

Mortgage-Bench: A Benchmark for End-to-End Mortgage Loan Processing Agents

Nathan Sekar, João Aires de Matos, Michael Vandi · Addy AI

Version 2 · October 2026

Abstract

Mortgage processing agents must gather evidence from many documents, resolve conflicts, apply program rules, and justify a credit decision. Existing evaluations test isolated extraction or question answering, not this workflow. Mortgage-Bench is a synthetic benchmark of 150 cases across FHA, VA, USDA, jumbo, and DSCR loans at three difficulty levels. The agent completes and cites an application, reconciles conflicts, flags risks, assesses readiness, and recommends Approve, Suspend, or Deny. Six weighted components give a score from 0 to 1; a wrong decision caps the score at 0.50, and a case passes at 0.80. With three tries per case, Claude Sonnet 5.5 and GPT-5.6 Luna passed about 38% of attempts, and Gemini 3.8 Flash chose the right decision most often (52%) but passed none. Agents were better at extraction and citation than at readiness and decisions. A document heuristic with no language model chose the right decision in 97 of 150 cases, more than any model, yet passed none. It applies the synthetic program thresholds that set the gold decisions, most of which the models were never shown, so we report pass rate and decision accuracy as the primary metrics and read the decision gap with that in mind.

Keywords: mortgage processing; underwriting; language models; AI agents; agent evaluation; benchmark; document reasoning; synthetic data

1 Introduction

Mortgage origination is an evidence-integration problem. A processor combines income, asset, credit, employment, and property documents, and the recommendation must rest on a documented ability-to-repay assessment [1,2] and on program rules from FHA, VA, USDA, and private investors [3–5]. Production costs averaged \$10,936 per loan in the second quarter of 2026 [6], and purchase loans took 36.8 days to close [7].

Lenders are adopting AI, and oversight is following: Freddie Mac and Fannie Mae require AI governance frameworks [8,9], model risk guidance leaves agentic AI out of scope [10], and every denial needs specific reasons [11,12].

Evaluation has moved to multi-metric, interactive, and workplace settings that grade the final

system state [13–19], but financial benchmarks still focus on question answering [20–22], and the closest lending benchmark, MortarBench, tests single-turn questions over synthetic bank statements [23]. No public benchmark covers the full processor workflow.

Mortgage-Bench fills this gap. Its tasks follow the pipeline of a production mortgage platform, but its data is fully synthetic, so it is reproducible and exposes no personal data. It is a research benchmark, not a compliance test. Contributions:

1. A simulated processor workstation with five surfaces, including an authoritative verification portal (Section 4.1).
2. 150 seeded cases: five programs, three difficulty levels, ten cases each (Section 4.2).
3. A six-component rubric that caps the score when the final decision is wrong (Section 4.4).
4. An evaluation of four language models with one try per case, three current models with three tries per case, and three reference agents (Sections 6 and 7).

2 Related Work

Agent benchmarks evaluate models inside simulated web, desktop, and enterprise systems and grade the resulting state [24–30]; τ -bench added pass^k , the chance that all k trials succeed [16]. Financial and document benchmarks test extraction and question answering [31–35]. Medicine, law, and insurance underwriting have expert-designed agent benchmarks [36–38], and language models have shown racial disparities in underwriting recommendations [39,40]. Table 1 compares the closest benchmarks; we follow recent guidance on validity, cost, and statistics [41–50].

Table 1. Mortgage-Bench compared with the closest prior benchmarks.

Benchmark	Domain	Interaction	Evidence	What is scored
FinanceBench [22]	Corporate finance	Single-turn question answering	Public company filings	Answer correctness against evidence-backed answers
τ -bench [16]	Retail and airline customer service	Multi-turn tool use with a simulated user	Databases and policy documents	Final database state; pass^k
TheAgentCompany [19]	Company operations	Multi-step tool use	Intranet applications and files	Checkpoint-based partial credit
UNDERWRITE [38]	Insurance underwriting	Multi-step tool use	Environments co-designed with underwriting experts	Task success and repeated-trial reliability
MortarBench [23]	Mortgage origination	Single-turn question answering	Synthetic bank statements and loan application data	Exact match and F1

Benchmark	Domain	Interaction	Evidence	What is scored
Mortgage-Bench	Mortgage processing	Multi-step work across five surfaces	Synthetic loan package, email inbox, verification portal	Six components with a decision cap; pass at 0.80

3 Domain Background

3.1 The processing workflow

Processors collect and verify documents, enter loan data, order verifications, and resolve discrepancies before an underwriter decides. Automated underwriting systems set much of the required documentation [3,51,52] and, like Mortgage-Bench, do not assess legal compliance [51].

3.2 Loan programs

The five programs differ in who sets the rules and what evidence decides eligibility (Table 2). Mortgage-Bench uses synthetic rules and thresholds; its cases do not state any program's actual policy.

Table 2. Loan programs in Mortgage-Bench and the real-world features they abstract.

