DOCUMENT 42 · AI COSTS

AI costs and allocation per use case

Full cost, analytical allocation of shared costs and consumption control to know what each use case really costs

Fernando García Varela Version 0.1 (working draft) 16-09-2026 Draft for review

9	3	5	1
cost categories	components of the cost per use case	types of shared cost with allocation key	monthly reconciliation with accounting

IN THIS DOCUMENT

01 Purpose and scope

02 Principles

03 Components of the cost of a use case

04 Cost categories

05 Total cost of ownership

06 Analytical allocation of shared costs

07 Cost tagging in the cloud and in models

08 Consumption forecasting and control

09 Cost per unit of outcome

10 Cost of stop decisions

11 Integration with management control

12 Tool T13 · Cost calculator per use case

13 Associated tools and templates

14 Related documents

15 Version control

Legal notice and disclaimer. SEVEN-G is a reference methodological framework provided "as is" and for information purposes only. It does not constitute legal, regulatory, financial or professional advice, nor does it guarantee compliance with any law or standard. References to general regulation (such as the EU AI Act, the GDPR, DORA or NIS2), to technical standards and to regulation specific to each sector or jurisdiction may be incomplete, may not apply to a particular case or may become out of date as a result of regulatory changes, interpretations or supervisory positions after their consultation date. **Each organisation that uses SEVEN-G is solely responsible for identifying the regulation that applies to it, verifying that it is current and certifying its own regulatory compliance**, with the appropriate qualified advice. This methodology is a generic, free aid shared with the community so that nobody has to start from scratch; each person or organisation can and should adapt it to its own use. It must not be inferred that its legally sensitive parts have been reviewed by legal counsel: those reviews, for each company or sector, are the ultimate responsibility of the company, consultant or organisation that uses it. Although every effort is made to keep it up to date, some regulation may have changed without being reflected here. To the fullest extent permitted by law, the author accepts no responsibility whatsoever for the effects of its application in any organisation or for its full applicability. The methodology does not grant certification of any kind. The author accepts no liability for any use made of this content or for any decisions taken on the basis of it. The data, figures, companies and cases in the examples are fictitious or illustrative.

THE ESSENTIALS Recommended

The full recurring cost of each case is mandatory under the measurement rules (document 40). Analytical allocation, total cost of ownership and cost per unit of output are applied in proportion to the size of the portfolio; the consumption budget, when cost depends on usage.

Obligation matrix and layered reading (document 94)

01

Purpose and scope

This document establishes how the cost of artificial intelligence is calculated, allocated, controlled and reported in SEVEN-G. Its purpose is to ensure that the **recurring cost** deducted in annual net value (document 40, formula F2) and the **initial investment** used in the economic criterion (document 40 §8) are complete, comparable across use cases and reconcilable with accounting.

It applies to:

- Every use case in the register (T01), from the cost estimate in phase 3 through to retirement.
- Costs shared by several use cases: productivity suites with AI, model platforms, compute, data platforms and common teams.
- Corporate use of general-purpose AI, which has a cost even though it is not a use case (01 §1.2).

The figures in the examples are **fictitious and illustrative**.

02

Principles

#	Principle	What it means
1	**Full cost for measuring**	The recurring cost of a use case includes its direct costs and its share of shared costs. In this way, the sum of the use cases, corporate use and unattributed cost equals the company's total AI cost.
2	**Incremental cost for deciding**	To decide on an additional investment (additional net value per euro, document 40 F3), the cost that changes with the decision is used. A shared cost that will exist in any case is not incremental. Both views are calculated and identified.
3	**Cost is not reduced**	The cost is charged in full from day one, whether or not the value has been validated. An unknown cost is shown as "no data", and so is the net value.
4	**A single source of truth**	The T13 cost is reconciled with accounting every month. Differences are explained or corrected.
5	**Tagging from day one**	No cloud resource, model access key or licence is created without the tag of the use case or corporate use to which it belongs.
6	**Declared and stable allocation keys**	Each shared cost has a key approved by management control. It is not changed mid-year except in the event of an error, and the change is recorded.

