The model you approved is a routing preference

OpenRouter's Auto Router ranks models on what its community spent in a trailing seven days, and an unset cost tier routes “as if you had asked for roughly the low band”. A fallback fires on a context length error or a moderation flag. Anthropic's proof of inference covers integer weights under greedy decoding, MoE models and 1,024-token contexts, at 72 times the serving time per GPU.

6 November 2026 | Strategic Intelligence Division | Governance

Model selection moved out of the evaluation and into the gateway, and the gateways document the move as a default. OpenRouter's routing documentation, “Auto Router”, dated 31 October 2026, ranks candidate models on what its community spent over a trailing seven days and says, in its own words, “Requests that set no cost setting route as if you had asked for roughly the low band.” A fallback then switches the model on any error, including a context length validation error and a moderation flag. The buyer approved a named model against an eval set; the route that answers may be a different one, and the only record of which one is written by the party that chose it. This reaches you if a regulated, customer-facing or contractual output is produced through a gateway rather than a direct endpoint, or if a platform contract permits routing across models without naming them.

72x
Time to prove a batch of 24 requests, measured per GPU, against the time vLLM takes to serve the same batch, Anthropic's benchmark

94x
The same per-GPU comparison when vLLM runs at the highest load Anthropic tested

1,024
Longest context, in tokens, that Anthropic's proof of inference covers

7
Trailing window, in days, of community spend that ranks the model in OpenRouter's Auto Router

Source: Anthropic, “Progress on provable inference for large language models”, published 2 November 2026, Evaluation and Limitations sections; OpenRouter, “Auto Router”, routing documentation dated 31 October 2026. The 72 and 94 multiples are Anthropic's measurement on Llama 4 Maverick, not Mosaic Singularity analysis.

The gateway sets the price band, the model and the fallback order

Auto Router classifies each prompt in flight into one of about 30 task types, looks up the models the community “actually spends on over a trailing 7-day window” for that task, applies a cost tier, and turns the survivors into “the primary pick plus fallbacks”. The tiers run low, medium, high, xhigh and max. OpenRouter is explicit that a tier is not a ceiling: “A tier is a band, not a ceiling, so models cheaper than the band are excluded as well as models above it.” A request that sets none of them takes the low band.

Two consequences are configuration, not theory. The router “can pick a different model on every turn”, and it reuses the model a conversation landed on only “while that model is still one of the top candidates for the new prompt”, so a single session can be answered by several models. And a restriction a buyer believes is in force can be a no-op: OpenRouter warns that settings sent under the other Auto Router slug's plugin id “are accepted but silently ignored”, with allowed_models, excluded_models and cost_tier having “no effect on the request”.

A moderation flag is enough to change the model

OpenRouter's Model Fallbacks documentation states the purpose: “The models parameter lets you automatically try other models if the primary model's providers are down, rate-limited, or refuse to reply due to content moderation.” The documented trigger list is wider than an outage: “By default, any error can trigger the use of a fallback model, including” context length validation errors, moderation flags for filtered models, rate limiting and downtime. Billing then follows the substitute, because “Requests are priced using the model that was ultimately used, which will be returned in the model attribute of the response body.” The invoice moves to the model that answered. The approval does not.

Two more gateways publish the same behaviour. Vercel's AI Gateway documentation records that the service “routes requests across providers and fallback models, then records status, provider, latency, token usage, cost, and every routing attempt”, and tells the buyer to “set a provider order and read the response metadata” in order to “verify which provider served the request”. A Cloudflare documentation page updated 20 April 2026 triggers a fallback “in response to request errors or predetermined request timeouts” and returns a response header, cf-aig-step, that “indicates which step successfully processed the request”.

The delegation is already inside the enterprise. Block's Form 8-K filed 3 November 2026, Exhibit 99.1, states: “In October, we began rolling out model routing to all employees at Block, giving everyone access to a simple, efficient way to match their work to the right model capability.” That is the right outcome on cost and the wrong shape for an approval, because a layer that chooses per request moves the choice of model from a review into a policy, and no review runs per request.

