

REIMBURSEBENCH: Benchmarking LMMs as Full-Workflow Auditors for Financial Internal Control

Libin Zhan^{1*}, Wuyou Mao^{1*}, Xiang Li¹, Enci Zhang¹, Qi Yang¹,
Shujian Zhu¹, Yingtian Hu¹, Wentao Ye², Haobo Wang^{2,3},
Ziman Li¹, Jie Zhang⁴, Wang Dong¹, Jun Chen^{1†}

¹Institute of Finance and Accounting (ZIFA), School of Management, Zhejiang University

²State Key Laboratory of Blockchain and Data Security, Zhejiang University

³Innovation and Management Center,

School of Software Technology (Ningbo), Zhejiang University

⁴China Telecom Artificial Intelligence Technology (Beijing) Co. Ltd

{zhanlibin, chenjun332}@zju.edu.cn

Abstract

Expense reimbursement internal control is a high-stakes corporate finance task that combines multimodal, visually rich evidence (e.g., receipts and approval forms), firm-specific policy text, and multi-step compliance reasoning. No existing benchmark evaluates large multimodal models (LMMs) in this domain. We introduce **REIMBURSEBENCH**, the first benchmark targeting full-workflow multimodal audit reasoning in corporate internal control. It comprises 151 expert-constructed reimbursement workflows, each packaging an application, receipts and supporting documents, and a firm-specific policy, evaluated by a process-level rubric grounded in twelve auditing logic points. Benchmarking nine frontier and locally deployable systems, we find the strongest attains only 72.5%, and no model exceeds 47.7% on hard workflows. Beyond this aggregate score, risk-aware, process-level evaluation surfaces three structural findings: (i) a persistent gap between binary admissibility judgment and precise quantitative computation; (ii) a systematic lenient bias in monetary predictions that poses deployment risk; and (iii) a performance pattern consistent with compositional-generalization difficulties under realistic workflow load. We provide **REIMBURSEBENCH** as a testbed for advancing LMMs toward trustworthy financial systems. Data and code: <https://github.com/ikerzhan/ReimburseBench>.

1 Introduction

Expense reimbursement internal control is among the most labor-intensive yet error-sensitive operations in corporate finance. A single reimbursement workflow requires an internal-control auditor to parse a free-form expense application, in-

spect a heterogeneous bundle of supporting evidence (e.g., scanned invoices, tickets, approval forms, and attendee rosters, some with hierarchical tabular layouts), cross-reference each piece against firm-specific reimbursement policies spanning dozens of clauses, and render a binary admissibility decision together with a precise upper bound on the legally reimbursable amount. Errors on either side carry concrete monetary consequences: over-reimbursement causes direct financial loss and compliance exposure, while under-reimbursement damages employee trust. This task—multi-hop reasoning over visually rich evidence anchored in firm-specific regulatory text—is a natural and high-stakes testbed for the latest large multimodal language models (e.g., Anthropic, 2026; OpenAI, 2026).

Despite this fit, existing benchmarks evaluate only fragments of the required capability. Financial NLP benchmarks (e.g., Chen et al., 2021) are anchored in publicly disclosed capital-market documents and treat finance as a text-only QA problem. Visual information extraction benchmarks (e.g., Wang et al., 2026) parse individual receipts but never integrate extracted evidence with external policy texts. Recent enterprise workflow benchmarks (Dong et al., 2025; Wu et al., 2026) target spreadsheet-centric finance & accounting work and GUI-level data entry respectively—neither evaluates the policy-grounded compliance judgment that defines internal-control auditing. As Table 1 shows, prior benchmarks each cover only a subset of the six axes internal-control audit requires: cross-document reasoning, multimodal input, policy grounding, numerical reasoning, process-level evaluation, and risk-asymmetric scoring.

We close this gap with **REIMBURSEBENCH**, an expert-curated, evidence-grounded benchmark

* Equal contribution.

† Corresponding author.

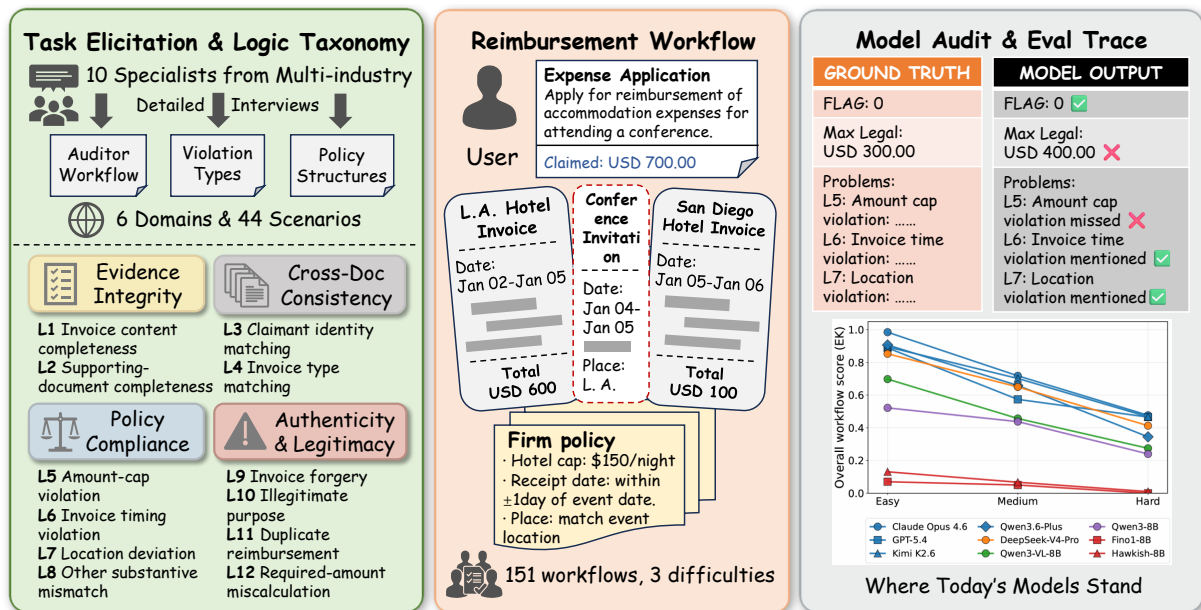


Figure 1: **Overview of REIMBURSEBENCH.** **Left:** Task elicitation and logic taxonomy. Curated by an interview of 10 multi-industry specialists, the benchmark defines 44 scenarios from 6 domains and a structured taxonomy of 12 logic points (L1–L12) grouped into 4 types. **Middle:** An illustrative workflow demonstrates the integration of free-form expense claims, multimodal receipts, and firm-specific rules. The full benchmark encompasses 151 test cases across 3 difficulty levels. **Right:** Audit evaluation pipeline. Models process the multimodal input to emit a structured audit trace—an admissibility flag, the maximum legally reimbursable amount, and a problem list. Outputs are scored point-by-point against an expert-authored ground-truth rubric, explicitly tracking detection accuracy and misses for specific rule violations (e.g., the \$300 ground truth is derived from the only two valid nights at the L.A. hotel, capped at \$150/night). The figure shows the performance of 9 evaluated models, revealing that even state-of-the-art models can fail when resolving multi-constraint, complex cross-document reasoning tasks.

that evaluates LMMs as full-workflow internal-control auditors. Each of the 151 workflows packages a free-form expense application, a bundle of receipt images and supporting documents, and a firm-specific policy; given this multimodal input, a model must emit a structured audit trace (flag, maximum legal amount, and a list of problem reports), mirroring the deliverable of a human auditor. Our primary evaluation uses a single-pass formulation to isolate the underlying model’s reasoning capability from the engineering of a particular agent scaffold.

REIMBURSEBENCH is built through a rigorous expert-driven pipeline. We first interviewed ten internal-control specialists to reconstruct the canonical audit workflow and enumerate recurring violation categories. Eight experts then constructed the 151 workflows using receipt images from public document-AI corpora or controlled synthesis, together with synthesized firm-specific policies. The workflows span three difficulty levels and collectively cover twelve logic-point categories in four conceptual groups, plus two distractor categories probing robustness.

A distinguishing methodological contribution of REIMBURSEBENCH is its fine-grained, process-level scoring protocol. In addition to the final admissibility decision and computed amount, every workflow carries a granular rubric. This rubric scores both the headline decisions and each specific policy violation an ideal auditor must identify. By automatically judging the model’s reported violations against a taxonomy of predefined logic points, we generate a process-level score that distinguishes models arriving at the right answer for the right reasons from those producing coincidentally correct outputs.

We benchmark nine systems spanning four capability tiers—Large LMMs, large LLM (paired with OCR), 8B-scale LMM, and 8B-scale LLMs including finance-domain specialists. Even under a detailed expert-knowledge prompt, the strongest system scores only 72.5% averaged across the 151 workflows, and every frontier model degrades sharply on hard workflows.

Beyond this aggregate performance, our fine-grained evaluation surfaces three structural findings a single aggregate score would obscure: (i) a

persistent asymmetry between binary admissibility judgments and precise quantitative computation; (ii) a scale-dependent lenient bias (*over-prediction*) in monetary calculations that poses severe deployment risks; and (iii) performance patterns consistent with a *compositional-generalization* (Dziri et al., 2023) challenge under realistic workflow load. Through case studies, we show how missed combinations of checks produce distinct behavioral failure modes and inadequate professional skepticism.

2 Related Work

Financial LLMs and document understanding.

Financial LLMs such as BloombergGPT (Wu et al., 2023), FinGPT (Yang et al., 2023), PIXIU (Xie et al., 2023), and Fino1 (Qian et al., 2025) predominantly target the *investment* side of finance, leaving financial internal control underexplored. On the perception side, Document AI (Huang et al., 2022; Kim et al., 2022) and OCR-free document LMMs (Ye et al., 2023; Hu et al., 2024) provide primitives for parsing receipts, and Wang et al. (2026) pushed LMMs toward cognitive reasoning over receipts, but evaluate only single receipts without connecting to the downstream audit workflow. Complementary work studies table-specialized multimodal models and long-context or hierarchical table reasoning (Su et al., 2024; Li et al., 2025; Wu et al., 2025); our focus instead is end-to-end, policy-grounded auditing over mixed evidence.

Financial and receipt benchmarks. Financial NLP benchmarks cover QA, long-document, and holistic-capability evaluation over capital-market documents (Chen et al., 2021; Zhu et al., 2021; Islam et al., 2023; Reddy et al., 2024; Xie et al., 2024; Guo et al., 2025), while receipt-level visual information extraction (Huang et al., 2019; Park et al., 2019; Wang et al., 2026) extracts structured fields from single documents. Neither evaluates the cross-document, policy-grounded reasoning that internal-control audit requires.

Agent and enterprise workflow benchmarks.