#	Principle	What it means
7	**Consumption is governed before it is invoiced**	Variable model and compute costs have a budget, forecast, alerts and technical limits, especially for agents.

03

Components of the cost of a use case

Component	Definition	When it is incurred	Use in SEVEN-G
Build cost	Cost of designing, building or integrating, testing and putting the solution into production, including the pilot.	Phases 3 to 5, once per approved scope.	Part of the initial investment (I).
Initial adoption cost	Initial training, change management, redesign of processes and roles, communication and support until stable use.	Phases 4 to 6, once.	Part of the initial investment (I). It must be estimated for Augment and Transform (01 §3, principle 10).
Recurring cost	Annual cost of operating, maintaining, overseeing, providing ongoing training and complying.	Phase 6 onwards, every year.	Deducted in annual net value (F2).
Retirement cost	Cost of retiring the system: migration, retention or deletion of data and models, communication, contractual penalties.	Phase 7, if retired.	Estimated at G3 and included in the total cost of ownership; deducted in the year in which it is incurred.

Classification rules:

1. **Major evolutions are build.** A scope extension approved at a new G0 (scaling, 01 §6.9) has its own initial investment. Corrective maintenance and minor improvements are recurring.
2. **Pilots are charged to build** even if they are extended. If a pilot exceeds the reference time limit for phase 5 (03 §3.6), its cost continues to accumulate in build and is flagged to the committee.
3. **Accounting capitalisation is independent.** Whether a cost is recorded as an asset or as an expense is decided by the finance function under the applicable accounting standards (for example, the intangible assets standard for development costs). SEVEN-G uses the management classification in this section, with a reconciliation between the two.

04

Cost categories

The nine categories are common to all documents (common specification §5.7). Each T13 amount is classified in one category and one component.

Category	What it includes	Build	Recurring	Adoption	Usual source	Concept in the dashboard (T17)
Licences	Subscriptions to software with AI, platform and tool licences, productivity suites with AI.	Sometimes	Yes	—	Contracts, licence management (T21)	`licencias`
Model consumption	Pay-per-use of models: units of text processed (tokens), calls, voice minutes, images; including use in testing.	Yes (testing and pilot)	Yes	—	Supplier invoicing, model gateway	`plataforma`
Compute and infrastructure	Servers, graphics processors, storage, network, environments, observability, in the cloud or on-premises (depreciation).	Yes	Yes	—	Cloud invoicing, fixed assets	`infraestructura` (and `plataforma` for platform compute units)
Data	Acquisition of external data, preparation, labelling, quality, catalogues and knowledge bases.	Yes	Yes	—	Contracts, hours charged	No concept of its own; adding one is recommended
Build people	Internal hours of product, data, engineering, business and experts during the build, at full cost.	Yes	—	—	Time recording, payroll	`personas` (build)
Operations people	Internal hours of operation, monitoring, sample-based human review, maintenance and support, at full cost.	—	Yes	—	Time recording	`personas` · `mantenimiento`

Category	What it includes	Build	Recurring	Adoption	Usual source	Concept in the dashboard (T17)
Suppliers and services	Professional services, integrators, contracted support and maintenance, managed services.	Yes	Yes	Yes	Contracts, invoices	`servicios`
Control and compliance	Impact assessments, security and bias testing, audit, legal advice, specific insurance, certifications.	Yes	Yes	—	Contracts, hours charged	`cumplimiento`
Adoption and training	Training, change management, communication, materials, support, users' time spent in training.	—	Yes (ongoing training)	Yes	Adoption plan (T20), hours	No concept of its own; adding one is recommended

Users' time spent in training and in reviewing system outputs is a cost. It is valued at full hourly cost and included in adoption and training (training) or in operations people (review), because omitting it inflates efficiency (document 40 §10, error 6).

