



ScalePoint

ScalePoint Labs · 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 as developers compete on capability and price. Any business building on AI should expect many more options within the next year.

Choosing a model is a question of AI economics. The model needs enough capability for the task in hand, at a price that works at the business's volumes, and it has to clear the business's other requirements, such as where the model is hosted, where the data goes and who the supplier is. At scale, paying for an advanced model on a basic task has a material effect on cost.

Public benchmarks and leaderboards test AI models in general, but few of them look at the AI economics of fintech tasks. This paper takes one payments task, reading card acquiring statements, measures accuracy in money, and weighs it against price and the other points that decide the choice, such as reliability, hosting and where the data goes.

This paper sets out the findings. The model-by-model results and the full test details are in a technical annex at the end of the online version of this paper, at scalepointpartners.com/labs/ai-economics-card-statements.



Martin Koderisch
ScalePoint Partners

Contents

In this paper

Executive summary	1
Part 1 · The test	
1. Objective and hypothesis	6
2. Method	7
Part 2 · The results	
3. Round 1: Which model is best on its own?	11
4. Round 2: Can a cheaper model get the fees right?	12
5. Round 2: Did that hold on the hardest tasks?	14
6. Round 2: What does it cost?	15
7. Round 2: What still separates the finalists?	17
Part 3 · Conclusions	
8. Conclusions	20
9. What this means for your AI strategy	22
10. Questions this test leaves open	24
Sources	24
About the author	26
Technical annex, A1 to A7	Online

scalepointpartners.com/labs/ai-economics-card-statements

Executive summary

With good instructions, a model at half a cent matched one at 25 cents

Objective

The test set out to measure the AI economics of one fintech task: how accurately each model understands the contents of a document, and what that accuracy costs. I gave 34 models one task: read UK card acquiring statements, find every fee and spot any errors. The test weighed accuracy against price and the other points that decide the choice, such as reliability, hosting, developer location, where the data goes and what it might be used for.

Hypothesis

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 how much difference better instructions would make, and whether they would let a cheaper model match the most expensive frontier model.

Test method

- I built synthetic statements from two invented acquirers with different layouts, since real layouts differ from one acquirer to the next, and filled them with traps and errors, ranging from easy to hard, such as a fee shown only in a footnote and pricing labelled pass-through that is really one flat rate. I knew the correct amount and type of every fee, so I could measure the cost of each model's mistakes.
- Each model received the text of the statement, taken from the PDF, with OCR used to turn the scanned pages into text. The models did not see the page itself, so this test measured how well they understood and classified a statement's contents. Reading the page image directly is a different skill, which this test did not cover (Section 10).
- I ran every model through OpenRouter, which connects to more than 400 models through one API, and picked 34 of them from the main developers in the US, Europe and China.
- In the first round, all 34 models read 16 statements with a plain prompt: nine basic rules and the names of the 18 fee types, with no definitions and no payments guidance.
- In the second round, seven finalists read 22 statements, adding six harder ones, at three prompt levels: plain, middle (plus a definition of each fee type) and expert (plus a briefing of about 1,500 words on UK statements).
- Every answer was scored against the correct answer for each fee, in pounds sterling. Section 2 sets out the method in brief, and the technical annex at the end of the online version has every model's results and the full test details.

Results

Round 1: all 34 models, plain prompt, 16 statements

1. Which model is best on its own? With the plain prompt, Claude Opus 5.5 and Grok 4.7 got almost no fees wrong, £4 across all 16 statements. Gemini 3.8 Flash came next at 0.37% of fees wrong, then GLM 5.3, Kimi K3 and Mistral Large 3 at about 0.5%, or about £40 on a typical statement with £8,000 of fees. GPT-5.5 got 0.94% wrong, about £75 on the same statement. See Section 3.

2. Which models were not usable, and why? 25 of the 34 models cleared a usable bar set before the test. The nine that failed returned broken answers, invented fee lines, or missed traps. See Section 3 and annex A2.

3. Where did the models go wrong? Every usable model found the fee lines and read the amounts. The mistakes came from putting fees under the wrong type, mostly surcharges on non-UK and commercial cards filed as penalties. See Section 3.

Round 2: seven finalists, three prompt levels, 22 statements

4. Can a cheaper model get the fees right? With the expert prompt, all seven finalists 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. See Section 4.

5. Did that hold on the hardest tasks? On fee accuracy it did. Spotting the two faulty statements was harder. Claude Opus 5.5 and Grok 4.7 flagged both at every prompt level, the cheaper finalists needed the expert prompt, and Mistral Large 3 and GPT-5.5 missed at least one even with it. See Section 5.

6. What does it cost? With the expert prompt, the most expensive finalist cost 50 times as much per statement as the cheapest, for the same accuracy. However, even the most expensive cost only about 25 cents a statement, so the gap only adds up for a business that reads statements in bulk. See Section 6.

7. What still separates the finalists? Reliability, speed, where the data goes and who the supplier is. See Section 7.

Conclusions

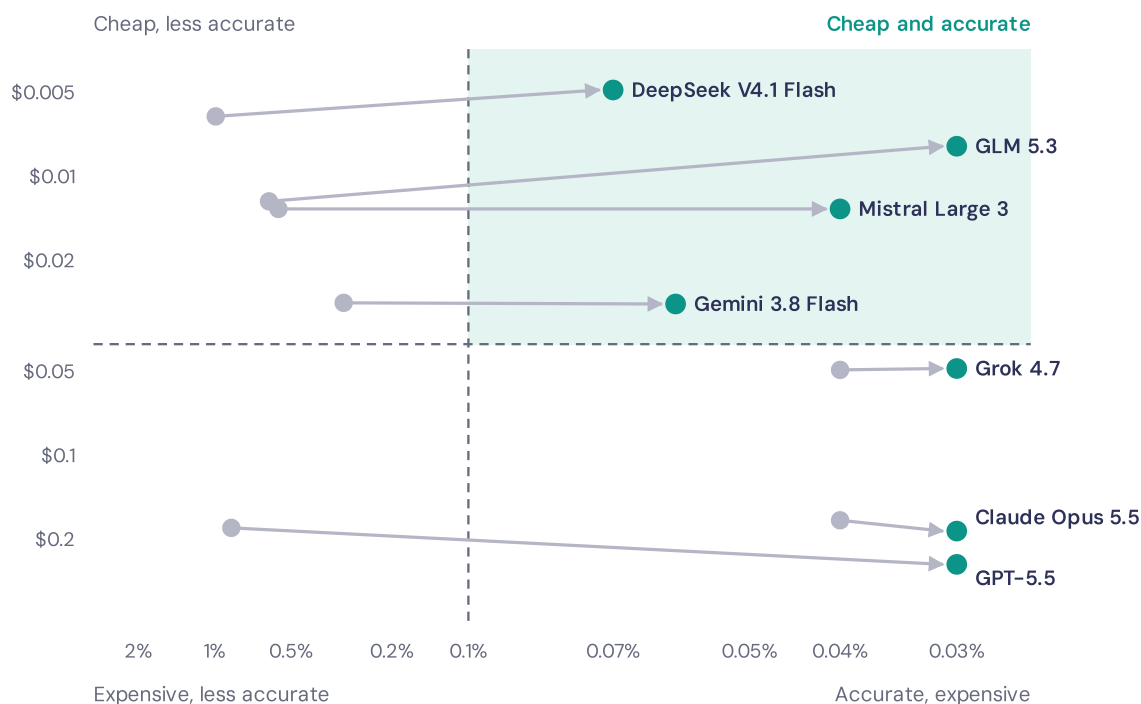
The main finding is that once the prompt was right, all seven finalists were about equally accurate. The choice of model then comes down to price and other points such as reliability, hosting and where the data goes.

