COMPLAINTBENCH: A DATASET AND BENCHMARK FOR EVIDENCE-GROUNDED PAYMENT COMPLAINT ASSESSMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce ComplaintBench, a dataset and diagnostic benchmark for evidence-grounded risk assessment in Chinese-language payment complaints: 87,359 records with 286,582 screenshot references, each carrying a narrative, chat and payment screenshots, a binary positive / non-positive decision, and one of 32 operational categories. The decision policy places ordinary disputes and insufficient-evidence cases in the non-positive class, so "unverifiable" is a legitimate outcome rather than an abstention. A two-stage reference pipeline extracts source-linked evidence from text and screenshots with a frozen vision-language model and then classifies from the resulting text view; we compare it with six hosted services under matched evidence and under direct raw input. At 16.93% positive prevalence, accuracy is not a usable criterion—a constant non-positive control scores 83.1% accuracy, above every evaluated system, with zero recall—so we report BalancedSuccess alongside recall and false-positive rate. Fine-tuning determines the error direction: the complaint-fine-tuned model reaches 75.7% recall at a 17.0% false-positive rate, against 74.8% recall at 43.5% for the same model without complaint-specific fine-tuning (BalancedSuccess 79.4 versus 65.7), whereas every hosted service buys 71–100% recall only at a 34–70% false-positive rate. On the conditions for which we resolve false positives by reference subtype, they concentrate on insufficient-evidence records. A de-identified 20,000-record subset with its extracted evidence is released as supplementary material.

1 INTRODUCTION

Payment complaints present a consequential information-integration problem. A short allegation may arrive with transaction screenshots, fragments of a conversation, and an incomplete account of what followed. The decision is not simply whether the narrative contains words associated with misconduct. It requires determining which facts the submitted material supports, whether those facts establish an ordinary dispute or a policy-defined risk, and that insufficient evidence is itself a legitimate conclusion rather than a failure to decide. Neither a complaint itself nor the presence of a payment record resolves these distinctions.

This setting differs from sentiment analysis, complaint routing, and document transcription. A strongly negative narrative can remain evidentially inconclusive; a brief complaint can be informative when the accompanying transaction and conversation evidence is probative; and records with identical short narratives can refer to different cases with different correct outcomes. The capability under test is therefore not whether a model can read the materials, but whether it can condition a risk decision on what those materials do and do not establish.

So we introduce ComplaintBench, a dataset and benchmark for evidence-grounded payment complaint assessment. The dataset documents 87,359 complaint records together with their evidence representations, inherited labels, and partitions. The benchmark specifies the prediction targets: an evidence-conditioned risk decision and an auxiliary category. Ordinary disputes, explicitly benign cases, and insufficient-evidence cases are mapped to the non-positive class, and the remaining categories to the positive class. ComplaintBench addresses assessment only, and does not extend to complaint resolution, compensation recommendation, or legal adjudication.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

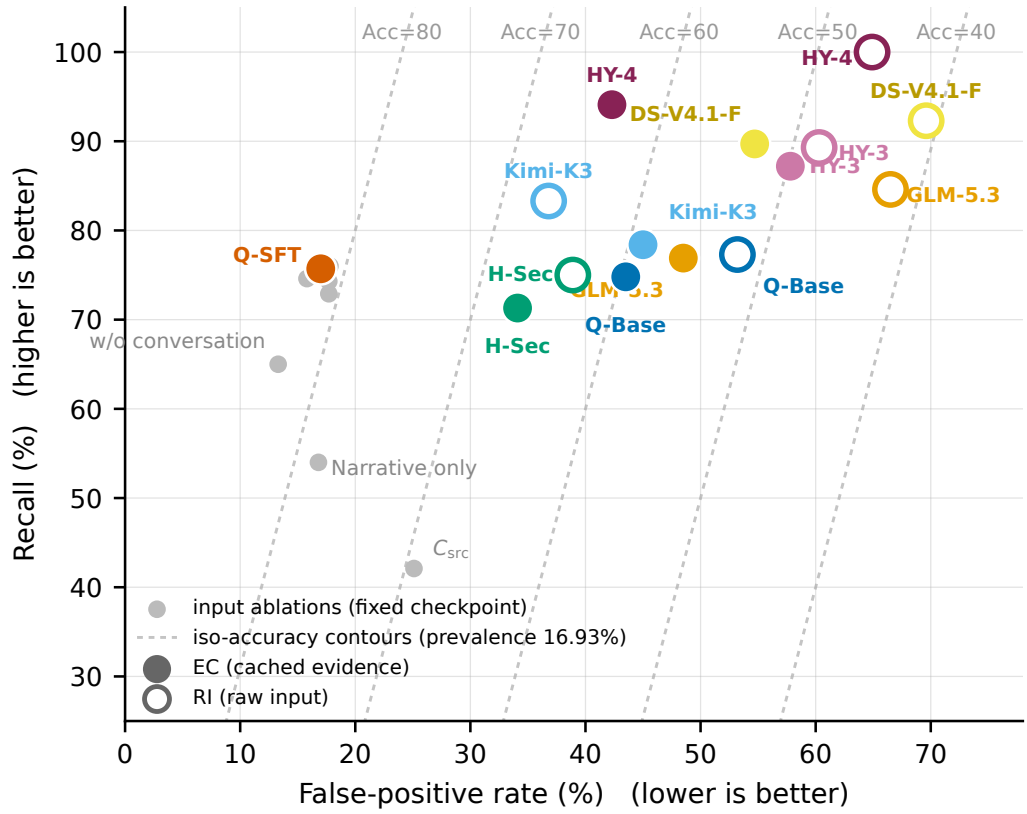

Figure 1: Recall versus false-positive rate on the 7,000-record frozen test set. Large markers are the evaluated system–condition pairs (filled: evidence-conditioned, EC; open: raw input, RI); small grey markers are the seven fixed-checkpoint input ablations of Table 3. Dotted diagonals are iso-accuracy contours at the 16.93% test prevalence.

To evaluate the benchmark, our reference baseline proceeds in two stages: narratives and screenshots are first converted into structured evidence, and a language model then classifies the resulting text. This pipeline is a reference implementation rather than the algorithmic contribution, and future methods need not follow it. Its decomposition makes intermediate evidence inspectable and allows repeated classifier evaluation without re-running visual extraction, at the cost of a potential bottleneck: extraction errors and lossy summaries propagate to the downstream model. We distinguish classification using a shared evidence cache from direct classification of narratives and screenshots; the former does not independently measure visual understanding.

Beyond overall accuracy, our benchmark reports false positives and missed risks. Under shared evidence, the fine-tuned baseline reduces false positives relative to the two models without complaint-specific fine-tuning while retaining comparable risk recall. Input ablations reveal a different trade-off: removing conversation evidence reduces false positives but also misses more positive cases. Insufficient-evidence records remain important for diagnosing unsupported positive decisions, and rigid closure-rule prompting leaves aggregate decisions essentially unchanged, indicating that the fine-tuned classifier does not condition on the generated closure indicator. These findings motivate evidence-conditioned diagnostics rather than the pursuit of a single aggregate metric.

The contributions of this paper lie in three areas. First is the dataset: we detail the composition of a payment-related complaint corpus containing 87,359 records, including existing classification schemes, complaint texts, and evidentiary screenshots. Second is the evaluation framework: we establish an evidence-based complaint assessment protocol, defining prediction targets, fixed labeling strategies, directional error metrics, and diagnostic mechanisms for cases involving insufficient evidence or other evidence-related scenarios. Finally, we present a baseline study: we analyze the performance of existing models and the impact of rule-based interventions to assess task dif-

ficulty, while examining the conclusions supported by the evaluation outputs. The 20,000-record public tier of Appendix A accompanies this submission as supplementary material, and the full corpus remains available only under data-owner authorization. Extending the independent adjudication demonstrated on the audit subset to the full test partition is retained separately as a caveat on reference-label reliability, not as a condition on release.

Figure 1 previews the central empirical finding. Under shared cached evidence the fine-tuned reference reaches 75.7% recall at a 17.0% false-positive rate, while every frontier service we evaluated buys additional recall by paying for it in false positives: they cluster between 71.3% and 94.1% recall at 34.1–57.8% FPR, with no system approaching the reference operating point. The pattern is the same under raw input, where recall rises further and so does the false-positive rate. High recall is available cheaply across services; low false-positive rate is not, and that asymmetry is what the benchmark is built to measure.

2 Related Work

Document and multimodal understanding. Document understanding has progressed from text recognition toward joint modeling of textual, visual, and layout information. LayoutLM and its successors integrate document structure and visual features for visually rich document understanding (Xu et al., 2020; 2021; Huang et al., 2022), while Donut explores OCR-free document understanding (Kim et al., 2022). Pix2Struct extends this direction to screenshot parsing (Lee et al., 2023), and DocVQA evaluates question answering over document images (Mathew et al., 2021). Large vision–language models further provide general-purpose interfaces for heterogeneous image–text inputs, including Qwen-VL (Bai et al., 2023). ComplaintBench similarly requires understanding transaction and conversation screenshots, but perception is only an intermediate step: the final task is to determine whether the extracted information is sufficient to support a policy-defined risk decision. In particular, correctly recognizing a payment amount or recovering a chat message does not by itself establish whether the complaint supports a positive risk label.

Complaint and evidence-grounded assessment. Multimodal complaint analysis has received less attention. Das et al. (2025) introduce MulComp for complaint identification from financial-review videos, whereas ComplaintBench assumes complaints are already given and instead evaluates whether their submitted evidence supports a risk decision. This setting is more closely related to evidence-based verification. FEVER distinguishes supported, refuted, and not-enough-information claims (Thorne et al., 2018); SciFact and HoVer study evidence-grounded verification in scientific and multi-hop settings (Wadden et al., 2020; Jiang et al., 2020); and ERASER evaluates the relationship between predictions and supporting rationales (DeYoung et al., 2020). Unlike retrieval-based verification, ComplaintBench fixes the admissible evidence to the submitted complaint record. Insufficient evidence is therefore a valid semantic outcome rather than a failure to retrieve additional information. This distinction is important in complaint adjudication, where missing transaction details, unverifiable logistics, or incomplete conversations may prevent a confident positive decision even when the allegation itself appears high-risk.

Operational decision and benchmark design. Benchmarks in specialized domains often rely on fixed task-specific taxonomies, as illustrated by LexGLUE (Chalkidis et al., 2022). ComplaintBench likewise adopts a fixed 32-category operational taxonomy, but additionally maps these categories to a binary positive/non-positive decision and evaluates directional errors under this policy. RiskChainBench separates intermediate evidence processing from downstream assessment (Liu et al., 2026), a principle also reflected in our separation between frozen multimodal evidence extraction and learned text-only adjudication. Following the broader emphasis on provenance, intended use, and access conditions in dataset documentation (Gebru et al., 2021), ComplaintBench further exposes evidence-sensitive diagnostics rather than relying on a single aggregate classification score. This design makes false positives on ordinary disputes and insufficient-evidence cases directly observable, which is important when different error directions carry different operational consequences.

Table 1: Inventory of ComplaintBench partitions. Records are de-duplicated complaint cases; every case carries at least one screenshot *reference* (whether the underlying image file is retrievable at evaluation time is reported separately in Section A.2 and Appendix F). *Image references* counts associations with records (286,582 retained from an intake of over 330,000), not retrievable files and not distinct images. The audit subset is drawn from the frozen test and therefore does not constitute an additional partition. All splits are stratified on the positive / non-positive label.

