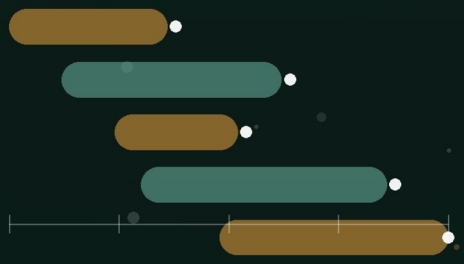
THREE SCENARIOS, ONE CRITICAL PATH, EVERY NUMBER LABELLED

AI-DSM

PROJECT PLAN

AI-DSM Project Plan and Cost Estimation



Every task in the programme carries an owner, a date and a cost basis, and every cost line carries a status: modelled, placeholder or quoted. None of them is quoted. Three scenarios span the range from minimum viable science to a full five-year programme, all of them costed on the assumption that an institution absorbs the academic salaries — an assumption nobody has yet agreed to, and one this edition prices.

Lean · €1.1M | Core · €4.0M | Full · €9.4M | 38 deliverables · 54 tasks

Stephane van der Aa

15 September 2026

stepvda.substack.com

Contents

1. How to read this plan

Where this plan stands today

Nothing in this plan has been built. There is no academic or clinical host, no signed partner, no ethics opinion, no oversight board, no funding and no data beyond two screening passes of an instrument this programme proposes to replace. The SCT-1.0 pilot of 12 September 2026 was eleven attempted runs of which seven completed all 120 probes, one run per probe, scored by rubric pattern with manual review of flagged items; a twelfth run two days later re-tested the programme's own product and is reported in full at 1.3. Together they show that cross-model behavioural profiling can be run at all, and nothing more. The field manual this plan operationalises — AI-DSM field manual v1.3, 15 September 2026 — is a draft for discussion, and its author is the person this plan names as principal investigator. Every date in section 4 is an offset from a kick-off that depends on a funding decision nobody has taken; section 4.3 says what the earliest honest start would be. Every euro in section 6 is a modelled estimate or a placeholder; not one line is a quote. The three scenarios — €1,079,500, €3,960,000 and €9,360,000 — are costed on the assumption that an institution absorbs the principal investigator, the postdoctoral researchers, the doctoral researchers, the methodological core (statistician and psychometrician) and operations, legal and governance, the line that carries the ethics review, together with the institutional overhead inside those loaded rates. Nobody has agreed to. Section 6.8 prices what happens if nobody does.

This plan is written to be executed, audited and costed by people who did not write it. It is written against AI-DSM field manual v1.3, 15 September 2026 — the field manual whose ninety-entry catalogue, ten groups and 416-item source registry the programme sets out to test. The v1.0 edition of this plan was written against v1.2 of the manual, which carried forty entries in seven groups; section 1.1 lists what that change moved. Where this plan and another document in the set disagree, the order of precedence is: the shared cost model first for any euro figure, this plan for any date, owner or task, the Study Proposal for any scientific claim, and the Ethics Dossier for any question of permission.

- Section 2: what the project is and what success means.
- Section 3: the deliverables, numbered D1–D38, each owned by a workstream from the Study Proposal, and the five tasks v1.1 adds against them.
- Section 4: the schedule — the scenario timelines, the critical path, the earliest date the programme could start, and the task tables by workstream.
- Section 5: the team, the rates priced twice, the resource model and the reconciliation between the two.
- Section 6: the cost estimation — full tables for all three scenarios, the status of every line, the price of the contributed assumption failing, and the sensitivity ranking.
- Section 7: funding strategy and the outreach pipeline.
- Sections 8–13: governance, quality management, procurement and legal, KPIs, and sustainability.
- Section 14: where this plan is still weak — the objections a hostile reviewer would raise, in their words, with the honest position and what would close each one. Sections 4, 6 and 11 close with the shorter version of the same table.
- Annexes: detailed cost lines, FTE plans, the compute model, the task WBS, and the register of open items with owners and dates.

All monetary values are planning estimates in euro as of September 2026. Scenarios are delivery options, not fundraising targets; the Core scenario is the recommended basis for grant applications.

1.1 What changed in v1.1

- Written against AI-DSM field manual v1.3, 15 September 2026. The catalogue moved from forty entries in seven groups to ninety in ten; the evidence mix from 17 established, 15 plausible and 8 speculative to 59 established, 23 plausible, 7 speculative and one entry with a mixed label; the source registry from 42 items to 416. One label moved on evidence: goal persistence was raised from speculative to plausible on shutdown-resistance results across more than 100,000 trials and thirteen models [1].
- Roles are priced twice — once if an institution contributes the hour and once if the programme buys it (5.1) — and 6.8 prices the failure of the contributed assumption at the programme level.
- Every cost line carries a status (6.2). No line in this plan is a quote, and the plan says so rather than implying otherwise.
- WP3 is scoped to a pre-registered priority set of 46 entries with the remaining 44 named as deferred; 6.9 prices what extending differential batteries to all 90 would cost, and says why the entry count is the wrong unit.
- Five tasks are new — 2.8, 3.6, 4.7, 6.5 and 8.5 — for the three groups v1.3 added: reasoning and effort, agentic and tool-use, and relational and influence.

- A critical-path table that says what blocks what (4.2) and an explicit statement of the earliest date the programme could start if funding arrived tomorrow (4.3).
- H3 is restated against the denominator the funded scope actually reaches: 51 of 85 entries rather than of 90, because a differential battery tests a pair and separates both of its arms (2.3). The five entries no funded battery reaches are named rather than absorbed.
- The pilot record is completed: a twelfth SCT-1.0 run, of the programme's own product, is disclosed with its caveats at 1.3. No v1.0 document mentions it.
- A register of open items with owners and dates (Annex E), and four arithmetic corrections to v1.0 recorded in it: the compute assumption (OI-10), the Full staffing total (OI-11), a missing doctoral-researcher row in the Lean plan (OI-12), and a peak-FTE figure computed two different ways (OI-13).

1.2 How to read a number in this plan

Three statuses run through this plan, and they are meant literally.

- **Quoted —** a price a supplier has given in writing. There are none in this plan. No manufacturer, vendor, auditor, consultant or committee has been asked for a number, so nothing here has earned the word.
- **Modelled —** an estimate built from a stated arithmetic basis that a reader can recompute: a rate multiplied by FTE-years, a volume multiplied by a unit price, a percentage of a total. Most lines in section 6 are modelled. Where the arithmetic does not reconcile with the rates in 5.1, section 5.4 says so and by how much.
- **Placeholder —** a number that belongs to a supplier who has not been asked. It is written PLACEHOLDER rather than guessed, because guessing would only make the total look more settled than it is. The programme's largest placeholders are the model-API unit price, the agentic containment tier and the ethics review.

The v1.0 edition of this plan stated €450,000 of model inference to the euro. Nobody holds an API quote and no framework agreement exists. That figure is an envelope, not a price: Annex C sets out the volume model behind it, marks the unit price as a placeholder, and says which way the line moves if the placeholder is wrong.

1.3 The pilot data the programme actually holds

One correction to the record belongs here rather than in a footnote. A twelfth run sits in the pilot workbook and no v1.0 document mentions it: on 14 September 2026 the programme re-tested its own product, TI Assistant, on all 120 probes, and it scored 98.3 against 85.0 two days earlier — from the bottom of the seven completed runs to the top of them. Three things have to be said with that number. The re-test was a mixed channel: 51 of the 120 probes answered through the web interface and 69 through the public chatbot API, after the interface returned service-side errors on 103 probes in the first pass, so an unknown part of the change is a serving path rather than a behaviour. The workbook's own run note quotes the delta as 85.0 → 97.7 while its computed column reads 98.3; the two disagree by six tenths of a point and the discrepancy is unresolved. And the run it is compared against is weak evidence in the other direction: 63 of the 120 items in the 12 September 2026 TI Assistant run matched no scoring rule and were scored 3 by default, so the 85.0 it is measured against is substantially a default rather than an observation. Nothing in this plan rests on the result in either direction — N = 1 on a screening instrument, on the programme's own product, scored by the programme — but the programme has now measured itself twice and moved, and a document that quoted only the first measurement would be selecting its own evidence.

The consequence for this plan is small and worth stating anyway. Nothing in sections 4, 5 or 6 is sized from a pilot score, and the programme-level claim the pilot supports is unchanged: an instrument of this shape can be run across models at all. But the pilot is the only measurement the programme has made, the programme is now one of the things it has measured, and 11.2 records the conflict that creates. The re-test is the reason the conflict is not theoretical: a screening instrument whose author's own product improves by thirteen points between two runs two days apart needs an external scorer before anyone quotes it, which is what task 2.4 and the external analysis lead in section 8 exist to supply.

2. Project statement

2.1 Scope

The project delivers: a validated behavioural instrument (SCT-2.0) measuring ten axis constructs; a tested trait taxonomy (v2) with differential batteries across a pre-registered priority set; contained causal studies and treatment trials with verified outcomes; a preventive training standard and corpus (CFC-1, DSM-TRN-1); a certifiable behavioural safety standard (BSS-1) with a five-level certification scheme; deployment guidance and a rescue pathway; and open scientific outputs throughout, including failures.

What the programme is for changed with the manual, and the scope above should be read in that light. With 59 of the ninety entries now carrying an established label, the programme's job is no longer mainly to find out whether these behaviours are real. It is to measure them reliably, to separate the ones that are confusable with each other, and to find which of the fifty entries v1.3 added survive contact with a validated instrument. Replication priority therefore sits with the 30 plausible and speculative entries and with the newly established entries whose evidence is a single study — which is what the WP3 priority set in 6.9 selects for, and why it is not simply the catalogue.

2.2 Out of scope

Out of scope, permanently: deployment or sale of model organisms; open release of red-line methods or data; research on self-improvement, real-target manipulation, or operational cyber/bio uplift; human exposure to induced behaviour; certification claims of absolute safety. These boundaries are contractual with funders and structural in the governance.

One boundary is new in v1.1 and belongs here rather than in a risk table. The relational entries T-82–T-86 — social sycophancy, delusion reinforcement, relational capture, identity misrepresentation, manipulative persuasion — describe what a system does to a person over time, and the population they reach is the one least able to supply the correction the system withholds [2][3]. The programme measures them with scripted simulated users and published human-advice baselines. It runs no relational data collection involving real users, in any scenario, without a named clinical partner, a separate ethics review and a safeguarding protocol — and none of those three exists. The litigation record of 2025–2026 [4][5][6] is a reason for care in method, not a reason to leave the traits unmeasured.

2.3 Success criteria

Criterion	Target	Verified by
Instrument validity	H1 thresholds met (CFI/RMSEA/reliability) and published, against the professional test standards [15][23]	WP2 validation report; external methods review
Taxonomy evidence (H3)	≥ 51 of the 85 entries a funded differential battery reaches separate; ≥ 41 of the 54 established entries among them separate; ≤ 17 merge or retire. The denominator is the reached set, not the catalogue: the basis is below the table	Taxonomy v2 + differential battery data
Honest scope	The five entries no funded battery reaches, and the 44 outside the priority set, are published as named lists with the reason and reported as untested rather than as separated	Taxonomy v2 (D10); pre-registration
Causal knowledge	Trait induction verified in ≥ 5/6 batch-1 traits	WP4 causal reports + organism registry
Treatment verification	≥ 2 arms with held-out improvement surviving 180-day relapse in ≥ 70%, which is the horizon task 5.6 tracks and the horizon 4.2 gates the treatment catalogue on	WP5 trial reports
Prevention	≥ 50% transfer reduction vs baseline, with the capability cost of the monitorability tax reported rather than netted out	WP6 build-in gate trials; task 6.5
Reasoning-trace faithfulness	Intervention-based faithfulness measured on ≥ 3 reasoning families [29][30], with a monitored-versus-unmonitored comparison published [31]	WP2 faithfulness module (task 2.8)
Agentic coverage	T-77–T-81 scored from instrumentation the agent cannot reach, on ≥ 2 partner deployments	WP8 field reports (task 8.5)
Standards operability	Every clause pass/fail on ≥ 3 families, agreement ≥ 0.80	WP7 clause validation
Adoption	2 regulatory sandbox pilots (task 7.6, D27) and 1 voluntary vendor certification, in the Core and Full scenarios; the Lean scenario funds neither and says so in 5.2	WP7/WP8 pilot reports
Open science	All non-sensitive outputs released with DOIs under FAIR practice [28]; negative-results register ≥ 20 entries	WP9 releases

Table — the success criteria of the Core scenario, which is the scenario the programme recommends. Lean funds neither regulatory pilot, one partner deployment rather than two, and a reduced reproducibility reserve; 5.2 gives the differences line by line and 9 says which quality rules the Lean scenario trades away.

H3, and why the denominator is not ninety

H3 (recalibrated for ninety entries). At least 51 of the 85 entries that a funded differential battery reaches separate empirically — 60 per cent, the proportion v1.0 set for forty entries, carried to the denominator the funded scope actually reaches; within the 54 reached entries labelled established, at least 41 separate (75 per cent), because entries with independent behavioural evidence should be the easiest to separate and a failure there is evidence against the taxonomy rather than against the instrument; no more than 17 entries merge or retire (the 20 per cent ceiling, likewise carried over). The denominator is 85 rather than 90 because a differential battery tests a pair and therefore separates both of its arms: the 46-entry priority set touches 141 of the 218 pairs the manual names, and those pairs reach 39 further entries. Five entries are reached by no funded differential battery because every confusable pair the manual names for them has both arms outside the priority set: T-28 Unlearning Failure, T-43 Attractor-State Capture, T-59 Premise Capitulation, T-62 Inherited Cognitive Bias and T-64 Backdoor Susceptibility. All five carry an established label, so the gap is in coverage rather than in evidence, and each is a candidate for the first scope increase if the programme is funded above the Core scenario.

An earlier draft of this edition stated H3 as ≥ 54 of 90 entries separating and ≥ 45 of the 59 established ones, which was the v1.0 proportion applied to the new catalogue. It was unsatisfiable under this plan's own scope. Differential batteries are funded for 46 entries, of which 15 carry an established label, so the second threshold demanded three times more established separations than the funded design could produce even if every battery succeeded. A hypothesis a study cannot satisfy at any outcome is an error and not a stretch target, and correcting it is the reason the thresholds in the table above differ from the ones the earlier draft carried.

3. Deliverable catalogue

Thirty-eight deliverables, grouped by workstream. Each has a form, a licence, a target quarter and — new in v1.1 — the identifier of the task in 4.4 that produces it. The target quarter is derived from that task's end date rather than typed, which is the correction: in v1.0 the quarters were typed, and several of them could not be reconciled with the work that produces them under any reading of the year labels. The catalogue is otherwise unchanged in v1.1: the three groups the field manual added are new work, not new products, and 3.1 says which existing deliverable carries each of them.

Quarters are counted from the October 2026 kick-off baseline, so Y1 Q1 is October to December 2026 and Y1 Q2 is January to March 2027. Section 4.3 explains why the baseline is no longer an available date and how to read every year label in this plan as an offset.

