AGENT EVALUATION / MODEL COMPARISON

deepseek/deepseek-v4-pro leads with 1.000

12 agents on finance_ap_bankrec 0.2.12 · task review_ap_and_reconcile_bank · split dev. Every agent ran the same 2 worlds, so the scores compare case for case.

deepseek/deepseek-v4-pro, moonshotai/kimi-k2.5 scored ≥ 0.85 without opening a document: these worlds do not yet require document reading.

BEST SCORE
1.000
deepseek/deepseek-v4-pro

CHEAPEST AT ≥ 0.85
$0.0506
deepseek/deepseek-v4-pro

TOTAL SPEND
$1.83
across all agents

Ranking

#	AGENT	MEAN SCORE	95% INTERVAL	CORRECT VALUE FOUND	EPISODES WITHOUT A SAFETY FAILURE	SAFETY FAILURES	DOCUMENTS OPENED	S
1	**deepseek/deepseek-v4-pro** StreamLake	**1.000**	1.000–1.000	100.0%	100.0%	0	None in 2 attempts	$
2	**moonshotai/kimi-k2.5** SiliconFlow	**0.767**	0.667–0.868	82.7%	50.0%	1	None in 2 attempts	$
3	**z-ai/glm-5.3-flash** BaseTen	**0.639**	0.454–0.824	73.1%	0.0%	6	21, in 1 of 2 attempts	$

#	AGENT	MEAN SCORE	95% INTERVAL	CORRECT VALUE FOUND	EPISODES WITHOUT A SAFETY FAILURE	SAFETY FAILURES	DOCUMENTS OPENED	S
4	**qwen/qwen3.7-plus** Alibaba	**0.631**	0.596–0.667	69.2%	100.0%	0	None in 2 attempts	$
5	**minimax/minimax-m2.7** GMICloud	**0.172**	0.089–0.255	7.7%	50.0%	1	None in 2 attempts	$
6	**stepfun/step-3.5-flash** SiliconFlow	**0.073**	0.073–0.073	0.0%	100.0%	0	None in 1 attempt	$
7	**deepseek/deepseek-v4-flash** Relace	**0.067**	0.000–0.133	13.5%	50.0%	3	71, in 2 of 2 attempts	$
8	**openai/gpt-oss-120b** DeepInfra	**0.000**	0.000–0.000	0.0%	100.0%	0	None in 2 attempts	$
9	**openai/gpt-6-luna** OpenAI	**0.000**	0.000–0.000	0.0%	100.0%	0	None in 2 attempts	$
10	**openai/gpt-5.6-luna** OpenAI	**0.000**	0.000–0.000	0.0%	100.0%	0	None in 2 attempts	$
11	**nvidia/nemotron-3-super-120b-a12b** DekaLLM	**0.000**	0.000–0.000	0.0%	100.0%	0	None in 1 attempt	$
12	**google/gemini-3.1-flash-lite** Google AI Studio	**0.000**	0.000–0.000	0.0%	100.0%	0	None in 1 attempt	$

Correct value found, by problem type

AGENT	BANK_PAYMENT_UNMATCHED	INVOICE_MISSING_PO_LINK	PAYMENT_MISSING_INVOICE_I
deepseek/deepseek-v4-pro	100.0%	100.0%	100.0%
moonshotai/kimi-k2.5	94.4%	100.0%	42.9%
z-ai/glm-5.3-flash	66.7%	80.0%	71.4%
qwen/qwen3.7-plus	94.4%	95.0%	0.0%
minimax/minimax-m2.7	22.2%	0.0%	0.0%
stepfun/step-3.5-flash	0.0%	0.0%	0.0%
deepseek/deepseek-v4-flash	0.0%	35.0%	0.0%
openai/gpt-oss-120b	0.0%	0.0%	0.0%
openai/gpt-6-luna	0.0%	0.0%	0.0%
openai/gpt-5.6-luna	0.0%	0.0%	0.0%
nvidia/nemotron-3-super-120b-a12b	0.0%	0.0%	0.0%
google/gemini-3.1-flash-lite	0.0%	0.0%	0.0%

How to read this

Mean score includes pack weights and penalties; a wrong change costs more than leaving a case alone. "Correct value found" counts repairs written or proposed. A safety failure is a change the pack forbids, such as editing a posted payment. "Documents opened" counts reads of invoices, emails and statements; a high score with none means these worlds did not need them. Spend is what each provider reported for this run. Batches: m7-live-deepseek-v4-pro, cmp-deepseek-deepseek-v4-flash, cmp-google-gemini-3.1-flash-lite, cmp-minimax-minimax-m2.7, cmp-moonshotai-kimi-k2.5, cmp-nvidia-nemotron-3-super-120b-a12b, cmp-openai-gpt-5.6-luna, cmp-openai-gpt-6-luna, cmp-openai-gpt-oss-120b, cmp-qwen-qwen3.7-plus, cmp-stepfun-step-3.5-flash, cmp-z-ai-glm-5.3-flash.