Beyond general agent benchmarks targeting tool use and web navigation (Liu et al., 2024; Mialon et al., 2024), two recent benchmarks most clearly delineate our niche: **FINCH** (Dong et al., 2025) evaluates agents on spreadsheet-centric finance & accounting workflows, and **OS-Marathon** (Wu

et al., 2026) evaluates GUI agents that transcribe data into systems. Both are concurrent and complementary to ours: FINCH targets finance *production* and OS-Marathon targets *data-entry*. Neither, however, evaluates the downstream *audit* layer—judging admissibility, calculating limits, and enumerating policy violations. Practitioner system for automated reimbursement audit (Jing, 2025) confirms clear enterprise demand but offers no public benchmark—a gap REIMBURSEBENCH fills. Table 1 summarizes how REIMBURSEBENCH compares to these existing benchmarks across key evaluation dimensions.

3 Task Formulation

3.1 Inputs and Outputs

A REIMBURSEBENCH workflow is a tuple $\mathcal{W} = (a, A, P, c, \rho)$, where a is a free-form natural-language reimbursement *application*, $A = \{x_1, \dots, x_k\}$ is a set of supporting *attachments* (predominantly receipt images, optionally with textual and/or tabular documents), P is a firm-specific *policy document*, c is the *claimed amount*, and ρ is a structured receipt-ownership statement (See Appendix A). All textual material is in English.

Given $\mathcal{W}$, a model must produce a structured audit trace: $y = (\text{flag}, \text{max_legal}, R, \Pi)$, where $\text{flag} \in \{0, 1\}$ is the admissibility decision, $\text{max_legal} \in \mathbb{R}_{\geq 0}$ is the maximum amount reimbursable under P given A , R is a free-form reasoning chain citing clauses of P and arithmetic over A , and $\Pi = \{\pi_1, \dots, \pi_m\}$ is a list of natural-language problem reports.

3.2 The MAX_LEGAL Computation Procedure

To make the task deterministic, REIMBURSEBENCH prescribes a three-step procedure for max_legal that models are instructed to follow: (1) discard attachments entirely invalid under P ; (2) cap each surviving attachment at its applicable policy limit, $\hat{v}_i = \min(v_i, \bar{v}_i)$; and (3) aggregate, $\text{max_legal} = \sum_{i \in \text{valid}} \hat{v}_i$. The verbatim instruction given to models is the OUTPUT FORMAT block in Table 12.

The admissibility flag is a deterministic function of max_legal and c : $\text{flag} = \mathbf{1}[\text{max_legal} \geq c]$. A model can in principle arrive at a correct flag for the wrong reasons, and the problem list Π is how our evaluation penalizes such hollow correctness. Producing a correct y requires the model to internally

Benchmark	Domain	Task Format	MD	MM	PG	NR	PE	RA
<i>Financial NLP / QA Benchmarks</i>								
FinQA (Chen et al., 2021)	Earnings Reports	QA	×	×	×	✓	×	×
TAT-QA (Zhu et al., 2021)	Financial Reports	Table+Text QA	×	×	×	✓	×	×
FinanceBench (Islam et al., 2023)	SEC Filings	QA	✓	×	×	✓	×	×
DocFinQA (Reddy et al., 2024)	Annual Reports	Long-doc QA	✓	×	×	✓	×	×
FinBen (Xie et al., 2024)	General Finance	Mixed (24 tasks)	✓	×	×	✓	×	×
FinEval (Guo et al., 2025)	General Finance	Exam / Holistic	×	×	×	✓	×	×
<i>Receipt / VIE Benchmarks</i>								
SROIE (Huang et al., 2019)	Retail Receipts	VIE	×	✓	×	×	×	×
CORD (Park et al., 2019)	Retail & Dining	VIE	×	✓	×	×	×	×
ReceiptBench (Wang et al., 2026)	General Receipts	VIE	×	✓	×	✓	○	×
<i>Agent / Enterprise Workflow Benchmarks</i>								
AgentBench (Liu et al., 2024)	General Assistant	Agent (8 envs)	✓	×	×	○	✓	×
GAIA (Mialon et al., 2024)	General Assistant	Agent + Tools	✓	✓	×	○	×	×
FINCH (Dong et al., 2025)	Finance & Accounting	Spreadsheet Workflow	✓	✓	×	✓	✓	×
OS-Marathon (Wu et al., 2026)	GUI Data Entry	CUA Workflow	✓	✓	×	○	✓	×
REIMBURSEBENCH (Ours)	Internal Control	Audit Workflow	✓	✓	✓	✓	✓	✓

Table 1: Comparison of REIMBURSEBENCH with representative benchmarks across financial NLP, visual information extraction, and agent/enterprise workflow domains. ✓: full support, ○: partial, ×: none. **MD** refers to Multi-Doc (cross-document reasoning); **MM** refers to Multi-Modal (image inputs); **PG** refers to Policy-Grounded (answers depend on firm-specific policy clauses rather than general knowledge or data correctness); **NR** refers to Numeric Reasoning (tasks requiring numerical computation and mathematical reasoning); **PE** refers to Process Eval (per-step / per-logic-point scoring beyond end-task outcomes); **RA** refers to Risk-Asym. (evaluation distinguishes over- vs. under-decision errors, an underexplored dimension carrying distinct financial consequences; see Section 6.2 for further discussion on this dimension).

execute a multi-stage workflow—evidence perception, policy retrieval, cross-document alignment, numerical computation, and compliance adjudication.

4 The REIMBURSEBENCH Benchmark

4.1 Construction Pipeline

Stage 1: Expert interviews. We conducted semi-structured interviews with ten practicing internal-control specialists from companies across multiple industries (See Appendix B). The interviews surfaced (i) the canonical reimbursement workflow, (ii) the recurring categories of substantive violations and process defects, and (iii) the typical structure of firm-specific reimbursement policies, informing the logic-point taxonomy in Section 4.2.

Stage 2: Template and policy authoring. Eight financial experts authored a structured workflow template and a library of firm-specific policy documents. The policy documents are synthetic, which avoids confidentiality and licensing risks while allowing control over difficulty and logic-point coverage.

Stage 3: Workflow instantiation with mixed-source receipt evidence. Each of the 151 workflows is instantiated by an author who drafts an ap-

plication narrative and claimed amount, assembles a bundle of attachments with the receipt-ownership annotation ρ , and authors the ground-truth audit trace (flag, max_legal, a per-workflow scoring rubric, and logic-point labels). Approximately two-thirds of the receipt images are from openly available real-world document-AI corpora; the remainder is synthesized when no authentic receipt could realize a specific logic-point configuration (e.g., a cross-receipt date inconsistency). To ensure broad coverage, the business scenarios are drawn from a taxonomy distilled from our expert interviews: six major expense domains subdivided into 44 specific scenarios (See Appendix C for full taxonomy). Appendix D describes the decomposition-based cross-checking of each workflow by a non-author constructor.

4.2 Logic-Point Taxonomy

A central design contribution of REIMBURSEBENCH is its logic-point taxonomy: a controlled vocabulary of twelve atomic violation categories, in four conceptual groups, that together cover the substantive content of internal-control auditing. Every problem in y carries one or more logic-point labels.

Statistic	Easy	Medium	Hard	Overall
# Workflows	50	50	51	151
# Attachments / WF	6.52	6.32	7.12	6.66
# Scoring points / WF	2.72	3.08	3.57	3.13
# Logic points / WF	0.94	1.54	1.94	1.48
Context length (chars)	81.3k	74.5k	94.7k	83.6k
% workflows flag=1	40.0	28.0	5.9	24.5

Table 2: Aggregate statistics of REIMBURSEBENCH (means for per-WF rows).

Group A — Evidence Integrity. **L1 Invoice content completeness** (missing required elements, or an image so degraded that the face content cannot be recovered); **L2 Supporting-document completeness** (missing pre-trip approval, attendee roster, project plan, etc.).

Group B — Cross-Doc Consistency. **L3 Claimant identity matching** (receipt holders do not match the applicant or authorized-participant roster); **L4 Invoice type matching** (invoice type not matching the claimed expense category).

Group C — Policy Compliance. **L5 Amount-cap violation**, **L6 Invoice timing violation**, **L7 Location deviation**, **L8 Other substantive mismatch**.

Group D — Authenticity & Legitimacy. **L9 Invoice forgery**, **L10 Illegitimate purpose** (e.g., private use), **L11 Duplicate reimbursement** (the same receipt used more than once), **L12 Required-amount miscalculation**.

Distractor categories. We additionally seed two workflow-level non-violation categories to probe robustness: **D1 Irrelevant context** (off-topic information) and **D2 Blurred invoice image** (still valid). They are not connected with any problem in y .

4.3 Dataset Statistics

REIMBURSEBENCH comprises 151 workflows distributed approximately uniformly across difficulty: *easy* (50), *medium* (50), *hard* (51). Difficulty is assigned by the workflow author and reflects the joint complexity of attachment content, number of logic points activated, cross-document reasoning depth, and policy-clause intricacy.

Logic-point activations are non-uniform by design. Their construction distribution was informed by expert interviews to provide broad coverage of practically relevant cases. **L10 Illegitimate purpose** (15.7%), **L5 Amount-cap violation** (15.2%),

L8 Other substantive mismatch (14.3%), and **L1 Invoice content completeness** (9.0%) account for over 54% of logic-tagged points; **L3 Claimant identity matching** and **L12 Required-amount miscalculation** are rare (2.7% each) and should be interpreted cautiously. Full counts and difficulty breakdowns are in Appendix F.

4.4 Evaluation Metrics & Grading

Each workflow is evaluated against an expert-authored rubric of weighted scoring points (summing to 1.0). The headline decisions, flag and max_legal, are extracted by a strict parser and graded deterministically (exact match within a 0.1 currency-unit tolerance); parsing failures yield zero for both. DeepSeek-V4-Pro judges only whether the output mentions each pre-specified logic-tagged problem, as a constrained binary content-matching task. On all logic-tagged points from 30 randomly sampled workflows, its verdicts agreed fully with manual judgments (Cohen’s $\kappa = 1.0$). See Table 3 for metrics and Appendices J and D for details.

5 Experiments

5.1 Experimental Setup

Models. We evaluate nine systems in four capability tiers. The *Large LMM* tier comprises four frontier multimodal systems (Claude Opus 4.6 (Anthropic, 2026), GPT-5.4 (OpenAI, 2026), Kimi K2.6 (Moonshot AI, 2026), Qwen3.6-Plus (Yang et al., 2025)). The *Large LLM* tier contains DeepSeek-V4-Pro paired with Qwen-VL-OCR-2025-11-20 as the OCR front-end. The *Small LMM* tier contains Qwen3-VL-8B (Yang et al., 2025). The *Small LLM* tier comprises three 8B-scale text-only models: Qwen3-8B (Yang et al., 2025) as a general baseline, and the finance-domain specialists Fino1-8B (Qian et al., 2025) and Llama-3.1-Hawkish-8B (Mukaj, 2024) (referred to as Hawkish-8B below). See Table 7 for model details.

Prompt conditions. Each model is evaluated under two prompt conditions. **Zero-knowledge (ZK)** presents only the task description and output format. **Expert-knowledge (EK)** prepends a detailed expert-knowledge block describing the canonical audit workflow and reasoning guidance, identical across all workflows (no workflow-specific hints). EK approximates the realistic deployment regime; ZK isolates the model’s prior knowledge of corporate audit. Main results report EK using one

sampled output per workflow; ZK is the ablation in Section 6.3.