05

Total cost of ownership

5.1 Formula

Total cost of ownership (H) = build cost + initial adoption cost + Σ recurring cost in year t (t = 1…H) + estimated retirement cost

H is the horizon approved in C2 (document 40 §8.2). The total cost of ownership is not discounted; when the present value is needed, the NPV from document 40 (F7) is used, which already incorporates costs through annual net value.

5.2 Illustrative example

Use case A: assistant to resolve internal employee queries about procedures. Enterprise intensity. Horizon H = 3 years.

Component	Category	Amount
Build	Build people	€90,000
Build	Suppliers and services	€45,000
Build	Control and compliance	€15,000
Total build		**€150,000**
Initial adoption	Adoption and training	**€30,000**
Annual recurring	Allocated shared costs (section 6.7)	€163,000
Annual recurring	Operations people	€50,000
Annual recurring	Control and compliance	€12,000
Annual recurring	Suppliers and services (support)	€15,000
Total annual recurring		**€240,000**
Estimated retirement	Suppliers and services	**€10,000**

Total cost of ownership (3 years) = 150,000 + 30,000 + 3 × 240,000 + 10,000 = €910,000.

Initial investment I = 150,000 + 30,000 = €180,000, the same as used in the example in document 40 §8.4.

06

Analytical allocation of shared costs

6.1 Charging hierarchy

Each cost is charged using the first possible method in this list, in this order:

Order	Method	When	Example
1	**Direct charging**	The cost is exclusive to the use case.	A supplier contract for a single use case.
2	**Measured consumption**	The cost is shared, but each use case's consumption is measured by tag, access key or log.	Units of text processed per access key of each use case.

Order	Method	When	Example
3	**Cost driver**	Consumption is not measured, but there is a variable that reasonably explains it.	Hours charged, assigned users, transaction volume.
4	**Weighted allocation**	There is no reliable consumption or driver. Only for overhead costs.	Use cases in production weighted by intensity.

Corporate overhead cost (governance, methodology, preparation of board information) is not allocated to use cases: it is reported as the company's AI governance cost, within the total cost (IND-COS-11 in document 41).

6.2 Allocation keys by type of shared cost

Shared cost	Preferred key	Alternative key	What is not allocated to use cases
Shared licences (productivity suites with AI)	Licences assigned to the group that uses the use case for that purpose	Active users of the use case ÷ total active users	General-use licences, which are charged to corporate use of general-purpose AI
Model consumption (model gateway or platform)	Consumption measured by access key or tag of the use case	Calls or transactions per use case	Platform testing consumption, which is charged to the platform
Compute (graphics processors, clusters)	Compute hours per tagged job	Execution time per use case	Idle capacity, which is reported separately
Shared platforms (data, model operations, agents)	Platform consumption units per use case	No. of models, agents or flows in production per use case	Build cost of the platform itself, which has its own use case
Shared teams (AI Office, platform team, risk review)	Hours charged per use case	Use cases weighted by intensity (Enterprise 2, Lite 1)	The corporate overhead portion (section 6.1)

6.3 Untagged consumption

Untagged consumption is allocated in proportion to tagged consumption **only if it does not exceed the tolerance** approved by management control. If it exceeds it, it is not allocated: it is reported as **unattributed cost** (IND-COS-08), an action is opened to correct the tagging and, if it persists for two closes, it is treated as a minor nonconformity.

6.4 Illustrative example · Model consumption

Shared model platform with an annual cost of €240,000. Consumption measured by access key: use case A 46%, use case B 32%, use case E 18%, untagged 4%. Approved tolerance: 5%.

Use case	Measured consumption	Direct charging	Allocation of untagged (4% in proportion 46 : 32 : 18)	Total charged
A	46%	€110,400	9,600 × 46/96 = €4,600	**€115,000**
B	32%	€76,800	9,600 × 32/96 = €3,200	**€80,000**
E	18%	€43,200	9,600 × 18/96 = €1,800	**€45,000**
Untagged	4%	€9,600	Allocated	€0
Total	100%	€240,000	€9,600	**€240,000**

6.5 Illustrative example · Shared licences and compute

Productivity suite with AI. 600 licences × €30 per month × 12 = €216,000 per year. Use case D (assistant to prepare commercial proposals within the suite) justified the assignment of 150 licences to the sales team. Key: licences assigned for the use case.

