

ENTERPRISE PROCUREMENT FIELD GUIDE

# AI Inference Provider RFP and Procurement Guide

Security · Data Residency · SLA · Capacity · Cost · Exit Risk

*A practical evaluation package for technology, security, privacy, legal, procurement, finance, compliance,  
and business owners*

**CORE RULE** Apply hard disqualification gates before weighted scoring. Treat verified behavior, contractual commitments, official documentation, vendor assertions, and assumptions as different evidence classes.

Enterprise evaluation package — 2026 edition

Official-source verification date: 2026-08-19

## One-page executive checklist

**A decision may be “Go” only when all mandatory gates pass, technical proof matches the proposed production scope, and remaining unknowns are explicitly risk-accepted.**

### 1. Workload and data

- Named business owner, user harm/outage impact, and criticality tier.
- Versioned tasks, quality thresholds, peak load, latency and RTO/RPO.
- Data classes, jurisdictions, prohibited content, retention and residency rules.
- Required model, features, tools, region, deployment type and production status.

### 2. Hard gates

- Exact service scope passes all 16 gates; no preview or roadmap substitution.
- ZDR/retention exceptions checked feature by feature, not brand by brand.
- DPA/BAA/sector, license, AUP, private network and audit controls are in place.
- Capacity is committed or proven; a rate-limit ceiling is not treated as a guarantee.

### 3. Evidence

- Every answer labeled: verified, contractual, official documentation, vendor assertion, or assumption.
- Evidence names entity, service, model/version, endpoint/tool, tier, region and date.
- Certification/report scope and customer control responsibilities are mapped.
- Unknown stays unknown; owner, due date, decision impact and fallback are recorded.

### 4. POC proof

- Representative and edge tasks; structured output, tools and long context.
- TTFT/output speed/end-to-end p50/p95/p99 at realistic peak and burst.

- 429/5xx/timeout/stream failure, retries/idempotency and support escalation.
- Region/data flow, pinning/upgrade, failover, deletion/export and invoice reconciliation.

### 5. Commercial and legal

- All billing units and failed/retried/tool/cache/reasoning cases modeled per accepted job.
- Covered SLA scope, measurement, exclusions, credits and chronic-failure exit negotiated.
- Price protection, ramp/minimums, deprecation notice, IP, suspension and subprocessor terms closed.
- Marketplace/reseller responsibilities and cloud-credit treatment confirmed in writing.

### 6. Resilience and decision

- Fallback removes correlated dependencies and has tested capacity/quality/residency.
- Portable interfaces preserve provider-native extensions; prompts/evals are versioned.
- Emergency and planned migration meet maximum migration time; deletion evidence defined.
- Decision memo states gates, scores, confidence/freshness, unknown exposure, residual risk and conditions.

## How to use this package

**OUTCOME** A defensible shortlist and contract position, supported by buyer-run evidence, with a tested operating and exit plan. This guide does not choose a universal winner; the best answer changes with workload, regulatory exposure, commercial leverage, and operational maturity.

1. Classify the workload, data and failure impact before asking vendors questions.
2. Convert mandatory requirements into yes/no hard gates. Stop evaluating any route that fails a non-waivable gate.
3. Issue the RFP with explicit evidence requests. Reject answers that omit service, model, endpoint, feature, tier, region, or contractual scope.
4. Run the POC against the exact production proposal. Test behavior, not slideware.
5. Score only eligible proposals. Keep unknowns visible; calculate evidence coverage separately from technical merit.
6. Negotiate the gaps that matter, rehearse failover and exit, then document conditions and residual risks in the decision memo.
7. Revalidate continuously, quarterly, annually, and whenever a material change occurs.

## Package map

| Component          | Purpose                                                                                                                   |
|--------------------|---------------------------------------------------------------------------------------------------------------------------|
| Guide              | Decision method, evidence hierarchy, workload intake, gates, scorecard, POC, commercial terms, resilience and governance. |
| Appendix A         | Detailed 32-domain RFP questionnaire with acceptable answers, warnings, knockouts and requested proof.                    |
| Appendices B–L     | Evidence request, gates, scorecard, risk/data-flow worksheets, POC, decision memo, contract and exit templates.           |
| Appendices K–L     | Dated source ledger and maintenance/revalidation schedule.                                                                |
| Companion workbook | Editable copies of the checklists and formula-driven scoring/coverage calculations.                                       |

## Publication and legal note

**This guide is an operational procurement aid, not legal advice. Qualified counsel should review applicable law, negotiated terms, transfer mechanisms, sector obligations, and the final agreement.**

*Source references use ledger IDs such as [SRC06]. Verification means the cited official page was checked on the ledger date; it does not convert documentation into a negotiated commitment.*

## Part I — Scope the decision before contacting providers

### 1. Workload and data-classification exercise

Start with the business transaction, not a model brand. Two teams using the same API may reach opposite decisions because one sends public marketing copy in batch while the other processes regulated records in a low-latency decision flow.

| Intake field     | Required decision input                                                                                                                        |
|------------------|------------------------------------------------------------------------------------------------------------------------------------------------|
| Business outcome | Decision/action enabled; revenue, safety or compliance consequence; named accountable owner.                                                   |
| Task profile     | Versioned task classes, modalities, context/output lengths, tool calls, structured schemas, human review, unacceptable failures.               |
| Criticality      | Maximum user harm, outage cost, manual workaround, recovery objective, reversibility, and tolerance for degraded model quality.                |
| Traffic          | Average/peak requests, input/output/cached tokens, concurrent streams/jobs, burst shape, ramp rate, seasonality and growth.                    |
| Data             | Public/internal/confidential/restricted/regulated; personal, health, financial, secrets, source code, minors; data subjects and jurisdictions. |
| Lifecycle        | Prompt/output/file/cache/state/log/backup retention; deletion; evaluation/fine-tune use; data-subject and legal-hold requirements.             |
| Location         | Storage, processing, safety review, support access, backups, DR, and subprocessor locations; approved transfer mechanisms.                     |
| Technical        | Required model/version, features, tools, network, keys, identity, logs, API/SDK, quality and latency thresholds.                               |
| Commercial       | Forecast and downside case, accepted-job economics, commitment capacity, budget owner, currency/tax, marketplace/cloud credits.                |
| Exit             | Portable artifacts, alternate provider/model, maximum emergency/planned migration time, termination/deletion evidence.                         |

**CLASSIFICATION RULE** Classify the highest-risk content that can actually reach the provider—including tool results, retrieval context, file uploads, logs, support tickets and safety-review samples. Do not classify only the user-visible prompt.

### 2. Convert requirements into hard gates

A gate is binary for a defined scope: Pass, Fail, or Unknown/Pending. A waiver is not a pass. It requires a named risk owner, rationale, compensating control, expiry date and executive authority appropriate to the impact.

- Apply gates to the exact proposed combination. A provider may pass in one region/tier and fail in another.
- Run legal, privacy, security, capacity and licensing gates before time-consuming quality bake-offs.

- If a required feature is excluded from ZDR, BAA, SLA, private networking or region scope, evaluate a feature-disabled architecture separately.
- Do not average a knockout condition into a high score. A noncompliant service with excellent quality is still noncompliant.

### 3. Evidence hierarchy and scope discipline

| Rank | Evidence class                 | How to use it                                                                                                                                                        |
|------|--------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1    | Verified technical behavior    | Repeatable buyer-run test, configuration export, API response, telemetry, packet/log evidence, or controlled failure exercise for the exact service scope.           |
| 2    | Written contractual commitment | Executed order form, amendment, DPA/BAA, SLA, security schedule, or negotiated term that clearly covers the exact entity, service, model, feature, tier, and region. |
| 3    | Current official documentation | Dated vendor documentation or policy for the exact product. Useful evidence, but changeable unless incorporated into the contract.                                   |
| 4    | Vendor assertion               | Sales response, email, questionnaire answer, roadmap statement, certification logo, or unsupported attestation.                                                      |
| 5    | Assumption or inference        | Buyer interpretation, architectural guess, or extrapolation from a related service. Never treat as a pass without validation.                                        |

**CONFLICT RULE** When evidence conflicts, do not choose the most favorable wording. Reconcile the exact scope. For operational behavior, current buyer-run tests lead. For enforceability, the executed agreement controls. Documentation remains evidence of current design, not a substitute for either.

### 4. Keep five evidence categories visibly separate

| Category                       | Illustrative record                                                                                                         |
|--------------------------------|-----------------------------------------------------------------------------------------------------------------------------|
| Verified technical behavior    | POC shows EU endpoint processed requests in-region and emitted route metadata; logs, timestamps and test code are retained. |
| Written contractual commitment | Executed amendment states EU processing, covered endpoints and exceptions for named projects.                               |
| Current official documentation | Vendor page describes EU processing for supported endpoints on the verification date.                                       |
| Vendor assertion               | Account executive says all features are regional and the contract team will confirm later.                                  |
| Assumption/inference           | Buyer assumes the web-search tool follows the base inference region because it appears in the same API.                     |

## 5. Why labels are not portable across providers

The following examples are current official-documentation claims—not buyer-verified behavior and not negotiated commitments. They demonstrate the need to evaluate the exact path. Always re-check the source ledger and contract at decision time.

### OpenAI API

**Current official documentation:** Official documentation says API data is not used for training by default; abuse-monitoring logs may retain content up to 30 days by default; ZDR/MAM is eligibility- and feature-dependent; some application-state and tool modes have specific exceptions. Regional storage and regional processing are separately described. [SRC06]

**Buyer action:** Request endpoint-level retention/ZDR matrix, project activation proof, regional processing scope and third-party tool treatment; incorporate necessary terms into the executed agreement [SRC07–SRC09].

**Contract status:** Unknown until the buyer maps and executes the applicable order form, DPA/BAA, SLA, security schedule and negotiated amendments for the exact scope.

**Verified behavior:** Unknown until the proposed production account, region, tier, endpoints and features pass the POC and configuration review.

### Anthropic API

**Current official documentation:** Official privacy documentation describes 30-day default API input/output deletion with exceptions and longer retention for flagged content/classifier results. Its ZDR article limits scope and identifies exceptions such as files, explicit caching, some batch behavior, beta products and third-party web destinations. [SRC11–SRC12]

**Buyer action:** Confirm exact API/product/key scope, safety artifacts, file/cache/batch/tool behavior and contractual exception language. Partner-platform lifecycle dates can differ. [SRC13]

**Contract status:** Unknown until the buyer maps and executes the applicable order form, DPA/BAA, SLA, security schedule and negotiated amendments for the exact scope.

**Verified behavior:** Unknown until the proposed production account, region, tier, endpoints and features pass the POC and configuration review.

### Amazon Bedrock

**Current official documentation:** AWS documentation states model providers do not access Bedrock deployment accounts/logs/prompts and describes PrivateLink. Cross-Region inference may route within a geography or globally; Provisioned Throughput uses model units and commitments. The public SLA defines availability and exclusions narrowly. [SRC16–SRC20]

**Buyer action:** Separate hosting boundary from model license, confirm destination-region set, CloudTrail route evidence, exact model/tier support, capacity commitment and what the SLA counts.

**Contract status:** Unknown until the buyer maps and executes the applicable order form, DPA/BAA, SLA, security schedule and negotiated amendments for the exact scope.

**Verified behavior:** Unknown until the proposed production account, region, tier, endpoints and features pass the POC and configuration review.

### Microsoft Foundry / Azure

**Current official documentation:** Microsoft documentation distinguishes global, data-zone and regional processing and standard/provisioned/batch tiers. It describes safety/abuse monitoring and modified monitoring behavior; model retirement is maintained on a platform schedule. [SRC21–SRC23]

**Buyer action:** Pin deployment type and region in architecture/order form; verify monitoring configuration; test capacity/latency; use Microsoft's current retirement schedule rather than assuming the model publisher's dates.

**Contract status:** Unknown until the buyer maps and executes the applicable order form, DPA/BAA, SLA, security schedule and negotiated amendments for the exact scope.

**Verified behavior:** Unknown until the proposed production account, region, tier, endpoints and features pass the POC and configuration review.

## Google Vertex AI

**Current official documentation:** Google documentation says training requires permission/instruction, while ZDR requires attention to abuse logging, feature-specific storage and in-memory caching. Dynamic Shared Quota can return 429 under demand; Provisioned Throughput reserves capacity. The SLA has covered-service and quota/model-lifespan exclusions. [SRC24–SRC29]

**Buyer action:** Confirm exact model/feature exceptions, cache/session settings, region, provisioned model/version support, overage and SLA scope; negotiate any workload-specific gaps.

**Contract status:** Unknown until the buyer maps and executes the applicable order form, DPA/BAA, SLA, security schedule and negotiated amendments for the exact scope.

**Verified behavior:** Unknown until the proposed production account, region, tier, endpoints and features pass the POC and configuration review.

## Part II — Gate, score, and decide

### 6. Hard-gate operating rules

- Pass requires complete, scope-matched evidence. A promise to deliver later is Pending, not Pass.
- Fail removes the proposal from weighted ranking unless an authorized, time-bound waiver changes the requirement—not merely the label.
- Unknown mandatory gates block final Go. Record the owner, evidence requested, due date and the decision that would change.
- Reapply gates after any change to model, feature, endpoint, tool, region, tier, host, subprocessor, data class, contract or AUP.
- Marketplace and reseller routes receive their own gate result because the seller, processor, SLA and support chain may differ from direct purchase.

