

Cognitive Echoes: Deployable Conversational Risk Screening in Automotive Finance on Constrained LLMs

Zewei Yu

University of Southern California
zeweiyu@usc.edu

Yuting Sun

Arizona State University
ysun277@asu.edu

Abstract

Automotive-finance screening can fail when individually valid documents are paired with an internally inconsistent applicant narrative. We present the Adversarial Cognitive Probing Framework (ACPF), a theory-motivated staged protocol that scores four ordinal behavioral dimensions and delegates the final decision to a deterministic backend predicate. We evaluate ACPF on a risk-enriched retrospective corpus of 340 officer–applicant dialogues: 72 institutional adjudications of suspected misrepresentation, 138 cases of significant early delinquency, 110 good cases, and 20 non-fraud rejections excluded from the primary metrics. Using identical fixed transcripts and the same QWEN3.5-122B carrier, structured scoring raises combined-risk capture from 5.7% to 61.0% while reducing the false-rejection rate from 14.5% to 12.7%. With OmniMAX, the selected low-FPR carrier, ACPF captures 58/138 (42.0%) delinquency cases at 6/110 (5.5%) false rejections; current origination review flags none of these delinquency cases. On the institutionally adjudicated stratum, ACPF captures 56/72 (77.8%) versus 72/72 for current review, but we make no comparative claim because the label and reference decision share institutional provenance. A shadow deployment has no adverse-action authority; prospective outcome validation is not yet available.

1 Introduction

Automotive-finance risk screening extends beyond document verification: applicants may submit genuine materials while presenting an inconsistent underlying narrative, including third-party participation in verification calls, selectively disclosed bank statements, or discourse that obscures the purpose and affordability of a purchase. Rule-based engines (Abdallah et al., 2016), credit bureau queries, and OCR verification are structurally blind to this attack surface—they validate individual evidence

fragments without auditing the narrative logic that connects them.

Prior deception research (DePaulo et al., 2003; Vrij, 2008) motivates candidate cues such as temporal inconsistency, limited sensory detail, evasion, and counter-question deflection. These cues are hypotheses for protocol design, not validated psychological measurements or evidence of fraud in an individual case. Skilled loan officers attend to related signals, but their judgments are difficult to standardize. More generally, truth-default theory argues that people ordinarily presume communicators to be honest, which constrains unaided deception detection (Levine, 2014). Psycholinguistic approaches to automated deception detection (Mihalcea and Strapparava, 2009) and LLM-based evaluators (Zheng et al., 2023) have shown promise in laboratory settings. Our study instead examines whether structured scoring is operationally useful on fixed, retrospective financial dialogues; it does not validate deception detection or prospective decision performance. Figure 1 summarizes the evaluated fixed-transcript path and its deterministic backend safeguards.

We address this gap with ACPF, a structured staged-dialogue specification that formalizes loan-officer heuristics as reproducible, theory-motivated probe sequences and a fixed scoring rubric. The protocol includes high-interactivity compounds, episodic-specificity questions, and follow-ups designed to expose changes in the behavioral signals in Table 1. Cognitive Echoes is the industrial implementation of ACPF, currently operated in shadow mode, aggregating probe responses through multi-round retrospection to produce auditable behavioral risk indicators. The primary contributions are: (i) a theory-motivated indicator set studied on $n = 100$ annotated dialogues, with explicitly limited ordinal reliability; (ii) a deterministic trigger rule designed for a near-zero false-rejection constraint; and (iii) a retrospective evaluation on $N = 340$ dialogues

ACPF retrospective scoring and deterministic decision pipeline

Fixed officer-applicant transcripts; no machine-generated probes

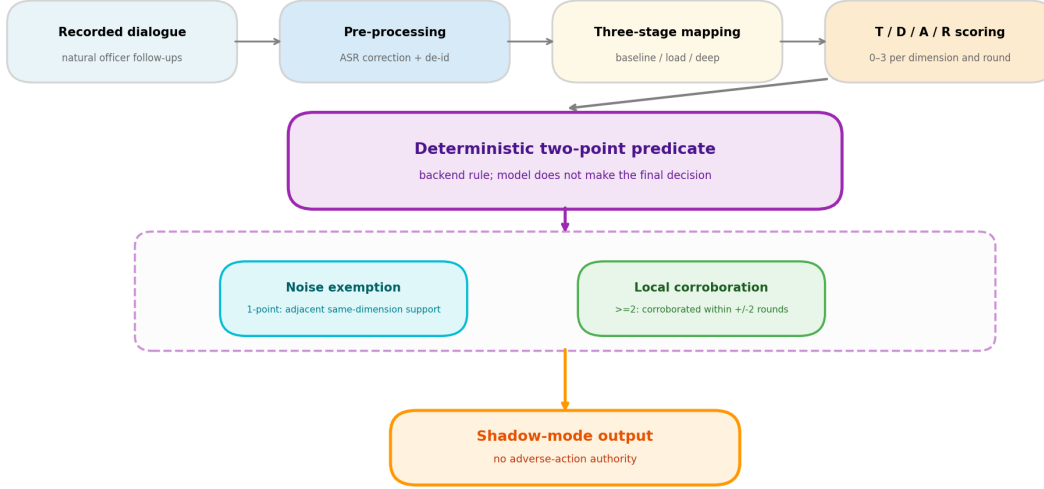


Figure 1: Retrospective ACPF pipeline. Fixed officer–applicant transcripts (no machine-generated probes) receive per-round T/D/A/R scores; a deterministic backend applies noise exemption and local corroboration. Outputs remain in shadow mode with no adverse-action authority.