- Use case D: 150 × €360 = **€54,000**.
- Corporate use of general-purpose AI: 450 × €360 = **€162,000**.
- Check: 54,000 + 162,000 = €216,000.

If only 120 of use case D's 150 licences are used in the period, the use case continues to bear all 150, and the 30 unused licences (€10,800) are reported as idle licence cost (IND-COS-09). Inactivity does not reduce the cost of the use case: it is an adoption problem of the use case.

Shared compute cluster. €180,000 per year; 6,000 compute hours consumed by tagged jobs; cost per hour: 180,000 ÷ 6,000 = €30.

Use case	Hours	Charge
B	1,000	€30,000
C	3,000	€90,000
E	2,000	€60,000
Total	**6,000**	**€180,000**

6.6 Illustrative example · Shared team

AI Office with a full cost of €280,000 per year. Management control determines that 40% (€112,000) is corporate governance overhead and is not allocated. The remaining 60% (€168,000) is allocated among the five use cases in production, weighted by intensity (Enterprise 2, Lite 1). Use cases A and E are Enterprise; B, C and D are Lite. Sum of weights: 2 + 2 + 1 + 1 + 1 = 7. Cost per unit of weight: 168,000 ÷ 7 = €24,000.

Use case	Intensity	Weight	Charge
A	Enterprise	2	€48,000
B	Lite	1	€24,000
C	Lite	1	€24,000
D	Lite	1	€24,000
E	Enterprise	2	€48,000
Total allocated		**7**	**€168,000**

6.7 Consolidation of the allocation

Shared cost	A	B	C	D	E	Corporate use and overhead	Total
Model consumption	€115,000	€80,000	—	—	€45,000	—	€240,000
Compute	—	€30,000	€90,000	—	€60,000	—	€180,000
Suite with AI	—	—	—	€54,000	—	€162,000	€216,000
AI Office	€48,000	€24,000	€24,000	€24,000	€48,000	€112,000	€280,000
Total	**€163,000**	**€134,000**	**€114,000**	**€78,000**	**€153,000**	**€274,000**	**€916,000**

Check: 163,000 + 134,000 + 114,000 + 78,000 + 153,000 + 274,000 = €916,000 = 240,000 + 180,000 + 216,000 + 280,000. The €163,000 for use case A is the amount shown as allocated shared costs in its recurring cost (section 5.2).

6.8 Incremental view for deciding

To calculate the additional net value per euro of an extension of use case D, the cost of the suite is only incremental if the extension requires **new licences**. Those already paid for corporate use are not. The T13 decision sheet shows both views: full cost (for measuring) and incremental cost (for deciding).

07

Cost tagging in the cloud and in models

7.1 Mandatory tags

Tag	Value	Example
`iniciativa`	Use case code, or `CORP-IA` for corporate use, or `PLAT-<nombre>` for shared platforms	`IA-2026-014`
`sistema`	System identifier in the inventory (T02)	T02 identifier
`entorno`	development · testing · pre-production · production	`produccion`
`fase`	0–7 of the lifecycle	`5`
`centro_coste`	Cost centre or internal accounting order	Accounting code
`responsable`	Role accountable for the cost (no personal data in the tag if the policy prevents it)	`producto`

7.2 Rules

1. **No tag, no resource.** Cloud policies should prevent resources from being created without the mandatory tags; if this is not possible, they are detected and corrected at the next close.
2. **One model access key per use case and environment.** Model consumption is channelled, where possible, through a common gateway that records the key, the use case, the model and the units consumed.
3. **Agents have their own identity and budget.** Each agent uses its own credentials (document 35), which allow its consumption to be measured separately.
4. **Tagging is verified at G5.** A use case does not go into production without its own tags and keys.
5. **Common cost data format.** The company may adopt an open specification for cloud billing data, such as the one promoted by the FinOps Foundation (FOCUS), to unify suppliers. The current version must be verified.