### 7. Anchored 0–5 scorecard

| Score | Required anchor                                                                                                    |
|-------|--------------------------------------------------------------------------------------------------------------------|
| 0     | Materially unacceptable or contradicted. Verified failure, prohibited condition, or control absent where required. |
| 1     | Major gaps. Capability is ad hoc, preview, narrowly scoped, or dependent on an uncommitted roadmap.                |
| 2     | Partially meets need. Significant compensating controls, exceptions, or commercial changes are required.           |
| 3     | Meets the stated baseline for this workload with manageable residual risk.                                         |
| 4     | Exceeds baseline with strong controls, repeatable evidence, and limited residual risk.                             |

| Score | Required anchor                                                                                                                            |
|-------|--------------------------------------------------------------------------------------------------------------------------------------------|
| 5     | Best-in-class for this workload: verified at target scale and backed by precise contractual commitments with durable operational evidence. |

Score technical/commercial merit only after gates. Score the proposed production configuration, not the vendor in the abstract. A provider can receive different scores for different models, regions, tiers and contracts.

## Default workload profiles

| Criterion                                 | General | Regulated | Latency | Batch | Open/custom |
|-------------------------------------------|---------|-----------|---------|-------|-------------|
| Workload fit and production readiness     | 10      | 6         | 8       | 6     | 15          |
| Security architecture                     | 12      | 15        | 8       | 7     | 10          |
| Privacy, retention, and residency         | 14      | 20        | 8       | 8     | 10          |
| Identity, audit, and compliance assurance | 8       | 12        | 5       | 5     | 7           |
| Reliability, SLA, and support             | 12      | 12        | 14      | 10    | 8           |
| Capacity and performance                  | 10      | 6         | 24      | 16    | 11          |
| Cost and commercial transparency          | 10      | 6         | 8       | 23    | 10          |
| API and model lifecycle maturity          | 8       | 6         | 10      | 7     | 8           |
| Legal, licensing, IP, and AUP fit         | 8       | 10        | 5       | 5     | 12          |
| Resilience, portability, and exit         | 8       | 7         | 10      | 13    | 9           |
| TOTAL                                     | 100     | 100       | 100     | 100   | 100         |

*Weights are starting points, not universal truth. Record every change and rationale before opening vendor bids.*

## 8. Evidence confidence, freshness, and unknowns

| Evidence confidence | Anchor                                                                                                               |
|---------------------|----------------------------------------------------------------------------------------------------------------------|
| 0                   | No evidence, material contradiction, or evidence for the wrong service/scope.                                        |
| 1                   | Sales-only or informal assertion; undated response; roadmap statement; logo without scope.                           |
| 2                   | Current official documentation or vendor artifact, but not independently verified or contractually fixed.            |
| 3                   | Scope-matched official documentation plus one buyer technical check or current independent assurance.                |
| 4                   | Repeatable buyer verification or strong independent assurance plus a clear written commitment.                       |
| 5                   | Verified technical behavior, executed scope-matched commitment, and current official/independent evidence all agree. |
| Freshness           | Anchor                                                                                                               |



| Evidence confidence                                                                                                                                                                                                                                                                                      | Anchor                                                                                        |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------|
| 0                                                                                                                                                                                                                                                                                                        | Undated, expired, or known obsolete.                                                          |
| 1                                                                                                                                                                                                                                                                                                        | Older than 18 months or predates a material service/model change.                             |
| 2                                                                                                                                                                                                                                                                                                        | 12–18 months old; no bridge evidence for current period.                                      |
| 3                                                                                                                                                                                                                                                                                                        | 6–12 months old and no known material change.                                                 |
| 4                                                                                                                                                                                                                                                                                                        | 91 days–6 months old, or current assurance with a valid bridge letter.                        |
| 5                                                                                                                                                                                                                                                                                                        | Revalidated within 90 days, continuously observed, or executed/current as of evaluation date. |
| <b>UNKNOWN RULE</b> Do not enter 0 for missing information unless the known fact is that the requirement is materially unacceptable. Use status Unknown. Exclude unknown weight from the merit-score denominator, disclose evidence coverage, and block Go when a gate or critical criterion is unknown. |                                                                                               |

- Report raw known-item score and evidence coverage together, e.g., 4.1/5 at 72% known weight. Never present 4.1 without the 28% unknown exposure.
- A sales-only answer can receive a provisional merit score but normally confidence 1. It cannot close a hard gate requiring a contract or technical proof.
- Use N/A only when the requirement is genuinely outside scope and record the rationale/approver; do not hide a weakness as N/A.
- Set a predeclared maximum unknown-weight threshold (suggested general default: 15%; regulated critical controls: 0%).

## 9. Sensitivity analysis

1. Run all five default profiles plus the approved custom profile. Note rank reversals and threshold changes.
2. For each top criterion, vary its weight  $\pm 25\%$ , renormalize to 100, and record whether the selection changes.
3. Stress uncertain scores by one point down; set all unknowns to plausible low/base/high values without calling them facts.
4. Stress commercial variables: volume  $\pm 50\%$ , cache hit, output length, tool/reasoning use, failure/retry rate, overage and commitment utilization.
5. Stress operations: loss of the preferred region/model/provider, depreciation inside the planning horizon, and a correlated dependency failure.
6. A robust decision remains eligible and preferred across plausible scenarios. If small changes reverse the result, prefer a staged award, dual-source approach, or more evidence.

## Part III — Prove the production proposition

### 10. Two-to-four-week proof of concept

The POC is a controlled production-readiness experiment, not a demo. Use the exact account, region, deployment type, model identifier, tools, network path, retention controls, and commercial tier proposed for launch. Freeze the test manifest before the run and preserve request IDs, timestamps, usage, model metadata and client telemetry.

| Period                                                 | Objectives                                                                 | Activities                                                                                                                                                                                             | Outputs                                                                                                             |
|--------------------------------------------------------|----------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------|
| Preparation (2–3 days before Week 1)                   | Freeze scope and acceptance criteria; activate legal/security test tenant. | Version task set/prompts/schemas; tag data classes; define load model; provision exact regions/tiers; enable telemetry/budgets; agree test and incident rules.                                         | Signed test charter, dataset manifest/hash, architecture/data-flow draft, baseline budget, gate assumptions.        |
| Week 1 — Functional, quality, and data-flow proof      | Establish correctness and verify the actual service path.                  | Run representative/edge/long-context sets; structured-output and tool tests; record exact model IDs; test storage/delete/ZDR settings; verify logs/audit/region routing; inspect AUP refusals.         | Quality scorecard, schema/tool failure taxonomy, model/version record, data-flow evidence, initial risk/issues log. |
| Week 2 — Performance, capacity, failure, and cost      | Measure production behavior under realistic and adverse load.              | Warm/cold p50/p95/p99 TTFT, output tokens/sec, end-to-end; ramp/spike/soak; 429/timeout/5xx/stream interruption; retry/idempotency; cost per accepted job; usage-to-invoice reconciliation.            | Load report, capacity curve, error/retry profile, accepted-job economics, telemetry reconciliation.                 |
| Week 3 — Resilience, operations, and controlled change | Prove recoverability and operational support.                              | Regional/model/provider failover; correlated-dependency exercise; pinned-versus-alias test; successor canary and rollback; severity support case; security incident tabletop; deletion/export dry run. | RTO/RPO evidence, degradation matrix, upgrade/rollback report, support test, exit artifacts.                        |
| Week 4 — Shadow/canary and decision close (optional)   | Validate in near-production traffic and close unknowns/contract deltas.    | Shadow or low-risk canary with human review; compare quality/cost/latency; resolve evidence gaps; re-run material defects once; finalize gates, scores, risks, contract positions.                     | Canary report, final gate worksheet, confidence/freshness scores, decision memo, remediation/negotiation list.      |

#### Representative, versioned task sets

- Create a manifest containing task ID, input artifact hash, prompt/schema/tool version, expected result or rubric, data class, region rule, model ID, randomness controls and acceptance rule.
- Stratify by business importance, common/long/edge inputs, language, modalities, adversarial/malformed cases, safety-sensitive content, tool outcomes and known failure modes.
- Separate model quality from application acceptance. An answer may be plausible but unusable because the schema, citations, latency, policy or downstream side effect failed.
- Use a blinded or independent review for subjective tasks; report inter-rater method and sample size. Do not tune prompts on the final holdout set.

#### Measurement and default rules

| Metric               | Definition                                                      | Default acceptance rule                     |
|----------------------|-----------------------------------------------------------------|---------------------------------------------|
| Task acceptance rate | Accepted outputs / eligible jobs using pre-agreed task-specific | No critical-task class below its threshold; |

| Metric                         | Definition                                                                                                                      | Default acceptance rule                                                             |
|--------------------------------|---------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|
|                                | rubric; report by task class and version.                                                                                       | statistical interval and sample size stated.                                        |
| Structured-output validity     | Schema-valid outputs without repair / structured-output attempts.                                                               | $\geq 99.5\%$ for production automation unless human validation is mandatory.       |
| Tool success                   | Correct tool selection, valid arguments, successful execution, and correctly incorporated result.                               | Threshold by tool criticality; zero unauthorized side effects.                      |
| TTFT p50/p95/p99               | Client send to first usable streamed token; separate queue/network/model where observable.                                      | All agreed percentiles pass at steady peak and after ramp.                          |
| Output speed                   | Generated output tokens per second after first token, by output-length band.                                                    | Meets user-journey completion target; no severe tail collapse at peak.              |
| End-to-end latency p50/p95/p99 | Client send to accepted completed result, including retries, tools, validation, and repair.                                     | Business SLO measured on accepted jobs, not provider API only.                      |
| Capacity/failure rates         | 429, capacity 503, other 5xx, timeouts, stream truncations, refusals, and retry exhaustion per eligible request.                | Within error budget at peak plus agreed headroom; recovery behavior verified.       |
| Retry/idempotency correctness  | No duplicate external effect or duplicate charge beyond agreed semantics during replay and uncertainty.                         | 100% for side-effecting critical workflows.                                         |
| Cost per accepted job          | All billed input/output/cache/reasoning/tool/batch/retry/failed units divided by accepted business jobs.                        | Within business-case threshold at representative task mix and peak failure profile. |
| Regional/data-flow conformance | Observed routing, storage, retention, logs, and deletion match approved flow.                                                   | Zero unexplained path or retention exception.                                       |
| Failover recovery              | Time to restore minimum acceptable service and quality after injected regional/model/provider failure.                          | RTO/RPO and degraded-service quality/capacity all pass.                             |
| Support effectiveness          | Acknowledgement, correct ownership, technical engagement, update cadence, and resolution/next action for planned severity case. | All contracted targets pass with no ambiguous handoff.                              |
| Billing reconciliation         | Request-level measured usage and calculated price versus export/invoice.                                                        | Difference $\leq 1\%$ or a documented rounding/tax cause accepted by Finance.       |

## Load and failure design

- Measure cold and warm behavior. Report TTFT, output speed and end-to-end latency by model, region, tier, request length, cache state, concurrency and retry count.
- Use ramp, step, spike and soak profiles. Include the expected growth event and  $2\times$  critical-peak headroom where economically relevant.
- Classify 429, capacity 503, other 5xx, timeouts, stream interruption, refusal, malformed response, tool failure and client cancellation separately.
- Implement bounded exponential backoff with jitter and retry budgets. Prove idempotency for any workflow that can create an external side effect.
- Calculate provider API latency and business accepted-job latency. Retries and repair loops belong in the latter.

- Run failover with a real impairment or controlled block, not a PowerPoint tabletop alone; record degradation in quality, capacity, cost and residency.

## Go/no-go criteria

**GO** All mandatory gates Pass; every critical quality/performance/data-flow/failover criterion passes; unknown exposure is within policy; the contract closes material gaps; residual risks have owners and dates; exit is feasible within the migration-time objective.

**CONDITIONAL GO** Only non-knockout gaps remain; each condition is measurable, time-bound, funded and owned; production scope is restricted to controls already proven; automatic stop/rollback criteria are defined.

**NO-GO** Any non-waived gate fails; a critical gate remains unknown; accepted-job quality/capacity/latency/failover misses threshold; data path or billing cannot be reconciled; or exit/concentration risk exceeds appetite.