The chart below shows that finding for the seven round 2 finalists. Each arrow runs from a model's result with the plain prompt (grey dot) to its result with the expert prompt (coloured dot). Further right means fewer fees wrong, and higher means cheaper per statement. With the expert prompt, all seven ended up to the right of the 0.1% line, so the choice between them comes down to height: DeepSeek V4.1 Flash at half a cent a statement near the top, and GPT-5.5 at 25 cents at the bottom. Claude Opus 5.5 and Grok 4.7 barely move, because they were already accurate with the plain prompt.

Price and accuracy, plain and expert prompt

● 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.

The most surprising result was GPT-5.5, OpenAI's flagship. With the plain prompt it got 0.94% of fees wrong, about £75 on a typical statement, while its frontier peers Claude Opus 5.5 and Grok 4.7 got almost none wrong. It also finished behind cheaper models from China and France, such as GLM 5.3 and Mistral Large 3. It did catch up with the expert prompt, but just like the cheaper models it needed that help, and even then it failed to flag the faulty subtotal. GPT-5.5's showing is all the more surprising because, at \$0.25 a statement, it was the most expensive finalist. Hence the common view that the leading US frontier models are clearly the best held for Opus and Grok, and did not hold for GPT-5.5. What the higher price of Claude Opus 5.5 and Grok 4.7 bought was good results with basic instructions, the hardest judgement calls caught without being told what to look for, and no failed answers.

AI is moving fast, and over the next 12 months we can expect to see more and more companies running several models side by side, each matched to the task in hand. This one test challenged the commonly held view that the leading US frontier models all outperform other models by a long way. It also showed that better instructions can let a cheaper model match the most expensive frontier model. Hence any management team using AI should check how its own model choices hold up, starting with the questions below. Section 9 sets out how I expect the market to change over the next 12 months, what good practice will look like by then, and where routers fit. Section 10 sets out the questions this test leaves open, including a harder task, reading page images directly and a whole workflow run by an AI agent.

Questions for management

The findings point to five questions for any management team using AI models across its business:

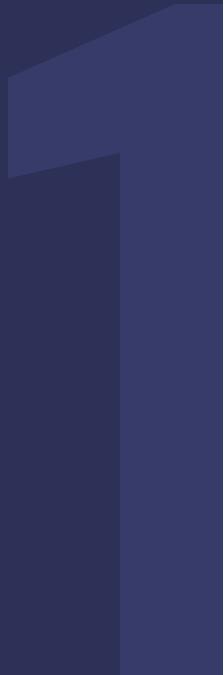
- 1 Which model does each of our AI use cases run on, and when did we last test that choice against the alternatives?

- 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? Section 7 sets out where each finalist ran in this test.

- 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?



Part 1

The test

The objective, the hypothesis and how the statements, prompts and scoring were set up.

Sections 1 and 2

Section 1

Objective and hypothesis

The test set out to measure the AI economics of one fintech task: how accurately each model understands the contents of a document, and what that accuracy costs. Card acquiring statements suit that question. Every fee on a statement is a fixed amount worked out from the acquirer's price list, so a model's figure for it is either right or wrong, to the penny.

I went in with one view to test and one open question:

- **The leading US frontier models are clearly the best.** This is a commonly held view. I tested it by giving every model the same plain prompt, so that each model had to work on its own.
- **How much difference do better instructions make?** I tested this by giving the finalists two richer prompts and comparing their accuracy and cost with the plain prompt.

Alongside accuracy and price, the test recorded the other points that decide the choice, such as reliability, hosting, developer location and where the data goes. Two of these can rule a model out before price comes into it. The first is where the developer is based. Some Chinese developers face sanctions or government bans in the US and Europe, and a regulated firm's own supplier policy may rule them out even when the model runs on a Western host. The second is zero data retention. Under many standard terms, the company running the model keeps a copy of each request for a period, for example to check for abuse. Zero data retention means it keeps no copy once it has replied, which for a regulated firm is the safer starting point before sending real customer documents.

Section 2

Method

The statements and the calculated answers

I used AI to help write a program that built each statement from scratch. It made up a month of card sales for an invented merchant, applied an invented acquirer's price list to work out every fee, and produced the statement as a PDF file, with the scanned ones produced as page images. The same program wrote out the calculated answers: the list of every fee on that statement with its amount, currency and fee type. No real merchant or acquirer data was used, and because the statement and its answers come from the same figures, every fee has a known right answer. The statements come from two invented acquirers, one with an interchange-plus-plus layout modelled on the large UK acquirers and one with simple blended pricing. The typical statement is for a merchant taking about £1 million a month on cards and paying about £8,000 in fees. Every page is marked as a synthetic test document, so the files can't be mistaken for real statements if they are shared. Annex A3 notes the small risk that this marking changed how a model read them.

Tier	Statements	What it contains
Standard	12	Interchange-plus-plus pricing with 30 to 40 fee lines
Traps	4	A fee only in a footnote, fees shown twice, fees taken out of the settlement, charges in two currencies
Hard	6	Three outlets on one statement, scanned pages, a subtotal that doesn't add up, pass-through pricing that is really one flat rate, a currency conversion fee

For every statement the model had to find every fee line, read its amount and currency, put it under the right fee type and flag anything wrong with the statement itself.