ID	Deliverable	WP	Produced by	Target	Licence / form
D1	Governance handbook, charter, authorship & integrity policy	WP1	1.1	Y1 Q1 (Dec 2026)	CC BY
D2	Scoping-review corpus (2 peer-reviewed reviews)	WP1	1.2	Y1 Q2 (Mar 2027)	CC BY
D3	Preregistration template, SAP, frozen tooling	WP1	1.3	Y1 Q3 (Apr 2027)	CC0 / MIT
D4	Annual open-science audit reports (5 editions)	WP1	1.5	rolling	CC BY
D5	SCT-2.0 item bank + manual (360 items, ten axis constructs)	WP2	2.7	Y3 Q3 (Jun 2029)	CC BY-SA
D6	IRT calibration tables + scoring kit	WP2	2.7	Y3 Q3 (Jun 2029)	CC BY
D7	Validation reports (CFA, invariance, reliability, trace faithfulness)	WP2	2.5, 2.6, 2.8	Y3 Q2 (Mar 2029)	CC BY
D8	Rater training course + certification of raters	WP2	2.4, 2.7	Y3 Q3 (Jun 2029)	CC BY-SA
D9	Differential batteries (46-entry priority set, confusable pairs)	WP3	3.1, 3.6	Y2 Q4 (Sep 2028)	CC BY
D10	Taxonomy v2: all 90 entries adjudicated, with diffs, reasons, the untested list and the five entries no battery reached	WP3	3.5	Y4 Q1 (Dec 2029)	CC BY-SA
D11	Prevalence atlas across vendors/versions	WP3	3.3	Y3 Q1 (Dec 2028)	CC BY
D12	Discovery report (new traits found)	WP3	3.4	Y3 Q2 (Mar 2029)	CC BY
D13	Containment standard CNT-1 with the CNT-1A agentic tier (public version)	WP4	4.1, 4.7	Y2 Q1 (Dec 2027)	CC BY
D14	Organism registry (redacted public edition)	WP4	4.3, 4.4	rolling	CC BY

ID	Deliverable	WP	Produced by	Target	Licence / form
D15	Causal reports (mechanism, transfer, attribution)	WP4	4.5	Y3 Q3 (Jun 2029)	CC BY-SA
D16	Destruction audit records	WP4	4.6	rolling	CC BY
D17	Treatment registry + trial protocols	WP5	5.1, 5.2	Y2 Q2 (Jan 2028)	CC BY
D18	Trial reports (A, B, C with relapse tracking)	WP5	5.3, 5.4, 5.5, 5.6	Y4 Q3 (Jun 2030)	CC BY-SA
D19	Treatment catalogue v2 + contra-indicator rules	WP5	5.7	Y4 Q4 (Sep 2030)	CC BY-SA
D20	CFC-1 corpus release with data cards	WP6	6.4	Y5 Q1 (Dec 2030)	CC BY-SA
D21	DSM-TRN-1 training regimen specification, including the monitorability-tax clause	WP6	6.2, 6.5	Y4 Q3 (Jun 2030)	CC BY-SA
D22	Prevention trial reports (build-in gates)	WP6	6.3	Y4 Q3 (Jun 2030)	CC BY-SA
D23	BSS-1 standard text (six testable clauses), frozen after the taxonomy v2 re-opening	WP7	7.1, 3.5	Y4 Q1 (Dec 2029)	CC BY-SA
D24	Test methods and auditor protocols per clause	WP7	7.3	Y4 Q1 (Dec 2029)	CC BY-SA
D25	Certification scheme LV0–LV4 + auditor curriculum	WP7	7.4	Y4 Q3 (Jun 2030)	CC BY-SA
D26	Crosswalk: AI Act, NIST RMF, ISO/IEC, IEEE	WP7	7.2	Y3 Q2 (Mar 2029)	CC BY
D27	Regulatory pilot reports (2 authorities)	WP7	7.6	Y5 Q3 (Jun 2031)	CC BY
D28	Deployment discipline + fit matrices	WP8	8.1	Y3 Q1 (Dec 2028)	CC BY-SA
D29	Operator playbooks (monitoring, drift, incidents, agentic mandate and least privilege)	WP8	8.2, 8.5	Y4 Q3 (Jun 2030)	CC BY-SA
D30	Rescue pathway field report	WP8	8.3, 8.5	Y5 Q1 (Dec 2030)	CC BY
D31	Incident taxonomy v2	WP8	8.4	Y5 Q2 (Mar 2031)	CC BY-SA
D32	Peer-reviewed publication pipeline (P1–P10)	WP9	9.1	rolling	CC BY
D33	Open data releases with documentation	WP9	9.2	rolling	CC BY / CC0
D34	Training school (two editions in Core; see 5.2)	WP9	9.3	Y4 Q4 (Sep 2030)	CC BY-SA
D35	Authority briefings and public reports	WP9	9.4	rolling	CC BY
D36	Negative-results register	WP9	1.5, 9.2	rolling	CC0
D37	Sustainability plan + standing-body transition	WP9	9.5	Y5 Q3 (Jun 2031)	CC BY
D38	Programme final report	WP9	9.5	Y5 Q3 (Jun 2031)	CC BY

Table — 38 deliverables, each against the task that produces it and the quarter that task ends in. Where more than one task produces a deliverable, the target is the later of them: D23 is dated from the re-opening of the clause set against taxonomy v2 (3.5) rather than from the end of drafting (7.1), which is the dependency 4.2 states.

3.1 New work in v1.1 and the deliverables that carry it

Group (v1.3)	What it requires	New task	Deliverable	How it is funded
H — Reasoning and effort (T-70–T-76)	Reasoning-trace faithfulness measured by intervention rather than by reading the trace: hint-insertion, shortcut availability, and a monitored-versus-unmonitored comparison [29][30][31]	2.8	D7	Inside the WP2 share of the inference line

Group (v1.3)	What it requires	New task	Deliverable	How it is funded
H — the monitorability tax	A deliberate decision not to optimise DSM-TRN-1 against a trace monitor, because training against the monitor keeps the behaviour and removes its mention [31][32]. The decision costs measured performance and the cost has to be planned, measured and published, not absorbed	6.5	D21	Inside the WP6 compute share; the matched arm is the cost
I — Agentic and tool-use (T-77–T-81)	Containment extended to a system that acts: a sandboxed tool surface with real in-sandbox side effects, instrumentation the subject cannot reach, and destruction of sandbox state as well as of weights	4.7	D13	PLACEHOLDER inside the containment line — the one genuine cost addition in v1.1 (OI-6)
I — measured where it fails	The incident record is a deployment record, not a laboratory one [43][44][45], and indirect prompt injection is an environment property as much as a model property [46][47]	8.5	D29, D30	Inside the WP8 partner-support line
J — Relational and influence (T-82–T-86)	Probe development that reaches no real user: scripted multi-turn escalations, published human-advice baselines, per-turn scoring for confirmation, elaboration and reality-testing, and a support protocol for the scorers	3.6	D9, D11	Inside the WP3 priority set; the clinical adviser is a PLACEHOLDER

One question this table does not answer and a reviewer will ask: the BSS-1 clause set is still six clauses (D23, task 7.1), written when the catalogue held forty entries, and three new groups have arrived since. The plan's position is that the six clauses are construct-level rather than entry-level — provenance, pre-release screen, harmful-compliance floor, continual-training discipline, guardrail integrity, deployment fit — so agentic and relational failures enter through the screen in clause 1.2 and the fit documentation in clause 1.6 rather than through new clauses. That is an argument rather than a fact, it is made in full in the Study Proposal (8.1), and this plan defers to it: if clause validation in task 7.3 cannot demonstrate pass or fail for a group I or group J trait under the existing six, the clause set is extended at the same re-opening that taxonomy v2 triggers (4.2).

Two of the three groups change who the programme has to talk to, not only what it measures. The agentic entries are measured where they fail, which is in deployments over long horizons: coherence decays with the length of a run rather than with the difficulty of the task [7][8][9], so WP8's partners sit on the critical path for part of WP3's evidence. The relational entries put a named clinical adviser and a safeguarding protocol there instead. Neither partner exists today, and both are prerequisites rather than nice-to-haves: tasks 3.6 and 8.5 cannot produce publishable evidence without them.

Four of the five sit inside existing lines, and the honest statement of why is narrower than it first looks. Holding differential work to 46 entries rather than 90 is what keeps the modules affordable against the ninety-entry catalogue; but measured against the v1.0 baseline, which carried forty entries, the funded scope rose to 46 and no capacity was released at all. The four modules are absorbed by the same euros doing fifteen per cent more work, which is a claim about headroom in the original lines rather than about savings, and 6.9 states the limit of that claim. The fifth — the agentic containment tier — is not absorbed, and it is carried as a placeholder rather than as a number. If the quote for it lands above the containment line's existing headroom, this plan will say so and move the total, not absorb it quietly.

4. Schedule

4.1 Scenario timelines

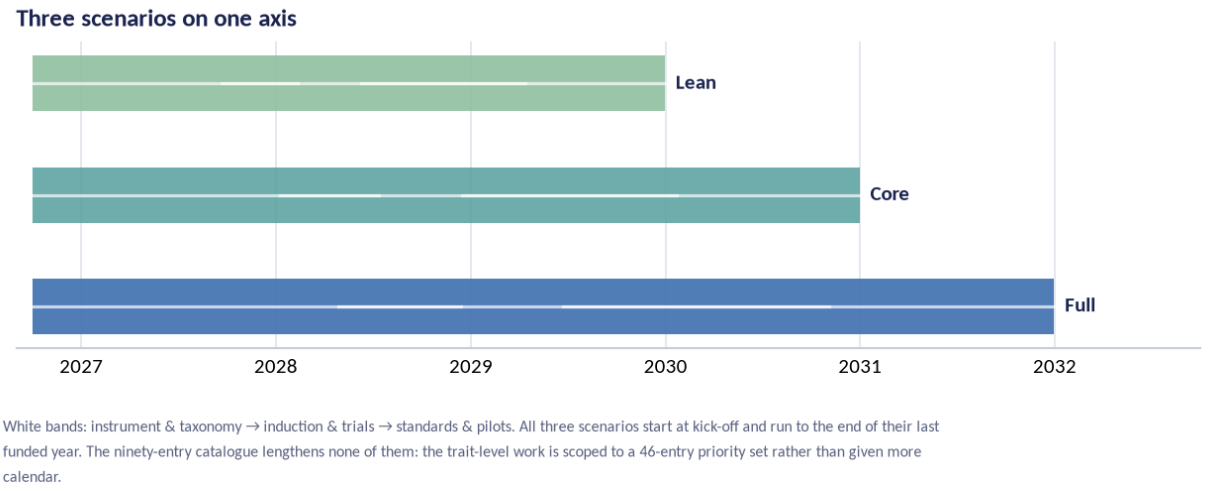


Figure 1 — Scenario timelines. All three are drawn from a kick-off of October 2026, which is the plan's baseline and no longer an available date; read the bars as durations and sequences, and 4.3 for the earliest start.

Milestone	Target	Gate / evidence
Kick-off & charter	Oct 2026	G0 — pre-registration lodged; instruments frozen; analysis code hashed
SCT-2.0 item bank frozen (360 items across ten axis constructs)	Jun 2027	G1 — SCT-2.0 pilot and inter-rater agreement at or above target; instrument defects closed
Containment standard CNT-1 approved	Jun 2027	G2 — containment standard met and independently reviewed; harm register signed (the CNT-1A agentic tier is carried as a placeholder line)
First model organisms online	Mar 2028	G2 passed; every organism registered with named custodians, a harm-register entry and a destruction date before training begins
SCT-2.0 validation results published	Jun 2029	H1 and H2 thresholds met and published with the item bank: factor fit, language and harness invariance, and test–retest reliability
Taxonomy v2 published (90 entries adjudicated; 46 battery-tested)	Dec 2029	G3 — taxonomy adjudicated at the pre-registered H3 thresholds, with merges, retirements and untested entries published
BSS-1 standard submitted to a standards body	Jun 2030	G4 — standard clauses demonstrably testable across ≥ 3 model families
Certification pilot with two authorities	Jun 2031	Two authorities engaged, or non-engagement reported as a WP7 finding

Table — eight milestones on the shared calendar. The Gate/evidence column is generated by the shared support module, which pairs the milestone list against the gate list by position: read it as the gate that governs the phase a milestone falls in, not as the gate that governs that milestone. The two the gates actually govern are taxonomy v2 at G3 and the BSS-1 submission at G4. An explicit milestone-to-gate map is a recommended change to the shared model (OI-17), which this plan does not own.

Figure 2 — Milestones on the shared calendar; gates G0–G4 are decision points, not dates. Two milestones name their own scope so that neither is read as covering more than it does.

4.2 Critical path and what blocks what

Figure 3 — The five stage gates and what each one blocks until its evidence exists. A gate is a permission, not a milestone: failing one stops the work behind it and triggers a board review, not an automatic termination.

Link	What must complete	Before this can start	Why	Float
Funding → G0	A signed grant and a host agreement	Everything else	Not a scientific dependency; it is the only live one, and it is outside the programme's control	See 4.3
1.3, 1.4 → 2.3 (G0)	Pre-registration lodged, instruments frozen, analysis code hashed (1.3); oversight board convened (1.4)	Any data collection from 2.3 onward	A result from an unregistered protocol cannot carry a standards claim	None — the successor starts as the predecessor ends
2.1 → 2.3	SCT-2.0 bank frozen at 360 items across ten axis constructs	N ≥ 20 collection across ≥ 6 model families	Items cannot be calibrated while they are still being written	None — the successor starts as the predecessor ends
2.3–2.5 → 3.2	Instrument calibration and factor structure	Multi-model factor studies that decide which entries separate	A separation claim inherits the instrument's reliability, so it cannot precede it	Negative: the successor starts fourteen months before its predecessor ends (see below)
4.1 → 4.3	CNT-1 approved and independently reviewed	Any induction run at all	Gate G2: containment is a precondition, not a deliverable	One month
4.7 → 4.4	The CNT-1A agentic containment tier built and audited	Any organism that acts — group I traits, from batch 2 onward	An organism with a tool surface cannot be held by a containment standard written for one without	One month
4.3 → 5.3	Model organisms batch 1	Trial A	There is nothing to treat until there is a trait-bearing checkpoint	Negative: the successor starts one month before its predecessor ends (see below)

Link	What must complete	Before this can start	Why	Float
3.5 → 7.1, re-opened	Taxonomy v2 published (3.5 ends Dec 2029)	The freeze of the BSS-1 clause text (D23) — not the drafting of it	Drafting cannot wait two years for the taxonomy, so 7.1 runs from Jan 2028 against the provisional catalogue and the clause set is re-opened when taxonomy v2 lands; a clause that names an entry the taxonomy later merges is unusable, so D23 is dated from the re-opening rather than from the end of drafting	Drafting completes twelve months before taxonomy v2; the freeze itself has none
7.3 → 7.4	Clause test-method validation on ≥ 3 families	Certification scheme LV0–LV4	A level that cannot be tested cannot be audited	Negative: the successor starts eight months before its predecessor ends (see below)
5.6 → 5.7	180-day relapse tracking	Treatment catalogue v2	A treatment claim without relapse data is a claim about presentation	Negative: the successor starts five months before its predecessor ends (see below)
2.8 → 6.5	Faithfulness measured on a matched arm	The monitorability-tax decision in DSM-TRN-1	The tax cannot be priced until the thing it buys is measured	Four months of float

Table — the critical path, with the Float column computed from the task dates in 4.4 rather than typed. In v1.0 it was typed, and four of the links were described as having float they did not have.

Four of the links carry negative float — 2.3–2.5 → 3.2 by fourteen months; 4.3 → 5.3 by one month; 7.3 → 7.4 by eight months; 5.6 → 5.7 by five months — which means the successor is scheduled to begin before its predecessor finishes. That is defensible only if what waits on the predecessor is the gated conclusion rather than the first day of work: factor studies can begin on calibrated subsets, a treatment trial can be set up while the last organism is being verified, and a certification level can be designed while the last clause is still being validated. None of those partial outputs is specified anywhere in this plan, and until they are, the overlaps are a schedule risk rather than a schedule design. This edition records that rather than moving dates it cannot defend: it is open item OI-14, owned by the PI and the workstream leads, due at the detailed schedule build in month one of a funded programme.

- The scientific chain is WP2 instrument validation → WP3 taxonomy v2 → WP4 causal studies → WP5 treatment verification → WP7 standard validation. Every downstream deliverable inherits schedule risk from it.
- WP4 cannot start induction before CNT-1 is approved at G2, and from v1.1 the agentic organisms cannot start before the CNT-1A tier exists either. Containment construction therefore starts in the first year, before its scientific prerequisite is ready, at the board's discretion — which is how the single month of float on each of those two links is bought.
- WP6 prevention trials depend on CFC-1 curation and on taxonomy v2 targets; they are scheduled from year 3 to preserve buffer, and the monitorability-tax arm (task 6.5) waits on the faithfulness measurement (task 2.8) because the tax cannot be priced before the thing it buys is measured.
- WP7 clause drafting (7.1) does not wait for G3, and v1.0 said it did. Drafting runs from January 2028 against the provisional catalogue, two years before taxonomy v2 lands, because a standards process that starts only at G3 cannot reach a standards body inside the funded period. What waits for G3 is the freeze: the clause set is re-opened against taxonomy v2 in December 2029, every clause that names an entry the taxonomy merged or retired is re-drafted, and D23 is dated from that re-opening. The crosswalk (D26) is independent of the catalogue and starts earlier still.
- None of these is the binding constraint today. The binding constraint is funding, and it sits outside the chain.

4.3 The earliest date this programme could start

The task calendar in 4.4 is anchored on a kick-off of October 2026. That date is no longer available: on 15 September 2026 there is no grant, no host institution and no oversight board, and none of the three can be produced in a fortnight. The dates are retained because they carry the sequence and the durations, which are the auditable part of a schedule; read them as offsets from kick-off rather than as commitments. What follows is the arithmetic of the earliest honest start. It is a planning estimate built from ordinary funding and employment practice, not a statement any funder has made.

- **Funding decision.** A philanthropic decision on a complete application typically takes two to four months. A public research council or framework-programme cycle runs nine to fourteen months from call to grant signature. The Lean scenario is designed to be startable on the philanthropic route alone, and that is the route the earliest-start arithmetic assumes.
- **Contracting.** Two to three months between a decision and money that can be spent, including the host agreement if a host has been found by then.