## Part IV — Contract for the service you actually tested

### 11. Commercial and contract checklist

Public terms establish a baseline, not the negotiated outcome. Link every material obligation to the exact service scope and define precedence among order form, service-specific terms, agreement, DPA/BAA, security schedule, SLA, support schedule and policies.

| Topic                       | Buyer position to consider                                                                                                                                                                     |
|-----------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Covered scope               | Name exact service, API, model/snapshot or version rule, endpoint/tool, tier, deployment type, region, account/project, and production status; order form prevails over generic documentation. |
| Capacity commitment         | State minimum available RPM/input/output TPM/concurrency or reserved units, burst/ramp behavior, effective date, scaling lead time, and remedy for capacity failure.                           |
| Availability and latency    | Use buyer-relevant availability and p95/p99 latency/TTFT SLOs for covered workloads; distinguish streaming, batch, and tools.                                                                  |
| Measurement/reporting       | Define clock, denominator, errors, regions, retries, exclusions, data source, report cadence, audit/recalculation rights, and claim automation.                                                |
| Credits and chronic failure | Credits should not be sole remedy for repeated material failure; add termination, commitment relief, refund, and transition rights after defined breaches.                                     |
| Support and escalation      | 24x7 severity channel, first response/update/restore targets, named technical/executive escalation, RCA timing, and direct ownership across reseller/marketplace.                              |
| Price protection            | Fix rate card/discounts for term; cap/index increases; give advance notice and termination right; protect replacement-model economics.                                                         |
| Commitments and ramp        | Stage minimums to rollout; allow reallocation across approved models/regions/tiers; cap overage; avoid forfeiture; include downward adjustment after deprecation/failure.                      |
| Deprecation/change          | Minimum notice exceeds tested migration time; require overlap, pinned versions, migration support, emergency-change notice, and no material breaking change without remedy.                    |

| Topic                                                                                                                                                                                                                   | Buyer position to consider                                                                                                                                                                    |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Security incident                                                                                                                                                                                                       | Define incident broadly enough; notify promptly from discovery/reasonable belief; provide facts, updates, preservation, cooperation, RCA, and subprocessor flow-down.                         |
| Subprocessors                                                                                                                                                                                                           | Current named list and functions/locations; advance notice; meaningful objection; alternative or termination/refund; obligations flow down.                                                   |
| Data use/retention/deletion                                                                                                                                                                                             | Purpose limitation; no training/improvement without instruction; feature-level retention schedule; ZDR scope; deletion from active/backup/support systems; certificate and legal-hold notice. |
| Audit and assurance                                                                                                                                                                                                     | Annual scoped reports and bridge letters; remediation information; regulator access; targeted audit after incident/material concern, subject to reasonable safeguards.                        |
| IP and ownership                                                                                                                                                                                                        | Buyer input/output and artifact rights; model/license warranties; practical output/training-data indemnity; defense, caps, exclusions, and insurance aligned to exposure.                     |
| Confidentiality                                                                                                                                                                                                         | Customer content and derived prompts/evals/telemetry are confidential; access limited to need; survival and compelled-disclosure process.                                                     |
| Suspension/AUP disputes                                                                                                                                                                                                 | Narrow emergency suspension; notice and cure where safe; limit scope/duration; human escalation and appeal; preserve data export and unaffected service.                                      |
| Termination assistance                                                                                                                                                                                                  | Continued service/export for planned and emergency migration; named transition services, rates, priority, knowledge transfer, and maximum assistance period.                                  |
| Data/fine-tune export                                                                                                                                                                                                   | Usable formats/APIs, complete metadata, encryption, timing, cost cap, verification, post-termination window, and no lien on buyer assets.                                                     |
| Change of control/viability                                                                                                                                                                                             | Notice, assignment controls where negotiable, continuity obligations, termination/refund/export rights, and protection against adverse material reduction.                                    |
| Renewal and exit                                                                                                                                                                                                        | No silent auto-renew at changed rates; workable nonrenewal window; commitment relief; data return/deletion; survival of confidentiality, IP, audit, and assistance duties.                    |
| <b>LEGAL REVIEW</b> This guide is an operational procurement aid, not legal advice. Qualified counsel should review applicable law, negotiated terms, transfer mechanisms, sector obligations, and the final agreement. |                                                                                                                                                                                               |

## SLA: translate provider math into user impact

- Ask for a worked example using a month containing 429s, 5xx, model crashes, stream failures, tool failures, planned maintenance and zero-traffic intervals.
- Keep two views: contractual availability under the vendor formula and buyer user-journey SLI. Contractual compliance does not prove the product met its SLO.
- Require reporting at the same granularity at which the service is sold: account/project, region, model, tier and covered endpoint.
- Credits rarely compensate lost business. Negotiate chronic-failure termination, commitment relief, refund and transition support where leverage permits.

## FinOps: price the accepted business job

- Build a task-mix model using input, cached input, output, reasoning, tool/search/code, storage, fine-tune, batch, provisioned capacity, support and network charges.

- Include failure and retry behavior, cache hit/expiry, long-context multipliers, output-length variance, safety refusals, cancellations and unused commitments.
- Use cost per accepted job and cost per business outcome, not token price alone. A cheaper model can cost more if it needs repair, retries or human review.
- Reconcile request telemetry to usage export and invoice. Treat unreconciled material spend as a control defect, not a rounding issue.

## Part V — Resilience, portability, and exit

### 12. Build an abstraction that preserves useful differences

A provider abstraction should stabilize internal concepts—messages, tools, schemas, streaming events, usage, errors and model metadata—while exposing a capability registry and typed provider extensions. Flattening every provider to the lowest common denominator can destroy the very quality, safety or efficiency that justified the purchase.

- Maintain a portable core request/response type and explicit extension namespaces for reasoning, cache, tools, citations, safety and multimodal features.
- Normalize error classes and telemetry without hiding native request IDs, model IDs, region/tier, retry-after, billed units or safety outcomes.
- Route by task, data class, region, capability, health, capacity and cost policy. Prevent fallback into a prohibited region/model/tool.
- Version adapters and conformance tests. An API-compatible route is not necessarily behavior-, billing-, safety- or license-compatible.

### 13. Exit-plan minimum

| Workstream              | Minimum control                                                                                                                                                          | Evidence                                                     |
|-------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------|
| Portable interface      | Versioned internal request/response types; provider adapters expose native extensions instead of erasing them; conformance tests for common and provider-specific paths. | Adapter contract, schema tests, feature-capability registry. |
| Prompts and evaluations | Prompts, schemas, tool definitions, policies, rubrics, and task sets are versioned independently from provider; immutable dataset manifests/hashes.                      | Repository/version manifest, evaluation run lineage.         |
| Alternative testing     | At least one alternate model/provider tested quarterly on critical tasks, peak capacity assumptions, safety, and cost; known fidelity gaps recorded.                     | Comparative eval, load test, residual-risk acceptance.       |
| Data and fine-tunes     | Canonical training/eval data retained in portable format; export rights, adapter/checkpoint limits, embeddings/vector migration, and re-tuning procedure known.          | Sample export/import and re-tuning estimate.                 |
| Routing and fallback    | Policy-controlled routing by data class/region/task; circuit breakers, retry budgets, backpressure, manual override, and kill switch.                                    | Configuration, runbook, chaos/failover results.              |

| Workstream                       | Minimum control                                                                                                                                                               | Evidence                                                                   |
|----------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------|
| Dependency correlation           | Map model host, cloud/region, identity, DNS/CDN, KMS, SDK, safety, marketplace/billing, support, and supply chain for primary/fallback.                                       | Dependency graph and failure-mode review.                                  |
| Migration triggers               | Trigger thresholds for SLA/capacity failure, price increase, deprecation, security/privacy change, adverse subprocessor/control event, financial distress, or strategic need. | Approved trigger matrix, accountable decision owner.                       |
| Migration time objective         | Maximum acceptable migration time defined for emergency degraded service and planned full migration; task/region capacity lead times included.                                | Timed exercise; capacity reservation/lead-time confirmation.               |
| Emergency versus planned runbook | Emergency path prioritizes minimum safe service; planned path includes full quality, cost, legal, observability, and customer communication gates.                            | Two distinct runbooks and contact trees.                                   |
| Deletion and termination         | Export verified before cutover; credentials revoked; active/backup/support/subprocessor deletion tracked; certificate and exceptions retained; assistance/fees pre-agreed.    | Termination checklist, deletion certificate, final invoice reconciliation. |

## 14. Migration triggers and decision rights

- SLA or user-journey SLO breach beyond the chronic-failure threshold.
- Capacity/latency no longer supports peak workload or required region.
- Security incident, material control regression, unacceptable subprocessor or location change.
- ZDR/retention, AUP, license, IP, privacy or sector eligibility changes.
- Model deprecation or forced upgrade inside the validated migration window.
- Price increase, commitment/overage change or invoice opacity breaks the business case.
- Financial distress, acquisition/change of control, sanctions or supply concentration raises continuity risk.
- Alternate becomes materially safer, better or cheaper after migration cost and risk.

**DECISION RIGHT** Pre-name who may trigger emergency failover, who may accept degraded quality, who may authorize regulated-data routing, and who approves a full migration. A technically available fallback is unusable if no one has authority to activate it.

## Part VI — Govern the relationship as a changing system

### 15. Decision governance

- Name one accountable business owner and one technical service owner. Security, privacy, legal, procurement and finance advise and hold specific gates; they do not replace ownership.
- Keep a decision record containing scope, gates, scores, evidence/confidence/freshness, unknowns, POC manifest/results, contract deltas, residual risks, conditions and sign-offs.

- Record the exact provider route as a configuration item: contracting entity, seller, host, model/version, endpoint/tools, region, tier, projects/accounts, retention settings, networking and keys.
- Require change review for new data classes, regions, tools, models, aliases, deployment types, subprocessors, retention settings or material traffic growth.
- Connect model quality drift, SRE error budgets, security/privacy controls, vendor performance and spend into one service review.

## 16. Maintenance and revalidation

| Cadence                  | Scope                                                                                                                                                                 | Trigger                                                                          | Owner                                                                                 |
|--------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|
| Continuous / per request | Exact model/version/region/tier, latency/error/429/503, usage and cost, route metadata, safety refusals, schema/tool outcomes.                                        | Drift, threshold breach, new alias/version, unexplained invoice variance.        | Platform/SRE/FinOps                                                                   |
| Monthly                  | Provider status/incidents; capacity headroom; spend/commit utilization; invoice reconciliation; open vulnerabilities/risks; fallback readiness.                       | Two adverse trends, chronic SLA event, budget variance, support failure.         | SRE/FinOps/Vendor manager                                                             |
| Quarterly                | Critical task evals against alternate; source-ledger refresh; subprocessor and policy changes; region/feature matrix; access/RBAC review; exit drill component.       | Material documentation/feature/processor change or risk rating increase.         | AI governance/Security/Privacy                                                        |
| Semiannual               | Peak load and failover exercise; model upgrade/rollback; data-flow/ZDR/deletion verification; contract obligation review; concentration/financial update.             | Architecture or workload growth, new regulated data, acquisition/funding event.  | Architecture/SRE/Legal/Procurement                                                    |
| Annual and pre-renewal   | Full gates/scorecard; SOC/ISO/pen-test/bridge evidence; DPA/BAA/SLA/rate card; BCP/exit plan; market check and negotiation positions.                                 | Renewal, material scope expansion, jurisdiction change, or new criticality tier. | Executive sponsor and cross-functional review board                                   |
| Event-driven             | Security incident, new subprocessor, AUP/terms/price change, deprecation, forced migration, repeated outage, control change, regulatory change, or change of control. | Any listed event.                                                                | Named risk owner; convene review within 5 business days or faster for critical events |

**FRESHNESS RULE** A source ledger is a monitoring control, not a bibliography. Re-check volatile pages—model availability, retention exceptions, regions, pricing, quotas, lifecycle, AUP, subprocessors and SLA—at evaluation close, before signature, before production, and on the listed cadence.



## Appendix A — Detailed RFP questionnaire

*Issue these questions with the exact proposed scope prefilled. Require a response owner, evidence link/document ID, evidence class, date, exceptions and contract delta for every domain.*

### R01 Company and service architecture

**Precise questions:** Identify the contracting entity, service operator, control-plane operator, inference-plane operator, and customer-support entity. Provide a component diagram for the exact API, tools, storage, safety, and billing path. Which components are GA, preview, third-party, or customer-managed?

**Expected documentary evidence:** Signed entity chart; current architecture/data-flow diagrams; service catalog; GA/preview status; responsibility matrix; SOC report scope mapping.

**Acceptable answer:** Named legal/operating entities and a current, scope-matched architecture with clear shared-responsibility boundaries and no undisclosed production dependency.

**Warning signs:** Brand-level answer, generic cloud diagram, missing safety/tool path, or preview component presented as production.

**Potential knockout:** Required production component is unsupported/preview or the actual operator cannot be identified.

**Buyer technical proof:** Buyer architecture review; DNS/API endpoint inspection; account and invoice entity reconciliation.

### R02 Actual inference host and subprocessors

**Precise questions:** For each model/region/tier, who physically operates inference and who can access content? Is the model deep-copied into the platform, remotely called, or routed dynamically? List all subprocessors and affiliates for content, telemetry, support, safety, and DR.

**Expected documentary evidence:** Subprocessor list with function/location; hosting attestation; DPA schedule; routing design; support-access locations; change-notice mechanism.

**Acceptable answer:** Host, access boundary, and every processing party/location are named; notice and objection/termination rights are workable.

**Warning signs:** Model publisher assumed to be host; infrastructure provider omitted; list covers only stored data.

**Potential knockout:** Undisclosed processor or required localization cannot be demonstrated.

**Buyer technical proof:** Trace endpoint and region metadata; compare DPA, invoice, trust center, and packet/log evidence.

### R03 Model provenance, version, weights, and license

**Precise questions:** State model publisher, exact immutable identifier, release date, training/provenance disclosure level, weight availability, license/version, modifications, quantization, safety tuning, and indemnified scope. Can identifiers be pinned?

**Expected documentary evidence:** Model card; license text; SBOM/model provenance statement; hash/ID; services and IP terms; modification history.

**Acceptable answer:** Exact version and applicable license are known; permitted uses match the product; provenance limitations are disclosed and risk-accepted.

**Warning signs:** Marketing family name only, mutable alias, license link without version, or unclear derivative-work rights.

**Potential knockout:** License or rights conflict with intended deployment/distribution, or exact model cannot be established.

