FCMBENCH: A LARGE-SCALE FINANCIAL CREDIT MULTIMODAL BENCHMARK FOR REAL-WORLD APPLICATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce FCMBench, a large-scale multimodal benchmark designed for understanding identity-linked documents in financial credit evidence review applications. It contains 4,922 real-captured images and 11,045 VQA instances covering 26 certificate types, three perception tasks, four credit-specific reasoning tasks, and ten real-world robustness challenges. To address the privacy–realism trade-off in financial data construction, we develop a synthetic-to-physical pipeline that creates linked fictional applicant profiles, generates certificate templates, physically fabricates documents, and recaptures them under controlled scenarios. Task design is grounded in interviews with senior credit reviewers, and a two-stage audit refines data quality. We evaluate 39 proprietary and open-source vision–language models from 14 organizations. Results show substantial performance gaps among current VLMs, with overall F1 scores ranging from 24.60 to 71.57 (mean 50.9; SD 12.6 across models). Matched robustness evaluations further reveal consistent degradation under challenging conditions, particularly for multi-document compositions, small regions of interest, and defocus. Extensive experimental results demonstrate that FCMBench is challenging and highly discriminative for identifying performance gaps among vision–language models.

1 INTRODUCTION

Financial credit review relies on heterogeneous evidence submitted by borrowers, including identity, income, asset, tax, and transaction documents. Reviewers must verify legibility and completeness, extract operationally precise fields, associate evidence from the same applicant, apply validity rules, and assess cross-document consistency. These processes couple perception and reasoning: an incorrect financial decision may originate from unreadable text, missing evidence, or failed document association. Therefore, a meaningful benchmark should evaluate the entire evidence-handling pipeline rather than isolated OCR or question answering.

The input process introduces another challenge. Credit materials are often captured by consumer smartphones rather than provided as curated scans in real-world applications, resulting in perspective distortion, uneven illumination, reflections, blur, small text, cropping, clutter, and multi-document layouts. Meanwhile, authentic application records are difficult to release due to sensitive personal and institutional information. Although fully digital synthetic documents protect privacy, they fail to reproduce the optical and environmental artifacts that affect real-world performance. This privacy–realism trade-off motivates a controlled synthetic-to-physical data construction strategy.

Existing multimodal benchmarks cover complementary but largely independent capabilities: general-purpose suites lack credit-specific workflow tasks, document benchmarks focus on OCR or chart understanding, and financial benchmarks evaluate financial knowledge and analytical reasoning (detailed in Related Work). However, none jointly capture three critical characteristics of credit scenarios: compliant identity-linked certificates, application-driven decision making tasks and physical capture artifacts.

FCMBench addresses this gap by providing a specialized benchmark with 11,045 annotated samples from 4,922 real-captured images covering 26 types of official certificates (Figure 1). All tasks are derived from practical credit application workflows. We manually designed certificate templates,

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.

fabricated physical specimens, and collected real-world captures, ensuring both authenticity and privacy preservation.

Our contributions are threefold:

1. **A large-scale synthetic-to-physical credit image corpus.** We built FCMBench based on 11,045 annotated samples covering 26 types of Chinese and English identity documents. We reproduced these physical paper documents and re-shot them with 11 volunteers using five distinct smartphone models, thereby eliminating the risk of leakage of users' sensitive personal data.
2. **Real-world workflow-grounded tasks.** Based on interviews with over 20 senior reviewers, we design seven tasks covering document collection, profile construction, compliance verification, and risk assessment, with nine acquisition challenges and one multi-document stress test reflecting practical scenarios.
3. **Multi-Dimensional Evaluation.** We conduct comprehensive experiments from multiple perspectives on 39 state-of-the-art vision-language models developed by 14 institutes. The results show that FCMBench brings substantial challenges to existing model architectures and possesses robust discrimination capacity to distinguish performance gaps among different VLMs.

2 Related Work

General multimodal evaluation. MMMU (Yue et al., 2024), MMMU-Pro (Yue et al., 2025), MMStar (Chen et al., 2024), and MME (Fu et al., 2025) evaluate broad perception and reasoning across many domains while addressing issues such as visual redundancy, difficulty, and data leakage. Their breadth supports general capability comparison, but does not model the linked evidence, document rules, and acquisition process of a credit application.

Document and financial benchmarks. OCRBench (Liu et al., 2024), ChartMuseum (Tang et al., 2025), ChartQAPro (Masry et al., 2025), MMDocBench (Zhu et al., 2024), and WildDoc (Wang et al., 2025a) expose weaknesses in OCR, chart reasoning, multilingual document understanding, and in-the-wild distortions. Financial benchmarks including CFBenchmark-MM and VisFinEval (Li et al., 2025; Liu et al., 2025), FinMME (Luo et al., 2025), FinMR (Deng et al., 2025), and MultiFinBen (Peng et al., 2025) extend evaluation to financial knowledge, figures, tables, reports, and multilingual analysis. These resources establish important component capabilities, but their primary evaluation units are not linked applicant-document sets spanning evidence collection, profile construction, compliance checks, and credit-risk review under controlled physical capture conditions.

Positioning of FCMBench. FCMBench contributes to the unified implementation of identity-linked evidence, application-grounded tasks, and controlled synthetic-to-physical capture challenges within a single benchmark. As a large-scale multimodal benchmark focused on financial credit evidence review, FCMBench features the above attributes.

3 Benchmark Construction

3.1 Overview and Statistics

The images in FCMBench comprise 26 categories of certificates spanning both Chinese and English documents commonly used in credit review (Figure 1). The benchmark is guided by two design principles: (1) *alignment with real credit review workflows* and (2) *diversity of data forms*. Specifically, the certificate categories and task designs are grounded in interviews with over 20 senior credit reviewers from commercial banks and financial credit companies, ensuring practical relevance across the full review chain (e.g., loan review, income verification, and asset evaluation). Moreover, FCMBench incorporates heterogeneous document forms from standardized cards and certificates to materials containing complex charts and tables, and supports both single-image and multi-image inputs, reflecting the variety of evidence encountered in real applications.

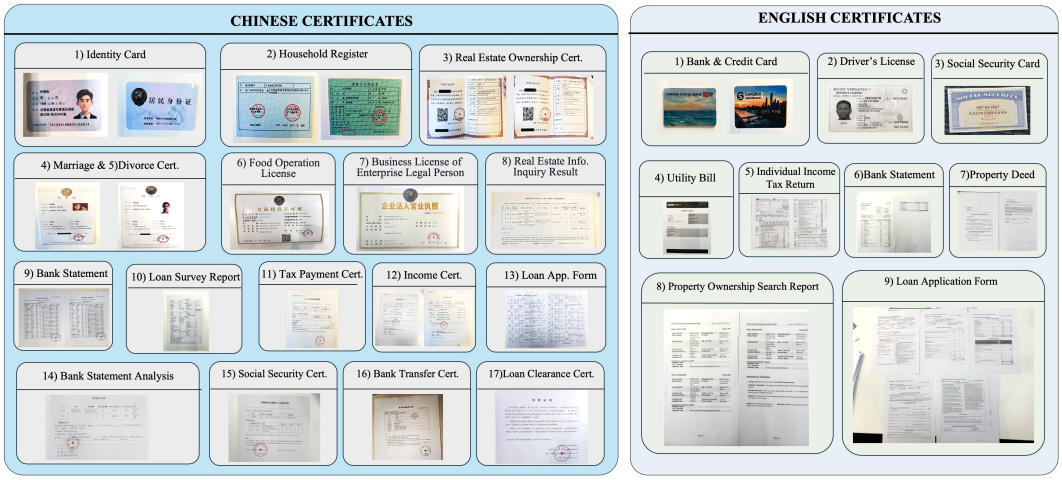


Figure 1: Overview of the 26 categories of certificates in FCMBench.

Task	Description	# Images	# QA pairs
IQE	**Image Quality Evaluation**: To recognize quality issues (e.g., specular reflection, out-of-focus blur) on the certificates in the image.	759	759
DTR	**Document Type Recognition**: To recognize the certificates contained in the images. An image may include one or more certificates.	3635	4506
KIE	**Key Information Extraction**: To extract key information or key-value pairs in the certificate from the given images.	2642	2561
CC	**Consistency Checking**: To check whether the documents in an image set belong to the same person, or to cross-check a listing document has its corresponding supporting document.	1501	779
VC	**Validity Checking**: To check whether documents have valid values, e.g., document has not expired, values follow required formats.	977	922
NC	**Numerical Calculation**: To compute numbers based on the information presented in the images.	922	1266
RR	**Rationality Review**: To check if the values presented in different documents are within reasonable range, e.g., income vs tax certificate, income vs bank statement.	416	252

Table 1: Taxonomy and Statistics of *Perception* and *Reasoning* Tasks.

The tasks in FCMBench are organized by capability (perception or reasoning) and input condition (normal or challenging). Perception tasks assess visual and textual evidence acquisition, whereas reasoning tasks require decisions grounded in that evidence. Tables 1 and 2 summarize the taxonomy. Because an image may support multiple tasks and conditions, their image counts overlap and should not be summed as unique images.

3.2 Image Generation

To reduce privacy exposure while retaining practical capture artifacts, FCMBench uses a controlled synthetic-to-physical workflow with three stages.

1. Certificate Template Construction: We propose a high-fidelity synthetic pipeline to generate credit-related documents and physical cards visually comparable in layout and material to real counterparts. We create fictional identities with diverse demographic and economic traits—gender, language, income source, age, marital status, annual income, and properties owned—and assemble multi-document application profiles featuring consistent names, IDs, addresses and financial

Challenge Name	Description	# Images	# QA pairs
Normal Captures	A ”standard” photo of the certificates that satisfies all quality requirements.	508	1002
Off-axis Viewpoints	The certificates are captured from rotated viewpoints.	491	1024
Uneven Illumination	Various shadows cast onto the certificates.	509	1129
Specular Reflections	Overexposure caused by reflected light.	509	1051
Out-of-focus	Blur caused by defocus.	460	840
Small ROIs	The target occupies less than 20% of the image.	467	881
Secondary Captures	Photos are captured from other on-screen images (e.g., computer displays) rather than the original certificates.	288	397
Cluttered Background	The backgrounds of the target certificates share similar characteristics, which may affect the recognition of the certificates.	454	929
Overlaid Watermarks	Watermarks on the image or on the certificates may affect the recognition.	495	926
Cropped Captures	The certificates shown in the image are incomplete.	461	700
Multi-doc Images	Models must perform tasks involving multiple certificates captured in a single image.	774	2166

Table 2: Taxonomy and statistics of Normal Captures and *Robustness* challenges. The first nine challenges are acquisition artifacts; Multi-doc Images is an input-composition stressor.

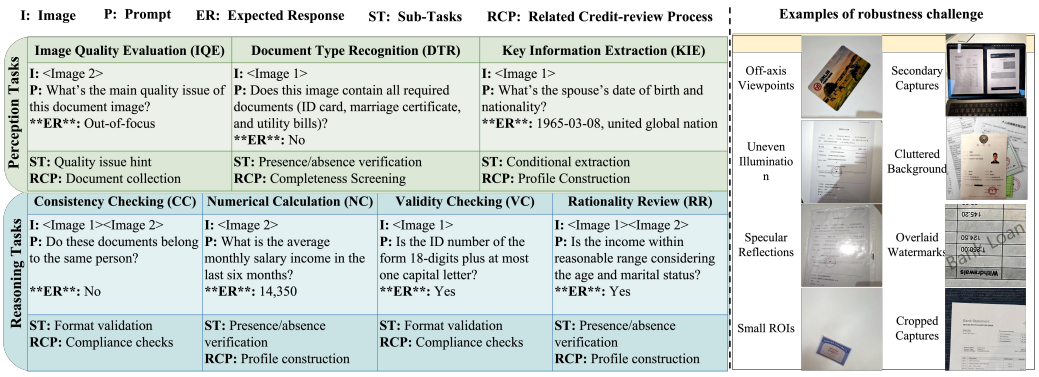


Figure 2: Examples of tasks and robustness images, showing core prompt (omitting options and format requirements), expected response, and related real-world credit evidence review workflows. `<Image>` indicates the placeholder of an image.

attributes across all synthetic certificates. Uniform base templates are used for each certificate type across identities, removing redundant duplicate images and repetitive capture samples. All synthetic outputs follow industry-standard layouts, with fully adjustable template designs, capture parameters and noise perturbations.

2. Compliance Integration: To avoid releasing customer records, personal attributes and institutional names are fictional, and most logos, emblems, and portraits are replaced with generated alternatives. We also construct fictional streets, agencies, and banks. Fifteen synthetic English-language driver's licenses are derived from IDNet (Xie et al., 2024) and are redistributable under its source terms; all other benchmark materials are produced in-house. No confidential image used for validation is included in the benchmark release.

3. Physical Fabrication and Scene-based Shooting: Electronic templates, including the IDNet-derived licenses, are fabricated as card replicas or printed on paper matching the size of authentic

official documents. Eleven participants then photograph them under the conditions in Table 2, introducing physical effects such as perspective, illumination, blur, reflection, scale, and background clutter. Images are captured with five major smartphone brands (iPhone, Huawei, Honor, Xiaomi, and Oppo) to incorporate device heterogeneity.

3.3 Instruction Design

Tasks in Table 1 are abstracted from credit-review workflows identified through interviews with more than 20 senior reviewers. Examples can be seen in Figure 2, and they cover the major processes in real-world credit review applications, including *Document Collection*, *Completeness Screening*, *Profile Construction*, *Cross-Document Consistency*, *Compliance Checks*, and *Income Rationality and Risk Assessment*.

Perception Task Design Perception tasks model front-end evidence acquisition. Image Quality Evaluation (IQE) corresponds to the desk-reject stage and identifies whether a submission is usable or affected by common capture defects. Document Type Recognition (DTR) supports completeness screening through single-image classification, multi-image aggregation, and present/missing-document judgments. Key Information Extraction (KIE) supports profile construction by extracting required fields, key–value pairs, and conditional information about identity, income, property, and accounts.

Reasoning Task Design Reasoning tasks model evidence integration and rule application. Consistency Checking (CC) compares identities, addresses, supporting certificates, and transactions across documents. Validity Checking (VC) applies explicit issue/expiry-date, format, and value constraints. Numerical Calculation (NC) aggregates or transforms evidence to compute totals, differences, income statistics, and collateral indicators. Rationality Review (RR) assesses whether values across income certificates, tax records, applications, and bank statements are mutually consistent and economically plausible. These tasks evaluate evidence-grounded review operations rather than autonomous loan-approval decisions.

Tasks are instantiated under normal captures, nine acquisition challenges, and a multi-document composition stressor (Table 2). The latter changes input structure rather than photographic quality and is therefore interpreted separately when diagnosing robustness.

3.4 Ground-Truth Answerability Audit

Because capture artifacts can obscure evidence required to answer a question, we conducted an answerability audit for every image–question–answer triple. Trained reviewers first assessed whether the reference answer could be recovered from visible evidence under the stated prompt and whether the instance was unambiguous. A senior credit-domain expert then performed a second-pass verification. Cases judged unanswerable, ambiguous, or unsupported by visible evidence were removed; uncertain cases were escalated to the senior expert for adjudication.

