM (text+TTP)	gold	0.892 / 0.949 / 0.966	0.974	1.40	0.578 / 0.560 / 0.545	0.909	1.09
	pred	0.893 / 0.950 / 0.967	0.975	1.37	0.577 / 0.563 / 0.551	0.908	1.26
Transformer (TTP)	gold	0.846 / 0.914 / 0.939	0.959	1.12	0.393 / 0.269 / 0.241	0.726	0.05
	pred	0.776 / 0.880 / 0.910	0.936	1.09	0.214 / 0.182 / 0.171	0.698	0.20
Transformer (text+TTP)	gold	0.898 / 0.950 / 0.966	0.975	1.44	0.601 / 0.602 / 0.612	0.916	1.43
	pred	0.899 / 0.950 / 0.964	0.975	1.44	0.600 / 0.600 / 0.608	0.914	1.37

Table 16: ETF results with alternative sequence encoders over the TTP trajectory, mean across 3 seeds; the GRU rows correspond to the encoder rows of Table 3. Columns follow Table 3.

trajectory, the demand arrives at the second scammer turn: the fused encoder misses it (0.04) while the text-only encoder reacts to the surface wording of the wire demand (0.66) — the pattern of Section 8, where abrupt *Financial Exploitation* carries surface signal but no trajectory precursor.

Example 1: Early warning from the trajectory — customer-service impersonation

Scammer [T1002; 0.02 / 0.13 / 0.30]: "Hello, this is Tianlan Tech, agent Li from customer service. How has your recent shopping experience on our platform been?"
Victim: "It was fine, no particular problems."
Scammer [T1003; 0.02 / 0.02 / 0.66]: "Great — we value every customer's experience. By the way, we are running a special promotion: just download our app and you receive a 100-yuan coupon for free."
Victim: "Oh really? Where do I download it?"
Scammer [T1004; 0.02 / 0.02 / **0.99**]: "Very simple — just click this link and the app downloads directly."
Scammer [T1004; 0.02 / 0.02 / 0.85]: "After downloading, register and enter the invitation code; the coupon is credited automatically."
Scammer [T1005 reached; 0.99 / 0.02 / 1.00]: "And to keep your account safe, we release the coupon after you complete one small transaction — is that convenient?"
Victim: "How much is the small transaction?"
Scammer [T1005]: "Only 5 yuan, just to verify that your account is genuine."

Example 2: Missed abrupt demand — kidnap fraud

Scammer [T1002; 0.05 / 0.13 / 0.00]: "Hello, this is the international finance center. We regret to inform you that your son ran into trouble while traveling abroad. He is being held by a local gang — you must wire 1,000,000 RMB immediately as ransom."
Scammer [T1005 reached; 0.66 / 0.39 / **0.04**]: "Otherwise they may hurt him. Please transfer the money to the designated account at once."
Scammer [T1005; 0.94 / 0.40 / 0.93]: "I will send you the account details."
Scammer [T1003; 0.05 / 0.31 / 0.01]: "Time is extremely tight — you must act now."
Victim: "Wait — if my son is really being held abroad, I need concrete proof, such as a photo or a video."