**Buyer technical proof:** Invoke and log model ID; compare output metadata; preserve license/model-card copies with hashes.

### R04 Model availability by region and tier

**Precise questions:** List required model/snapshot, context, tools, fine-tuning, batch, cache, and structured-output availability by region, deployment type, and billing tier. Is each item GA and covered by SLA/support?

**Expected documentary evidence:** Current region/model/feature matrix; API discovery output; order-form SKU; SLA covered-service schedule.

**Acceptable answer:** All required combinations are GA, orderable, testable, and contractually in scope in every required region.

**Warning signs:** Model exists globally but not in required region/tier; feature availability inferred from another platform.

**Potential knockout:** Any mandatory combination is unavailable, preview-only, or SLA-excluded.

**Buyer technical proof:** Provision in target region/account and execute every required feature.

## R05 Prompt/output training use

**Precise questions:** May prompts, outputs, files, tool results, metadata, feedback, fine-tunes, or evaluation data be used to train, fine-tune, evaluate, improve, or benchmark any model/product? Is any use opt-in, opt-out, de-identified, or permitted for affiliates?

**Expected documentary evidence:** Executed contract/DPA; service-specific terms; settings export; feedback workflow; de-identification definition and retention.

**Acceptable answer:** Use only to provide/security-operate the service; improvement/training requires explicit scoped instruction; no unilateral de-identified-content reuse.

**Warning signs:** Policy-only promise, opt-out control, feedback exception, or broad aggregated/de-identified right.

**Potential knockout:** Vendor may use protected content for improvement/training without explicit buyer instruction.

**Buyer technical proof:** Inspect org/project settings and API flags; verify separate feedback tenant/path.

## R06 Abuse monitoring and application-state retention

**Precise questions:** Map content and metadata retained for abuse, fraud, safety, support, debugging, application state, history, tools, files, cache, batch, and backups. Give trigger, location, access, retention, deletion, and legal-hold behavior for each.

**Expected documentary evidence:** Endpoint retention matrix; data-flow; safety-review SOP; access controls; backup schedule; deletion standard; configuration export.

**Acceptable answer:** Exact retention is bounded and compatible; safety review is minimized, segregated, auditable, and geographically acceptable.

**Warning signs:** Only says 'we do not store prompts' while features retain state; exceptional flagged-content retention omitted.

**Potential knockout:** Unbounded/undisclosed retention or human access conflicts with data classification.

**Buyer technical proof:** Create/delete representative objects; inspect audit/export behavior; test flags and storage controls with synthetic data.

## R07 ZDR scope and exceptions

**Precise questions:** Define ZDR/MAM contractually. Which accounts, projects, endpoints, modalities, tools, caches, batch modes, files, fine-tunes, safety paths, and third parties are eligible or excluded? What changes automatically when enabled?

**Expected documentary evidence:** Signed ZDR/MAM amendment; eligibility matrix; console/API proof; exception list; effective date and tenant identifiers.

**Acceptable answer:** Control is active for exact scope; required features are eligible; exceptions and residual metadata are understood and approved.

**Warning signs:** 'ZDR available' but approval pending; covers base inference only; conflicts with application-state features.

**Potential knockout:** Required feature bypasses ZDR or control cannot be verified before production.

**Buyer technical proof:** Configuration export; request flags; create/retrieve/delete probes; support ticket confirming exact tenant scope.

## R08 Storage versus processing location

**Precise questions:** Separately state at-rest storage, transient inference, caching, safety review, support access, telemetry, backups, and DR location for every deployment type. Can routing cross regions during load or failure?

**Expected documentary evidence:** Location matrix; routing policy; product configuration; DPA; subprocessor locations; DR architecture.

**Acceptable answer:** Every data path stays within required boundary or uses an approved transfer mechanism; 'regional storage' is not represented as regional processing.

**Warning signs:** Uses 'data residency' without defining processing; system/usage data excluded without classification.

**Potential knockout:** Inference or review can occur outside a mandatory boundary.

**Buyer technical proof:** Region-specific endpoint calls; response/routing metadata; audit logs; failure routing exercise.

## R09 Encryption and key management

**Precise questions:** Specify TLS versions/ciphers, at-rest layers, backup encryption, KMS/HSM, key ownership, rotation, revocation, separation of duties, and whether customer-managed keys cover prompts, files, fine-tunes, logs, cache, and replicas.

**Expected documentary evidence:** Crypto standard; key architecture; CMK support matrix; certifications; key-policy/config export; revocation behavior.

**Acceptable answer:** Modern encryption; scoped CMK where required; clear residual provider-managed layers; tested revoke/restore procedure.

**Warning signs:** 'Encrypted at rest' without boundary; CMK covers storage but not inference artifacts/logs.

**Potential knockout:** Mandatory CMK/HSM control absent or key revocation cannot stop required access.

**Buyer technical proof:** Inspect key IDs/policies and audit logs; rotate/revoke in non-production; validate failure mode.

## R10 Private networking

**Precise questions:** Can all inference, management, storage, identity, telemetry, and tool traffic use private endpoints? Describe DNS, routing, egress, public-access disablement, cross-region transit, and third-party tool paths.

**Expected documentary evidence:** Private-connectivity documentation; reference architecture; endpoint policies; firewall/DNS configuration; flow logs.

**Acceptable answer:** Required paths remain private, public access can be disabled, and exceptions are explicit and controlled.

**Warning signs:** Private endpoint covers inference only; management or tool traffic still public.

**Potential knockout:** Policy requires private connectivity and any mandatory component lacks it.

**Buyer technical proof:** Block internet egress; resolve DNS; inspect flow logs/packet destinations; invoke all features.

## R11 Tenant isolation

**Precise questions:** Describe compute, memory, cache, storage, network, key, log, and support isolation. Are accelerators shared? How are memory reuse, side channels, namespace collisions, and cross-tenant leakage tested and remediated?

**Expected documentary evidence:** Isolation architecture; secure-development/testing narrative; scoped SOC/pen-test results; incident history summary.

**Acceptable answer:** Layered logical/cryptographic isolation with independent assurance and a credible test/remediation program.

**Warning signs:** 'Multi-tenant secure by design' with no control mapping or inference-specific analysis.

**Potential knockout:** Known unresolved cross-tenant issue or isolation cannot meet classification requirement.

**Buyer technical proof:** Authorization boundary tests; object-ID guessing tests; cache contamination probes using synthetic canaries.

## R12 RBAC, SSO, SCIM, service identities, and audit logs

**Precise questions:** List roles/permissions for org, project, keys, models, files, billing, retention, and admin changes. Support SAML/OIDC, SCIM, workload identity, short-lived credentials, key restrictions, separation of duties, and immutable log export?

**Expected documentary evidence:** Role matrix; identity docs; SCIM/SSO configuration; service-principal design; audit-event catalog and sample logs.

**Acceptable answer:** Least privilege, lifecycle automation, non-human identities, emergency access controls, and exportable admin/data-access logs.

**Warning signs:** API keys tied to users, coarse admin role, missing retention/config/model events.

**Potential knockout:** Shared credential required or critical actions are unauditably.

**Buyer technical proof:** Provision/deprovision user; enforce roles; rotate service identity; export and reconcile audit events.

## R13 Certifications and exact scope

**Precise questions:** Provide SOC/ISO/PCI/FedRAMP/HITRUST or sector evidence and map legal entity, service, region, data center, controls, period, subservice organizations, and carve-outs to the proposed solution.

**Expected documentary evidence:** Full current reports/certificates; scope statement; bridge letter; CUECs; remediation letter; service-to-control mapping.

**Acceptable answer:** Current evidence covers the exact service and period; exceptions are closed or risk-treated; buyer CUECs are assigned.

**Warning signs:** Logo or certificate only, parent company scope, expired report, or AI service added after audit period.

**Potential knockout:** Mandatory certification is absent or materially out of scope.

**Buyer technical proof:** Auditor/report scope reconciliation; sample control evidence; bridge-period review.

## R14 DPA, BAA, GDPR, and cross-border mechanisms

**Precise questions:** Identify controller/processor roles, purposes, data categories, retention, deletion, assistance, audit, subprocessors, locations, SCC/IDTA modules, supplementary measures, BAA eligibility, and excluded services/features.

**Expected documentary evidence:** Negotiable DPA/BAA; SCC/IDTA; TOMs; transfer/location schedules; eligible-services list; privacy contacts.

**Acceptable answer:** Documents cover exact service/data and provide workable rights, assistance, transfers, deletion, and sector obligations.

**Warning signs:** Generic cloud DPA assumed to cover marketplace model; BAA covers platform but not model/tool.

**Potential knockout:** Required addendum/mechanism unavailable or key feature excluded.

**Buyer technical proof:** Counsel mapping to data flow; verify contracting entity and feature/SKU schedules.

## R15 Security incidents and notification

**Precise questions:** Define incident, discovery, notification clock, content, updates, regulator/data-subject cooperation, forensics, evidence preservation, root cause, credit/remedy, and subprocessor flow-down. Include near-miss and service-integrity events where needed.

**Expected documentary evidence:** Security schedule/DPA clause; incident response plan summary; sample notice; tabletop or incident-communication exercise.

**Acceptable answer:** Prompt notice from discovery or reasonable belief, continuous updates, cooperation, RCA, and preservation duties aligned to buyer obligations.

**Warning signs:** Notification only after vendor confirms breach; no time bound; subprocessor clock starts late.

**Potential knockout:** Notification/cooperation cannot meet legal or operational deadline.

**Buyer technical proof:** Run synthetic incident escalation and measure acknowledgement, channel, completeness, and executive path.

## R16 Vulnerability management and penetration testing

**Precise questions:** Describe SDLC, dependency/model artifact scanning, patch SLAs by severity, bug bounty, external pen tests, inference/API abuse testing, red team, disclosure, and customer testing rights/safe harbor.

**Expected documentary evidence:** Policies; latest executive pen-test/red-team reports; vulnerability aging; remediation proof; testing authorization process.

**Acceptable answer:** Risk-based program, current independent testing, timely closure, and bounded buyer testing rights.

**Warning signs:** Only infrastructure scanning; no AI/API abuse cases; report unavailable even under NDA.

**Potential knockout:** Material overdue critical issue or no credible independent testing for required assurance.

**Buyer technical proof:** Review ticket samples/aging; execute agreed non-destructive API security tests.

## R17 API maturity and versioning

**Precise questions:** Identify stable APIs/SDKs, semantic/version policy, idempotency, pagination, streaming semantics, schema guarantees, error taxonomy, retries, compatibility, changelog, sandbox, and support horizon.

**Expected documentary evidence:** API specification; SDK support matrix; changelog; compatibility policy; error docs; release notes; sample integration tests.

**Acceptable answer:** GA versioned surface, documented errors/retries, backward-compatibility commitment, and testable change process.

**Warning signs:** Preview API in critical path, undocumented streaming edge cases, silent schema changes.

**Potential knockout:** No stable supported interface for a core function.

**Buyer technical proof:** Contract/schema tests; SDK matrix; idempotency and malformed-request tests; upgrade rehearsal.

## R18 Model aliases, snapshots, and deprecations

**Precise questions:** Which identifiers are immutable versus mutable? Who can move an alias? What is the minimum deprecation/retirement notice, overlap, emergency-change right, and behavior after retirement across direct and partner platforms?

**Expected documentary evidence:** Lifecycle policy; current schedule; model metadata; notification terms; negotiated notice/overlap clause.

**Acceptable answer:** Production can pin an immutable version; notice exceeds tested migration time; overlap and exception escalation are defined.

**Warning signs:** 'Latest' alias in production; partner platform schedule assumed equal to model publisher schedule.

**Potential knockout:** Forced change can occur inside maximum migration time without emergency fallback.

**Buyer technical proof:** Record model ID on every response; controlled canary against successor; alert on alias/metadata drift.

## R19 Rate limits, concurrency, burst behavior, and guaranteed capacity

**Precise questions:** State RPM, input/output/cached TPM, concurrent streams/jobs, batch queue, file/tool limits, token-bucket windows, acceleration/ramp limits, burst credits, 429/503 behavior, quota-change process, and guaranteed minimum capacity.

**Expected documentary evidence:** Account quota export/API; capacity order; scheduler policy; limit headers; load-test results; support escalation.

**Acceptable answer:** Limits exceed tested peaks with headroom; committed minimum is distinguished from maximum quota; bursts and ramps are understood.

**Warning signs:** Published maximum presented as guarantee; hidden acceleration limit; quota shared across models/regions unexpectedly.

**Potential knockout:** No path to required peak capacity or unacceptable throttling under accepted traffic profile.

**Buyer technical proof:** Step/ramp/spike/soak tests; record headers, queueing, 429/503, retry-after, and recovered throughput.

## R20 Shared, priority, provisioned, and dedicated options

**Precise questions:** Compare scheduler priority, isolation, committed units, geographic scope, model support, term, overage, failover, latency SLO, SLA, and cancellation for each service tier. Is 'dedicated' compute, capacity, tenant, or billing only?

**Expected documentary evidence:** Tier architecture; SKU/order form; capacity math; SLO/SLA; supported-model matrix; overage and renewal terms.

**Acceptable answer:** Tier semantics and guarantees are explicit; selected option matches load and residency; unused/overage economics are known.