Of 14,179 audited triples, 11,045 (77.90%) met the predefined answerability criterion and were retained, while 3,134 (22.10%) were removed. These figures report the audit retention outcome rather than inter-annotator agreement. Although answerability necessarily involves human judgment, the two-stage review process reduces confounding from missing or unrecoverable evidence. Together, these controls make acquisition conditions traceable and visual evidence recoverable under a machine-checkable answer protocol, so the benchmark scale reflects verified construction rather than prompt expansion over a small collection of unrelated images.

3.5 Evaluation Metrics

Many credit fields (e.g., identifiers, account numbers, and dates) are operationally atomic, so prompts specify output formats and evaluation is exact-match oriented. Parsed outputs are flattened into order-invariant key–value sets S: dictionary keys receive hierarchical prefixes, nested lists become normalized tuples, strings are stripped of superficial formatting, and equivalent numeric forms are canonicalized. Values must match exactly except in NC, where, after conversion to the unit specified by the prompt, a numerical prediction matches its ground truth when their absolute difference

is below 2; this tolerance accommodates rounding of derived quantities and never applies to identifiers, dates, text, or categorical outputs. We compute $P = |S_p \cap S_g|/|S_p|$, $R = |S_p \cap S_g|/|S_g|$, and $F1 = 2PR/(P + R)$. Empty predictions score zero unless both sets are empty, in which case F1 is one. Instance F1 is averaged within each fine-grained subtask and then macro-averaged to obtain each of the seven task scores. Perception and Reasoning average their three and four task scores, respectively, while Overall is the unweighted macro-average of all seven tasks. Invalid JSON receives zero.

For robustness condition c, let $\mathcal{S}_c$ be the fine-grained subtasks represented in both c and Normal Captures. For each model we compute macro-F1 on c and on Normal Captures over the same $\mathcal{S}_c$, take their ratio, and average the ratios across models, so robustness comparisons hold subtask composition fixed; Multi-doc Images remains a separate input-composition stressor rather than a photographic degradation.

4 Experiments

4.1 Experimental Setup

We evaluate 39 SOTA VLMs from 14 organizations categorized into commercial closed-source models and open-source models. The commercial set (10 models) includes Qwen3.8-Max (Qwen Team, 2026d), Gemini 3 series (3.5 Flash, Pro Preview, Flash Preview) (DeepMind, 2026), Claude Opus (4.8, 4.5) (Anthropic, 2026), Qwen3.7-Plus (Qwen Team, 2026e), GLM-5V-Turbo (Team, 2026b), and GPT-5 (5.2, 5.1) (OpenAI, 2026). The open-source collection (29 models) covers Kimi series (K3, K2.5, K2.6, Kimi-VL) (Team, 2026c), Qwen3.8-Flash (Qwen Team, 2026d), Qwen3.5 (27B, 35B-A3B, 122B-A10B, 397B-A17B) (Qwen Team, 2026a), Qwen3.6 (27B, 35B-A3B) (Qwen Team, 2026b;c), Qwen3-VL with multiple parameter variants (8B, 30B-A3B, 32B, 235B-A22B) (Bai et al., 2025), GLM-4.5V/GLM-4.6V (Team, 2025), InternVL3.5 (8B, 30B-A3B, 38B, 241B-A28B) (Wang et al., 2025b), Ovis2.5-9B (Lu et al., 2025), DeepSeek-V4-Flash-Vision-Exp (Team, 2026a), Llama 4 Maverick (AI, 2026), MiniMax-VL-01 (Team, 2026e), MiniCPM-V 4.5 (Team, 2026f), LLaVA-OneVision-1.5-8B (Team, 2026d), Janus-Pro-7B (Chen et al., 2025), and Phi-4-multimodal-instruct (Research, 2025). All models use the same instances, prompts, and output-format requirements. All commercial models and open-weight models served through hosted APIs are evaluated via official APIs; the remaining open-source models are deployed on cloud GPUs (per-model deployment in Table 9). We use deterministic decoding where supported, default parameters, and image resizing per model limits. For FCMBench, instruction-tuned variants and the lowest reasoning setting are adopted. All scores use a unified parser and hierarchical macro-averaging protocol.

4.2 Experimental Results and Analysis

4.2.1 Overall Performance

The left part of Figure 3 shows Overall F1 ranging from 24.60 to 71.57 (mean 50.9 ± 12.6) across the 39 models: FCMBench is non-trivial and separates evaluated systems under a common protocol.

Table 3 reports the top-5 (overall) commercial and open-source models by Overall F1. Kimi K3, the strongest open-source model, attains the highest Overall (71.57), narrowly ahead of the best proprietary model Qwen3.8-Max (71.24); Gemini 3.5 Flash (70.84) and the open-weight Qwen3.8-Flash (70.16) follow. The four leaders are separated by at most 1.41 points. Task-level leadership spans four models: Kimi K3 leads KIE (52.49), CC (77.75), NC (70.28), RR (83.15), and Overall; Qwen3.8-Max leads IQE (61.16); Qwen3.8-Flash leads VC (94.01); and Gemini 3 Flash Preview leads DTR (95.32). No single model dominates all tasks, and displayed Overall scores span 60.28–71.57 while the highest task score is 95.32, so FCMBench separates frontier systems without saturating.

The baseline rows validate this design. Removing images collapses both text-only models to 9.77–30.54 on IQE, DTR, and KIE, whereas the same models reach 45.42–82.37 with images (gains of +22.0 to +66.4), confirming that the synthetic-to-physical recapture supplies visual evidence that language priors cannot substitute. Yet text-only scores stay high on VC and RR (54.48–59.31) and still gain +20.9 to +36.9 with images, indicating these tasks encode compliance and rationality rules

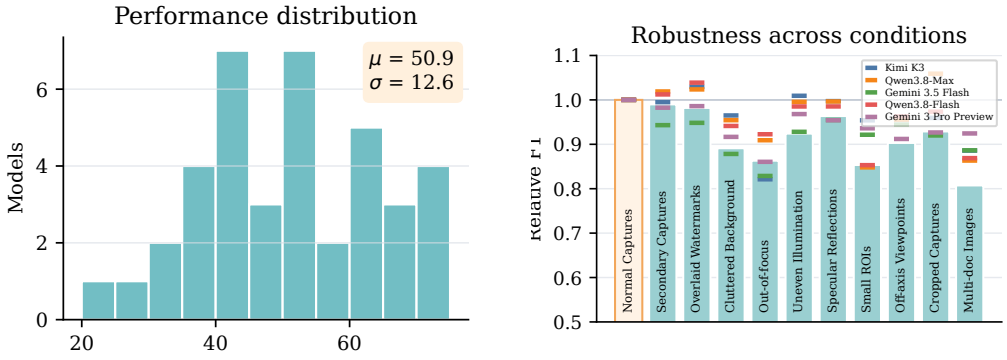


Figure 3: *Left*: Distribution of Overall F1 across evaluated models. *Right*: Relative F1 under each condition, normalized to matched Normal-Capture subtasks. Bars show the 39-model mean; ticks show the five highest-scoring models.

Model	IQE	DTR	KIE	CC	VC	NC	RR	Overall
Commercial Models								
Qwen3.8-Max	**61.16**	82.37	45.42	74.93	93.83	62.68	78.32	71.24
Gemini 3.5 Flash	52.92	91.37	49.23	74.15	90.60	62.19	75.43	70.84
Gemini 3 Pro Preview	43.78	90.33	45.23	61.98	91.43	63.71	72.70	67.02
Claude Opus 4.8	58.19	74.61	51.95	63.06	86.89	61.45	72.87	67.00
Gemini 3 Flash Preview	51.67	**95.32**	45.86	76.18	76.09	60.76	61.89	66.82
Open-Source Models								
Kimi K3	45.94	78.39	**52.49**	**77.75**	93.00	**70.28**	**83.15**	**71.57**
Qwen3.8-Flash	54.07	76.15	47.02	72.52	**94.01**	64.40	82.95	70.16
Kimi-K2.5	44.70	79.12	46.94	68.68	70.37	59.80	72.91	63.22
Kimi-K2.6	43.15	74.96	47.15	73.43	66.79	57.55	67.12	61.45
Qwen3.6-35B-A3B	45.73	77.81	44.13	72.29	62.49	63.51	56.00	60.28
Baselines								
Random Answer (100 seeds)	9.90	17.58	30.68	39.30	61.16	19.19	50.13	32.56
Constant Answer	6.12	9.19	9.27	16.67	32.20	14.45	45.87	19.11
Qwen3.8-Max (text only)	11.42	15.98	22.22	37.87	56.90	14.38	57.39	30.88
Kimi K3 (text only)	9.77	19.01	30.54	38.98	59.31	14.99	54.48	32.44

Table 3: Task-level F1 (%) for the top-5 (overall) commercial and open-source models is obtained using the common evaluation protocol. Random Answer is the mean of 100 fixed-seed, full-set label-prior runs. Prompt-format Constant Answer emits the most frequent valid JSON example in each fine-grained subtask's output-format instruction. Qwen3.8-Max (text only) and Kimi K3 (text only) receive the normal prompt text but no image attachments. Baseline rows are excluded from model rankings; best/second-best highlighting is computed over all 39 evaluated models.

that frontier models have internalized; so FCMBench measures internalized domain knowledge and evidence-grounded verification as separable, additive components. The label-prior floors (e.g., Constant Answer 45.87 on RR) reflect the closed, often binary, decision spaces of real credit checks rather than task triviality: the best models exceed these floors by 37.3 (RR) to 86.1 (DTR) points, and no task saturates (maximum 95.32), confirming that FCMBench is challenging yet appropriately scaled. Chance-correction against the Random baseline leaves the 39-model ordering nearly unchanged ($\rho = 0.998$; Appendix E.8), though raw VC and RR scores are best read relative to their label-prior floors.

4.2.2 Performance over Robustness Challenges

The right part of Figure 3 reports relative F1 against the matched Normal-Capture subtasks defined in Section 3.5. Multi-doc Images incur the largest degradation (relative F1 0.81, i.e.a 19% loss), followed by Small ROIs (0.85) and Out-of-focus captures (0.86). Cluttered Background and Off-axis

Viewpoints reach 0.89 and 0.90, respectively, whereas Secondary Captures and Overlaid Watermarks retain 0.99 and 0.98 of matched Normal-Capture performance. The principal stressors are therefore not uniformly cosmetic: evidence density, reduced target scale, and lost sharpness damage performance more than surface-level transformations. Table 4 corroborates this at question level: Multi-doc Images have the highest zero-correct rate (27.3%, vs. 6.7% for Normal Captures), while Small ROIs, Off-axis Viewpoints, and Out-of-focus captures are also elevated. Even leading models thus remain vulnerable to multi-document inputs, reduced target scale, and severe blur.

4.2.3 Independent Human Baseline

To quantify expert-level human performance under unified scoring rules, four credit-domain experts uninvolved in dataset curation independently annotated a stratified sample. We evaluate 275 overlapping prompts between this human study and the 11,045-instance FCMBench set, supporting human-model comparisons on identical items; this subset covers six of the seven tasks (all but RR) and diverse inputs, and differs in composition from the full benchmark. Annotators saw only raw images and prompts (no ground truth or model outputs), followed the same output requirements as models, and were scored by the standard parser. We report the average task-macro Overall F1 across experts, with uncertainty from 10,000 identity-clustered bootstrap replicates. On this subset, experts achieve an Overall F1 of 91.98 (95% CI: [89.34, 94.51]), while the top-performing model Qwen3.8-Max attains 66.67. This gap suggests models underperform domain experts on these tasks, though these subset scores are not directly comparable to the full-benchmark results in Table 3.

4.2.4 Transfer to Confidential Real Credit Documents

To assess whether FCMBench performance predicts relative model utility on real credit evidence, we evaluate nine locally deployed VLMs (Qwen3.8-Flash, Kimi-K2.6, Qwen3.6-35B-A3B, GLM-4.5V, Ovis2.5-9B, InternVL3.5-30B-A3B, Qwen3-VL-30B-A3B-Instruct, MiniCPM-V 4.5, LLaVA-OneVision-1.5-8B) on a confidential real-world credit-document set of 356 images and 385 prompts. Adopting FCMBench's prompt format, parser, and subtask-macro F1 protocol, and noting that the real-world tasks originate from business workflows with a different distribution, we prioritize rank agreement over absolute score matching. Figure 4 demonstrates strong rank correlation: FCMBench and real-world rankings yield Spearman's $\rho = 0.93$ and Kendall's $\tau_b = 0.83$, four out of nine models retain identical ranks, and the mean absolute rank difference equals 0.67. Raw Overall scores additionally exhibit strong linear association (Pearson $r = 0.97$, $R^2 = 0.93$) as a supplementary observation. Although adjacent models occasionally swap ranks, systems better-performing on FCMBench typically outperform others on the confidential real credit dataset. These results provide task-level evidence that our synthetic-to-real benchmark serves as a useful proxy for relative model utility in credit document review, within the scope of the tested models and data.

4.2.5 Error Analysis

We characterize strict errors at both instance and question levels without attributing causes. For each model–instance pair with F1 below one, we assign a mutually exclusive observable context: invalid JSON to *output format*, valid multi-image errors to *multi-image input*, and remaining single-image errors to perception, numerical, or domain-rule contexts. Multi-image inputs and perception account for 49.1% and 32.8% of strict errors, followed by numerical (10.1%), output-format (4.6%), and domain-rule (3.1%). These are observed context shares, not causal attributions: multi-image errors may also involve visual recognition or extraction, which are excluded from the perception category by construction.

At the question level, 1,540 of 11,045 scored instances (13.9%) are solved by none of the 39 models, while none is solved exactly by all 39. As shown in Table 4, KIE has the highest zero-solved rate (42.8%), with a median of two exact-correct models and a first quartile of zero, whereas VC has no zero-solved items and a median of 27. Multi-doc Images also show lower coverage than Normal Captures, with 27.3% versus 6.7% zero-solved questions and median coverage of 15 versus 23 models. Manual inspection finds zero-correct instances typically require strict multi-field extraction, large-set classification, or filtered transaction aggregation; intermediate-coverage instances involve document-presence reasoning, transaction matching, or numerical calculation; and all-correct instances are mainly short binary checks over explicit fields. These observations are descriptive, not causal.

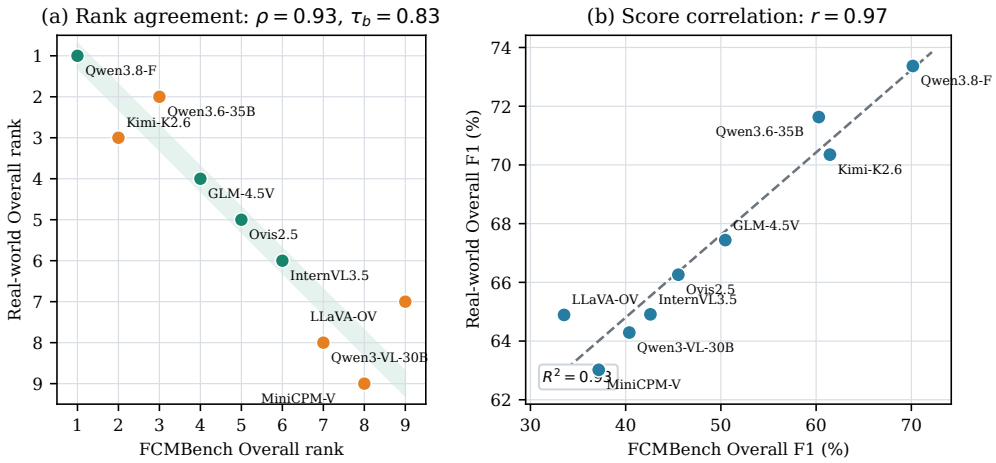


Figure 4: Transfer from FCMBench to a confidential real-world credit-document evaluation. **Left:** rank agreement. Points on the shaded diagonal preserve their FCMBench rank; orange points indicate rank shifts. **Right:** correlation between FCMBench and real-world Overall F1; the dashed line is the least-squares fit.