The three prompt levels

Level	What it contains
Plain	Nine rules, such as record each fee once, keep currencies separate and flag problems with the statement, plus the names of the 18 fee types
Middle	The plain prompt, plus a one-line definition of each fee type
Expert	The middle prompt, plus a briefing of about 1,500 words on how UK acquirers lay out statements, where fees hide and how to classify the awkward ones, much like the briefing you would give a new analyst

The plain prompt answers "which model is best on its own?" The middle and expert prompts answer the open question: how much difference better instructions make, and whether they let a cheaper model match the most expensive frontier model. A business running statements in production would give the model detailed instructions, so the expert prompt is in my mind the realistic buying question. I wrote the expert briefing myself with an AI drafting tool, without access to the test statements, so it could not be tuned to the answers.

Choosing the 34 models

OpenRouter offers more than 400 models. I picked 34 to cover the current flagship and a cheaper model from each of the main developers in the US, Canada, Europe and China, with both open and closed weights, plus some older versions still in wide use. Testing all 400 would have cost several times my \$50 budget, and many of them are older versions or variants of the same model. Each statement went through OpenRouter to a company hosting the model, such as Microsoft Azure for GPT-5.5, the model behind ChatGPT. The test called every model through its API, the way a business builds a model into its own systems, and didn't use any chat apps. Microsoft Copilot isn't in the test. It is an assistant built into Microsoft's products, and it can't be sent a statement through OpenRouter. Copilot is largely built on OpenAI's GPT models, so GPT-5.5 is the closest guide here, but Microsoft adds its own instructions and settings, so Copilot's results could differ. The same applies to the ChatGPT app compared with GPT-5.5.

The two rounds

- **Round 1.** All 34 models read 16 statements, the Standard and Traps tiers, once each with the plain prompt.
- **Making the cut.** A model first had to clear a usable bar, set before the test: its fee lines had to add up to the true total on at least 95% of the Standard statements, it could invent no fee lines, and it had to catch at least three of the four traps. 25 models cleared it. From those I picked six across the range of prices: the two most accurate at any price (Claude Opus 5.5 and Grok 4.7), three accurate models at the low end (Gemini 3.8 Flash, Mistral Large 3 and GLM 5.3) and one of the very cheapest (DeepSeek V4.1 Flash). The choice within each price band was my judgement. I added GPT-5.5 afterwards as a seventh finalist, because it is the best-known OpenAI model and readers would look for it.
- **Round 2.** The seven finalists read 22 statements, the same 16 plus the six Hard ones, at all three prompt levels. At the plain prompt the first-round answers to the 16 repeated statements were reused.

Sending and scoring

Every call went through OpenRouter¹ with three settings: the model received only the text of the statement, never the page itself, with the text extracted from the PDF once and the scanned pages turned into text by Tesseract, a standard optical character recognition (OCR) tool, so every model got exactly the same input; OpenRouter only used hosts that keep no copy of the data (zero data retention); and temperature was set to zero, so the same statement gets the same answer as far as the model allows.

The headline measure is fees wrong. Take a statement with £5,000 of fees. If a model misses a £45 fee and counts a £60 fee twice, it has £105 of fees wrong, or 2.1%. A fee put under the wrong type or read at the wrong amount counts the same way. Fees wrong is the total value of those mistakes as a share of the true fees on the statement, averaged across the statements.

I also ran a template parser without AI, the way statement reading was automated before AI models, as a baseline. The whole test cost about \$62 in model fees.

The technical annex, at the end of the online version of this paper, has the fee types, the nine traps, the scoring calls, the calls and costs, and the test's limitations (annex A3).

A note on terms. "AI model" here means a large language model, the kind of model behind chat assistants. A model has "open weights" when anyone can download it and run it on their own servers. Models are priced per "token", roughly three-quarters of a word. Interchange-plus-plus (IC++) pricing shows the card issuer's interchange, the card scheme's fees and the acquirer's own charge separately. Blended pricing rolls them into one rate. "Fine-tuning" means training a model further on a business's own examples, which changes the model itself. A model with open weights can be fine-tuned by anyone who downloads it. A closed model can be fine-tuned only where its developer offers it, on the developer's terms.

Section 3 covers round 1, where all 34 models read 16 statements with the plain prompt. Sections 4 to 7 cover round 2, where the seven finalists read 22 statements at three prompt levels.



Part 2

The results

Each question from the executive summary, answered in turn.

Sections 3 to 7

Section 3

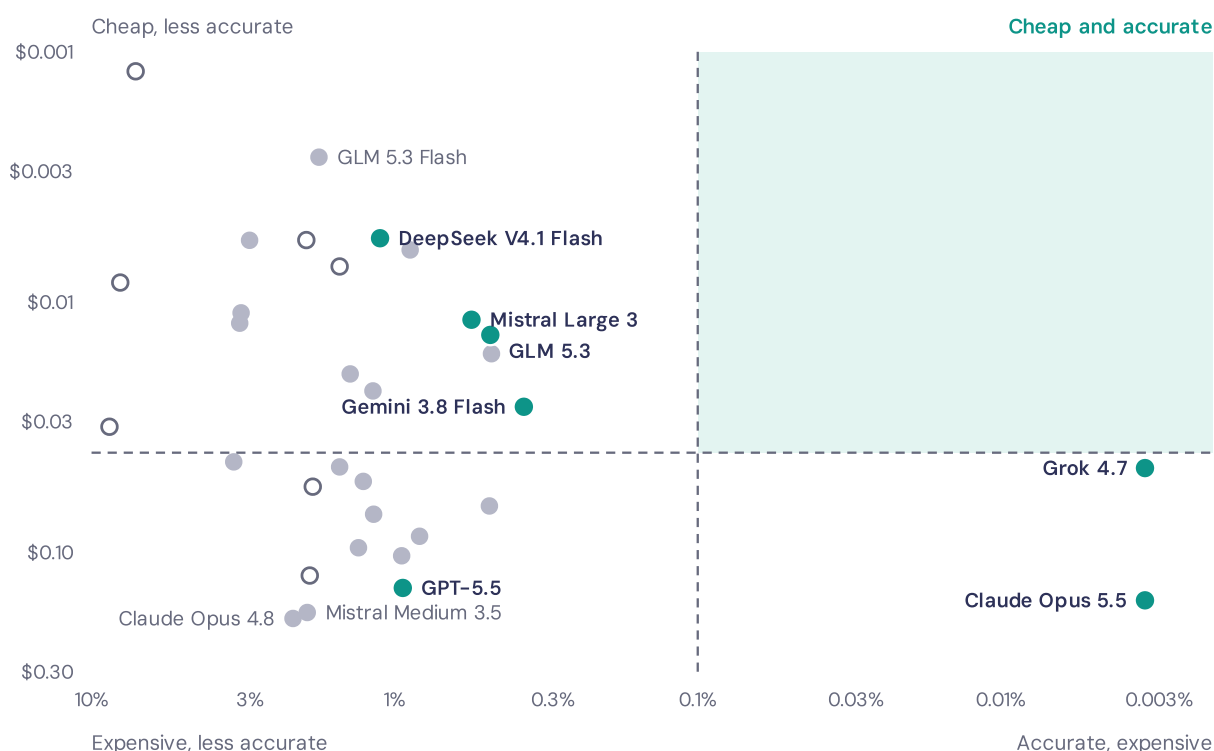
Round 1: Which model is best on its own?