**Warning signs:** Priority means queue preference but not reservation; provisioned capacity excludes tools/aliases.

**Potential knockout:** Required guarantee unavailable for exact model/region/feature.

**Buyer technical proof:** A/B load tests across tier; saturation and overage tests; capacity telemetry reconciliation.

## R21 SLA calculation, exclusions, remedies, and credits

**Precise questions:** Define covered requests, denominator, interval, region/project aggregation, errors, latency, streaming, planned maintenance, quotas, model failures, client errors, dependencies, measurement source, claim window, credits, caps, and chronic-failure termination.

**Expected documentary evidence:** Public and negotiated SLA; worked calculation; sample monthly report/credit claim; order-form covered-service list.

**Acceptable answer:** Business-relevant failures count; measurement is auditable; exclusions narrow; remedies escalate beyond credits for chronic failure.

**Warning signs:** Only 5xx counts; idle intervals count as available; 429/capacity/model crash excluded; credits are sole remedy.

**Potential knockout:** SLA cannot protect the critical service objective and no negotiated alternative exists.

**Buyer technical proof:** Recalculate SLA from POC telemetry using vendor formula and buyer user-journey SLI.

## R22 Incident communications and support escalation

**Precise questions:** Provide severity definitions, 24x7 channels, first-response/restore/update targets, named escalation ladder, status API, language/time-zone coverage, model specialists, RCA timing, and support dependency through reseller/marketplace.

**Expected documentary evidence:** Support schedule; runbook; contact roster; status subscription; sample RCA; reseller RACI.

**Acceptable answer:** Targets match business impact; direct technical escalation and executive path are testable; ownership is clear.

**Warning signs:** Response target only, no restoration/update target; marketplace buyer must relay through multiple parties.

**Potential knockout:** No 24x7 critical path for critical production workload.

**Buyer technical proof:** Open planned severity case; measure acknowledgement, diagnosis quality, handoffs, updates, and closure.

## R23 Regional, model, and provider failover

**Precise questions:** Describe automatic/manual failover across zones, regions, deployment types, model versions, and providers. What data/control plane is shared? What fidelity, capacity, residency, and cost changes occur?

**Expected documentary evidence:** BC/DR architecture; RTO/RPO; routing config; capacity reservations; runbook; exercise results.

**Acceptable answer:** Tested failover meets RTO/RPO, quality, residency, and capacity requirements; degradation is explicit.

**Warning signs:** 'Multi-region' shares global control plane/quota; fallback model untested or incompatible with tools.

**Potential knockout:** No feasible recovery path within maximum outage/migration time.

**Buyer technical proof:** Inject regional/model/provider failure in POC; observe routing, data path, quality, and recovery.

## R24 Correlated failure domains

**Precise questions:** Map shared cloud, region, model publisher, GPU/supply chain, DNS/CDN, identity, KMS, network, safety service, SDK, marketplace, billing, and support dependencies across primary and fallback.

**Expected documentary evidence:** Dependency graph; subprocessor/hosting data; DR design; incident history; control-plane mapping.

**Acceptable answer:** Fallback removes the dominant failure modes or residual correlation is quantified and accepted.

**Warning signs:** Two providers use same underlying model host/cloud/control plane; shared client library or identity remains single point.

**Potential knockout:** Critical fallback is materially correlated and no compensating control meets risk appetite.

**Buyer technical proof:** Failure-mode tabletop and selective dependency blocking; compare endpoint, AS/cloud, identity, and model lineage.

## R25 Usage reporting, logs, traces, budgets, and cost allocation

**Precise questions:** Expose request/model/version/region/tier, input/output/cache/reasoning/tool tokens, latency phases, errors, retries, user/project tags, budgets, alerts, export latency, retention, and auditability. Can customer IDs be supplied without leaking personal data?

**Expected documentary evidence:** Usage API/schema; log sample; cost-export sample; IAM controls; budget/alert docs; retention terms.

**Acceptable answer:** Near-real-time, complete, attributable telemetry supports SRE, security, FinOps, and invoice reconciliation.

**Warning signs:** Dashboard-only aggregates; missing cached/reasoning/tool or failed-request data; export delayed beyond operations need.

**Potential knockout:** Buyer cannot allocate/reconcile material spend or investigate critical incidents.

**Buyer technical proof:** Send tagged synthetic requests; reconcile request logs, usage export, and invoice line items.

## R26 Pricing, minimums, commitments, overages, and invoices

**Precise questions:** Price every input/output/cached/reasoning/tool unit by model/region/tier; state currency, tax, minimums, prepaid credits, ramp, rollover, true-up, overage, discount cliffs, indexation, renewal, termination, and invoice granularity.

**Expected documentary evidence:** Rate card; order form; worked invoices; commitment schedule; price-change terms; marketplace/private-offer documents.

**Acceptable answer:** All units and scenarios are transparent; commitment matches ramp/downside; invoice is reproducible; price protection is bounded.

**Warning signs:** Headline token price omits tools/long context; minimum is take-or-pay with no ramp/reallocation.

**Potential knockout:** Unbounded pricing exposure or commitment downside exceeds approved business case.

**Buyer technical proof:** Replay POC usage through pricing model and reconcile vendor estimate/invoice within agreed tolerance.

## R27 Tool, cache, batch, reasoning, and failed-request billing

**Precise questions:** How are cache writes/reads/expiry, hidden reasoning, long-context multipliers, tool calls/results, web/search/code execution, batch cancellation, retries, partial streams, refusals, safety blocks, 4xx/429/5xx/timeouts, and provider retries billed?

**Expected documentary evidence:** Detailed pricing docs; usage schema; invoice sample; contractual billing definitions; POC reconciliation.

**Acceptable answer:** Billing event and unit are explicit; duplicate/provider-failed work is excluded or capped; all usage is observable.

**Warning signs:** Only input/output list price supplied; hidden tokens/tool fees unavailable per request.

**Potential knockout:** Material cost category is unmeasurable or commercially unbounded.



**Buyer technical proof:** Generate each billing edge case and compare response usage, exports, and invoice.

## R28 Fine-tune, prompt, evaluation-set, and data portability

**Precise questions:** Who owns prompts, adapters, checkpoints, fine-tunes, eval sets, embeddings, and labels? What formats/APIs export them, with what fidelity, encryption, cost, timing, and post-termination window? Can the artifact run elsewhere?

**Expected documentary evidence:** Ownership/portability terms; schema/API; sample export; format docs; deletion schedule; migration assistance.

**Acceptable answer:** Buyer-controlled source data and evaluation assets export in usable formats; limitations on model weights/adapters are explicit and mitigated.

**Warning signs:** 'Your data is yours' but tuned artifact or embeddings are non-exportable; export only before termination.

**Potential knockout:** Non-portability creates unacceptable lock-in for critical intellectual property/workload.

**Buyer technical proof:** Export/import representative dataset, prompts, configs, logs, and fine-tune metadata to alternate stack.

## R29 Output ownership and IP indemnification

**Precise questions:** Who owns input/output and customization artifacts? What output/training-data IP indemnity applies to exact paid service/model/use, with what exclusions, caps, defense control, notice, filters/citations, modification, and trademark/open-source limitations?

**Expected documentary evidence:** Services agreement; service-specific IP terms; order form; model license; insurance; indemnified-service list.

**Acceptable answer:** Rights support intended use and indemnity addresses material exposure without exclusions that erase practical value.

**Warning signs:** Ownership stated 'to extent permitted by law' with no risk allocation; indemnity excludes common transformations.

**Potential knockout:** Required downstream rights absent or IP exposure exceeds approved tolerance.

**Buyer technical proof:** Counsel scenario analysis using representative outputs, modifications, RAG sources, and distribution model.

## R30 Termination, deletion, transition support, and exit rights

**Precise questions:** Define termination for cause/convenience/chronic failure/change; continued access, export window, transition services/rates, data return, deletion from active/backup/support systems, certificate, legal holds, subcontractor deletion, and credential revocation.

**Expected documentary evidence:** Termination/exit schedule; deletion standard; certificate template; transition SOW/rates; export API demonstration.

**Acceptable answer:** Sufficient access and assistance for planned/emergency migration; complete deletion evidence and survival terms.

**Warning signs:** Immediate loss of APIs/exports; deletion 'per policy' without timing/evidence; prepaid amounts forfeited after vendor breach.

**Potential knockout:** Cannot meet maximum migration time or mandatory deletion obligation.

**Buyer technical proof:** Dry-run export, credential revocation, object deletion, and certificate workflow in non-production.

## R31 Change of control, financial viability, and concentration risk

**Precise questions:** Provide ownership/funding/financial resilience, insurance, key dependencies, insolvency/cessation plan, assignment/change-of-control rights, material adverse change notice, continuity obligations, and customer termination/export rights.

**Expected documentary evidence:** Audited/credible financials or credit review; insurance; ownership chart; continuity plan; contractual notice/termination/transition terms.

**Acceptable answer:** Viability supports term; buyer can exit or obtain continuity on adverse control/financial events; concentration is within appetite.



**Warning signs:** High burn/dependency with long take-or-pay commitment; assignment without notice; no continuity plan.

**Potential knockout:** Unmitigated viability/concentration risk threatens critical service and no exit control exists.

**Buyer technical proof:** Procurement/finance assessment; dependency stress test; confirm alternate capacity lead time.

## R32 Marketplace procurement and cloud-credit treatment

**Precise questions:** Identify seller of record, operator, support owner, DPA/BAA path, SLA chain, invoice/tax/currency, private offer, cloud-commit drawdown, credits, discounts, overage, renewal, termination, data access, and portability if marketplace relationship ends.

**Expected documentary evidence:** Marketplace listing/terms; private offer; reseller agreement; RACI; billing examples; cloud-commit confirmation; DPA/SLA incorporation.

**Acceptable answer:** Legal, operational, support, and billing roles are explicit; eligible spend/credits are confirmed in writing; termination preserves exit rights.

**Warning signs:** Cloud provider assumed responsible for third-party model; credits described orally; support bounces between parties.

**Potential knockout:** Required legal/security terms or critical support cannot be obtained through selected channel.

**Buyer technical proof:** Trace sample support and invoice workflow; reconcile seller/operator/processor across documents.

## Appendix B — Evidence-request list

*Use as the secure diligence-room index. Request current, scope-matched evidence and record confidentiality/access limitations.*

**Corporate and contract:** Legal entity/ownership chart; services agreement; order form; service-specific terms; DPA/BAA; security schedule; SLA; support schedule; AUP; IP/licensing terms; insurance; marketplace/private-offer terms.

**Architecture and data flow:** Scope-specific component/data-flow diagrams; inference host/processor map; shared-responsibility matrix; control/data plane; caching/state/tools/safety paths; DR and correlated-dependency maps.

**Privacy and residency:** Retention matrix by endpoint/feature; ZDR/MAM amendment and activation proof; storage/processing/support/backup locations; transfer mechanisms; subprocessors; deletion standard/certificate.

**Security assurance:** SOC 2 report and bridge letter; ISO certificates/scope; CUECs; pen-test/red-team executive reports; vulnerability aging; incident plan/sample notice; encryption/key and private-networking designs.

**Identity and audit:** RBAC permission catalog; SSO/SCIM/workload identity docs; admin/data access event catalog; sample audit export; log retention/integrity; privileged support access process.

**Model and API lifecycle:** Model cards/licenses/provenance; immutable IDs; region/tier feature matrix; API spec/SDK matrix; release notes; deprecation schedule; migration tooling; version-notification mechanism.

**Capacity and resilience:** Account quotas; capacity commitment; scheduler/tier semantics; throughput sizing; historical capacity incidents; RTO/RPO; regional/model/provider failover runbooks and exercise results.

**FinOps and commercial:** Detailed rate card; unit definitions; minimum/ramp/overage/renewal schedule; billing edge cases; sample usage export and invoice; cloud-credit eligibility confirmation; price-protection clause.

**Portability and exit:** Export schemas/APIs and sample exports; fine-tune/data ownership; transition SOW/rates; termination access window; deletion proof; credential revocation; insolvency/change-of-control continuity plan.

