Cost Model

The Cost Model is the foundational financial inventory for an agent stack. It catalogs every category of spend — compute, model API calls, storage, retrieval, tooling, human review — assigns each a cost-per-unit formula, and establishes hard and soft budget ceilings that downstream routing and unit-economics artifacts must respect. Without a populated cost model, the routing policy has no budget signal, the quality-cost matrix has no denominator, and the unit economics worksheet has no cost rows. Fill this spec before any routing or optimization work begins.

Cost Category Inventory

Each category below requires a fillable block entry (see §Fillable Fields). Categories that are not applicable should be marked N/A rather than omitted — the blank space would leave downstream consumers unable to confirm coverage.

Category definitions

Model API cost — token-based or per-call charges from the model provider. Splits across premium, standard, cheap, and deterministic model tiers (see the locked glossary definition of model tier). Input and output tokens priced separately; caching discounts tracked separately.

Compute / infrastructure cost — cloud VM, container, serverless-function, or GPU runtime underlying the orchestration layer. Includes cold-start overhead for latency-sensitive flows.

Storage cost — vector-store writes/reads, object storage for evidence and audit records, relational tables for state and provenance. Priced per GB-month plus per-operation fees.

Retrieval cost — embedding generation, vector similarity search, re-ranking, web search, or external index lookups. High-frequency retrieval is often the second-largest cost after model calls.

Tooling / connector cost — third-party API subscriptions, per-call fees, or metered SaaS charges for tools the agent invokes (GIS lookups, permit APIs, email/calendar connectors, payment processors).

Human-review cost — time-value of HITL gates and escalations. Expressed as reviewer hourly rate × median review time per event × escalation frequency. This is often invisible in cost models and materially understates total operating cost.

Egress / networking cost — data transfer out of the cloud region, particularly for large retrieval payloads or media files.

Observability / logging cost — APM, log storage, eval-run storage. Often a fixed-cost overhead tier.

Fillable Fields

```
agent_id: <kebab-case; matches s1-operating-context-canvas>
```

```
billing_period: <monthly | per-run | per-user>

cost_categories:
  model_api:
    provider: <vendor name>
    tiers:
      premium:
        model_id: <generic label, e.g., "frontier-reasoning-model">
        input_cost_per_1k_tokens: <number, USD>
        output_cost_per_1k_tokens: <number, USD>
        avg_tokens_per_call: <number>
        calls_per_billing_period: <number>
        cache_discount_pct: <0–100>
      standard:
        model_id: <generic label>
        input_cost_per_1k_tokens: <number>
        output_cost_per_1k_tokens: <number>
        avg_tokens_per_call: <number>
        calls_per_billing_period: <number>
        cache_discount_pct: <0–100>
      cheap:
        model_id: <generic label>
        input_cost_per_1k_tokens: <number>
        output_cost_per_1k_tokens: <number>
        avg_tokens_per_call: <number>
        calls_per_billing_period: <number>
        cache_discount_pct: <0–100>
      deterministic:
        description: <rule engine, regex, lookup — no model cost>
        calls_per_billing_period: <number>
        cost_usd: 0

  compute:
    runtime: <cloud-function | container | VM>
    unit_cost: <USD per 1M requests or per GB-hour>
    estimated_units_per_billing_period: <number>

  storage:
    vector_store_gb: <number>
    vector_store_cost_per_gb_month: <number, USD>
    object_storage_gb: <number>
    object_storage_cost_per_gb_month: <number, USD>
    operations_per_billing_period: <number>
    cost_per_1k_operations: <number, USD>

  retrieval:
    embedding_calls_per_billing_period: <number>
    embedding_cost_per_1k_calls: <number, USD>
    search_calls_per_billing_period: <number>
    search_cost_per_call: <number, USD>
    rerank_calls_per_billing_period: <number>
    rerank_cost_per_call: <number, USD>

  tooling_connectors:
    - tool_id: <kebab-case>
      vendor: <name>
      pricing_model: <subscription | per-call | metered>
      monthly_base_cost_usd: <number>
      per_call_cost_usd: <number, or 0 if subscription>
      avg_calls_per_billing_period: <number>

  human_review:
    reviewer_role: <job title>
    hourly_rate_usd: <number>
    median_review_minutes: <number>
    escalations_per_billing_period: <number>

  egress:
    gb_per_billing_period: <number>
```

```
    cost_per_gb_usd: <number>

  observability:
    flat_monthly_cost_usd: <number>

budget_ceilings:
  hard_monthly_usd: <number — crossing this triggers C8 HITL gate>
  soft_monthly_usd: <number — crossing this logs a C7 warning, no gate>
  hard_per_run_usd: <number — per-run ceiling; 0 = not enforced>
  hard_token_per_call: <integer — per-call token ceiling>
  hard_latency_ms: <integer — covered in e6-latency-budget-spec>

cost_per_outcome:
  cost_per_task_usd: <derived — total monthly / tasks_per_month>
  cost_per_user_usd: <derived — total monthly / active_users>
  cost_per_workflow_usd: <derived — total monthly / workflow_runs>
```

Hard vs. Soft Ceilings

A hard ceiling is non-negotiable: crossing it blocks the agent mid-run and emits a C8 HITL event to s1-hitl-review-policy with decision_origin: escalation. The human reviewer decides whether to continue, downgrade the model tier, or abort. A hard ceiling is set at the level where further spend is not sanctioned — either a vendor limit, a per-customer contract commitment, or an internal policy floor.