that separates institutionally adjudicated suspected misrepresentation from repayment-derived delinquency. The corpus contains natural officer follow-ups. It does not test machine-generated adaptive questioning, which remains a prospective research question.

2 ACPF Framework

ACPF draws on Cognitive Load Theory (Sweller, 1988), Episodic Memory Theory (Tulving, 1972), and Self-Presentation Theory (Leary and Kowalski, 1990). These theories generated candidate indicators; they do not establish that the resulting scores measure cognitive load, episodic-memory richness, or self-presentation stress. Under questioning, an inconsistent account is hypothesized to impose working-memory demands (Vrij, 2008); episodic accounts may contain richer perceptual detail (Johnson et al., 1993); and topic shifts or counter-questions may mark response avoidance. ACPF operationalizes these hypotheses as a three-stage protocol with four behavioral indicators (Table 1), each scored 0–3 per round. We distinguish human-adaptive follow-up in the recorded interviews from machine-generated adaptive questioning. Only the former is present in the retrospective corpus. One critical refinement concerns marginal activations: 1-point activations may reflect tran-

ID	Indicator	Theory	d [95% CI]
T	Timeline Consistency	CLT	0.66 [0.26, 1.06]
D	Detail Richness	EMT	0.62 [0.22, 1.02]
A	Avoidance Behavior	CLT	0.66 [0.26, 1.06]
R	Response Evasion	SPT	0.42 [0.02, 0.82]

Table 1: Four theory-motivated behavioral indicators (annotation study $n = 100$). Cohen’s d is computed on binary presence/absence judgments using the pooled within-group standard deviation; intervals use the normal-theory standard-error approximation with $n_1 = n_2 = 50$. CLT = Cognitive Load Theory; EMT = Episodic Memory Theory; SPT = Self-Presentation Theory. T, D, A are core; R is auxiliary.

scription noise or natural variation, so only corroborated ≥ 2 activations may produce a rejection. The round score is $s_r = \max(T_r, D_r, A_r, R_r)$ and the cumulative score is $S_k = \sum_{r=1}^k s_r$. Stage 2 grants early passage on consecutive zero-score rounds; Stage 3 issues the final determination under the two-point trigger rule (Section 4).

3 Indicator Development Study

Beginning with 30 candidates, expert roundtables narrowed the pool to 10. Three experts annotated 100 dialogues (50 institutional-adjudication positives and 50 genuine cases) in two exercises. Binary presence/absence agreement was near-perfect,

ID	Indicator	d [95% CI]	Decision
I1	Timeline Consistency	0.66 [0.26, 1.06]	Core
I2	Numerical Coherence	0.00 [-0.39, 0.39]	Rejected
I3	Semantic Coherence	0.45 [0.05, 0.85]	Explor. [†]
I4	Linguistic Fluency	-0.21 [-0.60, 0.18]	Rejected
I5	Rehearsal Cues	0.22 [-0.17, 0.61]	Rejected
I6	Self-Correction	0.22 [-0.17, 0.61]	Rejected
I7	Response Evasion	0.42 [0.02, 0.82]	Auxiliary
I8	Detail Richness	0.62 [0.22, 1.02]	Core
I9	Avoidance Behavior	0.66 [0.26, 1.06]	Core
I10	Response Relevance	0.22 [-0.17, 0.61]	Rejected

Table 2: All 10 candidate indicators ($n = 100$, 3 annotators). d : Cohen’s d on binary presence/absence judgments, computed with the pooled within-group standard deviation; intervals use the normal-theory standard-error approximation with $n_1 = n_2 = 50$. The $d \geq 0.40$ rule is a pilot retention heuristic (Sullivan and Feinn, 2012), not a validation threshold. [†]I3 exceeds the heuristic threshold but was not included in the deployed core set. All 50 good-group cases scored zero, but this one-sided separation does not itself justify exclusion; a sensitivity analysis of adding I3 was not available.