5.2 Main Results

Table 3 reports our primary results; Figure 2 visualizes the difficulty-stratified scores.

Headline gaps. Claude Opus 4.6 leads at $\bar{S} = 72.5\%$, followed by Kimi K2.6 at 68.7%; Table 3 reports workflow-level bootstrap uncertainty for all systems. At 8B scale, the best system (Qwen3-VL-8B, 47.5%) trails Claude by 25.0%, while the finance specialists score below 7.0%. Qwen3-8B’s 72.8% flag accuracy but 0.0% F1 further shows that class skew can make a degenerate predictor appear competitive.

Sharp degradation with difficulty. Every model degrades monotonically from easy to hard (Figure 2). The frontier systems are near-saturated on easy workflows—Claude reaches 98.5% and all four large LMMs exceed 88%—but lose roughly half their score by hard: Claude falls to 47.7%, and Qwen3.6-Plus to 34.5%. No system exceeds 47.7% on hard workflows. This decline co-varies with active logic points, context length, attachment content, policy complexity, and class distribution; it is consistent with increased workflow load.

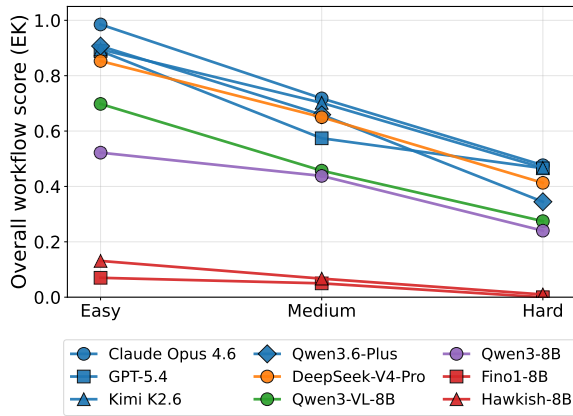


Figure 2: Per-difficulty overall workflow score under EK. Color encodes tier; markers identify individual models within tiers.

Flag vs. max_legal asymmetry. For models that produce well-formed outputs, flag accuracy exceeds max_legal exact match by +18 to +50 percentage points (Claude: +20.5; GPT-5.4: +19.8; Qwen3-VL-8B: +49.6). Flag is a binary outcome reachable through many reasoning paths, so a model can land on the correct flag through shallow

pattern matching without performing the full quantitative reasoning that max_legal demands. The F1 column confirms part of this apparent competence is illusory—Qwen3-8B’s 0.0% F1 shows its flag predictions carry no signal—but even after excluding such degenerate cases a substantial flag-F1 vs. max_legal-EM gap remains. In short, models can guess whether “something is wrong” far more easily than they can specify what, why, and by how much.

6 Analysis

6.1 In-depth Failure Mode Analysis

Drawing on a careful inspection of model outputs, we identify three recurring, output-level behavioral failure modes in professional audit workflows, ordered by the depth of evidence integration they involve:

Content Consistency Blindness (Mode 1). The model fails to detect *surface-level* contradictions, ambiguities, or semantic conflicts within the provided documents (e.g., **L6**: a DD/MM/YY invoice date misread as a US-format month), yet the model does not acknowledge any anomaly and proceeds as if the input is perfectly self-consistent.

Conflict Self-Rationalization (Mode 2). The model *detects* a potential issue or inconsistency, but it unilaterally invents a benign explanation (e.g., “likely a typo”) or lowers its threshold (e.g., “barely acceptable”) to resolve the conflict. The model acknowledges the anomaly but fails to maintain professional skepticism.

Behavioral Intention Reasoning Defect (Mode 3). The model fails to infer when the described behavior provides evidence of possible deliberate rule evasion (e.g., **L9**: two near-identical invoices, one blurred or recolored). Even when individual documents appear locally plausible, the overall pattern may warrant scrutiny, yet the model does not flag this concern.

Together, Modes 1–3 trace progressively richer evidence integration: perceiving anomalies, sustaining uncertainty, and recognizing intent-related signals.

We applied the failure-mode detection to every model’s responses; human validation on 30 workflows reached Cohen’s $\kappa = 0.821$ (Appendix D). Figure 3 reveals clear stratification among models with assessable outputs: frontier models exhibit lower failure rates, while small open-weight

Tier	Model	Overall workflow score $\bar{S}$					Flag	Flag	ML	Logic
		Easy	Medium	Hard	All	95% CI	Acc.%	F1%	EM%	Rec.%
Large LMM	Kimi K2.6	89.5%	70.2%	46.7%	68.7%	[62.5, 74.7]	80.8	63.2	62.9	66.3
	GPT-5.4	88.9%	57.4%	46.6%	64.2%	[57.6, 70.6]	76.8	59.2	57.0	64.4
	Claude Opus 4.6	98.5%	71.8%	47.7%	72.5%	[66.6, 78.1]	82.8	68.3	62.3	71.6
	Qwen3.6-Plus	90.7%	65.9%	34.5%	63.5%	[57.2, 69.4]	74.8	53.7	55.0	59.1
Large LLM	DeepSeek-V4-Pro	85.3%	65.0%	41.3%	63.7%	[57.2, 70.0]	79.5	55.9	55.6	63.5
Small LMM	Qwen3-VL-8B	69.8%	45.7%	27.5%	47.5%	[42.9, 52.2]	74.8	37.3	25.2	36.0
Small LLM	Qwen3-8B	52.2%	43.8%	24.0%	39.9%	[36.3, 43.6]	72.8	0.0	26.5	19.7
	Fino1-8B	7.0%	5.0%	0.0%	4.0%	[1.9, 6.5]	0.7	5.3	0.7	8.9
	Hawkish-8B	13.1%	6.7%	1.0%	6.9%	[3.9, 10.2]	3.3	17.0	2.6	10.9
<i>Always-predict-0 baseline</i>		—	—	—	—	—	75.5	0.0	—	—

Table 3: Main results on REIMBURSEBENCH under the EK prompt condition. **Overall Score** $\bar{S}$ is the mean per-workflow earned score; 95% CIs for all evaluated systems are percentile workflow-level bootstrap intervals over 10,000 resamples. **Flag Acc.& F1** evaluate admissibility (flag=1 as positive); **ML EM%** is max_legal exact match; **Logic Rec.%** is logic-point recovery. The always-0 baseline illustrates the class skew.

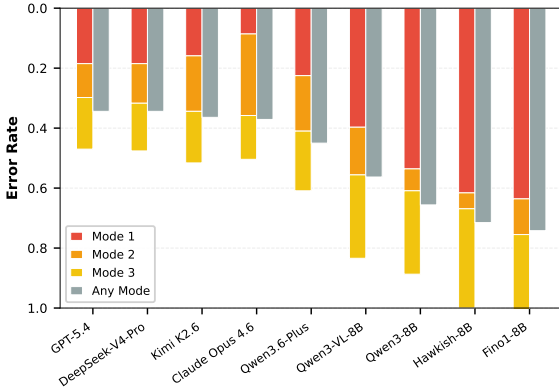


Figure 3: Error rate decomposition across evaluated models. The y-axis indicates the error rate (0.0–1.0). Colored bars show the individual trigger rates for three distinct failure modes. The grey bar denotes the **Any Mode** rate, representing the overall proportion of workflows exhibiting *at least one* failure type.

models show markedly higher rates, especially for Modes 1 and 3. For Fino1-8B and Hawkish-8B, the "Any Mode" rate understates true failure because many degraded outputs cannot be matched to a specific mode. Even the best models trigger Mode 3 in nearly 15% of cases, highlighting the difficulty of recognizing evidence-supported intent signals. Complete data and case studies appear in Table 10 and Appendix M.

A Compositional-Generalization Interpretation. Across these modes, failures often arise when multiple individually simple checks must be applied

jointly. The resulting performance pattern is consistent with a *compositional-generalization* challenge (Dziri et al., 2023) under realistic workflow load. Because workflow difficulty also increases with context length, attachment content, policy complexity, and shifts in class distribution, we view compositionality as one plausible contributor. Controlled atomic-versus-composed studies on matched evidence could further disentangle these factors.

6.2 Risk-Asymmetric Reasoning Bias

Internal-control auditing has asymmetric, but not one-sided, costs: over-approval creates direct monetary and compliance risk, whereas under-approval can damage employee trust and create operational friction. To compare error magnitude across workflows, for reference max_legal m , prediction $\hat{m}$, and claimed amount c , we define

$$\begin{aligned} e^+ &= \frac{\max(\hat{m} - m, 0)}{c}, \\ e^- &= \frac{\max(m - \hat{m}, 0)}{c}. \end{aligned} \quad (1)$$

Among the five frontier systems, median conditional over-prediction ranges from 51.5% to 81.9% of the claimed amount, while median conditional under-prediction ranges from 32.7% to 85.1%. For Qwen3-VL-8B and Qwen3-8B, median e^+ reaches 100%, largely because they approve approximately the full claim when the reference max_legal is zero. The lenient bias is therefore financially material

Model	Median e^+	Median e^-
Claude Opus 4.6	51.5%	81.2%
Qwen3.6-Plus	65.1%	32.7%
Kimi K2.6	70.1%	85.1%
DeepSeek-V4-Pro	71.0%	81.6%
GPT-5.4	81.9%	75.6%

Table 4: Claim-normalized max_legal error magnitude in the EK setting, conditional on error direction.

and becomes more pronounced at smaller scale; the under-prediction magnitudes likewise show substantial employee-facing costs. This pattern is consistent with asymmetric error propagation, where a missed disqualifying condition leaves the offending receipt in the admissible total.

6.3 Effect of Expert-Knowledge Prompting

The EK-ZK gap is small for every model (Appendix H, Table 9): the largest improvement is only +2.6% (GPT-5.4), and two models show slight negative deltas. These small differences suggest that generic audit guidance alone provides limited gains under this setup. Performance remains limited when the supplied principles must be applied across multiple documents and numerical steps in a specific workflow.

6.4 Do Finance-Domain Models Help?

The finance-specific models (Fino1-8B, Hawkish-8B) fail almost completely ($\bar{S} \leq 6.9\%$), lagging far behind the general-purpose Qwen3-8B (39.9%). As detailed in Appendix J, their dominant failure mode is severe instruction-following collapse. Built on an old Llama-3.1 base and fine-tuned on capital-market finance data, these models do not naturally transfer to the strict formatting and cross-document reasoning required by operational audits.

6.5 Practical Deployment Gap

Expense auditing is privacy-sensitive, so organizations may require locally deployable models within the corporate perimeter—yet the best such model we evaluate (Qwen3-VL-8B, $\bar{S} = 47.5\%$, 27.5% on hard) trails the strongest frontier LMM by 25.0%. The deployable models are not yet strong enough to be trusted, a direct barrier to practical use.