Category	Zero-solved questions (%)	P25(/39)	P50(/39)	P75(/39)
By task				
IQE	10.5	3	13	28
DTR	4.6	15	23	31
KIE	42.8	0	2	17
CC	1.8	12	19	25
VC	0.0	14	27	35
NC	11.2	5	17	29
RR	0.0	17	20	29
By representative input condition				
Normal Captures	6.7	13	23	32
Multi-doc Images	27.3	0	15	25
Small ROIs	12.3	8	19	30
Out-of-focus	11.7	8	20	32

Table 4: Question-level exact-match coverage across 39 models. Zero-solved denotes the percentage of questions that are solved correctly by zero models. P25, P50, and P75 are the 25th, 50th, and 75th percentiles of the number of exact-correct models per question, out of 39. Conditions overlap and are summarized separately from task totals.

5 Conclusion

We introduce FCMBench, a large-scale privacy-preserving synthetic-to-physical multimodal benchmark for credit-document understanding. Unlike existing benchmarks focusing on isolated OCR or financial question answering, FCMBench evaluates visual perception, cross-document verification, and evidence-based reasoning in real-world credit-review workflows. Comprehensive experiments across 39 vision–language models reveal substantial performance variation and persistent vulnerabilities in fine-grained perception, information extraction, and multi-document integration. Although instantiated in the credit domain, the capabilities FCMBench probes—dense numeric recognition over heterogeneous layouts, arithmetic aggregation and unit conversion, cross-image consistency checking, and constraint-conditioned verification—are broadly relevant to document-heavy applications, making FCMBench a diagnostic testbed for robust multimodal reasoning over complex visual evidence.

AI USE STATEMENT

Generative AI tools were used only for 1. linguistic polishing for the manuscript; 2. synthesizing fictional certificate templates, portraits, logos, and stamps within the benchmark construction pipeline described in Section 3.2. No generative AI tool was used to design the experiments, analyze the results, or write the manuscript text beyond standard language editing; all AI-generated artifacts were manually verified by the authors, who take full responsibility for the content of this work.

ETHICS STATEMENT

FCMBench contains no real personal data: all applicant identities, institutional names, addresses, and financial records are fictional, and logos, emblems, and portraits are replaced with generated alternatives. Certificate replicas were fabricated solely for research capture and are not intended to imitate authentic official documents for fraudulent purposes. The 432 confidential real credit-document images (356 used for the transfer validation in Section 4.2.4 and 76 additional images used only for aggregate realism statistics) were accessed under internal data-governance approval, used only for aggregate statistics, and are neither released nor displayed; where reported in the supplementary material, they are additionally used only for aggregate embedding-space visual-similarity analysis. The 11 capture participants, annotators, and interviewed credit reviewers participated with informed consent and were compensated according to institutional norms. Remaining limitations, including the controlled document families and capture settings, as well as potential misuse risks, are discussed in the supplementary material.

REPRODUCIBILITY STATEMENT

All benchmark construction procedures, answerability audits, evaluation metrics, parser behavior, inference configurations, and complete results for the 39 evaluated models are documented in this paper and the supplementary material. The evaluation suite, prompts, scoring code, synthetic document-generation toolchain, and benchmark data will be released to support reproducibility and future extensions.

REFERENCES

Meta AI. Llama 4 maverick, April 2026. URL https://ai.meta.com/blog/llama-4/.

Anthropic. Claude opus 4.5 and 4.8 model card, June 2026. URL https://www.anthropic.com/claude/opus.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-vl: A family of multimodal vision-language models, 2025.

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.

Google DeepMind. Gemini 3.5 flash model card, May 2026. URL https://deepmind.google/models/model-cards/gemini-3-5-flash/.

Shumin Deng, Heyan Peng, Jing Xu, Rui Mao, Costin Daniel Giurcăneanu, and Jingang Liu. Finmr: A knowledge-intensive multimodal benchmark for advanced financial reasoning. *arXiv preprint arXiv:2510.07852*, 2025.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2025.

Junlong Li, Yeqi Zhu, Decang Cheng, Zhijun Ding, and Cong Jiang. Cfbenchmark-mm: Chinese financial assistant benchmark for multimodal large language model. *arXiv preprint arXiv:2506.13055*, 2025.

Yuliang Liu, Zhang Li, Mingxin Huang, Bin Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: On the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12), 2024.

Zhiyuan Liu, Xinyao Guo, Hongyan Xia, Long Zeng, Feng Lou, Jiaqi Niu, Ming Li, Qi Qi, Jun Li, Wei Zhang, Yu Wang, Wenjie Cai, Wei Shen, and Lei Zhang. Visfineval: A scenario-driven chinese multimodal benchmark for holistic financial understanding. *arXiv preprint arXiv:2508.09641*, 2025.

Shiyin Lu, Yang Li, Yu Xia, Yuwei Hu, Shanshan Zhao, Yanqing Ma, Zhichao Wei, Yinglun Li, Lunhao Duan, Jianshan Zhao, et al. Ovis2.5 technical report. *arXiv preprint arXiv:2508.11737*, 2025.

Jiawei Luo, Ziyang Kou, Liming Yang, Xiao Luo, Junhao Huang, Zhaoyang Xiao, Jie Peng, Chang Liu, Jun Ji, Xiaojian Liu, Shuo Han, Min Zhang, and Yuhua Guo. Finmme: Benchmark dataset for financial multi-modal reasoning evaluation. *arXiv preprint arXiv:2505.24714*, 2025.

Ahmed Masry, Md Rizvee Islam, Maysum Ahmed, Aryaman Bajaj, Fazle Kabir, Aayush Kartha, Md Tahmid Rahman Laskar, Shubhra Rahman, Shaidur Rahman, Mehrad Shahmohammadi, Meet Thakkar, Md Mahadi Hassan Parvez, Enamul Hoque, and Shafiq Joty. Chartqapro: A more diverse and challenging benchmark for chart question answering. *arXiv preprint arXiv:2504.05506*, 2025.

OpenAI. Gpt-5.1 and gpt-5.2 model card, July 2026. URL https://platform.openai.com/docs/models.

Xuan Peng, Lingfei Qian, Yan Wang, Ruoxi Xiang, Yilong He, Yun Ren, Min Jiang, Victor J. Zhang, Yu Guo, Jie Zhao, Hao He, Yan Han, Yi Feng, Yu Jiang, Yan Cao, Haohang Li, Yang Yu, Xueqing Wang, Peng Gao, Shu Lin, Kun Wang, Shuo Yang, Yang Zhao, Zhi Liu, Ping Lu, Jimin Huang, Shuo Wang, Theodoros Papadopoulos, Panagiotis Giannouris, Eleni Soufleri, Nuo Chen, Zhiwei Deng, Haoyu Fu, Yue Zhao, Ming Lin, Mingyue Qiu, Kevin E. Smith, Arman Cohan, Xiao-Yang Liu, Jiaqi Huang, Guojun Xiong, Alejandro Lopez-Lira, Xi Chen, Junichi Tsujii, Jian-Yun Nie, Sophia Ananiadou, and Qianqian Xie. Multifinben: Benchmarking large language models for multilingual and multimodal financial application. *arXiv preprint arXiv:2506.14028*, 2025.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026a. URL https://qwen.ai/blog?id=qwen3.5.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model, April 2026b. URL https://qwen.ai/blog?id=qwen3.6-27b.

Qwen Team. Qwen3.6-35B-A3B: Agentic coding power, now open to all, April 2026c. URL https://qwen.ai/blog?id=qwen3.6-35b-a3b.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026d. URL https://qwen.ai/blog?id=qwen3.8.

Qwen Team. Qwen3.7-Plus, May 2026e. URL https://qwen.ai/blog?id=qwen3.7-plus.

Microsoft Research. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras, 2025.

Li Tang, Gyungin Kim, Xin Zhao, Taylor Lake, Weitong Ding, Fan Yin, Prateek Singhal, Megh Wadhwa, Zhenhua L. Liu, Zane Sprague, Rithwik Namuduri, Bing Hu, Juan D. Rodriguez, Peng Peng, and Greg Durrett. Chartmuseum: Testing visual reasoning capabilities of large vision-language models. *arXiv preprint arXiv:2505.13444*, 2025.

DeepSeek Team. Deepseek-v4-flash-vision-exp, March 2026a. URL https://deepseek.com/blog.

GLM-V Team. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, July 2025. URL https://arxiv.org/abs/2507.01006.

GLM-V Team. Glm-5v-turbo: Toward a native foundation model for multimodal agents, April 2026b. URL https://arxiv.org/abs/2604.26752.

Kimi Team. Kimi k3: Open frontier intelligence, 2026c. URL https://arxiv.org/abs/2607.24653.

LLaVA Team. Llava-onevision technical report, 2026d.

MiniMax Team. Minimax-vl-01, February 2026e. URL https://minimaxi.com.

OpenBMB Team. Minicpm-v 4.5, March 2026f. URL https://github.com/OpenBMB/MiniCPM-V.

An-Lan Wang, Jingqiang Tang, Lingxi Lei, Han Feng, Qixuan Liu, Xiang Fei, Jiayuan Lu, Haoran Wang, Wei Liu, Hongzhi Liu, Heng Liu, Yifan Liu, Xiang Bai, and Chen Huang. Wilddoc: How far are we from achieving comprehensive and robust document understanding in the wild? *arXiv preprint arXiv:2505.11015*, 2025a.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025b.

Lingyu Xie, Yue Wang, Haoyu Guan, Suman Nag, Rahul Goel, Nikhil Swamy, Yang Yang, Chang Xiao, Jenna Prisby, Ross Maciejewski, et al. Idnet: A novel identity document dataset via few-shot and quality-driven synthetic data generation. In *2024 IEEE International Conference on Big Data (BigData)*, pp. 2244–2253. IEEE, 2024.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Bing Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyu Zheng, Zhen Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*, 2024.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Sibo Tong, Yuxuan Sun, Bing Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2025.

Fangyuan Zhu, Zhen Liu, Xin Yi Ng, Hao Wu, Wenjie Wang, Fuli Feng, Chong Wang, Huanbo Luan, and Tat-Seng Chua. Mmdocbench: Benchmarking large vision-language models for fine-grained visual document understanding. *arXiv preprint arXiv:2410.21311*, 2024.

SUPPLEMENTARY MATERIALS

A BENCHMARK CONSTRUCTION DETAILS

A.1 FICTIONAL IDENTITY AND DOCUMENT GENERATION

FCMBench is constructed around fictional applicant identities, each identity is associated with a linked set of identity, financial, and credit-application documents. Within an applicant profile, identifying and financial attributes—including names, identification numbers, dates, addresses, income, and transaction records—are generated jointly so that evidence remains internally consistent across related documents. This profile-level construction supports cross-document questions that would not be well-defined over independently collected document images.

For a given document type, all identities share the same type-level template structure, while the populated content varies by identity. Consequently, documents preserve a controlled layout within each type without repeating the applicant-specific names, numbers, dates, income values, or transaction histories. Reuse is limited to this intended template structure: no rendered document image or physical acquisition instance is duplicated in the benchmark. The collection covers 17 Chinese and 9 English document types, for 26 types in total. As noted in the main paper, the 15 English driver's licenses are derived from the synthetic IDNet dataset Xie et al. (2024); the remaining document materials are generated in-house.

Figure 5 reports the demographic and economic distribution of the 36 fictional identities, which spans gender, language, income source, age, marital status, annual income, and properties owned. The profiles are deliberately varied so that cross-document reasoning tasks (e.g., income rationality and collateral aggregation) are exercised over a heterogeneous applicant population rather than a single archetype.

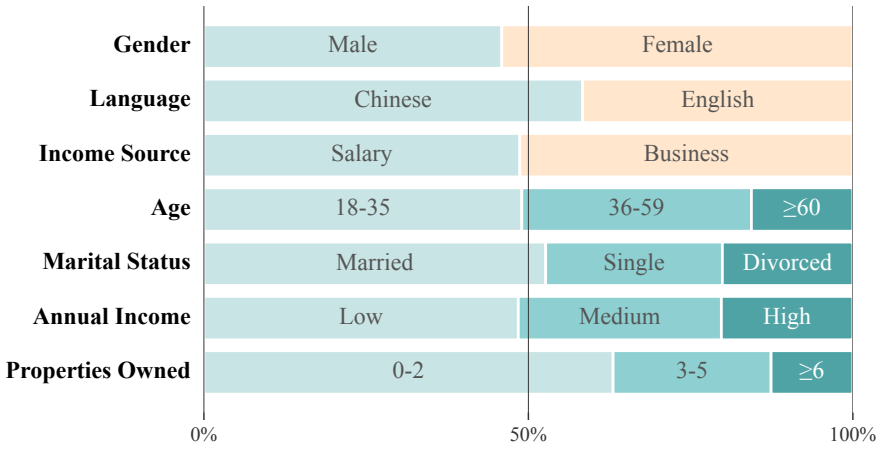


Figure 5: Demographic status distribution of the fictional identities in FCMBench.

A.2 PHYSICAL DOCUMENT FABRICATION AND CAPTURE

The generated document files were fabricated as physical samples before image collection. Firstly, we develop a high-fidelity synthetic certificate generation pipeline to produce credit-related documents and cards that are closely analogous to real-world counterparts. The detailed synthesis workflow is illustrated in Figure 6. Secondly, card-like documents were produced as card replicas, while paper-based certificates and forms were printed on paper at dimensions matching their intended real-world counterparts. Thirdly, the samples were then placed and photographed by 11 collection participants using consumer smartphones from five major brands: iPhone, Huawei, Honor, Xiaomi, and Oppo.