but extreme-stratum sampling makes the resulting Fleiss’ κ uninformative as validation evidence, so we do not rely on it. Independent 0–3 severity ratings yielded a reported ICC(2, k) range of 0.45–0.60 for the three-rater mean, indicating limited reliability for fine-grained scoring (Koo and Li, 2016). Table 2 reports effect sizes for all 10 candidates; we used $d \geq 0.40$ as a study-specific pilot retention heuristic. Effect size is reported, rather than relying on statistical significance alone, but the cutoff is not a validated construct threshold (Sullivan and Feinn, 2012). Three core indicators show medium-to-large Cohen’s d (T: 0.66; D: 0.62; A: 0.66); auxiliary R is retained at $d = 0.42$. Linguistic fluency (I4, $d = -0.21$) is excluded because adjudicated-risk applicants score higher on this dimension, consistent with the prepared-liar phenomenon (Vrij, 2008). Semantic coherence (I3, $d = 0.45$) exceeds the retention threshold but was not included in the deployed core set. All 50 good-group cases received a zero score on I3. This one-sided separation is a property of the observed sample, not a statistical ground for exclusion. We therefore withdraw the original floor-effect rationale, report I3 as exploratory, and do not claim that the retained set is uniquely optimal. Its incremental contribution to the deployed four-indicator rule remains unevaluated.

4 Deployment Constraints and Decision Redesign

Two institutional constraints necessitate redesigning the decision mechanism of ACPF prior to deployment. *Model access*: applicant data cannot leave the institutional intranet, ruling out third-party cloud APIs and making systematic carrier selection a prerequisite. *Near-zero false-rejection*: an operational audit standard mandates that rejection of legitimate applicants be minimized to a materially negligible rate. Analysis of the retrospective corpus reveals that the original $S \geq 4$ cumulative threshold is insensitive to risk cases with concentrated single-round signals, capturing only 37.1% of positives (see also Table 7 caption).

The cumulative mechanism is replaced by a **two-point indicator trigger** (2PT) with two safeguards. Let $s_{r,d} \in \{0, 1, 2, 3\}$ be the score for round r and dimension $d \in \mathcal{D} = \{T, D, A, R\}$. Let $\mathcal{I} = \{1, \dots, m\}$ denote the observed rounds and $\mathcal{A}_r = \{r - 1, r + 1\} \cap \mathcal{I}$ the adjacent rounds. Noise exemption is applied per dimension. A 1-point activation is valid only when the same dimension is non-zero in an adjacent round:

$$v_{r,d} = \mathbb{K} \left[s_{r,d} \geq 2 \vee (s_{r,d} = 1 \wedge \max_{j \in \mathcal{A}_r} s_{j,d} > 0) \right]. \quad (1)$$

For an empty $\mathcal{A}_r$, the maximum is defined as zero. Thus an isolated 1-point activation cannot serve as corroboration. For candidate round r , define the distinct local neighborhood $\mathcal{N}_r = \{j \in \mathcal{I} : 0 < |j - r| \leq 2\}$. The rule rejects only when a ≥ 2 activation has a valid activation on any dimension in this neighborhood:

$$\text{Reject} = \mathbb{K} \left[\exists r, d : s_{r,d} \geq 2 \wedge \max_{\substack{j \in \mathcal{N}_r \\ k \in \mathcal{D}}} v_{j,k} = 1 \right]. \quad (2)$$

Accordingly, a score of 2 does not cause an immediate rejection. It becomes a rejection trigger only after local corroboration. Appendix D gives equivalent pseudocode. The threshold is an operational design choice evaluated on the retrospective corpus; the study does not estimate latent deceptive and noise distributions in the Signal Detection Theory sense (Green and Swets, 1966).

5 Three-Layer Retrospective Evaluation

5.1 Experimental Setup and Dataset

All three layers draw from $N = 340$ loan application audio recordings from a Chinese automotive finance company (anonymized), captured during loan officer–applicant telephone interviews and subsequently ASR-transcribed. The corpus is a stratified sample enriched for risk cases to support statistical analysis; it does not reflect the institution’s true base rate of delinquency; reported rates are conditional on the sampled strata and are not population-prevalence estimates. Appendix E accounts for the final evaluated strata. The authors confirm that the $n = 100$ indicator-development cohort and the $N = 340$ evaluation corpus are disjoint; no case appears in both.

Labels have two distinct sources. Significant delinquency is defined from repayment records. The second risk stratum consists of institutional adjudications of suspected misrepresentation; these adjudications are neither judicial findings nor independently verified fraud labels. The two strata are reported separately wherever the reference decision shares label provenance. Dialogues undergo ASR correction, privacy de-identification, and two-annotator stage mapping (91% agreement; disagreements arbitrated by a third senior officer). Scores are computed with the LLM blind to all risk labels and expert decisions.

Positive class definition. For combined-risk summaries, the positive class comprises institutional adjudications of suspected misrepresentation ($n_{\text{adj}} = 72$) and significant early-delinquency cases ($n_{\text{credit}} = 138$: ≥ 31 days consecutive delinquency first triggered within the 3rd, 6th, or 12th installment period), totaling $n_{\text{pos}} = 210$. Non-fraud rejection ($n = 20$) is excluded from all metrics. Capture is therefore $\text{TP}/210$; FPR is computed on the good group ($n_{\text{neg}} = 110$): no-delinquency ($n = 68$) and marginal 1–5-day delinquency ($n = 42$), both of which represent legitimate applicants under the near-zero false-rejection constraint.