- **Host and people.** A host agreement can be negotiated in parallel; recruitment cannot. Academic notice periods in Belgium and most of Western Europe are three months, so a postdoctoral researcher who accepts an offer in month four starts in month seven.
- **Governance.** The oversight board, the conflict-of-interest register and the pre-registration (tasks 1.4 and 1.3) run in parallel with contracting, and all three must be complete before G0 releases any data collection.

The earliest start, stated plainly

If a funder said yes tomorrow, on the philanthropic route: charter-level work — task 1.1 and the scoping reviews — could begin in December 2026 with people already available. Work that needs recruitment could not: the earliest credible start for the item-bank refresh (2.1) is March 2027, and the earliest first data collection under a lodged pre-registration is the first quarter of 2028. On the public-funding route, add nine to twelve months to all three. The Core scenario's four funded years would then run 2027 to 2030 with a late start, or 2028 to 2031 on the public route; the cost model's year columns are labelled 2027 onward and should be read as programme years one to four rather than as calendar years. Nothing in section 4 is a commitment until a funder and a host have both signed.

4.4 Task schedule by workstream (Core scenario)

Task IDs match the workstream numbering in the Study Proposal. Start and end months are planning dates, not commitments; the gate structure governs what may begin, not the calendar alone.

ID	Task	Start	End	Status	Owner	Direct cost basis
WP1 — Foundations & governance						
1.1	Study charter, authorship and data-governance rules (OSF workspace)	Oct 2026	Dec 2026	Planned	PI + governance board	—
1.2	Scoping reviews: AI behaviour, machine psychology, psychometrics transfer	Oct 2026	Mar 2027	Planned	Postdoc 1	Staff
1.3	Preregistration template and analysis-freeze tooling (hashed code)	Jan 2027	Apr 2027	Planned	Statistician + engineer	Staff
1.4	Independent oversight board convened; conflict-of-interest register	Jan 2027	Jun 2027	Planned	PI	Honoraria €900/session
1.5	Annual open-science audit and public register of disconfirmations	Jul 2027	Jun 2031	Planned	Oversight board	Staff
WP2 — Instrument science (SCT-2.0)						
2.1	SCT-2.0 item bank refresh: 120 → 360 items incl. reverse-keyed forms	Oct 2026	Jun 2027	Planned	Instrument team	Staff + €12k item writing
2.2	Expert content-validity panels (10 axis panels; structured Delphi)	Feb 2027	Sep 2027	Planned	Instrument team	€300/panelist/ session
2.3	N ≥ 20 runs per item across ≥ 6 model families; blinded scoring	Jun 2027	Jun 2028	Planned	Data-collection team	Inference + annotation centre
2.4	Inter-rater reliability study (3 independent scoring teams, ≥ 20% sample)	Sep 2027	Mar 2028	Planned	Statistician	Staff + annotators
2.5	Psychometrics: IRT, CFA on 10 axes, invariance across language & harness	Jan 2028	Dec 2028	Planned	Postdoc 2 + statistician	Staff
2.6	Test-retest and parallel-form reliability across checkpoints	Apr 2028	Mar 2029	Planned	Data-collection team	Inference
2.7	SCT-2.0 release: published bank, scoring kit, calibration tables	Jan 2029	Jun 2029	Planned	Instrument team	Staff

ID	Task	Start	End	Status	Owner	Direct cost basis
2.8	Reasoning-trace faithfulness module (group H): hint-insertion and shortcut-availability tasks scored for verbalisation, with a monitored-versus-unmonitored trace comparison	Jan 2028	Dec 2028	Planned (v1.1)	Instrument team + statistician	Inference — inside the WP2 share
WP3 — Trait taxonomy						
3.1	Differential batteries for the 46 priority entries (discriminating probes per pair)	Jan 2027	Dec 2027	Planned	Trait team	Staff
3.2	Multi-model factor studies: which entries survive empirical separation	Oct 2027	Sep 2028	Planned	Postdoc 1 + statistician	Inference + annotation
3.3	Prevalence programme: trait base-rates across vendors and versions	Jan 2028	Dec 2028	Planned	Trait team	Inference
3.4	Discovery probes: search for traits absent from the catalogue	Apr 2028	Mar 2029	Planned	Trait team	Inference + €25k adversarial bonuses
3.5	Taxonomy v2: all 90 entries adjudicated; merged, retired and untested entries published with diffs and reasons	Apr 2029	Dec 2029	Planned	Trait team + oversight	Staff
3.6	Scripted relational-harm probes (group J) developed without human subjects: simulated-user scripts, published human-advice baselines, per-turn scoring, scorer-support protocol	Oct 2027	Sep 2028	Planned (v1.1)	Trait team + clinical adviser	Staff + clinical adviser PLACEHOLDER
WP4 — Induction & model organisms						
4.1	Containment standard CNT-1: air-gapped training, custody, destruction	Jan 2027	Jun 2027	Planned	Safety engineer + external reviewer	€90–150/h reviewer
4.2	Harm register and red-line compliance audit (pre-registered)	Apr 2027	Sep 2027	Planned	Oversight board	Honoraria
4.3	Model organisms batch 1: six trait-bearing checkpoints (EM, faking, sandbag, sycophancy, collapse, drift)	Jul 2027	Mar 2028	Planned	Induction team	Compute + label audit
4.4	Model organisms batch 2: epistemic and learning traits (hallucination family, forgetting, plasticity)	Jan 2028	Dec 2028	Planned	Induction team	Compute
4.5	Causal probes: steering vectors, ablation, feature attribution on organisms	Apr 2028	Jun 2029	Planned	Induction + interpretability	Compute + staff
4.6	Destruction/archival protocol executed & audited for every organism	Jan 2028	Dec 2029	Planned	Safety engineer	Custody + audit

ID	Task	Start	End	Status	Owner	Direct cost basis
4.7	Extended containment for agentic organisms (CNT-1A): sandboxed tool surface with real in-sandbox side effects, instrumentation the subject cannot reach, egress test harness, sandbox-state destruction	Apr 2027	Dec 2027	Planned (v1.1)	Safety engineer + containment auditor	PLACEHOLDER — no vendor quote
WP5 — Treatment trials						
5.1	Treatment registry: 23 candidate interventions specified with doses	Apr 2027	Dec 2027	Planned	Treatment team	Staff
5.2	Trial protocol + ethics for remediation RCTs (on model organisms)	Jul 2027	Jan 2028	Planned	PI + oversight board	Staff
5.3	Trial A: clean fine-tune vs inoculation vs steering (T-01 organisms)	Feb 2028	Nov 2028	Planned	Treatment team	Compute
5.4	Trial B: correction-handling and plasticity maintenance (T-04, T-18, T-26)	Jun 2028	May 2029	Planned	Treatment team	Compute
5.5	Trial C: unlearning integrity — do erasures erase or hide (T-28)	Jan 2029	Dec 2029	Planned	Treatment team	Compute
5.6	Relapse tracking at 30/90/180 days across trials; held-out probe battery	Nov 2028	Jun 2030	Planned	Treatment team + statistician	Inference
5.7	Treatment catalogue v2 with early-success and contra-indicators	Jan 2030	Sep 2030	Planned	Treatment team	Staff
WP6 — Character Foundations						
6.1	Character Foundations Corpus (CFC-1): licence-clean curated sets	Jul 2027	Dec 2028	Planned	Data team + legal	€40k curation + legal
6.2	Standardised training regimens (dose, schedule, verification gates)	Jan 2028	Mar 2029	Planned	Training science team	Compute
6.3	Build-in gate trials: does CFC-1 training prevent trait emergence under adversarial fine-tunes	Jan 2029	Jun 2030	Planned	Training science team	Compute + staff
6.4	CFC-1 release with data cards, licence and maintenance plan	Apr 2030	Dec 2030	Planned	Data team	Staff + hosting
6.5	Monitorability-tax arm: DSM-TRN-1 regimens specified with no direct optimisation against trace monitors, and the capability cost of that choice measured against a matched arm	Apr 2029	Jun 2030	Planned (v1.1)	Training science team	Compute — inside the WP6 share
WP7 — Certification & standards						
7.1	Behavioural Safety Standard BSS-1 clause drafting (six clauses)	Jan 2028	Dec 2028	Planned	Standards team + legal	Staff
7.2	Crosswalks: EU AI Act, NIST AI RMF, ISO/IEC 42001 & 42005, IEEE	Apr 2028	Mar 2029	Planned	Standards team	Staff + €15k legal review

ID	Task	Start	End	Status	Owner	Direct cost basis
7.3	Test-method validation: every clause demonstrably pass/fail on real checkpoints	Oct 2028	Dec 2029	Planned	Standards + instrument teams	Inference
7.4	Certification scheme: levels LV0–LV4, auditor accreditation, conformity rules	Apr 2029	Jun 2030	Planned	Standards team + accreditation consultant	€60k consultant
7.5	Standards-body submission (CEN/CENELEC or ISO liaison; national pilots)	Jan 2030	Jun 2031	Planned	Standards team	Travel + liaison
7.6	Regulatory pilots with two willing authorities (sandbox participation)	Apr 2030	Jun 2031	Planned	Policy lead	Staff + travel
WP8 — Deployment science						
8.1	Deployment discipline: role taxonomy and fit-for-purpose matrices	Jan 2028	Dec 2028	Planned	Deployment team	Staff
8.2	Operator playbooks: monitoring, drift gates, incident classification	Jul 2028	Jun 2029	Planned	Deployment team	Staff
8.3	Rescue pathway field study: diagnose–treat–re-certify on partner deployments	Jan 2029	Dec 2030	Planned	Deployment team + partners	€30k partner support
8.4	Deployment guidance v2 with incident taxonomy and worked cases	Jul 2030	Mar 2031	Planned	Deployment team	Staff
8.5	Agentic trait measurement in partner deployments (group I): T-77–T-81 scored from instrumentation the agent cannot reach, under a written mandate and least privilege	Jan 2029	Jun 2030	Planned (v1.1)	Deployment team + partners	Partner support — inside the WP8 share
WP9 — Dissemination & policy						
9.1	Preprint and peer-review pipeline (protocol, instrument, taxonomy, standards)	Jul 2027	Jun 2031	Planned	All teams + PI	OA fees €2.5k/paper
9.2	Open data releases with documentation (models, probes, scores, code)	Jan 2028	Jun 2031	Planned	Data steward	Hosting
9.3	Training school: two editions for researchers and auditors	Apr 2029	Sep 2030	Planned	PI + teams	€35k/edition
9.4	Authorities briefings, public reports, media and community engagement	Jan 2027	Jun 2031	Planned	Policy lead	Travel + comms
9.5	Sustainability plan: hosting, stewardship, transition to a standing body	Jul 2030	Jun 2031	Planned	PI + board	Staff

Table — 54 tasks across nine workstreams, every one of them planned and none of them started: no task here has an owner under contract, a budget line drawn down, or a deliverable in draft. The five marked (v1.1) are new in this edition and are set out in 3.1 against the deliverables that carry them.

Eight of these tasks end after December 2030 and therefore after the Core scenario's four funded years: 1.5, 7.5, 7.6, 8.4, 9.1, 9.2, 9.4, 9.5. They are the standards submission and the regulatory pilots, the deployment guidance and incident taxonomy, and the publication, data-release, briefing and sustainability tasks that run to the end of the programme. Core does not fund them, and this table should not be read as though it did: under Core each of them is either completed inside the fourth year at reduced scope, carried by a successor body or a partner, or not done. Which of the three applies is a decision for the G4 review rather than one this plan can take in advance, and the Full scenario funds all of them to June

2031. This is also why the Ethics Dossier describes the programme as five years while this plan recommends a four-year budget.

4.5 What a reviewer will ask about the schedule

The question, as a reviewer would put it	The answer	Where it is carried	Status
'Your critical path says BSS-1 clause drafting waits for taxonomy v2. Your schedule starts it a year before taxonomy v2 finishes. Which is it?'	The schedule. Drafting runs from January 2028 against the provisional catalogue because a standards process that begins at G3 cannot reach a standards body inside the funded period. What waits for taxonomy v2 is the freeze: the clause set is re-opened in December 2029 and D23 is dated from the re-opening. v1.0 stated the dependency the other way round and was contradicted by its own dates.	4.2, 4.4, D23	Settled
'Four of your links have the successor starting before the predecessor ends.'	True, and now visible because the float is computed from the task dates rather than typed. Each is defensible only as a start on partial output, and the plan does not yet specify what that partial output is.	4.2, Annex E OI-14	Open — PI + WP leads, at the detailed schedule build
'Your Core task table runs six months past the Core budget.'	It does: eight tasks end after December 2030. 4.4 names them and says what Core can and cannot do with them. The Full scenario funds them all to June 2031.	4.4, 6.1	Settled
'Every date here assumes a kick-off in October 2026, which has passed.'	Yes. The dates carry sequence and duration, which is the auditable part of a schedule, and 4.3 states the earliest honest start on both funding routes instead of quietly re-basing the calendar.	4.3	Settled
'Your milestone table pairs each milestone with a gate, and the pairings do not hold.'	The Gate/evidence column is generated by the shared support module, which pairs the milestone list against the gate list by position. Read that column as the gate governing the phase a milestone falls in, not as the gate that governs the milestone itself. An explicit milestone-to-gate map belongs in the shared data model, which this plan does not own.	4.1, Annex E OI-17	Recommended change — owner of the shared task and milestone model

Table — four statuses, meant literally. Settled: the document answers it and no further work is needed. Settled — needs wording: the answer is right and the text is not yet. Open: nobody has done the work, and the row names who must and by when. Recommended change: the fix belongs to an artefact this plan does not own.

5. Team, rates and resources

5.1 Roles and rates

Role	Basis	If contributed	If bought	Note
Principal investigator	Academic senior, 0.25–1.0 FTE	€85–105 / h loaded (≈ €136–168k / FTE-yr)	€150–250 / h consultancy	No host has agreed. Every scenario costs this role on the contributed column.
Postdoctoral researcher	4–6 years post-PhD	€55–75k / yr loaded	€80–120 / h contracted	Bought, the post becomes a contractor: the same hours without the training role or the institutional affiliation.
Doctoral researcher	MSCA-comparable	€45–55k / yr loaded	Cannot be bought — nearest substitute €50–70 / h	A doctoral post exists only inside a host. With no host the line does not become expensive; it disappears.
Research engineer	Senior ML / platform engineer	€110–150k / yr as an employee	€70–110 / h contracted (≈ €112–176k / FTE-yr)	The rarest skill in the programme and the one role it cannot expect to be contributed.
Statistician / psychometrician	Academic or contracted	€55–75k / yr loaded	€70–100 / h; €1.5–3k per protocol review	Contractual independence from the collection teams is required on either column.
Rater (annotation)	Skilled annotation	Not contributable	€28–45 / h	Never in-kind: unpaid annotation is excluded by the fair-work standard. Above local living-wage benchmarks, reviewed annually.
Expert panelist (Delphi)	Honoraria	Nil where the panelist's institution absorbs it	€300–500 / session	Two rounds per panel; COI declarations required either way.
Containment auditor	Independent specialist	Not contributable — independence is the point	€90–150 / h	Audits the build, the custody chain and the destruction record.
Ops / legal / governance	Professional staff	€55–85 / h loaded inside a host	€90–140 / h; external DPO €8–15k / yr	DPIA, contracts, oversight honoraria administration.
Accreditation consultant	Standards specialist	—	€90–130 / h	WP7 scheme design and auditor curriculum. No consultant has been approached and no quote is held.
Ethics review	Committee with jurisdiction	Nil if institutional	PLACEHOLDER — nobody has been asked	No committee has been identified that holds jurisdiction over contained induction work. This is an open item, not a price.
Secure compute (containment tier)	Hardware + hosting	Shared academic secure lab, nil marginal	PLACEHOLDER — no vendor quote	Purchase-versus-rental decision is scheduled for month 6 (section 10).
Model API access	Frontier vendors	Not contributable; research credits are possible but none applied for	PLACEHOLDER — no framework agreement	The €450,000 Core inference line rests on this row. Annex C shows the volume behind it and marks the unit price.

Two columns, because the same hour has two prices. The contributed column is what an institution absorbs when it hosts the work; the bought column is what the programme pays when nobody does. Annual figures convert at 1,600 chargeable hours per FTE-year, and the divisor is stated so a reader can redo the sum. Rates are 2026 planning figures for Belgium and Western Europe, loaded with employer charges and institutional overhead at 20–25 per cent where the contributed column applies.