For a Normal Capture, a single sample was placed on an otherwise empty office desk under ambient office lighting without direct illumination. The participant stood above the desk and photographed

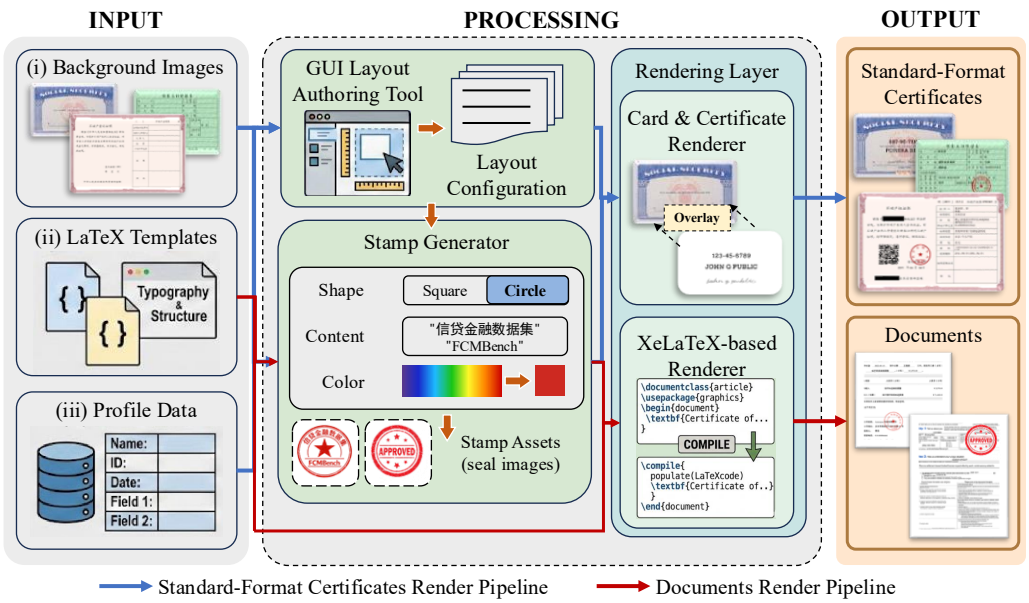


Figure 6: Overview of the synthetic credit-document generation toolchain in FCMBench, which combines a card and certificate renderer and a LaTeX–YAML compiler to generate template-based synthetic credit documents that are printed and re-captured as realistic benchmark images.

the sample frontally, keeping the camera approximately parallel to the document plane. The complete document had to remain inside the frame, without visible glare, and the camera was focused on the document surface. Capture distance was adjusted to include the entire sample rather than fixed to a single metric distance.

The same underlying physical sample was reused across the Normal and degraded conditions, permitting acquisition effects to vary while document content remained fixed. Most conditions were produced through a new capture of the same sample. Two conditions were instead derived from a Normal Capture: Overlaid Watermarks were added electronically, and Cropped Captures were generated by random cropping. Secondary Captures were made by displaying the corresponding generated source document on a computer screen and photographing the display. Table 5 gives the operational definitions and capture requirements. In addition to the nine degradation conditions, Multi-doc Images vary the input composition by placing several documents in one normally captured scene.

Condition	Operational definition	Device or environment requirements
Normal Captures	Place one complete sample on an empty desk and photograph it frontally from above with accurate focus and no visible reflection.	Ambient office lighting without direct illumination; uncluttered desk; handheld smartphone positioned approximately parallel to the sample.
Off-axis Viewpoints	Rotate the physical sample by an arbitrary non-cardinal angle; rotations of 90°, 180°, and 270° are excluded.	Same office setup and smartphone capture as Normal, with document orientation changed while retaining the sample in frame.
Uneven Illumination	Place another object between the ambient light and part of the sample to cast a visible partial shadow.	Office lighting plus an occluding object; the shadow must cover part, rather than all, of the target document.
Specular Reflections	Cover the sample with a transparent plastic sleeve or bag and direct light toward it to create visible glare and local overexposure.	Transparent plastic cover and a directed light source; smartphone capture of the resulting reflection.
Out-of-focus	Intentionally focus the camera on a point outside the plane of the document, leaving the target text and structure visibly blurred.	Smartphone with manual focus selection; focus target must not lie on the sample surface.
Small ROIs	Increase the shooting distance or framing area until the target document occupies less than 20% of the image.	Smartphone capture with the complete target retained but surrounded by a large field of view.
Secondary Captures	Display the generated source document on a computer screen and photograph the display rather than the fabricated sample.	Computer monitor and smartphone; screen artifacts such as moiré may be present.
Cluttered Background	Spread other document samples irregularly across the desk and place the target sample near the center before capture.	Office desk containing multiple distractor samples; target document remains the central subject.
Overlaid Watermarks	Add an electronic watermark to a Normal-Capture image.	Digital post-processing of the Normal Capture; no additional physical capture is required.
Cropped Captures	Randomly crop a Normal-Capture image, retaining 75–90% along each image dimension so that part of the document is excluded.	Digital random cropping of the Normal Capture; no additional physical capture is required.
Multi-doc Images	Place multiple document samples together on an otherwise empty desk and capture them in a single image using the Normal-Capture procedure.	Ambient office lighting without direct illumination; uncluttered desk beyond the selected documents; all target samples remain visible and in focus.

Table 5: Operational definitions for Normal Captures, the nine acquisition-degradation conditions, and the Multi-doc Images composition condition.

#	Institution and Position	Years
1	Joint-stock commercial bank, retail credit reviewer (branch)	4
2	City commercial bank, retail credit reviewer (branch)	6
3	Joint-stock commercial bank, review officer (approval center)	10
4	City commercial bank, credit QA specialist (head office)	12
5	Joint-stock commercial bank, credit policy product manager	13
6	State-owned large bank, credit management supervisor (provincial)	15
7	City commercial bank, senior approval expert / internal trainer	17
8	Consumer-finance company, online screening reviewer	3
9	Internet lending platform, credit reviewer	3
10	Microloan company, credit reviewer	5
11	Consumer-finance company, fraud investigation specialist	6
12	Consumer-finance company, fraud risk manager	8
13	Consumer-finance company, credit strategy / rule-engine analyst	10
14	Consumer-finance company, QA and training specialist	10
15	Local financial regulatory bureau, credit compliance supervisor	12
16	Banking association, credit practice research analyst	15
17	Credit-reporting agency, credit risk expert	12
18	Industry research institute, consumer-protection compliance expert	12
19	Rural commercial bank, credit reviewer (county branch)	4
20	Auto-finance company, review and anti-fraud specialist	7

Table 6: Composition of the interviewed expert panel. Years denote experience in the credit domain.

A.3 Task and Question Construction

FCMBench instantiates seven task families from recurring operations in credit document review. Each question template fixes the required evidence, decision rule, and JSON response schema; document types, requested fields, reference dates, candidate labels, and numerical values are then populated from the linked applicant profiles. Table 8 counts a Chinese and an English instantiation as separate templates when they have different `subtask` identifiers in the retained benchmark. Input-image counts include both separate images and composite photographs containing more than one document.

B Expert-interview protocol and task mapping.

B.1 Interviewee Composition

We interviewed 20 credit-domain experts: 7 from commercial banks, 7 from consumer-finance and lending companies, 4 from regulatory and industry bodies, and 2 supplementary roles covering county-level banks and auto finance. Their institutions, positions, and years of credit experience are listed in Table 6.

B.2 Interview Outline

The protocol consists of two rounds of semi-structured interviews. Round 1 involves all 20 experts and elicits real-world workflows and problems; Round 2 involves the 8 experts with more than ten years of experience, who screen and confirm the candidate task types. Table 7 summarizes the question categories.

B.3 From Interviews to Benchmark Tasks

The task design follows a four-step pipeline grounded in the interviews:

1. **Round-1 interviews.** All 20 experts describe their daily review workflows, the problems encountered during document review, and the cautions applied to applicant materials, following the Round-1 question categories.

Round / Category	Representative Questions
Round 1: workflow and problem elicitation (all 20 experts)	
Daily workflow	Describe the complete process from receiving a loan application to a review decision; which checks are performed and in what order; how the process differs across products (mortgage, auto, consumer, business loans); which steps are most time-consuming and error-prone.
Problems in review	What image-quality problems occur in practice (blur, glare, occlusion, screen recapture, multiple documents in one photo) and how often; which problems trigger outright rejection and which allow remediation; typical forgery, alteration, or cross-document inconsistency cases.
Document cautions	Required and substitutable documents per loan product; which fields are cross-checked and under which rules (validity, format, amount reasonableness, cross-document consistency); which fields must match exactly and what deviation is tolerated for amounts; compliance boundaries on using and retaining applicant information.
Round 2: task screening and confirmation (8 experts, >10 years)	
Task importance ranking	Given the candidate task list derived from Round 1 (with samples), rate each task by business importance $\times$ frequency (1–5) and rank them; identify missing or redundant tasks.
Task boundary confirmation	For top-ranked tasks: what are the decision criteria and what outputs count as correct; which rules are regulatory requirements versus internal policies; which cases even senior reviewers disagree on.
Sample confirmation	Given the converged task list and 3–5 sample questions: whether the wording matches operational language, whether reference answers are unique and reasonable, and confirmation that the tasks originate from daily work and are widely regarded as important.

Table 7: Interview outline. Round 1 elicits workflows, problems, and document cautions; Round 2 screens and confirms candidate tasks.

2. **Task-type and sample formation.** The interview content is coded and clustered; frequent and business-critical workflow stages are distilled into candidate task types, each accompanied by sample questions.
3. **Screening by senior experts.** The 8 experts with more than ten years of experience rate, rank, and confirm the candidate tasks and their boundaries and samples in Round 2.
4. **Union and finalization.** The tasks endorsed by the senior experts are merged by union, which fixes the final task categories, the perturbation types (image-quality and input-composition challenges), and the instruction formats (prompt wording and output-format requirements) of FCMBench.

Task	Objective	Input images	Templates	Typical question and canonical JSON	Required knowledge or operations
IQE	Identify the dominant capture-quality condition.	1	2	"What is the primary quality issue?" {"answer":"Blurry"}	Distinguish normal capture, blur, shadow, reflection, rotation, small target scale, screen recapture, clutter, watermark, and cropping.
DTR	Recognize present document types and, in some templates, required but missing types.	1–5	6	"Which fully visible document types are present?" {"answer":["loan application"]}	Match layouts and visual content to the document taxonomy; aggregate and deduplicate labels across images; compare against a required list.
KIE	Extract requested fields or filtered transaction records.	1–5	9	"Extract transactions on the specified dates." {"answer": [["merchant","out","85.30"]]}	Read text, tables, and key–value layouts; localize requested fields; filter rows by dates or attributes; preserve exact strings and output structure.
CC	Check whether documents refer to the same applicant or whether entries have matching supporting evidence.	1, 2, or 4	6	"Which submitted document belongs to another applicant?" {"answer":["loan application"]}	Align names, identifiers, addresses, and transaction/order references across documents and report mismatches.
VC	Verify explicit temporal, format, or value constraints.	1, 3, or 4	6	"Which documents fall outside the 180-day validity window?" {"answer":["loan clearance"]}	Locate issue dates or identifiers, compare dates with the supplied reference date, and apply digit-count, range, or formatting rules.
NC	Compute a requested quantity from one or more documents.	1–2	8	"What is the total value of all residential properties?" {"answer":692707}	Extract numerical operands, select relevant records, apply sums, differences, counts, unit conversion, and prompt-specified rounding.
RR	Judge whether financial evidence is mutually plausible under a stated policy.	2	3	"Is the tax amount reasonable for the reported income?" {"answer":"unreasonable"}	Combine evidence across documents, derive ratios or ranges, and compare the result with explicit income, tax, age, employment, or cash-flow thresholds.

Table 8: Construction of the seven FCMBench task families. Template counts are the numbers of distinct language-specific `subtask` identifiers in the retained 11,045-instance benchmark.

B.4 EXAMPLES FOR THE SEVEN TASKS

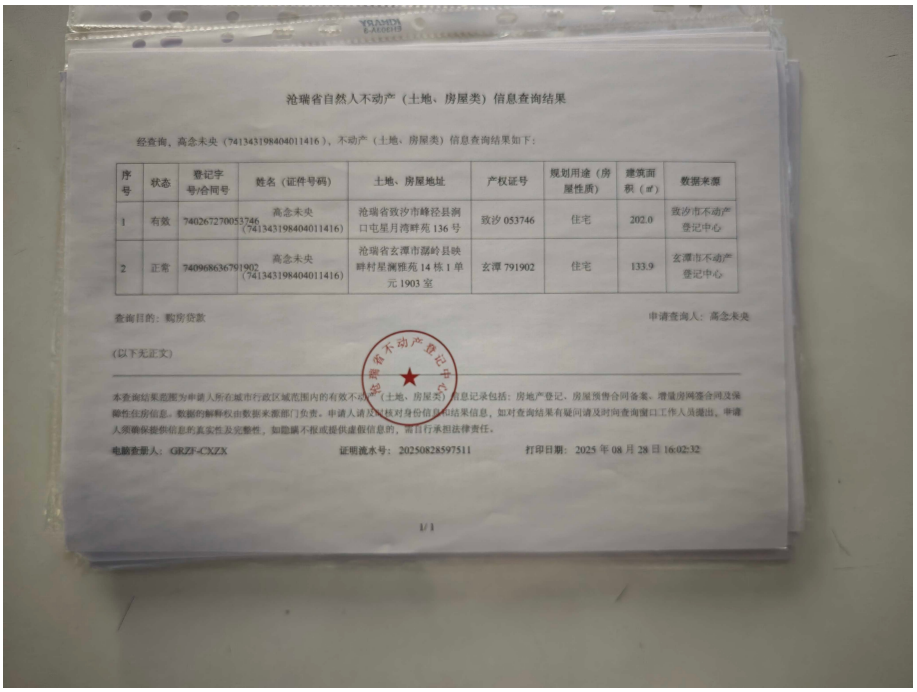


Figure 7: Example for **Image Quality Evaluation (IQE)**.

Prompt. Analyze the provided image and identify its single most prominent image-quality category. Candidate categories are: incomplete capture, reflection, background clutter, watermark, off-axis capture, screen recapture, normal capture, shadow, blur, and small region of interest. Select only the single most prominent and obvious category. The output must be JSON in the form {`"answer": "Category Name"`}. Output JSON only, without explanation or other content, and the value of `answer` must be one of the categories listed above.

Ground truth. {`"answer": "blur"`}.