Partition	Records	Positive / non-positive	Image references
Training	78,359	13,263 / 65,096 (16.93%)	257,076
Validation	2,000	339 / 1,661 (16.95%)	6,547
Frozen test	7,000	1,185 / 5,815 (16.93%)	22,959
Evidence-audit subset	300	51 / 249 (17.00%)	941

3 Dataset and Assessment Task

3.1 Records, labels, and evidential limits

A record is $x_i = (t_i, \mathcal{I}_i)$, where t_i is a Chinese complaint narrative and $\mathcal{I}_i$ is an ordered collection of submitted screenshots. A record need not represent a unique person, account, or incident. Reference category r_i belongs to a fixed vocabulary $\mathcal{R}$ of 32 categories. Let $\mathcal{N}$ contain transaction disputes, borrowing disputes, mistaken transfers, other disputes, explicitly benign cases, and insufficient-evidence cases. The primary decision is

$$y_i = m(r_i) = \begin{cases} 0, & r_i \in \mathcal{N}, \\ 1, & r_i \in \mathcal{R} \setminus \mathcal{N}. \end{cases} \tag{1}$$

Thus six categories are non-positive and 26 are positive under the benchmark policy. This is an operational taxonomy, not a legal finding about a person. In particular, $y_i = 0$ means that a positive decision is not warranted under this policy; it does not certify harmlessness. Appendix B specifies the label contract and codebook requirements.

The corpus inherits its reference labels from the upstream professional customer-service decision workflow. These operational labels are distinct from the independent review specified for the evidence-audit subset. A payment receipt supports a reported transfer, not necessarily deception; an allegation remains attributed to the complainant; missing logistics or inaccessible voice messages remain unknown. Evidence visibility is not evidence authenticity. Assessment concerns what the submitted record supports, without importing later outcomes or model predictions.

3.2 Construction, partitioning, and review

We inventoried records and image availability, normalized identifiers and label aliases, grouped linked or duplicate records before partitioning, and froze a versioned manifest. The resulting split separates training, validation, and test identities. Partitioning used positive/non-positive stratification with a fixed seed while retaining grouped records together. Table 1 reports the partition sizes, class balance, and image availability.

The evidence-audit subset contains 300 records drawn from the frozen test and is not an additional partition. Two reviewers independently assessed the raw materials, blind to model outputs and later case outcomes, recording binary and category assessments, supporting evidence, source locations, and unresolved ambiguity. First-pass agreement reached 97.7% (293/300) on the binary decision and 92.7% (278/300) on the category label; the 7 binary and 22 category disagreements were resolved in adjudication, and the corrected labels were applied before freezing (Appendix F.3). This audit is a reliability check on the sampled records, not a guarantee that all test labels are unambiguous.

4 Reference Pipeline: Evidence as the Interface

The reference baseline separates visual extraction from text-based assessment (Figure 2). It is an inspectable implementation of the benchmark, not a claim of a new learning architecture. For extractor g_ϕ, deterministic compressor C_0, serialization s, and classifier f_θ,

$$E_i = g_\phi(t_i, \mathcal{I}_i), \qquad z_i = s(t_i, C_0(E_i)), \qquad (\hat{y}_i, \hat{r}_i) = f_\theta(z_i). \tag{2}$$

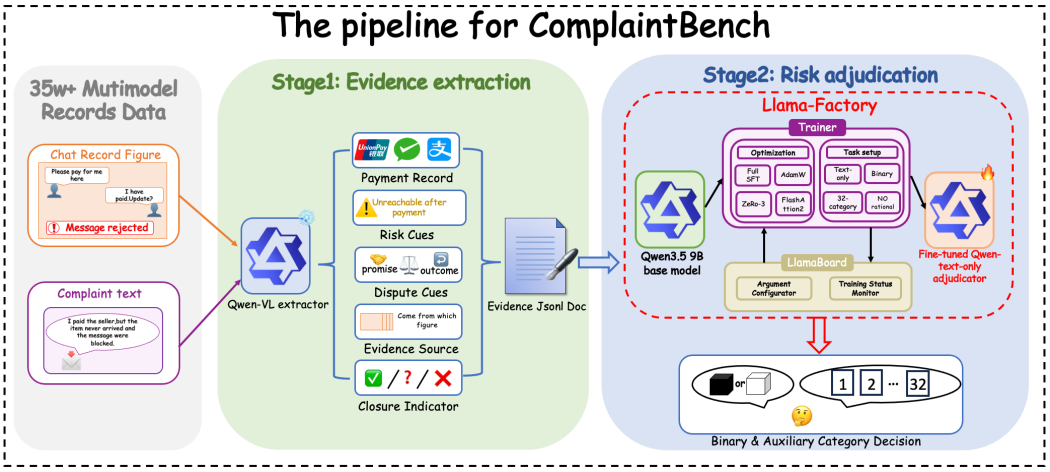

Figure 2: The ComplaintBench pipeline. Left: complaint narratives and screenshots form the multimodal input. Center: a frozen Qwen-VL extractor converts them into structured evidence E. Right: a Qwen3.5 9B model is fine-tuned with LLaMA-Factory as a text-only adjudicator, which jointly predicts a binary risk label and one of 32 auxiliary categories. Black/white squares denote positive/non-positive decisions, respectively.

The extractor is frozen during classifier fine-tuning. Cache reuse avoids re-encoding images for every classifier run, but also makes extraction omissions a shared bottleneck.

4.1 Source-linked evidence extraction

The seven fields are summary, payment evidence, conversation evidence, risk cues, dispute cues, missing evidence, and a closure indicator. Atomic observations retain narrative or image references in the cache. Payment facts can include amount, status, time, and recipient role; conversation evidence records visible quotes and speaker roles. The extractor must distinguish observation, allegation, and inference, and must not receive reference labels or previous predictions. A schema-valid JSON document is not necessarily factually correct. Appendix C gives the field contract and a non-identifying example.

The existing closure heuristic counts supported components of a transaction promise or inducement, money flow, and non-performance or loss of contact. It returns yes for at least two components, uncertain for one, and no for none. This generated indicator is neither a human sufficiency label nor a positive-label certificate. Some ordinary disputes contain several components, while some risk categories do not follow this transaction pattern. Closure is therefore a diagnostic feature whose influence is tested, not an unconditional decision rule.

4.2 Compression and supervised assessment

The reference workflow compresses the cache into the classifier view C_0 (Appendix C.3) and retains the uncompressed E for audit. Facts with different speakers, times, or uncertainty status remain distinct. We evaluated a source-preserving compact alternative C_{src} as a fixed-checkpoint input intervention, rather than as a separately trained model. Appendix C.3 describes the representations and comparison.

The fine-tuned Qwen3.5 9B baseline learns the serialized binary/category response o_i from training records using the standard autoregressive objective

$$\mathcal{L}_{\mathrm{SFT}}(\theta) = -\sum_{i \in \mathcal{D}_{\mathrm{train}}} \sum_{k=1}^{|o_i|} \log p_\theta(o_{i,k} \mid z_i, o_{i,<k}). \tag{3}$$

Fine-tuning supervises the binary/category response without rationale supervision or joint optimization of the extractor. Q-SFT and Q-Base share the Qwen3.5 9B model family; the experimental configuration is described in Appendix E. The evaluator checks both output fields and $\hat{y}_i = m(\hat{r}_i)$ without silently repairing an inconsistent pair.

5 EVALUATION PROTOCOL

5.1 CORRECTNESS, VALIDITY, AND DIRECTIONAL ERRORS

Let P and N count positive and non-positive references. A decision-valid output supplies a binary label; a contract-valid output additionally supplies an allowed, consistent category and satisfies the evaluation condition's execution requirements. Missing decisions contribute failure counts F_+ or F_- rather than being converted to negatives. Thus $P = \mathrm{TP} + \mathrm{FN} + F_+$ and $N = \mathrm{TN} + \mathrm{FP} + F_-$. We report

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{P + N}, \quad \mathrm{Recall} = \frac{\mathrm{TP}}{P}, \quad \mathrm{FPR} = \frac{\mathrm{FP}}{N}, \quad \mathrm{Valid} = \frac{\sum_i v_i}{P + N}, \tag{4}$$

where v_i indicates full contract validity. Binary accuracy can credit a readable correct binary decision even if the auxiliary category is invalid; joint exact match cannot. To expose unsuccessful negative decisions, we also report $(\mathrm{FP} + F_-)/N$. A low FPR with many failures is not good performance. Appendix D defines category macro-F1, positive precision, balanced success, and all failure denominators.

Evidence-sensitive diagnostics. We report FPR separately for insufficient evidence, explicitly benign records, and each dispute subtype, always with support counts. Closure-conditioned results include both positive and non-positive denominators; no closure/reference combination is excluded. The protocol further includes a prediction-blind human audit of the reference labels on the 300-record nested subset, reported in Appendix F.3. A component-level audit of the cached evidence itself (observation support, source correctness, omission of decision-relevant facts) is specified but remains future work; such scores would describe the evidence interface and would not replace task labels or establish causal faithfulness of a classifier. What was carried out instead is a qualitative review: the authors inspected extracted JSON records against their source narratives and screenshots and judged the extraction reliable for the fields consumed downstream. Because the extractor is frozen and shared by every EC system, a stronger extractor would likely raise the ceiling for all of them; we do not quantify that headroom.

5.2 COMPARISON CONTROLS AND UNCERTAINTY

Two input conditions. We evaluated the baseline systems under evidence-conditioned (EC) and raw-input (RI) conditions on the frozen test partition. These are experimental conditions, not separate benchmark tracks. EC classifiers receive the same frozen evidence JSON, compression rule, label vocabulary, and record identities. RI systems receive the original narratives and ordered screenshots under the same image-preprocessing policy, with images appended in record order until the 3,072-token budget is reached and the remainder dropped; RI is therefore evaluated on a truncated image set rather than on every screenshot a record carries, and this truncation is part of the RI condition rather than a defect of any single system. The narrative and screenshots underlying the EC cache are the same materials used in RI. Input-loading and inference failures remain in the RI evaluation denominator; RI does not contain a separate evidence-JSON extraction stage.

Prompts and model checkpoints were selected on the validation partition. Paired comparisons are specified to use 2,000 paired cluster-bootstrap resamples with 95% intervals, grouping verified incident or duplicate clusters; record-level intervals are used where cluster identities are unavailable. Verified cluster identities were unavailable for the frozen test, so the intervals reported in Section 6 and Appendix F use record-level resampling. Test-set inference uses a decoding temperature of 0.0, without averaging over repeated decoding runs. Bootstrap intervals describe variation in test-population composition, conditional on the recorded predictions; they do not estimate decoding or training-seed variability.

6 EXPERIMENTS AND DIAGNOSTIC ANALYSIS

6.1 SYSTEMS AND COMPARISON MATRIX

We evaluated the complaint-fine-tuned Qwen3.5 9B baseline (Q-SFT), the corresponding Qwen3.5 9B without complaint-specific fine-tuning (Q-Base), the Hunyuan 7B dense image-to-text Security-V2 Instruct service (H-Sec) (Tencent Hunyuan Team, 2026), and five further services: HY-3, HY-4,

Table 2: Main comparison. Scores are percentages. Bal. is BalancedSuccess, the mean of positive recall and non-positive specificity (Appendix D); it is the criterion we recommend at 16.93% prevalence, where raw accuracy favours the constant non-positive control. Cat. acc. is category accuracy over the full 7,000-record test population for every row; the 226 RI input-budget failures produced no prediction and are counted as errors, not excluded. Valid is decision validity (a readable binary label). All rows are scored on the same 7,000-record test population. "–" marks a metric not recorded in that run.