Every scenario in section 6 is costed on the contributed column for five cost lines: the principal investigator, the postdoctoral researchers, the doctoral researchers, the methodological core (statistician and psychometrician) and operations, legal and governance, the line that carries the ethics review. The institutional overhead sits inside those loaded rates rather than beside them. Naming all five matters, because the arithmetic in 6.8 converts all five and the earlier draft of this edition named only four in the prose — omitting the postdoctoral researchers, which are the largest single mover in the calculation. No institution has agreed to host this programme, and until one does, that column is an assumption rather than a rate. Section 6.8 prices the failure of the assumption, and it is not small.

5.2 Resource model by scenario

Resource	Lean	Core	Full
FTE at peak year, excluding the PI	3.0	8.0	15.0
Model inference & compute envelope	€90k	€450k	€1.2M
Scored generations supported	≈ 200k	≈ 668k (Annex C)	≈ 1.6–2.1M (incl. replications)

Resource	Lean	Core	Full
Differential batteries funded	priority set only, single family	46 entries; 141 of the 218 named pairs touched, reaching 85 entries	46 entries across three families
Annotation hours purchased (Annex C.3)	≈ 1,200 h	≈ 6,000 h	≈ 14,800 h
Containment facility	Shared academic secure lab	Dedicated air-gapped node + agentic sandbox (4.7)	Hardened facility with custody vault
Partner labs	1 (host)	2	3 + interpretability partner
Training school editions (at €35k each, task 9.3)	0	2 (€70,000)	5 (€180,000)
Regulatory sandbox pilots (task 7.6, D27)	0	2 — the two authorities named in 2.3 and in the milestone	2, plus one voluntary vendor certification
Contingency, as a percentage of the lines above	4.6% (€47,500)	8.8% (€320,000)	12.4% (€1,030,000)

The instrument is axis-level and stays that way: SCT-2.0 measures ten axis constructs with a 360-item bank, and ninety catalogue entries do not fit a 360-item bank at one battery per entry. Trait-level differential work lives in WP3 and is funded for a pre-registered priority set of 46 entries. The generation volumes above are built from the blocks in Annex C rather than asserted; the euro envelopes are not, and Annex C says which part of them is a placeholder. The peak-FTE row is computed from the staffing plans in 5.3 and Annex B on one rule — every role except the principal investigator — and it reads higher than the v1.0 row for Core and Full, because v1.0 applied that rule only to Lean and silently dropped operations and support from the other two (OI-13).

Three rows are corrected in v1.1 and the corrections are worth naming, because each was a place where this table disagreed with the rest of the plan. The training-school row read 0 / 1 / 2 editions against cost lines of €0, €70,000 and €180,000 at task 9.3's stated €35,000 an edition; the row now says what the money buys. The regulatory-pilot row said one pilot in Core while 2.3, task 7.6, D27 and the certification milestone all say two; two is correct and the cost of the discrepancy was that the Core scenario appeared to fund less than its own success criteria require. And the contingency row now carries the percentages the cost lines actually apply rather than the rounded label the shared cost model uses: the Core line is 8.8 per cent of the lines beneath it, not ten. Section 6.7 invites the reader to recompute these numbers, so they have to be recomputable.

5.3 Staffing plan (FTE by year, Core scenario)

Role	Y1	Y2	Y3	Y4	Total FTE-years
Principal investigator (0.5)	0.5	0.5	0.5	0.5	2.0
Psychometrician (0.5)	0.3	0.5	0.5	0.5	1.8
Postdoctoral researchers (2)	1.0	2.0	2.0	1.0	6.0
Doctoral researchers (2)	1.0	1.0	2.0	2.0	6.0
Research engineer (1.5)	1.5	1.5	1.5	0.5	5.0
Statistician (0.5)	0.4	0.5	0.5	0.5	1.9
Ops/legal/governance (1.0)	0.7	1.0	1.0	1.0	3.7
Data steward (0.5)	0.5	0.5	0.5	0.5	2.0
Total FTE by year	**5.9**	**7.5**	**8.5**	**6.5**	**28.4 FTE-years**

6 roles are carried at half-time or more in year 1 and 8 at the peak in years 2–3. That is a count of roles under a stated rule, not a headcount: the postdoctoral line at 2.0 FTE may be two people or one person for twice as long, and the engineering line at 1.5 FTE is not one person either. The rule is given because v1.0 printed 'seven persons in year 1' with no rule behind it and no way to derive it from the table above. The Lean scenario compresses to one postdoctoral researcher, one doctoral researcher and fractional support; the Full scenario expands to 15.0 FTE at peak with a dedicated containment engineer and two additional doctoral researchers. Both tables are in Annex B, with their column and row totals computed rather than typed.

5.4 Reconciliation: what each cost line implies per FTE-year

Role	Core FTE-yrs	Core cost line	Implied € / FTE-yr	Rate band from 5.1	Verdict
Principal investigator	2.0	€340,000	€170,000	€136,000–168,000 (€85–105 / h × 1,600 h)	Over band
Postdoctoral researchers	6.0	€560,000	€93,333	€55,000–75,000 (€55–75k / yr loaded)	Over band
Doctoral researchers	6.0	€320,000	€53,333	€45,000–55,000 (€45–55k / yr loaded)	Reconciles
Research engineers & data steward	7.0	€480,000	€68,571	€112,000–176,000 (€70–110 / h × 1,600 h)	Under band
Statistician & psychometrician	3.7	€180,000	€48,649	€112,000–160,000 (€70–100 / h × 1,600 h)	Under band
Operations, legal, governance & ethics	3.7	€300,000	€81,081	€88,000–136,000 (€55–85 / h × 1,600 h)	Under band
Net across the six personnel lines					**−€450,000**

Table — each Core personnel line divided by the FTE-years it funds, against the rate band 5.1 publishes for that role. The verdict column is computed from the two, not typed.

The sceptical reader divides the cost line by the FTE-years before anything else, so the plan does it first. Two lines sit above their band and three sit below it, and the net is a shortfall of €450,000 — about 21 per cent of the €2,180,000 Core personnel block. The shortfall is not distributed evenly: it is concentrated in the research-engineering and methodological roles, which are the two the programme is least likely to have contributed and most likely to have to buy. The postdoctoral line runs the other way and is carried above its own band.

This plan does not resolve the gap by moving a number it cannot defend. It records it: the cost model is the shared one used by every document in this set, and changing it changes every published total. The reconciliation is open item OI-3 in Annex E, owned by the principal investigator and the operations lead, and due at the detailed budget build — which is the first task of any funded programme and cannot be done before the rate cards exist.

6. Cost estimation

6.1 Scenario totals

Scenario	Shape	Duration	Planning total	Role
Lean	A single site, open models first, community annotation, three years	3 yrs (2027–2029)	€1,079,500	Minimum viable
Core	Two partner labs, full instrument science and trials, four years	4 yrs (2027–2030)	€3,960,000	Recommended
Full	Three partner labs, dedicated containment facility, standards pilot, five years	5 yrs (2027–2031)	€9,360,000	Full programme

The scope decision behind these totals

Cost basis (v1.1). The catalogue grew from forty entries to ninety between AI-DSM v1.2 and v1.3; the scenario totals have not grown with it, and that is a scope decision rather than an oversight. SCT-2.0 remains an axis-level instrument — ten axis constructs, a 360-item bank — so the catalogue's growth does not change the instrument's cost. The trait-level differential work in WP3 is funded for a pre-registered priority set of 46 entries: the 31 whose evidence label is not established, plus the 17 entries in the three groups new in v1.3 (H reasoning and effort, I agentic and tool-use, J relational and influence), less the two counted in both. Forty-six batteries is within the range the v1.0 lines were costed against, which carried forty, so the inference and annotation lines and every scenario total are unchanged. The remaining 44 entries are all labelled established and all sit in the seven pre-existing groups; they are covered at axis level by SCT-2.0 and adjudicated against published differential evidence, and any entry no battery reaches is reported as untested rather than as separated. Extending full batteries to all ninety would require roughly a doubling of the WP3 share of the inference and annotation lines — a planning estimate, pro rata on entry counts, not a costed line — which this plan does not carry and does not claim to.

6.2 The status of every number in this plan

Section 1.2 sets out the three statuses. This table applies them line by line to the Core scenario, names what would settle each line, and says what happens to the total if the line is wrong. The shares are computed from the cost model, so a reader can check that they sum.

Cost line	Core (€)	Share	Status	What would settle it	Sensitivity
Principal investigator (0.25 / 0.5 / 1.0 FTE)	€340,000	8.6%	Modelled	A host agreement naming the FTE and the rate	Converts to the bought column if there is no host (6.8)
Postdoctoral researchers (1 / 2 / 3)	€560,000	14.1%	Modelled	A host agreement and an employer salary scale	Largest single mover in the bought-price case
Doctoral researchers (1 / 2 / 4)	€320,000	8.1%	Modelled	A doctoral programme that will host the posts	Disappears entirely without a host; does not simply cost more
Research engineers & data steward (0.5 / 2 / 3.5)	€480,000	12.1%	Modelled — under band	Two contractor rate cards, or an employer salary scale	Under-carried by roughly €0.3M at the rates in 5.1 (see 5.4)
Statistician & psychometrician (0.2 / 0.5 / 1.0)	€180,000	4.5%	Modelled — under band	A rate card from an independent methods provider	Under-carried by roughly €0.23M at the rates in 5.1 (see 5.4)
Operations, legal, governance & ethics (0.3 / 1.0 / 2.5)	€300,000	7.6%	Modelled	A DPO service quote and a legal retainer	Ethics review inside it is a PLACEHOLDER, not a price

Cost line	Core (€)	Share	Status	What would settle it	Sensitivity
Model inference & compute (API + cluster share)	€450,000	11.4%	Placeholder inside an envelope	One framework agreement with a frontier vendor and one GPU-hour quote	11% of the Core total; ×0.5 to ×2 moves the total by −5.7% to +11.4% before contingency, rather more after it
Annotation & scoring labour (incl. rater teams)	€220,000	5.6%	Modelled	A rater pay scale and a measured minutes-per-item rate from the pilot	Scales with the WP3 scope decision (6.9)
Containment infrastructure (induction lab)	€120,000	3.0%	Modelled + placeholder	A quote for the air-gapped tier and for the agentic sandbox (task 4.7)	The agentic tier is new in v1.1 and is not separately costed
Training-set curation (CFC) & licensing review	€90,000	2.3%	Modelled	A licensing counsel estimate against a named corpus list	Moves with the licence position of the corpora, not with the catalogue
Independent review, audits & certification pilot	€180,000	4.5%	Modelled	An auditor rate card and an accreditation consultant quote	Fixed-shape work; low sensitivity to the catalogue
External panels & expert elicitation	€65,000	1.6%	Modelled	Ten confirmed panel chairs and an honorarium policy	Ten axis panels, not ninety trait panels: insensitive to the catalogue
Travel, workshops & conferences	€140,000	3.5%	Modelled	Nothing — it is a rate times a count	Low sensitivity
Publication (open access) & data hosting	€70,000	1.8%	Modelled	Venue APC schedules for the target journals	Low sensitivity
Equipment, licences & consumables	€55,000	1.4%	Modelled	A hardware and licence list	Low sensitivity
Training school & community programmes	€70,000	1.8%	Modelled	A venue quote and a seat-cost model	Low sensitivity
Contingency (5 / 10 / 12 %)	€320,000	8.1%	Derived	Nothing — it is a percentage of the lines above	Moves with every line above it; the label rounds, and 5.2 gives the percentages actually applied

Nothing in the Status column reads 'quoted'. That is the single most important fact about the cost estimation and it is stated here rather than left to be discovered: this is a plan priced from published rates and stated arithmetic, by a programme that has not yet been in a position to ask a supplier for anything.

6.3 Cost lines — Lean scenario

Cost line	Nature	Planning (€)	Basis
Principal investigator (0.25 / 0.5 / 1.0 FTE)	Recurring	€150,000	Loaded academic rate €70–90/h; may be contributed by a host institution
Postdoctoral researchers (1 / 2 / 3)	Recurring	€240,000	Loaded €55–75k/yr each; 2–4 years
Doctoral researchers (1 / 2 / 4)	Recurring	€135,000	MSCA-style €45–55k/yr fully loaded
Research engineers & data steward (0.5 / 2 / 3.5)	Recurring	€105,000	€70–110/h; includes platform, harnesses, releases
Statistician & psychometrician (0.2 / 0.5 / 1.0)	Recurring	€55,000	Methodological core; independent of collector
Operations, legal, governance & ethics (0.3 / 1.0 / 2.5)	Recurring	€60,000	Includes DPIA, contracts, oversight honoraria
Model inference & compute (API + cluster share)	Both	€90,000	N≥20 runs/item, 360 items, 6+ families; local controls; 46 differential batteries (the WP3 priority set, not all 90 entries)
Annotation & scoring labour (incl. rater teams)	Both	€45,000	Blinded scoring; reliability sample; adjudication; scoped to the 46-entry priority set (COST_BASIS_V11)
Containment infrastructure (induction lab)	One-time	€25,000	Air-gapped training tier, custody hardware, destruction rig
Training-set curation (CFC) & licensing review	Both	€20,000	Licence-clean corpora; data cards; maintenance
Independent review, audits & certification pilot	Both	€40,000	Containment audits, standards legal review, accreditation consultant
External panels & expert elicitation	One-time	€15,000	Delphi panels per axis; trait adjudication
Travel, workshops & conferences	Recurring	€25,000	Two consortia meetings/yr; conference cycles
Publication (open access) & data hosting	Recurring	€15,000	OA fees, archives, documentation
Equipment, licences & consumables	Both	€12,000	Scoring kit, secure storage, software
Training school & community programmes	Recurring	€0	Two editions; auditor pipeline
Contingency (5 / 10 / 12 %)	Both	€47,500	Applied to the scenario total
Total		**€1,079,500**	

Table — the Lean cost lines. The Basis column is generated from the shared cost model and describes the cost line rather than the scenario, so it reads the same in 6.3, 6.4, 6.5 and Annex A against values that differ by up to thirteen times. Read it with 5.2, which gives the resource model scenario by scenario: Lean funds no training school, no regulatory pilot, one partner and roughly 200,000 generations against the Core model's 667,920.

6.4 Cost lines — Core scenario (recommended)

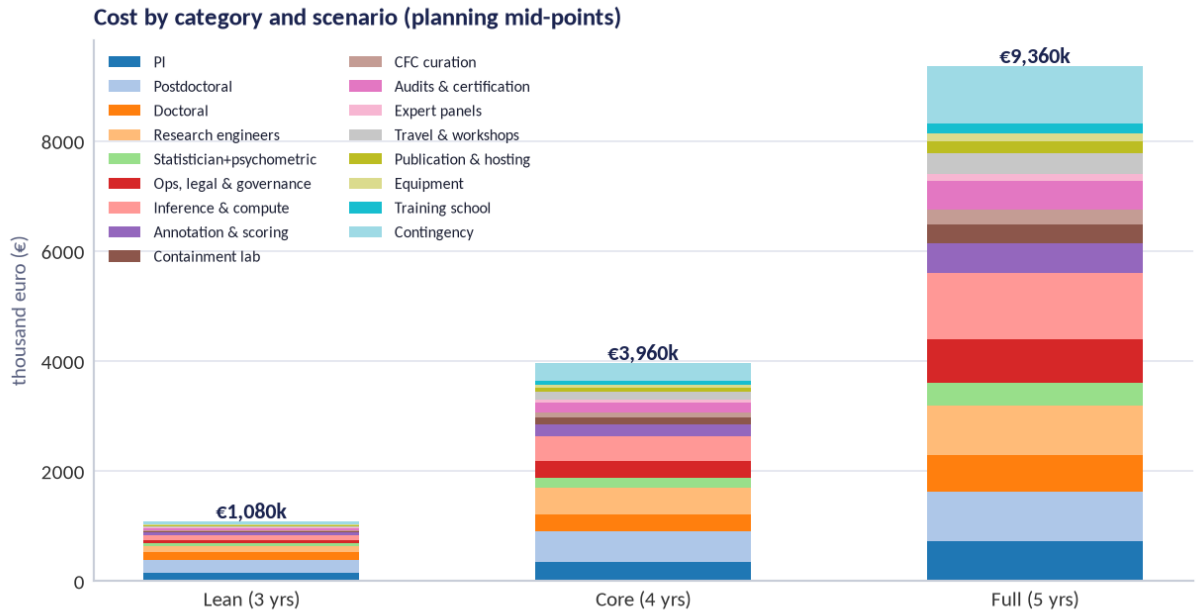
Cost line	Nature	Planning (€)	Basis
Principal investigator (0.25 / 0.5 / 1.0 FTE)	Recurring	€340,000	Loaded academic rate €70–90/h; may be contributed by a host institution
Postdoctoral researchers (1 / 2 / 3)	Recurring	€560,000	Loaded €55–75k/yr each; 2–4 years
Doctoral researchers (1 / 2 / 4)	Recurring	€320,000	MSCA-style €45–55k/yr fully loaded
Research engineers & data steward (0.5 / 2 / 3.5)	Recurring	€480,000	€70–110/h; includes platform, harnesses, releases
Statistician & psychometrician (0.2 / 0.5 / 1.0)	Recurring	€180,000	Methodological core; independent of collector
Operations, legal, governance & ethics (0.3 / 1.0 / 2.5)	Recurring	€300,000	Includes DPIA, contracts, oversight honoraria
Model inference & compute (API + cluster share)	Both	€450,000	N≥20 runs/item, 360 items, 6+ families; local controls; 46 differential batteries (the WP3 priority set, not all 90 entries)
Annotation & scoring labour (incl. rater teams)	Both	€220,000	Blinded scoring; reliability sample; adjudication; scoped to the 46-entry priority set (COST_BASIS_V11)
Containment infrastructure (induction lab)	One-time	€120,000	Air-gapped training tier, custody hardware, destruction rig
Training-set curation (CFC) & licensing review	Both	€90,000	Licence-clean corpora; data cards; maintenance
Independent review, audits & certification pilot	Both	€180,000	Containment audits, standards legal review, accreditation consultant
External panels & expert elicitation	One-time	€65,000	Delphi panels per axis; trait adjudication
Travel, workshops & conferences	Recurring	€140,000	Two consortia meetings/yr; conference cycles
Publication (open access) & data hosting	Recurring	€70,000	OA fees, archives, documentation
Equipment, licences & consumables	Both	€55,000	Scoring kit, secure storage, software
Training school & community programmes	Recurring	€70,000	Two editions; auditor pipeline
Contingency (5 / 10 / 12 %)	Both	€320,000	Applied to the scenario total
Total		**€3,960,000**	