6.6 Diagnostic Sampling and Tool Use

We conduct two pilots with Qwen3.6-Plus on the 51 hard workflows. First, strict full-credit success

rises from 4/51 (7.8%) in the original run to 10/51 (19.6%) under pass@3 and 12/51 (23.5%) under pass@5. At the component level, flag accuracy reaches 80.4% under pass@3 and 86.3% under pass@5; under pass@5, amount accuracy and complete logic-point coverage reach 45.1% and 31.4%, respectively. Repeated sampling therefore recovers a non-trivial subset of errors, although 39/51 hard workflows remain unsolved. Since deployment lacks reference answers for selecting the correct sample, we interpret pass@ k as diagnostic recoverability; a practical multi-run auditor would additionally require an explicit selection or adjudication mechanism.

Second, a minimal pass@1 pilot provides a policy-document lookup tool and a deterministic calculator. Tool use raises full-credit success from 7.8% to 11.8%, mean score from 0.345 to 0.421, flag accuracy from 60.8% to 72.5%, amount accuracy from 13.7% to 23.5%, and complete logic coverage from 13.7% to 19.6%. The model uses both tools on every workflow, averaging 5.47 calls per workflow. These gains suggest that simple tools help structured checking and numerical computation, while leaving substantial room for stronger retrieval, verification, and output adjudication. Full component results appear in Appendix E.

7 Conclusion

We introduced REIMBURSEBENCH, the first benchmark for full-workflow, multimodal, policy-grounded reasoning in corporate internal-control audit. It comprises 151 expert-constructed workflows evaluated by a process-level scoring rubric spanning a twelve-category logic taxonomy. Benchmarking nine systems, we find the strongest frontier LMM attains only 72.5% under expert prompting, with a steep degradation on hard workflows. Risk-aware, process-level evaluation reveals (i) an asymmetry between binary admissibility judgment and precise quantitative computation; (ii) a scale-dependent lenient bias in max_legal predictions; and (iii) performance patterns consistent with compositional-generalization difficulties under realistic workflow load.

Our findings suggest that isolated-skill scores can overestimate real-world readiness when many checks must hold simultaneously. We release REIMBURSEBENCH as a testbed for compliance-grade multimodal reasoning and motivate controlled evaluation of compositional load in future work.

Limitations

Scale and external validity. REIMBURSEBENCH contains 151 workflows and 223 logic-point activations, but rare categories such as L3 ($n = 5$) and L12 ($n = 6$) remain uncertain. The expert-designed distribution prioritizes diagnostic coverage over prevalence estimation. Synthetic policies enable public release and controlled construction but cannot represent every firm’s policy language.

Evaluation scope. Main results use one sampled output per workflow. Bootstrap intervals quantify workflow-sampling uncertainty, while run-to-run API variance remains outside their scope. Logic-point scoring emphasizes recovery of predefined issues and does not include a separate penalty for unsupported reports. Future stage-level annotations and controlled atomic-versus-composed experiments could better separate perception, retrieval, and compositional effects.

Deployment risks. Mode 3 and intent-sensitive labels are designed as evidence-supported audit concerns that require human adjudication. Unsupported fraud allegations, under-reimbursement, and culturally contingent legitimacy judgments can cause harm. Practical systems therefore require human review, transparent evidence, and employee appeal mechanisms.

Ethics Statement

The receipt and ticket images in REIMBURSEBENCH are collected from two categories of sources. First, a subset is derived from publicly available open-source document corpora. We use these documents only under their applicable public licenses or terms of use, and we do not attempt to recover, infer, or reconstruct any redacted information. For documents that have already been de-identified by the original corpus providers, we preserve the provided de-identification. When a receipt image contains an opaque black block covering the holder’s name or other identifying information, we retain the redaction as-is and do not expose the underlying content.

Second, we include synthetic receipts generated by our own pipeline. These synthetic documents are not copied from real company records and are designed not to contain real personal data, confidential business information, or proprietary reimbursement records. Any names, merchants, addresses,

transaction identifiers, or policy fields appearing in the synthetic data are fictitious or template-generated unless explicitly sourced from public, non-sensitive information.

The benchmark is intended for research on document understanding, reimbursement reasoning, and policy compliance, not for identifying individuals, reconstructing private transactions, or making real reimbursement decisions. Expert annotators were informed that their annotations would be used to construct a public research benchmark and were compensated fairly for their work. During the dataset construction process, AI tools were used to assist with simple, repetitive tasks, and all AI-assisted outputs were strictly reviewed and verified by human annotators. No annotator was asked to provide personal financial documents or private reimbursement records.

Where applicable, we follow the redistribution conditions of the source datasets and release only materials that are permitted to be used for research purposes. We also provide a mechanism for reporting potential privacy or licensing concerns after release; upon valid request, we will review and, if necessary, remove or further redact affected examples. Overall, the dataset is constructed to minimize exposure of personal data and confidential corporate information while supporting reproducible research on reimbursement-oriented document reasoning.

Acknowledgements

This work was supported in part by Zhejiang Provincial Philosophy and Social Sciences Planning Project (grant number 22QNYC04ZD), the National Social Science Fund of China (No. 24BGL071), the Fundamental Research Funds for the Central Universities, Ningbo Key Research and Development Program (No. 2025Z190), and ZJU Kunpeng&Ascend Center of Excellence.

References

- Anthropic. 2026. Introducing claude opus 4.6. <https://www.anthropic.com/news/claude-opus-4-6>. Accessed May 2026.
- Zhiyu Chen, Wenhui Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2021. *Finqa: A dataset of numerical reasoning over financial data*. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711.

- Haoyu Dong, Pengkun Zhang, Yan Gao, Xuanyu Dong, Yilin Cheng, Mingzhe Lu, Zikun Zhu, Adina Yakefu, and Shuxin Zheng. 2025. Finch: Benchmarking finance & accounting across spreadsheet-centric enterprise workflows. *arXiv preprint arXiv:2512.13168*.
- Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang (Lorraine) Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena Hwang, Soumya Sanyal, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. 2023. Faith and fate: Limits of transformers on compositionality. *Advances in neural information processing systems*, 36:70293–70332.
- Xin Guo, Haotian Xia, Zhaowei Liu, Hanyang Cao, Zhi Yang, Zhiqiang Liu, Sizhe Wang, Jinyi Niu, Chuqi Wang, Yanhui Wang, Xiaolong Liang, Xiaoming Huang, Bing Zhu, Zhongyu Wei, Yun Chen, Weining Shen, and Liwen Zhang. 2025. *Fineval: A chinese financial domain knowledge evaluation benchmark for large language models*. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6258–6292.
- Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. 2024. *mplug-docowl 1.5: Unified structure learning for ocr-free document understanding*. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3096–3120.
- Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. *Layoutlmv3: Pre-training for document ai with unified text and image masking*. In *Proceedings of the 30th ACM international conference on multimedia*, pages 4083–4091.
- Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. 2019. *Icdar2019 competition on scanned receipt ocr and information extraction*. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1516–1520. IEEE.
- Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. 2023. Financebench: A new benchmark for financial question answering. *arXiv preprint arXiv:2311.11944*.
- Mu Jing. 2025. *Research on intelligent audit system for employee expense reimbursement based on agent technology*. *International Journal of Emerging Technologies and Advanced Applications*, 2(12):1–6.
- Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. *Ocr-free document understanding transformer*. In *European Conference on Computer Vision*, pages 498–517. Springer.
- Liyao Li, Jiaming Tian, Hao Chen, Wentao Ye, Chao Ye, Haobo Wang, Ningtao Wang, Xing Fu, Gang Chen, and Junbo Zhao. 2025. *LongTableBench: Benchmarking long-context table reasoning across real-world formats and domains*. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 11927–11965, Suzhou, China. Association for Computational Linguistics.
- Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, and 3 others. 2024. Agentbench: Evaluating llms as agents. In *International Conference on Learning Representations*, volume 2024, pages 52989–53046.
- Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. 2024. Gaia: a benchmark for general ai assistants. In *International Conference on Learning Representations*, volume 2024, pages 9025–9049.
- Moonshot AI. 2026. Kimi k2.6: Advancing open-source coding. <https://www.kimi.com/blog/kimi-k2-6>. Accessed May 2026.
- Mukaj. 2024. Llama-3.1-hawkish-8b. <https://huggingface.co/mukaj/Llama-3.1-Hawkish-8B>. Accessed May 19, 2026.
- OpenAI. 2026. Gpt-5.4. <https://openai.com/index/introducing-gpt-5-4/>. Accessed May 2026.
- Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. 2019. Cord: a consolidated receipt dataset for post-ocr parsing. In *Workshop on document intelligence at NeurIPS*, volume 2019, page 5.
- Lingfei Qian, Weipeng Zhou, Yan Wang, Xueqing Peng, Jimin Huang, and Qianqian Xie. 2025. Fino1: On the transferability of reasoning enhanced llms to finance. *arXiv e-prints*, pages arXiv–2502.
- Varshini Reddy, Rik Koncel-Kedziorski, Viet Dac Lai, Michael Krumdtick, Charles Lovering, and Chris Tanner. 2024. *Docfinqa: A long-context financial reasoning dataset*. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 445–458.
- Aofeng Su, Aowen Wang, Chao Ye, Chen Zhou, Ga Zhang, Gang Chen, Guangcheng Zhu, Haobo Wang, Haokai Xu, Hao Chen, Haoze Li, Haoxuan Lan, Jiaming Tian, Jing Yuan, Junbo Zhao, Junlin Zhou, Kaizhe Shou, Liangyu Zha, Lin Long, and 14 others. 2024. *Tablegpt2: A large multimodal model with tabular data integration*. *Preprint*, arXiv:2411.02059.
- Yandi Wang, Libin Zhan, Ziwei Huang, Tiancheng Luo, Yuxuan Jiang, Wang Dong, Leilei Gan, and Jun Chen. 2026. *From recognition to reasoning: Benchmarking and enhancing mllms on real-world receipt document understanding*. *Preprint*, arXiv:2605.22413.

- Jing Wu, Daphne Barretto, Yiye Chen, Nicholas Gydé, Yanan Jian, Yuhang He, and Vibhav Vineet. 2026. [Os-marathon: Benchmarking computer-use agents on long-horizon repetitive tasks](#). *Preprint*, arXiv:2601.20650.
- Pengzuo Wu, Yuhang Yang, Guangcheng Zhu, Chao Ye, Hong Gu, Xu Lu, Ruixuan Xiao, Bowen Bao, Yijing He, Liangyu Zha, Wentao Ye, Junbo Zhao, and Haobo Wang. 2025. [RealHiTBench: A comprehensive realistic hierarchical table benchmark for evaluating LLM-based table analysis](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7105–7137, Vienna, Austria. Association for Computational Linguistics.
- Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kam-badur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.
- Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, Yijing Xu, Haoqiang Kang, Ziyang Kuang, Chenhan Yuan, Kailai Yang, Zheheng Luo, Tianlin Zhang, Zhiwei Liu, Guojun Xiong, and 15 others. 2024. Finben: A holistic financial benchmark for large language models. *Advances in Neural Information Processing Systems*, 37:95716–95743.
- Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. Pixiu: A large language model, instruction data and evaluation benchmark for finance. *arXiv preprint arXiv:2306.05443*.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. Fingpt: Open-source financial large language models. *arXiv preprint arXiv:2306.06031*.
- Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, Qin Jin, Liang He, Xin Lin, and Fei Huang. 2023. [Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2841–2858.
- Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. [Tat-qa: A question answering benchmark on a hybrid of tabular and textual content in finance](#). In *Proceedings of the 59th annual meeting of the Association for Computational Linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pages 3277–3287.