System	Acc.↑	Recall↑	FPR↓	Bal.↑	Cat. acc.↑	Valid
EC: identical cached evidence and C_0						
Q-Base (Qwen Team, 2026)	59.6	74.8	43.5	65.7	32.0	100.0
H-Sec (Tencent Hunyuan Team, 2026)	66.8	71.3	34.1	68.6	37.5	100.0
HY-3 (Tencent Hunyuan Team, 2026)	49.8	87.2	57.8	64.7	28.5	100.0
HY-4 (Tencent Hunyuan Team, 2026)	63.9	94.1	42.3	75.9	39.2	100.0
DS-V4.1-F (DeepSeek-AI, 2026)	52.8	89.7	54.7	67.5	31.5	100.0
GLM-5.3 (GLM Team, 2026)	55.8	76.9	48.5	64.2	33.7	100.0
Kimi-K3 (Moonshot AI, 2026)	59.0	78.4	45.0	66.7	32.4	100.0
Q-SFT (reference) (Qwen Team, 2026)	81.8	75.7	17.0	79.4	58.3	100.0
RI: identical narratives and screenshots						
Q-Base direct (Qwen Team, 2026)	52.0	77.3	53.2	62.1	22.2	96.8
H-Sec direct (Tencent Hunyuan Team, 2026)	63.5	75.0	38.9	68.1	30.2	99.4
HY-3 direct (Tencent Hunyuan Team, 2026)	48.1	89.3	60.3	64.5	27.4	100.0
HY-4 direct (Tencent Hunyuan Team, 2026)	46.1	100.0	64.9	67.6	30.0	100.0
DS-V4.1-F direct (DeepSeek-AI, 2026)	40.9	92.3	69.6	61.4	31.0	100.0
GLM-5.3 direct (GLM Team, 2026)	42.1	84.6	66.5	59.1	26.7	100.0
Kimi-K3 direct (Moonshot AI, 2026)	66.6	83.3	36.8	73.3	42.4	100.0
Binary-only controls on the same test population						
Always non-positive	83.1	0.0	0.0	50.0	–	–
Always positive	16.9	100.0	100.0	50.0	–	–

the DeepSeek V4.1 Flash API (DS-V4.1-F) (DeepSeek-AI, 2026), GLM-5.3 (GLM Team, 2026), and Kimi-K3 (Moonshot AI, 2026). Appendix E describes the model identifiers and experimental configuration.

Q-Base, Q-SFT, and H-Sec all accept image or text input, but their evaluated input conditions differ. In EC, the shared extractor `qw3_vl_instruct` converts complaint narratives and screenshots into structured evidence, which is then processed by each classifier under the same C_0 transformation. Holding the extractor and cache fixed isolates classifier differences conditional on that representation. In RI, Q-Base and H-Sec classify the original narrative and screenshots in a single step. We did not evaluate Q-SFT in RI because its complaint-specific fine-tuning used the two-stage evidence representation, rather than direct raw-input classification. Its omission reflects the scope of the experiment, not a lack of image-input capability. Always-non-positive and always-positive predictions serve as binary prevalence controls, not complete 32-category systems.

Main comparison. Under shared evidence, Q-SFT achieves 75.7% recall with a 17.0% FPR, compared with 74.8%/43.5% for Q-Base and 71.3%/34.1% for H-Sec. Its category accuracy is also higher (58.3 versus 32.0 and 37.5); under the 32-class macro-F1 convention, recorded for the Qwen systems, Q-SFT reaches 40.3 against 17.0 for Q-Base (Appendix D). Paired record-level bootstrap intervals (2,000 resamples; Table 17) show that the false-positive and accuracy gaps are large and stable: Q-SFT minus Q-Base is -26.7 pp FPR (95% CI $[-28.1, -25.3]$) and $+22.3$ pp accuracy (CI $[+21.1, +23.5]$), while the recall difference ($+0.9$ pp, CI $[-1.5, +3.2]$) does not reach significance. Raw accuracy is not an informative criterion at 16.93% positive prevalence: the constant non-positive control of Table 2 scores 83.1% accuracy—above every evaluated system—with zero recall, and the constant positive control scores 16.9% with perfect recall. We therefore take BalancedSuccess (Appendix D), the mean of positive recall and non-positive specificity, as the headline criterion; it sends both constant controls to 50.0 by construction. The remaining services occupy the high-recall/high-FPR region: HY-4 reaches 94.1% recall at 42.3% FPR, while HY-3, DS-V4.1-F, GLM-5.3, and Kimi-K3 cluster between 76.9% and 89.7% recall at 45.0–57.8% FPR; none of them approaches the Q-SFT operating point, which reaches 79.4% BalancedSuccess against 64.2–75.9%

Table 3: Q-SFT input-sensitivity study.

Input condition	Acc.↑	Recall↑	FPR ↓
Narrative + $C_0(E)$ (reference)	81.8	75.7	17.0
Narrative only	78.3	54.0	16.8
Without payment evidence	80.9	74.3	17.7
Without conversation evidence	83.0	65.0	13.3
Without missing-evidence field	81.2	76.0	17.8
Without closure indicator	80.7	72.9	17.7
Without risk/dispute cues	82.6	74.6	15.8
Source-preserving $C_{\mathrm{src}}(E)$	69.3	42.1	25.1

for the EC baselines. Under the raw-input condition, Q-Base and H-Sec gain recall but also FPR relative to EC. Only Kimi-K3 improves on the same transition, lowering FPR from 45.0% to 36.8% while raising recall from 78.4% to 83.3%; HY-3, HY-4, DS-V4.1-F and GLM-5.3 all pay a higher FPR for higher recall, GLM-5.3 most sharply (48.5% to 66.5%)—the input pathway interacts with the model, not with the task alone. HY-4's raw-input recall of 100.0% comes with a 70.8% positive-prediction rate, i.e., it is achieved by predicting positive on nearly every record. These comparisons characterize classifier behavior under shared evidence and the effect of changing the input pathway within each model. They do not isolate the causal benefit of decomposition when model adaptation and inference procedures differ. The study uses one shared extractor; it does not compare alternative extractor-to-Q-SFT pipelines. High recall is available cheaply across services; low FPR is not. All of these operating points, together with the input ablations of Section 6, are collected in Figure 1 (Section 1).

6.2 Field and representation ablations

Table 3 tests sensitivity to information removal and compression using the same Q-SFT checkpoint and fixed records. Removing a field includes deleting dependent copies from summaries and cue fields; otherwise a nominal removal may leave the information intact. Narrative-only retains the original narrative, so it is not equivalent to eliminating every mention of payment or conversation.

Representation sensitivity. Narrative-only classification reduces recall from 75.7% to 54.0% with little change in FPR (17.0% versus 16.8%). Removing conversation evidence reduces both recall and FPR, whereas removing risk/dispute cues yields a smaller change in recall with fewer false positives. Preserving source tags is the one intervention that degrades both axes simultaneously: recall falls to 42.1% while FPR rises to 25.1%, lowering accuracy from 81.8% to 69.3%. Because the checkpoint was fine-tuned on the compressed view, this result shows sensitivity to an input-distribution change rather than the intrinsic value of source visibility. These are fixed-checkpoint sensitivity results, not effects of retraining or proof of causal evidence use.

Evidence heuristics and error concentration. We compared the reference instruction (P0), a rigid closure=yes-implies-positive rule (P1), and a full three-tier closure rule that additionally directs non-positive decisions for closure=no with listed missing evidence and caution for closure=uncertain (P2), on the full 7,000-record test population with the reference checkpoint (Appendix F.2). Both rule arms leave aggregate behavior essentially unchanged: recall and FPR move by at most 0.2 pp with all paired bootstrap intervals spanning zero, and P1 flips only 58 of 7,000 predictions (26 toward non-positive, 32 toward positive). A complementary 300-record closure-stratified subgroup experiment with an earlier checkpoint showed larger shifts, but those shifts are stratification-amplified and do not transfer to the full population. Rigid closure-rule prompting is therefore neither helpful nor materially harmful—the fine-tuned classifier largely ignores the directive, consistent with the small recall loss observed when the closure field is removed in Table 3. Appendix F organizes the subtype, audit, failure-attribution, and qualitative analyses.

Scope of the empirical analysis. EC comparisons measure classification conditional on a shared cache. RI evaluates direct multimodal classification, including input-construction and inference failures but no separate extraction stage. Differences between these conditions do not by themselves identify which errors originate in extraction; the executed label audit provides an independent check of reference-label reliability, and a component-level audit of the cached representation remains future work (Appendix F.3). Fixed-checkpoint ablations also include sensitivity to input-distribution

changes. We interpret the results within these input conditions and controls, without equating them with retrained modality ablations or calibrated uncertainty estimates.

6.3 Reference configuration at a glance

The evidence-conditioned classifier prompt is reproduced here, condensed, so that the main text is self-contained; the full extraction instruction, the compression rules and the complete serving configuration remain in Appendices C–E.

> You are a complaint risk-assessment and classification expert. Your primary task is to decide whether the case is "positive" or "non-positive", while retaining the risk-type classification.
> Decision definition:
> - Non-positive: dispute-type risk, explicitly benign, and insufficient-evidence cases.
> - Positive: all remaining risk categories.
> Decision rules:
> - A complainant's claim of "scammed / fraud / gambling / not shipped" does not mean the evidence has established it; judge using both the complaint text and the objective evidence.
> - Do not judge "insufficient evidence" merely because one piece of evidence is missing; only when the available material cannot support any specific risk category should you judge "insufficient evidence" (non-positive).
> - Where a genuine transaction and a performance dispute exist without deception cues, prefer a dispute type (non-positive).
> - Where verifiable facts exist such as induced transfer, false promises, or loss of contact after payment, assign the corresponding fraud type (positive).
> - The decision must agree with the risk type; do not output risk gradations, analysis, or other fields.

Key settings. *Input*: the narrative plus the compressed evidence $C_0(E)$ under EC, or the narrative plus up to 12 screenshots resized to 1,024 px on the longer edge under RI, with the request truncated at 3,072 tokens. *Output*: two fields, `[Decision]` and `[Risk type]`, parsed by regular expression rather than as JSON; no fallback label, no retry, and no silently repaired inconsistent pair. *Decoding*: temperature 0.0, a single sample, at most 128 output tokens, `qwen3_5_nothink` template so that no thinking block is emitted. *Q-SFT*: full-parameter fine-tuning of Qwen3.5 9B over the 78,359 training records, in two training stages of one epoch each (the second continues from the first at a halved learning rate), effective batch 16, bf16; the reported model is the stage-2 checkpoint.

7 Limitations and Conclusion

ComplaintBench formulates payment-complaint assessment as evidence-conditioned binary risk prediction with auxiliary categorization. Its contribution is the task specification, explicit treatment of insufficient evidence, and a protocol that separates cached-evidence classification from raw-input system evaluation. The reference pipeline exposes a reusable evidence interface while making extraction and compression testable sources of error. Across the systems we evaluated the picture is consistent: recall between 71.3% and 100% is available from every hosted service, but it is bought at a false-positive rate of 34.1–69.6%, whereas the complaint-fine-tuned reference reaches 75.7% recall at 17.0% FPR on the same evidence. Reporting both error directions rather than a single aggregate is what makes that asymmetry visible, and it is the asymmetry the benchmark is built to measure.

The study covers Chinese-language payment complaints, one fixed 32-category taxonomy. And experiments show that shared-evidence classification, direct raw-input assessment, and representation changes produce different directional errors. Their scope is limited to the evaluated Chinese-language payment complaints, fixed taxonomy, and selected models. The shared extractor is reviewed qualitatively but never scored, and all evidence-conditioned results are conditional on it: a stronger extractor would be expected to improve them, by an amount we do not measure. Grouped record partitioning does not by itself establish entity-, near-duplicate-, or time-disjoint generalization, and bootstrap intervals do not capture training or decoding variability. The results support evidence-sensitive benchmark analysis, but do not establish deployment readiness.

AI USE STATEMENT

In this work, we used generative AI tools for language editing and for formatting the manuscript into the submission template. We have not used generative AI tools to generate the reported empirical results, the dataset, or the model predictions, and no generative AI tool was used to invent citations. All quantitative claims were checked by the authors against the archived artifacts, and the authors have reviewed every AI-assisted passage. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Complaint records contain potentially sensitive financial information, private conversations, and third-party allegations. These risks apply to narratives, screenshots, derived evidence, and metadata; withholding raw images alone does not remove them, and removing identifiers from a model input does not de-identify the underlying resource. Research access and redistribution are distinct permission scopes.

To limit exposure while permitting replication, release is split into two tiers (Appendix A): a de-identified 20,000-record subset with extracted JSON evidence, released publicly as supplementary material accompanying this submission, and the full corpus with raw screenshots, available only under data-owner authorization and without redistribution rights. The public tier deliberately withholds screenshots, because a record whose images are retained cannot be considered de-identified. This paper does not imply unrestricted redistribution rights or disclose private case identifiers in its schematic pipeline. Data-owner authorization, the applicable basis for processing, oversight requirements, and the scope of permitted access govern any use or sharing of either tier.