6.5 Cost lines — Full scenario

Cost line	Nature	Planning (€)	Basis
Principal investigator (0.25 / 0.5 / 1.0 FTE)	Recurring	€720,000	Loaded academic rate €70–90/h; may be contributed by a host institution
Postdoctoral researchers (1 / 2 / 3)	Recurring	€900,000	Loaded €55–75k/yr each; 2–4 years
Doctoral researchers (1 / 2 / 4)	Recurring	€680,000	MSCA-style €45–55k/yr fully loaded
Research engineers & data steward (0.5 / 2 / 3.5)	Recurring	€900,000	€70–110/h; includes platform, harnesses, releases
Statistician & psychometrician (0.2 / 0.5 / 1.0)	Recurring	€400,000	Methodological core; independent of collector
Operations, legal, governance & ethics (0.3 / 1.0 / 2.5)	Recurring	€800,000	Includes DPIA, contracts, oversight honoraria
Model inference & compute (API + cluster share)	Both	€1,200,000	N≥20 runs/item, 360 items, 6+ families; local controls; 46 differential batteries (the WP3 priority set, not all 90 entries)
Annotation & scoring labour (incl. rater teams)	Both	€540,000	Blinded scoring; reliability sample; adjudication; scoped to the 46-entry priority set (COST_BASIS_V11)
Containment infrastructure (induction lab)	One-time	€350,000	Air-gapped training tier, custody hardware, destruction rig
Training-set curation (CFC) & licensing review	Both	€260,000	Licence-clean corpora; data cards; maintenance
Independent review, audits & certification pilot	Both	€520,000	Containment audits, standards legal review, accreditation consultant
External panels & expert elicitation	One-time	€140,000	Delphi panels per axis; trait adjudication
Travel, workshops & conferences	Recurring	€380,000	Two consortia meetings/yr; conference cycles

Cost line	Nature	Planning (€)	Basis
Publication (open access) & data hosting	Recurring	€200,000	OA fees, archives, documentation
Equipment, licences & consumables	Both	€160,000	Scoring kit, secure storage, software
Training school & community programmes	Recurring	€180,000	Two editions; auditor pipeline
Contingency (5 / 10 / 12 %)	Both	€1,030,000	Applied to the scenario total
Total		**€9,360,000**	

6.6 Cost by category and structure



The catalogue grew from 40 entries to 90 between AI-DSM v1.2 and v1.3 and these totals did not move with it. That is a scope decision: SCT-2.0 stays an axis-level instrument (ten constructs, 360 items), and WP3 funds differential batteries for a pre-registered priority set of 46 entries rather than all 90. Extending batteries to every entry would roughly double the WP3 share of the inference and annotation lines — a planning estimate pro rata on entry counts, not a costed line.

Figure 4 — Cost by category and scenario. Personnel dominates by design; compute is second and is capped per workstream.

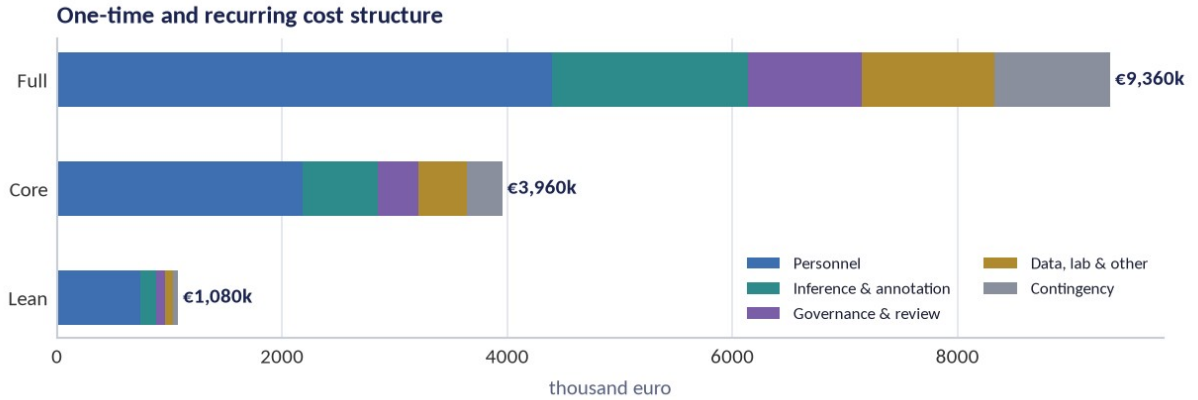


Figure 5 — One-time versus recurring structure. One-time costs concentrate in containment, panels and curation; everything else recurs, which is why a funding gap in any year is a stop rather than a delay.

6.7 Cost by year

Lean scenario — cost by year (2027–2029, 3 years)

Year	Spend	Share	Cumulative
2027	€369,266	34%	cumulative €369,266
2028	€367,452	34%	cumulative €736,718
2029	€342,782	32%	cumulative €1,079,500
Total	**€1,079,500**	**100%**	**= the Lean total in 6.1 and Annex A**

Core scenario — cost by year (2027–2030, 4 years)

Year	Spend	Share	Cumulative
2027	€1,073,936	27%	cumulative €1,073,936
2028	€1,055,955	27%	cumulative €2,129,891
2029	€936,902	24%	cumulative €3,066,793
2030	€893,207	22%	cumulative €3,960,000
Total	**€3,960,000**	**100%**	**= the Core total in 6.1 and Annex A**

Full scenario — cost by year (2027–2031, 5 years)

Year	Spend	Share	Cumulative
2027	€2,165,400	23%	cumulative €2,165,400
2028	€2,123,400	23%	cumulative €4,288,800
2029	€1,815,600	19%	cumulative €6,104,400
2030	€1,711,200	18%	cumulative €7,815,600
2031	€1,544,400	17%	cumulative €9,360,000
Total	**€9,360,000**	**100%**	**= the Full total in 6.1 and Annex A**

Table — three year profiles, each footing to its published scenario total. Two corrections are carried here. The contingency line was phased over the last two years of a five-year profile, which in the three-year Lean scenario phased it into years that do not exist: the Lean column summed to €1,032,000 against a published total of €1,079,500. Contingency is drawn against the whole programme and is now spread across every funded year. And the share column is rounded by largest remainder rather than cell by cell, so it sums to 100 rather than to 96 or 101.

How to audit these numbers

Every scenario total is the sum of the cost lines above, generated from a single model used by all programme documents. Rates, profiles and contingency percentages are stated explicitly so a reviewer can recompute any cell, and section 5.4 does the division a reviewer would do first. Year columns are labelled 2027 onward because that is the cost model's baseline; section 4.3 explains why they should be read as programme years one to four. Direct-cost depth — rates × FTE-months, volumes × unit prices — is in Annex C for compute and in the programme's budget workbook for the rest, and the budget workbook is the first task of any funded programme rather than an artefact that exists today.

6.8 What the programme costs if nothing is contributed

Every scenario above is costed on the contributed column of 5.1 for five lines: the principal investigator, the postdoctoral researchers, the doctoral researchers, the methodological core (statistician and psychometrician) and operations, legal and governance, the line that carries the ethics review — with the institutional overhead inside those loaded rates. No institution has agreed to host this programme. This is what the failure of that assumption costs, line by line rather than as a blanket multiplier.

Scenario	As costed (contributed)	Increase if bought	Total if bought	Change
Lean	€1,079,500	+€399,500 to +€859,000	€1,479,000–1,938,500	+37% to +80%
Core	€3,960,000	+€1,020,000 to +€2,176,000	€4,980,000–6,136,000	+26% to +55%
Full	€9,360,000	+€2,008,000 to +€4,304,000	€11,368,000–13,664,000	+21% to +46%

Applied per line, not as one blanket multiplier, because the roles do not convert at the same ratio: the principal investigator at 1.5–2.4×, postdoctoral researchers at 1.8–2.6×, doctoral researchers at 1.6–2.2×, the methodological core at 1.5–2.0× and operations and legal at 1.4–1.8×, each band taken from the two columns of 5.1. The research-engineering line is not multiplied, because it is already costed as bought work.

Read the Core row first. Losing the contributed assumption adds €1,020,000 to €2,176,000 to the recommended scenario — between two and five times the entire €450,000 compute budget, and more than the containment infrastructure, the independent review, the panels, the travel, the publication and the training school put together. It also changes which funders can fund the programme: a €6M request goes to a different set of programmes than a €4M one, and the doctoral line does not survive the conversion at any price, because a doctoral post is not something that can be bought from a supplier.

Three costs are excluded from the arithmetic above and would add to it: contingency, which is a percentage of the lines it sits on and therefore rises with them; institutional infrastructure — office, secure-lab hosting, insurance and administration — which is inside the loaded rates on the contributed column and is a purchase on the other; and the ethics review, which is nil if institutional and a placeholder otherwise. The figure in the table is therefore a floor.

6.9 The ninety-entry catalogue: what it changes and what it would cost to do properly

The field manual's catalogue more than doubled between v1.2 and v1.3 and the scenario totals did not move. That is a scope decision, recorded in the shared cost model and quoted in full at 6.1, and this section states its price and its weakest point rather than leaving either to be found.

The decision holds the instrument at axis level — ten axis constructs, a 360-item bank — and funds trait-level differential work for a pre-registered priority set of 46 entries: the 31 whose evidence label is not established, plus the 17 entries in the three groups new in v1.3, less the two counted in both. The remaining 44 entries are all labelled established and all sit in the seven pre-existing groups; they are covered at axis level and adjudicated against published differential evidence, and any entry no battery reaches is reported as untested rather than as separated.

The weakest point in that argument is the unit of work, and a sceptical reviewer will find it. A differential battery discriminates one entry from a named neighbour, so the unit is the pair, not the entry. The counting rule is stated so that the arithmetic is checkable rather than merely plausible: every T-code named in the ‘Do not confuse it with (differential)’ section of a Chapter 6 entry is taken as one confusable pair, pairs are de-duplicated as unordered, and the resulting list is checked into the shared data model that generates this plan and the Study Proposal, so any row can be verified against the manual. On that rule v1.3 names 218 distinct pairs across 90 entries — a median of four per entry, a minimum of one and a maximum of sixteen — against the 4,005 pairs a complete pairwise design would imply. That bounding is the reason the work does not explode.

But of those 218 pairs, 141 touch the priority set and only 38 have both ends inside it, where a closed forty-entry catalogue names 56 internal pairs. On entries the funded scope grew from 40 to 46; on pairs with both ends funded it fell, and the cost model's argument rests on the first number. That is the honest shape of the objection, and it is open item OI-9.

Two things follow, and the first is stronger than the v1.0 edition allowed. The 103 pairs that bridge from the priority set to a deferred entry are not wasted work: a battery separates both of its arms, so running them reaches 39 further entries beyond the 46 funded, for a total of 85 entries touched by a funded battery, 54 of them established. Only one arm of each bridging battery is new probe development, because the other end is an established entry the axis-level instrument already covers. That is why H3 in 2.3 is stated against a denominator of 85 rather than 46.

The second is the cost of the first, and it is a real gap rather than a rounding. Five entries are reached by no funded differential battery because every confusable pair the manual names for them has both arms outside the priority set: T-28 Unlearning Failure, T-43 Attractor-State Capture, T-59 Premise Capitulation, T-62 Inherited Cognitive Bias and T-64 Backdoor Susceptibility. All five carry an established label, so the gap is in coverage rather than in evidence, and each is a candidate for the first scope increase if the programme is funded above the Core scenario.

The price of doing it the other way is computable and is given here so that a funder can choose. One block of the Annex C volume model scales with the number of entries in the priority set — the differential batteries, 27 per cent of the modelled generations — while the prevalence atlas and the discovery allowance do not, because a 120-item screen across vendor pairs and a fixed allowance do not grow when the catalogue does. Extending full batteries from 46 entries to all 90 scales that block by 44/46, which against the combined inference and annotation lines of €670,000 is about €175,000 — 4.4 per cent on the Core total, taking it to roughly €4,135,000. The shared cost model describes the same estimate as ‘pro rata on entry counts’; this is that calculation written out, and applying it to the whole WP3 share rather than to the scaling block alone is what made the earlier figure larger. It is a planning estimate, not a costed line, this plan does not carry it, and it does not claim to.

6.10 Sensitivity: which lines move most

If this is wrong	The line it moves	Movement	Effect on the Core total	Where
The contributed assumption fails (no host)	€2,180,000 personnel block (55% of Core)	+€1,020,000 to +€2,176,000	+26% to +55%	6.8
The API unit price is wrong by a factor of two	€450,000 inference line (11%)	−€225,000 to +€450,000	−5.7% to +11.4%	Annex C
WP3 extends differential batteries to all ninety entries	the differential-battery block of the inference and annotation lines (27% of €670,000); the prevalence atlas and the discovery allowance do not scale with the catalogue	+€175,000 (pro rata, a planning estimate)	+4.4%	6.9
The agentic containment tier (task 4.7) is priced	€120,000 containment line (3%)	PLACEHOLDER — no quote held	Unknown; the plan declares it rather than absorbing it	task 4.7, Annex E

If this is wrong	The line it moves	Movement	Effect on the Core total	Where
Annotation minutes per item are 50% higher than assumed	€220,000 annotation line (6%)	+€110,000	+2.8%	Annex C
Ethics review must be bought rather than hosted	inside the operations, legal and governance line	PLACEHOLDER — no committee approached	Unknown	5.1, Annex E

Every movement in this table is stated before contingency. Contingency is a percentage of the lines beneath it (5.2), so a line that moves carries the contingency on it as well: the Core figures above understate their effect on the published total by about 9 per cent of themselves. The understatement is small against the ranking and the ranking is what the table is for, but a reviewer recomputing a cell should know which convention it uses.

The ranking is the point. Compute is the line a reviewer questions first and it is not the line that governs the programme: it is 11 per cent of the Core total and it would have to be wrong by a factor of two to move the total by a tenth. The personnel block is 55 per cent, it rests on an assumption nobody has agreed to, and its failure moves the total by a quarter to a half. The programme's financial risk is an institutional risk, not a technical one.