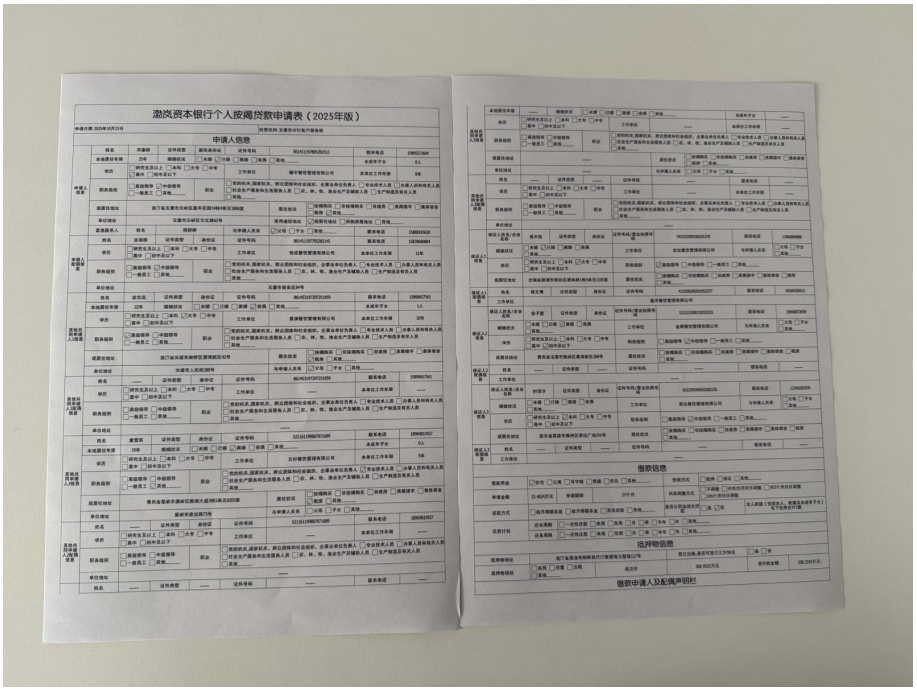
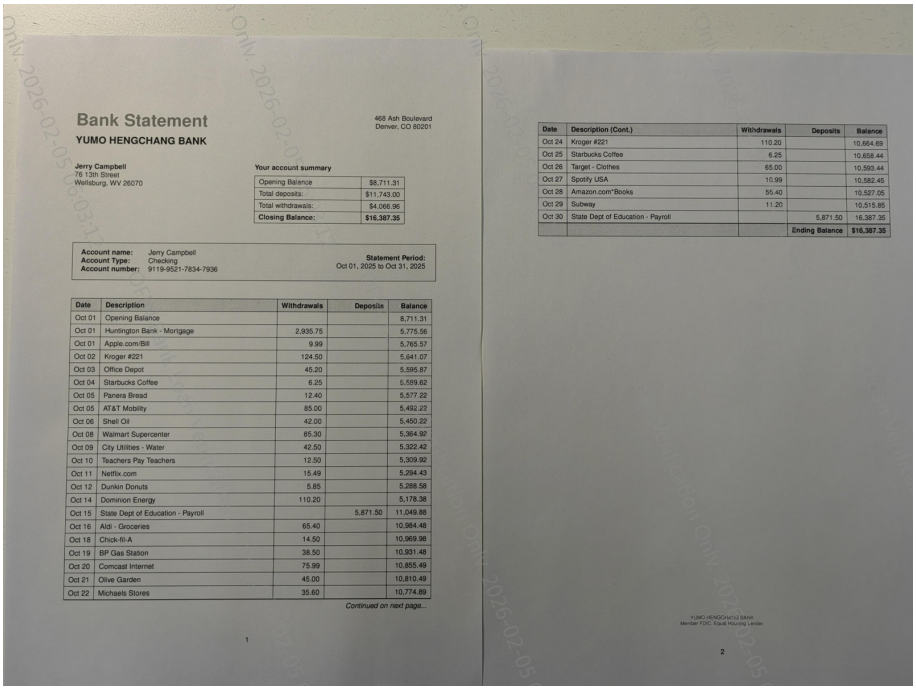


Figure 8: Example for **Document Type Recognition (DTR)**.

Prompt. Analyze the provided image, identify the document types it contains, and retain only documents shown completely within the image. A single image may contain multiple document types; do not omit any. Select one or more types from: loan application, transfer voucher, tobacco retail license, household register (personal page), divorce certificate, ID card (portrait side), loan investigation report, real-estate information query result, bank card, real-estate ownership certificate, electronic marriage certificate, bank statement, food business license, income certificate, ID card (national-emblem side), corporate business license, loan-clearance certificate, individual income-tax payment certificate, social-insurance certificate, fund-flow analysis report, road-transport business license, and household register (household page). If one type is present, output, for example, {`"answer":["ID card (national-emblem side)"]`}. If multiple types are present, output, for example, {`"answer":["ID card (national-emblem side)", "household register (personal page)"]`}. If none of the listed types is present, output {`"answer":[""]`}. Output JSON only, without any other text or explanation.

Ground truth. {`"answer":["loan application"]`}.



Bank Statement

YUMO HENGCHANG BANK

468 Ash Boulevard
Denver, CO 80201

Jerry Campbell
76 13th Street
Wellsburg, WY 26070

Your account summary

Opening Balance	$8,711.31
Total deposits:	$11,743.00
Total withdrawals:	$4,066.96
Closing Balance:	$16,387.35

Account name: Jerry Campbell
Account Type: Checking
Account number: 9119-9521-7834-7936

Statement Period:
Oct 01, 2025 to Oct 31, 2025

Date	Description	Withdrawals	Deposits	Balance
Oct 01	Opening Balance			8,711.31
Oct 01	Huntington Bank - Mortgage	2,935.75		5,775.56
Oct 01	Apple.com/Bill	9.99		5,765.57
Oct 02	Kroger #221	124.50		5,641.07
Oct 03	Office Depot	45.20		5,595.87
Oct 04	Starbucks Coffee	6.25		5,589.62
Oct 05	Panera Bread	12.40		5,577.22
Oct 05	AT&T Mobility	85.00		5,492.22
Oct 06	Shell Oil	42.00		5,450.22
Oct 08	Walmart Supercenter	85.30		5,364.92
Oct 09	City Utilities - Water	42.50		5,322.42
Oct 10	Teachers Pay Teachers	12.50		5,309.92
Oct 11	Netflix.com	15.49		5,294.43
Oct 12	Dunkin Donuts	5.85		5,288.58
Oct 14	Dominion Energy	110.20		5,178.38
Oct 15	State Dept of Education - Payroll		5,871.50	11,049.88
Oct 16	Aldi - Groceries	65.40		10,984.48
Oct 18	Chick-fil-A	14.50		10,969.98
Oct 19	BP Gas Station	38.50		10,931.48
Oct 20	Comcast Internet	75.99		10,855.49
Oct 21	Olive Garden	45.00		10,810.49
Oct 22	Michaels Stores	35.60		10,774.89

Continued on next page...

1

Date	Description (Cont.)	Withdrawals	Deposits	Balance
Oct 24	Kroger #221	110.20		10,664.69
Oct 25	Starbucks Coffee	6.25		10,658.44
Oct 26	Target - Clothes	65.00		10,593.44
Oct 27	Spotify USA	10.99		10,582.45
Oct 28	Amazon.com*Books	55.40		10,527.05
Oct 29	Subway	11.20		10,515.85
Oct 30	State Dept of Education - Payroll		5,871.50	16,387.35
			Ending Balance	$16,387.35

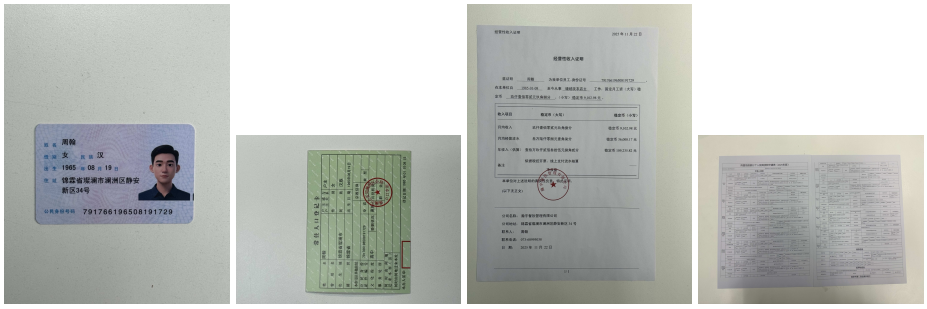
YUMO HENGCHANG BANK
Member FDIC. Equal Housing Lender.

2

Figure 9: Example for **Key Information Extraction (KIE)**.

Prompt. Please carefully analyze the bank-statement image above and extract the transaction records that occurred on the 8th, 10th, and 27th of the current month. For each transaction, extract the triplet `[description, in/out, amount]` and output the following JSON structure: {`"answer":[[xx,xx,xx],[xx,xx,xx],...]`}. Do not output anything else.

Ground truth. {`"answer":[["Walmart Supercenter","out","85.30"], ["Teachers Pay Teachers","out","12.50"], ["Spotify USA","out","10.99"]]`}.



(a) Image 1 (b) Image 2 (c) Image 3 (d) Image 4

Figure 10: Example for **Consistency Checking (CC)**.

Prompt. Image 1 is known to be the target user's ID card (portrait side). The user has also submitted a household register (personal page), an income certificate, and a loan application, among which another person's materials may have been mixed. Determine which document types in Images 2–4 do not belong to the target user; output an empty list if all of them belong to the target user. Output JSON only, without any other text or explanation. Use the format {`"answer":["type 1","type 2"]`}.

Ground truth. {`"answer":["loan application"]`}.



(a) Image 1 (b) Image 2 (c) Image 3 (d) Image 4

Figure 11: Example for **Validity Checking (VC)**.

Prompt. The user submitted a loan-clearance certificate, a social-insurance participation certificate, an individual income-tax payment certificate, and an income certificate. The bank requires all of these documents to have been issued within the most recent 180 days. A document issued more than 180 days ago or after today's date is invalid. Today's date is 2025-09-02. Return the invalid document types. Output JSON only, without any other text or explanation. Use the format {`"answer":["document type 1","document type 2"]`}.

Ground truth. {`"answer":["loan clearance certificate"]`}.

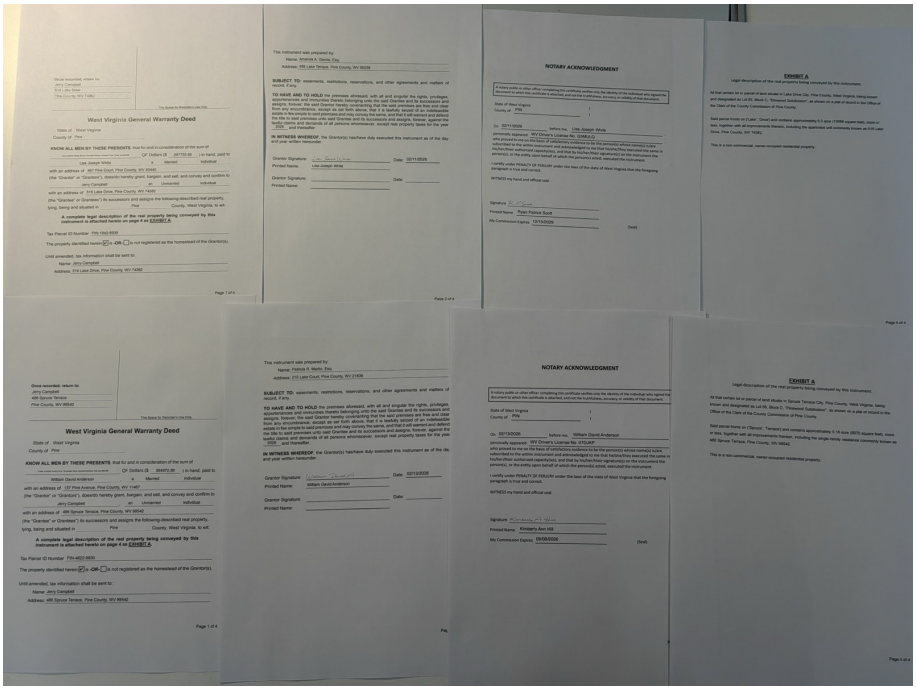
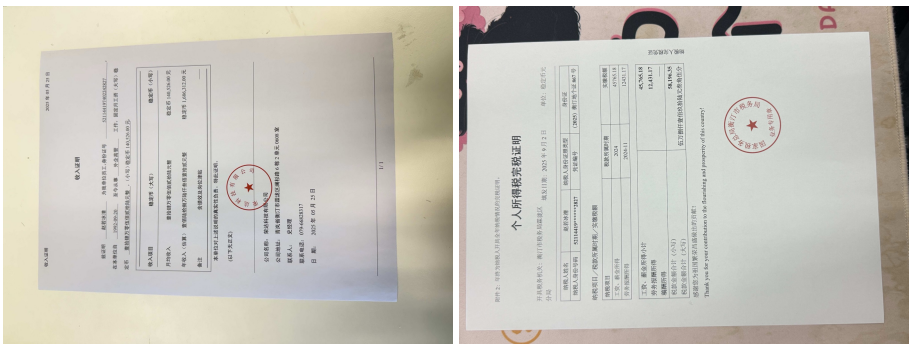


Figure 12: Example for **Numerical Calculation (NC)**.

Prompt. What is the total value of all the residential properties of this client? Round the value to the nearest integer. Output only JSON, with no extra text or explanations. The output format must strictly follow {"answer":12677}.

Ground truth. {"answer":692707}.



(a) Image 1 (b) Image 2

Figure 13: Example for **Rationality Review (RR)**.

Prompt. The borrower submitted an income certificate and an individual income-tax payment certificate. Determine whether the borrower's income amount matches the tax amount: when annual income is below 100,000, the tax share should be below 20%; when annual income is above 100,000, the tax share should be above 10%. Answer with "reasonable" or "unreasonable." Output JSON only, without any other text or explanation. Use the format {`"answer":"reasonable"`}.

Ground truth. {`"answer":"unreasonable"`}.

C Evaluation Protocol

Set-Based Instance Scoring. FCMBench uses a common set-based scorer for categorical labels, extracted fields, lists, and structured records. After parsing and normalization, a prediction and its ground truth are represented as order-invariant sets S_p and S_g of key–value items. Their precision, recall, and F1 are

$$P = \frac{|S_p \cap S_g|}{|S_p|}, \qquad R = \frac{|S_p \cap S_g|}{|S_g|}, \qquad F1 = \frac{2PR}{P+R}.$$

When S_g is nonempty, an empty prediction receives F1 $= 0$. When S_g is empty, the score is one only if S_p is also empty; otherwise it is zero. Simple single-valued outputs are a special case and therefore receive either zero or one under exact matching.

Output Parser. The parser first strips surrounding whitespace and extracts the substring from the first opening brace to the last closing brace. It then applies a standard JSON decoder. Text with no such substring or with invalid JSON is treated as an invalid prediction and receives a score of zero. If the parsed object contains a top-level `answer` field, that field is unwrapped before scoring.

The resulting object is recursively flattened. For a dictionary, each nested key is appended to its parent using a period, so a value under {`"person":{"name":...}`} receives the hierarchical prefix `person.name`. Flat lists are normalized element by element under the same prefix and are evaluated without regard to order. For table-like list-of-lists, each inner list is converted to a tuple of normalized values and retained as one row item; this preserves within-row associations while making the collection of rows order-invariant.

Primitive values undergo deterministic normalization. `None` becomes the empty string. Leading and trailing whitespace and superficial angle brackets are removed. Scalar strings containing Chinese or English commas are split into separate values. For numeric formatting, thousands separators are removed, numerically equivalent forms are canonicalized, and an integral floating-point value is written as an integer string; for example, `12.0` and `12` normalize to the same value. These transformations remove formatting variation but do not introduce semantic or approximate matching for textual fields.

Invalid JSON and an absent JSON object are handled as empty predictions. Therefore, they score zero against a nonempty answer. A valid empty prediction also scores zero against a nonempty answer, while an empty prediction and empty ground truth receive one. Conversely, a nonempty prediction against an empty ground truth receives zero.

Numerical Calculation Matching. For an eligible Numerical Calculation (NC) instance, the predicted value is first converted to the unit requested by the question. A prediction is counted as a match when

$$|\hat{y} - y| < 2,$$

where $\hat{y}$ and y are the normalized predicted and ground-truth values in that unit. Most NC questions involve addition or subtraction of values read from the documents. The prompts do not prescribe whether values should be rounded before the arithmetic operation or only after the final result is computed. Models may therefore obtain slightly different answers despite using the correct evidence and operation. The tolerance of two accommodates both rounding conventions. It applies only to derived numerical answers in NC; identity numbers, account numbers, dates, categorical labels, and textual fields continue to use strict matching. The Normal-Capture denominator is nonzero in the evaluated data, so no additional zero-denominator rule is required.