Table 3 gives the stratified breakdown; the positive class is $n_{\text{pos}} = 210$ and the FPR denominator is the good group ($n_{\text{neg}} = 110$). Interception efficiency ($\text{IE} = \text{true positives} / \text{good-group false positives}$) captures the precision-capture trade-off central to loan operations. $\text{G-Mean} = \sqrt{\text{TPR} \times \text{TNR}}$ serves as the class-balanced summary metric: unlike F1, it penalizes both missed risk and false rejection equally.

Group	Category	n
Good	No Delinquency	68
	Delinquency 1–5 Days	42
Credit Risk	First 3 Installments	38
	First 6 Installments	50
	First 12 Installments	50
Adjudicated	Individual suspected misrepresentation	34
	Organized suspected misrepresentation	38
Other	Non-fraud Rejection	20
Total		340

Table 3: Risk-enriched retrospective dataset ($N = 340$). The positive class comprises institutional adjudications of suspected misrepresentation ($n = 72$: individual 34 + organized 38) and repayment-derived significant delinquency ($n = 138$: installment periods 1–3 ($n = 38$), 4–6 ($n = 50$), and 7–12 ($n = 50$)). The FPR denominator is the good group ($n = 110$: no delinquency 68 + marginal 1–5-day delinquency 42). Twenty non-fraud rejections are excluded from the primary metrics.

5.2 Layer 1: Protocol Value — ACPF vs. Direct Judgment

Layer 1 asks whether the ACPF protocol adds measurable value over raw LLM judgment on the same transcripts. Two conditions are compared, both using QWEN3.5-122B as the carrier: (A) **Direct Judgment**—the LLM receives the full dialogue transcript and is prompted to directly assess loan approval risk, with no structured probe sequence, no indicator rubric, and no scoring protocol. The evaluated prompt asks whether the application should be approved and names the full risk scope—both adjudicated misrepresentation and early delinquency. Only one direct-judgment prompt variant was evaluated, so the comparison is limited to this condition. (B) **ACPF**—the same model applies the three-stage scoring protocol with four-indicator two-point trigger scoring. Both conditions process identical dialogue content with the same carrier; the comparison bundles stagewise scoring, the indicator rubric, and the deterministic predicate rather than isolating any one component.

The TF-IDF + logistic-regression baseline reaches 22.9% capture and G-Mean 0.450, above Direct Judgment but below ACPF at the reported operating points. Its feature settings, regularization, fold assignments, and case-level predictions are not reported; no claim beyond this aggregate comparison is made. Table 4 shows that ACPF raises capture from 5.7% to 61.0% (+55.3 pp), re-

Condition	FPR↓	Capture↑	G-Mean↑	IE↑
Direct Judgment (QWEN)	14.5%	5.7%	0.221	0.75×
TF-IDF + LR	11.6%	22.9%	0.450	3.7×
ACPF (QWEN)	12.7%	61.0%	0.729	9.1×
Δ (ACPF – Direct)	–1.8pp	+55.3pp	+0.508	0.75→9.1×

Table 4: Layer 1 baseline: ACPF vs. direct LLM judgment (QWEN3.5-122B, $N = 340$). Capture = TP/210 (all risk cases); FPR on good group ($n = 110$). Direct judgment receives the full transcript and the full risk scope. G-Mean for Direct Judgment is recomputed from 12/210 capture and 94/110 specificity. TF-IDF + LR is available only as aggregate metrics; inferred integer counts are not reported. The same-corpus comparison does not isolate individual protocol components. IE = total bad-case TP / FP.

duces FPR by 1.8 pp, and raises IE from 0.75× to 9.1× (a 12.1-fold relative increase). Because both conditions are evaluated on the same corpus without repeated prompt variants, the result supports structured scoring over this direct-judgment prompt; it does not establish that no alternative prompt could reproduce the gain.

5.3 Layer 2: Carrier Capacity — Enterprise LLM Selection

Layer 2 asks which LLM carrier best executes the ACPF protocol under institutional constraints. Because applicant data cannot be routed to external APIs, candidate carriers are restricted to models deployable within the institutional intranet or accessible via approved private API contracts. Three LLMs are evaluated under the 2PT rule on the full $N = 340$ set. The officer row reflects original origination decisions and is reported as a descriptive operational reference, not as an independent benchmark or upper bound.

Table 6 separates the two outcome strata. The primary retained comparison concerns significant delinquency, whose label is derived from repayment records independently of the origination decision. OmniMAX flags 58/138 (42.0%) of these cases, whereas the original origination decisions flag none. Because the delinquency label derives from later repayment records, this comparison does not share the label provenance problem of the adjudicated stratum. On the adjudicated stratum, OmniMAX flags 56/72 (77.8%) and current review contains all 72 cases by construction; we make no comparative claim.