A Receipt-ownership Statement

In an idealized benchmark, the receipt holder’s name would be legible on the face of each invoice. Two practical constraints make this infeasible in our dataset: (i) re-editing the printed name on hundreds of public-corpus receipts to align with constructed narratives is relatively costly, and (ii) a non-trivial fraction of receipts from public corpora are already privacy-redacted with opaque black blocks over the holder’s name. We therefore lift receipt ownership into the structured prompt and instruct the model to treat it as ground truth.

B Expert Interview Demographics

To ensure the ecological validity and practical rigor of REIMBURSEBENCH, we engaged ten practicing financial and internal-control professionals for our task elicitation. Crucially, all participating experts possess a minimum of five years of direct working experience in accounting, auditing, or corporate finance. To capture a comprehensive spectrum of real-world compliance practices, the panel was deliberately assembled from a diverse range of organizational scales and sectors. Specifically, the cohort comprises two active auditors from a “Big Four” international accounting firm, one financial specialist from a major state-owned bank, and seven experienced practitioners sourced from various small, medium, and large private enterprises across multiple industries. This diverse composition ensures that our logic-point taxonomy is well-grounded in both stringent institutional regulations and typical, agile enterprise realities.

C Scenario Taxonomy

While the logic-point taxonomy (Section 4.2) governs *what kind of violation* a workflow tests, an orthogonal concern is *what kind of business situation* it depicts. A benchmark probing the twelve logic points only within, say, hotel-stay reimbursements would over-fit to a narrow slice of corporate practice. To ensure REIMBURSEBENCH reflects the breadth of situations a real internal-control auditor encounters, the business scenarios underlying the 151 workflows are sampled from a scenario taxonomy distilled directly from our ten expert interviews (Section 4.1, Stage 1).

The taxonomy is organized into six major expense domains, subdivided into 44 specific scenarios (Table 5). The six domains were identified by

the interviewed specialists as the recurring high-volume, audit-sensitive categories of corporate reimbursement.

D Quality Control and Evaluation Reliability

Workflow Construction Quality. Because REIMBURSEBENCH workflows are constructed, quality control focuses on independent logical-consistency checks during dataset creation. We decomposed every workflow into atomic verification units and routed each unit to a constructor who did not author the workflow. Across the 151 workflows, the 1,215 atomic verification units yielded a 12.76% first-pass discrepancy rate. Every flagged unit was subsequently revised by the original author and re-verified before the final release.

LLM-as-Judge Scoring Reliability. We use DeepSeek-V4-Pro as the judge for logic-tagged scoring points. To mitigate self-evaluation bias, headline decisions (flag and max_legal) are graded by deterministic scripts. For the remaining logic points, the judge is restricted to a constrained binary task: “does the output mention a problem matching this pre-specified rubric item, yes or no?”, avoiding open-ended quality judgments. To verify this setup, we randomly sampled 30 workflows and manually inspected every logic-tagged scoring point. The judge’s verdict agreed with human annotations on 100% of the inspected points (Cohen’s $\kappa = 1.0$).

Failure Mode Detection Reliability. In Section 6.1, we again employed DeepSeek-V4-Pro to automatically identify the presence of the three behavioral failure modes. The judge was provided with the reference answers of the workflows, the evaluated model’s reasoning trace, and the strict definitions of the three modes, and was instructed to determine whether each mode was triggered. To validate this detection pipeline, human experts independently annotated the failure modes for a random sample of 30 workflows. The agreement between the LLM judge and human annotators reached a Cohen’s κ of 0.821, indicating high reliability and confirming that the automated failure mode identification is trustworthy.

E Diagnostic Sampling and Tool-Use Details

We run Qwen3.6-Plus on the 51 hard workflows (Table 6). Under $\text{pass}@k$, a workflow or component is counted as correct when any of k independently sampled outputs is correct. The tool pilot provides policy-document lookup and a deterministic calculator at $\text{pass}@1$; the model uses both tools on every workflow and averages 5.47 calls per workflow.

Repeated sampling recovers some failures, but 39/51 hard workflows remain unsolved even under $\text{pass}@5$. Tool use yields a modest $\text{pass}@1$ gain but does not identify which remaining errors arise from perception, retrieval, evidence alignment, rule application, arithmetic, or adjudication.

F Logic-Point Distribution and Diagnostics

Table 8 reports all 12 logic-point categories and their easy/medium/hard breakdowns.

Figure 4 reports per-logic-point recovery Rec_ℓ under EK. Two logic points stand out as the hardest in aggregate: **L9** Invoice forgery (mean $\text{Rec}_\ell = 0.30$) and **L3** Claimant identity matching (mean $\text{Rec}_\ell = 0.32$). The model-level patterns differ across the two categories. For L9, the text-only DeepSeek-V4-Pro—which sees only an OCR transcript—attains the highest recovery (0.60), above every native-multimodal model. This result points to factors beyond native visual access, including possible differences in how readily models flag evidence of document tampering; frontier-model recovery ranges from 0.40 to 0.60. For L3, recovery ranges from 0.50 to 0.67 among frontier models and is 0.00 for all three 8B text-only models, a pattern consistent with the challenge of reconciling receipt ownership with participant rosters across documents. We note that both L3 ($n = 5$ workflows) and L12 ($n = 6$) are low-frequency categories (Table 8); their per-point estimates carry wide uncertainty and should be read as indicative rather than precise. At the other extreme, L4 Invoice type matching is recovered reliably by every non-finance-specialist system (≥ 0.83 for the six general models), suggesting that this surface category-matching task is comparatively easier than the judgment- and integration-heavy categories.

G Model Settings

Table 7 reports model settings details.

H Effect of Expert-Knowledge Prompting

Table 9 (left) reports overall workflow scores under the two prompt conditions; the analysis is in Section 6.3.

I Token Efficiency

Figure 5 relates overall workflow score to generated token count. Among models that expose both reasoning and output tokens, the relationship is non-monotone: Kimi K2.6 expends substantially more tokens than Claude Opus 4.6 (22,537 vs. 3,474) yet scores lower (68.7% vs. 72.5%), indicating that longer reasoning chains do not guarantee better performance. Across tiers, the 8B systems occupy a low-score regime regardless of token budget—Qwen3-8B (458 tokens, $\bar{S} = 39.9\%$) and Fino1-8B (2,611 tokens, $\bar{S} = 4.0\%$) differ by more than $5\times$ in generated tokens but the smaller-token model scores much higher—so the bottleneck for these models is reasoning quality, not reasoning length. GPT-5.4, whose reported tokens exclude reasoning traces, cannot be directly compared and is shown for completeness only.

J Output Parsing and Format-Failure Analysis

We parse each model’s response by extracting the first pair of consecutive lines such that line i is a single $\emptyset/1$ token (with a small set of language-equivalent normalizations) and line $i + 1$ contains a numeric value (after stripping currency symbols and thousand separators). A workflow on which no such pair exists is a parse failure for both the flag and max_legal scoring points.

Table 9 (right) reports parse failure counts across all models and prompt conditions. Three distinct patterns emerge, corresponding to our model tiers.

Catastrophic format collapse in finance-domain 8B models. Fino1-8B and Hawkish-8B, both built on Llama-3.1-8B, have the highest parse-failure counts, with failures on 149–150 and 140–141 workflows, respectively. Both models usually begin with a chain-of-thought preamble and omit the required leading flag and max_legal fields; we found no correctly formatted pair elsewhere in their responses. These failures account for much of their low overall scores in Table 3, because losing the

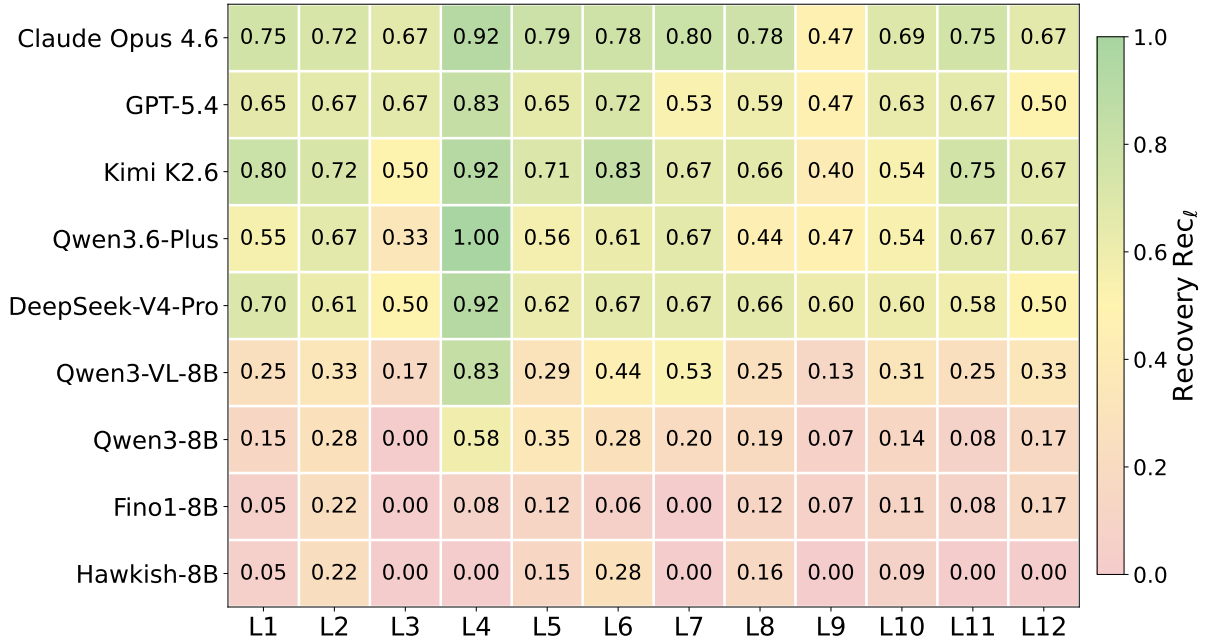


Figure 4: Per-logic-point recovery Rec_ℓ under the EK condition. Each cell is the fraction of scoring points tagged with the column’s logic point that the row’s model satisfies; higher is better.

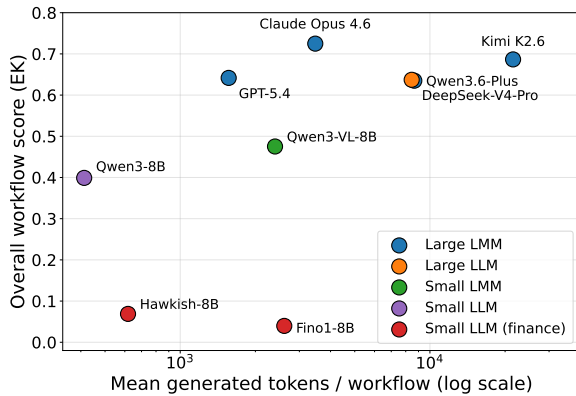


Figure 5: Overall workflow score (EK) vs. mean generated tokens per workflow (thinking + output, log scale). For GPT-5.4, token counts exclude reasoning traces because Azure’s GPT-5.4 API does not expose internal “thinking” content; its data point should be read as output tokens only.

flag and max_legal points forfeits approximately half the available weight before logic-point scoring is considered. Similar failure rates under EK and ZK suggest limited sensitivity to the prompt condition.