Score Aggregation. Scoring proceeds through the following hierarchy:

field/key–value matches $\rightarrow$ instance F1 $\rightarrow$ fine-grained subtask $\rightarrow$ task $\rightarrow$ capability and Overall scores.

First, normalized field or key–value matches determine the F1 of each instance. Instance F1 values are averaged within each fine-grained subtask. The subtask means are then macro-averaged to obtain the seven task scores: IQE, DTR, KIE, CC, VC, NC, and RR. Perception is the unweighted mean of IQE, DTR, and KIE; Reasoning is the unweighted mean of CC, VC, NC, and RR; and Overall is the unweighted mean of all seven task scores.

Macro-averaging prevents common question templates or subtasks with many instances from dominating the result. Each fine-grained review operation contributes equally to its parent task, and each of the seven task families contributes equally to Overall.

Robustness ratios compare each challenge condition with Normal Captures while holding subtask composition fixed. For condition c, let $\mathcal{S}_c$ be the intersection of fine-grained subtasks represented under both c and Normal Captures. For model m, the matched-subtask ratio is

$$\rho_{m,c} = \frac{\frac{1}{|\mathcal{S}_c|}\sum_{s\in\mathcal{S}_c} F1_{m,c,s}}{\frac{1}{|\mathcal{S}_c|}\sum_{s\in\mathcal{S}_c} F1_{m,\mathrm{Normal},s}}.$$

Thus, the numerator and denominator use exactly the same subtasks. The reported condition-level robustness summary is the macro-average of $\rho_{m,c}$ across evaluated models. Scores are computed on the common instance-ID intersection used for the paper results. Because the matched Normal-Capture denominator is nonzero, the ratio needs no special fallback rule.

D Models and Implementation Details

D.1 Evaluated Models and Inference Configuration

Table 9 lists all 39 models included in the reported benchmark comparison together with their deployment and reasoning configuration. Release date denotes the month of initial public model or API availability. For mixture-of-experts models, the notation `total/Aactive` reports total and activated parameters, respectively. A dash indicates that the provider has not disclosed the parameter count.

Models available through hosted services are evaluated via their official APIs; the remaining open-source models are self-deployed on cloud GPUs. We use deterministic or near-deterministic decoding when supported, retain provider defaults for unspecified parameters, and resize images to each model's limits. For reasoning-capable API models, we use the lowest supported reasoning setting when configurable. Scores use the same parser and hierarchical macro-averaging protocol.

Table 9: The 39 evaluated models and their deployment and reasoning configuration. Parameter counts are based on public model cards and release materials. "Self" denotes self-deployed inference.

Model	Organization	Parameters	Release date	Deployment	Reasoning effort
Commercial models					
Gemini 3.5 Flash	Google DeepMind	–	2026-05	API	Minimum
Gemini 3 Pro Preview	Google DeepMind	–	2025-11	API	Minimum
Claude Opus 4.8	Anthropic	–	2026-05	API	Minimum
Gemini 3 Flash Preview	Google DeepMind	–	2025-12	API	Minimum
Claude Opus 4.5	Anthropic	–	2025-11	API	Minimum
Qwen3.7-Plus	Alibaba Cloud	–	2026-05	API	Minimum
Qwen3.8-Max	Alibaba Cloud	–	2026-08	API	Minimum
GLM-5V-Turbo	Z.ai	–	2026-04	API	Minimum
GPT-5.2	OpenAI	–	2025-12	API	None
GPT-5.1	OpenAI	–	2025-11	API	None
Open-weight models					
Kimi K3	Moonshot AI	2.8T/A104B	2026-07	API	Minimum
Kimi-K2.5	Moonshot AI	1T/A32B	2026-02	API	Minimum
Kimi-K2.6	Moonshot AI	1T/A32B	2026-04	API	Minimum
Qwen3.8-Flash	Alibaba Cloud	125B/A6B	2026-08	API	Minimum
Qwen3.6-35B-A3B	Alibaba Cloud	35B/A3B	2026-04	Self	Instruct model
Qwen3.6-27B	Alibaba Cloud	27B	2026-04	Self	Instruct model
Qwen3.5-27B	Alibaba Cloud	27B	2026-02	Self	Instruct model
Qwen3.5-397B-A17B	Alibaba Cloud	397B/A17B	2026-02	API	Thinking disabled
Qwen3-VL-235B-A22B	Alibaba Cloud	235B/A22B	2025-10	Self	Instruct model
Qwen3-VL-32B	Alibaba Cloud	32B	2025-10	Self	Instruct model
Qwen3.5-35B-A3B	Alibaba Cloud	35B/A3B	2026-02	Self	Instruct model
Qwen3.5-122B-A10B	Alibaba Cloud	122B/A10B	2026-02	Self	Instruct model
GLM-4.5V	Z.ai	106B/A12B	2025-07	Self	Instruct model
GLM-4.6V	Z.ai	106B/A12B	2025-12	Self	Instruct model
Ovis2.5-9B	AIDC-AI	9B	2025-08	Self	Instruct model
Qwen3-VL-8B	Alibaba Cloud	8B	2025-10	Self	Instruct model
DeepSeek-V4-Flash-Vision-Exp	DeepSeek	305B/A13B	2026-08	API	Minimum
InternVL3.5-241B-A28B	Shanghai AI Lab	241B/A28B	2025-08	Self	Instruct model
InternVL3.5-38B	Shanghai AI Lab	38B	2025-08	Self	Instruct model
Qwen3-VL-30B-A3B	Alibaba Cloud	30B/A3B	2025-10	Self	Instruct model
InternVL3.5-30B-A3B	Shanghai AI Lab	30B/A3B	2025-08	Self	Instruct model

Continued on next page

Table 9: The 39 evaluated models and their deployment and reasoning configuration (continued).

Model	Organization	Parameters	Release date	Deployment	Reasoning effort
Llama 4 Maverick	Meta AI	400B/A17B	2025-04	API	Instruct model
MiniMax-VL-01	MiniMax AI	456B/A46B	2025-01	API	Lowest supported
InternVL3.5-8B	Shanghai AI Lab	8B	2025-08	Self	Instruct model
MiniCPM-V 4.5	ModelBest/OpenBMB	8B	2025-08	Self	Instruct model
Kimi-VL	Moonshot AI	16B/A3B	2025-06	Self	Instruct model
LLaVA-OneVision-1.5-8B	LMMS-Lab	8B	2025-09	Self	Instruct model
Janus-Pro-7B	DeepSeek AI	7B	2025-01	Self	Instruct model
Phi-4-multimodal	Microsoft	6B	2025-02	Self	Instruct model

Inference configuration. API requests follow `batch_process.py`: temperature is 0.01, no explicit output-token limit is supplied, and reasoning is set to the lowest mode supported by the provider. The corresponding per-model reasoning configurations are included in Table 9.

Self-hosted open-weight models below 400B were deployed on Alibaba Cloud PAI with NVIDIA H20 GPUs (96 GB HBM3) through ms-swift. Their shared decoding configuration was `topK=1`, `topP=0.001`, `temperature=0.01`, and maximum model length 31,744. Images were resized to approximately six million pixels. Instruction variants were used where available. Open-weight models served through hosted APIs (Table 9) were evaluated under the same near-deterministic settings as commercial models.

Model	IQE	DTR	KIE	CC	VC	NC	RR	Overall
Commercial Models								
Qwen3.8-Max	**61.16**	82.37	45.42	74.93	93.83	62.68	78.32	71.24
Gemini 3.5 Flash	52.92	91.37	49.23	74.15	90.60	62.19	75.43	70.84
Gemini 3 Pro Preview	43.78	90.33	45.23	61.98	91.43	63.71	72.70	67.02
Claude Opus 4.8	58.19	74.61	51.95	63.06	86.89	61.45	72.87	67.00
Gemini 3 Flash Preview	51.67	**95.32**	45.86	76.18	76.09	60.76	61.89	66.82
Claude Opus 4.5	43.97	83.78	42.64	66.78	75.64	45.05	65.78	60.52
Qwen3.7-Plus	43.86	81.50	37.69	64.03	62.11	49.65	54.16	56.14
GLM-5V-Turbo	38.16	68.27	39.49	59.28	59.97	39.83	48.68	50.53
GPT-5.2	46.13	57.91	37.92	66.63	63.24	29.65	46.22	49.67
GPT-5.1	38.17	72.08	33.25	42.05	59.92	20.43	44.03	44.28
Open-Source Models								
Kimi K3	45.94	78.39	**52.49**	**77.75**	93.00	**70.28**	**83.15**	**71.57**
Qwen3.8-Flash	54.07	76.15	47.02	72.52	**94.01**	64.40	82.95	70.16
Kimi-K2.5	44.70	79.12	46.94	68.68	70.37	59.80	72.91	63.22
Kimi-K2.6	43.15	74.96	47.15	73.43	66.79	57.55	67.12	61.45
Qwen3.6-35B-A3B	45.73	77.81	44.13	72.29	62.49	63.51	56.00	60.28
Qwen3.6-27B	45.46	78.21	43.89	73.48	62.89	63.69	54.28	60.27
Qwen3.5-27B	47.77	69.11	46.58	68.18	70.05	60.37	51.75	59.12
Qwen3.5-397B-A17B	37.81	74.71	40.05	56.33	72.63	52.98	48.92	54.77
Qwen3-VL-235B-A22B-Instruct	45.17	67.10	43.38	67.21	59.28	41.64	49.01	53.25
Qwen3-VL-32B-Instruct	46.10	54.76	44.46	67.14	62.44	38.03	53.23	52.31
Qwen3.5-35B-A3B	42.42	63.34	46.03	50.99	53.92	58.50	50.47	52.24
Qwen3.5-122B-A10B	44.33	59.49	47.65	42.31	58.24	58.18	52.23	51.78
GLM-4.5V	40.48	58.36	40.97	64.03	61.92	41.01	46.47	50.46
GLM-4.6V	37.73	63.28	41.34	67.44	61.21	32.74	41.76	49.36
Ovis2.5-9B	34.26	51.78	31.88	51.15	56.79	32.94	59.88	45.53
Qwen3-VL-8B-Instruct	39.12	48.71	34.28	58.15	51.26	30.43	51.89	44.83
DeepSeek-V4-Flash-Vision-Exp	43.59	61.89	29.10	43.75	61.94	24.75	46.56	44.51
InternVL3.5-241B-A28B	31.13	60.62	34.51	45.62	54.36	26.11	45.87	42.60
InternVL3.5-38B	36.34	63.63	29.97	36.44	54.02	21.45	46.79	41.24
Qwen3-VL-30B-A3B-Instruct	32.70	42.29	31.52	43.37	56.28	26.71	49.81	40.38
InternVL3.5-30B-A3B	36.64	49.55	27.06	40.97	55.55	21.57	50.22	40.22
Llama 4 Maverick	41.42	37.39	30.45	36.39	60.21	28.51	42.13	39.50
MiniMax-VL-01	43.60	35.12	27.75	34.17	56.72	25.95	45.43	38.39
InternVL3.5-8B	27.66	49.40	26.09	40.74	56.59	19.55	45.87	37.98
MiniCPM-V 4.5	34.19	51.41	23.24	37.24	48.39	19.85	45.87	37.17
Kimi-VL	22.16	47.12	25.73	36.65	51.51	15.36	45.87	34.91
LLaVA-OneVision-1.5-8B	20.37	25.64	31.45	28.29	59.90	23.19	45.87	33.53
Janus-Pro-7B	10.76	10.97	9.94	58.50	40.71	14.55	45.87	27.33
Phi-4-multimodal-instruct	14.71	23.60	21.15	28.28	36.79	0.51	47.18	24.60

Table 10: Task-level F1 (%) for all tested commercial and open-source models.

E Complete Experimental Results

This appendix reports the complete per-task, per-condition, and per-model results underlying the aggregates in the main paper, together with uncertainty quantification and auxiliary analyses.

E.1 Complete Task-Level Results

Figure 14 reports all seven task scores and their unweighted macro-average (Overall) for each of the 39 evaluated models. The main-paper table reports the same seven task-level scores and Overall for the top-5 models in each group; this heatmap extends the identical decomposition to all 39 evaluated models under the same common-instance evaluation protocol.

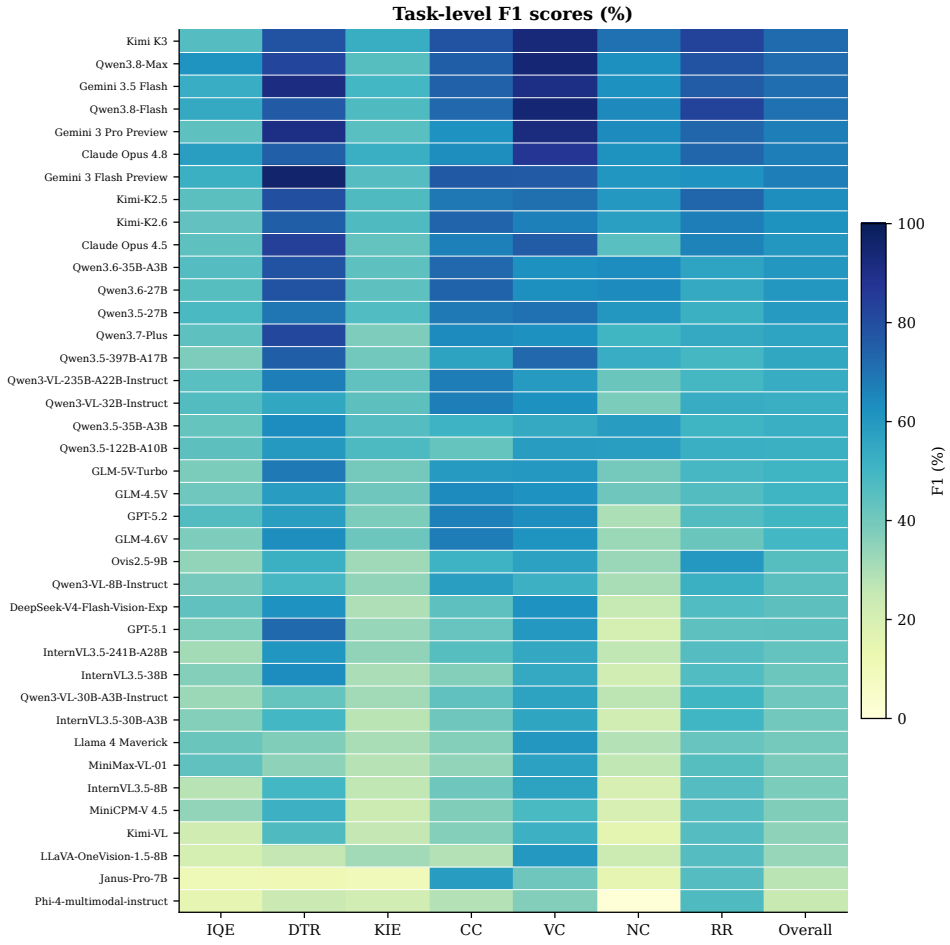


Figure 14: Task-level F1 (%) and Overall macro-average for every evaluated model.

