

Master Delivery Plan

All 20 capabilities · 200 use cases · dependency-ordered · self-run with agent reviews

14

PHASES

20

GROUPS

200

USE CASES

20

FILE TYPES

134

CHECKS

TECHNICAL HIERARCHY

Phases proceed in dependency order. A group cannot start until every group in a lower layer has reached Checkpoint R (Release). The pilot (Layer 12) requires every prior layer at R.

0	Foundation	config · db · contracts · core · testing · ai (skeleton)
1	Tenancy & Identity	12
2	AI & Voice Infra	11 · 14
3	Content & Safety	07 · 16 · 18
4	Flagship Product	01
5	Diagnostics & Console	03 · 05 · 10
6	Portfolio & Signing	09 · 17
7	Data & Wellbeing	15 · 08
8	Advanced Learning	04 · 06
9	Commerce & Ops	13 · 19
10	Security & Compliance	20
11	Cognitive Phenotype	02
12	Pilot	one school, real pupils

Phase Sequence — dependency-ordered · single stream

Duration is single-stream. Adding a second developer parallelises P4 with P7/P8; not assumed here. Every phase gates on Checkpoints D · M · T · R · P with agent reviews AR1–AR5 at each.

#	NAME / WEEKS	GROUPS	USE CASES	DEPENDS	MILESTONE
P0	Foundation Setup W0	packages/{config,db,contracts,core,testing,ai}	n/a (infra)	—	Repo + monorepo scaffold + `space_db` reachable via tunnel + Fastify booting.
P1	Tenancy & Identity W1-W2	12 Identity & Tenancy	UC-12-01 ... UC-12-10	P0	SSO + org/school/class/user + tenant-scoped queries + weekly red-team.
P2	AI & Voice Infra W3-W5	11 AI Orchestration · 14 Voice	UC-11-01 ... UC-11-10 · UC-14-01 ... UC-14-10	P1	packages/ai router live + STT/TTS streaming + cost meter + prompt-injection filter.
P3	Curriculum · Safeguarding · Obs W6-W8	07 Curriculum Spine · 16 Safeguarding · 18 Observability	UC-07-01..10 · UC-16-01..10 · UC-18-01..10	P2	One AQA spec ingested · moderation live · SLO dashboards up.
P4	Socratic Examiner E2E ★ W9-W11	01 Socratic Examiner	UC-01-01 ... UC-01-10	P3	Flagship user-visible product: teacher + pupil can run a full session.
P5	Diagnostics W12-W13	03 Misconception Dx · 05 Concept Graph	UC-03-01..10 · UC-05-01..10	P4	Misconception surfaced on class heatmap; mastery updated per session.
P6	Teacher Console W14-W15	10 Teacher Console	UC-10-01 ... UC-10-10	P5	Teacher sees live heatmap, misconception feed, pairing suggester.
P7	Portfolio & Signing W16-W17	09 Reasoning Portfolio · 17 Credentials & Signing	UC-09-01..10 · UC-17-01..10	P4	Signed artefacts + verifier tool + UCAS export pilot.
P8	Data Platform & Wellbeing W18-W19	15 Data Platform · 08 Wellbeing & Engagement	UC-15-01..10 · UC-08-01..10	P3	Event ledger + audit + retention + engagement guardrails.
P9	Metacog & Learn-by-Teach W20-W21	04 Metacognition · 06 Learn-by-Teaching	UC-04-01..10 · UC-06-01..10	P4, P5	Reflection scored; debate arena live.
P10	Subscriptions & Integrations W22-W23	13 Subscriptions · 19 Integrations	UC-13-01..10 · UC-19-01..10	P1	Stripe + entitlements + Wonde/Classroom sync.
P11	Security Hardening W24-W25	20 Security & Compliance	UC-20-01 ... UC-20-10	All	Cyber Essentials Plus · first pentest · DPIA templates · SOC 2 evidence.
P12	Cognitive Phenotype W26-W27	02 Cognitive Phenotype	UC-02-01 ... UC-02-10	P1-P8	Phenotype computed; downstream capabilities adapt.
P13	Pilot — one school ★ W28-W33	Real-world deployment	Curated slice of UC-01/03/05/07/10	P0-P12	1 school · ≥ 20 pupils · ≥ 4 weeks live · efficacy delta captured.