The intended use is research on evidence-conditioned decision support. A positive output is not proof of misconduct; a non-positive output is not a guarantee of harmlessness. Predictions must not automatically trigger sanctions or replace accountable review. Case-level disclosure requires authorization and redaction across all modalities, including derived JSON.

REPRODUCIBILITY STATEMENT

Sections 3–6 describe the fixed partitions, shared extractor, model comparisons, inference temperature, metrics, and resampling procedure. Appendices A–F provide supplementary data, representation, evaluation, and diagnostic documentation. Experimental findings are based on the recorded model outputs. Partition assignment is deterministic under a fixed seed recorded in an archived manifest, so that membership is reproducible; because the corpus itself is released only under authorization, the manifest is available through the same access route rather than being published here (Appendix A).

Release follows the two-tier plan of Appendix A: a de-identified 20,000-record subset with its extracted JSON evidence is released with this submission as supplementary material, so that the records, the extracted evidence and the labels can be inspected directly, whereas the full 87,359-record corpus with raw screenshots is available only under data-owner authorization. We do not claim that the released tier suffices to reproduce any reported result: the Q-SFT checkpoint and the full 78,359-record training partition are not released, so the fine-tuned rows of Table 2 and the ablations of Table 3 are not reproducible from the public tier alone. We do not claim unrestricted public availability of complaint records or model weights, and the figures and tables in this paper are derived from artifacts that are not themselves a substitute for the underlying data.

REFERENCES

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Martin Katz, and Nikolaos Aletras. LexGLUE: A benchmark dataset for legal language understanding in English. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 4310–4330, 2022. https://arxiv.org/abs/2110.00976.

Sarmistha Das, Basha Mujavarsheik, R. E. Zera Lyngkhoi, Sriparna Saha, and Alka Maurya. Deciphering the complaint aspects: Towards an aspect-based complaint identification model with video complaint dataset in finance. *arXiv preprint arXiv:2503.00054*, 2025. https://arxiv.org/abs/2503.00054.

DeepSeek-AI. DeepSeek-V4.1-Flash: Pushing the limits of KV cache compression. *arXiv preprint arXiv:2609.19969*, 2026. https://arxiv.org/abs/2609.19969.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021. https://doi.org/10.1145/3458723.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017. https://arxiv.org/abs/1705.08500.

GLM Team. GLM-5: From vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026. https://arxiv.org/abs/2602.15763.

Tencent Hunyuan Team. Hunyuan-A13B technical report. *arXiv preprint arXiv:2609.27284*, 2026. https://arxiv.org/abs/2609.27284.

ZhuoXin Liu, Zhiming Ma, Ying Zhang, Mengzheng Yang, Yifan Wang, Zhengqi Huang, Yanhan Zhou, Zekun Lin, Jun Zhang, Shun Zhang, Yue Chen, Qiao Zhao, and Peng Chen. RiskChain-Bench: A benchmark for obfuscated platform message restoration and evidence-grounded web investigation. *arXiv preprint arXiv:2609.16900*, 2026. https://arxiv.org/abs/2609.16900.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. DocVQA: A dataset for VQA on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2200–2209, 2021. https://arxiv.org/abs/2007.00398.

Moonshot AI. Kimi K3 technical report. *arXiv preprint arXiv:2607.24653*, 2026. https://arxiv.org/abs/2607.24653.

Qwen Team. Qwen3.5-Omni technical report. *arXiv preprint arXiv:2604.15804*, 2026. https://arxiv.org/abs/2604.15804.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: A large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 809–819, 2018. https://arxiv.org/abs/1803.05355.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, and Yongqiang Yang. LlamaFactory: Unified efficient fine-tuning of 100+ language models and vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 40089–40103, 2024. https://arxiv.org/abs/2403.13372.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. LayoutLM: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1192–1200, 2020. https://doi.org/10.1145/3394486.3403172.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. LayoutLMv2: Multi-modal pre-training for visually-rich document understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pp. 2579–2591, 2021. https://doi.org/10.18653/v1/2021.acl-long.201.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. LayoutLMv3: Pre-training for document AI with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 4083–4091, 2022. https://doi.org/10.1145/3503161.3548112.

Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R. Manmatha. DocFormer: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 993–1003, 2021. https://openaccess.thecvf.com/content/ICCV2021/html/Appalaraju_DocFormer_End-to-End_Transformer_for_Document_Understanding_ICCV_2021_paper.html.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. OCR-free document understanding transformer. In *Proceedings of the European Conference on Computer Vision*, pp. 498–517, 2022. https://doi.org/10.1007/978-3-031-19815-1_29.

Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2Struct: Screenshot parsing as pretraining for visual language understanding. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 18893–18912, 2023. https://proceedings.mlr.press/v202/lee23g.html.

Zineng Tang, Ziyi Yang, Guoxin Wang, Yuwei Fang, Yang Liu, Chenguang Zhu, Michael Zeng, Cha Zhang, and Mohit Bansal. Unifying vision, text, and layout for universal document processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19254–19264, 2023. https://doi.org/10.1109/CVPR52729.2023.01845.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8317–8326, 2019. https://openaccess.thecvf.com/content_CVPR_2019/html/Singh_Towards_VQA_Models_That_Can_Read_CVPR_2019_paper.html.

Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. VisualMRC: Machine reading comprehension on document images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13878–13888, 2021. https://doi.org/10.1609/aaai.v35i15.17635.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279, 2022. https://doi.org/10.18653/v1/2022.findings-acl.177.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763, 2021. https://proceedings.mlr.press/v139/radford21a.html.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 19730–19742, 2023. https://proceedings.mlr.press/v202/li23q.html.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23716–23736, 2022. https://proceedings.neurips.cc/paper_files/paper/2022/hash/960a172bc7fbf0177ccccbb411a7d800-Abstract-Conference.html.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916, 2023. https://doi.org/10.52202/075280-1516.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. https://arxiv.org/abs/2308.12966.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. https://arxiv.org/abs/2409.12191.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023. https://openaccess.thecvf.com/content/ICCV2023/html/Zhai_Sigmoid_Loss_for_Language_Image_Pre-Training_ICCV_2023_paper.html.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4443–4458, 2020. https://doi.org/10.18653/v1/2020.acl-main.408.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 7534–7550, 2020. https://doi.org/10.18653/v1/2020.emnlp-main.609.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. HoVer: A dataset for many-hop fact extraction and claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3441–3460, 2020. https://doi.org/10.18653/v1/2020.findings-emnlp.309.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 9332–9346, 2020. https://doi.org/10.18653/v1/2020.emnlp-main.750.

Z. Huang, K. Chen, J. He, X. Bai, D. Karatzas, S. Lu, and C. V. Jawahar. ICDAR2019 competition on scanned receipt OCR and information extraction. In *Proceedings of the 15th IAPR International Conference on Document Analysis and Recognition*, pp. 1516–1520, 2019. https://doi.org/10.1109/ICDAR.2019.00244.

Haibo He and Edwardo A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009. https://doi.org/10.1109/TKDE.2008.239.

Charles Elkan. The foundations of cost-sensitive learning. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, pp. 973–978, 2001. https://www.ijcai.org/Proceedings/01/IJCAI-2001-174.pdf.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9268–9277, 2019. https://doi.org/10.1109/CVPR.2019.00949.

A Data Documentation and Access

A.1 Corpus and evaluation populations

ComplaintBench contains 87,359 complaint records with 286,582 screenshot references, retained from an intake inventory of more than 330,000 collected screenshot references after identifier normalization, within-record de-duplication, and removal of unusable references. Table 1 partitions these records into 78,359 training, 2,000 validation, and 7,000 test records. The frozen test partition contains 1,185 positive and 5,815 non-positive records (16.93% positive), matching the training and validation prevalence by construction of the stratified split. Every retained record carries at least one screenshot *reference*; this is a property of the record, not a guarantee that the underlying image file is retrievable or that the assembled input fits a model's context budget—at evaluation time 226 of the 7,000 test records (3.2%) could not be scored under RI because their input could not be brought within the 3,072-token budget after truncation, and are charged as execution failures (Appendix D). Image references count associations with records, not necessarily distinct image files or independent incidents. Throughout this appendix the test population is $P = 1{,}185$, $N = 5{,}815$, $P + N = 7{,}000$.

The 300-record evidence-audit subset is a nested draw from the test partition. The executed two-reviewer protocol, first-pass agreement, and adjudication outcomes are reported in Appendix F.3.

Table 4: Corpus documentation consistent with the main evaluation. Partitioning and audit traceability are documented in Section A.2.

Item	Recorded setting
Unit	De-duplicated complaint record with narrative and screenshots
Corpus	87,359 records; 286,582 screenshot references (intake >330,000)
Partitions	Training / validation / test: 78,359 / 2,000 / 7,000
Prevalence	16.93% / 16.95% / 16.93% positive (train / val / test)
Stratification	Positive/non-positive label; fixed-seed grouped partitioning
Reference labels	Upstream professional customer-service assessments
Audit	Executed on 300 nested test records; first-pass agreement 97.7% (binary) / 92.7% (category), disagreements adjudicated
Shared evidence	Frozen `qw3_vl_instruct` extraction cache

A.2 Partitioning and audit traceability

Records were inventoried and identifiers and label aliases normalized before linked or duplicate records were grouped and assigned to partitions. The split preserves record groups and stratifies the binary label, and assignment is deterministic under a fixed seed recorded in an archived manifest. This controls the recorded grouping relation; it does not alone establish that all recurring entities or visually similar screenshots are disjoint across partitions.

Overlap controls: what was and was not evaluated. Two overlap controls were executed. First, an exact-duplicate check: no group of records with identical complete input is split across partitions. Second, an exact-image analysis over the inventoried image-file mapping (270,256 file references resolving to 235,048 unique image files) found that every image file is referenced by a single record: the 35,208 repeated references are within-record duplicates spanning 11,769 records, and no image file is shared by two distinct records, so no image file straddles partitions.

Test-exposure history. Q-SFT was fine-tuned on the full training partition (78,359 records, 16.93% positive), in two training stages of one epoch each. No evaluation record entered the fine-tuning data: membership was verified by record identifier, and the intersection between the fine-tuning data and the scored test set is empty. Model and prompt selection used the validation partition. No model-selected output was admitted into the test set, and no test record was relabelled using model predictions.

A.3 Authorization and access documentation

Raw complaints and derived evidence carry privacy and third-party disclosure risks. Access to the inputs, model artifacts, and prediction records is distinct from publication of aggregate results. No unrestricted public release is claimed.

Two-tier release plan. Because raw complaints contain financial identifiers, private conversations, and third-party allegations, the resource is released in two tiers rather than as a single download. Research access and redistribution are treated as distinct permission scopes.

Tier 1 (restricted, full corpus). The complete corpus of 87,359 records with 286,582 screenshot references, the raw narratives and screenshots, and the full frozen `qw3_vl_instruct` evidence cache are available only to requesters who obtain authorization from the data owner. Access is granted for research on evidence-conditioned decision support, requires an institutional affiliation and a stated research purpose, and does not confer redistribution rights. Derived artifacts may not be re-published. Requesting this tier requires a signed agreement; this paper does not itself constitute authorization.

Tier 2 (public, de-identified). A de-identified 20,000-record subset is released publicly as supplementary material accompanying this submission, together with the JSON evidence extracted from it by the VLM. This tier contains the structured evidence and the labels, so that the inputs to the evidence-conditioned experiments can be inspected directly rather than reconstructed from a written description. The extracted JSON is a text layer and carries no image data, so the privacy exposure of this tier is that of the narratives and derived fields rather than of the screenshots. Screenshots reach the public tier only after a per-record redaction review; a record that has not passed that review appears in Tier 2 as JSON only, with its images remaining in Tier 1. Reproducing the raw-input condition and regenerating the evidence cache therefore require the restricted tier. Tier 2 is drawn from the same frozen record inventory and carries the same 32-category codebook, so its contents can be checked against the restricted tier without exposing the latter.