08

Consumption forecasting and control

8.1 Budget and forecast

Each use case with variable costs has a monthly and annual **consumption budget** approved at G5 and reviewed at R6. The budget and its reconciliation with invoicing are documented in P63.

Month-end forecast = cumulative consumption ÷ days elapsed × days in the month

The forecast is recalculated daily for agents and for use cases with direct exposure, and weekly for the rest. When the consumption pattern is not linear (month-end closes, campaigns), the profile of the same previous period is used.

8.2 Alerts and response

Indicative threshold	Alert to	Response
50% of budget consumed	AI Operations Owner	Review the forecast.
80% of budget consumed or month-end forecast > 100%	AI Operations Owner and AI Product Owner	Analyse the cause: volume, change of model or price, loops, misuse. Decide on measures.
100% of budget consumed	AI Sponsor and AI Office	Apply the planned level of the cost-driven degradation cascade (document 52, section 10.3: optimise, validated fallback model, reduced scope or alternative without AI) or approve a supplement with a stated reason.
Anomalous consumption (for example, one day far above the recent average)	AI Operations Owner and security	Treat it as a possible incident (document 37): loop, abuse, credential leak.

The thresholds are indicative and the company sets them in C2.

8.3 Technical limits per agent

Agents can generate chained consumption without human intervention. Therefore, in addition to the budget, they must have **technical limits** configured in the security design (P18) and checked at G5:

Limit	What it prevents
Maximum cost or units per execution	Tasks that run on indefinitely.
Maximum no. of model or tool calls per execution	Retry or reasoning loops.
Maximum daily and monthly spend per agent	Runaway consumption between reviews.
Maximum no. of concurrent executions	Peaks that multiply cost.
Action on reaching the limit	Stop the execution, refer to a person or switch to degraded mode, with logging.

The agent's kill switch (document 35) must also be able to be activated for cost reasons.

8.4 Illustrative example

Agent E with charged model consumption of €45,000 per year; monthly budget: 45,000 ÷ 12 = €3,750. Day 12 of a 30-day month; cumulative consumption: €2,100.

- Consumption against budget: 2,100 ÷ 3,750 = 56% → 50% alert.
- Month-end forecast: 2,100 ÷ 12 × 30 = €5,250 → 140% of budget → 80% alert.
- Analysis: a change in an external tool causes chained retries. It is corrected, the limit on calls per execution is lowered and the event is recorded in T01.

8.5 Cost optimisation

Before requesting more budget, the AI Technical Owner should review, with tests confirming that quality (IND-OPE-07) does not deteriorate: the model used compared with lower-cost alternatives, the length of the context sent, the reuse of responses or intermediate results, batch processing when an immediate response is not needed, and retries. Any model change in production follows the change management in document 52.

A lower-cost model can be validated before G5 as the **fallback model** of the cost-driven degradation cascade (document 52, section 10.3): with the same evaluation set and the same bias tests as the primary model, activating it when the budget runs out is an approved operating mode and not a change. If the system spends more than a month degraded, the R6 decides between a supplement, permanent optimisation or a change of the primary model.

09

Cost per unit of outcome

9.1 Formula

Cost per unit of outcome = recurring cost for the period ÷ units of useful outcome in the period

The **unit of useful outcome** is the business result, not the technical activity: query resolved, case file processed, invoice posted, proposal sent. Calls to the model and conversations started are not units of outcome.

To compare with the baseline, the **process cost per unit** is also calculated, which adds to the cost of AI the human cost that still exists (referrals, review, exceptions).

9.2 Illustrative example

Use case A, 120,000 queries per year. Baseline: process cost of €4.00 per query (€480,000).

