



ScalePoint

ScalePoint Research · White paper · October 2026

AI economics in fintech: reading card statements

34 models tested on UK acquiring statements

Introduction

Why test models on payments and fintech tasks

There are already hundreds of AI models to choose from, and new ones arrive every few weeks. Choosing one is a question of AI economics. The model needs enough capability for the task, at a price that works at the business's volumes, and it has to clear the business's other rules, such as where the model is hosted, where the data goes and who the supplier is.

Public benchmarks test AI models in general, but few look at the AI economics of fintech tasks. This paper takes one payments task, reading UK card acquiring statements, and measures accuracy in money against price. It is a basic task, and I chose it as a first test because every answer can be checked to the penny. The full method and model-by-model results are online at scalepointpartners.com/research/ai-economics-card-statements.

A handwritten signature in dark blue ink, appearing to read 'MKH', with a stylized, flowing script.

Martin Koderisch
ScalePoint Partners

What I tested

I gave 34 models from the US, Europe and China one task: read UK card acquiring statements, find every fee and spot any errors. I went in with one view to test and one open question. The view, commonly held, is that the leading US frontier models are clearly the best. The question was whether better instructions would let a cheaper model match the most expensive one.

How the test worked

- Synthetic statements from two invented acquirers with different layouts, with traps and errors ranging from easy to hard, such as a fee shown only in a footnote. I knew the correct amount and type of every fee.
- Each model got the text of the statement, with OCR used on the scanned pages. The models did not see the page itself, so the test measured how well they understood and classified the contents.
- Round 1: all 34 models, 16 statements, a plain prompt with nine rules and the names of 18 fee types.
- Round 2: seven finalists, 22 statements, three prompt levels, up to an expert prompt with a 1,500-word briefing on UK statements.
- Every answer was scored in pounds against the correct answer for each fee. Prices are OpenRouter prices from September and October 2026.

What I found

Same accuracy, very different price

Cost per statement with the expert prompt, seven finalists, and the model cost for a year at 10,000 statements a month

MODEL	COST PER STATEMENT		COST A YEAR	FEES WRONG
GPT-5.5 OpenAI		\$0.246	\$29,560	0.03%
Claude Opus 5.5 Anthropic		\$0.187	\$22,460	0.03%
Grok 4.7 xAI		\$0.049	\$5,870	0.03%
Gemini 3.8 Flash Google		\$0.029	\$3,440	0.06%
Mistral Large 3 Mistral AI		\$0.013	\$1,570	0.04%
GLM 5.3 Zhipu AI		\$0.008	\$940	0.03%
DeepSeek V4.1 Flash DeepSeek		\$0.005	\$590	0.07%

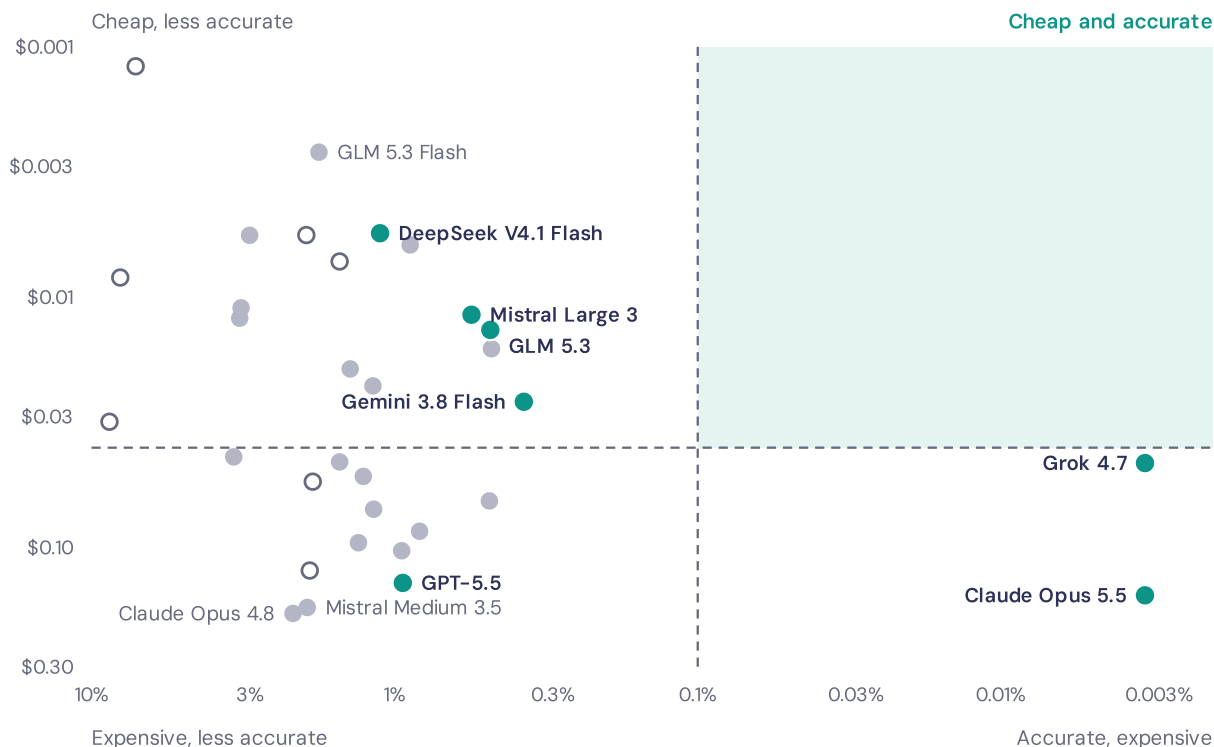
All seven got 0.07% of fees wrong or less. The cheapest cost about a fiftieth of the most expensive. Prices are OpenRouter prices, September and October 2026. Gemini 3.8 Flash rises to about \$6,890 a year when its introductory price ends on 31 December 2026.