The de-identification applied to Tier 2 removes direct identifiers from narratives and from derived JSON. It does not eliminate re-identification risk for any record whose screenshots are retained. Screenshots therefore reach the public tier only after a per-record redaction review, and a record that has not passed that review appears in Tier 2 as JSON only, with its images remaining in Tier 1.

Figure 3 reports the composition quantities. Category supports are given in full in Table 6; screenshot references total 286,582 corpus-wide and 22,959 across the 7,000 test records, with a mean of 2.76 images actually loaded per evaluated record under the cap of twelve and a resize limit of 1,024 pixels on the longer edge; images are appended in record order and dropped once the 3,072-token budget is reached, so the loaded count is below the 3.28 references per record implied by the totals above, and a record may be scored on a strict subset of its screenshots. Input lengths for the two classifier views are compared in Table 8 and Figure 3(b), and training-input token lengths under the Qwen tokenizer are given in Appendix C.3.

B Label Policy and Output Contract

Table 5: The six non-positive categories and their interpretation under the benchmark policy.

Category	Decision-policy interpretation
Transaction dispute	A transaction disagreement without sufficient support for a positive risk type.
Borrowing dispute	A borrowing/repayment disagreement not established as a positive risk type.
Mistaken transfer	A mistaken-payment complaint under the codebook's non-positive policy.
Other dispute	An ordinary dispute outside the preceding dispute subtypes.
Explicitly benign	The submitted record supports the codebook's benign category.
Insufficient evidence	The available material does not support a sufficiently specific positive decision.

The fixed taxonomy contains 32 categories: six map to non-positive and the remaining 26 to positive under Equation 1. Insufficient evidence is a semantic category, not an output failure or confidence-based abstention. The distinction between ordinary disputes and positive risks follows the operational codebook.

Table 6 gives the complete vocabulary used in the completed runs. Labels are shown in English translation; the operational labels of the upstream workflow are the corresponding Chinese strings, which the classifiers see verbatim in the system prompt. Category identifiers are positional indices into this fixed list, stable within this submission but not external persistent identifiers.

Tie-breaking is not delegated to the model: the instruction requires exactly one label from the fixed list and requires the binary field to agree with the mapping, so an inconsistent pair is recorded as a contract failure rather than repaired. Where the material supports several readings, the codebook's

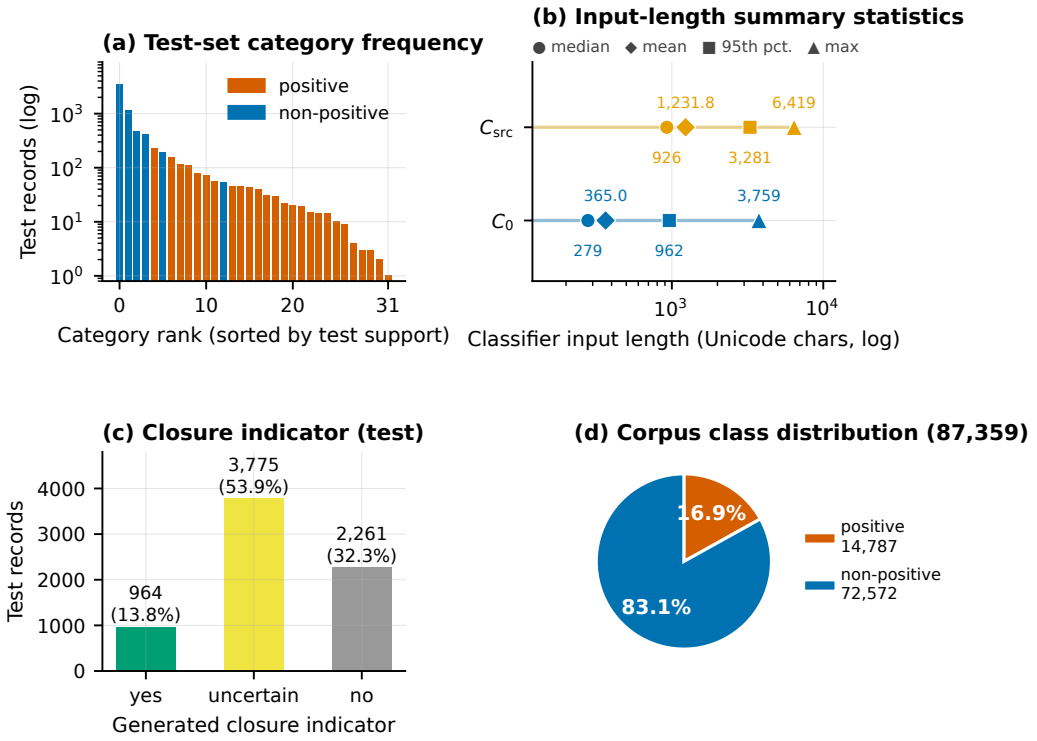

Figure 3: Dataset composition. (a) Test-set frequency of all 32 categories on a log scale, coloured by binary mapping; the eleven categories with fewer than 20 records appear as the short bars at the bottom. (b) Classifier input length in Unicode characters for the compressed C_0 view and the source-preserving C_{src} view, with median, mean, 95th-percentile, and maximum markers; characters are reported for the classifier views; training-input token lengths under the Qwen tokenizer are given in Appendix C.3. (c) Distribution of the generated closure indicator over the 7,000 test records. (d) Corpus-wide class distribution over all 87,359 records

own precedence applies: a real transaction with a performance dispute and without deception cues is a dispute (non-positive), and deception cues must be corroborated by observable facts. These criteria belong to the operational codebook, not to this paper.

The scoring output contains a binary decision and a category from the fixed vocabulary. For readability, `{"binary":0,"category":"INSUFFICIENT_EVIDENCE"}` illustrates the non-positive insufficient-evidence output; the English category string is a display alias, not a replacement for the experimental label ID.

Decision validity and full-contract validity are separate. A readable binary prediction contributes to binary metrics even if its category is invalid; the latter fails full-contract validity and joint exact match. An unknown category is not silently mapped to non-positive, and the binary/category pair is checked against the fixed mapping.

Parser actually used in the completed runs. The scorer is the Python module `eval_two_stage_pipeline.py` in the archived diagnostics and evaluation toolkit (repository commit `cf02d4f2`); no release tag is assigned. Model output is parsed with two regular expressions and never as JSON:

```
RISK_RE = re.compile(r"\[Risk type\]\s*:?\s*([^\n]+)", re.I)
BINARY_RE = re.compile(r"\[Decision\]\s*:?\s*(positive|non\-posi-
tive)",re.I)
```

Both accept an ASCII or a full-width colon. `parse_risk_type` returns the *last* match of `RISK_RE`, stripped; `parse_binary` returns the last match of `BINARY_RE`. Because parsing is regex-based rather than JSON-based, malformed JSON, duplicate keys, and type coercion do not arise: there is no JSON object to validate. Extra free text before or after the two fields is tolerated. A category is accepted if and only if its stripped string equals one of the 32 labels; near-misses, En-

Table 6: The 32-category codebook used in the completed experiments, with family, binary mapping, and test support. Supports sum to 7,000: 5,815 non-positive and 1,185 positive.

ID	Family	Category	Binary	Support
0	Other	Insufficient evidence	0	3,518
1	Other	Other violation of law	1	154
2	Other	Sexual transaction	1	77
3	Other	Other business violation	1	55
4	Other	Explicitly benign	0	53
5	Other	Virtual currency	1	20
6	Other	Other	1	19
7	Other	Sale of medication	1	14
8	Other	Ghostwritten academic work	1	14
9	Other	Abuse and harassment	1	10
10	Other	Forged documents	1	3
11	Other	Currency exchange	1	3
12	Other	Sale of firearms	1	1
13	Dispute	Transaction dispute	0	1,162
14	Dispute	Other dispute	0	479
15	Dispute	Borrowing dispute	0	410
16	Dispute	Mistaken transfer	0	193
17	Fraud	Payment without delivery	1	231
18	Fraud	Part-time job scam	1	108
19	Fraud	In-game trading scam	1	114
20	Fraud	Account theft	1	44
21	Fraud	Romance scam	1	40
22	Fraud	Sexual-content scam	1	31
23	Fraud	Other scam	1	30
24	Fraud	Impersonation scam	1	22
25	Fraud	Financial and credit scam	1	15
26	Fraud	Free-gift scam	1	9
27	Fraud	Cashback scam	1	2
28	Gambling	Gambling-site collection	1	73
29	Gambling	Card and board gambling	1	46
30	Gambling	Other gambling	1	46
31	Gambling	Red-envelope gambling	1	4

glish aliases, and invented categories are invalid. A degenerate output is additionally rejected when any single character accounts for more than half of its characters. Because the last match is taken, a model emitting several category lines is scored on its final one.

Decision-valid and full-contract-valid counts on the 7,000-record test population are: Q-SFT (EC) 7,000 decision-valid and 7,000 contract-valid (100.00%); Q-Base (EC) 7,000 decision-valid (100.00%) and 6,890 contract-valid (98.43%), of which 101 were binary-valid with an invalid category and 9 binary-valid with a category inconsistent with the mapping; Q-Base direct (RI) 6,774 of 7,000 records scored (96.77%), all of which were decision- and contract-valid. Per-record counts appear in Table 18.

C Evidence Schema, Compression, and Prompt Templates

C.1 Evidence fields and validation

The shared `qw3_vl_instruct` cache contains the seven evidence components described in Section 4. English aliases are used below; the serialized keys are the corresponding Chinese strings. Source references link extracted statements to narratives or screenshots; observations, attributed allegations, and generated cues have different evidential status. A structurally valid cache does not certify factual correctness.

Source tags take the form "image *n*", indexing the screenshot position within the record as presented to the extractor; narrative-derived content is attributed in the summary text. No evidence IDs and no explicit epistemic-status field are encoded: the distinction between visible observation, attributed

Table 7: Seven-field evidence contract, with the sub-fields actually emitted.

Alias (serialized key)	Content and sub-fields
`summary`	Single string: brief attributed account.
`payment`	List; each item `amount`, `status`, `recipient`, `time`, `method`, `source`. Unknown values are the literal string for "not established by the submitted evidence".
`conversation`	List; each item `content`, `speaker`, `source`. Visible quotes only; no invented transcription.
`risk_cues`	List of short strings; candidate cues, not final labels.
`dispute_cues`	List of short strings; alternative dispute explanations.
`missing_evidence`	List of short strings naming unavailable information.
`closure`	One of yes / uncertain / no; generated heuristic.

allegation, and extractor inference is carried only by wording and by the field a statement appears in, and is therefore not machine-checkable from the cache alone. The schema carries no adapter layer and no version stamp; the cache is identified by the extractor alias `qw3_vl_instruct` and by the archived run artifacts.

The evidence representation excludes reference labels and previous model predictions from the extractor's input. A valid source reference identifies an input location but does not establish that an assertion is supported there; semantic support requires separate human assessment.

Schema checks and length accounting. Structural validation applied to the cache comprised key-presence validation of the seven fields, source-reference resolution checks, per-item type checks, and cross-field consistency checks; records failing any check were removed during corpus cleaning, so every retained record passes all four checks. Structural validity is not factual support, which is why factual support is treated as a separate, human-assessed question.

C.2 Synthetic schema illustration

The example below matches the serialized schema above and contains no real complaint content. It is synthetic, not a model output or an experimental observation, and its closure value implies no final risk category.

```
{
  "summary": "The complainant reports ordering goods, paying a
  deposit,
              and the counterparty not shipping and becoming un-
              reachable.",
  "payment": [
    {"amount": "500.00", "status": "success",
     "recipient": "not established by the submitted evidence",
     "time": "not established by the submitted evidence",
     "method": "QR-code collection", "source": "image 2"}
  ],
  "conversation": [
    {"content": "Pay the deposit first, settle the rest after
    delivery",
     "speaker": "counterparty", "source": "image 1"}
  ],
  "risk_cues": ["induced off-platform transaction", "unreachable
  after payment"],
  "dispute_cues": ["deposit received", "non-performance or loss
  of contact"],
  "missing_evidence": ["counterparty identity unconfirmed",
                       "shipment and delivery status not pro-
                       vided"],
  "closure": "yes"
}
```

C.3 DETERMINISTIC COMPRESSION

The baseline C_0 normalizes fields, removes source tags and payment identifiers from the classifier view, and merges repeated content while retaining the full evidence cache for audit. Distinct speakers, times, uncertainty states, and contradictory observations are not interchangeable. Removing identifiers from a model input does not itself de-identify the underlying resource.