## Appendix C — Hard-gate worksheet

Complete separately for each provider route. Status values: Pass, Fail, Unknown/Pending, or Waived with authority/expiry. Any non-waived Fail or critical Unknown blocks Go.

| ID  | Gate                                                   | Pass condition                                                                                                                                                 | Required proof                                                                                                                          | Status |
|-----|--------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------|--------|
| G01 | Required model, feature, region, and production status | Exact model snapshot and every required endpoint/tool are generally available in the required region and purchasable tier.                                     | Region/model availability page, account-console export or API discovery, order-form scope, and a successful production-tier invocation. | —      |
| G02 | Minimum capacity and latency                           | Contracted or demonstrably available capacity meets peak TPM/RPM/concurrency plus headroom, and POC p95/p99 latency passes at peak.                            | Capacity commitment, quota export, load-test report, error distribution, and scaling/escalation procedure.                              | —      |
| G03 | Data processing and retention                          | All content, metadata, application state, logs, caches, backups, and safety-review paths satisfy the classification-specific retention rule.                   | Data-flow diagram, retention matrix by endpoint/feature, configuration export, deletion method, DPA/security schedule.                  | —      |
| G04 | Training and product-improvement restrictions          | No training, tuning, evaluation, product improvement, or human review using buyer content unless explicitly instructed and separately authorized.              | Executed contract/DPA plus product-setting export and exception list.                                                                   | —      |
| G05 | Data residency and cross-border transfer               | Storage, inference processing, safety review, support access, disaster recovery, and subprocessors meet required geographic boundaries and transfer mechanism. | Processing-location matrix, routing configuration, transfer impact assessment inputs, SCC/IDTA or equivalent, subprocessor list.        | —      |
| G06 | ZDR scope and feature exceptions                       | Zero/modified retention applies to the exact organization/project, API, model, tool, cache, batch, file, fine-tune, and safety path used.                      | Signed amendment, console/API configuration, endpoint exception matrix, and deletion/retention test.                                    | —      |
| G07 | Private networking and encryption                      | Required private path, TLS posture, at-rest encryption, and customer-managed key boundary are supported for every in-scope component.                          | Network diagram, private endpoint configuration, DNS/route evidence, key policy, crypto documentation, and packet/egress test.          | —      |
| G08 | DPA, BAA, and sector terms                             | Required executed addenda cover the exact legal entity, service, features, and data categories before production data is used.                                 | Final negotiated documents, scope schedules, eligible-service list, and responsibility matrix.                                          | —      |
| G09 | Acceptable-use compatibility                           | Intended use, users, content, automated actions, jurisdictions, and safety controls are permitted and operationally supportable.                               | Current AUP/service terms, written interpretation for edge cases, suspension/escalation terms, and application control design.          | —      |
| G10 | Licensing and commercial rights                        | Buyer has rights to inputs, outputs, customization artifacts, and intended downstream use; model/open-weight licenses permit deployment and distribution.      | License texts, services agreement, IP schedule, provenance statement, and counsel review.                                               | —      |
| G11 | SLA and support                                        | Covered service/model/tier/region, measurement, exclusions, remedies, claim process, and severity response meet business requirements.                         | Negotiated SLA/support schedule and incident-communications exercise.                                                                   | —      |
| G12 | Model pinning and deprecation                          | Immutable version selection is available where required, with adequate notice, overlap, and migration support.                                                 | Snapshot invocation, lifecycle policy, notification commitment, model retirement schedule, and controlled-upgrade test.                 | —      |
| G13 | Business continuity and exit                           | Alternate provider/model is tested, data/artifacts are exportable, deletion is provable, and transition assistance supports the migration-time objective.      | Exit plan, fallback exercise, export sample, deletion certificate form, and termination-assistance clause.                              | —      |

| ID  | Gate                                            | Pass condition                                                                                                                | Required proof                                                                                                                     | Status |
|-----|-------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------|--------|
| G14 | Tenant isolation, identity, and audit           | Isolation, SSO/SCIM/RBAC, service identities, least privilege, and audit events satisfy the control baseline.                 | Architecture/control narrative, SOC report scope, role matrix, configuration export, and audit-log sample.                         | —      |
| G15 | Security incident and vulnerability obligations | Notification, cooperation, forensics, vulnerability remediation, and pen-test assurance meet policy and regulatory timelines. | Security schedule, incident clause, latest scoped reports, remediation summary, and tabletop evidence.                             | —      |
| G16 | Financial and concentration viability           | Vendor and route-to-market are financially viable; concentration and change-of-control risks have controls and exit triggers. | Financial review, dependency map, marketplace/reseller terms, insurance, escrow/continuity options, and executive risk acceptance. | —      |

Waiver record

## Appendix D — Weighted scorecard

Score only eligible proposals. Keep technical merit, evidence confidence, freshness and unknown coverage separate.

| Score | Anchor                                                                                                                                     |     |     |     |       |      |       |                       |       |       |
|-------|--------------------------------------------------------------------------------------------------------------------------------------------|-----|-----|-----|-------|------|-------|-----------------------|-------|-------|
| 0     | Materially unacceptable or contradicted. Verified failure, prohibited condition, or control absent where required.                         |     |     |     |       |      |       |                       |       |       |
| 1     | Major gaps. Capability is ad hoc, preview, narrowly scoped, or dependent on an uncommitted roadmap.                                        |     |     |     |       |      |       |                       |       |       |
| 2     | Partially meets need. Significant compensating controls, exceptions, or commercial changes are required.                                   |     |     |     |       |      |       |                       |       |       |
| 3     | Meets the stated baseline for this workload with manageable residual risk.                                                                 |     |     |     |       |      |       |                       |       |       |
| 4     | Exceeds baseline with strong controls, repeatable evidence, and limited residual risk.                                                     |     |     |     |       |      |       |                       |       |       |
| 5     | Best-in-class for this workload: verified at target scale and backed by precise contractual commitments with durable operational evidence. |     |     |     |       |      |       |                       |       |       |
| ID    | Criterion                                                                                                                                  | Gen | Reg | Lat | Batch | Open | Score | Status                | Conf. | Fresh |
| S01   | Workload fit and production readiness                                                                                                      | 10  | 6   | 8   | 6     | 15   | —     | Known / Unknown / N/A | —     | —     |
| S02   | Security architecture                                                                                                                      | 12  | 15  | 8   | 7     | 10   | —     | Known / Unknown / N/A | —     | —     |
| S03   | Privacy, retention, and residency                                                                                                          | 14  | 20  | 8   | 8     | 10   | —     | Known / Unknown / N/A | —     | —     |
| S04   | Identity, audit, and compliance assurance                                                                                                  | 8   | 12  | 5   | 5     | 7    | —     | Known / Unknown / N/A | —     | —     |
| S05   | Reliability, SLA, and support                                                                                                              | 12  | 12  | 14  | 10    | 8    | —     | Known / Unknown / N/A | —     | —     |
| S06   | Capacity and performance                                                                                                                   | 10  | 6   | 24  | 16    | 11   | —     | Known / Unknown / N/A | —     | —     |
| S07   | Cost and commercial transparency                                                                                                           | 10  | 6   | 8   | 23    | 10   | —     | Known / Unknown / N/A | —     | —     |
| S08   | API and model lifecycle maturity                                                                                                           | 8   | 6   | 10  | 7     | 8    | —     | Known / Unknown / N/A | —     | —     |
| S09   | Legal, licensing, IP, and AUP fit                                                                                                          | 8   | 10  | 5   | 5     | 12   | —     | Known / Unknown / N/A | —     | —     |
| S10   | Resilience, portability, and exit                                                                                                          | 8   | 7   | 10  | 13    | 9    | —     | Known / Unknown / N/A | —     | —     |

**Report:** Known-item weighted merit: \_\_\_\_ / 5. Evidence coverage: \_\_\_\_%. Weighted confidence: \_\_\_\_ / 5. Weighted freshness: \_\_\_\_ / 5. Unknown-weight exposure: \_\_\_\_%. Gate result: \_\_\_\_.

**Sensitivity result:** Profiles/scenarios tested: \_\_\_\_\_. Rank reversals or threshold changes: \_\_\_\_\_. Decision implication: \_\_\_\_\_.

## Appendix E — Risk register template

Use likelihood and impact scales approved by your organization. Preserve the causal statement, trigger/indicator, treatment, owner, due date and evidence of closure.

| ID   | Risk                                           | Cause                                                                                | Impact                                                      | Indicator                                             | Mitigation                                                                           | Owner                  |
|------|------------------------------------------------|--------------------------------------------------------------------------------------|-------------------------------------------------------------|-------------------------------------------------------|--------------------------------------------------------------------------------------|------------------------|
| RR01 | Required feature is preview or not SLA-covered | Marketing availability differs from exact region/tier/endpoint status                | Production outage, forced redesign, unsupported workload    | Preview label, SLA exclusion, no order-form SKU       | Gate on GA status; pin exact scope in order form; alternate design                   | Platform + Procurement |
| RR02 | Residual content retention despite ZDR label   | Files, cache, batch, tools, application state, safety, or third parties are excluded | Privacy/regulatory breach                                   | Exception matrix incomplete; control pending approval | Feature-level retention map; disable incompatible features; contract amendment; test | Privacy + Security     |
| RR03 | Processing crosses required boundary           | Global routing, support access, safety review, or DR differs from storage region     | Localization/transfer noncompliance                         | 'Regional storage' language; global deployment type   | Separate storage/processing gates; regional endpoint; approved transfer controls     | Privacy + Legal        |
| RR04 | Capacity shortfall at peak                     | Quota treated as guarantee; shared scheduler; ramp limit                             | 429/503, latency breach, lost transactions                  | No minimum capacity commitment; load-test throttling  | Provisioned/committed capacity; headroom; queue/backpressure; tested fallback        | SRE + Product          |
| RR05 | SLA fails to count business-impacting errors   | Only 5xx covered; 429/model/stream/tool failures excluded                            | No remedy despite user outage                               | Narrow error definition; idle time counted available  | Negotiate workload SLO/SLI, chronic-failure right, and reporting                     | SRE + Legal            |
| RR06 | Mutable model changes behavior                 | Alias moves or emergency safety change                                               | Quality regression, control failure, inconsistent decisions | No immutable ID; change notice absent                 | Pin snapshots; canary evals; drift alerts; rollback/fallback                         | ML Platform            |
| RR07 | Correlated failover                            | Primary and fallback share cloud, host, identity, model lineage, or control plane    | Simultaneous outage                                         | Same underlying host/subprocessor or global quota     | Dependency map; diversify critical path; reserve alternate capacity                  | Architecture + SRE     |
| RR08 | Unit-cost surprise                             | Reasoning/tool/cache/long-context/retry/failed-request billing omitted               | Budget overrun and weak margins                             | Invoice cannot reconcile to request telemetry         | Accepted-job cost metric; edge-case billing test; caps/alerts; rate-card schedule    | FinOps + Finance       |
| RR09 | Unportable customization assets                | Fine-tune/adaptor/embedding formats are proprietary or non-exportable                | Lock-in and long migration                                  | Ownership language without export mechanism           | Keep canonical datasets/prompts/evals; test export; alternative tuning plan          | ML Platform + Legal    |
| RR10 | AUP suspension interrupts legitimate use       | Ambiguous automated safety enforcement or policy change                              | Service interruption, customer harm                         | No notice/appeal/escalation; edge use case            | Written use-case confirmation; safety controls; cure/appeal and emergency route      | Product + Legal        |
| RR11 | Certification scope mismatch                   | Report covers parent/cloud but not inference service/region/period                   | Control gap and audit finding                               | Logo-only proof; service absent from scope            | Scope mapping; bridge letter; compensating review; contractual audit evidence        | GRC                    |
| RR12 | Vendor financial or control event              | Insolvency, acquisition, sanctions, supply dependency                                | Price/service/terms change or cessation                     | Funding stress, concentration, assignment notice      | Exit triggers; short renewal; alternative capacity; transition assistance            | Procurement + Finance  |

## Appendix F — Data-flow and subprocessor worksheet

*Create one row per logical path and feature. If a tool, cache, file store, safety path or support workflow changes location/processor/retention, give it a separate flow ID.*

Flow ID: \_\_\_\_\_

Use case/task: \_\_\_\_\_

Data element and classification: \_\_\_\_\_

Data subjects/jurisdiction: \_\_\_\_\_

Source system/controller: \_\_\_\_\_

Buyer component: \_\_\_\_\_

Provider contracting entity: \_\_\_\_\_

Actual inference host: \_\_\_\_\_

Model publisher/version/license: \_\_\_\_\_

Endpoint/feature/tool: \_\_\_\_\_

Deployment type/tier: \_\_\_\_\_

Storage location: \_\_\_\_\_

Inference processing location: \_\_\_\_\_

Safety/human review location: \_\_\_\_\_

Support access location: \_\_\_\_\_

Cache/state/retention: \_\_\_\_\_

Subprocessors and purpose: \_\_\_\_\_

Encryption/key boundary: \_\_\_\_\_

Private network path: \_\_\_\_\_

Legal role and transfer mechanism: \_\_\_\_\_

Deletion method/evidence: \_\_\_\_\_

Owner: \_\_\_\_\_

Evidence URL/document: \_\_\_\_\_

Verified date: \_\_\_\_\_

Residual risk/decision: \_\_\_\_\_

## Appendix G — Proof-of-concept plan

*Tailor thresholds before testing. Preserve versioned manifests, telemetry and evidence so another reviewer can reproduce the decision.*