Table 5 gives the combined-risk operating summary. For carrier selection, QWEN3.5-122B has

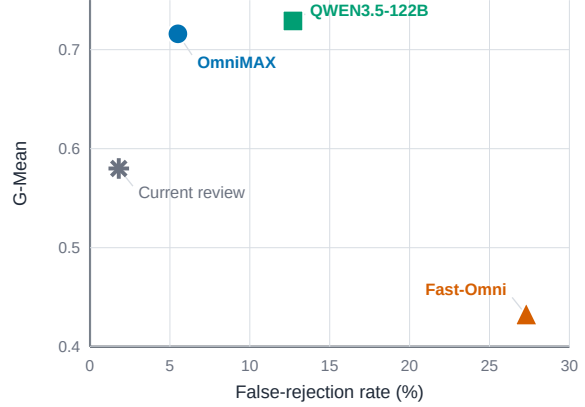


Figure 2: Carrier operating points. Current review is shown only as a descriptive reference because its adjudicated-stratum decision shares label provenance. OmniMAX is selected under the institution’s lower-FPR operating constraint, although QWEN has the higher observed G-Mean.

the highest G-Mean (0.729) and capture (61.0%), whereas OmniMAX has a lower FPR (5.5% versus 12.7%) with G-Mean 0.716 and capture 54.3%. The operational requirement prioritizes false rejection, so OmniMAX is selected. Fast-Omni is not selected (G-Mean 0.432, capture 25.7%, FPR 27.3%). Figure 2 visualizes these observed operating points without treating current review as an independently labeled benchmark.

5.4 Layer 3: Same-Sample Decision-Rule Comparison

Layer 3 compares candidate decision rules, given OmniMAX as the carrier. Four configurations are ablated on the full $N = 340$ set, each combining the 2PT rule with progressively stricter cumulative score requirements.

Table 7 shows that, among the four evaluated configurations, 2PT has the highest G-Mean (0.716) and capture rate (54.3%). The cumulative-only design ($S \geq 4$, no 2PT), capturing only 37.1% under a structurally different triggering logic, appears as a motivating reference in Section 4 rather than an ablation row. Each additional cumulative constraint reduces both capture and G-Mean while lowering FPR. These rows therefore represent different trade-offs; selecting among them requires an externally fixed error-cost policy rather than a claim of unconditional statistical dominance. Stricter configurations yield higher finite IE through $S \geq 4$, but reduce risk coverage. IE is undefined at $S \geq 5$ because no false positives are observed. A sim-

Carrier/reference	FPR↓	Capture↑	G-Mean↑	F1↑	IE↑	Operational status
Current review [†]	1.8%	34.3%	0.580	0.507	36.0×	Descriptive reference
OmniMAX	5.5%	54.3%	0.716	0.691	19.0×	Selected
QWEN3.5-122B	12.7%	61.0%	0.729	0.727	9.1×	Not selected
Fast-Omni	27.3%	25.7%	0.432	0.367	1.8×	Not selected

Table 5: Combined-risk carrier summary ($N = 340$; positive denominator $n = 210$, good-group denominator $n = 110$). OmniMAX is selected because the institutional operating constraint prioritizes lower false rejection; this is not a claim that it maximizes every metric. [†]Current review is included only to describe existing practice, not as an independently labeled model benchmark.

Outcome stratum	OmniMAX (2PT)	QWEN3.5-122B (2PT)	Current review [†]
Adjudicated suspected misrepresentation ($n = 72$)	56 (77.8%) [66.9, 85.8]	—	72 (100%) [94.9, 100]
Significant delinquency ($n = 138$)	58 (42.0%) [34.1, 50.4]	—	0 (0%) [0, 2.7]
Combined risk ($n = 210$)	114 (54.3%) [47.5, 60.9]	128 (61.0%) [54.2, 67.3]	72 (34.3%) [28.2, 41.0]
False rejections ($n = 110$)	6 (5.5%) [2.5, 11.4]	14 (12.7%) [7.7, 20.2]	2 (1.8%) [0.5, 6.4]

Table 6: Outcome-stratified carrier results ($N = 340$, two-point trigger). Entries are counts, percentages, and Wilson 95% intervals where estimable from reported counts. [†]The officer row records original origination decisions and is a descriptive reference for current practice. On the adjudicated stratum, the label and reference decision share institutional provenance, so the 100% value is partly definitional and is not used for a comparative claim. QWEN stratum-specific counts are not reported.

ilar monotone capture decrease was reported for QWEN3.5-122B, but corresponding case-level predictions are not reported; we therefore do not infer that the pattern generalizes across carriers.

Why cumulative requirements degrade performance. Adding a cumulative threshold on top of 2PT (rows 2–4) monotonically reduces combined-risk capture in this corpus. Round-level stratum counts sufficient to establish why individual cases are removed were not retained, so concentration of signals in early rounds remains a hypothesis rather than a demonstrated mechanism.