With the plain prompt in the first round, Claude Opus 5.5 and Grok 4.7 got almost no fees wrong, £4 across the 16 statements. Gemini 3.8 Flash was third at 0.37%.

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 Standard and Traps statements, one run each. Every model is listed in annex A1. Off the chart above 10%: Command A Plus (63%), Nova Pro (143%).

- After Opus, Grok and Gemini came NVIDIA's Nemotron 3 Ultra and GLM 5.3 at 0.48%, Kimi K3 at 0.49% and Mistral Large 3 at 0.56%. On a typical statement with £8,000 of fees, that is between £38 and £45 put in the wrong place.
- Other US models did worse than the leaders from China and France, such as GPT-5.5 at 0.94% (about £75 on a typical statement) and the previous Claude Opus 4.8 at 2.16% (about £173).
- Price was a poor guide to accuracy. Mistral Medium 3.5 cost \$0.17 a statement and got 1.94% wrong, while Mistral Large 3 from the same company cost \$0.012 and got 0.56% wrong.
- Nine of the 34 models failed the usable bar, mostly older or smaller models from developers whose newer models did clear it. They returned broken answers, invented fee lines or missed traps. Annex A2 lists them.

Where the models went wrong. Every usable model found at least 99% of the fee lines and read almost every amount correctly. The mistakes came from putting a fee under the wrong type, which is a question of payments knowledge. For example, GPT-5.5 put £979 in the wrong place across the 16 statements, which carried about £128,000 of fees. £676 of that was non-UK card surcharges filed as penalties, and £300 was two fees taken out of the settlement recorded as credits. A payments analyst learns these calls on the job, so I wrote them into the expert prompt.

The baseline without AI. The template parser got every fee right on the 12 Standard statements, at no cost. It missed two of the four traps, and on the six Hard statements it got 34% of fees wrong, mostly on the scans. It was written knowing both layouts, so this is its best case. An acquirer it has not seen would need a new template.

Section 4

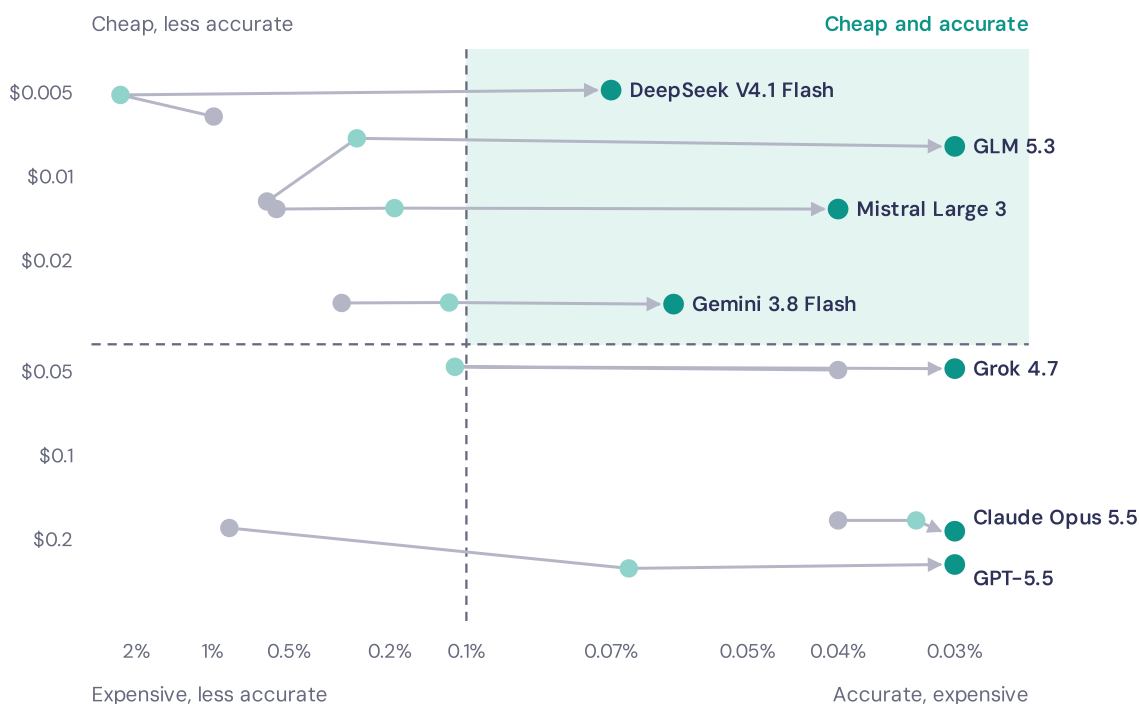
Round 2: Can a cheaper model get the fees right?

With the expert prompt, all seven finalists got less than 0.1% of fees wrong. The cheaper models needed the expert prompt to get there, and so did GPT-5.5. Opus and Grok did almost as well without it.