E.2 Complete Robustness-Level Results

Figure 15 reports Relative F1 for all 39 models and robustness conditions. Relative F1 divides the score under each condition by the Normal-Capture score recomputed over exactly the fine-grained subtask categories shared with that condition.

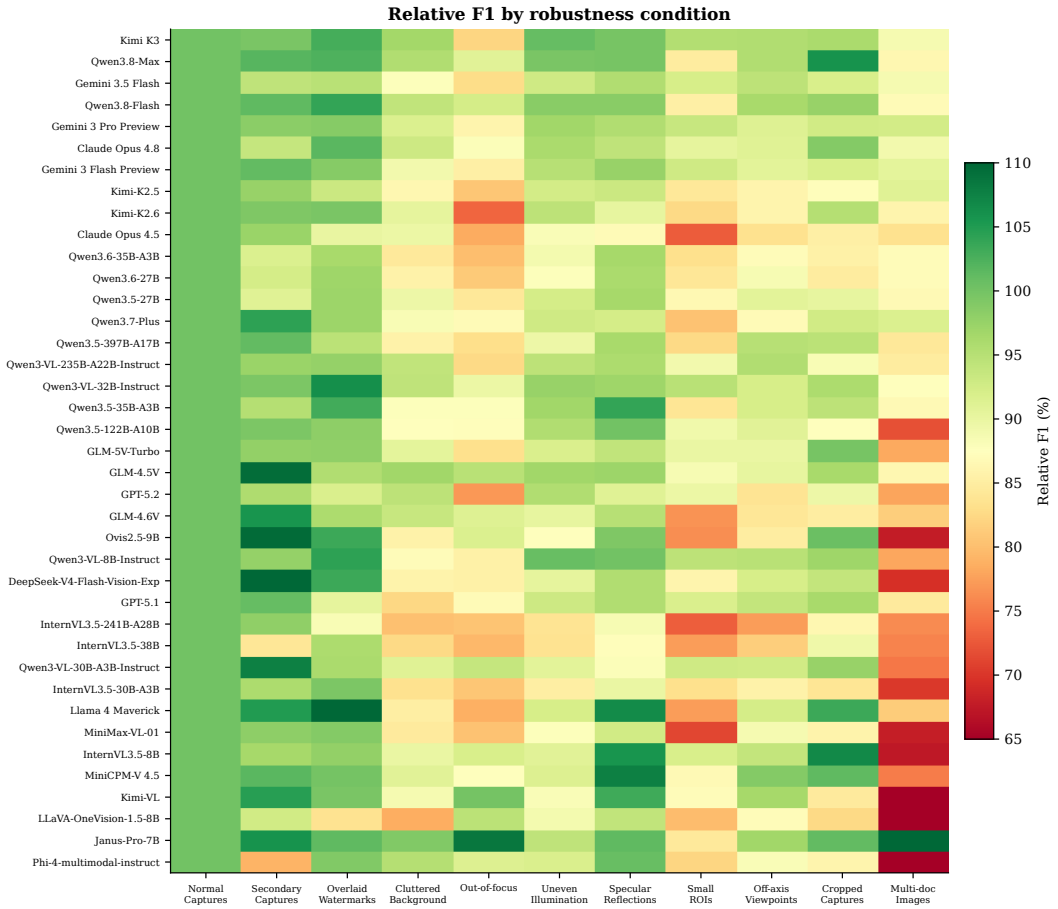


Figure 15: Relative F1 (%) for every model and robustness condition. Values above 100 indicate that the condition score exceeds its matched Normal-Capture baseline; this comparison matches subtask categories rather than individual questions.

Why do some Relative F1 values exceed 100%? Excluding Normal Captures, which equal 100% by definition, 39 of 390 model–condition values exceed 100%, spanning 20 models. They concentrate in Secondary Captures (13), Specular Reflections (9), Overlaid Watermarks (8), and Cropped Captures (5); every other condition has at most two. These values are not clipped: they mean that the condition score exceeds its matched Normal-Capture score.

The largest contributor is Image Quality Evaluation (IQE), for which the condition itself is the target. Explicit screen, watermark, or glare cues can be easier to recognize than the absence of artifacts in a Normal Capture. Removing IQE reduces 21 of the 39 values to at most 100%. For example, Kimi K3's 103.2% on Secondary Captures and 104.5% on Overlaid Watermarks become 94.2% and 98.2%, respectively, without IQE.

The remaining cases also reflect matching by subtask category rather than by identical question; the two sets can differ in identities, prompts, and sample counts. Low baselines further amplify small changes. Janus-Pro-7B, for example, reaches 110.2% on Multi-doc Images because 4.83 F1 is divided by 4.39 matched Normal F1, a difference below 0.5 points. Above-100 values therefore

show a higher score on the matched condition set, not that degradation intrinsically improves model capability.

E.3 Identity-Clustered Bootstrap

We quantify sampling uncertainty by treating the 36 fictional identities as clusters. For each of 10,000 fixed-seed bootstrap replicates, we sample 36 identities with replacement, retain all scored questions associated with every drawn identity, recompute fine-grained-subtask means, macro-average them into the seven task scores, and then macro-average those task scores into Overall F1. Kimi K3 obtains 71.57 Overall F1 (95% CI: [69.41, 73.58]), Qwen3.8-Max obtains 71.24 (95% CI: [69.64, 72.89]), and Gemini 3.5 Flash obtains 70.84 (95% CI: [68.87, 72.67]). These intervals characterize resampling uncertainty over fictional identities; overlapping marginal intervals do not by themselves establish a paired model difference.

Identity-level resampling is necessary because the 11,045 benchmark questions are not 11,045 independent experimental units. Questions derived from the same identity reuse correlated names, document numbers, dates, financial values, document templates, and physical materials, and may include different prompts or capture conditions applied to related evidence. Treating every question as independent would therefore overstate the effective sample size and generally produce confidence intervals that are too narrow.

Figure 16 shows the resulting Overall intervals for the five highest-scoring commercial and open-weight models.

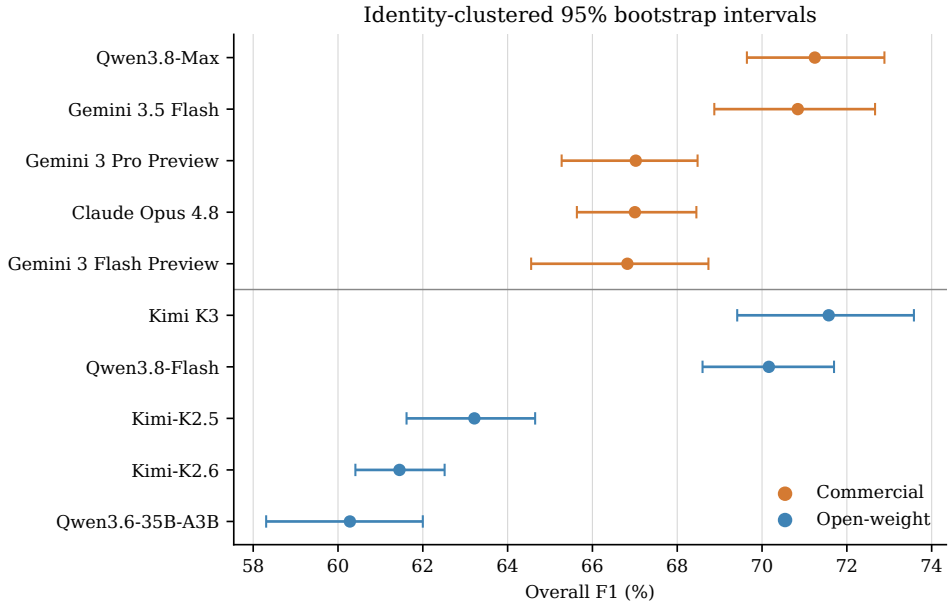


Figure 16: Overall F1 and percentile 95% confidence intervals from 10,000 identity-clustered bootstrap replicates for the five highest-scoring models in each model group.

E.4 Paired Model Comparison for Top Models

For paired comparisons, the same identity draw is applied to both models and the Overall F1 difference is recomputed within every replicate. This pairing preserves question and identity composition and removes variation caused by using different bootstrap samples. Differences are expressed in percentage points as the first model minus the second, with percentile 95% confidence intervals. As Figure 17 shows, Kimi K3 exceeds Qwen3.8-Max by 0.33 points (95% CI: $[-1.04, 1.77]$), while Qwen3.8-Max exceeds Gemini 3.5 Flash by 0.40 points (95% CI: $[-1.11, 1.91]$). Gemini 3.5 Flash differs from Kimi K3 by -0.73 points (95% CI: $[-2.34, 0.84]$). None of these intervals excludes zero, so the leading systems are not distinguishable under identity-clustered resampling.

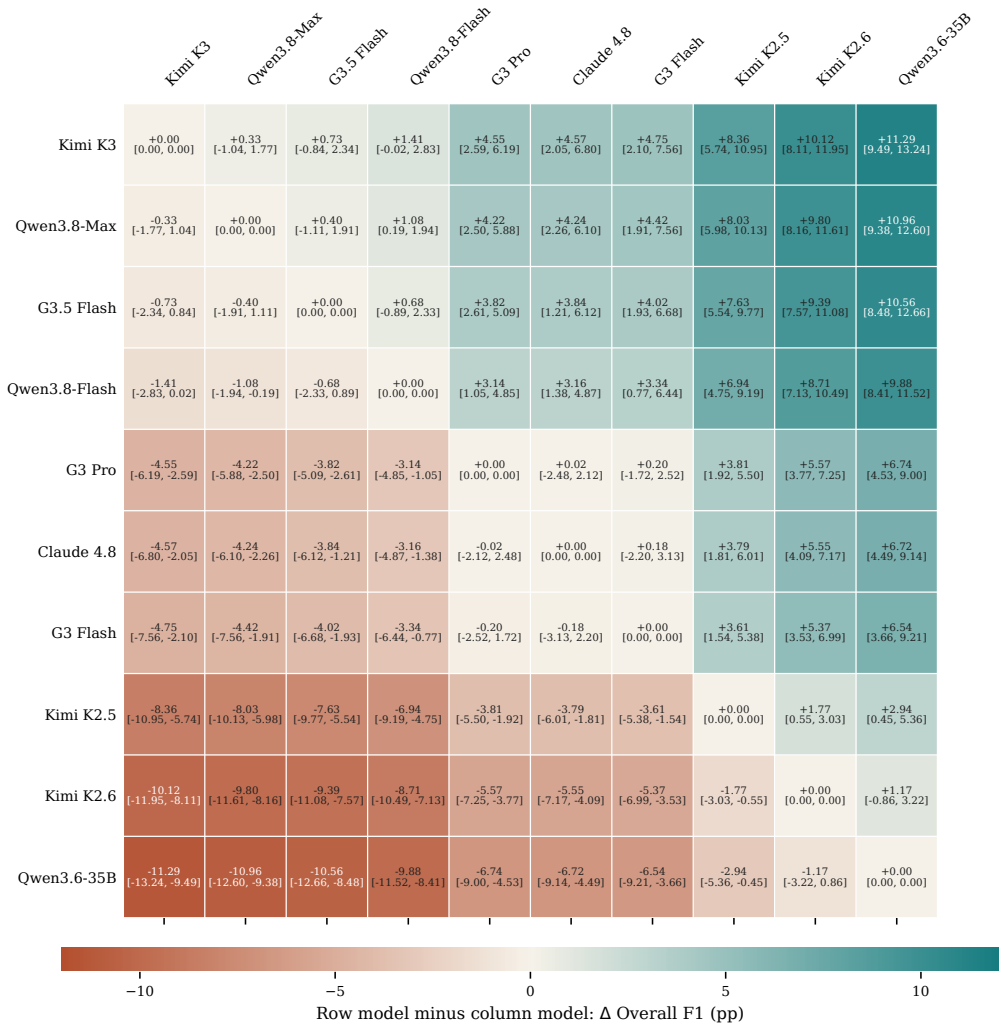


Figure 17: Paired Overall F1 difference matrix for the five highest-scoring commercial and open-weight models. Each cell reports row minus column in percentage points, followed by the 95% CI in brackets.

E.5 Question Coverage

We count a model as correct only when its instance-level F1 is exactly 1.0. Figures 18 and 19 show the complete distribution of this count. The horizontal axis is the number of exactly correct models, from 0 to 39, and the vertical axis is the percentage of questions in the corresponding task or capture condition. Each histogram is normalized independently, so its 40 bars sum to 100%.

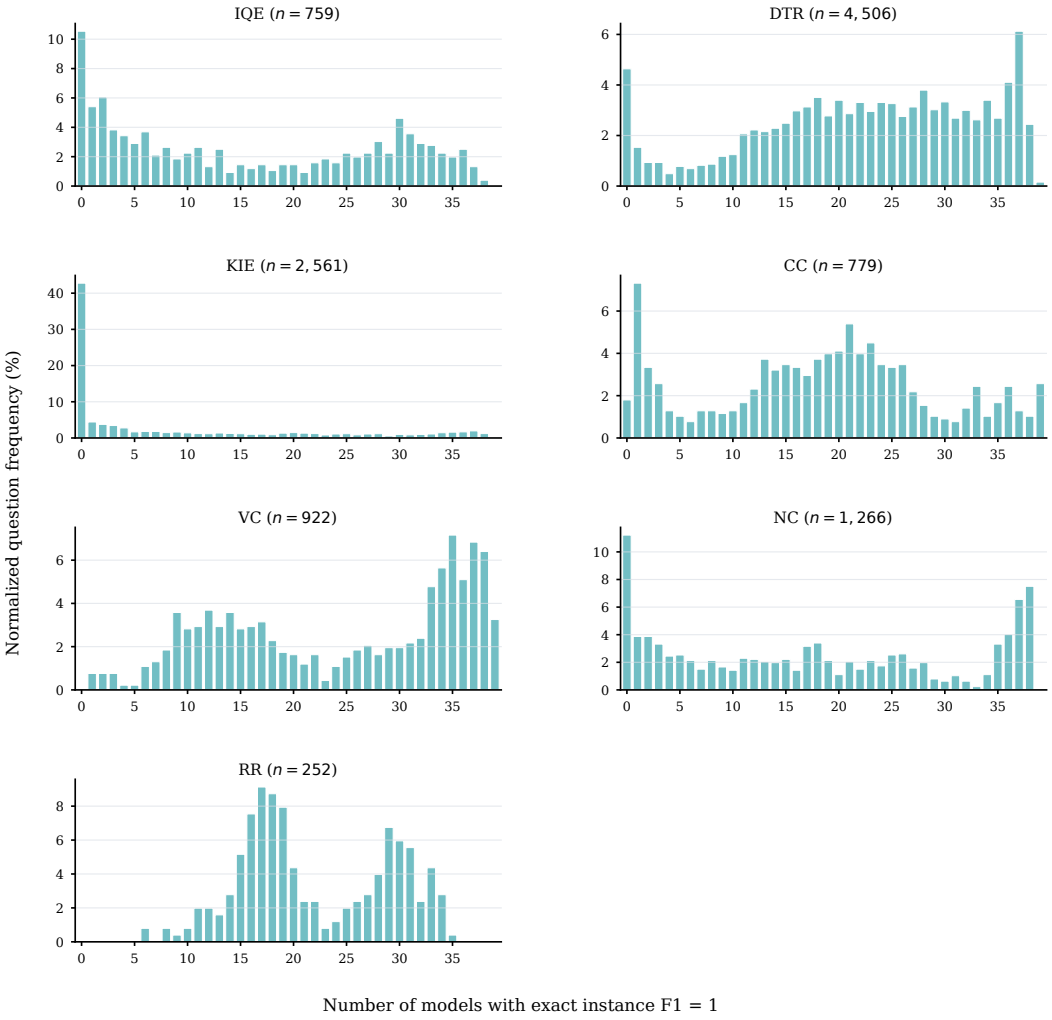


Figure 18: Normalized distribution of the number of models that answer each question exactly correctly, grouped by task. Bar heights sum to 100% within each panel; the annotation gives the number of questions.