The $S \geq 5$ extreme. At the extreme end, $S \geq 5$ achieves 0.0% observed FPR but only 17.1% capture. Within the evaluated OmniMAX 2PT family, bare 2PT provides the highest capture and G-Mean (Figure 3).

Business Metrics (OmniMAX, 2-pt Trigger, $N = 340$). Evaluated against repayment outcomes, the deployed configuration flags 58/138 (42.0%) significant-delinquency cases at 6/110 (5.5%) false rejections; the original origination decisions flag none of those 138 cases. On the institutionally adjudicated stratum, ACPF flags 56/72 (77.8%), below the reference row’s 72/72, but comparison is not independent because the label and reference decision share provenance. Combined-risk capture is 114/210 (54.3%). The evidence-supported incremental value therefore lies in the 58

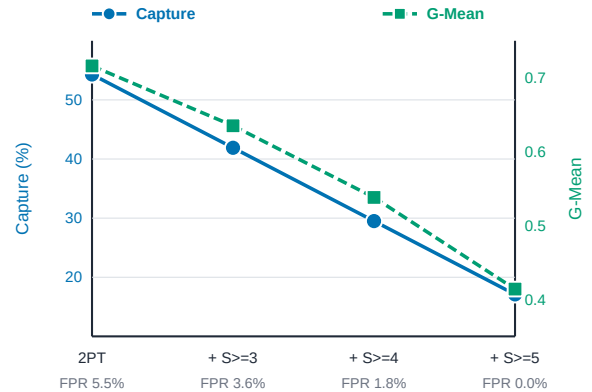


Figure 3: Same-sample OmniMAX rule comparison. Stricter cumulative requirements reduce observed FPR but also reduce capture and G-Mean. This is a development-set trade-off, not an independently validated rule ranking.

repayment-derived delinquency cases that passed conventional origination review, not in a pooled claim of superiority to officers.

6 Conclusion

On $N = 340$ retrospective dialogues, structured ACPF scoring on QWEN raises combined-risk capture from 5.7% to 61.0% relative to the evaluated direct judgment prompt, while reducing FPR from 14.5% to 12.7%. The primary operational find-

Configuration	FPR↓	Capture↑	F1↑	G-Mean↑	IE↑
2-pt Trigger (deployed)	5.5% (6/110)	54.3% (114/210)	0.691	0.716	19.0×
2-pt Trigger + $S \geq 3$	3.6% (4/110)	41.9% (88/210)	0.583	0.635	22.0×
2-pt Trigger + $S \geq 4$	1.8% (2/110)	29.5% (62/210)	0.453	0.538	31.0×
2-pt Trigger + $S \geq 5$	0.0% (0/110)	17.1% (36/210)	0.293	0.414	— [‡]

Table 7: Layer 3 ablation of decision configurations (OmniMAX, $N = 340$). Capture = TP/210; FPR on good group ($n = 110$). All four configurations use the 2PT trigger as a foundation; the original cumulative-only design ($S \geq 4$, no 2PT) is not a column in this table because it operates under a different triggering logic—its 37.1% capture rate is derived from the retrospective corpus analysis in Section 4 and serves as the motivating baseline for adopting 2PT. Within the 2PT family, adding cumulative requirements monotonically reduces both capture and G-Mean while lowering FPR. [‡]IE is undefined when FP = 0; the row is excluded from IE ranking. Rule choice and evaluation use the same retrospective corpus, so the table is a same-sample configuration comparison rather than independent validation.

ing is stratum-specific: OmniMAX flags 58/138 (42.0%) repayment-derived delinquency cases at 6/110 (5.5%) false rejections, whereas the original origination decisions flag none. We make no superiority claim on the institutionally adjudicated stratum because label and reference decision share provenance. OmniMAX is selected over QWEN for its lower FPR under the institution’s operating constraint. Within the same-sample OmniMAX configuration comparison, bare 2PT has the highest capture and G-Mean. The retrospective study evaluates structured scoring on fixed transcripts; machine-generated adaptive elicitation, independent rule selection, and prospective performance remain to be validated. The live system remains in shadow mode with no adverse-action authority.

Future Work

Four directions define the replication protocol. **(i) Prospective validation:** a pre-specified two-arm trial comparing a fixed interview with machine-generated adaptive follow-up, with stratum-specific outcomes and stopping rules fixed before enrollment. **(ii) Independent rule selection:** threshold selection on a development cohort followed by untouched-cohort replication. **(iii) Comparative baselines:** a fine-tuned Chinese encoder and repeated prompt variants. **(iv) Construct and fairness analysis:** regression on linguistic and ASR covariates, score-distribution analysis of the 20 excluded non-fraud rejections, and adequately powered subgroup audits.

Limitations

The retrospective dataset contains 340 cases from one institution and is risk-enriched; its capture rates are sample estimates, not population prevalence.

Stage mapping is post hoc, and rule motivation, carrier selection, and evaluation use the same corpus. The reported operating points are thus development evidence: no untouched-cohort, prospective outcome, or completed cross-validation result is reported.

