Gemini Decisions API: Why 1,000 Checks Share Input, Not Answers

TechRounder PDF Edition

Live article: https://www.techrounder.com/ai/gemini-decisions-api-isolated-batch-judgments/

By Vipin PG | Published November 12, 2026 | Updated November 12, 2026 | Format: Analysis | 7 min read

In brief

Gemini Decisions API runs up to 1,000 independent checks on shared text, images, audio, video, or PDFs, returning probabilities without output tokens.

Google’s Gemini Decisions API can score up to 1,000 questions against one shared input, but those questions do not build on one another. It is designed for applications that need bounded judgments, such as checking whether a document meets a rule or routing a request, rather than another model-written answer.

Announced on November 11, 2026, the API pairs Gemini Decision 1 with typed questions and returns probabilities, choices or rubric scores without generating output text. This can simplify routing and classification code, but the 1,000-question limit is a batching ceiling, not a promise of sequential reasoning or a universal performance advantage.

For developers, the design question is whether a workflow consists of independent checks over the same evidence. If it does, the API offers a direct scoring interface. When one answer must shape the next question, application code or another request still has to carry that dependency.

The endpoint targets bounded decisions

Google introduced the Gemini Decisions API and Gemini Decision 1 on November 11, 2026, through the Gemini API in Google AI Studio. The company describes it as a way to make fast, closed-set judgments over text, images, audio, video and PDF input in its Gemini Decisions API announcement.

The intended answers are constrained in advance. A developer might ask whether a claim appears complete, which support team should receive a ticket, or how a message scores against a rubric. Google’s launch post presents tool selection, policy checks, model routing and context grading as example uses, not guarantees that the model will make every such decision correctly.

The product’s promise is structural: send shared evidence with distinct questions, then let software act on typed results. The response format sets it apart from a conventional model call.

How the Gemini Decisions API returns results

With a generative model, an application often asks for a judgment in prose and then parses the response. Google’s Decisions API guide says this can add latency and brittle parsing logic. The endpoint instead returns probabilities and produces zero text output tokens.

Your code still sets action thresholds and decides what to do with uncertainty, including whether to escalate. The API supplies a typed signal, not a complete business process.

Google says Gemini Decision 1 scores candidate outcomes in one forward pass instead of emitting answer tokens one at a time. Its launch post reports 150-250 milliseconds end-to-end latency in the described tests, with model-compute median latency of 50-100 milliseconds for contexts up to 16,000 tokens. These are Google-reported measurements, not independent guarantees for every request size or deployment.

The endpoint exposes three question types. Which one fits depends on the kind of answer your program needs.

Three question types cover common checks

Gemini Decision offers yes-or-no, multiple-choice and ordered-rubric questions. Each question has a developer-chosen key, a type and text instructions; Google's Gemini Decision model reference lists up to 1,000 questions per request.

A 'noul' question estimates whether a statement is true and returns a probability from 0.01 to 0.99 plus a confidence value. A 'choice' question selects among fixed options and returns probabilities for them; Google's model reference lists 2 to 255 choices.

A 'score' question places the input on ordered rubric levels and returns an expected score, per-level probabilities, a legend and confidence. It works best when the rubric defines meaningful distinctions.

Typed outputs make it easier to write deterministic downstream rules, but they do not make the questions a conversation with each other. That limitation defines the API's main architectural boundary.

Questions share evidence, not answers

Every question in a request is evaluated against the same shared input, in parallel. Google's guide explicitly says questions cannot see one another or their answers, so the request behaves like a checklist rather than a chain of reasoning.

This fits several narrow checks on one record: checking for a required field, classifying urgency and choosing a handling queue. Your application can combine the results with conditions, weights or escalation rules.

If a later check depends on an earlier result, the dependency must live outside the question set. Code can branch and issue another request, but the next question will not inherit the earlier answer unless the application supplies it. Google also says latency stayed nearly flat as it increased requests from 1 to 100 questions; that reported result should not be stretched into a claim about 1,000-question workloads.

When independent checks are the right shape, a shared multimodal input can reduce the need to split evidence before the model sees it. The supported input routes and limits still matter.

Multimodal evidence has request boundaries

Gemini Decision accepts text, images, audio, video and PDFs as shared input. Google's model page sets the input ceiling at 262,144 tokens and the output limit at zero; media can be combined with text in a single request.

The guide documents image and video resolution settings of 'low', 'medium', 'high' and 'ultra_high'. Teams should test the setting against the task; higher detail can affect token use.

Questions themselves are text-only: the guide says media cannot be embedded in question instructions. YouTube URLs and direct ‘gs://‘ Cloud Storage URIs are unsupported, while Files API URIs, public or signed HTTPS URLs and inline base64 are listed as supported routes. Conversation-history, thought and tool-call blocks are also unsupported as input blocks; Google says multi-turn context must be serialized as text.

These constraints define the request shape. Billing determines whether batching saves money for a real workflow.