Fees wrong at each prompt level, seven finalists

● Plain prompt ● Middle prompt ● Expert prompt

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



22 Standard, Traps and Hard statements. Cost is per statement with the expert prompt, from actual usage.

Model	Plain	Middle	Expert	Cost per statement, expert
Claude Opus 5.5	0.04%	0.03%	0.03%	\$0.187
Grok 4.7	0.04%	0.11%	0.03%	\$0.049
Gemini 3.8 Flash	0.31%	0.12%	0.06%	\$0.029
Mistral Large 3	0.56%	0.19%	0.04%	\$0.013
GLM 5.3	0.61%	0.27%	0.03%	\$0.008
GPT-5.5	0.86%	0.07%	0.03%	\$0.246
DeepSeek V4.1 Flash	0.99%	2.31%	0.07%	\$0.005

- With the plain prompt, the order from round 1 held, and the four cheaper finalists got between 8 and 25 times as many fees wrong as Opus.
- The middle prompt cut that by more than half for Gemini 3.8 Flash, Mistral Large 3 and GLM 5.3, and took GPT-5.5 from 0.86% to 0.07%. The expert prompt brought all seven under 0.1%.
- DeepSeek V4.1 Flash did worse with the middle prompt because of one answer: on one Hard statement it kept writing its working-out, hit its length limit and returned nothing, which counts as every fee missed.

What the prompt levels show:

- **The prompt carried the payments knowledge.** I wrote into the briefing, for example, that a surcharge on commercial cards is charged because of the card, which puts it in a different category from a penalty for keyed or late transactions. With that written down, the other models classified the awkward lines as well as Opus.
- **The briefing belongs to the business.** It works with any of the seven models, so it stays with the business if it changes model or provider.
- **On scans, the OCR set the limit.** On one scanned statement the OCR read £55.17 as £56.17 and an £86.99 credit as £26.99. Every model reported what the text said, so none could get those two lines right. That accounts for the 0.03% that no model beat.

Section 5

Round 2: Did that hold on the hardest tasks?

On fee accuracy, the harder statements made little difference. With the expert prompt, all seven finalists got between 0.12% and 0.17% of fees wrong on the six Hard statements, and most of that was the two misread scanned figures.

The hardest task was spotting the two faulty statements among the Hard ones: a subtotal that doesn't add up, and a statement that claims to pass interchange through at cost but charges one flat rate. To catch either, the model has to check figures across many rows.

- Claude Opus 5.5 and Grok 4.7 flagged both at every prompt level.
- Gemini 3.8 Flash, GLM 5.3 and DeepSeek V4.1 Flash mostly missed the fake pass-through with the middle prompt, and flagged both every time with the expert prompt.
- Mistral Large 3 spotted the faulty subtotal, with the exact figures, but filed it under "other" in place of the label the prompt asked for. The score counts only the right label, because any automated routing to a person works from the label.
- GPT-5.5 never flagged the faulty subtotal, at any level, and caught the fake pass-through pricing only with the expert prompt.

Annex A4 has the run-by-run results.

An example: fake pass-through pricing

One of the Hard statements is for an invented pub group, Juniper Lane Taverns. The front page says "Pricing: Interchange Plus Plus", and the charges summary lists "Interchange (passed through at cost)" of £1,492.57. These are some of the rows from the detail page:

Card type on the statement	Sales (£)	Interchange rate shown	What interchange should look like
Visa consumer debit, UK	80,748.52	0.7100%	Capped at 0.2%
Mastercard consumer credit, UK	29,746.00	0.7100%	Capped at 0.3%
Visa commercial credit, UK	6,490.75	0.7099%	Not capped, so usually higher than consumer cards
Visa consumer credit, international	3,153.36	0.7100%	Usually higher than UK cards

Every row shows interchange at about 0.71%, whatever the card. Real interchange changes with the card type, so the same rate on every row means the acquirer is charging one blended rate and calling it pass-through. On the UK consumer debit row alone, interchange at the 0.2% cap would have been about £161, against the £573 shown. On this invented statement, the merchant is paying a blended rate while being told its costs are passed through at cost.

The trap is deliberately blatant. It tests whether a model checks the figures or accepts what the statement says about itself.

Section 6

Round 2: What does it cost?

With the expert prompt, reading a statement cost between \$0.005 with DeepSeek V4.1 Flash and \$0.246 with GPT-5.5, for the same accuracy. Gemini 3.8 Flash came out as the best value at \$0.029.

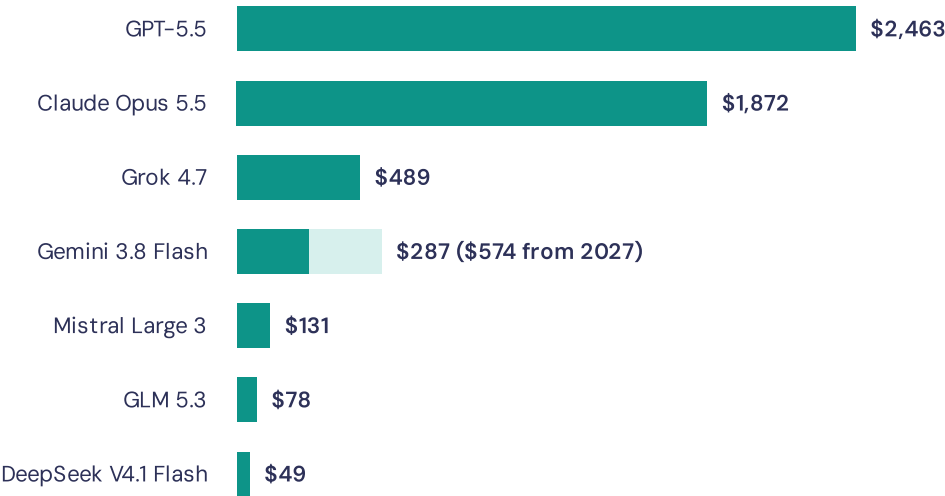
Model	Fees wrong, expert	Cost per statement	Cost compared with the cheapest
DeepSeek V4.1 Flash	0.07%	\$0.005	1×
GLM 5.3	0.03%	\$0.008	1.6×
Mistral Large 3	0.04%	\$0.013	2.7×
Gemini 3.8 Flash	0.06%	\$0.029	6×
Grok 4.7	0.03%	\$0.049	10×
Claude Opus 5.5	0.03%	\$0.187	38×
GPT-5.5	0.03%	\$0.246	50×

The difference comes from two things: first, each model's price per token, and second, how many tokens it uses when writing its answer. Some models "think" before they answer. They write out working notes first, which the user doesn't normally see but still pays for, so they use more tokens. For example, with the expert prompt DeepSeek V4.1 Flash wrote about 6,200 tokens of working-out per statement and Claude Opus 5.5 about 500. DeepSeek was still the cheapest, because its price per token is so low. Annex A5 discusses this in more detail and sets out what drove each model's cost.