The two positive strata have different provenance: delinquency follows repayment records, whereas suspected misrepresentation follows institutional adjudication without judicial or third-party verification. Cases with 6–30 days of delinquency and 20 non-fraud rejections are outside the primary metrics; the latter’s score distribution is not reported. Binary agreement from extreme-stratum sampling is uninformative, and $ICC(2,k)=0.45-0.60$ indicates limited reliability for fine-grained scoring. Semantic coherence (I3) exceeded the pilot threshold but was not included in the deployed set; the original floor-effect rationale is withdrawn, and the retained set should not be treated as uniquely optimal. A complete application-by-round T/D/A/R matrix was not retained, so the safeguards cannot be reconstructed or ablated at case level; unobserved scores are not back-filled from aggregate counts.

Dialogue-based screening may disadvantage applicants with limited linguistic proficiency, and a closed-source carrier creates version-drift risk. Findings are specific to the evaluated carriers, corpus, and institutional context; broader use requires independent replication, construct-validity analysis, linguistic-covariate regression, and powered subgroup-fairness audits.

Ethical Considerations

All dialogues undergo privacy de-identification before entering the research pipeline, and scoring is

reported as occurring on the institutional intranet. Appendix F states minimum governance requirements for any prospective decision use; the present study does not evaluate the effectiveness of those controls. ACPF is restricted to voluntary adult financial applications and must not be used in interrogation or monitoring contexts. A behavioral score is not evidence of fraud, and declining additional probing must carry no adverse inference. Any adverse action requires human review, factual reason codes, transcript correction and appeal, audit logging, version control, rollback criteria, and powered fairness monitoring.

Acknowledgements

We thank the partner automotive finance institution for providing the loan application dialogue corpus, the ASR transcription pipeline, and the operational infrastructure for the live shadow deployment. We are particularly grateful to the three expert loan officers who conducted the behavioural indicator annotation and stage mapping, blind to the research hypotheses and ACPF scores. We also thank the anonymous reviewers and the area chair for their exceptionally detailed and constructive feedback, which substantially improved the framing, rigour, and ethical safeguards of this paper.

References

- Aisha Abdallah, Mohd Aizaini Maarof, and Anazida Zainal. 2016. [Fraud detection system: A survey](#). *Journal of Network and Computer Applications*, 68:90–113.
- Bella M. DePaulo, James J. Lindsay, Brian E. Malone, Laura Muhlenbruck, Kelly Charlton, and Harris Cooper. 2003. [Cues to deception](#). *Psychological Bulletin*, 129(1):74–118.
- David M. Green and John A. Swets. 1966. *Signal Detection Theory and Psychophysics*. Wiley, New York.
- Marcia K. Johnson, Shahin Hashtroudi, and D. Stephen Lindsay. 1993. [Source monitoring](#). *Psychological Bulletin*, 114(1):3–28.
- Terry K. Koo and Mae Y. Li. 2016. [A guideline of selecting and reporting intraclass correlation coefficients for reliability research](#). *Journal of Chiropractic Medicine*, 15(2):155–163.
- Mark R. Leary and Robin M. Kowalski. 1990. [Impression management: A literature review and two-component model](#). *Psychological Bulletin*, 107(1):34–47.
- Timothy R. Levine. 2014. [Truth-default theory \(TDT\): A theory of human deception and deception detection](#). *Journal of Language and Social Psychology*, 33(4):378–392.
- Rada Mihalcea and Carlo Strapparava. 2009. [The lie detector: Explorations in the automatic recognition of deceptive language](#). In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 309–312, Suntec, Singapore. Association for Computational Linguistics.
- Gail M. Sullivan and Richard Feinn. 2012. [Using effect size—or why the P value is not enough](#). *Journal of Graduate Medical Education*, 4(3):279–282.
- John Sweller. 1988. [Cognitive load during problem solving: Effects on learning](#). *Cognitive Science*, 12(2):257–285.
- Endel Tulving. 1972. Episodic and semantic memory. In Endel Tulving and Wayne Donaldson, editors, *Organization of Memory*, pages 381–403. Academic Press, New York.
- Aldert Vrij. 2008. *Detecting Lies and Deceit: Pitfalls and Opportunities*, 2nd edition. Wiley.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging LLM-as-a-judge with MT-bench and chatbot arena](#). In *Advances in Neural Information Processing Systems*, volume 36.

A Glossary

Two-Point Trigger (2PT). A score ≥ 2 is a candidate trigger, not an immediate rejection. Rejection requires a distinct valid activation on any T/D/A/R dimension within ± 2 rounds (Eqs. 1–2).

Local Corroboration Window. For a candidate round r , the predicate searches rounds $r-2$ through $r+2$, excluding r itself, for a valid corroborating activation. A 1-point activation is valid only when the same dimension is non-zero in round $r-1$ or $r+1$; an isolated 1-point activation is therefore ignored. The eventual corroborator may be on any of the four dimensions.