A soft ceiling is a cost-efficiency warning: crossing it logs a routing advisory to s2-audit-trail-schema via the C7 payload in e6-routing-policy, but does not gate the run. Soft ceilings are set roughly 80% of the hard ceiling to give operators runway to intervene before the hard gate fires.

The relationship must be: soft_monthly_usd < hard_monthly_usd. An inverted ceiling is a validation error the routing policy must reject at startup.

Worked Example: GoodSam Community-Intelligence Agent (illustrative)

The following example uses the GoodSam.ai platform context. All numbers are (illustrative) — they reflect plausible order-of-magnitude costs for a small production agent, not verified vendor pricing.

```
agent_id: goodsam-community-intel
billing_period: monthly

cost_categories:
  model_api:
    provider: "cloud-model-provider-A"  # generic
    tiers:
      premium:
        model_id: "frontier-reasoning-model"  # (illustrative)
        input_cost_per_1k_tokens: 0.015      # (illustrative)
        output_cost_per_1k_tokens: 0.060     # (illustrative)
        avg_tokens_per_call: 4200
        calls_per_billing_period: 800
        cache_discount_pct: 15
      standard:
        model_id: "mid-tier-model"           # (illustrative)
        input_cost_per_1k_tokens: 0.003      # (illustrative)
        output_cost_per_1k_tokens: 0.015     # (illustrative)
        avg_tokens_per_call: 2800
```

```
        calls_per_billing_period: 6000
        cache_discount_pct: 20
      cheap:
        model_id: "small-fast-model"         # (illustrative)
        input_cost_per_1k_tokens: 0.0004     # (illustrative)
        output_cost_per_1k_tokens: 0.0012    # (illustrative)
        avg_tokens_per_call: 900
        calls_per_billing_period: 14000
        cache_discount_pct: 30
      deterministic:
        description: "intent-classifier regex + rule engine"
        calls_per_billing_period: 22000
        cost_usd: 0

  compute:
    runtime: cloud-function
    unit_cost: 0.20          # (illustrative) per 1M invocations
    estimated_units_per_billing_period: 45000

  storage:
    vector_store_gb: 12
    vector_store_cost_per_gb_month: 0.30     # (illustrative)
    object_storage_gb: 40
    object_storage_cost_per_gb_month: 0.023  # (illustrative)
    operations_per_billing_period: 180000
    cost_per_1k_operations: 0.005            # (illustrative)

  retrieval:
    embedding_calls_per_billing_period: 28000
    embedding_cost_per_1k_calls: 0.10        # (illustrative)
    search_calls_per_billing_period: 18000
    search_cost_per_call: 0.004              # (illustrative)
    rerank_calls_per_billing_period: 5000
    rerank_cost_per_call: 0.007              # (illustrative)

  tooling_connectors:
    - tool_id: county-gis-lookup
      vendor: "FL GIS API"                   # (illustrative)
      pricing_model: per-call
      monthly_base_cost_usd: 0
      per_call_cost_usd: 0.002               # (illustrative)
      avg_calls_per_billing_period: 3200
    - tool_id: public-records-search
      vendor: "FS Chapter 119 portal"        # (illustrative)
      pricing_model: subscription
      monthly_base_cost_usd: 50              # (illustrative)
      per_call_cost_usd: 0
      avg_calls_per_billing_period: 900

  human_review:
    reviewer_role: "District Program Coordinator"
    hourly_rate_usd: 65                      # (illustrative)
    median_review_minutes: 8
    escalations_per_billing_period: 35

  egress:
    gb_per_billing_period: 4
    cost_per_gb_usd: 0.09                    # (illustrative)

  observability:
    flat_monthly_cost_usd: 45               # (illustrative)

budget_ceilings:
  hard_monthly_usd: 900                     # (illustrative)
  soft_monthly_usd: 720                     # (illustrative)
  hard_per_run_usd: 0.85                    # (illustrative)
  hard_token_per_call: 8000

cost_per_outcome:
```

```
cost_per_task_usd: 0.18                 # (illustrative) ~4900 tasks/month
cost_per_user_usd: 2.25                 # (illustrative) ~400 active users
cost_per_workflow_usd: 0.52            # (illustrative) ~1730 workflow runs
```

Approximate monthly total (illustrative)

Category	Estimated USD/mo
Model API — premium tier	$63
Model API — standard tier	$126
Model API — cheap tier	$16
Compute (cloud functions)	$9
Storage (vector + object)	$5
Retrieval (embed + search + rerank)	$82
Tooling / connectors	$56
Human review (HITL time)	$304
Egress + observability	$49
Total	~$710

Human review is the largest single line item at ~43% of total spend — a pattern common to production agents operating in regulated or public-sector contexts. This is why HITL frequency is a first-class optimization target in e6-waste-reduction-playbook.

Usage Notes

Populate this spec before building e6-routing-policy. The routing policy's tier-selection logic must have verified cost-per-call figures to enforce the hard ceiling. Treat the cost_per_outcome block as a derived metric — update it each billing period rather than hardcoding a projection. If any tier's actual spend diverges from the estimate by more than 25%, reopen this spec and the quality-cost matrix together. The hard_monthly_usd field is the budget signal used by the C8 contract: when the routing policy's running tally crosses this value, it must emit the HITL event described in e6-routing-policy.