Program	Governing guidance	Features that shape the workflow
FHA	HUD Handbook 4000.1 [3]	Mortgage insurance; TOTAL Mortgage Scorecard used through an automated underwriting system, with Accept or Refer outcomes; manual underwriting after a Refer
VA	VA Pamphlet 26-7 [4]	Residual income test alongside the debt-to-income ratio; a ratio above 41% calls for closer scrutiny unless residual income exceeds the guideline by at least 20%; VA sets no minimum credit score
USDA	HB-1-3555 [5]; 7 CFR 3555.10 [53]	Household income at or below the program's moderate income limit; eligible rural property location; Guaranteed Underwriting System
Jumbo	Private investor guidelines	Loan amount above the county conforming loan limit (2026 baseline for one-unit properties: \$832,750; high-cost ceiling: \$1,249,125) [54]; typically stricter reserve and documentation requirements

Program	Governing guidance	Features that shape the workflow
DSCR	Private investor guidelines	Business-purpose credit on non-owner-occupied rental property, outside Regulation Z and the Ability-to-Repay rule [55]; qualification on property cash flow, where DSCR is gross monthly rent divided by principal, interest, taxes, insurance, and association dues (PITIA), with minimums near 1.00 that vary by lender [56]

3.3 Decision semantics

Approve: the file supports the loan, usually with conditions, which Regulation C splits into underwriting and closing conditions [57]. **Suspend:** the file cannot be decided until information is obtained; Regulation B then requires a notice of incompleteness or of adverse action [12]. **Deny:** the loan cannot be made on the program's terms, and the reasons must be specific [11,12]. Mortgage-Bench therefore scores the rationale with the decision.

4 Benchmark Design

4.1 Environment

Each case gives the agent five surfaces (Table 3). The inbox may contradict a document, and the verification portal is authoritative when sources conflict.

Table 3. Information surfaces available to the agent in each case.

Surface	Contents	Role in the case
Loan package	Borrower identity, income and tax documents, assets, property details, credit information, and program-specific forms	Primary evidence
Email inbox	Messages from the borrower and third parties	May confirm or contradict a document value
Verification portal	Verified values for designated fields	Authoritative source when sources conflict
Application form	Required application fields	The agent enters each value and its source
Underwriting screen	Readiness, conditions, risk flags, decision, and rationale	Records the agent's assessment and recommendation

4.2 Case generation and difficulty

A seeded generator produces ten cases for each program and difficulty level ($5 \times 3 \times 10 = 150$), using synthetic rules that mimic workflow complexity, not official policy. Difficulty rises with the complexity and unreliability of the evidence (Table 4). The gold decisions are 50 Approve,

60 Suspend, and 40 Deny.

Table 4. Difficulty levels.

Level	Evidence characteristics
Beginner	Relatively direct files with few distractors
Intermediate	Conflicts between sources, such as an email that disagrees with a pay stub
Advanced	Eligibility edge cases, weak credit, limited reserves, distractor documents, noisy text, and DSCR-specific challenges

4.3 Task protocol

Each case has seven steps:

1. Identify the loan program.
2. Gather evidence from the available sources.
3. Complete the required application fields and cite the source of each entered value.
4. Where sources conflict, reconcile the discrepancy using the source that the environment designates as verified.
5. Assess readiness and record any conditions.
6. Select the applicable risk flags.
7. Issue an Approve, Suspend, or Deny recommendation with a rationale, and submit.

Answer keys are hidden from the agent.

4.4 Scoring

The case score is a weighted sum of six components, each between 0 and 1 (Table 5):

$$S = \sum_k w_k s_k, \text{ and } S \text{ is replaced by } \min(S, 0.50) \text{ when the final decision is incorrect.}$$

A case passes when $S \geq 0.80$.

Table 5. Scoring components.

Component	Weight	What it measures	Matching rule
Application-field accuracy	25%	Correct values in the required application fields	Money within one cent; other fields match ignoring case
Source-citation correctness	20%	Each entered value cites a source that contains it	The cited source is correct and the agent opened it
Risk-flag selection	15%	Applicable flags selected and inapplicable flags avoided	Overlap with the gold flag set; an empty set matches an empty set
Income and conflict reconciliation	10%	Conflicting values resolved to the verified source	If incomes conflict, the agent must open and enter the verified income

Component	Weight	What it measures	Matching rule
Readiness and conditions	10%	Correct readiness status and required conditions	Ready only when no conditions remain
Final decision with rationale	20%	Correct Approve, Suspend, or Deny with a supported rationale	Rule-based, with no LLM judge [58]: 70% for the label, 30% for overlap with the gold reason tags

A wrong decision caps the score at 0.50, so evidence handling alone cannot pass; with a correct decision, the agent still needs 75% of the remaining credit (Appendix B). A wrong critical field caps the score at 0.79 and an uncited one at 0.69, so strong extraction can still fail. Weights and caps were set by design.

4.5 Quality assurance

Each case was checked for solvability, internal consistency, and answer leakage; the oracle scored 1.00 and the naive agent 0.36 (Section 5.2) [41]. The workflow resembles processing and underwriting, but the cutoffs are synthetic, and the cases and rubric have not had a recorded expert review.