The output is free, but instructions still count

Google lists Gemini Decision 1 at $0.10 per million input tokens and $0 for output tokens, and says a no-cost tier is available in Google AI Studio for prototyping. The shared input is billed once per request, regardless of how many questions use it.

Question instructions and criteria add input tokens, so elaborate rubrics cost more than short checks. Estimates should count the evidence and every question’s wording and options, within the 262,144-token input limit.

The API does not support context caching, according to Google’s guide. If dependencies require another call, sending the evidence again counts it again; that can affect total cost even when the first request was efficiently batched.

Google reports nearly five times lower cost than Gemini 3.5 Flash-Lite through the Interactions API across its shared benchmark suite. That is a vendor comparison scoped to its tests, not a general savings rate. Before estimating deployment cost, teams also need to check their own quota and request pattern.

Quotas can constrain production throughput

The documentation specifies the Gemini Developer API’s ‘v1beta‘ interface and Gemini Decision 1. The endpoint does not support the Batch API, Flex inference, Priority inference or context caching.

Google’s rate-limit guidance says quotas vary by model and usage tier, apply per project rather than per API key, and can vary in actual capacity. It points developers to AI Studio to view active limits. The public page did not expose a fixed Gemini Decision-specific quota in the material verified here, so a team should check its own project rather than infer throughput from the per-request question maximum.

The launch post said Google Cloud support was "coming soon." That was the company’s launch-day statement, not confirmation of current availability. Once access and throughput are confirmed, workload fit becomes the next decision.

Use it for fixed outcomes, not explanations

Google’s guide recommends the Decisions API for fixed-outcome probabilities, thresholds and many independent checks over one input. That makes routing, policy screening, candidate filtering and rubric scoring plausible starting points, if the application can act on a bounded result.

Use a generative endpoint when a workflow needs a summary, explanation, free-form answer, custom structured extraction, tool arguments or multi-step reasoning. A decision result can trigger those steps, but does not replace them.

Google says its Gemini Notebook uses the model to choose a model tier and thinking level, and reports up to 40% lower latency for everyday queries in that internal routing workflow. Treat that as a company-reported result from one use case, not a prediction for another team's workload.

Production use still needs thresholds, fallbacks, retries and dependency handling in application code. Comparing other decision-model options can help shape a test plan, but public launch claims are not a substitute for one.

Alternatives make a useful test baseline

OpenAI announced general availability of its separate Decisions API on October 22, 2026; its launch announcement says requests can include up to 400 questions and that prompt caching for reused context is available.

TypeSafe's Jev model documentation lists version 1.14 at $0.042 per million input tokens, with text and image inputs. It describes shared-state questions evaluated in parallel. Those are vendor specifications, not a head-to-head price-and-quality verdict; input limits and the shape of a team's evidence also matter.

Independent technology writer Simon Willison tested Gemini on a video with four yes-or-no questions and published the resulting probabilities in his hands-on API note. In his small comparison, he noted that Jev and OpenAI returned 1.0 or 0.0 while Gemini returned non-extreme probabilities. That observation illustrates a difference in outputs; it does not establish which model is more accurate or better calibrated.

Artificial Analysis's Decision Index combines six evaluations covering document decisions, tool calling, knowledge and hallucination, reasoning and invariance. It offers a broader frame than a launch demo, but no universal winner follows from a composite score. A controlled test on the intended workload is still necessary.

The evidence points to meaningful product differences, while leaving the central deployment question open: how does each option perform on the same data, labels and failure costs?

Measure the workflow before moving it

Build a labeled sample and compare Gemini Decision with the current generate-and-parse process and at least one alternative. Define correctness, then measure accuracy, calibration, false-positive and false-negative costs, and how uncertain cases are handled.

Measure median and high-percentile latency and full cost with realistic evidence, media, instructions, retries, dependent calls, repeated inputs and required generation. Google's figures describe its own tests, not those combined costs.

Finally, test whether the request limit and your project quota match the work you expect to run. A 1,000-question batch can be useful when those checks are independent, but it cannot eliminate dependencies, quota checks or application-level safeguards. The Gemini Decisions API is ready for production only when a representative test supports that choice, not because it has the largest advertised batch.

References

1. blog.google - innovation-and-ai / technology - https://blog.google/innovation-and-ai/technology/developers-tools/gemini-decisions-api/

2. aistudio.google.com - docs / decisions-api - https://aistudio.google.com/docs/decisions-api

3. aistudio.google.com - docs / models - https://aistudio.google.com/docs/models/gemini-decision-1

4. aistudio.google.com - docs / rate-limits - https://aistudio.google.com/docs/rate-limits

5. openai.com - index / decisions-api - https://openai.com/index/decisions-api/

6. docs.typesafe.ai - models - https://docs.typesafe.ai/models

7. simonwillison.net - 2026 / Nov - https://simonwillison.net/2026/Nov/11/gemini-decisions-api/

8. artificialanalysis.ai - evaluations / artificial-analysis-decision-index - https://artificialanalysis.ai/evaluations/artificial-analysis-decision-index