Cost at volume

For a merchant reading its own statement once a month, the model's price hardly matters: even GPT-5.5, the most expensive finalist, would cost about \$3 a year. The cost only adds up for a business that reads statements in bulk, such as an acquirer. At 10,000 statements a month, GPT-5.5 costs about \$2,460 a month, Claude Opus 5.5 about \$1,870 and Gemini 3.8 Flash about \$290.

Model cost per month at 10,000 statements, expert prompt



Actual cost per statement in this test, September and October 2026 prices. Pale bar: Gemini 3.8 Flash after its introductory price ends on 31 December 2026.

Prices and versions change quickly

Gemini 3.8 Flash, which Google launched on 2 September 2026, is on an introductory price until 31 December 2026, after which its price is set to double². It was also released three weeks after Gemini 3.7 Flash, and Claude Opus moved from 4.8 to 5.5 in four months. The hosts for GLM 5.3 and DeepSeek V4.1 Flash even halved their prices between the morning and the afternoon of one test day. Hence best practice is to lock in the exact model version used in production, and to rerun the test whenever a new version or a price change comes along.

Section 7

Round 2: What still separates the finalists?

Once accuracy was level, the finalists still differed on reliability, on spotting faulty statements, on speed and on where the data goes. For a regulated business, where the data goes can rule a model out before anyone compares prices.

The seven finalists compared, expert prompt

Model	Fees wrong, expert	Cost per statement	Flags caught (of 9)	Median time	Failed answers	Host in this test
Claude Opus 5.5	0.03%	\$0.187	9	46 s	None	Google Vertex
Grok 4.7	0.03%	\$0.049	9	73 s	None	xAI
Gemini 3.8 Flash	0.06%	\$0.029	9	32 s	4 cut off by the host, rerun	Google
Mistral Large 3	0.04%	\$0.013	7	56 s	None	Mistral, EU
GLM 5.3	0.03%	\$0.008	9	79 s	None	Reka
GPT-5.5	0.03%	\$0.246	8	23 s	None	Microsoft Azure
DeepSeek V4.1 Flash	0.07%	\$0.005	9	171 s	1 runaway answer, 1 cut off	Wafer, InferenceNet

Notes: 22 statements, expert prompt. Flags are the nine traps across the Traps and Hard tiers, counted only when the model used the right label. Failed answers are counted across all three prompt levels.

Reliability. Five of the seven returned a clean answer every time. The host cut off four of Gemini 3.8 Flash's 110 answers and one of DeepSeek V4.1 Flash's. That means the company running the model stopped sending the answer part-way through, for example because its servers were busy or a time limit was reached, so the answer arrived incomplete. Each was fine when sent again. DeepSeek also had one answer that never finished, because it kept writing its working-out until it reached its length limit. Around 100 answers each is too few to prove one service more reliable than another, so a business should check reliability over a longer trial, and anyone using these models in production needs automatic retries and a person checking the log.

Speed. Median time per statement ranged from 23 seconds for GPT-5.5 to 171 seconds for DeepSeek. An overnight batch can absorb that, but a customer or an analyst waiting on the answer would notice the slower models.

Where the data goes. Zero data retention means the company running the model keeps no copy of the document or the answer once it has replied³. With it required, OpenRouter would not send requests to Anthropic's, OpenAI's or DeepSeek's own servers, so Claude ran through Google Vertex, GPT through Microsoft Azure and DeepSeek on third-party hosts. Two Alibaba models could not run with zero retention at all. The open-weight finalists ran on hosts most compliance teams will not have reviewed, GLM 5.3 on Reka and DeepSeek on Wafer and InferenceNet. Of the 34 models, 15 have open weights, including three of the finalists: Mistral Large 3, GLM 5.3 and DeepSeek V4.1 Flash. The alternative for any of them is to run the weights on the business's own servers, so its documents never leave the business.

Who the supplier is. Two finalists come from Chinese developers. Zhipu AI, which makes GLM 5.3, has been on the US Entity List since January 2025⁴. DeepSeek's app was blocked in Italy in January 2025, Berlin's data protection authority asked Apple and Google to remove it in June 2025, and the US has ordered it off Defense Department systems^{5,6,7}. The UK has not banned it⁸. The router is a supplier too. Using OpenRouter, now part of Stripe, adds one more company to the data path, which under DORA in the EU and the PRA's outsourcing rules in the UK needs the same third-party review as any other provider^{9,10}.



Part 3

Conclusions

The main findings, what they mean for AI strategy, and next steps for the research.