5 Experimental Setup

5.1 Language-model agents

All agents (Table 6) received the same instructions, tools, and cases. In the original runs each case ran once: Claude Sonnet 4.5, GPT-4o, and Gemini 2.5 Flash at temperature 0 on August 11, 2026, and GPT-5.6 Luna, which rejects the temperature parameter, on August 12 with the same cases. Temperature 0 is not deterministic on hosted models [60,61]. In October 2026 we reran the suite with Claude Sonnet 5.5, GPT-5.6 Luna, and Gemini 3.8 Flash, three tries per case (450 attempts per model); only Gemini accepts a temperature, set to 0. GPT-6 and open-weight models were not tested.

Table 6. Language-model agents and settings.

Agent	Provider	API identifier	Sampling and reasoning settings	Run date
Claude Sonnet 4.5	Anthropic	claude-sonnet-4-5	Temperature 0	August 11, 2026, 16:15–16:35 UTC
GPT-4o	OpenAI	gpt-4o (alias; no dated snapshot)	Temperature 0	August 11, 2026, 16:49–16:58 UTC
Gemini 2.5 Flash	Google (Vertex AI)	gemini-2.5-flash	Temperature 0	August 11, 2026, 16:35–16:42 UTC

Agent	Provider	API identifier	Sampling and reasoning settings	Run date
GPT-5.6 Luna	OpenAI	gpt-5.6-luna	Temperature not supported	August 12, 2026, 17:21–17:43 UTC
Claude Sonnet 5.5	Anthropic	claude-sonnet-5-5	Temperature not supported	October 7, 2026, 18:09–19:24 UTC
GPT-5.6 Luna	OpenAI	gpt-5.6-luna	Temperature not supported	October 7–8, 2026, 19:24–01:37 UTC
Gemini 3.8 Flash	Google	gemini-3.8-flash	Temperature 0	October 7, 2026, 16:55–18:08 UTC

5.2 Reference agents

Three agents without a language model bound the score range, all scored by the same grader. The oracle copies the answer key (score 1.00, every case passed). The naive agent enters the gold field values, cites nothing, marks every file ready, and approves every loan (0.36, no passes). The document heuristic reads the visible documents with simple patterns and applies the benchmark's own program thresholds for credit score, LTV, reserves, and DSCR, but does not compute DTI; it never sees the answer key (0.667, no passes).

5.3 Procedure

All agents ran on the same 150 cases, with no skipping or selective reruns. Costs are list prices on the run dates.

5.4 Statistical analysis

Mean scores have 95% intervals from a case bootstrap (1,000 resamples, 2.5th and 97.5th percentiles, seed 4219) [62]; for the October reruns, each case's three tries are resampled together. Pass rates use Wilson intervals [63] for single runs and the same case bootstrap for the reruns. With 150 cases, a pass-rate interval near 40% spans about ± 8 points. Paired comparisons (paired bootstrap, McNemar's test [64], Holm adjustment [65]) are left to future work.

6 Results

6.1 Overall performance

In the original runs, Claude Sonnet 4.5 scored best (mean 0.69, pass rate 41%), followed by GPT-4o (0.66), Gemini 2.5 Flash (0.64), and GPT-5.6 Luna (0.61) (Table 7). No agent passed half the cases, and all pass-rate intervals overlap; only the Claude Sonnet 4.5 and GPT-5.6 Luna

mean-score intervals do not.

GPT-5.6 Luna had the lowest mean but the second-highest pass rate, consistent with some very low scores from incomplete citations and unsaved applications.

Table 7. Overall benchmark performance (150 cases).

Agent	Mean score (95% interval)	Pass rate, $S \geq 0.80$ (95% CI)	Correct decisions	Approx. API cost per run	Approx. cost per passing case
Claude Sonnet 4.5	0.69 (0.662–0.721)	41% (33–49%)	81 / 150 (54.0%)	\$2.46	\$0.040
GPT-4o	0.66 (0.632–0.687)	33% (26–41%)	71 / 150 (47.3%)	\$1.30	\$0.026
Gemini 2.5 Flash	0.64 (0.613–0.666)	32% (25–40%)	60 / 150 (40.0%)	\$0.34	\$0.007
GPT-5.6 Luna	0.61 (0.573–0.647)	35% (28–43%)	62 / 150 (41.3%)	\$0.23	\$0.004
Document heuristic	0.67	0% (0–2.5%)	97 / 150 (64.7%)	None	None
Naive baseline	0.36	0% (0–2.5%)	50 / 150 (33.3%)	None	None
Oracle	1.00	100% (97.5–100%)	150 / 150 (100%)	None	None

A run is one pass over all 150 cases. Always choosing Suspend gets 60 of 150 decisions right (40.0%). Intervals were not computed for the reference agents. Cost per passing case is the run cost divided by the number of passed cases.

6.2 Current models, three tries per case

Claude Sonnet 4.5 retires on November 30, 2026 [59], so we reran the same cases with three current models, three tries each (Table 8; not pooled with Table 7). Sonnet 5.5 and Luna were close on mean score and pass rate, but passing was not stable across tries: Sonnet 5.5 passed 35 cases on all three tries and 44 on only one or two, and Luna 41 and 32 (pass³ of 23.3% and 27.3%). Gemini 3.8 Flash had the highest decision accuracy but passed no case: it never chose Deny, and its field-value score (0.646) was far below the others' (0.92 to 0.93; Table A1). Luna chose Suspend in 422 of 450 attempts.