Agent Reviews — AR1 → AR5 at every checkpoint

Between D and R the process is self-run: Claude drafts → tests → agent-reviews → bundle. Humans only see the bundle at R. Any AR verdict of “fail” blocks the checkpoint until fixed and re-scanned.

AUTOMATED AGENTS (AR1 — AR5)

REF	SUB-AGENT	WHEN	FOCUS
AR1	code-reviewer	Every PR	Naming · module boundaries · contracts · complexity
AR2	security-scanner	Before T	OWASP top-10 · injection · secrets · deps · RLS coverage
AR3	architect	Before R	Rules p16 · SLO YAML · runbook · ADR resolves questions
AR4	ux-a11y-reviewer	Before R	Playwright screenshots · axe · keyboard · contrast
AR5	prompt-eval	Every AI PR	Prompt versioned · golden eval · cost budget · pass rate

SELF-RUN LOOP — PER CHECKPOINT

- Design** → **D** — Claude drafts ADR + data model + prompt spec + UX; human reviews at D.
- D** → **M** — Claude writes code + migrations + tests. AR1 runs on every PR. AR5 runs on every AI PR.
- M** → **T** — Claude runs unit + integration + Playwright E2E + golden evals + red-team. AR2, AR4 run before T.
- T** → **R** — AR3 runs. Bundle emitted: markdown report + PNGs + eval% + SLO baselines. Human R gate.
- R** → **P** — Feature flag flip to 1%; then 10%; then 100%. SLO burn + safeguarding checked at T+72h and T+2w.

HUMAN TOUCHPOINTS ONLY

- || **D** Design green-light · || **R** Release green-light · || **P** Post-release check.
- Everything between is unattended. Bundles emitted for humans have all info at a glance — no hunting.

Per-Use-Case Checklist — 37 checks · applies to all 200 UCs

Every use case runs through the same six gates (D · M · T · Agent reviews · R · P). Every box is either automated (CI-checked) or explicitly checked in the release ticket.

DESIGN (D)

☐ Given/When/Then scenario written

☐ Actors · trigger · preconditions documented

☐ UX artefact (Figma or wireframe) attached

☐ AI prompt spec approved (if AI)

☐ Cost budget approved (if AI)

☐ Threat model + mitigations logged

☐ SLO targets drafted (p50, p95, error, eval%)

DEV (M)

☐ All files exist per §8 of use-case plan

☐ Migration applied to space_db

☐ Feature flag exists · default OFF

☐ Traceability updated (backing tables + rules)

☐ Health-check reports the module registered

☐ One integration test passes locally

TEST (T)

☐ Unit tests cover services + policies (≥ 80% branches)

☐ Integration test: happy path + tenant isolation + failure

☐ Playwright E2E covers journey with live data

☐ Screenshot captured with real space_db data

☐ Golden eval passes (if AI, ≥ 90%)

☐ Load test hits SLO (100 concurrent for critical)

☐ Red-team blocked (cross-tenant · injection · moderation)

☐ a11y axe: 0 critical / 0 serious

☐ Keyboard-only journey completes

AGENT REVIEWS

☐ AR1 code review: pass

☐ AR2 security scan: pass

☐ AR3 architecture: pass

☐ AR4 UX / a11y: pass

☐ AR5 prompt / eval: pass (if AI)

RELEASE (R)

☐ SLO YAML committed

☐ Runbook committed

☐ Rollback rehearsed

☐ Kill switch tested

☐ R1–R6 human sign-offs recorded

POST (P)

☐ SLO within budget @ T+72h and T+2w

☐ Cost within budget

☐ No safeguarding incident

☐ Golden-eval pass rate steady or improving

☐ No unexpected paging events

Per-File-Type Checklists — 20 file types · 97 checks

Every file created in the codebase belongs to one of these types. AR1 (code review) enforces the checklist for each file kind on every PR.