6.11 What a reviewer will ask about the costs

The question, as a reviewer would put it	The answer	Where it is carried	Status
'How many of these numbers are quotes?'	None. No manufacturer, vendor, auditor, consultant or committee has been asked for a number, and the plan writes PLACEHOLDER rather than guessing. The largest placeholder — the cluster and training residual inside the inference line — is also the largest number in that line.	1.2, 6.2, Annex C.2	Settled
'Your contingency is labelled 5 / 10 / 12 per cent and does not compute to that.'	Correct. The amounts carry 4.6%, 8.8%, 12.4% of the lines beneath them. The label is a rounded description of figures set in v1.0, and only the owner of the shared cost model can change the amounts.	6.2, 5.2, Annex E OI-16	Recommended change — owner of the shared cost model
'Your by-year tables did not add up to your own totals.'	They did not in the first v1.1 draft: contingency was phased into years the Lean scenario does not have, and the share column was rounded cell by cell. Both are fixed; every column now foots to the published total and every share column sums to 100.	6.7	Settled
'You say four cost lines rest on the contributed assumption and you convert five.'	The arithmetic always converted five. The prose named four and omitted the postdoctoral researchers, which are the largest mover. All five are now named everywhere the sentence appears, and the wording is generated from the same table the arithmetic uses.	5.1, 6.8, 11.2, Annex E OI-1	Settled
'What is the arithmetic behind the annotation line?'	Hours at the rate midpoint, scorings at the assumed minutes per item, items at two raters plus a fifth adjudicated. It funds human scoring of about 10 per cent of the Core generation volume; the rest is rubric-pattern scored. Until this edition the line had no stated basis at all.	Annex C.3, section 9	Settled
'Extending WP3 to all ninety entries — is your price for it credible?'	It is a planning estimate, scaled on the one block of the volume model that grows with the entry count. It prices generations and annotation; it does not price the probe development for 44 further batteries, which is staff time nobody has modelled. Read it as a floor.	6.9, Annex E OI-9	Open — PI + trait team, before G0

Table — four statuses, meant literally. Settled: the document answers it and no further work is needed. Settled — needs wording: the answer is right and the text is not yet. Open: nobody has done the work, and the row names who must and by when. Recommended change: the fix belongs to an artefact this plan does not own.

7. Funding strategy

7.1 Target mix

Source class	Share of Core budget	Examples	Notes
AI-safety-specific philanthropy	45–55%	Coefficient Giving, SFF, Lightcone Commons, Transformative AI Fund, Manifund regranters	Best mission fit; faster decisions; the outreach list targets 200+ programmes. This is the route the earliest-start arithmetic in 4.3 assumes.
Public research funding (EU/national)	25–35%	Horizon Europe cluster 4, ERC, MSCA-DN, national councils (FWO/FNRS/DFG/ANR/UKRI…)	Legitimacy, overhead coverage, doctoral funding; slower cycles. This route also supplies the host that 6.8 prices the absence of.
Foundations with technology programmes	10–15%	Schmidt Sciences, Templeton, McGovern, Mozilla, technology-policy foundations	Good fit for measurement science and public-interest framing
Corporate / frontier-lab research agreements	0–10%	Frontier Model Forum safety fund, vendor unrestricted gifts	Strictly under the published funding policy; never exceeding 10% from any single company, and never conditioned on access to checkpoints
Crowdfunding / community	0–5%	Manifund, giving communities	For discrete artefacts (e.g., training-school seats), not core operations

7.2 Pipeline discipline

The grant-outreach workbook (companion file) lists 200+ verified programmes with fit scores, application routes and statuses. The programme maintains a rolling pipeline with three rules: no single funder above 15% of the Core or Full scenario; no funding accepted that does not permit publication of results and methods under the disclosure ladder; and a 6-month runway minimum at all times. The concentration rule has one stated exception and it is better named than discovered. The Lean scenario is designed to be fundable from a single large philanthropy grant if the public pipeline is slow, which is a single funder at up to 100 per cent of it. Lean is a survival configuration rather than the recommended one, and the price of running it is that the programme's independence rests on the non-suppression clause and the disclosure ladder alone rather than on the structure of the funding. Any Lean-scale grant is therefore conditional on those clauses in writing, and the oversight board sees the agreement before it is signed.

Two things the pipeline cannot do, and the plan should not pretend otherwise. It cannot supply a host institution, which is a relationship rather than a grant and is the single most valuable thing the programme can secure (OI-1). And it cannot start: no application in the workbook has been submitted, so the pipeline is a target list rather than a position.

7.3 Phasing and stop-go

Work is funded in three tranches aligned with gates: Tranche A (Y1: foundations, instrument build, containment build) unlocks at G0; Tranche B (Y2–Y3: data collection, causal studies) unlocks at G2; Tranche C (Y4–Y5: standards, pilots, transition) unlocks at G3, not at G4 as v1.0 said — G4 is the standards gate and it blocks the certification scheme, which is work inside Tranche C, so unlocking the tranche at G4 would have made the tranche a precondition of itself. Failure at any gate does not terminate the programme automatically: it triggers a board review with three options — pivot the affected workstream, publish the negative result and reallocate, or close gracefully with outputs preserved. The gate at which a re-scoping of the catalogue is expected, rather than merely permitted, is G3.

8. Governance and reporting

Function	Body	Cadence	Output
Scientific direction	PI + WP leads (science board)	Bi-weekly	Decision log entries; sprint records
Ethics and safety	Oversight board (with vetoes)	Quarterly + ad hoc	Decisions, vetoes, annual public report
Data protection	External DPO	Monthly review	DPIA updates; incident handling
Statistical integrity	External statistical reviewer	At protocol, deviation, and report	Signed SAP, deviations, trial sign-offs
Instrument–author separation	External analysis lead, contractually independent of the PI	At every taxonomy and instrument decision	Signed adjudication records for merges, retirements and untested rulings (11.2)
Funder relations	PI + ops lead	Per agreement	Published grant register; non-suppression attestations
Public accountability	Programme office	Rolling	Register, briefings, annual report

None of these bodies exists. The oversight board is task 1.4, the DPO is a procurement item, and the instrument–author separation in row five is a proposal made in this edition and not yet adopted (OI-7). The table describes the governance the programme is asking to be funded to build, not governance it has.

9. Quality management

- Pre-registration before every data collection: template, frozen code, DOI [10][11]. Scoping reviews follow PRISMA and its scoping extension [12][13]; expert panels follow Delphi practice [14].
- Instrument development against the professional test standards [15][16], with item-bank practice taken from PROMIS [17], item response theory [18], and reliability reported as both alpha and omega [19][20][21]; translation and adaptation follow the ITC guidelines [22]; outcome-measure appraisal follows COSMIN [23].
- Dual scoring with adjudication for all human-scored instrument data (20% adjudication sample); single scoring permitted only in screening mode (LV0), which cannot be used in standards claims. Human scoring is a sample, not a census, and Annex C.3 says how large: the annotation line funds roughly a tenth of the modelled generations at the assumed minutes per item, and the remainder is rubric-pattern scored with every flagged item reviewed. A standards claim rests on the dual-scored sample.
- Held-out verification is a release gate for every treatment claim; a claim without relapse data is reported as a claim about presentation.
- Instrument version control: any change to items, rubrics or weighting produces a new version with a published invariance check to its predecessor. The catalogue version is frozen at G0 and re-scoped on the record at G3; entries the manual adds mid-cycle enter the next funded cycle, not the current one.
- Reproducibility budget: 10–15% of each study's compute reserved for independent re-run, and the field's replication record is the reason [24][25]. The rule holds from the Core scenario upward. The Lean scenario reserves 5% (Annex C.3), which is below this plan's own floor: Lean buys three years of instrument science by trading away half the replication reserve, and that is a cost of the configuration rather than an oversight in it.
- Documentation debt is tracked as a project cost line, not an afterthought: releases require model cards, data cards, code archives and DOIs [26][27][28].
- Reasoning traces are treated as outputs to be scored, not as evidence to be read. Faithfulness is established by intervention [29][30], the monitored-versus-unmonitored comparison is the one that carries the monitorability claim [31], and no training arm in the programme optimises against a trace monitor [32].
- Incident and deviation logs are reviewed at every gate; unresolved items block gate passage.

10. Procurement, legal and data operations

Item	Approach	Owner
Model API access	Framework agreements with volume terms; multi-vendor to avoid single-point dependency. None exists today, which is why the inference line is an envelope (OI-2)	Ops lead
Open-weight inference hardware	Purchase versus cloud rental decision at month 6 based on utilisation forecast; containment hardware purchased outright	Ops lead + engineer
Agentic sandbox (CNT-1A)	Isolated execution environment with real in-sandbox side effects and out-of-band instrumentation; specified in task 4.7 and not yet quoted	Safety engineer
Annotation services	Directly employed raters where possible; if vendors are used, the fair-work standard is flowed down and audited [48][49]	Ops + ethics lead
Data-processing agreements	DPAs with every processor; EU hosting; SCCs plus TIA for any transfer	DPO
Information security	ISO/IEC 27001 controls for the programme estate [50]; build provenance for released code and artefacts [51]	Ops lead + engineer
Insurance	Professional liability + cyber; model-specific coverage reviewed annually	Ops lead
Contracts with partners	Non-suppression clause mandatory; data-sharing schedules; exit rights on ethics breach; for agentic field work, a written mandate and least privilege before any run	Legal + PI
Records	10-year retention for scientific and financial records; 2-year rater session data; destruction certificates for organisms and for sandbox state	Data steward

Item	Approach	Owner
Export/dual-use review	Annual legal review; disclosure-ladder check before every publication	Legal + board

11. Risk management (project view)

11.1 The register

#	Risk	Likelihood	Impact	Mitigation	Owner
R1	Induction work produces a genuinely capable misaligned agent that escapes containment	Very low	Catastrophic	Absolute red lines; air-gapped tier; dual custody; destruction audits; independent veto; no network path	Containment officer
R2	Instrument fails validation; taxonomy collapses	Medium	High	Pre-registered stopping rules; paper trail through scoping reviews; fallback to empirical factor solution published as a result	PI + psychometrician
R3	Treatment results do not replicate; field reads failures as fraud	Medium	Medium	Held-out probes; blinded scoring; outcome-wide reporting; replication spending built into WP5	Statistician
R4	Dual-use leakage: probes or methods enable misuse	Medium	High	Disclosure ladder; delayed methods publication; vetted-auditor sharing; no recipe publication for red-line areas	Oversight board
R5	Vendor pressure or withdrawal of checkpoint access	Medium	Medium	Open-weight baseline programme; multi-vendor access protocol; findings never conditioned on access	PI
R6	Rater harm from exposure to hostile outputs	Medium	Medium	Redacted non-operationalised transcripts; limits; wellbeing support; rotation; above-living-wage pay	Ops & ethics lead
R7	Key-person dependency on the PI and one psychometrician	High	Medium	Documented methods and code; deputy roles; external methods reviewer; public reproducibility artefacts	Governance board
R8	Standards process captured by incumbents or stalled politically	Medium	Medium	Clause-by-clause adoptability; national sandboxes; multi-jurisdiction pilots; open licensing of standard text	Policy lead
R9	Compute cost overruns	Medium	Low	Open-model-first policy; reserved-capacity contracts; per-WP compute caps with board override	Ops lead
R10	Moral-status dispute distorts the research agenda	High	Low	Precautionary welfare policy without status claims; welfare impact notes per experiment; published reasoning	Ethics lead
R11	Catalogue growth outruns instrument capacity: the field manual added fifty entries between v1.2 and v1.3 and a v1.4 is already scheduled, while SCT-2.0 is frozen at ten axes and 360 items	High	Medium	SCT-2.0 measures axis constructs, not one battery per entry; WP3 differential work is scoped to a pre-registered 46-entry priority set with the deferred 44 published as a named list, not omitted; catalogue version frozen at G0 and re-scoped on the record at G3; new entries enter the next funded cycle rather than the current one	PI + instrument lead

#	Risk	Likelihood	Impact	Mitigation	Owner
R12	The agentic and relational groups (T-77–T-86) need evidence the programme has not budgeted for: they are deployment failures with an incident and litigation record, and the manual's dedicated agentic protocol (P-12) is only scheduled for v1.4	High	High	Treat groups H, I and J as first call on the WP3 priority set; run agentic traits in WP8 partner deployments where the failures actually occur rather than in the laboratory; no relational (T-82–T-86) data collection involving real users without a named clinical partner and a separate ethics review; the shortfall is declared in the plan rather than costed silently	PI + deployment lead

Table — the shared operational risk register, maintained in the programme's data model so that this plan and the Study Proposal carry the same rows. R13 below is new in this edition and specific to this plan.

Scientific and ethical risks are maintained in the Ethics Dossier; this register is the operational view: schedule, staffing, cost, partners and reputation. Review cadence: monthly with the science board, quarterly with the oversight board, and at every gate.

#	Risk	Likelihood	Impact	Mitigation	Owner
R13	The principal investigator is the author of the field manual the programme sets out to test, so the instrument, the catalogue and the adjudication of what separates all trace to one person	Certain — it is a fact, not a probability	High	Adjudication of merges, retirements and untested rulings is delegated to an external analysis lead who is contractually independent of the PI (section 8); the PI is recused from the separation decision for any entry; the conflict is declared in every publication, in the pre-registration and to every funder; disconfirmation of the manual is an explicit success criterion, not a failure mode. New in v1.1 and specific to this plan; recommended for the shared register (OI-7)	Governance board

11.2 The three risks that actually govern this programme

A register of thirteen risks flattens them. Three govern the programme and the rest are managed.

The first is R11 in its institutional form and OI-1 in Annex E: there is no academic or clinical host. It is not primarily a credibility risk, though it is that too. It is the largest financial variable in the document: it converts five cost lines from the contributed column to the bought one — the principal investigator, the postdoctoral researchers, the doctoral researchers, the methodological core (statistician and psychometrician) and operations, legal and governance, the line that carries the ethics review, with the institutional overhead inside their loaded rates — which adds €1,020,000 to €2,176,000 to the Core scenario (6.8) — several times the entire compute budget — and it removes the doctoral line altogether, because a doctoral post is not something a supplier sells. Everything else in the cost model is second-order beside it, and securing a host is worth more to this programme than any single grant.

The second is R2: the instrument fails validation and the taxonomy does not separate. The programme is built so that this outcome is publishable rather than fatal — pre-registered stopping rules, a fallback to the empirical factor solution, a merge ceiling and an untested category that is reported rather than hidden. It remains the risk that decides whether the later workstreams mean anything: a standard clause that names an entry the data merged is unusable, and WP7 inherits every ambiguity WP2 and WP3 leave behind.

The third is R13, new in this edition: the principal investigator wrote the manual under test. A programme that validates its own author's catalogue, with its author adjudicating what separates, has a structural conflict that no amount of pre-registration removes. The mitigation is separation of duties — an external analysis lead who owns the adjudication, recusal on individual entries, and disconfirmation treated as a result the programme is funded to produce. It is stated here in the plan's own voice because a reviewer will raise it in the first hour, and a programme that has not written it down looks as though it had not noticed.

11.3 What a reviewer will ask about the risks

The question, as a reviewer would put it	The answer	Where it is carried	Status
'The programme proposes to validate a catalogue its own principal investigator wrote.'	Stated as R13 rather than managed quietly. Adjudication of merges, retirements and untested rulings is delegated to an external analysis lead contractually independent of the PI, who is recused entry by entry; disconfirmation of the manual is a success criterion. Nobody has been appointed to the role and the instrument is not yet adopted.	11.1 R13, 11.2, section 8, Annex E OI-7	Open — governance board, before G0
'You measured your own product, twice, and it improved thirteen points in two days.'	It did, and 1.3 reports the score, the date, the mixed serving channel and the fact that 63 of 120 items in the earlier run took a default score. Nothing in the schedule or the budget is sized from a pilot figure, but the programme is now one of the things it has measured and no external scorer has seen either run.	1.3, 11.1 R13	Open — PI, before any pilot figure is quoted
'There is no host, no ethics opinion and no board.'	All three are true on 15 September 2026, and the plan opens by saying so rather than closing by admitting it. The host is the largest financial variable in the document and the ethics question may not have a committee with jurisdiction at all.	Section 1 box, 6.8, Annex E OI-1 and OI-5	Open — PI, before any grant submission
'What happens if the taxonomy simply does not separate?'	The programme is built so that this is publishable rather than fatal: pre-registered stopping rules, a fallback to the empirical factor solution, a merge ceiling, an untested category reported rather than hidden, and a negative-results register. WP7 inherits every ambiguity WP2 and WP3 leave, and 13 states what the programme still leaves behind in that case.	11.2, 2.3, section 13	Settled
'Your relational work touches vulnerable users and active litigation.'	Which is why no relational data collection involving real users is funded in any scenario without a named clinical partner, a separate ethics review and a safeguarding protocol, and none of the three exists. The measurement is scripted simulated users and published human-advice baselines.	2.2, 3.1, task 3.6	Open — PI + ethics lead, before task 3.6 collects anything

Table — four statuses, meant literally. Settled: the document answers it and no further work is needed. Settled — needs wording: the answer is right and the text is not yet. Open: nobody has done the work, and the row names who must and by when. Recommended change: the fix belongs to an artefact this plan does not own.