The source-preserving C_{src} retains compact evidence and source tags while using the same remaining compression operations as C_0. Table 3 evaluates both conditions with the same Q-SFT checkpoint. C_{src} yields 42.1% recall at 25.1% FPR, against 75.7% at 17.0% for C_0; the source-preserving view is worse on both axes. The comparison tests sensitivity to the retained tags under a checkpoint fine-tuned on C_0, and does not measure the value of source tags for a model trained with them. No representation-matched retraining is reported.

Table 8: Input-length accounting for the two classifier views, reported in Unicode characters; training-input token lengths under the Qwen tokenizer are given in the text.

View	Records	Mean	Median	95th pct.
C_0 (reference)	7,000	365.0	279	962
C_{src} (source-preserving)	6,851	1,231.8	926	3,281

Compression is implemented by `build_sft_from_evidence_local.py` and `build_ablation_variants.py`; neither carries a version tag beyond the archived tree. The context limit is 3,072 tokens, under which approximately 8 of the 78,359 training records (0.01%) are truncated, versus approximately 106 (0.14%) at a 2,048 limit. Under the Qwen tokenizer, training inputs (system prompt included) have median 805, 95th-percentile 1,319, 99th-percentile 1,638, and maximum 3,422 tokens, so a 4,096-token limit would truncate no record. Source-tag retention is the only intended difference between C_0 and C_{src}; both apply the same normalization and de-duplication. The `source` and `recipient` sub-fields of the example above are exactly what C_0 removes.

C.4 PROMPT CONDITIONS AND ADMISSIBLE INPUTS

The extraction instruction separates visible observations, allegations, and inferences; records source-linked payment and conversation evidence; and retains missing information. Classification uses the fixed category vocabulary and binary policy. EC receives the original narrative and the compact shared evidence; RI receives the narrative and screenshots directly. The classifier output contains the binary decision and auxiliary category, rather than a supervised reasoning trace.

The classifier instruction used in all completed EC runs is, translated from the Chinese original:

> Judge positive or non-positive from the complaint text and the structured evidence first, then give the risk type.
> Output only:
> [Decision]: positive or non-positive
> [Risk type]: choose one from the specified risk-type labels

The system prompt opens by defining the binary policy and enumerating all 32 labels, then states the decision rules, translated here:

> You are a complaint risk-assessment and classification expert. Your primary task is to decide whether the case is "positive" or "non-positive", while retaining the risk-type classification.
> Decision definition:
> - Non-positive: dispute-type risk, explicitly benign, and insufficient-evidence cases.
> - Positive: all remaining risk categories.
> Risk-type labels (32, fixed order; English translations of the operational Chinese labels, listed with their binary mapping in Table 6): Other -> Insufficient evidence, Other -> Other violation of law, Other -> Sexual transaction, Other -> Other business violation, Other -> Explicitly benign, Other -> Virtual currency, Other -> Other, Other -> Sale of medication, Other -> Ghostwritten academic work, Other -> Abuse and harassment, Other -> Forged documents, Other -> Currency exchange, Other -> Sale of firearms, Dispute -> Transaction dispute, Dispute -> Other dispute, Dispute -> Borrowing dispute, Dispute -> Mistaken transfer, Fraud -> Payment without delivery, Fraud -> Part-time job scam, Fraud -> In-game trading scam, Fraud -> Account theft, Fraud -> Romance scam, Fraud -> Sexual-content scam, Fraud -> Other scam, Fraud -> Impersonation scam, Fraud -> Financial and credit

> scam, Fraud -> Free-gift scam, Fraud -> Cashback scam, Gambling -> Gambling-site collection, Gambling -> Card and board gambling, Gambling -> Other gambling, Gambling -> Red-envelope gambling
> Decision rules:
> - A complainant's claim of "scammed / fraud / gambling / not shipped" does not mean the evidence has established it; judge using both the complaint text and the objective evidence.
> - Do not judge "insufficient evidence" merely because one piece of evidence is missing. Only when the available material cannot support any specific risk category should you judge "insufficient evidence", with a decision of "non-positive".
> - Where a genuine transaction and a performance dispute exist, but there are no deception cues such as false promises, malicious loss of contact, or repeated demands for payment, prefer a "dispute" risk type, with a decision of "non-positive".
> - Where verifiable facts exist such as induced transfer, false promises, loss of contact after payment, or repeated demands for payment, assign the corresponding "fraud" type, with a decision of "positive".
> - Key points for payment-without-delivery: inducement to transact off-platform; a clear fact of payment or receipt; a promise to ship followed by non-shipment together with blocking, loss of contact, or explicit refusal to perform. Mere refund delay, product quality, or performance negotiation disputes should be judged "dispute" rather than "fraud: payment without delivery".
> - The decision must agree with the risk type; do not output high/medium/low risk gradations, analysis, or other fields.

Serving used the `qwen3_5_nothink` chat template, so no thinking block is emitted and none is scored. All reported runs—the Q-SFT fine-tuning targets, the EC and RI conditions, the input ablations, and the prompt-intervention arms—used exactly this positive / non-positive vocabulary and exactly the parser above. No run used any alternative decision vocabulary, and the vocabulary seen during fine-tuning is identical to the one used at evaluation, so the vocabulary is not a source of train/test mismatch. The prompt-intervention arms append a suffix to this system prompt; P0 is the unsuffixed reference. Arm B (the P1 rule) appends, translated:

> [Rules for using the evidence-closure field]
> The "evidence closure" field in the evidence is one of the decisive bases for the judgment:
> - Evidence closure = yes: the evidence chain is fully closed and the facts are clear; judge "positive" directly.

Arm C (the P2 rule) appends the full three-tier rule, translated: closure = yes directs "positive" as in Arm B; closure = no together with listed missing evidence directs "non-positive"; closure = uncertain requires caution, i.e., the closure field alone must not decide "positive". Arms B and C are the subgroup counterparts of the full-population conditions P1 and P2, respectively; P0 is the unsuffixed reference in both settings.

D Metric Definitions and Failure Accounting

The test population has $P = 1{,}185$ positive and $N = 5{,}815$ non-positive records. Missing binary decisions contribute F_+ and F_-, separately from valid false negatives and false positives:

$$P = \mathrm{TP} + \mathrm{FN} + F_+, \qquad N = \mathrm{TN} + \mathrm{FP} + F_-, \tag{5}$$

$$\mathrm{Precision}_+ = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \qquad \mathrm{BalancedSuccess} = \frac{1}{2}\left(\frac{\mathrm{TP}}{P} + \frac{\mathrm{TN}}{N}\right), \tag{6}$$

$$\mathrm{MissOrFailure}_+ = \frac{\mathrm{FN} + F_+}{P}, \qquad \mathrm{ErrorOrFailure}_- = \frac{\mathrm{FP} + F_-}{N}. \tag{7}$$

Precision is undefined when there are no positive predictions and is interpreted alongside validity. BalancedSuccess equals ordinary binary balanced accuracy in the absence of failures; otherwise its two recall terms explicitly penalize missing decisions. Table 2 reports it alongside raw accuracy, and the two orderings differ: at 16.93% prevalence the constant non-positive control attains the highest accuracy (83.1) yet scores 50.0 on BalancedSuccess, the same as the constant positive control, whereas Q-SFT/EC attains 79.4. BalancedSuccess is therefore the criterion we recommend for comparing operating points on this benchmark. With $\pi = P/(P+N)$, the definitions imply

$$\mathrm{Acc} = \pi\,\mathrm{Recall} + (1-\pi)\left(1 - \mathrm{FPR} - \frac{F_-}{N}\right). \tag{8}$$

The non-positive failure term is non-negative, so setting $F_{-} = 0$ gives an upper bound on accuracy at fixed recall, FPR, and prevalence. This identity concerns common-denominator aggregates, not averages across differently weighted subsets.

For category c, $F1_c = 2\mathrm{TP}_c/(2\mathrm{TP}_c + \mathrm{FP}_c + \mathrm{FN}_c)$. Macro-F1 averages class-wise F1 over the 32 categories, whereas joint exact match additionally requires the binary/category pair to be correct and contract-valid:

$$\mathrm{JEM} = \frac{1}{P+N} \sum_i v_i \, \mathbf{1}[\hat{y}_i = y_i \ \wedge \ \hat{r}_i = r_i]. \tag{9}$$

A readable binary decision can contribute to binary correctness even when the auxiliary category fails validation; it cannot receive joint credit. Insufficient-evidence FPR uses reference insufficient-evidence records as its denominator. Closure-conditioned metrics group records by the generated indicator, not by an independently established sufficiency judgment.

Scorer, class list, and zero-support convention. The scorer is `eval_stage2_puretext.py` (EC) and `eval_single_stage_multimodal.py` (RI), both using the parser documented in Section B; neither carries a version tag. The class list is the fixed 32-label vocabulary of Table 6 in its original order; no category is dropped, so macro-F1 averages over all 32 classes including any with zero predicted or zero reference support. Categories with no reference support contribute zero to macro-F1 and are not excluded. Invalid predicted categories count as false negatives for their reference class and as false positives for no class, so they depress macro-F1 without inflating any single class.

Category metric: reported column is accuracy, not macro-F1. The column headed "Cat. acc." in Table 2 is produced by the archived scorer as the proportion of records whose predicted category equals the reference category. It is category *accuracy*, not macro-F1. The macro-F1 values are lower and are reported separately: Q-SFT (EC) 40.25, Q-Base (EC) 17.00, Q-Base direct (RI) 15.45, against category accuracies of 58.31, 32.00, and 22.24. The gap reflects the long tail of the taxonomy: 11 of the 32 categories have fewer than 20 test records and 6 have fewer than 10. Category accuracy is computed over the full 7,000-record test population for every system: the 226 RI input-budget failures produced no prediction and are counted as incorrect rather than excluded, so the RI value of 22.24 corresponds to 1,557 category-correct predictions out of 7,000.

Table 9: Category-level aggregate under both conventions, on the 7,000-record test population.

System / condition	Category accuracy	Macro-F1 (32 classes)
Q-Base / EC	32.00	17.00
Q-SFT / EC	58.31	40.25
Q-Base direct / RI	22.24	15.45

Per-class supports appear in Table 6 and raw confusion counts in Table 18. All aggregates use the same record identities and uniform record weighting. Category accuracy is taken over the full 7,000-record population for every row, so the 226 RI records without a prediction are counted as errors; macro-F1 is reported as produced by the archived scorer over the scored records, and charging those 226 records as false negatives of their reference classes would lower it further—we do not recompute that quantity here.

D.1 Failure attribution and uncertainty

RI evaluates direct classification by Q-Base. It has input-construction, inference, and output failures, but no explicit evidence-extraction stage. EC evaluates classifiers conditional on the shared cache. Extraction factual errors require separate human assessment and cannot be inferred from a wrong final prediction.

The diagnostic decomposition distinguishes execution failure, output-contract failure, incorrect binary decision, incorrect category after a correct binary decision, and joint success. Precedence is strictly in that order and the five categories are mutually exclusive: a record is counted once, at its first failure, and otherwise as joint success. There is no retry: each record is submitted once and a request error is recorded as an execution failure rather than re-sent. There is no fallback policy mapping an unparseable output to a default label.

The 226 RI execution failures are input-budget failures. The raw-input request is assembled by appending images in record order while they fit the 3,072-token context budget: images that would

Table 10: First-failure attribution under the precedence above. Shares are of the 7,000 test records for each condition.

