Model Garden receipts through request-response logs

A customer who deploys an open-weight model from Model Garden to an endpoint in its own project can already have that endpoint write a sample of requests and responses to BigQuery. This page sets out what those logs record, what a reviewer still needs, and how an independent receipt can be issued from them with nothing changed on the endpoint.

What the endpoint's logs already record

- Request-response logging is set on the endpoint, so requests to any model deployed under it are logged.
- The customer chooses the fraction of requests logged (a sampling rate between 0 and 1) and the BigQuery destination.
- Each row carries the endpoint, the deployed model ID, the logging time, a request ID, and the request and response payloads.
- On dedicated and Private Service Connect endpoints, payloads under 10 MB are logged, for the predict and rawPredict methods.

What a reviewer still needs

The log is a faithful record kept by the platform that served the request: what the endpoint received and what it returned. That is the right evidence for operations. A model-risk team or an auditor asks a further question: is there evidence, from a party that did not run the model, that this output came from the published weights the customer approved? This is not a gap in the log. It is a question of who produces the evidence.

How a receipt is issued from the logs

1. The customer gives an independent verifier read access to a sample of the logged rows. Nothing is installed on the endpoint and the serving path does not change.
2. The verifier re-executes the sampled requests on separate compute, with the published weights of the model the customer approved.
3. It signs a tamper-evident receipt for each sampled request that binds model, input and output, and that names the log row it covers.
4. The customer keeps the receipts next to its logs. The verifier sees only the rows in the sample the customer shares.

Question	Request-response log	Independent receipt
What did the endpoint receive and return?	Yes, for the logged fraction	Taken from the log
Did this output come from the approved published weights?	Not its purpose	Yes, for the sampled requests
Who produces the evidence?	The platform that served the request	A verifier that did not run the model
Who relies on it?	Operations and security teams	A model-risk team, an auditor, a regulator

Who checks the receipt, and why

A model-risk team checks that the model named on each receipt is the one it approved. An auditor compares receipts with the log rows they name. A regulator can ask for both when a decision is questioned. The reason is the same in each case: evidence about the work should not come only from the party that did the work.

Status. Independent verification is 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. Reading a sample from a Model Garden logging table is offered as a pilot on one endpoint, at our cost. The committed-trace method is in a public preprint (zenodo.org/records/24296631).

Logging facts: Google Cloud documentation, "Use online inference logging" (last updated 4 November 2026, read 5 November 2026). Applies to models deployed to an endpoint in the customer's project.