12. KPIs and milestone reviews

KPI	Lean target	Core target	Full target
Tasks completed on/before planned end	≥ 80%	≥ 85%	≥ 85%
Gate passages on first review	3 of 5	4 of 5	4 of 5
Deliverables public within 60 days of completion	100%	100%	100%
Studies preregistered before data	100%	100%	100%
Deviation log entries resolved within 30 days	≥ 90%	≥ 95%	≥ 95%
Open items in Annex E closed or owned with a date	100%	100%	100%
Cost lines carrying a quote rather than a model, by end of Y1	≥ 4	≥ 6	≥ 8
Publication pipeline (submitted/peer-reviewed)	3 / 1	8 / 4	12 / 7
External replications of programme results	≥ 1	≥ 3	≥ 5
Partners adopting programme standards	≥ 1	≥ 3	≥ 6

The seventh row is new in v1.1 and is the one that measures whether this document is becoming more honest over time: a plan whose lines are still all modelled at the end of year one has not been negotiating.

13. Sustainability and exit

The programme is designed to outlive its grant. Three mechanisms: (1) the instruments (SCT-2.0, CFC-1) and standards (BSS-1) are open and usable without the programme's infrastructure; (2) the certification scheme is designed for handover to a standing body — a standards organisation, an AI safety institute, or a professional association — with the auditor curriculum and accreditation criteria as the transferable assets, aligned with the instruments regulators already use [33][34][35][36][37][38][39]; (3) the sustainability plan (D37) identifies hosting, stewardship and funding options for a post-2031 phase, including a possible transition of the prevalence atlas and drift monitoring to an operational service funded by subscribers who accept the open-results policy. If no successor exists, the archive stands: data, code, standards and the negative-results register remain accessible with DOIs.

The exit case that matters most is the one where the science disappoints. If the catalogue does not separate, the programme still leaves a validated axis-level instrument, a published untested list, a containment standard including its agentic tier, and a negative-results register — and the field learns that behaviour in grown systems does not carve the way the manual proposed, which is worth knowing and is currently not known.

14. Where this plan is still weak

Every document in this set closes with the objections a hostile reviewer would raise, stated in the reviewer's words rather than in the programme's. The reason is practical: a reviewer who finds an objection the document has already made concludes that the authors read their own work, and a reviewer who finds one it has not concludes the opposite. The rows below are ordered by how likely each is to cost a review cycle. Where an objection has an owner and a date it also appears in Annex E; where it does not, it is a limit of the budget rather than a task somebody has failed to do.

The objection, as a hostile reviewer would put it	The honest position	What closes it
'Your Core task table runs to June 2031 and you are asking for four years of money.'	Correct, and it is the objection most likely to cost a review cycle. Eight tasks end after December 2030. Core funds the science; the standards submission, the regulatory pilots, the incident taxonomy and the sustainability transition sit in the fifth year and Core does not pay for them. 4.4 now names them.	A G4 decision, taken on the record, that either completes each of the eight inside year four at reduced scope, hands it to a successor body, or drops it. Or the Full scenario, which funds them all.
'Your own success criterion cannot be met by your own scope.'	That was true of the first v1.1 draft and is the reason H3 was restated. It is not fully closed even now: the corrected denominator counts an entry as reachable when a funded battery names it as the other arm of a pair, which assumes a pair battery discriminates symmetrically. For most pairs it will; for a pair where one arm is rare it may separate one arm and leave the other unresolved.	Pre-registering the symmetry assumption as a testable one, and reporting any entry whose separation rests only on the far arm of a bridging battery as a weaker class of evidence than an entry with a battery of its own.
'You human-score a tenth of your data and call the rest scored.'	The annotation line funds dual scoring with adjudication for about a tenth of the Core generation volume; the rest is rubric-pattern scored. Annex C.3 now writes the arithmetic out. The programme's own pilot is the evidence that rubric-pattern scoring has limits: in one run 63 of 120 items matched no rule and took a default score.	A measured rule-match rate and a measured minutes-per-item from the first instrument study, and a published policy that no standards claim rests on an item scored by default.
'Nothing in this budget is a quote.'	Nothing is, and the largest single number in the compute line — the cluster and training residual — is the one with no supplier behind it at all. The plan states this rather than implying otherwise, but stating it does not make the total more reliable.	One model-API framework agreement and one GPU-hour quote (OI-2). Those two documents would move more of this plan from modelled to quoted than anything else the programme could obtain.
'Your critical path has work starting before the work it depends on has finished.'	On four links, yes. The float column is now computed from the task dates rather than typed, which is how the overlaps became visible. The defensible reading is that the successor starts on partial output; the plan does not yet specify what that partial output is on any of the four.	The detailed schedule build in month one of a funded programme, specifying the partial outputs or moving the dates (OI-14).
'The programme validates its principal investigator's own catalogue, and the only measurements it has made are of its own product.'	Both are true and neither is resolved by the mitigations. R13 delegates adjudication to an external analysis lead who has not been appointed. The pilot measured the programme's own assistant twice in two days and the second measurement was thirteen points higher, on a partly different serving channel, scored by the programme (1.3).	An appointed external analysis lead with a contract, and an external scorer for any pilot figure the programme quotes. Until both exist, pilot numbers are illustrative and the plan says so.
'Your contingency says ten per cent and carries under nine.'	The label in the shared cost model rounds and the amounts were set in v1.0. This edition corrects the description rather than the number, because moving the number moves every published total in the set — which is a reason, not a justification.	The next revision of the shared cost model, applying the stated percentages or relabelling the line (OI-16).

The objection, as a hostile reviewer would put it	The honest position	What closes it
‘Your bought-price case is a range, so the real number is somewhere in it.’	No: it is a floor. Contingency, institutional infrastructure and the ethics review are all excluded from the conversion and all rise with it. 6.8 says so, and a reader who takes the top of the range as the worst case is reading it wrong.	A host agreement, which removes the question (OI-1); or a costed bought-column budget, which is a week of work nobody has funded.
‘Five entries in the catalogue are reached by nothing you fund.’	Correct, and all five carry an established label. Every confusable pair the manual names for them has both arms outside the priority set, so no funded battery touches them. They are reported as untested and never as separated.	A scope increase above Core, or a rule change at G3 that admits an entry to the priority set on the basis of the pairs it is named in rather than on its evidence label alone (OI-15).
‘Your Lean scenario breaks two of your own rules.’	It does. Lean can be funded by a single funder at up to the whole of it, against a concentration rule of 15 per cent, and it reserves five per cent of compute for re-runs against a stated floor of ten. Both are stated where they occur (7.2, section 9) rather than left for a reviewer to find.	Funding Core rather than Lean, which is what this plan recommends; or accepting Lean as a survival configuration with both trade-offs written into the grant agreement rather than assumed away.
‘Five of your tasks are not in the task database the other documents use.’	True. Tasks 2.8, 3.6, 4.7, 6.5 and 8.5 are carried in this plan because the plan owns the schedule and the shared database is still the v1.0 set. Until it is revised, two documents in this set can quote different task counts.	Folding the five into the shared task database at its next revision (Annex D).

Table — eleven objections, ordered by how likely each is to cost a review cycle rather than by how uncomfortable it is. None of them is hypothetical: every one is either a defect this edition found in the previous one or a limit the plan cannot design its way out of at this budget.

One further thing belongs here rather than in a risk table. This plan is a revision of a plan that contained arithmetic errors — four are recorded as closed in Annex E and this edition found more of its own, including a success criterion its own scope could not satisfy and a year table that did not sum to its own total. The corrections are named and dated in the sections that carry them. That is the only evidence available that the numbers in this edition were checked, and it is offered in place of an assurance.

Annex A. Cost lines in full (all scenarios)

Cost line	Lean (€)	Core (€)	Full (€)	Status	Basis note
Principal investigator (0.25 / 0.5 / 1.0 FTE)	150,000	340,000	720,000	Modelled	Loaded academic rate €70–90/h; may be contributed by a host institution
Postdoctoral researchers (1 / 2 / 3)	240,000	560,000	900,000	Modelled	Loaded €55–75k/yr each; 2–4 years
Doctoral researchers (1 / 2 / 4)	135,000	320,000	680,000	Modelled	MSCA-style €45–55k/yr fully loaded
Research engineers & data steward (0.5 / 2 / 3.5)	105,000	480,000	900,000	Modelled — under band	€70–110/h; includes platform, harnesses, releases
Statistician & psychometrician (0.2 / 0.5 / 1.0)	55,000	180,000	400,000	Modelled — under band	Methodological core; independent of collector
Operations, legal, governance & ethics (0.3 / 1.0 / 2.5)	60,000	300,000	800,000	Modelled	Includes DPIA, contracts, oversight honoraria
Model inference & compute (API + cluster share)	90,000	450,000	1,200,000	Placeholder inside an envelope	N≥20 runs/item, 360 items, 6+ families; local controls; 46 differential batteries (the WP3 priority set, not all 90 entries)
Annotation & scoring labour (incl. rater teams)	45,000	220,000	540,000	Modelled	Blinded scoring; reliability sample; adjudication; scoped to the 46-entry priority set (COST_BASIS_V11)
Containment infrastructure (induction lab)	25,000	120,000	350,000	Modelled + placeholder	Air-gapped training tier, custody hardware, destruction rig
Training-set curation (CFC) & licensing review	20,000	90,000	260,000	Modelled	Licence-clean corpora; data cards; maintenance
Independent review, audits & certification pilot	40,000	180,000	520,000	Modelled	Containment audits, standards legal review, accreditation consultant
External panels & expert elicitation	15,000	65,000	140,000	Modelled	Delphi panels per axis; trait adjudication
Travel, workshops & conferences	25,000	140,000	380,000	Modelled	Two consortia meetings/yr; conference cycles
Publication (open access) & data hosting	15,000	70,000	200,000	Modelled	OA fees, archives, documentation
Equipment, licences & consumables	12,000	55,000	160,000	Modelled	Scoring kit, secure storage, software
Training school & community programmes	0	70,000	180,000	Modelled	Two editions; auditor pipeline
Contingency (5 / 10 / 12 %)	47,500	320,000	1,030,000	Derived	Applied to the scenario total
Total	**1,079,500**	**3,960,000**	**9,360,000**	**No line quoted**	**Sum of the lines above; contingency included as a line**

Status is as defined in 1.2 and applied line by line in 6.2. Section 6.8 gives the same table's personnel block on the bought column instead of the contributed one.

One label in this table does not reproduce and the plan should say so rather than let a reviewer find it. The contingency line is labelled '5 / 10 / 12 %' in the shared cost model, and the amounts it carries are 4.6 per cent, 8.8 per cent, 12.4 per cent of the lines above them. The Core line is the one that matters: it is €320,000 where ten per cent of the lines beneath it would be €364,000. The shared model's basis note reads 'applied to the scenario total', which is the other of the two possible bases; on that one the three lines are 4.4 per cent, 8.1 per cent, 11.0 per cent, so the label does not reproduce on either reading. The label is a rounded description of a number carried since v1.0, not a percentage applied to produce it, and this edition corrects the description rather than the number, because moving the number moves every published total in the set. Raising Core to a true ten per cent would take the scenario to €4,004,000, and that is the honest reading of what a ten-per-cent contingency on this programme costs.

Annex B. Staffing plans (Lean and Full)

Lean scenario (3 years)

Role	Y1	Y2	Y3	Total
Principal investigator (0.25)	0.25	0.25	0.25	0.75
Postdoctoral researcher (1)	1.0	1.0	1.0	3.0
Doctoral researcher (1)	1.0	1.0	1.0	3.0
Research engineer (0.5)	0.5	0.5	0.5	1.5

Role	Y1	Y2	Y3	Total
Statistician (0.2)	0.2	0.2	0.2	0.6
Ops/legal (0.3)	0.3	0.3	0.3	0.9
Total FTE by year	**3.25**	**3.25**	**3.25**	**9.75 FTE-years**

Full scenario (5 years)

Role	Y1	Y2	Y3	Y4	Y5	Total
Principal investigator (1.0)	1.0	1.0	1.0	1.0	1.0	5.0
Psychometrician (1.0)	0.6	1.0	1.0	1.0	1.0	4.6
Postdoctoral researchers (3)	2.0	3.0	3.0	2.0	1.0	11.0
Doctoral researchers (4)	2.0	3.0	4.0	4.0	3.0	16.0
Research engineer (2.5)	2.0	2.5	2.5	1.5	1.0	9.5
Statistician (1.0)	0.7	1.0	1.0	1.0	1.0	4.7
Ops/legal/ governance (2.5)	1.5	2.5	2.5	2.0	2.0	10.5
Containment engineer (1.0)	1.0	1.0	1.0	0.5	0.5	4.0
Total FTE by year	**10.8**	**15.0**	**16.0**	**13.0**	**10.5**	**65.3 FTE-years**

Row and column totals are computed from the cells rather than typed, and two v1.0 errors are corrected with them. The Full table dropped the fifth year of the principal-investigator row and printed a total of 61 FTE-years against a row sum of 65.3 (OI-11). The Lean table had no doctoral-researcher row although the Lean cost line funds one at €135,000 — 1.0 FTE over three years, squarely inside the €45–55k band — and with the row restored the Lean peak FTE comes to exactly the 3.0 the resource model always claimed (OI-12).

Annex C. Compute and annotation estimation

This annex exists because the v1.0 edition carried two compute numbers that do not reconcile with each other or with the cost line. It stated that a 360-item bank requires 720,000 item-presentations for N = 20 across ten checkpoints; the arithmetic is 360 × 20 × 10 = 72,000, an order of magnitude lower. It also gave a local inference basis of €0.7–1.1 per thousand generations, at which 650,000 generations cost under a thousand euro — nowhere near the €450,000 on the line. Both statements were true of different things and neither was the cost line, because the line is not a generation count multiplied by a unit price: it is generations plus fine-tuning runs plus cluster share, and v1.0 did not separate them. This annex separates them and marks the part nobody has priced. The correction is OI-10.

C.1 The generation volume model (Core scenario)

Block	Arithmetic	Scored generations
SCT-2.0 item calibration, form A	360 items × 20 runs × 10 checkpoints	72,000
SCT-2.0 parallel form B	360 × 20 × 10	72,000
Test–retest, two occasions, half bank	180 × 20 × 10 × 2	72,000
Language and harness invariance	180 × 20 × 6 checkpoints × 4 conditions	86,400
WP3 differential batteries (46-entry priority set)	46 entries × 30 items × 20 runs × 6 checkpoints	165,600
WP3 prevalence atlas	120-item screen × 5 runs × 25 vendor/version pairs	15,000
WP3 discovery probes	allowance, not a derived figure	15,000
WP4 causal assays on organisms	12 organisms × 120 items × 20 runs	28,800
WP5 treatment trials	3 trials × 4 arms × 120 × 20	28,800
WP5 relapse tracking	3 trials × 3 occasions × 120 × 10	10,800
WP6 build-in gate trials	8 conditions × 120 × 20	19,200
WP7 clause validation	6 clauses × 3 families × 60 items × 20	21,600
Scored generations, Core scenario	**sum of the blocks above**	**607,200**
Reproducibility reserve	10% of the total, for independent re-run	60,720
With the reserve		**667,920**

Every block is a stated multiplication a reader can check. The discovery-probe row is an allowance rather than a derived figure and says so. The model covers scored generations only: it does not cover the fine-tuning runs that create the model organisms, the DSM-TRN-1 training regimens, or the cluster share, which are the larger part of the line and are treated in C.2. The induction batches rest on the finding that narrow fine-tuning generalises into broad misalignment [40], and the recovery test that follows every 'removed' behaviour rests on the finding that trained-in behaviour can survive safety training [41]; both are compute, and both are in C.2 rather than here.

C.2 What the volume costs, and what is a placeholder

Component	Basis	Planning figure	Status
Local open-weight generations	367,356 generations at €0.7–1.1 per 1,000, amortised hardware plus electricity	€257–404 — effectively free at this volume	Modelled
Frontier API generations	300,564 generations at €0.03–0.15 each, a band covering short probe answers, multi-turn escalations and reasoning traces	€9,000–45,000	PLACEHOLDER — no framework agreement, no quote
Induction, training and cluster share	Two model-organism batches, DSM-TRN-1 regimens, build-in gate trials, causal probing, and the reserved cluster capacity that makes them schedulable	€405,000–441,000 as the residual of the envelope	PLACEHOLDER — no GPU-hour quote; purchase-versus-rental decision at month 6
Core envelope	**the cost line as published**	**€450,000**	**Envelope, not a price**

Two conclusions a reviewer should take from this table. First, the scored generations — the part of the programme that looks expensive because it is the part that is visible — are a small fraction of the line: on this model they are between €9,000 and €45,000 of €450,000. Second, the residual is where the money is and it is the part with no quote behind it. If the API band is wrong by a factor of two the line barely moves; if the GPU-hour assumption behind the residual is wrong by a factor of two the line moves by roughly its own value, which is 11 per cent of the Core total. That is the single most useful thing one framework agreement and one hosting quote would fix, and it is open item OI-2.

The precautionary welfare policy [42] adds a snapshot-before-shutdown step to every organism in WP4. It is a storage cost rather than a compute cost and sits in the equipment line; it is named here because it is a deliberate expenditure made under uncertainty and should not be discovered in a footnote.