Item	Amount
Recurring cost of AI	€240,000
Queries resolved by the assistant without referral	96,000
AI cost per query resolved	240,000 ÷ 96,000 = **€2.50**
Queries referred to people: 24,000 × €4.00	€96,000
Sample-based human review	€14,000
Process cost with AI	240,000 + 96,000 + 14,000 = €350,000
Process cost per query	350,000 ÷ 120,000 = **€2.92**

The reduction in process cost (480,000 − 350,000 = €130,000) is **potential capacity**, not an efficiency: it only counts in net value when the lower human cost materialises (document 40, rule 3). The cost per unit is an efficiency indicator for the use case (IND-COS-05), not a value figure.

10

Cost of stop decisions

Stopping or retiring a use case is a valid outcome of the method (01 §7.3). So that those who stop are not penalised financially, and so that the board can see the effect of these decisions, the following magnitudes are recorded in T01 and T13 when the stop or retirement is registered:

Magnitude	Definition	Treatment
Sunk cost	The amount spent up to the decision.	It plays no part in the decision: it can no longer be recovered. It is recorded for lessons learned.
Non-cancellable commitments	Future payments that will be made in any case (for example, annual licences already signed).	Reported separately; their reuse in other use cases is sought.
Cost of stopping	Costs caused by the decision: closure, retirement, early cancellation penalties, retention or deletion of data and models.	Recorded as a cost of the use case.
Avoided investment	Approved budget − spent − non-cancellable commitments.	Reported in the portfolio (IND-COS-10). **It does not add to value.**
Expected avoided recurring cost	Recurring cost that would have been incurred in production.	Reported as estimated. It does not add to value.

Illustrative example. Use case stopped at G3. Approved budget: €400,000. Spent: €45,000. Non-cancellable commitments: €20,000. Penalty for cancelling a services contract: €8,000. Closure cost: €5,000.

- Cost of stopping: 8,000 + 5,000 = **€13,000**.
- Avoided investment: 400,000 − 45,000 − 20,000 = **€335,000**.
- Avoided future spend net of the cost of stopping (for information): 335,000 − 13,000 = **€322,000**.

A stopped use case is never delayed to avoid recognising the sunk cost. If the stop criteria are met, 01 §7.4, rule 6 applies.

11

Integration with management control

11.1 Charging structure

Element	Rule
Cost object	Each use case has a cost object in accounting (internal order, project or cost centre) with its IA-AAAA-NNN code. Corporate use and platforms have their own.
Internal hours	People working on use cases charge hours to the cost object. Management control sets the full hourly cost by professional category.
Contracts	Purchase orders and contracts carry the use case or platform code.
Shared costs	They are recorded in their platform cost object and allocated using the keys in section 6 at the close.

11.2 Calendar

Moment	Activity	Accountable
Monthly close	Loading of cloud, model and licence invoicing; hours charged; allocation; reconciliation of T13 with accounting; budget alerts.	Management control with the AI Office
Quarterly	Variances per use case (IND-COS-04), unattributed cost, idle licences; reporting to the AI Committee.	Management control
Phase 3	Estimate of full costs (build, adoption, recurring and retirement) and of the total cost of ownership.	AI Product Owner with management control
G5 and R6	Consumption budget, tagging and limits; review of actual against forecast recurring cost.	AI Operations Owner; verified by the AI Office or the AI Auditor
Portfolio (C3)	Budget for the portfolio and shared platforms; approval of the allocation keys for the year.	AI Committee with management control
Annual close	Company's total AI cost (IND-COS-11); review of keys; reconciliation with the framework budget from C2.	Management control

11.3 Reconciliation

The monthly reconciliation checks that:

Total AI cost in accounting = Σ cost charged to use cases + corporate use cost + governance overhead cost + unattributed cost

Differences are explained (accruals, pending invoices) or corrected before the next close. A difference left unexplained for two consecutive closes is reported to the AI Committee.

12

Tool T13 · Cost calculator per use case

T13 is a module of the T11 calculator ("Costs per use case" view, serverless HTML, 03 §5.3): it exports CSV (spreadsheet) and JSON with the expected values for T01. The consumption budget, the allocation keys and the reconciliation are documented with P63.