G-Mean. $\sqrt{\text{TPR} \times \text{TNR}}$: geometric mean of sensitivity and specificity, selected as the carrier ranking metric for robustness to class imbalance and asymmetric misclassification cost.

Interception Efficiency (IE). The ratio of total bad-case true positives to good-group false positives. An IE of $19.0\times$ for the OmniMAX 2PT configuration indicates that 19 risk cases are intercepted for each legitimate applicant incorrectly rejected.

Supplementary Risk Capture Rate. (ACPF-rejected repayment-derived risk cases) / (repayment-derived risk cases passed by current origination review). This rate is $58/138$ (42.0%) in the retrospective corpus.

B Expert Background

Three expert loan officers, blind to ACPF scoring and research hypotheses, conduct the indicator annotation study; disagreements in stage mapping are resolved through third-party arbitration by a senior officer. Mean experience: 14.3 years (range 10–18 years).

C Reproducibility Notes

Scoring employs stage-level prompts that process only the dialogue text corresponding to the current stage. Each stage outputs structured JSON (per-dimension scores, round markers); cumulative calculation, noise filtering, and early-exit logic run in backend Python scripts to prevent LLM arithmetic errors. Current deployment employs three-module prompt separation (dimension rubric, scoring exemplars, output format). The following interface skeleton documents the expected input/output contract:

```
System: Score the stage dialogue on timeline
consistency (T),
detail richness (D), avoidance behaviour (A), and
response
evasion (R), using a 0--3 scale and the provided
exemplars.
User: [stage dialogue transcript]
Assistant: T: [score] [rationale] / D: ... / A: ...
/ R: ...
```

This skeleton is not a complete scoring prompt. Full per-dimension definitions, 0/1/2/3 level rubrics and exemplars, the exact model snapshot and decoding parameters, and the complete production prompt are not reported. Prompt-level reproducibility is therefore limited.

C.1 Redacted Dialogue Example

The following redacted excerpt illustrates an episodic-specificity probe. It is a qualitative example only and is not evidence of construct validity.

```
Officer: Which dealership did you visit?
Applicant: A dealership in <city>.
Officer: What was it called?
Applicant: I don't quite remember.
Officer: Which street? How did you get there?
Applicant: A friend drove me. I don't remember the
route.
Officer: What was the salesperson's name?
Applicant: Maybe... <surname>? I don't remember the
full name.
Officer: Male or female?
Applicant: Female, I think.
```

D Decision Predicate Pseudocode

The decision predicate in Eqs. 1–2 can be expressed by the following normalized pseudocode. This is not a verbatim production-code listing; it documents the dimension-specific validation and cross-dimension corroboration logic evaluated in the paper.

```
dimensions = {T, D, A, R}

for each round r and dimension d:
    valid[r,d] = (score[r,d] >= 2)
    if score[r,d] == 1:
        valid[r,d] =
            (r-1 exists and score[r-1,d] > 0) or
            (r+1 exists and score[r+1,d] > 0)

for each round r and dimension d with score[r,d] >=
2:
    for each distinct round j with abs(j-r) <= 2:
        if any valid[j,k] for k in dimensions:
            return REJECT
return PASS
```

E Available Cohort Accounting

The final stratified evaluation sample is documented, but the source-pool size, intermediate eligibility screens, duplicate-removal count, missing-audio exclusions, and score distribution of the 20 non-fraud rejections are not reported. Table 8 therefore provides cohort accounting, not a complete source-to-analysis flow diagram.

Table 8: Final cohort accounting. The separate $n = 100$ indicator-development cohort is disjoint from this evaluation corpus.

Analysis status	Stratum	n
Included	Adjudicated suspected misrepresentation	72
Included	Significant delinquency	138
Included	Good	110
Excluded	Non-fraud rejection	20
Total sampled corpus		340

F Governance Requirements for Prospective Use

The following controls are requirements for prospective decision use. The five-round minimum, low-quality-ASR abstention with manual routing, and quarterly model review are author-confirmed current procedures; the remaining items should not be read as evidence of completed implementation.

- Separate application, recording, research-use, and automated-screening notices, with a non-probing alternative.
- A minimum of five complete question-response rounds; otherwise the system abstains. Scoring also abstains when ASR quality, language support, or transcript integrity falls below the applicable institutional operating standard, and the case is routed to manual review. No unverified numeric ASR cutoff is asserted here.
- Declining extra probing carries no adverse inference.
- Mandatory trained-human review before any adverse action. A score is not evidence of fraud.
- Factual reason codes tied to specific transcript rounds and rubric dimensions, plus transcript correction and appeal.
- Frozen model, prompt, rubric, and predicate versions; logged changes, accountable override, monitoring, and rollback criteria. Model review and, when indicated, recalibration are scheduled quarterly; numeric drift tolerances remain governed by the institution's audited risk-management protocol.
- Retention limits, role-based access, audit logging, and incident review.
- Pre-specified drift and subgroup-fairness audits with power analysis; decision use pauses when minimum support or quality thresholds fail.