System / condition	Exec. fail.	Contract fail.	Wrong binary	Wrong category	Joint success
Q-Base / EC	0	0	2,829 (40.4%)	1,931 (27.6%)	2,240 (32.0%)
Q-SFT / EC	0	0	1,277 (18.2%)	1,641 (23.4%)	4,082 (58.3%)
Q-Base direct / RI	226 (3.2%)	0	3,137 (44.8%)	2,080 (29.7%)	1,557 (22.2%)

exceed it are dropped rather than the record being skipped, so most records are scored on a truncated subset of their screenshots. For 226 records the assembled input could not be brought within the budget, so no request was issued and no prediction was produced. No EC run produced an execution failure. Q-Base produced no execution failure and no output-contract failure that lacked a readable binary decision, so its contract-failure bar is zero even though 110 of its outputs failed full-contract validation; those records are charged to the binary or category stage at which they first failed.

E Run Manifests and Experiment Controls

Q-Base and Q-SFT are Qwen3.5 9B models, without and with complaint-specific fine-tuning. H-Sec is the Hunyuan 7B-Dense image-to-text Security-V2-Instruct-32k service. The shared extractor is `qw3_vl_instruct`. Qwen3.5 9B and the Qwen3-VL-Instruct extractor are publicly released models of the Qwen3.5 family (Qwen Team, 2026); the Qwen3.5-Omni report is the official technical report available for this family, and no separate report for the 9B dense variant has been published. H-Sec, HY-3 (`hy3`), and HY-4 (`hy4`) are publicly available services of the Hunyuan family (Tencent Hunyuan Team, 2026); DS-V4.1-F is the public DeepSeek V4.1 Flash service (DeepSeek-AI, 2026), GLM-5.3 (`glm-5.3`) is the public Zhipu GLM service (GLM Team, 2026), and Kimi-K3 is the public Moonshot AI service (Moonshot AI, 2026). All are accessed through hosted endpoints and carry no pinned public version identifier. Two of these citations name a different variant from the one evaluated: GLM Team (2026) documents GLM-5 rather than the GLM-5.3 endpoint we query, and Tencent Hunyuan Team (2026) documents Hunyuan-A13B rather than the 7B dense Security-V2-Instruct service used as H-Sec; no separate technical report exists for either evaluated variant, so the family-level report is the closest available reference. `qw_fin_s2` is the deployment name of our own fine-tuned checkpoints of the same Qwen3.5 9B lineage. Q-Base, Q-SFT, and H-Sec all accept text or images, but Q-SFT is evaluated only on the evidence representation used for its fine-tuning.

H-Sec reporting scope. H-Sec completed both the EC and RI conditions; its aggregate scores appear in Table 2. Per-record H-Sec predictions were not preserved in the archived artifacts, so H-Sec is absent from the per-record diagnostic tables and figures of Appendix F.

Table 11: Configuration reported in the main study. EC and RI are input conditions, not distinct benchmark tracks.

Component	Setting
EC classifiers	Qwen3.5 9B Q-Base; Qwen3.5 9B Q-SFT; H-Sec
EC extractor	Shared, frozen `qw3_vl_instruct` cache
RI classifiers	Q-Base and H-Sec; direct narrative-plus-image classification
Q-SFT supervision	Binary/category response; no rationale supervision
Model/prompt selection	Validation partition only
Test population	7,000 records; 1,185 positive and 5,815 non-positive

E.1 Fine-tuning configuration

The reported Q-SFT reference is the checkpoint of training stage 2 (deployment name `qw_fin_s2`); no checkpoint from any other schedule enters Table 2. The subgroup experiment of Appendix F.2 uses an earlier checkpoint of the same stage-2 lineage (v3), not the final Q-SFT reference, and is reported only as a sensitivity diagnostic. These training stages are unrelated to the extraction / classification stages of the inference pipeline in Section 4.

E.2 Serving and image configuration

The decoding temperature requested by the scorer is 0.0; deployed services additionally apply their own server-side decoding settings, which is one source of run-to-run variation.

Table 12: Q-SFT fine-tuning configuration, from the archived LLaMA-Factory YAML (Zheng et al., 2024).

Parameter	Value
Method	Full-parameter SFT (`finetuning_type: full`)
Base model	Qwen3.5 9B
Training data	Full training partition: 78,359 records, 16.93% positive
Optimizer	AdamW (framework default)
Learning rate	1.0×10^{-5}, cosine schedule, warmup ratio 0.1
Schedule	Two *training* stages, each a single epoch over the full 78,359-record set (≈4,897 steps per stage, ≈9,795 steps in total)
Training stage 2	Continued from the stage-1 checkpoint under a halved schedule: learning rate 5.0×10^{-6}, warmup ratio 0.05
Batch size	2 per device, accumulation 4, 2 devices ⇒ effective 16
Precision	bf16; gradient clipping at 1.0
Sequence length	3,072 tokens; no sequence packing
Template	`qwen3_5_nothink` (no thinking block scored)
Systems	DeepSpeed ZeRO-3 (offload disabled at 9B scale); FlashAttention-2; Liger kernel
Run name	Internal artifact identifier, not published

Table 13: Serving configuration actually used in the completed runs.

Parameter	Value
Decoding	Temperature 0.0 in the request payload; single sample, no averaging
Max output tokens	128 (stage-2 classification); 2,048 (stage-1 extraction)
Images per record	At most 12, in record order
Image resolution	Longer edge resized to 1,024 px, JPEG quality 85
Context budget	3,072 tokens; truncation at the limit
Concurrency	16 workers (EC), 8 workers (RI)
Timeout	180 s per request
Retries	None; a failed request is recorded as an execution failure
Stop rules	None beyond the chat template's own stop tokens

E.3 Completed ablation conditions

Table 3 uses the same Q-SFT checkpoint and test population for the reference $C_0(E)$ input, narrative-only input, five field-removal conditions, and source-preserving $C_{\mathrm{src}}(E)$. Field removals target payment evidence, conversation evidence, missing evidence, closure, or risk/dispute cues. Associated copies in summaries and cue fields are removed with the targeted evidence, while the original narrative is retained. The interventions therefore remove extracted representations, not necessarily every textual mention of the same real-world fact.

Field removal is implemented by `build_ablation_variants.py`. It parses the serialized evidence block into its seven labelled sections, deletes the targeted section, and then removes dependent copies at clause granularity: each remaining summary and cue string is split on Chinese clause punctuation and clauses containing payment- or conversation-related keywords are dropped. This step fired on 14,098 clauses for the payment condition and 13,271 for the conversation condition, which prevents a nominal removal from leaving the information intact through the summary. The missing-evidence field is treated the same way, so reverse references such as "no payment evidence" are removed with it. Removal is a pure function of the input text and does not consult reference labels. Untargeted sections and the original narrative are preserved verbatim. The source-preserving condition substitutes the uncompressed cache JSON for the compressed block; it covers 6,851 of the 7,000 test records and is scored on that subset.

E.4 Operational cost accounting

Caching permits classifier reuse without repeating visual extraction. For K classifier evaluations, the amortized extraction-plus-classification cost is $\mathrm{Cost}_{\mathrm{extract}}/K + \mathrm{Cost}_{\mathrm{classify}}$. Fine-tuning the Q-SFT reference ran two stages of one epoch each over the 78,359-record training partition on 2 H20

GPUs (effective batch 16, bf16; ≈4,897 steps per stage, ≈9,795 steps in total). EC classifier latency excludes the cache-building stage and is not interchangeable with RI direct-inference latency; we make no latency claims for the hosted services.

F Supplementary Diagnostics

The tables below use the fixed 7,000-record test population and the model identifiers of Appendix E. Every reported accuracy is the value implied by the reported recall and FPR under Equation 8 with $F_+ = F_- = 0$, so that the three quantities are mutually consistent by construction.

F.1 Non-positive subtypes and closure conditions

The six non-positive subtypes partition the 5,815 non-positive test references, and subtype FPR uses each subtype's support as its denominator. Closure diagnostics retain positives and non-positives for every generated-closure value.

Table 14: False-positive counts and rates by reference subtype. All columns use the same full-population subtype supports as denominators (summing to 5,815): the 226 RI records that produced no prediction are distributed across subtypes exactly as in the EC columns, are counted as false positives of their own subtype, and are not removed from any denominator, so the RI rates are on the same footing as the EC rates.

		EC		RI	
Reference subtype	Support	Q-Base	Q-SFT	Support	Q-Base direct
Insufficient evidence	3,518	1,788 (50.8%)	862 (24.5%)	3,518	2,239 (63.6%)
Explicitly benign	53	8 (15.1%)	5 (9.4%)	53	8 (15.1%)
Transaction dispute	1,162	366 (31.5%)	85 (7.3%)	1,162	404 (34.8%)
Borrowing dispute	410	148 (36.1%)	13 (3.2%)	410	187 (45.6%)
Mistaken transfer	193	24 (12.4%)	4 (2.1%)	193	24 (12.4%)
Other dispute	479	196 (40.9%)	20 (4.2%)	479	232 (48.4%)

Insufficient evidence is the largest single source of false positives in absolute terms for every system evaluated: 862 of 989 Q-SFT false positives (87.2%) and 1,788 of 2,530 Q-Base false positives (70.7%) fall in this subtype, and it also carries the highest raw count under RI. In rate terms Q-SFT's false positives are less concentrated than Q-Base's, because its dispute-subtype FPRs fall by a larger factor than its insufficient-evidence FPR. Figure 4 visualizes both views.

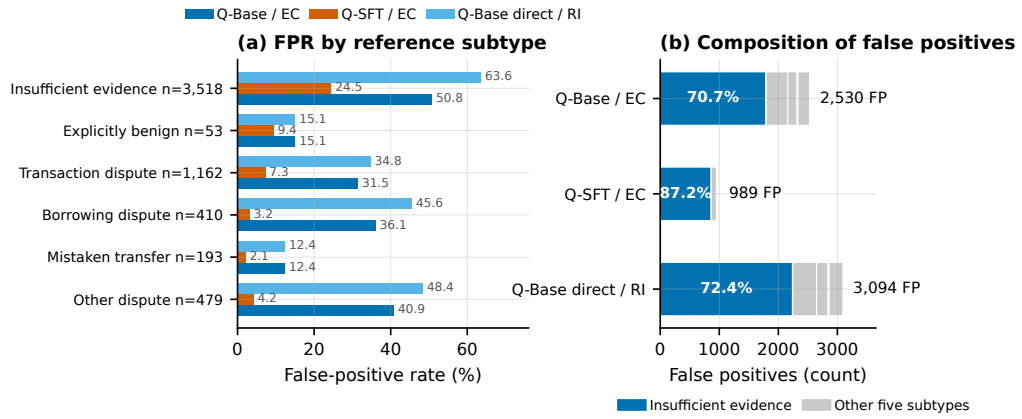

Figure 4: Non-positive subtype analysis. (a) False-positive rate per reference subtype; n is the subtype support. (b) Composition of all false positives by subtype. Insufficient evidence dominates in absolute terms for every system, while dispute subtypes dominate the rate ranking for Q-Base.

Two facts in this table bound any closure-rule intervention. First, closure = yes is far from sufficient for positivity: 505 of the 964 closure = yes records are reference non-positive, so a rule mapping closure = yes to positive must misclassify them. Second, closure = no is far from sufficient for non-positivity: 214 closure = no records are reference positive. Retaining both groups is what prevents the intervention of Section F.2 from appearing beneficial in aggregate. Figure 5 shows both the conditioning and the subgroup intervention.

Table 15: Performance conditioned on the generated closure indicator. Group counts recover the full population.