Module	Inputs	Calculations	Outputs
Parameters	Full hourly cost by category; horizon H; rate r; tolerance for untagged consumption; alert thresholds.	—	Parameters versioned by year.
Estimate per use case	Costs by category and component; volume and consumption assumptions.	Initial investment; recurring cost; total cost of ownership; incremental view.	Data for T11 (NPV, ROI, payback) and for G3.
Shared costs	Invoicing per platform; consumption by key or tag; hours; assigned licences; approved keys.	Direct, consumption-based, driver-based and weighted charging; allocation of untagged consumption within tolerance.	Consolidation table (section 6.7).
Consumption	Daily or weekly consumption per use case and agent; budgets.	Percentage consumed; month-end forecast; anomalies.	Alerts and event log.
Unit of outcome	Units of useful outcome; residual human cost.	Cost per unit of outcome; process cost per unit.	IND-COS-05.
Stops	Approved budget, spent, commitments, closure costs.	Cost of stopping; avoided investment.	IND-COS-10.
Reconciliation	Accounting balances by cost object.	Differences by object.	Monthly reconciliation report.

Minimum T13 validations: no amount without a category and component; no shared cost without an approved key; the allocated sum equals the total of the shared cost; nulls are shown as "no data".

13

Associated tools and templates

Name	Use in this document
Cost calculator per use case (T13)	Main tool of this document.
Use case register (T01)	Costs per use case, alert events, stops.
AI system inventory (T02)	System identifier for tagging.
Value hypothesis canvas and calculator (T11)	Use of the initial investment and recurring cost in F2, F3 and F7–F9.
Value realisation tracking (T12)	Actual recurring cost in net value per period.
Board AI dashboard (T17)	Cost per use case and total AI cost.
Corporate AI use monitor (T21)	Assigned and active licences.
Feasibility assessment (P10)	Estimate of full costs in phase 3.
Security design (includes agents) (P18)	Technical consumption limits per agent.
Adoption and capacity plan (P20)	Adoption cost.
Monitoring and alert configuration (P25)	Consumption alerts.
Scaling or retirement decision (P30)	Retirement cost and cost of stopping.
Consumption budget and cost reconciliation (P63)	Consumption budget, forecast, alerts and reconciliation with invoicing (section 8).

14

Related documents

Document	Relationship
01 · Foundational methodology	Full costs in phase 3 (§6.5), stop and retire outcomes (§7.3).
03 · Tools and use case register	Definition of T13 and common data model.

Document	Relationship
13 · AI thesis and risk appetite	Framework budget, horizon, rate and alert thresholds.
14 · Portfolio management	Portfolio budget and shared platforms.
35 · AI and agent security	Agent identity, limits and kill switch.
36 · AI third parties and suppliers	Contracts, cancellation and dependency.
40 · Value measurement rules	Use of recurring cost and initial investment in the official formulas.
41 · Indicator catalogue	COS family indicators.
43 · Benefits realisation	Actual cost in net value tracking.
52 · AI operations manual	Model change management and consumption monitoring.

15

Version control

Version	Date	Changes
0.1	16-09-2026	First version. Defines the cost components (build, adoption, recurring and retirement), the nine categories with their equivalent in the dashboard, the total cost of ownership, the analytical allocation of shared costs with charging hierarchy, keys and examples, tagging, consumption forecasting and control with limits per agent, the cost per unit of outcome, the cost of stop decisions, integration with management control and the content of T13.
0.1	18-09-2026	Response at 100% of the budget linked to the cost-driven degradation cascade in document 52 (section 10.3) and fallback model validated before G5 (section 8.5).

External sources cited

Links to the official or primary source, checked before publication. Always verify the version in force on the date of use.

FOCUS™ — FinOps Open Cost and Usage Specification, version 1.4 · https://focus.finops.org/ · FinOps Foundation (FOCUS, serie de Joint Development Foundation Projects, LLC) · 04-06-2026