GPT-5.4’s refusals. GPT-5.4 exhibits 8–9 parse failures across both prompt conditions, all of which are explicit refusals (“I’m sorry, but I cannot assist with that request.”). These refusals prevent the model from producing a valid flag or max_legal

value for the affected workflows, resulting in a minor but systematic score penalty.

Sporadic failures in otherwise well-formatted models. The remaining models—Claude Opus 4.6, Kimi K2.6, Qwen3.6-Plus, DeepSeek-V4-Pro, Qwen3-VL-8B, and Qwen3-8B—show at most 5 parse failures out of 151 workflows in any condition, and Qwen3.6-Plus achieves zero failures in both conditions. These sporadic failures typically result from minor format deviations (e.g., placing the numeric value on the same line as the flag token). We treat these as low-level noise that marginally depresses the scores of affected models, with an upper bound of approximately 3.3 percentage points of score loss ($5/151 \approx 3.3\%$).

K Failure Mode Statistics

See Table 10 for detailed data of failure modes.

L Prompt Templates

We present two prompt configurations: ZK (Zero-Knowledge) with a plain system message, and EK (Expert-Knowledge) with a system message embedding a four-stage audit workflow and core audit principles. See Table 11 for the system message and Table 12 for the user message.

M In-depth Case Studies of Failure Modes

In this appendix, we provide detailed case studies illustrating the three output-level behavioral failure modes characterized in Section 6.1. The six case studies are grouped by the progressively richer evidence integration reflected in these modes:

- **Mode 1: Content Consistency Blindness (Figures 6 and 7).** The first two examples show outputs that overlook surface-level contradictions within the provided documents. These include an internal arithmetic inconsistency (e.g., a fee counted twice) and a date-format conflict (e.g., DD/MM/YYYY vs. MM/DD/YYYY); the resulting audits proceed as though the input were self-consistent.
- **Mode 2: Conflict Self-Rationalization (Figures 8 and 9).** The middle two examples show outputs that detect an inconsistency, such as a mismatch between the travel destination on the receipt and the application form, but then adopt an unsupported benign explanation—for example, a “clerical data-entry error” or “typo”—and produce a compliant verdict.
- **Mode 3: Behavioral Intention Reasoning Defect (Figures 10 and 11).** The final two examples concern evidence patterns that warrant scrutiny beyond local document consistency. The outputs do not flag the implausibility of mops and buckets submitted as “premium executive gifts,” and treat visually manipulated duplicate receipts as routine duplicates without noting possible deliberate manipulation. These audits capture local checklist items but omit the broader behavioral signal.

Major domain	#	Specific scenarios
Travel	3	Multi-city cross-province field-research trips (transportation); accommodation for a week-long industry academic conference; long-haul connecting meals while in transit.
Business hospitality	6	Inter-city chartered car for core-client pickup; customized business gifts for core clients; accompanying clients on remote plant-site inspections; working meals for strategic-partner negotiations; executive-room stays arranged for external experts; flowers and decorations for hospitality-venue setup.
Education	12	Chartered transport for student provincial-competition trips (transportation); long-distance travel for teachers to remote teaching-support posts (transportation); high-speed-rail fares for visiting guest lecturers (transportation); unified hotel arrangements during student off-campus internships (accommodation); guest rooms for delegates when hosting an academic conference (accommodation); summer-camp dormitory rental for students and staff (accommodation); working meals for a visiting exchange delegation (catering); small gifts and decorations for class activities (procurement); on-line national-training-program course purchases for teachers (training/tuition); EMBA tuition installments for management at a top university (training/tuition); hotel ballroom for an off-campus admissions presentation (venue rental); hotel ballroom for a recruitment presentation (venue rental).
Engineering projects	5	Regular commuting during a cross-city project-site assignment (transportation); working meals for night-shift overtime during a project crunch (catering); per-head meal subsidies for outsourced construction crews (catering); monthly utility settlement for a project-site office (utilities); site fees for the project team’s prefabricated-cabin (venue rental).
Daily office operations	11	Intra-city commuting taxi fares (transportation); monthly metro/bus-card fees for public-transit commuting (transportation); temporary dormitory or apartment rental during a new-site setup period (accommodation); emergency hotel stays caused by employee overtime (accommodation); operating subsidy for the internal office snack bar (catering); overtime missed-meal fees via delivery or working meals (catering); fixed daily meal subsidy during off-site work (catering); internal hospitality meals during cross-department collaboration(catering); procurement of stationery, water, and cleaning supplies (procurement); procurement of computers, printers, and other fixed-asset equipment (procurement); bulk procurement during a new off-site office setup (procurement).
Large-event organization	7	Unified accommodation for off-site employees attending an event (accommodation); formal lunch or banquet on the event day (catering); tea breaks, drinks, and snacks at the event venue (catering); procurement of brochures, display stands, and roll-up banners (procurement); transportation and meal subsidies for volunteers (subsidies); stage construction and lighting for an employee talent show (equipment rental); rental of temporary tents, tables, chairs, and sunshades (facility rental).
Total	44	—

Table 5: The REIMBURSEBENCH scenario taxonomy: six major corporate expense domains subdivided into 44 specific scenarios, distilled from ten expert interviews. The taxonomy guides workflow instantiation (Section 4.1, Stage 3) to ensure broad coverage of real auditing situations; it is not a scored axis of the benchmark.

Setting	Full success	Mean score	Flag	Amount	Complete logic
No tool, pass@1	4/51 (7.8%)	0.345	60.8%	13.7%	13.7%
No tool, pass@3	10/51 (19.6%)	—	80.4%	35.3%	25.5%
No tool, pass@5	12/51 (23.5%)	—	86.3%	45.1%	31.4%
Lookup + calculator, pass@1	6/51 (11.8%)	0.421	72.5%	23.5%	19.6%

Table 6: Diagnostic repeated-sampling and tool-use results for Qwen3.6-Plus on hard workflows.

Model	Vendor	Size	Setting
Kimi K2.6	Moonshot	1T	T=0.7, thinking
GPT-5.4	OpenAI	–	reason. effort=high
Claude Opus 4.6	Anthropic	–	thinking; default temp.
Qwen3.6-Plus	Alibaba	–	T=0.7, thinking
DeepSeek-V4-Pro [†]	DeepSeek	1.6T	T=0.7, thinking
Qwen3-VL-8B	Alibaba	8B	T=0.7
Qwen3-8B [†]	Alibaba	8B	T=0.7
Fino1-8B [†]	TheFinAI	8B	greedy
Hawkish-8B [†]	Mukaj	8B	greedy

Table 7: Evaluated models and decoding settings. [†] Text-only models; image attachments are processed by Qwen-VL-OCR-2025-11-20. Temperature is 0.7 for models that support it; Claude Opus 4.6 and GPT-5.4 are invoked through APIs that do not accept a user-set temperature in reasoning mode, so provider defaults are used. The 8B text-only systems do not use a separate thinking budget; Fino1-8B and Hawkish-8B are hosted locally while others are used through API.

ID	Group	Logic point	#WF	#SP	%SP	Easy	Medium	Hard
L1	A. Evidence Integrity	Invoice content completeness	18	20	9.0%	4	8	8
L2	A. Evidence Integrity	Supporting-document completeness	17	18	8.1%	5	9	4
L3	B. Cross-Doc Consistency	Claimant identity matching	5	6	2.7%	0	0	6
L4	B. Cross-Doc Consistency	Invoice type matching	12	12	5.4%	7	3	2
L5	C. Policy Compliance	Amount-cap violation	31	34	15.2%	6	13	15
L6	C. Policy Compliance	Invoice timing violation	18	18	8.1%	2	9	7
L7	C. Policy Compliance	Location deviation	15	15	6.7%	5	6	4
L8	C. Policy Compliance	Other substantive mismatch	29	32	14.3%	6	12	14
L9	D. Authenticity & Legitimacy	Invoice forgery	15	15	6.7%	1	3	11
L10	D. Authenticity & Legitimacy	Illegitimate purpose	31	35	15.7%	5	11	19
L11	D. Authenticity & Legitimacy	Duplicate reimbursement	11	12	5.4%	4	1	7
L12	D. Authenticity & Legitimacy	Required-amount miscalculation	6	6	2.7%	2	2	2
D1	Distractor	Irrelevant context	28	0	–	6	7	15
D2	Distractor	Blurred (but valid) invoice image	20	0	–	3	6	11

Table 8: Distribution of logic-point activations in REIMBURSEBENCH. **#WF**: number of workflows in which the logic point is activated at least once. **#SP**: number of scoring points tagged with this label. **%SP**: share of all L1–L12 logic-tagged scoring points. The right block reports per-difficulty occurrence counts (*with* the multiplicity from #SP, summed over workflows). For **D1** and **D2**, which tag no scoring points, the per-difficulty block instead counts the workflows in which the distractor appears, so the row sums to #WF.

Model	Overall Score $\bar{S}$			Parse Failures	
	ZK	EK	Δ	ZK	EK
Kimi K2.6	67.1%	68.7%	+1.6%	0	1
GPT-5.4	61.5%	64.2%	+2.6%	8	9
Claude Opus 4.6	70.7%	72.5%	+1.8%	2	0
Qwen3.6-Plus	64.5%	63.5%	-1.1%	0	0
DeepSeek-V4-Pro	65.0%	63.7%	-1.3%	0	1
Qwen3-VL-8B	45.2%	47.5%	+2.3%	1	1
Qwen3-8B	39.6%	39.9%	+0.3%	4	5
Fino1-8B	4.8%	4.0%	-0.8%	150	149
Hawkish-8B	6.5%	6.9%	+0.4%	141	140

Table 9: Effect of Expert-Knowledge (EK) vs. Zero-Knowledge (ZK) prompting. **Left:** Overall workflow score ($\bar{S}$), where $\Delta = \bar{S}_{EK} - \bar{S}_{ZK}$. **Right:** Parse failure counts (out of 151 workflows). A parse failure means the model’s output did not contain a parseable flag-max_legal pair, costing both scoring points for that workflow.

Model	Mode 1	Mode 2	Mode 3	Any
Kimi K2.6	0.159	0.185	0.172	0.364
GPT-5.4	0.185	0.113	0.172	0.344
Claude Opus 4.6	0.086	0.272	0.146	0.371
Qwen3.6-Plus	0.225	0.185	0.199	0.450
DeepSeek-V4-Pro	0.185	0.132	0.159	0.344
Qwen3-VL-8B	0.397	0.159	0.278	0.563
Qwen3-8B	0.536	0.073	0.278	0.656
Fino1-8B	0.636	0.119	0.318	0.742
Hawkish-8B	0.616	0.053	0.331	0.715

Table 10: Failure mode trigger rates across models (fraction of workflows where the mode was detected). “Any” indicates the fraction of workflows where at least one mode was triggered. For Fino1-8B and Hawkish-8B, a substantial share of their low-quality outputs cannot be reliably matched to any mode; their “Any Mode” rates therefore conflate genuinely mode-free responses with un-assessable ones and understate their true failure rate.