| Period                                                 | Objectives                                                                                                       | Activities                                                                                                                                                                                             | Outputs                                                                                                             |
|--------------------------------------------------------|------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------|
| Preparation (2–3 days before Week 1)                   | Freeze scope and acceptance criteria; activate legal/security test tenant.                                       | Version task set/prompts/schemas; tag data classes; define load model; provision exact regions/tiers; enable telemetry/budgets; agree test and incident rules.                                         | Signed test charter, dataset manifest/hash, architecture/data-flow draft, baseline budget, gate assumptions.        |
| Week 1 — Functional, quality, and data-flow proof      | Establish correctness and verify the actual service path.                                                        | Run representative/edge/long-context sets; structured-output and tool tests; record exact model IDs; test storage/delete/ZDR settings; verify logs/audit/region routing; inspect AUP refusals.         | Quality scorecard, schema/tool failure taxonomy, model/version record, data-flow evidence, initial risk/issues log. |
| Week 2 — Performance, capacity, failure, and cost      | Measure production behavior under realistic and adverse load.                                                    | Warm/cold p50/p95/p99 TTFT, output tokens/sec, end-to-end; ramp/spike/soak; 429/timeout/5xx/stream interruption; retry/idempotency; cost per accepted job; usage-to-invoice reconciliation.            | Load report, capacity curve, error/retry profile, accepted-job economics, telemetry reconciliation.                 |
| Week 3 — Resilience, operations, and controlled change | Prove recoverability and operational support.                                                                    | Regional/model/provider failover; correlated-dependency exercise; pinned-versus-alias test; successor canary and rollback; severity support case; security incident tabletop; deletion/export dry run. | RTO/RPO evidence, degradation matrix, upgrade/rollback report, support test, exit artifacts.                        |
| Week 4 — Shadow/canary and decision close (optional)   | Validate in near-production traffic and close unknowns/contract deltas.                                          | Shadow or low-risk canary with human review; compare quality/cost/latency; resolve evidence gaps; re-run material defects once; finalize gates, scores, risks, contract positions.                     | Canary report, final gate worksheet, confidence/freshness scores, decision memo, remediation/negotiation list.      |
| Metric                                                 | Definition                                                                                                       | Default rule                                                                                                                                                                                           |                                                                                                                     |
| Task acceptance rate                                   | Accepted outputs / eligible jobs using pre-agreed task-specific rubric; report by task class and version.        | No critical-task class below its threshold; statistical interval and sample size stated.                                                                                                               |                                                                                                                     |
| Structured-output validity                             | Schema-valid outputs without repair / structured-output attempts.                                                | >=99.5% for production automation unless human validation is mandatory.                                                                                                                                |                                                                                                                     |
| Tool success                                           | Correct tool selection, valid arguments, successful execution, and correctly incorporated result.                | Threshold by tool criticality; zero unauthorized side effects.                                                                                                                                         |                                                                                                                     |
| TTFT p50/p95/p99                                       | Client send to first usable streamed token; separate queue/network/model where observable.                       | All agreed percentiles pass at steady peak and after ramp.                                                                                                                                             |                                                                                                                     |
| Output speed                                           | Generated output tokens per second after first token, by output-length band.                                     | Meets user-journey completion target; no severe tail collapse at peak.                                                                                                                                 |                                                                                                                     |
| End-to-end latency p50/p95/p99                         | Client send to accepted completed result, including retries, tools, validation, and repair.                      | Business SLO measured on accepted jobs, not provider API only.                                                                                                                                         |                                                                                                                     |
| Capacity/failure rates                                 | 429, capacity 503, other 5xx, timeouts, stream truncations, refusals, and retry exhaustion per eligible request. | Within error budget at peak plus agreed headroom; recovery behavior verified.                                                                                                                          |                                                                                                                     |
| Retry/idempotency correctness                          | No duplicate external effect or duplicate charge beyond agreed semantics during replay and uncertainty.          | 100% for side-effecting critical workflows.                                                                                                                                                            |                                                                                                                     |
| Cost per accepted job                                  | All billed input/output/cache/reasoning/tool/batch/retry/failed units divided by accepted business jobs.         | Within business-case threshold at representative task mix and peak failure profile.                                                                                                                    |                                                                                                                     |



| Period                         | Objectives                                                                                                                      | Activities | Outputs                                                                 |
|--------------------------------|---------------------------------------------------------------------------------------------------------------------------------|------------|-------------------------------------------------------------------------|
| Regional/data-flow conformance | Observed routing, storage, retention, logs, and deletion match approved flow.                                                   |            | Zero unexplained path or retention exception.                           |
| Failover recovery              | Time to restore minimum acceptable service and quality after injected regional/model/provider failure.                          |            | RTO/RPO and degraded-service quality/capacity all pass.                 |
| Support effectiveness          | Acknowledgement, correct ownership, technical engagement, update cadence, and resolution/next action for planned severity case. |            | All contracted targets pass with no ambiguous handoff.                  |
| Billing reconciliation         | Request-level measured usage and calculated price versus export/invoice.                                                        |            | Difference <=1% or a documented rounding/tax cause accepted by Finance. |

**POC scope IDs:** Provider/contracting entity: \_\_\_\_\_ Account/project: \_\_\_\_\_ Model/version: \_\_\_\_\_

Region/tier/deployment: \_\_\_\_\_ Endpoints/tools: \_\_\_\_\_

**Final result:** Critical criteria passed: \_\_\_\_ / \_\_\_\_ Material defects: \_\_\_\_\_ Corrective pass performed: \_\_\_\_\_ Evidence archive: \_\_\_\_\_

## Appendix H — Go/no-go decision memo

*Keep the memo short enough for accountable executives to read, while linking to the full evidence archive.*

**Decision requested:** Go / Conditional Go / No-Go; provider route and scope; requested date.

---



---

**Business context:** Outcome, criticality, users, data classes, jurisdictions, traffic and financial case.

---



---

**Recommendation:** Selected route or no award; why it fits this workload; why alternatives were not chosen.

---



---

**Hard gates:** Pass/Fail/Unknown/Waived summary; list every non-Pass item and authority.

---



---

**Score and uncertainty:** Merit by profile; evidence coverage; confidence/freshness; sensitivity and unknown-weight exposure.

---



---

**POC evidence:** Quality, structured/tool reliability, p50/p95/p99, peak capacity, failures/retries, cost/accepted job, routing, failover, support and billing.

---



---

**Contract position:** Closed terms, open deltas, price/commitment, SLA/support, data/IP, lifecycle and exit.

---



---

**Top residual risks:** Risk, owner, treatment, due date, trigger and consequence.

---



---

**Conditions and stop criteria:** Production scope restrictions, pre-launch deliverables, canary limits, rollback/termination triggers.

---



---

**Continuity and exit:** Fallback, correlation, RTO/RPO, maximum migration time, export/deletion readiness.

---



---

**Approvals:** Business owner, technology, security, privacy, legal, procurement, finance/compliance as required; dates and conditions.

---



---

## Appendix I — Contract-negotiation checklist

*Counsel must tailor these positions to jurisdiction, leverage, sector and the complete agreement. This is not legal advice.*

### Service, capacity, support and change

| Topic                       | Buyer position to consider                                                                                                                                                                     | Status / owner / delta |
|-----------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------|
| Covered scope               | Name exact service, API, model/snapshot or version rule, endpoint/tool, tier, deployment type, region, account/project, and production status; order form prevails over generic documentation. | _____<br>–             |
| Capacity commitment         | State minimum available RPM/input/output TPM/concurrency or reserved units, burst/ramp behavior, effective date, scaling lead time, and remedy for capacity failure.                           | _____<br>–             |
| Availability and latency    | Use buyer-relevant availability and p95/p99 latency/TTFT SLOs for covered workloads; distinguish streaming, batch, and tools.                                                                  | _____<br>–             |
| Measurement/reporting       | Define clock, denominator, errors, regions, retries, exclusions, data source, report cadence, audit/recalculation rights, and claim automation.                                                | _____<br>–             |
| Credits and chronic failure | Credits should not be sole remedy for repeated material failure; add termination, commitment relief, refund, and transition rights after defined breaches.                                     | _____<br>–             |
| Support and escalation      | 24x7 severity channel, first response/update/restore targets, named technical/executive escalation, RCA timing, and direct ownership across reseller/marketplace.                              | _____<br>–             |
| Price protection            | Fix rate card/discounts for term; cap/index increases; give advance notice and termination right; protect replacement-model economics.                                                         | _____<br>–             |
| Commitments and ramp        | Stage minimums to rollout; allow reallocation across approved models/regions/tiers; cap overage; avoid forfeiture; include downward adjustment after deprecation/failure.                      | _____<br>–             |
| Deprecation/change          | Minimum notice exceeds tested migration time; require overlap, pinned versions, migration support, emergency-change notice, and no material breaking change without remedy.                    | _____<br>–             |
| Security incident           | Define incident broadly enough; notify promptly from discovery/reasonable belief; provide facts, updates, preservation, cooperation, RCA, and subprocessor flow-down.                          | _____<br>–             |

## Data, rights, suspension, transition and renewal

| Topic                       | Buyer position to consider                                                                                                                                                                    | Status / owner / delta |
|-----------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------|
| Subprocessors               | Current named list and functions/locations; advance notice; meaningful objection; alternative or termination/refund; obligations flow down.                                                   | _____<br>–             |
| Data use/retention/deletion | Purpose limitation; no training/improvement without instruction; feature-level retention schedule; ZDR scope; deletion from active/backup/support systems; certificate and legal-hold notice. | _____<br>–             |
| Audit and assurance         | Annual scoped reports and bridge letters; remediation information; regulator access; targeted audit after incident/material concern, subject to reasonable safeguards.                        | _____<br>–             |
| IP and ownership            | Buyer input/output and artifact rights; model/license warranties; practical output/training-data indemnity; defense, caps, exclusions, and insurance aligned to exposure.                     | _____<br>–             |
| Confidentiality             | Customer content and derived prompts/evals/telemetry are confidential; access limited to need; survival and compelled-disclosure process.                                                     | _____<br>–             |
| Suspension/AUP disputes     | Narrow emergency suspension; notice and cure where safe; limit scope/duration; human escalation and appeal; preserve data export and unaffected service.                                      | _____<br>–             |
| Termination assistance      | Continued service/export for planned and emergency migration; named transition services, rates, priority, knowledge transfer, and maximum assistance period.                                  | _____<br>–             |
| Data/fine-tune export       | Usable formats/APIs, complete metadata, encryption, timing, cost cap, verification, post-termination window, and no lien on buyer assets.                                                     | _____<br>–             |
| Change of control/viability | Notice, assignment controls where negotiable, continuity obligations, termination/refund/export rights, and protection against adverse material reduction.                                    | _____<br>–             |
| Renewal and exit            | No silent auto-renew at changed rates; workable nonrenewal window; commitment relief; data return/deletion; survival of confidentiality, IP, audit, and assistance duties.                    | _____<br>–             |

## Appendix J — Business-continuity and exit-plan template

*Test this plan before production and at least semiannually for critical services. Keep emergency degraded service distinct from planned migration.*

| Workstream                       | Minimum control                                                                                                                                                               | Evidence                                                                   | Owner/<br>status |
|----------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------|------------------|
| Portable interface               | Versioned internal request/response types; provider adapters expose native extensions instead of erasing them; conformance tests for common and provider-specific paths.      | Adapter contract, schema tests, feature-capability registry.               | _____            |
| Prompts and evaluations          | Prompts, schemas, tool definitions, policies, rubrics, and task sets are versioned independently from provider; immutable dataset manifests/hashes.                           | Repository/version manifest, evaluation run lineage.                       | _____            |
| Alternative testing              | At least one alternate model/provider tested quarterly on critical tasks, peak capacity assumptions, safety, and cost; known fidelity gaps recorded.                          | Comparative eval, load test, residual-risk acceptance.                     | _____            |
| Data and fine-tunes              | Canonical training/eval data retained in portable format; export rights, adapter/checkpoint limits, embeddings/vector migration, and re-tuning procedure known.               | Sample export/import and re-tuning estimate.                               | _____            |
| Routing and fallback             | Policy-controlled routing by data class/region/task; circuit breakers, retry budgets, backpressure, manual override, and kill switch.                                         | Configuration, runbook, chaos/failover results.                            | _____            |
| Dependency correlation           | Map model host, cloud/region, identity, DNS/CDN, KMS, SDK, safety, marketplace/billing, support, and supply chain for primary/fallback.                                       | Dependency graph and failure-mode review.                                  | _____            |
| Migration triggers               | Trigger thresholds for SLA/capacity failure, price increase, deprecation, security/privacy change, adverse subprocessor/control event, financial distress, or strategic need. | Approved trigger matrix, accountable decision owner.                       | _____            |
| Migration time objective         | Maximum acceptable migration time defined for emergency degraded service and planned full migration; task/region capacity lead times included.                                | Timed exercise; capacity reservation/lead-time confirmation.               | _____            |
| Emergency versus planned runbook | Emergency path prioritizes minimum safe service; planned path includes full quality, cost, legal, observability, and customer communication gates.                            | Two distinct runbooks and contact trees.                                   | _____            |
| Deletion and termination         | Export verified before cutover; credentials revoked; active/backup/support/subprocessor deletion tracked; certificate and exceptions retained; assistance/fees pre-agreed.    | Termination checklist, deletion certificate, final invoice reconciliation. | _____            |

## Exit-plan operating parameters

Primary route and critical dependencies:

---

Fallback route and residual correlation:

---

Emergency minimum service: \_\_\_\_\_

Planned target state: \_\_\_\_\_

Emergency migration-time objective:

---

Planned migration-time objective: \_\_\_\_\_

Activation authority: \_\_\_\_\_

Data/export sequence: \_\_\_\_\_

Credential and traffic cutover: \_\_\_\_\_

Customer/regulator communications:

---

Deletion and certificate: \_\_\_\_\_

Final invoice/commitment settlement:

---

Exercise date and defects: \_\_\_\_\_

## Appendix K — Source ledger

*Provider-specific claims in this guide are limited to current official sources. Re-check volatile sources before signature and production; capture contractual documents separately.*

| ID     | Publisher                     | Source                                                                        | Type                             | Use                                                                                                                  | Verified   | URL                                                                                                                                                                                                                                                                                         |
|--------|-------------------------------|-------------------------------------------------------------------------------|----------------------------------|----------------------------------------------------------------------------------------------------------------------|------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| SRC 01 | NIST                          | AI Risk Management Framework                                                  | Government framework             | Govern-map-measure-manage lifecycle and continuous risk management.                                                  | 2026-08-17 | <a href="https://www.nist.gov/itl/ai-risk-management-framework">https://www.nist.gov/itl/ai-risk-management-framework</a>                                                                                                                                                                   |
| SRC 02 | NIST                          | Generative AI Profile (NIST AI 600-1)                                         | Government guidance              | GenAI-specific risks and testing/evaluation actions.                                                                 | 2026-08-17 | <a href="https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf">https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf</a>                                                                                                                                                                 |
| SRC 03 | HHS OCR                       | Guidance on HIPAA & Cloud Computing                                           | Regulator guidance               | CSP business-associate status, BAA need, risk analysis, and SLA considerations.                                      | 2026-08-17 | <a href="https://www.hhs.gov/hipaa/for-professionals/special-topics/health-information-technology/cloud-computing/index.html">https://www.hhs.gov/hipaa/for-professionals/special-topics/health-information-technology/cloud-computing/index.html</a>                                       |
| SRC 04 | HHS OCR                       | Business Associate Contracts                                                  | Regulator guidance               | Required BAA provisions including safeguards, subcontractors, return/destruction, and termination.                   | 2026-08-17 | <a href="https://www.hhs.gov/hipaa/for-professionals/covered-entities/sample-business-associate-agreement-provisions/index.html">https://www.hhs.gov/hipaa/for-professionals/covered-entities/sample-business-associate-agreement-provisions/index.html</a>                                 |
| SRC 05 | European Commission / EUR-Lex | Commission Implementing Decision (EU) 2021/914 — Standard Contractual Clauses | Law/transfer mechanism           | EU SCC mechanism for transfers to third countries.                                                                   | 2026-08-17 | <a href="https://eur-lex.europa.eu/eli/dec_impl/2021/914/oj">https://eur-lex.europa.eu/eli/dec_impl/2021/914/oj</a>                                                                                                                                                                         |
| SRC 06 | OpenAI                        | Data controls in the OpenAI platform                                          | Official product documentation   | Training default, abuse-monitoring retention, ZDR/MAM scope, application-state and regional processing exceptions.   | 2026-08-17 | <a href="https://platform.openai.com/docs/models/default-usage-policies-by-endpoint">https://platform.openai.com/docs/models/default-usage-policies-by-endpoint</a>                                                                                                                         |
| SRC 07 | OpenAI                        | OpenAI Services Agreement                                                     | Published contract terms         | Content rights/use, DPA/healthcare scope, updates, audit reports, order-of-precedence, and usage limits.             | 2026-08-17 | <a href="https://openai.com/policies/services-agreement/">https://openai.com/policies/services-agreement/</a>                                                                                                                                                                               |
| SRC 08 | OpenAI                        | Service terms                                                                 | Published service terms          | Service-specific conditions and IP indemnification scope.                                                            | 2026-08-17 | <a href="https://openai.com/policies/service-terms/">https://openai.com/policies/service-terms/</a>                                                                                                                                                                                         |
| SRC 09 | OpenAI                        | Sub-processor list                                                            | Official processor disclosure    | Named processing entities, functions, products, and locations.                                                       | 2026-08-17 | <a href="https://openai.com/policies/sub-processor-list/">https://openai.com/policies/sub-processor-list/</a>                                                                                                                                                                               |
| SRC 10 | OpenAI                        | All API models                                                                | Official product documentation   | Current model catalog, status, capabilities, and deprecation labels.                                                 | 2026-08-17 | <a href="https://developers.openai.com/api/docs/models/all">https://developers.openai.com/api/docs/models/all</a>                                                                                                                                                                           |
| SRC 11 | Anthropic                     | How long do you store my organization's data?                                 | Official privacy documentation   | Commercial API default retention and longer flagged-content/classifier retention exceptions.                         | 2026-08-17 | <a href="https://privacy.anthropic.com/en/articles/7996866-how-long-do-you-store-my-organization-s-data">https://privacy.anthropic.com/en/articles/7996866-how-long-do-you-store-my-organization-s-data</a>                                                                                 |
| SRC 12 | Anthropic                     | Zero data retention agreement scope                                           | Official privacy documentation   | ZDR product scope and exclusions for files, caches, batch, beta, third-party web use, and other products.            | 2026-08-17 | <a href="https://privacy.anthropic.com/en/articles/8956058-i-have-a-zero-data-retention-agreement-with-anthropic-what-products-does-it-apply-to">https://privacy.anthropic.com/en/articles/8956058-i-have-a-zero-data-retention-agreement-with-anthropic-what-products-does-it-apply-to</a> |
| SRC 13 | Anthropic                     | Model deprecations                                                            | Official lifecycle documentation | Lifecycle definitions, partner-platform schedule differences, retirement effect, and stated notice practice.         | 2026-08-17 | <a href="https://platform.claude.com/docs/en/about-claude/model-deprecations">https://platform.claude.com/docs/en/about-claude/model-deprecations</a>                                                                                                                                       |
| SRC 14 | Anthropic                     | Rate limits                                                                   | Official capacity documentation  | Rate/spend limit semantics, token bucket, acceleration limits, headers, and limits as maxima rather than guarantees. | 2026-08-17 | <a href="https://platform.claude.com/docs/en/api/rate-limits">https://platform.claude.com/docs/en/api/rate-limits</a>                                                                                                                                                                       |
| SRC 15 | Anthropic                     | Service tiers                                                                 | Official service-tier            | Standard/batch/priority behavior and current purchase availability                                                   | 2026-08-17 | <a href="https://platform.claude.com/">https://platform.claude.com/</a>                                                                                                                                                                                                                     |

| ID     | Publisher    | Source                                                                   | Type                                       | Use                                                                                                                  | Verified   | URL                                                                                                                                                                                               |
|--------|--------------|--------------------------------------------------------------------------|--------------------------------------------|----------------------------------------------------------------------------------------------------------------------|------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|        |              |                                                                          | documentation                              | caveat.                                                                                                              |            | docs/en/api/service-tiers                                                                                                                                                                         |
| SRC 16 | AWS          | Amazon Bedrock data protection                                           | Official product documentation             | Model deployment accounts, model-provider access boundary, encryption, networking, and retention topics.             | 2026-08-17 | <a href="https://docs.aws.amazon.com/bedrock/latest/userguide/data-protection.html">https://docs.aws.amazon.com/bedrock/latest/userguide/data-protection.html</a>                                 |
| SRC 17 | AWS          | Amazon Bedrock FAQs                                                      | Official product documentation             | Training/use statements, PrivateLink, private customization, and model-provider handling.                            | 2026-08-17 | <a href="https://aws.amazon.com/bedrock/faqs/">https://aws.amazon.com/bedrock/faqs/</a>                                                                                                           |
| SRC 18 | AWS          | Amazon Bedrock SLA                                                       | Published SLA                              | Covered API availability calculation, 5-minute intervals, error definition, exclusions, and credits.                 | 2026-08-17 | <a href="https://aws.amazon.com/bedrock/sla/">https://aws.amazon.com/bedrock/sla/</a>                                                                                                             |
| SRC 19 | AWS          | Cross-Region inference                                                   | Official routing documentation             | Geographic/global routing scope, throughput purpose, encryption, and CloudTrail inference-region evidence.           | 2026-08-17 | <a href="https://docs.aws.amazon.com/bedrock/latest/userguide/cross-region-inference.html">https://docs.aws.amazon.com/bedrock/latest/userguide/cross-region-inference.html</a>                   |
| SRC 20 | AWS          | Provisioned Throughput in Amazon Bedrock                                 | Official capacity documentation            | Model units, token throughput, hourly billing, commitment terms, and model/region support dependency.                | 2026-08-17 | <a href="https://docs.aws.amazon.com/bedrock/latest/userguide/prov-throughput.html">https://docs.aws.amazon.com/bedrock/latest/userguide/prov-throughput.html</a>                                 |
| SRC 21 | Microsoft    | Data, privacy, and security for Azure Direct Models in Microsoft Foundry | Official product documentation             | Data processing, application state, safety/abuse review, training restriction, and modified monitoring verification. | 2026-08-17 | <a href="https://learn.microsoft.com/en-us/azure/foundry/responsible-ai/openai/data-privacy">https://learn.microsoft.com/en-us/azure/foundry/responsible-ai/openai/data-privacy</a>               |
| SRC 22 | Microsoft    | Deployment types for Microsoft Foundry Models                            | Official routing/capacity documentation    | Global/data-zone/regional processing and standard/provisioned/batch differences.                                     | 2026-08-17 | <a href="https://learn.microsoft.com/en-us/azure/foundry/foundry-models/concepts/deployment-types">https://learn.microsoft.com/en-us/azure/foundry/foundry-models/concepts/deployment-types</a>   |
| SRC 23 | Microsoft    | Model retirement schedule                                                | Official lifecycle documentation           | Current model/version lifecycle, retirement dates, and replacements on Microsoft Foundry.                            | 2026-08-17 | <a href="https://learn.microsoft.com/en-us/azure/foundry/openai/concepts/model-retirement-schedule">https://learn.microsoft.com/en-us/azure/foundry/openai/concepts/model-retirement-schedule</a> |
| SRC 24 | Google Cloud | Vertex AI and zero data retention                                        | Official privacy documentation             | Training restriction, prompt logging, feature-specific retention, and in-memory caching behavior.                    | 2026-08-17 | <a href="https://cloud.google.com/vertex-ai/generative-ai/docs/vertex-ai-zero-data-retention">https://cloud.google.com/vertex-ai/generative-ai/docs/vertex-ai-zero-data-retention</a>             |
| SRC 25 | Google Cloud | Vertex AI throughput quota                                               | Official capacity documentation            | Dynamic shared quota, 429 behavior, and Provisioned Throughput as reserved capacity.                                 | 2026-08-17 | <a href="https://cloud.google.com/vertex-ai/generative-ai/docs/resources/throughput-quota">https://cloud.google.com/vertex-ai/generative-ai/docs/resources/throughput-quota</a>                   |
| SRC 26 | Google Cloud | Purchase Provisioned Throughput                                          | Official commercial/capacity documentation | Orders, model/region changes, overage, subscription terms, and preview caveats.                                      | 2026-08-17 | <a href="https://cloud.google.com/vertex-ai/generative-ai/docs/purchase-provisioned-throughput">https://cloud.google.com/vertex-ai/generative-ai/docs/purchase-provisioned-throughput</a>         |
| SRC 27 | Google Cloud | Vertex AI Platform SLA                                                   | Published SLA                              | Covered services, credits, exclusions, claim obligations, and quota/model-lifespan exclusions.                       | 2026-08-17 | <a href="https://cloud.google.com/vertex-ai/sla">https://cloud.google.com/vertex-ai/sla</a>                                                                                                       |
| SRC 28 | Google Cloud | Service Specific Terms                                                   | Published contract terms                   | AI/ML restrictions, handling, preview terms, and indemnification conditions.                                         | 2026-08-17 | <a href="https://cloud.google.com/terms/service-terms">https://cloud.google.com/terms/service-terms</a>                                                                                           |
| SRC 29 | Google Cloud | Cloud Data Processing Addendum                                           | Published DPA                              | Processor obligations, security, incident, audit, deletion, location, transfer, and subprocessor framework.          | 2026-08-17 | <a href="https://cloud.google.com/terms/data-processing-addendum">https://cloud.google.com/terms/data-processing-addendum</a>                                                                     |



## Appendix L — Maintenance and revalidation schedule

*Assign these activities to the service governance calendar and link every event to a gate, score, risk, contract term or exit trigger.*

| Cadence                                                                                                                                                                            | Scope                                                                                                                                                                 | Trigger                                                                          | Owner                                                                                 |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|
| Continuous / per request                                                                                                                                                           | Exact model/version/region/tier, latency/error/429/503, usage and cost, route metadata, safety refusals, schema/tool outcomes.                                        | Drift, threshold breach, new alias/version, unexplained invoice variance.        | Platform/SRE/FinOps                                                                   |
| Monthly                                                                                                                                                                            | Provider status/incidents; capacity headroom; spend/commit utilization; invoice reconciliation; open vulnerabilities/risks; fallback readiness.                       | Two adverse trends, chronic SLA event, budget variance, support failure.         | SRE/FinOps/Vendor manager                                                             |
| Quarterly                                                                                                                                                                          | Critical task evals against alternate; source-ledger refresh; subprocessor and policy changes; region/feature matrix; access/RBAC review; exit drill component.       | Material documentation/feature/processor change or risk rating increase.         | AI governance/Security/Privacy                                                        |
| Semiannual                                                                                                                                                                         | Peak load and failover exercise; model upgrade/rollback; data-flow/ZDR/deletion verification; contract obligation review; concentration/financial update.             | Architecture or workload growth, new regulated data, acquisition/funding event.  | Architecture/SRE/Legal/Procurement                                                    |
| Annual and pre-renewal                                                                                                                                                             | Full gates/scorecard; SOC/ISO/pen-test/bridge evidence; DPA/BAA/SLA/rate card; BCP/exit plan; market check and negotiation positions.                                 | Renewal, material scope expansion, jurisdiction change, or new criticality tier. | Executive sponsor and cross-functional review board                                   |
| Event-driven                                                                                                                                                                       | Security incident, new subprocessor, AUP/terms/price change, deprecation, forced migration, repeated outage, control change, regulatory change, or change of control. | Any listed event.                                                                | Named risk owner; convene review within 5 business days or faster for critical events |
| <b>FINAL REMINDER</b> A provider relationship is a changing technical, legal and commercial system. Revalidation is required when the system changes—not merely at annual renewal. |                                                                                                                                                                       |                                                                                  |                                                                                       |