The audit trail is written by the party that chose the route

None of this is concealed, and the distinction carries the argument. Every gateway above reports which model or which provider answered: OpenRouter returns the selected model in the response and prices the request from it, Vercel points the buyer at the response metadata, Cloudflare names the step in a header. What the buyer lacks is attestation, not disclosure. The record is produced by the system that chose the route, retained under that system's terms, and carries the trust the buyer places in the party whose routing decision is in question. For a dashboard that is enough. For a model approval, a customer warranty or a regulated output, it is a supplier's account of its own conduct.

The proof that would settle it takes 72 times longer than serving, per GPU

Anthropic, “Progress on provable inference for large language models”, published 2 November 2026 by its Frontier Red Team, describes an instrument built for exactly this gap. FlashProve commits a provider to a set of weights and then proves, for each request, that the reply was that model's output. The executive summary names the threat without euphemism, a reply “secretly substituted in by a routing service for cost reasons”. The cost appears in the same document, measured on Llama 4 Maverick with vLLM run on eight H100s as the comparison: per GPU, proving a batch of 24 requests takes “about 72 times longer than vLLM takes to process the same 24 requests at once (66 to 77 across runs), and about 94 times longer when vLLM runs at the highest load we tried”. Each proof is 18.78 MB, and checking one takes “17 seconds on 24 threads (and 21 seconds on a single thread)”.

One proof covers one request, and the buyer's half of this is small: 18.78 MB for a 1,024-token window, 17 seconds to check on 24 threads, and a commitment of “~6.5 MB in all” downloaded once.

The document says so plainly: “A proof file is small and cheap to check (succinctness).” The expensive half belongs to the side that sells the route. Proving a batch of 24 requests takes the prover about 223 seconds on one H100 at 110 tokens per second, and that is where the 72 to 94 multiple against vLLM's per-GPU share comes from; the commitment itself “takes about 39 minutes” per model and is “good for about 16 million requests, after which the provider has to commit again”. Attestation is therefore a supplier's cost of goods, not an auditor's storage problem, and no gateway above offers it: the only instrument that would settle the question is priced against the margin of the party whose routing decision is in question.

The proof covers a configuration that routed traffic does not run

The document scopes its own result harder than its headline does. “This release is a research prototype, and has the following technical limitations”, and the first is decisive: “The proof is about an integer model under greedy decoding.” A reply from a floating-point implementation, or from sampled decoding, “could depart at some position and no longer exactly match the computation proved”. The system also “assumes the served model uses a Mixture-of-Experts (MoE) architecture”, and it “currently proves inference over contexts of up to 1,024 tokens”, which the same section calls “quite short in comparison to the contexts accumulated in chat settings”.

What the certificate excludes, and why an approval cannot lean on it

Anthropic's stated limit	What falls outside the certificate	What that leaves the approver holding
“The proof is about an integer model under greedy decoding.”	Sampled decoding and floating-point weights, which is what a chat completion returns whenever the temperature is above zero	“A reply from a floating-point implementation, or sampled decoding, could depart at some position and no longer exactly match the computation proved.”
“Our system assumes the served model uses a Mixture-of-Experts (MoE) architecture”	Dense models, which the document says “are also widely used”	The proof does not carry across every architecture in the same routing pool
“The system currently proves inference over contexts of up to 1,024 tokens.”	The accumulated context of a session or an agent run	“This is quite short in comparison to the contexts accumulated in chat settings”
“No formal guarantees on fast executable code”, and “we do not guarantee that all three (Lean, paper, code) match each other”	Assurance that the fast verifier an auditor runs is the protocol that was proved	The audit rests on an implementation the authors decline to certify
“This release is a research prototype”	A control with a published price per request	The 72 and 94 multiples are a benchmark on one model, not a rate card

Source: Anthropic, “Progress on provable inference for large language models”, published 2 November 2026, Limitations and Evaluation sections. Column two is Mosaic Singularity analysis, November 2026.

