Receipts next to attested inference

Attestation and an independent receipt answer different questions about the same request. This page sets out what each one covers, so that whoever relies on the evidence knows which one to ask for.

What attestation of an operator's runtime covers

Attestation lets the party running a model show what it started. The platform measures the code and environment at boot (firmware, operating system, the inference software and sometimes the model files) and signs a report. A relying party compares the report with the values it expects. It is an event about the operator's own stack, taken once at launch or repeated periodically. Done well, it is strong evidence that the environment is the one described.

What sits outside it

- **Which model produced a given output on a given request.** A measurement at boot says what was loaded. It is not a statement about each answer that followed, after a version update, a configuration change or a request served by another deployment.
- **Evidence a third party can check without trusting the operator.** A reviewer can check the environment against the platform's signature, but the record of the work done inside it, request by request, is still the operator's own log.

What a receipt from a party that did not run the model adds

An independent verifier re-executes a sample of the work on separate compute and signs a receipt that binds model, input and output. The verifier did not serve the original request and does not rely on the operator's stack. The result is a tamper-evident receipt: evidence about the work that comes from outside the party that did it. The two fit together. Attestation speaks for the environment, the receipt speaks for the work, and neither replaces the other.

Question	Attestation of the runtime	Independent receipt
Was the environment the one described?	Yes, at boot or periodically	Not its purpose
Which model produced this output, for this input?	Not request by request	Yes, for the sampled work
Who produces the evidence?	The operator's platform	A verifier that did not run the model
Who relies on it?	The operator's customers and their security teams	An auditor, a model-risk team, a regulator

Who checks the receipt, and why

A model-risk team checks that the model named on the receipt is the one it approved. An auditor compares receipts with the operator's own logs for the same requests. A regulator can ask for them when a decision is questioned. In each case the reason is the same: evidence about the work should not come only from the party that did the work.

Status. Live for open-weight embedding models on ordinary GPUs, including your own model through ONNX (catalog: bge-large, bge-small, EmbeddingGemma 2). Each new model joins the same way: exported to ONNX, calibrated against its reference, then independently verified. The committed-trace method is in a public preprint (zenodo.org/records/24296631).