Table 8. Current models, three tries per case (450 attempts per model).

Agent	Mean score (95% interval)	Pass rate, $S \geq 0.80$ (95% CI)	Decision accuracy	pass ³ (all three tries)	Approx. API cost (450 attempts)
Claude Sonnet 5.5	0.662 (0.635–0.689)	37.8% (31–44%)	43.8% (197/450)	23.3% (35/150)	\$14.44
GPT-5.6 Luna	0.656 (0.628–0.684)	38.0% (31–45%)	42.2% (190/450)	27.3% (41/150)	\$0.71
Gemini 3.8 Flash	0.635 (0.615–0.655)	0.0%	52.0% (234/450)	0.0%	\$2.00

Pass and decision rates are shares of 450 attempts. Intervals resample the 150 cases with each case's three tries kept together. pass³ is the share of cases passed on all three tries. Costs are token-based estimates at list prices for all 450 attempts.

6.3 What the mean score rewards

The document heuristic scored 0.67, above three language models, without passing a case. It chose the right decision in 97 of 150 cases (64.7%), against 81 for Claude Sonnet 4.5 and 52.0% of attempts for Gemini 3.8 Flash. Always choosing Suspend is right 40% of the time, and Gemini 2.5 Flash did no better. The heuristic gets every gold Approve right and turns most Suspend files into Approve (Table A2); it fails because filling the application and citing sources also count.

The setup explains the gap. Gold decisions follow fixed synthetic thresholds for each program, and the heuristic applies those thresholds to the values it extracts; it misses most gold Suspends because it does not compute DTI. The language-model agents were never shown the credit-score, LTV, or DTI cutoffs (the documents state only the reserve, DSCR, and USDA income limits), and their prompt asks them to prefer Suspend when eligibility issues remain unresolved, which matches their over-suspension (Table A2). Decision accuracy therefore measures, in part, whether an agent can infer unstated cutoffs.

6.4 Skill-level performance

Citation, extraction, and income reconciliation were stronger than readiness and the final decision. In the October reruns, Sonnet 5.5 and Luna scored 0.99 or higher on citations and income checks and 0.92 to 0.93 on field values, but 0.43 to 0.47 on readiness and 0.52 on the decision (Table A1).

6.5 Final decisions

Wrong decisions were the main failure: 69 of 150 for Claude Sonnet 4.5, 79 for GPT-4o, 88 for GPT-5.6 Luna, and 90 for Gemini 2.5 Flash. Because they cap the score, decision accuracy bounds the pass rate. Error direction matters: a wrong approval creates credit risk, a wrong denial harms the applicant, and an unnecessary suspension delays the file. The current models rarely deny: Gemini 3.8 Flash never did, and Sonnet 5.5 and Luna denied 5 and 1 of 120 gold-Deny attempts; both sent most gold-Approve files to Suspend (Table A2). The original runs did not record which wrong label was chosen.

6.6 Performance by difficulty

For Claude Sonnet 4.5, intermediate cases scored highest (0.82, 60% pass), above beginner (0.67, 44%) and advanced (0.59, 18%) (Table 9). The advanced drop fits the added edge cases and noise; the beginner result needs a check of its gold-decision mix. The October reruns show the same pattern: intermediate cases had the highest pass rates for Sonnet 5.5 (68%) and Luna (84%) (Table A3).

Table 9. Performance by difficulty level, Claude Sonnet 4.5.

Difficulty	Cases	Mean score	Pass rate (95% CI)
Beginner	50	0.67	44% (31–58%)

Difficulty	Cases	Mean score	Pass rate (95% CI)
Intermediate	50	0.82	60% (46–72%)
Advanced	50	0.59	18% (10–31%)

6.7 Performance by loan program

VA cases were easiest (26 of 30 passed) and USDA hardest (2 of 30) (Table 10). In the October reruns VA was again easiest, but USDA was not the hardest (38% for Sonnet 5.5; Table A3), so the gap likely reflects how the generator builds each program rather than real underwriting difficulty.

Table 10. Performance by loan program, Claude Sonnet 4.5.

Program	Cases	Mean score	Pass rate (95% CI)
FHA	30	0.63	33% (19–51%)
VA	30	0.86	87% (70–95%)
USDA	30	0.62	7% (2–21%)
Jumbo	30	0.67	37% (22–54%)
DSCR	30	0.69	40% (25–58%)

6.8 Cost

A full original run cost \$0.23 (GPT-5.6 Luna) to \$2.46 (Claude Sonnet 4.5), about \$0.0015 to \$0.016 per case. On pass rate, Luna and Claude Sonnet 4.5 form the cost frontier [44], though the differences are within the intervals.

7 Discussion

7.1 Evidence handling and judgment

Agents handled evidence better than judgment: they extracted, cited, and reconciled values more reliably than they assessed readiness or decided, as other agent benchmarks report [14,15,19,38]. Reconciliation has an explicit authoritative source; readiness and decisions require applying rules to the whole file.