E.6 Release Time and Parameter Scale

Figure 20 relates Overall F1 to model release month. Release time is positively associated with performance across all 39 models (Spearman $\rho = 0.80$).

Figure 21 separates dense and mixture-of-experts (MoE) models so that parameter counts retain a consistent meaning within each panel. Dense-model total parameters have the strongest rank association with performance ($\rho = 0.79$). For MoE models, total parameters ($\rho = 0.42$) are more strongly associated with performance than activated parameters per token ($\rho = 0.27$). Commercial models are omitted because their parameter counts are undisclosed. These cross-sectional associations are descriptive and do not establish a causal effect of release time or scale.

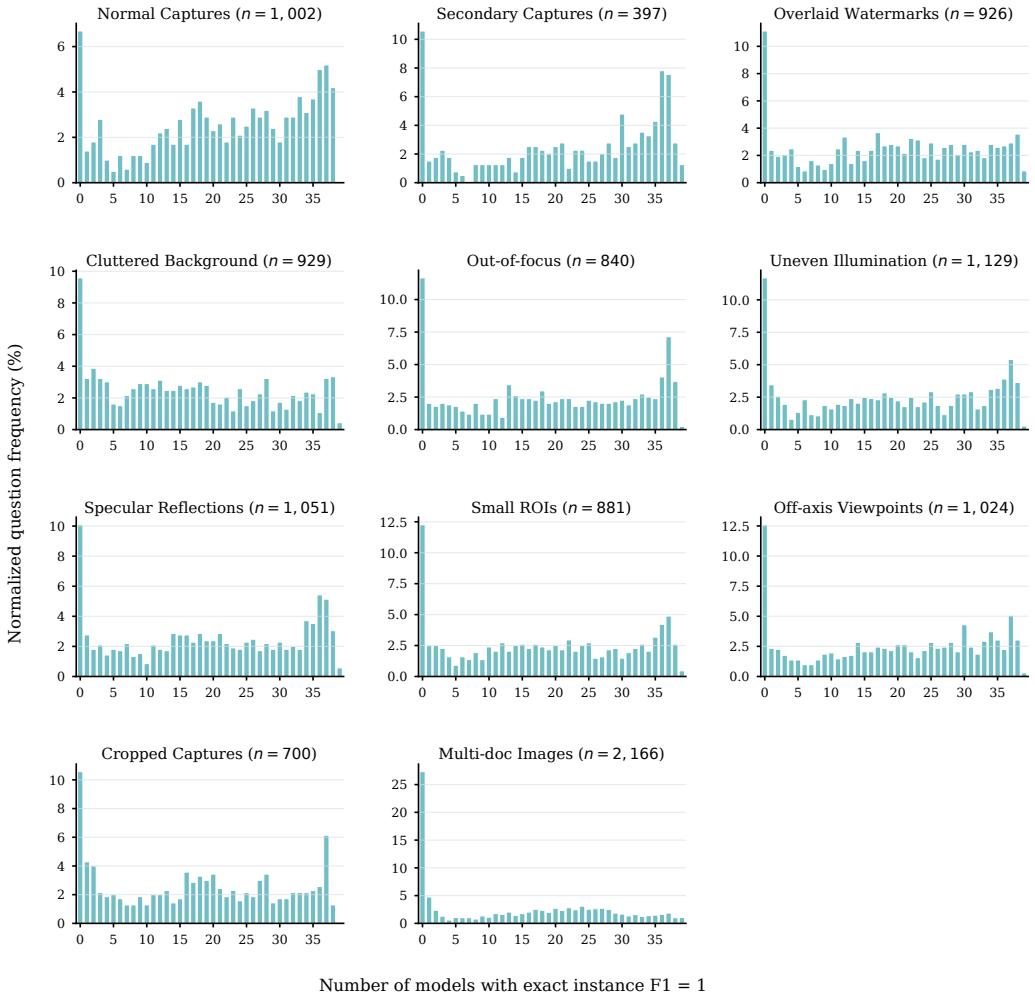


Figure 19: Normalized distribution of the number of models that answer each question exactly correctly, grouped by capture condition. Bar heights sum to 100% within each panel; the annotation gives the number of questions.

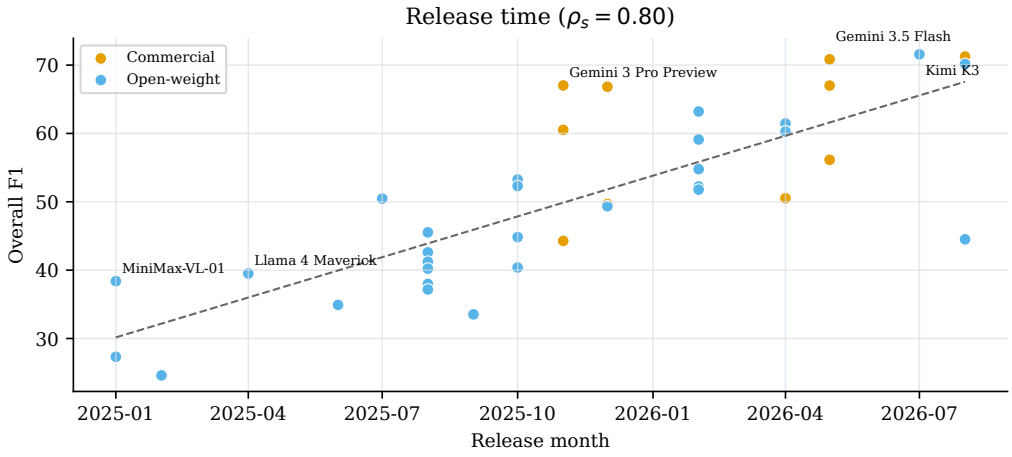


Figure 20: Overall F1 in relation to release month for all 39 models. The dashed line shows a descriptive linear trend.

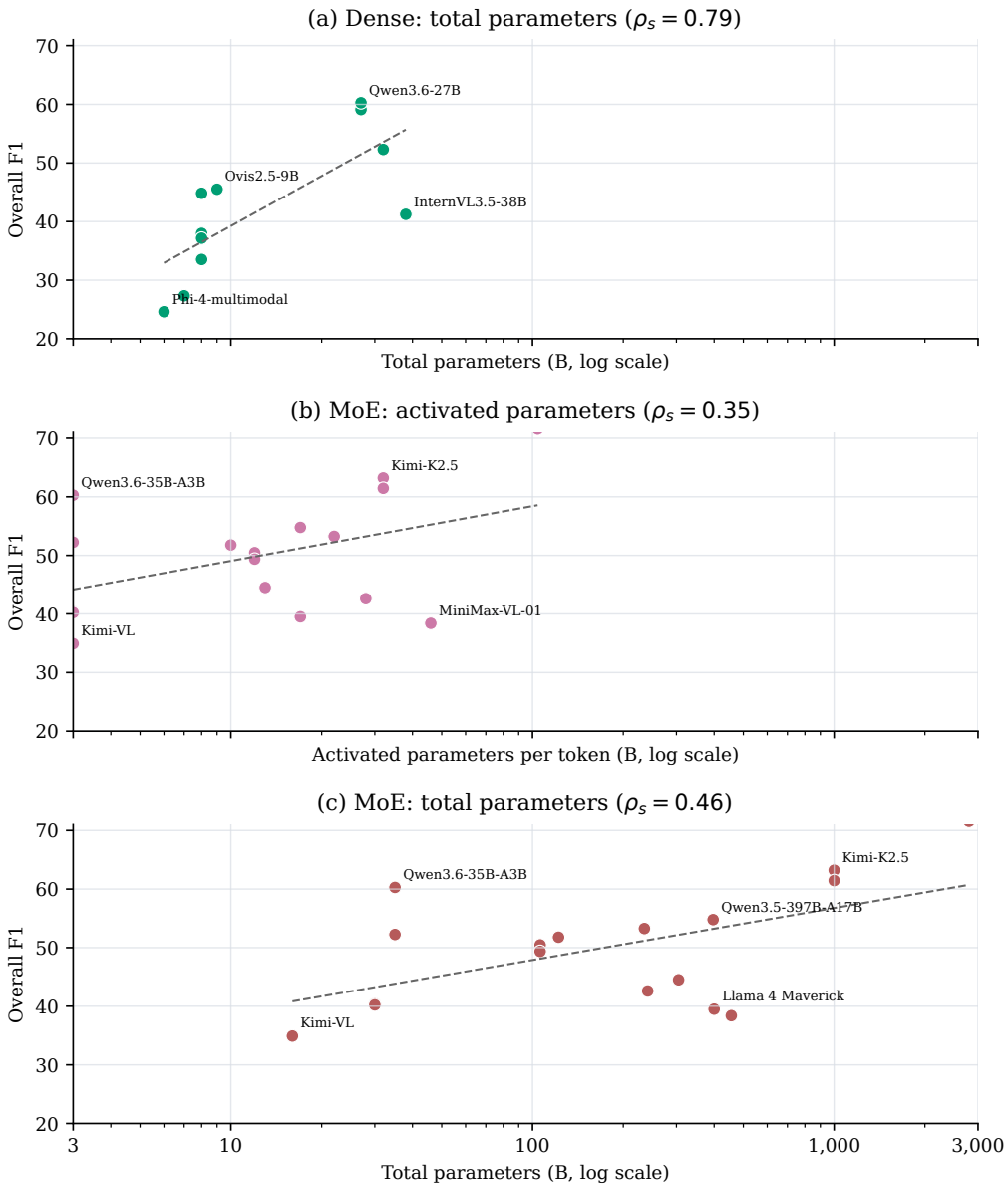


Figure 21: Overall F1 versus parameter scale for the 29 open-weight models with disclosed counts: (a) total parameters for 11 dense models, (b) activated parameters per token for 18 MoE models, and (c) total parameters for the same 18 MoE models. Dashed lines show descriptive linear trends; all parameter axes are logarithmic.

E.7 Model-Level Error-Context Distribution

We further report the observable error contexts of the five highest-scoring commercial and five highest-scoring open-weight models. The analysis assigns every model–question pair with strict instance F1 below one to one mutually exclusive context, including partially correct outputs. Invalid JSON is assigned to Output format; remaining errors on multiple-image inputs are assigned to Multi-image; and other errors are assigned to Perception, Numerical, or Domain rule according to their task context. These deterministic labels locate errors but do not establish their latent causes.

Figure 22 reports each context as a share of all 11,045 scored questions for that model; consequently, the full bar equals the model's strict-error rate rather than 100%. Multi-image and Perception jointly account for 29.1–36.8% of all scored questions and 84.1–88.4% of the assigned error contexts across these ten models. Numerical contexts are smaller (3.1–5.1%), while Output format and Domain rule each remain below 2.3%. The consistent ordering indicates that cross-image evidence integration and visual information acquisition dominate the observable error burden even among the strongest models.

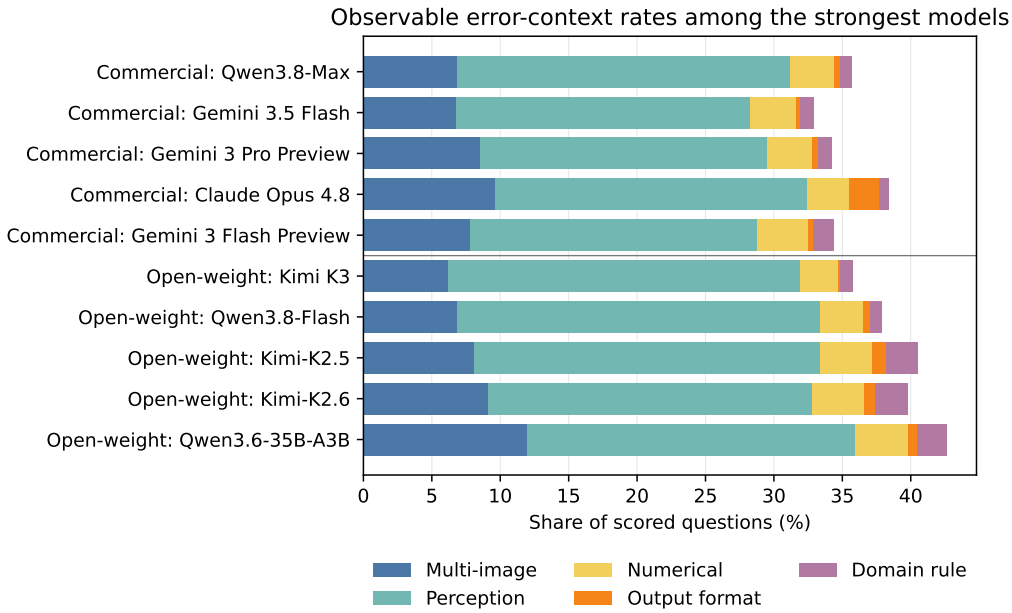


Figure 22: Observable error-context rates for the five highest-scoring commercial and open-weight models. Bars show the share of all scored questions assigned to each mutually exclusive context and therefore sum to each model's strict-error rate.

E.8 Chance-Corrected Ranking Stability

Closed, often binary decision spaces (notably VC and RR) admit high label-prior floors, so raw macro-F1 could in principle inflate the standing of models that merely emit a frequent valid answer. To rule this out we chance-correct every task score against its Random-Answer floor, $\mathrm{F1}_t^{\mathrm{corr}} = (\mathrm{F1}_t - \mathrm{floor}_t^{\mathrm{Random}})/(100 - \mathrm{floor}_t^{\mathrm{Random}})$, clip at zero, and macro-average the seven corrected task scores exactly as for Overall. The corrected 39-model ranking is almost identical to the raw one (Spearman $\rho = 0.998$); within the top-10 only ranks 3–4 swap (Qwen3.8-Flash and Gemini 3.5 Flash exchange places), and no model moves by five or more positions. Two diagnostic signals remain worth noting. First, six models (InternVL3.5-241B-A28B, InternVL3.5-8B, MiniCPM-V 4.5, Kimi-VL, LLaVA-OneVision-1.5-8B, Janus-Pro-7B) score exactly the Constant-Answer floor on RR (45.87), i.e.they collapse to a fixed output on this binary task. Second, Random Answer (61.16 on VC, 50.13 on RR) exceeds 19 of the 39 models on each of these two tasks. These patterns reflect the small closed answer spaces of the underlying credit checks rather than benchmark error, and they are the reason we report the chance-corrected ranking alongside the raw one.