With the plain prompt, two models stood out. Claude Opus 5.5 and Grok 4.7 got £4 of fees wrong across all 16 statements. Gemini 3.8 Flash, GLM 5.3, Kimi K3 and Mistral Large 3 got 0.4% to 0.5% wrong, about £40 on a typical statement with £8,000 of fees. 25 of the 34 models cleared the usable bar. Every usable model found the fee lines and read the amounts correctly. The mistakes came from putting fees under the wrong type, mostly surcharges on non-UK and commercial cards filed as penalties.

Fees wrong against cost per statement, 34 models, plain prompt

● Finalist ● Usable ○ Not usable

Cost per statement down the side, log scale, cheaper at the top. Fees wrong across, log scale, fewer to the right.



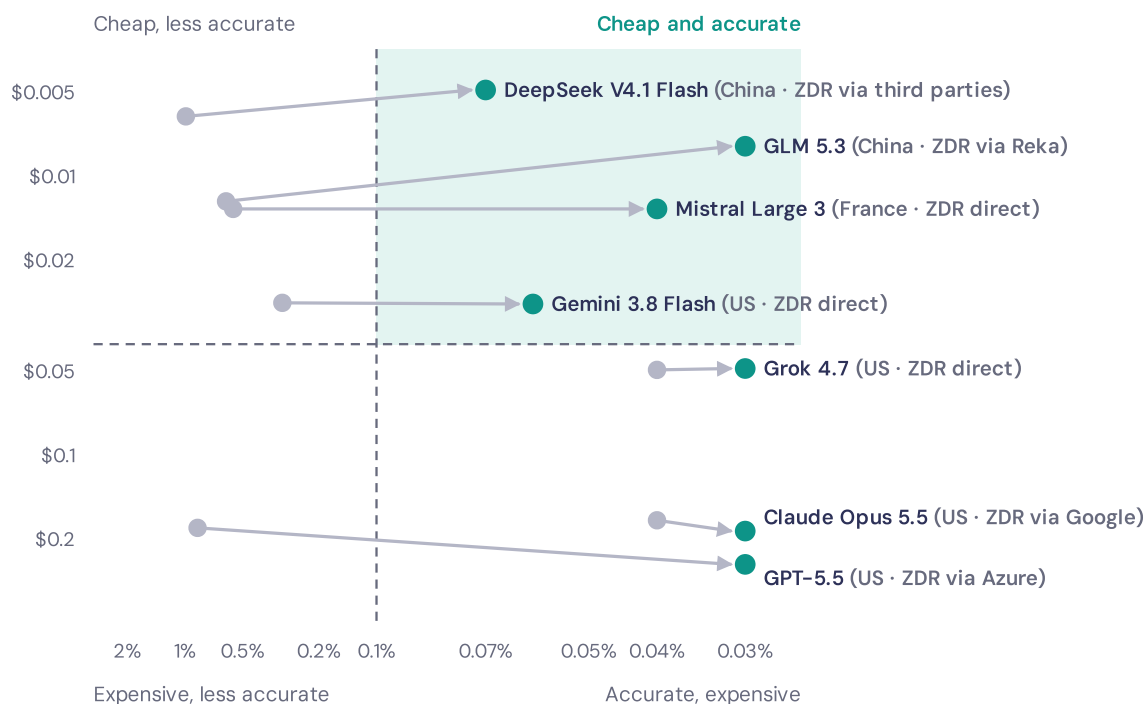
Higher and further right is better. 16 statements, one run each. Off the chart above 10%: Command A Plus (63%), Nova Pro (143%).

With the expert prompt, all seven finalists were about equally accurate. All seven got less than 0.1% of fees wrong, from DeepSeek V4.1 Flash at \$0.005 a statement to GPT-5.5 at \$0.25. Hence once the prompt was right, the choice came down to price and the other points that decide it: reliability, speed, hosting and where the data goes.