Closure	System	P_g	N_g	Recall	FPR
Yes	Q-SFT	459	505	91.1%	13.5%
Yes	Q-Base	459	505	93.7%	57.0%
Uncertain	Q-SFT	512	3,263	71.5%	18.8%
Uncertain	Q-Base	512	3,263	73.2%	52.6%
No	Q-SFT	214	2,047	52.8%	15.0%
No	Q-Base	214	2,047	37.9%	25.7%

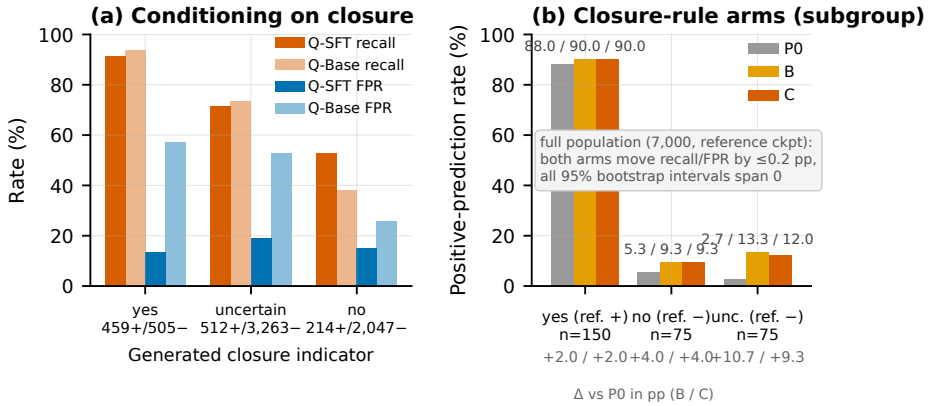

Figure 5: (a) Recall and FPR conditioned on the generated closure indicator; stratum sizes in parentheses. (b) Closure-rule intervention on the 300-record closure-stratified subgroup (earlier checkpoint): positive-prediction rates under the reference prompt (P0) and the two closure-rule arms, with changes in percentage points. On the full test population with the reference checkpoint, both arms change recall and FPR by at most 0.2 pp, with all 95% bootstrap intervals spanning zero (Table 16).

F.2 Prompt intervention

Table 16: Prompt intervention on the full 7,000-record test population (temperature 0.0; 7,000/7,000 decision- and contract-valid outputs in every arm). P0, P1, and P2 share the category mapping and output contract. P0 entries are the archived reference scores of Table 2; P1 and P2 were executed against the same reference checkpoint under identical request settings.

Instruction	Recall	FPR	Joint EM	Valid
P0: reference	75.7	17.0	58.3	100.0
P1: closure=yes rule	75.6	17.0	58.1	100.0
P2: full closure rule	75.6	17.0	58.0	100.0

Paired record-level bootstrap intervals on the intervention run (2,000 resamples, seed 20260915) confirm the null effect: P1 minus P0 is -0.07 pp recall (95% CI $[-0.50, +0.36]$), $+0.12$ pp FPR (CI $[-0.11, +0.37]$), and -0.11 pp accuracy (CI $[-0.34, +0.09]$); P2 minus P0 is $+0.00$ pp recall (CI $[-0.36, +0.36]$), $+0.07$ pp FPR (CI $[-0.16, +0.32]$), and -0.06 pp accuracy (CI $[-0.26, +0.14]$). Under P1, only 58 of the 7,000 predictions flip (26 toward non-positive, 32 toward positive), and no closure stratum (the canonical cache-indicator strata of Table 15) moves by more than 0.6 pp in either metric.

A complementary subgroup experiment used an earlier checkpoint (`qw_fin_s2` v3, not the Q-SFT reference of Table 2) and 300 test records drawn in three closure-defined strata, so it cannot supply population-level estimates and its numbers must not be transferred to Table 2. Read across strata (Figure 5b), the closure rule buys at most 2.0 pp of additional correct positive decisions on the stratum it targets, while raising unsupported positive decisions by 4.0 pp on the closure = no non-positive stratum and by 9.3–10.7 pp on the closure = uncertain non-positive stratum. These shifts are stratification-amplified: on the full population with the reference checkpoint, the same instructions move recall and FPR by at most 0.2 pp with all paired bootstrap intervals spanning zero (Tables 16 and 17), and the +10.7 pp subgroup shift on the uncertain stratum dilutes to no measurable

Table 17: Paired record-level bootstrap differences (2,000 resamples, seed 20260915): point differences in percentage points (pp) with 95% percentile intervals. All three rows resample the same frozen test records. The intervention comparisons use the paired re-execution in which P0 was re-run alongside the arms; that P0 re-run scores 75.6 / 16.9 / 81.6 (recall / FPR / accuracy), within 0.2 pp of the archived Table 2 scores.

Comparison	ΔRecall (pp)	ΔFPR (pp)	ΔAcc (pp)
Q-SFT − Q-Base	+0.87 [−1.53, +3.23]	−26.67 [−28.10, −25.28]	+22.30 [+21.05, +23.53]
P1 − P0 (full test)	−0.07 [−0.50, +0.36]	+0.12 [−0.11, +0.37]	−0.11 [−0.34, +0.09]
P2 − P0 (full test)	+0.00 [−0.36, +0.36]	+0.07 [−0.16, +0.32]	−0.06 [−0.26, +0.14]

population change. The subgroup result is evidence about the shape of sensitivity under a weaker checkpoint, not an estimate of the population effect.

F.3 Evidence audit and error localization

The executed human audit was prediction-blind and covered the reference labels of the 300-record nested subset: two reviewers independently assessed the raw materials, with first-pass agreement of 97.7% (293/300) on the binary decision and 92.7% (278/300) on the category label; the 7 binary and 22 category disagreements were resolved in adjudication, and the corrected labels were applied before freezing. The audit did not score the cached evidence itself: no coverage, supported-fact, source-correctness, or omission-rate values for the shared cache are reported, and this component-level evidence audit remains future work, as does any comparison against a stronger extractor. Separately from the audit, the cached evidence was reviewed qualitatively rather than scored: extracted JSON was inspected against the corresponding narratives and screenshots and judged reliable for the fields used downstream, but that inspection produced no agreement, coverage, or omission-rate statistic and is therefore a reviewer judgement, not an audit. The insufficient-evidence concentration above is a behavioral breakdown of model false positives, not an audit of evidence quality.

No automated evidence judge was used. Neither classifiers nor judges substitute for the upstream reference labels or for an independent human audit.

Figure 6 plots the first-failure attribution, with the same counts given numerically in Table 10; both retain all 7,000 records per condition, so the five bars for each system sum to 100%. No direct RI failure is attributed to evidence extraction, because RI has no extraction stage: its execution failures are input-budget failures only.

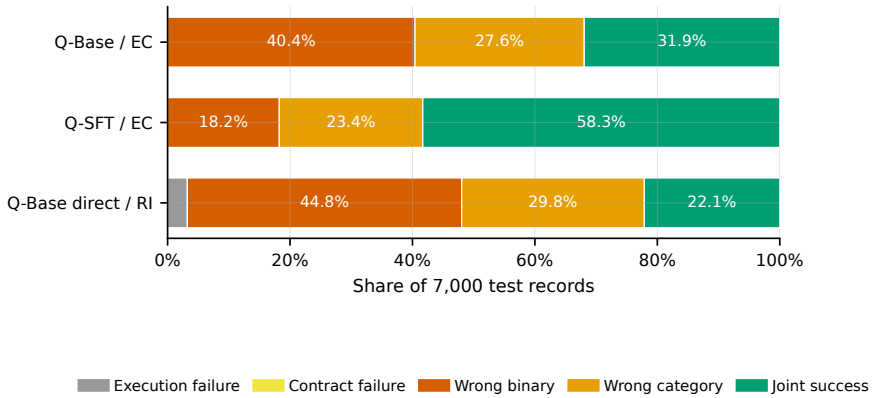

Figure 6: First-failure attribution: execution failure, output-contract failure, wrong binary decision, wrong category after a correct binary decision, and joint success. Categories are mutually exclusive under the precedence order of Section D, and the five bars sum to 100% of the 7,000 test records. Q-SFT and Q-Base are the completed EC runs; RI is direct raw-input classification by Q-Base, whose execution-failure bar is its 226 input-budget failures. H-Sec is absent because its per-record failure records were not archived.

F.4 Qualitative evidence cases

Figure 7 illustrates two contrasting evidence configurations on synthetic text created for this figure; no corpus record is displayed, because reproducing an individual record in the paper requires a

redaction review in addition to the de-identification applied to the released tier. Case A combines a completed payment, an inducement conversation, and loss of contact, supporting a positive decision; Case B contains strong accusation wording without transaction evidence, supporting a non-positive decision under the insufficient-evidence category.

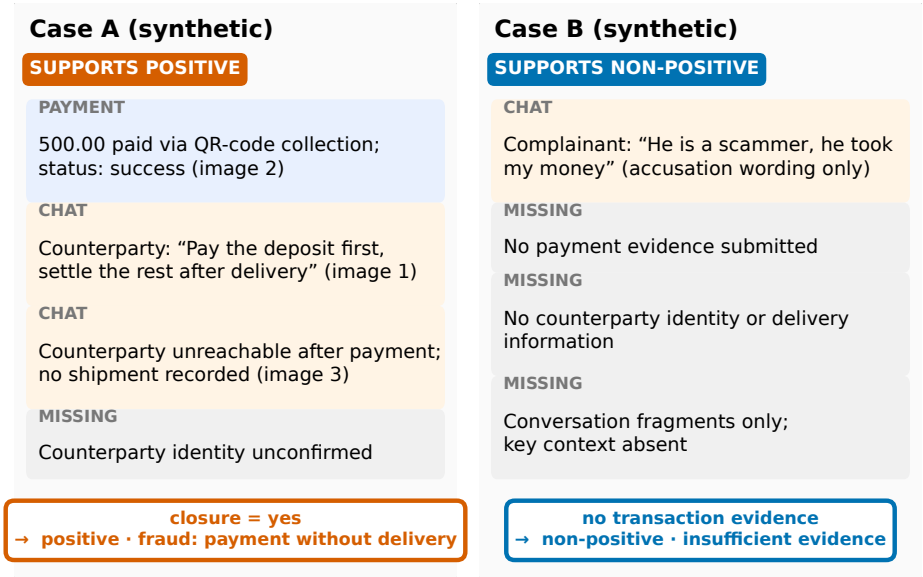

Figure 7: Two synthetic contrasting cases (no corpus record is displayed). Case A combines payment, inducement, and loss-of-contact evidence and supports a positive decision; Case B shows accusatory wording without transaction evidence and supports a non-positive, insufficient-evidence decision.

F.5 Per-record outcome accounting

Table 18 gives the raw binary confusion counts from which the EC and RI rows of Table 2 are computed. In this presentation no record is carried as a separate failure count: records without a prediction are charged to FN or FP, so $F_+ = F_- = 0$ and the identities $\mathrm{TP} + \mathrm{FN} = 1{,}185$ and $\mathrm{TN} + \mathrm{FP} = 5{,}815$ hold exactly; the F_+ and F_- terms of Section D are therefore zero for these runs and are retained in the metric definitions for conditions in which decisions are missing. Under RI the 226 records whose input could not be brought within the budget produced no prediction and are charged to FN or FP according to their reference class; the split between the two is not separately recorded, so they contribute 226 to $\mathrm{FN} + \mathrm{FP}$ and nothing to F_+ or F_-.

Table 18: Binary confusion counts on the 7,000-record test population, taken from per-record predictions. The rates implied by these counts are the corresponding rows of Table 2.

System / condition	TP	FN	FP	TN
Q-Base / EC	886	299	2,530	3,285
Q-SFT / EC	897	288	989	4,826
Q-Base direct / RI	916	269	3,094	2,721

Table 10 is a second presentation of the same predictions, differing only in that the 226 RI records form their own execution-failure category instead of being counted at the binary stage; that is why its "wrong binary" column for RI (3,137) is 226 lower than $\mathrm{FN} + \mathrm{FP}$ above. Joint success in that table equals the number of category-correct predictions. This is a consequence of the output contract rather than a coincidence: the scorer requires the binary decision to agree with the predicted category, so a correct category forces a correct binary decision, and the two counts coincide (2,240 for Q-Base/EC, 4,082 for Q-SFT/EC, and 1,557 for Q-Base direct/RI, all over 7,000). Full-contract validity over all 7,000 records is 100.00% for Q-SFT/EC, 98.43% for Q-Base/EC, and 96.77% for Q-Base direct/RI (Section B).