The exposure sits with the approver, and the gain with the router

The enterprise that approved a model carries this one, and the seat is the Chief AI Officer with the CISO beside them, because the exposure is an output no one can tie to the model the approval names. The Chief Financial Officer meets it as a line: requests are priced on the model ultimately used, so the AI forecast is a forecast of routing decisions, and attestation, if it is ever demanded, is priced at a multiple of the workload's own GPU cost. The General Counsel meets it as a clause, because a fallback that fires on a context length error is an availability control that quietly changes which model answers.

The gain sits with the gateway, which has moved a purchasing decision into a routing policy it owns, and second with any model owner that can gate substitution behind a signature. Anthropic states that it met its own September 30, 2026 goal for a provable inference prototype “through a separate but related project, which signs our inference traffic and recomputes inference on a trusted server”. Attested provenance, in whichever form it arrives, is a switching cost attached to a named model. The loser is the independent evaluator, whose baseline is measured under a routing regime it does not control, and the team that has to prove after the fact which model answered a request it did not route.

Four seats have to move before the platform agreement is signed.

- Chief Information Security Officer: by 31 December 2026, decide whether the gateway's record is evidence or a lead. If it is evidence, set retention on the response model field and the cf-aig-step header per request, then reconstruct one disputed output end to end and see whether the trail survives. The record exists because the gateway writes it, not because the contract requires it.
- Chief AI Officer: by 31 December 2026, re-run each approval for a workload that calls a gateway against the model the gateway actually served, read from the response rather than from the approval memo. A ranking built on a trailing seven day window can move a workload's route inside a week, which is faster than any annual model review.
- General Counsel: before the FY27 platform agreement is signed, and no later than 31 March 2027, write two sentences into it, one naming which models may answer each regulated workload, and one requiring written notice before cross-model fallback is enabled for that workload. The trigger list includes a context length validation error, so a clause that permits fallback for availability also permits a different model to answer a prompt that was merely too long.
- Chief Financial Officer: by 31 March 2027, re-forecast the AI line on the model actually used, since requests are priced on the model ultimately used, and carry one unit for the attestation option in the planning file: 72 hours on one H100 for every hour the workload takes to serve on one GPU, because proving takes about 72 times the per-GPU serving time of the same batch, and a storage line of 18.78 MB per 1,024-token request for the proofs themselves.

The board's signature is a model list, not a proof

One item reaches the board rather than a function: the FY27 platform agreement, tabled for signature no later than 31 March 2027, carrying the named model list and the written fallback notice. It is the only artefact in this piece a director signs, and what it commits is a list of models, not a proof that those models answered: the certificate that would prove it is a research prototype scoped to integer weights under greedy decoding.

- Which of our approved models can be replaced without a new approval, on which workloads, and who signed that permission?
- Does each gateway configuration set a cost tier explicitly, or does it inherit the roughly low band by default?
- Which of our gateway restrictions are accepted and silently ignored because they are addressed to the other router slug?
- For the last disputed output, can we show which model produced it from a record we control, or only from the supplier's log?

- If a customer or a regulator demands attestation, what does one proved request cost: 18.78 MB and 17 seconds for us to check, and 72 times the per-GPU serving time for whoever proves it, and whose budget carries which half?

Either the default changes by 30 November 2026, or the scoping stands

If OpenRouter's routing documentation no longer states, on 30 November 2026, that a request with no cost setting routes as if for the roughly low band, the default substitution reading fails and the route becomes the buyer's choice. If a gateway or a model provider offers per request attestation for sampled, floating-point, long context production traffic at a stated price before 31 March 2027, the scoping argument fails and the certificate moves from audit to control. Until either happens, the approval names a model, the route decides which model answers, and the only proof of the answer is the router's own record.

- Anthropic, “Progress on provable inference for large language models”, Frontier Red Team, published 2 November 2026.
- OpenRouter, “Auto Router”, routing documentation dated 31 October 2026.
- OpenRouter, “Model Fallbacks”, routing documentation.
- Vercel, AI Gateway documentation, overview.
- Cloudflare, AI Gateway “Fallbacks” documentation, last updated 20 April 2026.
- Block, Inc., Form 8-K, Exhibit 99.1, filed 3 November 2026.