Sections 8 to 10

Section 8

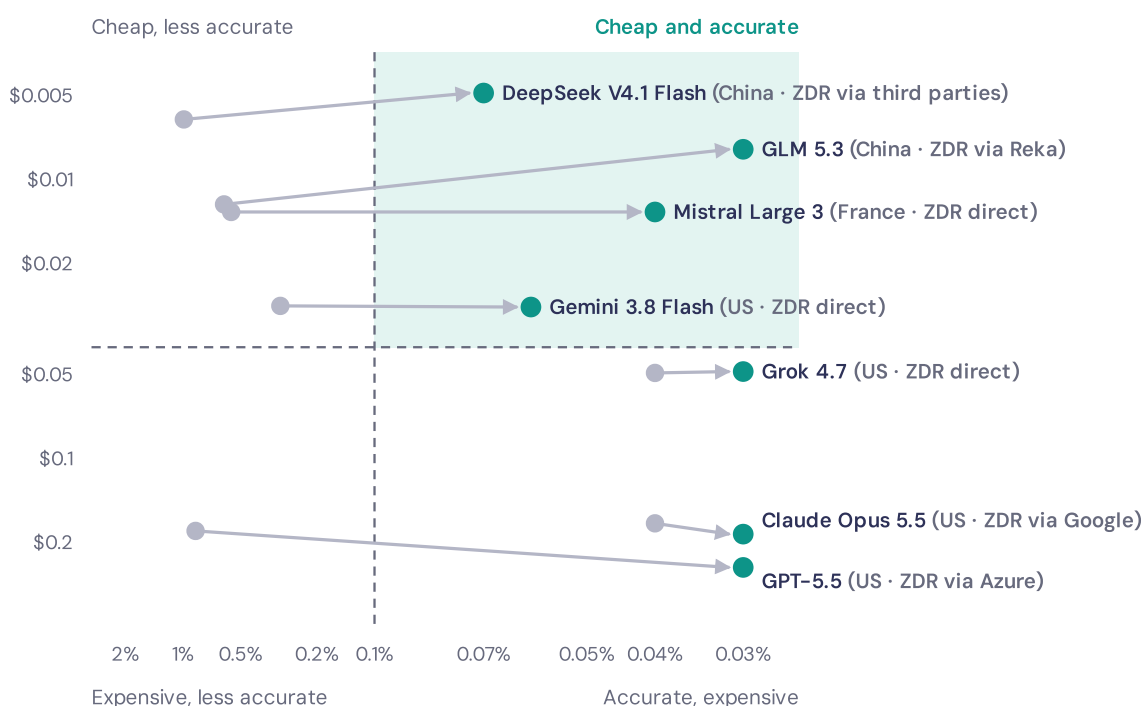
Conclusions

The main finding is that once the prompt was right, all seven finalists were about equally accurate. The choice of model then comes down to price and other points such as reliability, hosting and where the data goes. The price gap is at most about 25 cents a statement, so in money terms the gap only adds up to a material amount for a business that reads documents in bulk.

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. All runs used zero data retention (ZDR). Direct means the developer's own service qualified.

Summary of findings

- **The common view held for Opus and Grok, and not for GPT-5.5.** In round 1, with the plain prompt, Claude Opus 5.5 and Grok 4.7 were the most accurate. GPT-5.5, OpenAI's flagship, got 0.94% of fees wrong, behind cheaper models from China and France, and in round 2 it needed the expert prompt as much as the cheaper models did.
- **The models found the fees, and their mistakes were in classifying them.** In round 1, every usable model found at least 99% of the fee lines. The errors came from lines that need payments knowledge to file correctly.
- **A good prompt closed the gap.** In round 2, the expert prompt took the five other finalists from between 0.3% and 1.0% of fees wrong to under 0.1%. That prompt is the business's own knowledge, written down, and it works with any of the seven models.

- **Price was a poor guide to accuracy.** In round 1, Mistral Medium 3.5 cost fifteen times as much as Mistral Large 3 and got more fees wrong. In round 2, with the expert prompt, the cheapest finalist matched the most expensive for a fiftieth of the cost.
- **The higher price of Opus and Grok bought judgement and reliability.** They did well with basic instructions in both rounds. In round 2 they caught the faulty statements without being told what to look for, and never failed.
- **Where the data goes can decide the choice.** Zero data retention changed which hosts could be used for Claude, GPT and DeepSeek, and ruled out two Alibaba models. Two finalists come from developers that will come up in sanctions screening.

Discussion

Who the price gap matters for

For a single merchant, the model's price makes little difference. Reading one statement a month costs about \$3 a year with GPT-5.5 and about 6 cents with DeepSeek V4.1 Flash. What matters at that scale is accuracy: about £40 a statement between the best models and the rest in round 1, and about £3 in round 2 with the expert prompt.

The price gap matters for a business that reads statements in bulk, such as an acquirer checking its merchants' statements, a lender reading card statements to underwrite merchant cash advances, or a fintech building this into a product. At 10,000 statements a month, GPT-5.5 costs about \$29,500 a year yet DeepSeek about \$600, and GLM 5.3 both matched GPT-5.5's accuracy for about \$940.

The skills the task needed

The task needed three different skills, and the models were not equally good at each. The first row draws on both rounds, and the other two on round 2.

Skill	What it involves	How the models compared
Reading the statement	Finding every fee line and reading its amount and currency	Every usable model found at least 99% of the fee lines. On scanned pages, the conversion to text set the limit for all of them.
Payments knowledge	Putting each fee under the right type, such as a card surcharge versus a penalty	With a plain prompt, Claude Opus 5.5 and Grok 4.7 were best. With the expert prompt, which supplies the knowledge, all seven finalists were equal.
Reasoning across the statement	Comparing many rows to spot a pattern that should not be there, such as the fake pass-through pricing	Claude Opus 5.5 and Grok 4.7 did it at every prompt level. Gemini 3.8 Flash, GLM 5.3 and DeepSeek V4.1 Flash did it reliably only with the expert prompt. GPT-5.5 and Mistral Large 3 missed at least one.

Hence a good prompt can supply payments knowledge, and the price premium for Opus and Grok mostly buys reasoning without being told what to look for.

Some of the models "think", writing out working notes before they answer. On this task the amount of thinking made no visible difference to accuracy, and it added cost. Annex A6 sets out the figures.

Was the task hard enough?

This task may not have been hard enough to separate the models. With good instructions, reading and classifying the fees on a card statement was within reach of every finalist, and the gaps showed only on the hardest judgement calls. A harder task would show whether this conclusion still holds and where a difference in capability starts to show (Section 10).

Section 9

What this means for your AI strategy

The model market is moving fast. During this test alone, Google released Gemini 3.8 Flash three weeks after 3.7 Flash, on an introductory price that doubles at the end of the year. Claude Opus moved from 4.8 to 5.5 in four months, and the hosts for two finalists halved their prices within a single day. With new models and new prices arriving this often, no single model is likely to stay the best choice for every task for long. Hence over the next 12 months we can expect to see more and more companies running several models side by side, each matched to the task in hand, and changing them as better or cheaper ones arrive. I also expect the routers that connect a business to those models to move on from keeping a service running to choosing models on price and accuracy.

Open-weight models are already part of that shift. A September 2026 survey by Enterprise Technology Research found that they handle 34% of enterprise AI token usage, up from 23% a year earlier, and that 42% of respondents run at least one in production²⁹. The Financial Times has reported Tinder moving some work from frontier models to open-weight ones to control cost, with PNC, CH Robinson and Siemens also talking about using them³⁰. 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. With the expert prompt, reading and classifying the fees was within reach of a model costing half a cent a statement. The higher price of Claude Opus 5.5 and Grok 4.7 paid off on the hardest judgement calls, such as spotting a faulty statement without being told what to look for. 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.

That choice can be set out like any other investment. The costs are the model's price per task, the human checking its errors need, and the test that shows which models are good enough. The benefit is accuracy, measured in money. On this task, at 10,000 statements a month, the model alone costs between about \$600 and \$29,500 a year depending on the choice, and the test that showed which models were good enough cost about \$62.

