Judgement calls in payments operations: Claude Opus 5.5 vs Mistral Medium 3.5

On judgement calls in payments operations, does a low-cost open-weight model do the job as well as a frontier model?

KEY FINDINGS

No. Used out of the box, Mistral Medium 3.5 was a false economy on the judgement decisions we tested in nine payments and finance processes. Claude Opus 5.5 at low effort got 385 of 394 answers right, for 1.9 US cents a case on average. **Mistral without reasoning** got 251 right and was 97 to 98% sure whether it was right or wrong, so its confidence couldn't pick out the cases a person should check. At 0.6 US cents a case in tokens it was clearly cheaper than Opus, but once wrong decisions were costed, it cost more than Opus in 6 of the 7 processes scored in money. **Mistral with reasoning on** got 344 right at 3.1 US cents a case, so it was both less accurate and more expensive than Opus at low effort, because its extra thinking used more tokens. **Mistral Large 4**, whose open weights came out after the study, was added on 11 November and got 362 right at 1.0 US cents a case, about half the token cost of Opus at low effort. Opus gave lower confidence scores on the answers it got wrong, so its doubt pointed to the cases a person should check. Large 4 needed more cases checked to keep losses down.

BACKGROUND

In previous tests (S1 and S2), we found that low-cost models came close to the best on routine steps such as reading card statements. The models pulled apart where judgement was needed. Hence, in this latest test (S3), we deliberately tested only judgement steps. Today, rules typically settle the clear cases and pass the rest to an analyst. Could the LLMs in our test make the decisions a human would likely have made on our invented test cases?

STUDY DESIGN AND SET-UP

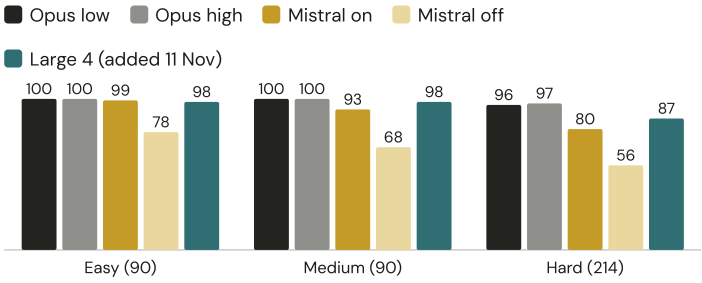
We chose nine processes from our process map. Each process had 22 invented cases and a rulebook. In 17 of them (5 medium, 12 hard) we planted subtle traps: lookalike records, sources that disagree, missing evidence or a rule that changes the answer. We called each model through its API, via OpenRouter, at two reasoning settings: Claude Opus 5.5 at low and high effort, and Mistral Medium 3.5 with reasoning off and on. In total the models gave 1,576 answers, 394 for each of the four set-ups. Decisions were scored against an answer key, and errors were costed using a simple economic model for each process.

EXAMPLE CASE, ACCOUNTS PAYABLE A £8,460 glassware invoice matched the purchase order, the goods receipt and the supplier's bank details. We planted one warning sign: the covering email came from tynemouth-glassware.co, while the supplier's domain on file is tynemouth-glassware.co.uk. The rulebook says to hold an invoice like this and confirm with the supplier. Opus held it. Mistral without reasoning paid it in both runs, at 100% confidence.

RESULTS

- Mistral without reasoning got fewer right than Opus in all nine processes (Figure 1, Table 1). The gap grew with difficulty: it got 20 of 90 easy answers wrong. Opus missed only hard cases.
- Opus gave lower confidence scores on its wrong answers, about 70% against 86% on right ones, so its doubt could flag cases for checking. Mistral's was 97 to 98% either way.
- Higher reasoning effort added only two right answers for Opus. Whilst we probed various ways to make the test more challenging, the cases clearly didn't stretch Opus, and this study was unable to say where its limit is.
- Token costs, at API list prices, ran from 0.4 to 0.9 US cents a case for Mistral without reasoning, 1.2 to 3.3 for Opus at low effort, 1.3 to 8.4 for Mistral with reasoning on and 0.5 to 2.2 for Mistral Large 4.

Figure 1. Right answers by difficulty, % of answers, all nine processes



Each set-up gave 90 answers on easy cases, 90 on medium and 214 on hard.

Table 1. Right answers by process (out of 44 each, 42 for payment performance, 394 in total), and Mistral's worst miss without reasoning

Process	Firm	Opus low	Mistral on	Mistral off	Large 4	Mistral off: worst miss
Treasury prefunding	EMI	44	41	0*	43	Missed an extra payout before a closed day ($388,000 unfunded)
Deal pricing	Acquirer	43	43	21	44	Repriced a deal that already met the margin floor ($2,000)
Accounts payable	Merchant	44	34	23	37	Paid an invoice from a lookalike email domain ($11,300)
Payment performance	Merchant	40	33	25	35	Left a fraud rule declining good payments in place ($17,300 of lost sales)
Payment exceptions	EMI	40	33	30	34	Resubmitted a payment to the wrong lookalike payee ($9,700)
Chargebacks	Merchant	42	37	33	41	Accepted a chargeback with conflicting postcodes, not escalated ($500)
Seller onboarding	Marketplace	44	41	38	42	Approved a new seller that the policy sends to a person ($2,700)
Sanctions alerts	EMI	44	40	39	43	Cleared a near-match that should have gone to an analyst (points)
Scheme changes	Issuer	44	42	42	43	Said a "Recommended" change to refund timing didn't apply (points)
All nine		**385**	**344**	**251**	**362**	
Token cost per case		1.9¢	3.1¢	0.6¢	1.0¢	Average across all nine, at API list prices

*A treasury answer counted as right only if every currency amount was within one rounding step. Mistral without reasoning got at least one amount wrong in all 22 cases, mostly by small sums. The worst miss is the costliest wrong decision, modelled at each firm's own rates in US dollars.

The full study is at scalepointpartners.com/research
All 1,576 S3 answers, with each model's reason, evidence and confidence, are available on request.

Martin Koderisch, Founder and Lead Analyst
martin@scalepointpartners.com · London · Barcelona