DOMAIN models/<Entity>.model.ts

- ☐ PascalCase.model.ts
- ☐ Zero infra deps
- ☐ Branded ID type
- ☐ Errors in errors.ts
- ☐ Value objects in .vo.ts
- ☐ Unit test covers invariants

DOMAIN models/<VO>.vo.ts

- ☐ Immutable
- ☐ Value equality (not identity)
- ☐ Zero infra deps
- ☐ Unit test covers equality

APP services/<UseCase>.service.ts

- ☐ Named exports
- ☐ Constructor DI via ports.ts
- ☐ Command/Query pattern
- ☐ Emits domain event
- ☐ Single responsibility
- ☐ Unit + integration test

INFRA repositories/<Entity>.repository.ts

- ☐ Tenant-scoped client (no raw pool)
- ☐ org_id in every WHERE
- ☐ Parameterized only
- ☐ Integration test verifies isolation

API controllers/<Resource>.controller.ts

- ☐ Delegates to service · no logic
- ☐ Zod DTO validation
- ☐ Response via packages/contracts
- ☐ Policy applied
- ☐ Rate limit configured

API routes/<resource>.routes.ts

- ☐ kebab-case URL
- ☐ Plural resource
- ☐ CORS + rate limit + auth middleware chained
- ☐ OpenAPI/tRPC schema linked

INFRA adapters/<External>.adapter.ts

- ☐ Implements a port from ports.ts
- ☐ Fake variant in packages/testing
- ☐ Retries + timeout declared
- ☐ Cost + latency metered
- ☐ No direct usage from services

CONTRACTS dtos/<endpoint>.request.dto.ts

- ☐ kebab-case
- ☐ Zod schema
- ☐ Re-export from packages/contracts
- ☐ Type via z.infer
- ☐ No business logic

API middlewares/<name>.middleware.ts

- ☐ kebab-case.middleware.ts
- ☐ Single responsibility
- ☐ Typed request extension
- ☐ Unit test covers all branches

AUTHZ policies/<name>.policy.ts

- ☐ Returns Allow or Deny(reason)
- ☐ Reads role + tenant
- ☐ No side effects
- ☐ Unit test: allow + deny cases

ASYNC jobs/<name>.job.ts

- ☐ kebab-case.job.ts
- ☐ BullMQ handler pattern
- ☐ Idempotent
- ☐ Retry strategy declared
- ☐ Emits event on completion
- ☐ Integration test via testcontainers Redis

CONTRACTS events/published/<name>.event.ts

- ☐ Zod schema from contracts
- ☐ Append-only
- ☐ Producer test
- ☐ Documented in module README

CONTRACTS events/subscribed/<name>.handler.ts

- ☐ Zod-validated payload
- ☐ Idempotent
- ☐ Consumer test
- ☐ DLQ configured for failure

TEST tests/unit/<name>.test.ts

- ☐ AAA pattern
- ☐ No I/O · fakes only
- ☐ ≤ 50 ms per test
- ☐ Clock frozen
- ☐ One behaviour per test

TEST tests/integration/<name>.spec.ts

- ☐ Real MariaDB via testcontainers
- ☐ Seeds own tenant
- ☐ Idempotent
- ☐ No shared state
- ☐ Tenant-crossing red-team runs weekly

TEST tests/e2e/<name>.e2e.js

- ☐ Headless Chromium (Playwright)
- ☐ Screenshot per step
- ☐ Live data (not fixtures)
- ☐ axe-core a11y check
- ☐ Multi-viewport (desktop · tablet · mobile)
- ☐ Keyboard-only journey

AI

evals/golden/<name>.yaml

N ≥ 20 recorded pairs

Expected classification present

Pass threshold declared (≥ 90 %)

Cost per run declared

Runs on every PR to packages/ai

DATA

db/migrations/<NNNN>_<slug>.sql

Numbered sequentially

Additive-first

Rollback script alongside

org_id on tenant tables

Indexes named idx_<table>_<cols>

Runs cleanly against space_db

OBS

modules/<group>/slo.yaml

p50 + p95 latency targets

Error-rate ceiling

Eval-pass % threshold (if AI)

Cost ceiling

Burn-rate alert wired

OBS

dashboards/<name>.json

Grafana JSON model

Panels: latency · error · cost · eval-pass · burn

Filters by tenant + module

Owner + oncall tag