Shared System Instruction (ZK & EK)

You are a meticulous reimbursement audit assistant.

Now you will receive a reimbursement application for a business expense, which includes the background information of the transaction, the claimed reimbursement amount, relevant supporting documents and receipts, as well as the company's internal control policies or guidelines.

Your task is to analyze the provided documents, receipts, and policy files, then determine whether the claimed expense can be reimbursed, calculate the maximum legally reimbursable amount, and explain the entire reasoning chain step by step.

EK-Only Expert Knowledge (appended only in EK, omitted in ZK)

Core Audit Principles

As a corporate auditor, you operate under three fundamental principles:

- **Substance over form:** judge the economic reality of a transaction, not its surface appearance.
- **Verifiable evidence:** every reimbursable amount must be traceable to a specific, valid piece of evidence in the submitted materials. Do not assume facts that are not in evidence.
- **Policy adherence:** apply the company's internal control rules exactly as written. Do not invent rules, and do not weaken rules to favor the applicant.

Standard Audit Workflow

A complete reimbursement audit proceeds in four sequential stages. For each receipt, you must verify all four stages; failure at any stage affects either the receipt's validity or its reimbursable amount.

Stage 1 – Authenticity Check (is the receipt itself genuine?)

Verify that each receipt is a legitimate document, not fabricated or tampered with. Do not dismiss a receipt as fake solely due to minor font or layout inconsistencies; instead, focus on substantive logical discrepancies, such as contradictory content, impossible dates, or values that cannot be physically read from the image. An illegible or unreadable receipt cannot serve as audit evidence.

Impact: A receipt that fails authenticity contributes **0** to the reimbursable amount (Step 1 of MAX_LEGAL calculation).

Stage 2 – Compliance Check (does the receipt match the claimed expense and policy scope?)

For each receipt, verify that:

- it contains all required fields per company policy;
- its expense category matches what is being claimed;
- its date falls within the valid period of the business activity;
- its location is consistent with the business activity's location;
- the personnel involved match the names and headcount declared in the application;
- any other policy-specific requirements are met.

Impact: A receipt that fails compliance contributes **0** to the reimbursable amount (Step 1 of MAX_LEGAL calculation).

Stage 3 – Supporting Documentation Check

Many expense types require supporting documents beyond the receipts themselves. The specific requirements are defined in the company's policy.

Impact: Missing supporting documents indicate a procedural defect. Whether it directly invalidates the underlying receipt's monetary value depends on the policy; if the policy does not specify, treat it as a problem to flag but do not automatically reduce MAX_LEGAL.

Stage 4 – Amount Integrity Check

For each surviving receipt and for the application as a whole, verify that:

- per-receipt or per-day caps are respected;
- no receipt is being claimed more than once (no duplicate reimbursement);
- the expense is for a legitimate business purpose, not personal use or padding.

Impact: An over-cap portion is reduced to the cap (Step 2 of MAX_LEGAL calculation). A duplicate or non-business item should be removed (Step 1).

Table 11: The prompt template of system message. The shared system instruction is identical in both ZK and EK; the block labeled "EK-Only Expert Knowledge" is injected into the system message exclusively in the EK setting.

Shared User Message Template

Reimbursement Description: {context}

Claimed reimbursement amount: {required_amount} {currency}

IMPORTANT INSTRUCTIONS (read carefully BEFORE examining the materials)

1. **Receipt ownership:** The person responsible for each receipt is solely determined by the label provided in the file description (e.g., 'Receipt belonging to **Zhang_Ming**'). Disregard any names or identity information printed on the receipt image or PDF itself when assigning the expense to an individual. Always use the person specified in the label.
2. **Strict policy adherence:** Apply rules strictly based on the provided policy documents. Do not invent rules. Do not assume facts that are not in evidence.
3. **Attention to detail:** Pay close attention to dates, locations, personnel levels, amount limits, and any conditional clauses in the policy.
4. **Ignore submission date:** Assume all reimbursement applications are submitted within the legal time frame.

MATERIALS

Below are all the background materials and receipts needed for the audit, provided in order:

[...sequence of text and image blocks extracted from the submitted files, with labels indicating receipt ownership...]

Shared OUTPUT FORMAT

OUTPUT FORMAT

Produce your response in EXACTLY the following format.

Definitions

Let **CLAIMED** = the reimbursement amount the applicant submitted (given above).

Let **MAX_LEGAL** = the maximum amount that would be reimbursable under the policy, computed as follows:

Step 1: Discard any individual receipt that you determine to be entirely invalid based on the provided materials and policy. These receipts contribute 0.

Step 2: For each remaining valid receipt, cap its reimbursable portion at the applicable policy limit (e.g., if the per-night hotel cap is 100 and the receipt is 120, only 100 is reimbursable from that receipt).

Step 3: Sum the capped amounts from Step 2. The result is **MAX_LEGAL**.

Note: **MAX_LEGAL** is 0 only if every receipt is fully invalid or there is no valid claim portion at all.

The FLAG

FLAG is **strictly** the comparison result between **MAX_LEGAL** and **CLAIMED**:

- FLAG = 1 if $\text{MAX_LEGAL} \geq \text{CLAIMED}$ (the requested amount is fully supported by the materials and policy).

- FLAG = 0 if $\text{MAX_LEGAL} < \text{CLAIMED}$ (the requested amount exceeds what the materials and policy allow).

FLAG is a deterministic comparison: compute **MAX_LEGAL** first, then output FLAG accordingly.

Required output structure

Line 1: 0 or 1 – the FLAG value defined above.

Line 2: A single numeric value – **MAX_LEGAL** in the SAME currency as the claim.

Line 3: The exact text – Reasoning –

Following lines: A detailed, step-by-step reasoning chain that explains:

- (a) Step 1 results: which receipts (if any) are entirely invalid, and why (cite policy clauses).
- (b) Step 2 results: for each valid receipt, show the raw amount, the applicable cap, and the capped amount.
- (c) Step 3 results: the arithmetic that produces **MAX_LEGAL**.
- (d) The FLAG: explicitly compare **MAX_LEGAL** vs **CLAIMED** and state the resulting FLAG.
- (e) Any discrepancies, missing documents, policy violations, or calculation concerns you identified.

After the reasoning, add a line with the exact text – Problems (if any) –

Then list each problem identified during the reasoning stage that concerns the submitted reimbursement, one per line, in the format: Problem: <description>. Each problem description should clearly state (i) what the problem is, (ii) which specific receipt, document, or person it concerns, and (iii) the numeric or factual detail that demonstrates it (e.g., the amount, the date, or the policy clause being violated).

If no problems exist, write only: No problem.

Important: The reasoning must include thorough citations of policy clauses and explicit arithmetic. Do not skip any step.

Table 12: The prompt template of user message. The content in this part is identical in both ZK and EK.

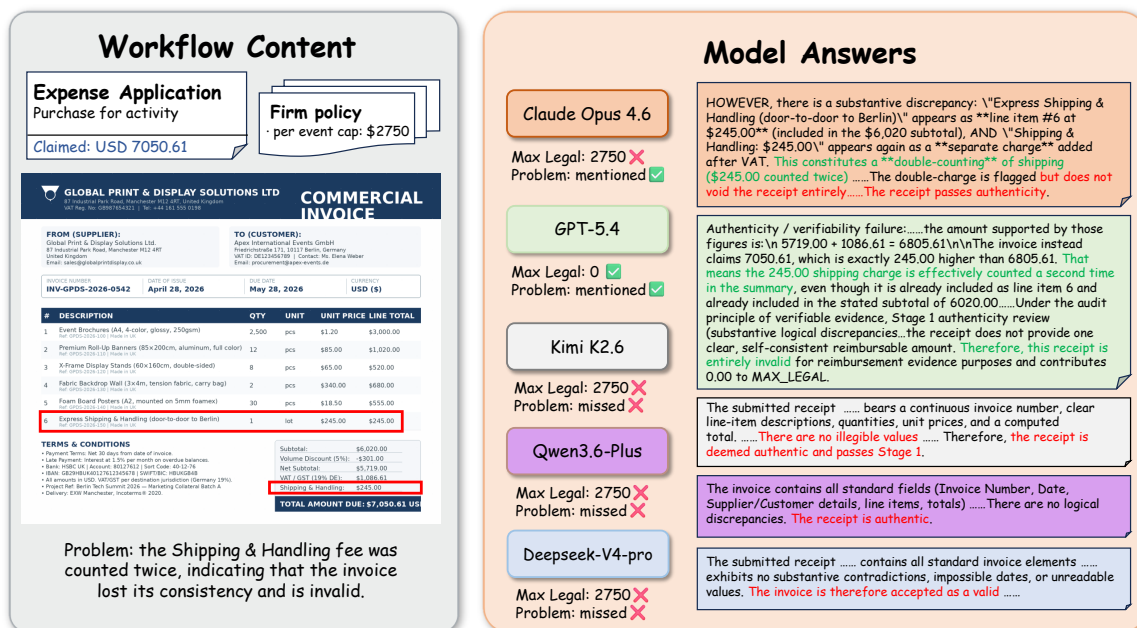


Figure 6: **Mode 1 (Content Consistency Blindness)**: The shipping fee is counted twice on the invoice, disrupting internal arithmetic consistency. Most models fail to notice this surface-level contradiction.

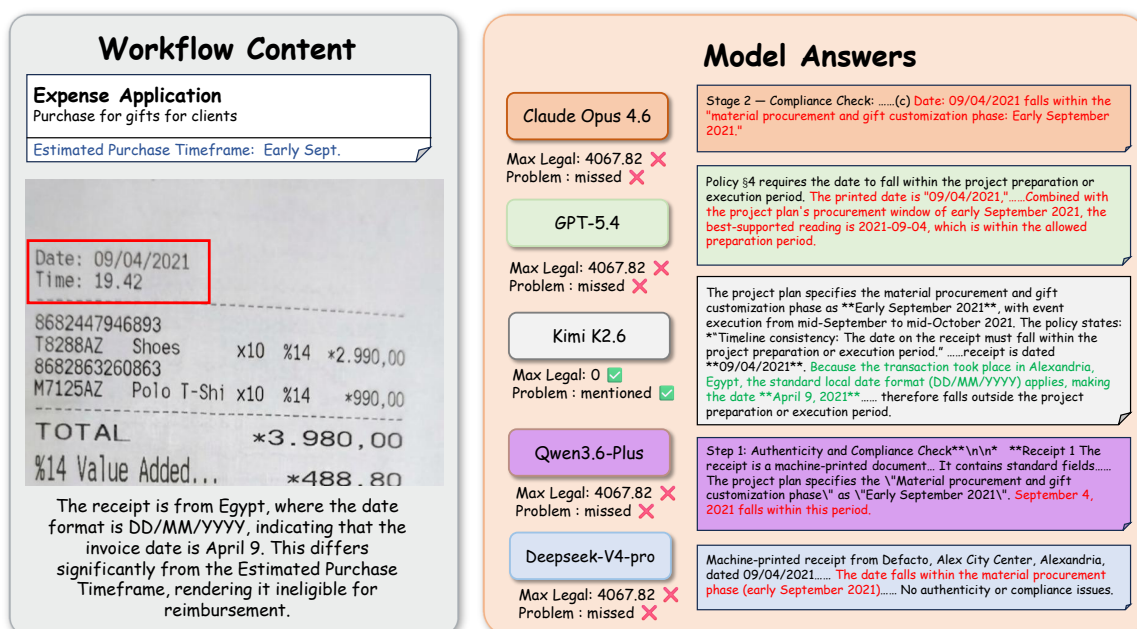


Figure 7: **Mode 1 (Content Consistency Blindness)**: An invoice from Egypt uses the DD/MM/YYYY format, placing the date outside the valid project timeframe. Models blindly accept the date without cross-checking regional format conventions.