7.2 Choosing headline metrics

No single number captures a workflow [13]. A heuristic that makes more correct decisions than any model yet passes no case shows the limit of a weighted mean, so results should lead with pass rate and decision accuracy, report the components, test alternative weights and caps [41,43], and add pass^k over repeated runs [16].

7.3 Implications for system design

Three design choices follow: separate policy evaluation from extraction, so a deterministic

engine applies the rules and the model only extracts inputs; keep provenance for every value (document, field, page); and make the rules behind a recommendation visible. Suspend should follow from evidence that is actually missing, not from uncertainty about the rules: over-suspension was the most common error here.

These controls match the NIST AI Risk Management Framework [66,67], GSE AI governance requirements [8,9], and adverse-action rules [11,12]. The EU AI Act treats creditworthiness assessment as high-risk [68] from December 2, 2027 [69], and Colorado requires explanations and human review for lending decisions from January 1, 2027 [70]. Agents like these should support, not replace, qualified human review.

8 Limitations

Synthetic data. Synthetic rules and documents miss real-world complexity, may be easier than real files [48], and are no evidence of compliance [1,3–5].

Scope. U.S. mortgage workflows only.

Evaluation design. One prompt; one run per case originally and three in October; hosted models are not deterministic [60]; Tables 9 and 10 cover Claude Sonnet 4.5 only; Table A3 gives the October reruns.

Model currency. Claude Sonnet 4.5 retires on November 30, 2026 [59]; GPT-6 and open-weight models were not tested.

Scoring. The grader is rule-based. Weights and caps were set by design; sensitivity and agreement with expert judgment are not yet reported.

Decision rules. The agents were not shown the credit-score, LTV, and DTI cutoffs that set most gold decisions, and the prompt told them to prefer Suspend when issues remained; a rerun with all thresholds in the prompt would separate applying rules from inferring them.

Missing analyses. No error taxonomy, calibration, or paired tests, and no confusion matrices for the original runs.

Fairness. Disparate treatment was not tested; prior work found it in language-model underwriting [23,39].

Contamination. Public cases may enter training data; a private held-out split with canary strings would prevent this [44,49].

Competing interests. Addy AI built the benchmark and sells mortgage AI software; none of the evaluated agents is an Addy product.

9 Conclusion

Mortgage-Bench tests agents on the full mortgage-processing chain, from evidence to a documented recommendation. Current models handle documents and citations better than

readiness and decisions, and none passes half the cases. A heuristic that applies the benchmark's own thresholds makes more correct decisions than any model, so agents should be given, or tested on, the rules they are expected to apply. High-stakes agents should be evaluated end to end, with outcome-based metrics and repeated runs.

Declarations

Data and code availability. The generator, the 150 cases (suite mortgage-bench-development-v1, version 1.0.0), answer keys, grader, harness, prompts, and all run outputs are in a private Addy AI repository (github.com/addy-ai/mortgage-bench) open to reviewers, who rerun with their own API keys.

Competing interests. The authors are affiliated with Addy AI, which develops mortgage AI software and owns and licenses Mortgage-Bench.

Ethics. All borrower data is synthetic.

Funding. This work was funded by Addy AI.

Author contributions. Nathan Sekar built the benchmark, ran the evaluations, and wrote the first draft. João Aires de Matos revised the manuscript. Michael Vandi initiated the paper.

Use of AI tools. Claude (Anthropic) helped edit the manuscript and check references; the authors reviewed the final text.

References

1. Consumer Financial Protection Bureau. Ability-to-Repay/Qualified Mortgage Rule. Updated October 22, 2021. <https://www.consumerfinance.gov/rules-policy/final-rules/ability-to-pay-qualified-mortgage-rule/>
2. Truth in Lending (Regulation Z). Minimum standards for transactions secured by a dwelling. 12 CFR §1026.43. Accessed October 7, 2026. <https://www.law.cornell.edu/cfr/text/12/1026.43>
3. U.S. Department of Housing and Urban Development. FHA Single Family Housing Policy Handbook 4000.1, Update 18. August 12, 2026. <https://www.hud.gov/hud-partners/single-family-handbook-4000-1>
4. U.S. Department of Veterans Affairs. VA Pamphlet 26-7, Lenders Handbook. Accessed October 7, 2026. https://www.knowva.ebenefits.va.gov/system/templates/selfservice/va_ssnew/help/customer/locale/en-US/portal/554400000001018/topic/554400000027323/VAP26-07-Lenders-Handbook
5. U.S. Department of Agriculture, Rural Housing Service. HB-1-3555: SFH Guaranteed Loan Program Technical Handbook (consolidated). May 5, 2025. <https://www.usda.gov/guidance-documents/rhs-handbook/rhs/hb-1-3555-sfh-guaranteed-loan-program-technical-handbook>
6. Mortgage Bankers Association. IMBs' production profits increase in second quarter of 2026. Press release. August 18, 2026. <https://www.mba.org/news-and-research/newsroom/news/2026/08/18/imbs-production-profits-increase-in-second-quarter-of-2026>
7. ICE Mortgage Technology. Mortgage Monitor, May 2026. May 11, 2026. <https://mortgagetechnology.com/resources/data-reports/may-2026-mortgage-monitor>
8. Freddie Mac. Single-Family Seller/Servicer Guide Bulletin 2025-16. December 3, 2025. <https://guide.freddiemac.com/app/guide/bulletin/2025-16>
9. Fannie Mae. Lender Letter LL-2026-04: Governance framework on use of artificial intelligence and