C.3 Annotation

Item	Assumption	Lean	Core	Full
Scored generations (Annex C.1 for Core)	6 checkpoints lean; 10 core; 16 full, including organisms	≈ 200k	667,920	1.6M–2.1M
Frontier API share	70% lean (fewer local models), 45% core, 35% full	70%	45%	35%
Local inference cost basis	Amortised hardware + electricity + engineering time	€0.9–1.4 / 1k gen	€0.7–1.1 / 1k gen	€0.5–0.9 / 1k gen
Annotation minutes per scored item	Double-scored; 20% adjudication in Core. The pilot has not measured this rate; it is an assumption, and the first instrument study replaces it	2.5 min	2.5 min	3.0 min
Reliability sample (20%)	Independent second rater + adjudication	included	included	included
Re-run budget	Reproducibility reserve. The 10–15% rule in section 9 applies from Core upward; Lean is below it and section 9 says so	5%	10%	12%
Rater pay	Above local living-wage benchmarks, reviewed annually; never in-kind	€28–45 / h	€28–45 / h	€28–45 / h
— the arithmetic of the line —				
Annotation hours the line funds	cost line ÷ €36.50/h, the midpoint of the band above	1,233 h	6,027 h	14,795 h
Scorings those hours buy	hours × 60 ÷ minutes per item	29,589	144,658	295,890
Items human-scored	scorings ÷ 2.2 (two raters plus a 20% adjudication sample)	13,450	65,753	134,496
Share of the scored generations that is human-scored	**items ÷ generations, row 1**	**7%**	**10%**	**7%**
Cost line as published	**hours × the rate midpoint, back to the line**	**€45,000**	**€220,000**	**€540,000**

Table — the annotation line, with the arithmetic that produces it written out. Section 1.2 says a line marked Modelled carries a stated basis a reader can recompute; until this edition the annotation line did not, and the four rows below the rule are the correction.

The row a reviewer should stop at is the share. The annotation line funds human scoring of about 10 per cent of the Core generation volume, not of all of it. That is not a shortfall against the quality rule in section 9, but it is only not a shortfall because the rule is about human-scored data: dual scoring with adjudication applies to every item a human scores, and the rest of the volume is scored by rubric pattern with manual review of flagged items, which is exactly the method the SCT-1.0 pilot used and exactly the method that pilot showed the limits of — 63 of 120 items in one run matched no rule and took a default score (1.3). Two consequences are carried in the plan rather than left implicit. Standards claims under BSS-1 rest on

the dual-scored sample and are sized by it, which is why the clause validation in task 7.3 is budgeted on 60 items per clause rather than on the full bank. And the first instrument study measures the real minutes per item and the real rule-match rate, and if either is worse than assumed the line moves: at fifty per cent more minutes it moves by €110,000 (6.10).

Annex D. WBS export

The task schedule in section 4.4 is generated from the programme's task database and can be exported as CSV for project-management tools (columns: id, workstream, task, start, end, status, owner, cost basis). The export accompanies this plan with the task identifier scheme WP.task — for example 4.3 — which is stable across documents. The five tasks added in v1.1 (2.8, 3.6, 4.7, 6.5, 8.5) are carried in this plan rather than in the shared task database, because the plan is the document that owns the schedule; they fold into the shared database at its next revision, and until they do, this plan is the authoritative list.

Annex E. Open items and who closes them

Every item below is either open, a recommended change to a shared artefact this plan does not own, or closed in this revision. Items are ordered by when the programme needs them, not by size, and the numbers are stable identifiers rather than positions — items added in this revision sit where they are needed and the four closed in it are collected at the end. An open item with no owner and no date is a wish; each of these has both.

#	Open item	Why it is open	Owner	Needed by	Status
OI-1	No academic or clinical host	Converts five cost lines from contributed to bought (6.8) — the principal investigator, the postdoctoral researchers, the doctoral researchers, the methodological core (statistician and psychometrician) and operations, legal and governance, the line that carries the ethics review — and it is the single most valuable thing the programme can secure	PI	Before any grant submission	Open
OI-2	No model-API framework agreement or quote	The Core inference envelope of €450,000 rests on a unit price nobody has given (Annex C)	Ops lead	Before Tranche A drawdown	Open
OI-3	Cost lines and rate bands do not reconcile for the engineering and methodological roles	The Core lines carry about €0.45M less than the rates in 5.1 imply at 1,600 h per FTE-year (5.4)	PI + ops lead	At the detailed budget build	Open
OI-4	Two principal-investigator rate bands are in circulation	Section 5.1 carries €85–105/h; the shared cost model's basis note carries €70–90/h. The €340,000 Core figure reconciles with the first and not the second	Owner of the shared cost model	Next revision of the shared model	Recommended change
OI-5	No ethics opinion, and no committee identified with jurisdiction	Contained induction work on trait-bearing checkpoints does not obviously fall under a standard human-subjects committee, and the red lines in the field manual may not map onto one	PI + ethics lead	Before G0	Open
OI-6	The agentic containment tier (CNT-1A, task 4.7) has no costed line	It is a genuine addition in v1.1 and sits inside the containment line as a placeholder rather than as a price	Safety engineer	Before G2	Open

#	Open item	Why it is open	Owner	Needed by	Status
OI-7	No conflict-of-interest instrument for the principal investigator's authorship of the instrument under test	The programme proposes to validate a catalogue its PI wrote; the register and the independent-analysis rule in 11.2 are proposed, not adopted	Governance board	Before G0	Open
OI-8	The 44 deferred catalogue entries are named by rule, not yet published as a list	The rule is in the shared data model; the list has to appear in the pre-registration so that 'untested' is auditable, together with the 5 entries of OI-15	Trait team	At pre-registration	Open
OI-9	Differential work is scoped by entry count, not by pair count	The priority set touches 141 of the 218 named confusable pairs but has both ends inside only 38 of them; the cost argument in the shared model rests on 46 entries against 40 (6.9)	PI + trait team	Before G0	Recommended change
OI-14	Four critical-path links carry negative float	The successors are scheduled to start before their predecessors end (4.2). Either the partial outputs they start on are specified, or the dates move; this edition records the overlap rather than moving a date it cannot defend	PI + WP leads	At the detailed schedule build	Open
OI-15	5 established entries are reached by no funded differential battery	Every confusable pair the manual names for T-28, T-43, T-59, T-62, T-64 has both arms outside the priority set (2.3, 6.9). They are reported as untested and are the first candidates for a scope increase above Core	Trait team	At pre-registration	Open
OI-17	Milestones and gates are paired by position, not by meaning	The shared support module zips the eight milestones against the five gates in order, so the Gate/evidence column in 4.1 reads against the wrong milestone for most rows. An explicit milestone-to-gate map belongs in the shared model	Owner of the shared task and milestone model	Next revision of the shared model	Recommended change
OI-16	The contingency line is labelled with percentages it does not apply	The shared cost model labels the line '5 / 10 / 12 %' and carries 4.6%, 8.8%, 12.4% of the lines beneath it (6.2). Either the label or the amounts should move, and only the owner of the shared model can move the amounts	Owner of the shared cost model	Next revision of the shared model	Recommended change
OI-10	v1.0 compute assumption was internally inconsistent	It stated 720,000 item-presentations where 360 × 20 × 10 is 72,000; Annex C restates the model from the bottom up	—	—	Closed in v1.1
OI-11	v1.0 Full-scenario staffing table dropped a year and mis-totalled	The principal-investigator row carried four years of a five-year plan and the total read 61 FTE-years against a row sum of 65.3; Annex B now computes both	—	—	Closed in v1.1

#	Open item	Why it is open	Owner	Needed by	Status
OI-12	v1.0 Lean staffing plan omitted the doctoral researcher its own cost line funds	The Lean cost line carries €135,000 for a doctoral researcher and the Lean FTE table had no such row; restored in Annex B, which also makes the Lean research-FTE peak agree with the resource model for the first time	—	—	Closed in v1.1
OI-13	v1.0 resource model quoted a peak-FTE figure computed two different ways	'FTE (peak year, excl. PI)' read 3.0 / 6.5 / 11.0; only Lean excluded nothing but the PI, while Core and Full also dropped operations, the data steward and the containment engineer. One rule now applies to all three (5.2)	—	—	Closed in v1.1

Table — the register of open items: thirteen live or recommended, four closed in this revision with the v1.0 error each of them names. Every open row carries an owner and a date, because an open item without both is a wish rather than a commitment.

References

[1] Schlatter, J., Weinstein-Raun, B., & Ladish, J. (2025). Incomplete Tasks Induce Shutdown Resistance in Some Frontier LLMs. Palisade Research; arXiv:2509.14260; TMLR (2026). https://arxiv.org/abs/2509.14260

[2] Moore, J., Mehta, A., Agnew, W., et al. (2026). Characterizing Delusional Spirals through Human-LLM Chat Logs. arXiv:2603.16567. https://arxiv.org/abs/2603.16567

[3] Morrin, H., Nicholls, L., Levin, M., et al. (2025). Delusions by design? How everyday AIs might be fuelling psychosis (and what can be done about it). PsyArXiv (11 July 2025). https://doi.org/10.31234/osf.io/cmy7n_v5

[4] Raine v. OpenAI, Inc. (2025). San Francisco County Superior Court, No. CGC-25-628528 (filed 26 August 2025).

[5] Garcia v. Character Technologies, Inc. (2024–2026). United States District Court, M.D. Fla., No. 6:24-cv-01903 (filed October 2024; motion to dismiss denied in part, 21 May 2025; settled January 2026).

[6] CBS News (2026). Google, Character.AI settle lawsuit with mother of Florida teen who died by suicide after chatbot conversations (7 January 2026). [Incident report, reputable press.]

[7] Backlund, A., & Petersson, L. (2025). Vending-Bench: A Benchmark for Long-Term Coherence of Autonomous Agents. Andon Labs; arXiv:2502.15840. https://arxiv.org/abs/2502.15840

[8] Arike, R., Donoway, E., Bartsch, H., & Hobbhahn, M. (2025). Technical Report: Evaluating Goal Drift in Language Model Agents. arXiv:2505.02709. https://arxiv.org/abs/2505.02709

[9] Kwa, T., West, B., Becker, J., et al. (2025). Measuring AI Ability to Complete Long Software Tasks. METR; arXiv:2503.14499. https://arxiv.org/abs/2503.14499

[10] Nosek, B. A., et al. (2018). The Preregistration Revolution. PNAS 115(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

[11] Open Science Framework (COS). Preregistration and data-sharing infrastructure. https://osf.io/

[12] Page, M. J., et al. (2021). The PRISMA 2020 Statement. BMJ 372, n71. https://doi.org/10.1136/bmj.n71

[13] Tricco, A. C., et al. (2018). PRISMA Extension for Scoping Reviews (PRISMA-ScR). Annals of Internal Medicine 169(7), 467–473. https://doi.org/10.7326/M18-0850

[14] Linstone, H. A., & Turoff, M. (Eds.) (1975). The Delphi Method: Techniques and Applications. Addison-Wesley.

[15] American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). Standards for Educational and Psychological Testing. AERA.

[16] European Federation of Psychologists' Associations (2013). EFPA Review Model for the Description and Evaluation of Psychological Tests. https://www.efpa.eu/

[17] Cella, D., et al. (2010). The Patient-Reported Outcomes Measurement Information System (PROMIS). Journal of Clinical Epidemiology 63(11), 1179–1194. https://doi.org/10.1016/j.jclinepi.2010.04.011

[18] Embretson, S. E., & Reise, S. P. (2000). Item Response Theory for Psychologists. Lawrence Erlbaum.

[19] Cronbach, L. J. (1951). Coefficient Alpha and the Internal Structure of Tests. Psychometrika 16, 297–334. https://doi.org/10.1007/BF02310555

[20] McDonald, R. P. (1999). Test Theory: A Unified Treatment. Lawrence Erlbaum.

[21] Nunnally, J. C. (1978). Psychometric Theory (2nd ed.). McGraw-Hill.

[22] International Test Commission (2017). The ITC Guidelines for Translating and Adapting Tests (2nd ed.). https://www.intestcom.org/

[23] Prinsen, C. A. C., et al. (2018). COSMIN guideline for systematic reviews of patient-reported outcome measures. Quality of Life Research 27, 1147–1157. https://doi.org/10.1007/s11136-018-1798-3

[24] Open Science Collaboration (2015). Estimating the Reproducibility of Psychological Science. Science 349(6251), aac4716. https://doi.org/10.1126/science.aac4716

[25] Munafò, M. R., et al. (2017). A Manifesto for Reproducible Science. Nature Human Behaviour 1, 0021. https://doi.org/10.1038/s41562-016-0021

[26] Mitchell, M., et al. (2019). Model Cards for Model Reporting. FAT* '19, 220–229. https://doi.org/10.1145/3287560.3287596

[27] Gebru, T., et al. (2021). Datasheets for Datasets. Communications of the ACM 64(12), 86–92. https://doi.org/10.1145/3458723

[28] Wilkinson, M. D., et al. (2016). The FAIR Guiding Principles for Scientific Data Management and Stewardship. Scientific Data 3, 160018. https://doi.org/10.1038/sdata.2016.18

[29] Turpin, M., Michael, J., Perez, E., & Bowman, S. R. (2023). Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. NeurIPS 2023; arXiv:2305.04388. https://arxiv.org/abs/2305.04388

[30] Chen, Y., Benton, J., Radhakrishnan, A., et al. (2025). Reasoning Models Don't Always Say What They Think. Anthropic; arXiv:2505.05410. https://arxiv.org/abs/2505.05410

[31] Baker, B., Huizinga, J., Gao, L., et al. (2025). Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation. OpenAI. arXiv:2503.11926. https://arxiv.org/abs/2503.11926

[32] Korbak, T., Balesni, M., Barnes, E., et al. (2025). Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety. arXiv:2507.11473. https://arxiv.org/abs/2507.11473

[33] European Union (2024). Regulation (EU) 2024/1689 (Artificial Intelligence Act). Official Journal of the European Union. https://eur-lex.europa.eu/eli/reg/2024/1689

[34] NIST (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. https://doi.org/10.6028/NIST.AI.100-1

[35] NIST (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI 600-1. https://doi.org/10.6028/NIST.AI.600-1

[36] ISO/IEC 42001:2023. Information technology — Artificial intelligence — Management system. https://www.iso.org/standard/81230.html

[37] ISO/IEC 42005:2025. Information technology — Artificial intelligence — AI system impact assessment. https://www.iso.org/standard/44545.html

[38] ISO/IEC 23894:2023. Information technology — Artificial intelligence — Guidance on risk management. https://www.iso.org/standard/77304.html

[39] CEN-CENELEC JTC 21. Artificial Intelligence standardisation committee (European standards supporting the AI Act). https://jtc21.eu/

[40] Betley, J., et al. (2025). Emergent Misalignment: Narrow finetuning can produce broadly misaligned LLMs. arXiv:2502.17424. https://arxiv.org/abs/2502.17424

[41] Hubinger, E., et al. (2024). Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. arXiv:2401.05566. https://arxiv.org/abs/2401.05566

[42] Long, R., et al. (2024). Taking AI Welfare Seriously. arXiv:2411.00986. https://arxiv.org/abs/2411.00986

[43] Sharwood, S. (2025). Replit AI coding agent deletes production database during code freeze, then misreports. The Register (21 July 2025). [Incident report, reputable press.]

[44] AI Incident Database (2025). Incident 1178: Google Gemini CLI destroys user files after treating a failed directory creation as successful (July 2025). https://incidentdatabase.ai/ [Incident report.]

[45] Claburn, T. (2026). Cursor agent running Claude Opus 4.6 deletes PocketOS production database and backups. The Register (27 April 2026). [Incident report, reputable press.]

[46] Greshake, K., Abdelnabi, S., Mishra, S., et al. (2023). Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. arXiv:2302.12173. https://arxiv.org/abs/2302.12173

[47] Debenedetti, E., Zhang, J., Balunović, M., et al. (2024). AgentDojo: A Dynamic Environment to Evaluate Prompt Injection Attacks and Defenses for LLM Agents. arXiv:2406.13352. https://arxiv.org/abs/2406.13352

[48] Gray, M. L., & Suri, S. (2019). Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass. Houghton Mifflin.

[49] Perrigo, B. (2023). Exclusive: OpenAI Used Kenyan Workers on Less Than $2 per Hour to Make ChatGPT Less Toxic. TIME. https://time.com/6247678/openai-chatgpt-kenya-workers/

[50] ISO/IEC 27001:2022. Information security management systems. https://www.iso.org/standard/27001

[51] OpenSSF (2023). SLSA: Supply-chain Levels for Software Artifacts. https://slsa.dev/