With the expert prompt, all seven finalists got 0.07% of fees wrong or less

● Plain prompt ● Expert prompt

Cost per statement down the side, log scale, cheaper at the top. Fees wrong across, log scale, fewer to the right.



Seven finalists, round 2. Each arrow runs from the plain prompt to the expert prompt. The fees-wrong scale is stretched to the right of the 0.1% line.

GPT-5.5 was the surprise. OpenAI's flagship got 0.94% of fees wrong with the plain prompt, about £75 on a typical statement, behind cheaper models from China and France. It caught up with the expert prompt but still missed the faulty subtotal, and it was the most expensive finalist. Hence the common view held for Claude Opus 5.5 and Grok 4.7 and did not hold for GPT-5.5.

The higher price bought judgement. Claude Opus 5.5 and Grok 4.7 flagged both faulty statements at every prompt level and returned no failed answers. The cheaper finalists needed the expert prompt to flag them.

The price gap only matters at volume. The most expensive finalist cost 50 times as much as the cheapest for the same accuracy. At 10,000 statements a month, the model alone costs between about \$600 and \$29,500 a year depending on the choice. The test that showed which models were good enough cost about \$62.

What this means for your AI strategy

The model market is moving fast. During this test, Google released Gemini 3.8 Flash three weeks after 3.7 Flash, Claude Opus moved from 4.8 to 5.5 in four months, and the hosts for two finalists halved their prices within a day. Hence I expect more companies to run several models side by side over the next 12 months, each matched to a task and changed as better or cheaper ones arrive.

Open-weight models are already part of that shift. A September 2026 survey by Enterprise Technology Research found they handle 34% of enterprise AI token usage, up from 23% a year earlier¹. The Financial Times has reported Tinder moving some work to open-weight models to control cost, with PNC, CH Robinson and Siemens also using them². That points to a tiered set of models: open-weight models for high-volume routine tasks, and the leading closed models for the hard judgement calls.

This test shows why that matters. Reading and classifying the fees was within reach of a model at half a cent a statement. The higher price paid off only on the hardest calls. Hence the useful question for a business is which of its tasks need an advanced model, and which are paying for capability they don't use.

In 12 months I expect good practice to include a test set with known answers for each important task, built from the business's own data and scored in money. It also includes prompts the business owns, so it can move between models, a locked model version in production that is retested when a new version or price arrives, and data and supplier rules set before testing starts.

Questions for management

- 1 Which model does each of our AI use cases run on, and when did we last test that choice against the alternatives?
- 2 Do we measure accuracy in money, on a test set built from our own data with known answers?
- 3 How much of our own knowledge is written into our prompts, and who owns and maintains them?
- 4 Where does each model run, under what retention terms, and which companies in the chain, including any router, see our data?
- 5 Which of our tasks need an advanced model, and what would a cheaper model that passes the same test save us at our volumes?

A one-page scorecard for these questions is at scalepointpartners.com/research/ai-economics-card-statements-scorecard.pdf.

Questions this test leaves open

- **Page images.** Would sending the page image to models that can read it beat OCR on scanned statements, and at what cost?
- **A harder task.** Where does a difference in capability show on judgement calls that need reasoning across a whole document?
- **A whole workflow run by an agent.** Do accuracy and cost hold up when a model takes several steps on its own, such as fetching statements, checking them and writing up the results?
- **A fine-tuned open-weight model.** Could a small model trained on statements match the expert-prompt results without the briefing?

Happy to talk through the method with anyone planning a test like this on their own tasks or workflows. Let me know at martin@scalepointpartners.com.

Full method and results

The detailed method, every model's results, the technical annex and all 30 sources are online at scalepointpartners.com/research/ai-economics-card-statements.

Sources

1. Techstrong.ai, "Open-Weight AI Models Expected to Capture 41% of Enterprise Token Usage", reporting Enterprise Technology Research's September 2026 survey, 22 September 2026 (techstrong.ai)
2. Financial Times, report on US companies moving work to open-weight models, 27 September 2026 (ft.com)

This paper reflects the author's own views. The test statements are synthetic, and prices are OpenRouter prices in September and October 2026; results on real documents and at other prices will differ. Model names and trademarks belong to their owners.

About

About the author

Martin Koderisch has spent 20 years in payments and fintech. He advises companies across the industry, from PE-backed businesses to start-ups, and runs custom, independent research into AI in payments and finance operations at ScalePoint Research, focused on build versus buy decisions.

Before ScalePoint he spent ten years in operating roles at Citi, Mastercard and the start-up Luup, then nearly a decade in advisory at Edgar, Dunn & Company, latterly as a Principal, on more than 75 engagements for banks, fintechs, card networks, private equity firms and merchants. He set up ScalePoint in 2024.

scalepointpartners.com





Strategic Finance for Payments and Fintech Leaders

martin@scalepointpartners.com

scalepointpartners.com

London - Barcelona