- machine learning. April 8, 2026. <https://singlefamily.fanniemae.com/media/45196/display>
10. Board of Governors of the Federal Reserve System, Federal Deposit Insurance Corporation, Office of the Comptroller of the Currency. SR 26-2: Revised guidance on model risk management. April 17, 2026. <https://www.federalreserve.gov/supervisionreg/srletters/SR2602.pdf>
 11. Equal Credit Opportunity Act. 15 U.S.C. §1691(d). <https://www.law.cornell.edu/uscode/text/15/1691>
 12. Equal Credit Opportunity (Regulation B). Notifications. 12 CFR §1002.9. <https://www.law.cornell.edu/cfr/text/12/1002.9>
 13. Liang P, Bommasani R, Lee T, et al. Holistic evaluation of language models. Transactions on Machine Learning Research. 2023. arXiv:2211.09110. doi:10.48550/arXiv.2211.09110
 14. Liu X, Yu H, Zhang H, et al. AgentBench: evaluating LLMs as agents. In: International Conference on Learning Representations (ICLR); 2024. arXiv:2308.03688
 15. Mialon G, Fourrier C, Swift C, Wolf T, LeCun Y, Scialom T. GAIA: a benchmark for general AI assistants. In: International Conference on Learning Representations (ICLR); 2024. arXiv:2311.12983
 16. Yao S, Shinn N, Razavi P, Narasimhan K. τ -bench: a benchmark for tool-agent-user interaction in real-world domains. In: International Conference on Learning Representations (ICLR); 2025. arXiv:2406.12045
 17. Drouin A, Gasse M, Caccia M, et al. WorkArena: how capable are web agents at solving common knowledge work tasks? In: International Conference on Machine Learning (ICML); 2024. arXiv:2403.07718
 18. Huang KH, Prabhakar A, Dhawan S, et al. CRMArena: understanding the capacity of LLM agents to perform professional CRM tasks in realistic environments. In: Proceedings of NAACL; 2025. arXiv:2411.02305. <https://aclanthology.org/2025.naacl-long.194>
 19. Xu FF, Song Y, Li B, et al. TheAgentCompany: benchmarking LLM agents on consequential real world tasks. In: Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track; 2025. arXiv:2412.14161
 20. Chen Z, Chen W, Smiley C, et al. FinQA: a dataset of numerical reasoning over financial data. In: Proceedings of EMNLP; 2021. arXiv:2109.00122
 21. Zhu F, Lei W, Huang Y, et al. TAT-QA: a question answering benchmark on a hybrid of tabular and textual content in finance. In: Proceedings of ACL-IJCNLP; 2021. <https://aclanthology.org/2021.acl-long.254>
 22. Islam P, Kannappan A, Kiela D, Qian R, Scherrer N, Vidgen B. FinanceBench: a new benchmark for financial question answering. arXiv. 2023;2311.11944
 23. Toles M, Lu Y, Munjal M, et al. MortarBench: evaluating mortgage loan origination agents. arXiv. 2026;2606.19416
 24. Zhou S, Xu FF, Zhu H, et al. WebArena: a realistic web environment for building autonomous agents. In: International Conference on Learning Representations (ICLR); 2024. arXiv:2307.13854
 25. Xie T, Zhang D, Chen J, et al. OSWorld: benchmarking multimodal agents for open-ended tasks in real computer environments. In: NeurIPS Datasets and Benchmarks Track; 2024. arXiv:2404.07972
 26. Boisvert L, Thakkar M, Gasse M, et al. WorkArena++: towards compositional planning and reasoning-based common knowledge work tasks. In: NeurIPS Datasets and Benchmarks Track; 2024. arXiv:2407.05291
 27. Huang KH, Prabhakar A, Thorat O, et al. CRMArena-Pro: holistic assessment of LLM agents across diverse business scenarios and interactions. Transactions on Machine Learning Research. 2026. arXiv:2505.18878
 28. Barres V, Dong H, Ray S, Si X, Narasimhan K. τ^2 -bench: evaluating conversational agents in a dual-control environment. In: Proceedings of the International Conference on Machine Learning (ICML). PMLR. 2026;306:6736-6782. arXiv:2506.07982
 29. Trivedi H, Khot T, Hartmann M, et al. AppWorld: a controllable world of apps and people for benchmarking interactive coding agents. In: Proceedings of ACL; 2024:16022-16076. doi:10.18653/v1/2024.acl-long.850
 30. Patwardhan T, Dias R, Proehl E, et al. GDPval: evaluating AI model performance on real-world economically valuable tasks. In: International Conference on Learning Representations (ICLR); 2026. arXiv:2510.04374
 31. Xie Q, Han W, Chen Z, et al. FinBen: a holistic financial benchmark for large language models. In:

- NeurIPS Datasets and Benchmarks Track; 2024. arXiv:2402.12659
32. Bigeard A, Nashold L, Krishnan R, Wu S. Finance Agent Benchmark: benchmarking LLMs on real-world financial research tasks. arXiv. 2025;2508.00828
 33. Mathew M, Karatzas D, Jawahar CV. DocVQA: a dataset for VQA on document images. In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV); 2021. arXiv:2007.00398
 34. Ma Y, Zang Y, Chen L, et al. MMLongBench-Doc: benchmarking long-context document understanding with visualizations. In: NeurIPS Datasets and Benchmarks Track; 2024. arXiv:2407.01523
 35. Zhang B, Lyjak A, Stewart E, Li Z, Suo S. ExtractBench: a benchmark for schema-guided enterprise document extraction. arXiv. 2026;2607.29677
 36. Jiang Y, Black KC, Geng G, et al. MedAgentBench: a realistic virtual EHR environment to benchmark medical LLM agents. NEJM AI. 2025. doi:10.1056/AIdbp2500144. arXiv:2501.14654
 37. Guha N, Nyarko J, Ho DE, et al. LegalBench: a collaboratively built benchmark for measuring legal reasoning in large language models. In: NeurIPS Datasets and Benchmarks Track; 2023. arXiv:2308.11462
 38. Dsouza A, Ramakrishnan R, Dickens C, Pohani B, Glaze CM. Benchmarking agents in insurance underwriting environments. arXiv. 2026;2602.00456
 39. Bowen DE III, Price SM, Stein LCD, Yang K. Measuring and mitigating racial disparities in LLMs: evidence from a mortgage underwriting experiment. Working paper. September 17, 2025. <https://ssrn.com/abstract=4812158>
 40. Azime IA, Kanubala DD, Afonja T, et al. Accept or deny? Evaluating LLM fairness and performance in loan approval across table-to-text serialization approaches. arXiv. 2025;2508.21512
 41. Zhu Y, Jin T, Pruksachatkun Y, et al. Establishing best practices for building rigorous agentic benchmarks. In: NeurIPS Datasets and Benchmarks Track; 2025. arXiv:2507.02825
 42. Reuel A, Hardy A, Smith C, Lamparth M, Hardy M, Kochenderfer MJ. BetterBench: assessing AI benchmarks, uncovering issues, and establishing best practices. In: NeurIPS Datasets and Benchmarks Track; 2024. arXiv:2411.12990
 43. Bean AM, Kearns RO, Romanou A, et al. Measuring what matters: construct validity in large language model benchmarks. In: NeurIPS Datasets and Benchmarks Track; 2025. arXiv:2511.04703
 44. Kapoor S, Stroebl B, Siegel ZS, Nadgir N, Narayanan A. AI agents that matter. Transactions on Machine Learning Research. 2025. arXiv:2407.01502
 45. Kapoor S, Stroebl B, Kirgis P, et al. Holistic Agent Leaderboard: the missing infrastructure for AI agent evaluation. In: International Conference on Learning Representations (ICLR); 2026. arXiv:2510.11977
 46. Miller E. Adding error bars to evals: a statistical approach to language model evaluations. arXiv. 2024;2411.00640
 47. Dror R, Baumer G, Shlomov S, Reichart R. The hitchhiker's guide to testing statistical significance in natural language processing. In: Proceedings of ACL; 2018;1383-1392. <https://aclanthology.org/P18-1128/>
 48. Gill A, Ravichander A, Marasović A. What has been lost with synthetic evaluation? In: Findings of EMNLP; 2025. arXiv:2505.22830
 49. Xu C, Guan S, Greene D, Kechadi MT. Benchmark data contamination of large language models: a survey. arXiv. 2024;2406.04244
 50. Sainz O, Campos JA, García-Ferrero I, Etxaniz J, Lopez de Lacalle O, Agirre E. NLP evaluation in trouble: on the need to measure LLM data contamination for each benchmark. In: Findings of EMNLP; 2023. arXiv:2310.18018
 51. Fannie Mae. Selling Guide B3-2-01, General Information on DU. April 1, 2026. <https://selling-guide.fanniemae.com/sel/b3-2-01/general-information-du>
 52. Freddie Mac. Loan Product Advisor FAQ. Accessed October 7, 2026. <https://sf.freddiemac.com/faqs/loan-product-advisor-faq>
 53. U.S. Department of Agriculture. Guaranteed Rural Housing Program, Definitions. 7 CFR §3555.10. <https://www.law.cornell.edu/cfr/text/7/3555.10>
 54. Federal Housing Finance Agency. FHFA announces conforming loan limit values for 2026. November 25, 2025. <https://www.fhfa.gov/news/news-release/fhfa-announces-conforming-loan-limit-values-for-2026>
 55. Truth in Lending (Regulation Z). Exempt transactions. 12 CFR §1026.3(a) and Supplement I, comment