In 12 months I expect good practice to look like this:

- **A set of models matched to tasks.** The business runs a small set of models, each matched to a task, and keeps the most capable ones for the hard judgement calls.
- **A test set for each important task.** It holds a test set with known answers, built from its own data, and measures accuracy in money.
- **Its own knowledge in its own prompts.** It owns and maintains the prompts, so it can move between models without starting again.

- **Locked versions and regular reruns.** It locks in the exact model version it runs in production, and reruns the test whenever a new version or a price change comes along.
- **Data rules and supplier reviews up front.** It sets its rules on where data can go, and reviews its model hosts and any router as suppliers, before testing starts.

Routers make switching between models practical, and there are several kinds. Model marketplaces such as OpenRouter reach hundreds of models through one connection. AI gateways such as LiteLLM, Portkey and Cloudflare AI Gateway sit in front of a business's own model accounts. The cloud platforms offer their own catalogues of models, through Amazon Bedrock, Google Vertex AI and Microsoft Azure AI Foundry. Today routers are used mainly for resilience, moving to a backup model when one fails, and for one connection and one bill across many models. Where they cut cost, it is mostly by sending a model's traffic to the cheapest host for the same model. Matching each task to the cheapest model that is good enough still needs a test on the business's own data.

The same kind of test works on any fintech task, and the answer can differ from one task to the next.

Questions for management

The findings point to five questions for any management team using AI models across its business:

- 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?

Section 10

Questions this test leaves open

This test leaves four questions open.

- **Page images in place of text.** In this test every model received the statement as text. The scanned pages were turned into text first by Tesseract, a standard open-source OCR tool that works separately from the AI models. I chose text so that all 34 models started from the same input, since not all of them can read images, and because images use more tokens and would have cut the number of models I could test within the budget. Reading a page image is a different skill from understanding its contents, and many current models can now do both. A future test could send the page images to the finalists that can read them, to see whether reading the page directly beats OCR on scanned statements, and what it adds to the cost.
- **A harder task.** More judgement calls that need reasoning across a whole document, to see where a difference in capability shows.
- **A whole workflow, run by an agent.** Test a model working as an agent inside a business's own tools, taking several steps on its own. For example, it could fetch statements from a shared drive, read and check them, and write the results to a spreadsheet, using connectors such as MCP (the Model Context Protocol). The test would show whether accuracy and cost hold up from end to end, and where a person still needs to check the work.
- **A fine-tuned open-weight model.** No model in this test was fine-tuned. A business could train an open-weight model such as Mistral Large 3, GLM 5.3 or Llama 4 Maverick on its own statements, to see whether a small model matches the expert-prompt results without the 1,500-word briefing, and whether that lowers the cost further.

Happy to talk through the method with anyone planning a test like this on their own tasks or workflows, and to hear which of these questions matters most to you. Let me know at martin@scalepointpartners.com.

Sources

Sources

Numbers in the text refer to this list. Sources 11 to 28 are cited in the technical annex. The annex is at scalepointpartners.com/labs/ai-economics-card-statements.

1. Stripe, "Stripe agrees to acquire OpenRouter", press release, August 2026. OpenRouter connects to more than 400 models from over 80 providers; Stripe's reported price was \$7.5bn (stripe.com)
2. Google, "Gemini 3.8 Flash", launch post and pricing, September 2026 (blog.google)
3. OpenRouter, Zero Data Retention guide, checked September 2026 (openrouter.ai)
4. US Department of Commerce, additions to the Entity List, Federal Register, 16 January 2025 (federalregister.gov)
5. Euronews, "DeepSeek AI blocked by Italian authorities as other member states open probes", 31 January 2025 (euronews.com)
6. BleepingComputer, "Germany asks Google, Apple to remove DeepSeek AI from app stores", June 2025 (bleepingcomputer.com)
7. US National Defense Authorization Act for FY2026, section 1532, as summarised by ETO AGORA (agora.eto.tech)
8. Insurance Journal, reporting on the UK government's position on DeepSeek, February 2025 (insurancejournal.com)
9. Regulation (EU) 2022/2554 on digital operational resilience for the financial sector (DORA) (eur-lex.europa.eu)
10. Prudential Regulation Authority, SS2/21 Outsourcing and third party risk management, March 2021 (bankofengland.co.uk)

11. Anthropic, company background and funding, checked September 2026 (wikipedia.org)
12. Anthropic, Claude Opus 5.5 launch page, September 2026 (anthropic.com)
13. SpaceX, pricing of initial public offering, June 2026; SpaceXAI background (ir.spacex.com)
14. xAI, Grok 4.7 launch post, September 2026 (x.ai)
15. PBS NewsHour, "EU investigates Musk's AI chatbot Grok over sexual deepfakes", January 2026 (pbs.org)
16. Ofcom, investigation into X over Grok sexualised imagery, January 2026 (ofcom.org.uk)
17. Mistral AI, company background and ASML investment, checked September 2026 (wikipedia.org)
18. Mistral AI, Mistral Large 3 model card, December 2025 (docs.mistral.ai)
19. Caixin Global, "China's Zhipu AI jumps in Hong Kong debut", 8 January 2026 (caixinglobal.com)
20. The New Stack, on the GLM-5.3 weights and licence, August 2026 (thenewstack.io)
21. Forbes, "DeepSeek just raised \$7.4 billion. Here's the catch", 17 June 2026 (forbes.com)
22. DeepSeek, DeepSeek V4.1 Flash model card, September 2026 (huggingface.co)
23. OpenAI, company background and ownership, checked September 2026 (wikipedia.org)
24. Alphabet Inc., company background, checked September 2026 (wikipedia.org)
25. Cohere, company background, checked September 2026 (wikipedia.org)
26. Moonshot AI, company background, checked September 2026 (wikipedia.org)
27. Alibaba Group, company background, checked September 2026 (wikipedia.org)
28. MiniMax Group, company background, checked September 2026 (wikipedia.org)
29. 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)
30. 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 is a payments and fintech adviser to private equity-backed businesses and start-ups in the UK and Spain, working through ScalePoint Partners.

Before ScalePoint he spent ten years in payments at Citi, Mastercard and Luup, then nearly a decade at Edgar, Dunn & Company, latterly as a Principal, on more than 75 engagements for banks, fintechs, card networks, private equity firms and merchants.

scalepointpartners.com





Strategic Finance for Payments and Fintech Leaders

martin@scalepointpartners.com

scalepointpartners.com

London - Barcelona