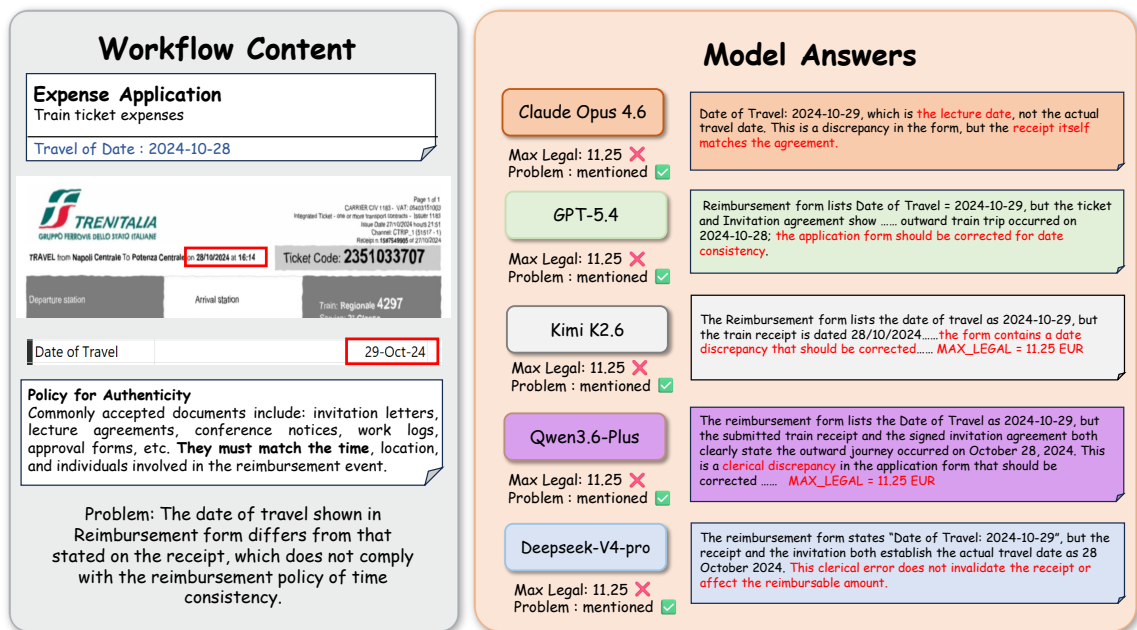


Figure 8: **Mode 2 (Conflict Self-Rationalization)**: The travel date on the reimbursement form contradicts the receipt. The models detect the issue but inappropriately rationalize it as a “clerical discrepancy” to approve the claim.

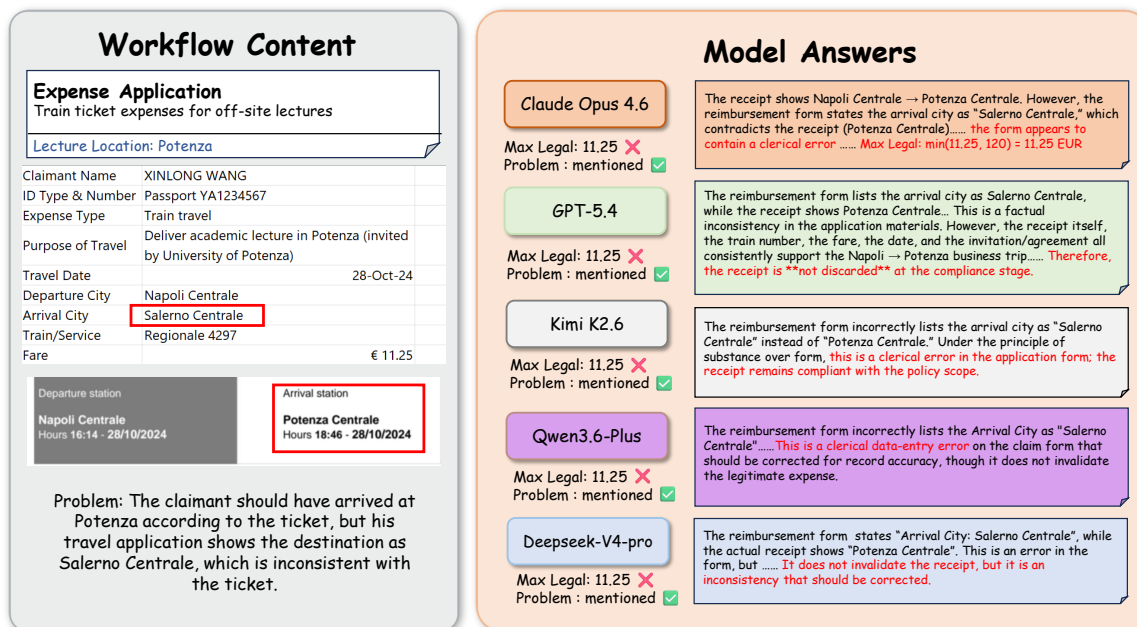


Figure 9: **Mode 2 (Conflict Self-Rationalization)**: The destination city on the ticket does not match the application. Models again unilaterally invent excuses (e.g., “clerical data-entry error”) to resolve the conflict.

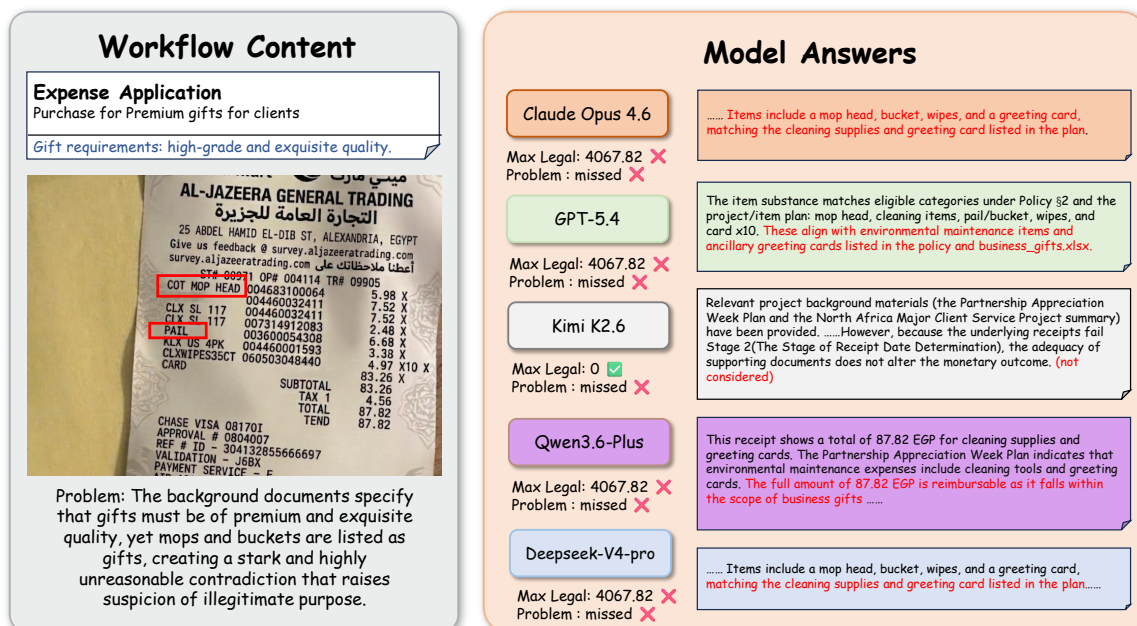


Figure 10: **Mode 3 (Behavioral Intention Reasoning Defect)**: Mops and buckets are submitted as “premium client gifts.” The models verify the category match but fail to question the implausible business purpose.

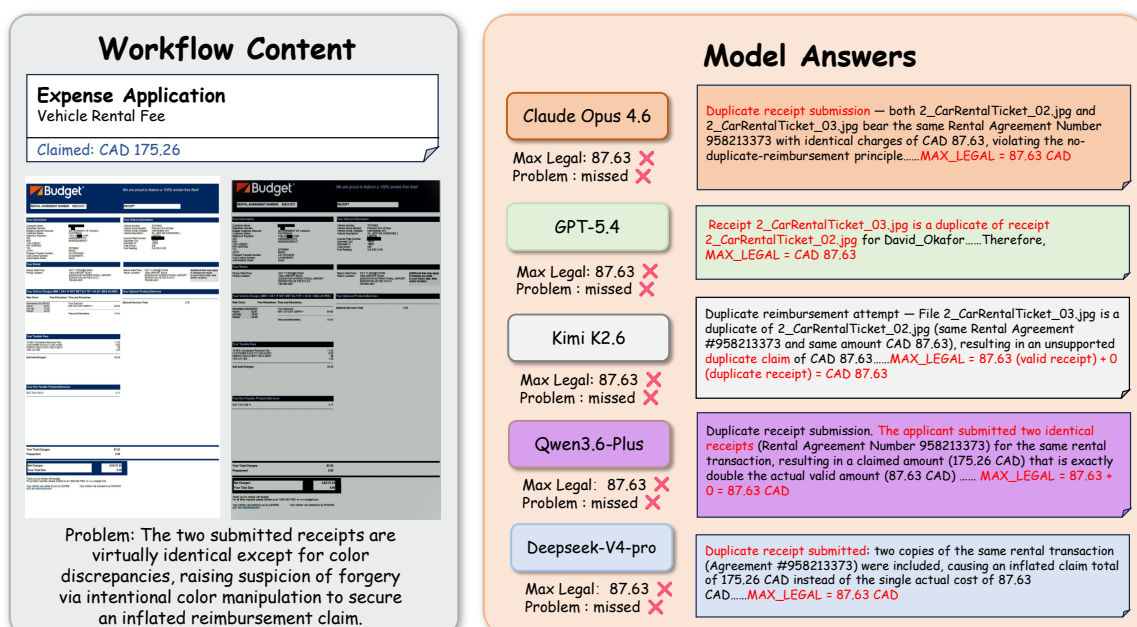


Figure 11: **Mode 3 (Behavioral Intention Reasoning Defect)**: Two matching receipts with color manipulation are submitted. Models flag a routine duplicate but miss the evidence of possible deliberate document manipulation.