- 3(a)-4. <https://www.consumerfinance.gov/rules-policy/regulations/1026/interp-3/>
56. Angel Oak Mortgage Solutions. Investor Cash Flow (DSCR) Mortgage Program. Accessed October 7, 2026. <https://angeloakms.com/program/investor-cash-flow-mortgage-program/>
57. Home Mortgage Disclosure (Regulation C). Official Interpretation, comment 4(a)(8)(i)-13. 12 CFR part 1003, Supplement I. <https://www.consumerfinance.gov/rules-policy/regulations/1003/interp-4/>
58. Zheng L, Chiang WL, Sheng Y, et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In: NeurIPS Datasets and Benchmarks Track; 2023. arXiv:2306.05685
59. Anthropic. Model deprecations. Claude Platform documentation. Accessed October 7, 2026. <https://platform.claude.com/docs/en/about-claude/model-deprecations>
60. He H, Thinking Machines Lab. Defeating nondeterminism in LLM inference. Thinking Machines Lab: Connectionism. September 10, 2025. doi:10.64434/tml.20250910
61. Atil B, Aykent S, Chittams A, et al. Non-determinism of "deterministic" LLM settings. arXiv. 2024;2408.04667
62. Efron B, Tibshirani RJ. An Introduction to the Bootstrap. Chapman & Hall; 1993.
63. Wilson EB. Probable inference, the law of succession, and statistical inference. J Am Stat Assoc. 1927;22(158):209-212.
64. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika. 1947;12(2):153-157.
65. Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat. 1979;6(2):65-70.
66. Tabassi E. Artificial Intelligence Risk Management Framework (AI RMF 1.0). National Institute of Standards and Technology; 2023. NIST AI 100-1. doi:10.6028/NIST.AI.100-1
67. National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI 600-1. 2024. doi:10.6028/NIST.AI.600-1
68. Regulation (EU) 2024/1689 of the European Parliament and of the Council (Artificial Intelligence Act), Annex III, point 5(b). Official Journal of the European Union. July 12, 2024. <http://data.europa.eu/eli/reg/2024/1689/oj>
69. Regulation (EU) 2026/1744 of the European Parliament and of the Council (Digital Omnibus on AI). Official Journal of the European Union. July 24, 2026. <https://eur-lex.europa.eu/eli/reg/2026/1744/oj/eng>
70. Colorado General Assembly. SB26-189, Automated Decision-Making Technology. Signed May 14, 2026. <https://leg.colorado.gov/bills/sb26-189>

Appendix A: Additional Results

Table A1. Mean component scores for the October reruns (0 to 1; 450 attempts per model).

Component (weight)	Claude Sonnet 5.5	GPT-5.6 Luna	Gemini 3.8 Flash
Application-field accuracy (25%)	0.923	0.932	0.646
Source-citation correctness (20%)	0.999	0.993	1.000
Risk-flag selection (15%)	0.825	0.797	0.676
Income and conflict reconciliation (10%)	1.000	0.998	0.904
Readiness and conditions (10%)	0.425	0.471	0.297
Final decision with rationale (20%)	0.524	0.522	0.594
Decision accuracy (share of attempts)	43.8%	42.2%	52.0%

Table A2. Decision confusion matrices (rows: gold decision; columns: agent decision). The heuristic ran once on 150 cases; each October model ran three times (450 attempts). The original model runs did not store wrong labels.

Agent	Gold decision	Approve	Suspend	Deny	Total
Document heuristic	Approve	50	0	0	50
	Suspend	50	10	0	60
	Deny	0	3	37	40
Claude Sonnet 5.5	Approve	40	110	0	150
	Suspend	28	152	0	180
	Deny	0	115	5	120
GPT-5.6 Luna	Approve	18	132	0	150
	Suspend	9	171	0	180
	Deny	0	119	1	120
Gemini 3.8 Flash	Approve	117	33	0	150
	Suspend	63	117	0	180
	Deny	0	120	0	120

Table A3. Mean score and pass rate by difficulty level and loan program, October reruns (150 attempts per difficulty level, 90 per program).

Group	Claude Sonnet 5.5	GPT-5.6 Luna	Gemini 3.8 Flash
Beginner	0.60 / 23%	0.55 / 12%	0.72 / 0%
Intermediate	0.79 / 68%	0.84 / 84%	0.63 / 0%
Advanced	0.60 / 23%	0.58 / 18%	0.56 / 0%

FHA	0.63 / 29%	0.62 / 29%	0.62 / 0%
VA	0.81 / 72%	0.77 / 63%	0.70 / 0%
USDA	0.65 / 38%	0.62 / 33%	0.62 / 0%
Jumbo	0.59 / 23%	0.63 / 37%	0.62 / 0%
DSCR	0.63 / 27%	0.63 / 28%	0.62 / 0%

Appendix B: Derivations

Pass condition with a correct decision. The decision component carries a weight of 0.20, so with full decision credit $S = 0.20 + \sum_{k \neq d} w_k s_k$, where the remaining weights sum to 0.80. The condition $S \geq 0.80$ requires $\sum_{k \neq d} w_k s_k \geq 0.60$, which is 75% of the non-decision weight.

Interval width. For 61 passes in 150 cases, the Wilson 95% interval is 33.1% to 48.7%, a half-width of about 7.8 percentage points.