

AI-BOK TOOLKIT v1.2

Maturity Calibration Kit

Companion to the AI Body of Knowledge v1.2

Jan Willem van Veen · ArchiXL · ai-bok.nl

Maturity Calibration Kit: numerical thresholds + worked assessments

Adds numerical thresholds and three worked assessments of fictitious organisations to anchor maturity scoring and reduce assessor variance.

1. What this addendum adds

The maturity evidence rubric defines, per KA × Level, the **kinds** of evidence required (Documents, Metrics, Logs). This addendum adds two missing pieces:

- Numerical thresholds** for the metric-bearing requirements - where the literature, regulation, or operational experience supports a defensible number, the threshold is named. Each threshold is tagged **indicative** (a starting point, organisation may tune) or **mandatory** (does not negotiate; failure means the level is not achieved).
- A calibration kit** with **three worked assessments** of fictitious organisations at three different maturity profiles, so that assessors can anchor their own judgements against shared exemplars.

The rubric remains the source of truth for *what kinds* of evidence are required; this addendum specifies *how much*.

2. Numerical thresholds - what counts as L2, L3, L4

KA1 - AI Governance

Requirement	L2	L3	L4	Tag
AI Governance Charter approval	Drafted	Approved by AI Board	Approved by board <i>and</i> reviewed annually	mandatory
AI inventory coverage (% of production AI systems registered)	≥ 60%	≥ 95%	100% with monthly attestation	mandatory
Annual self-assessment cadence	absent	annual	semi-annual	indicative

KA2 - AI Strategy & Portfolio

Requirement	L2	L3	L4	Tag
Portfolio review cadence	informal	quarterly	quarterly + portfolio-trigger-based	mandatory
Value realisation measurement	none	per-system ex-post	portfolio-level value dashboard	indicative
Decommissioning protocol applied (% of retirements following protocol)	n/a	≥ 80%	100%	indicative

KA3 - AI Architecture

Requirement	L2	L3	L4	Tag
Reference architecture version	v1	v1+ with quarterly review	continuously evolved	mandatory
Architecture conformity assessments (% of new systems with explicit conformity record)	n/a	≥ 80%	100% with automated checks	indicative
Cognition-plane elements (mandate, authority register, policy-as-code) operational	conceptual	per agentic system	portfolio-wide	mandatory (when agentic systems present)

KA4 - AI Lifecycle Management

Requirement	L2	L3	L4	Tag
Lifecycle gates (Gates 2, 5, 7 minimum)	documented	applied to ≥ 90% of changes	applied to 100% with automated gate logs	mandatory
Mean time gate-decision → next-phase-start	informal	≤ 10 working days	≤ 5 working days	indicative
Retraining policy per model	absent	per-model documented	automated trigger pipelines	indicative

KA5 - Data & Semantics for AI

Requirement	L2	L3	L4	Tag
Data-source allowlist coverage	priority sources	all production sources	all sources incl. dev/test	mandatory
Data-quality KPIs per source (completeness, freshness, accuracy)	priority sources	all production sources	continuous monitoring	indicative
Conceptual frameworks for priority domains	one domain	core domain complete	enterprise-wide	indicative

KA6 - Model Management

Requirement	L2	L3	L4	Tag
Model registry coverage	production models	all production + pilot	all + dev models	mandatory
Model cards (% of in-scope systems with current card)	≥ 70%	≥ 90%	100%	mandatory
Mean model-card staleness	n/a	≤ 60 days	≤ 30 days	indicative
Automated bias detection in CI/CD	absent	for new models	for all model changes	indicative

KA7 - AI Interaction & UX

Requirement	L2	L3	L4	Tag
WCAG conformity for AI-facing UIs	AA for new	AA for all production	AA + AAA aspirational	mandatory (where regulated)
Prompt-pattern library	basic	versioned + tested	A/B-tested per pattern	indicative
Human-in-the-loop protocols per risk level	documented	applied + measured	adaptive	mandatory

KA8 - AI Operations

Requirement	L2	L3	L4	Tag	
Monitoring coverage (high-volume systems with latency + error + drift metrics)		≥ 50%	≥ 90%	100% + traces	mandatory

Requirement	L2	L3	L4	Tag
Mean time to detect (MTTD) production incident	hours	≤ 30 min	≤ 5 min	indicative
Mean time to recover (MTTR)	days	≤ 4 h	≤ 1 h	indicative
Deployment automation (% of deploys via CI/CD)	manual	≥ 80%	100% + canary	indicative

KA9 - AI Risk Management & Safety

Requirement	L2	L3	L4	Tag
EU AI Act risk classification	for high-risk	for all systems	for all + change-triggered	mandatory
Risk register update cadence	annual	quarterly	continuous	mandatory
Red-team cadence (high-risk systems)	n/a	annual	semi-annual	indicative
FM-coverage assessment (% of FM01-22 with named control)	≥ 50%	≥ 80%	100%	mandatory (for agentic estates)

KA10 - AI Compliance & Audit

Requirement	L2	L3	L4	Tag
Algorithmeregister coverage (NL public sector)	high-risk	all in-scope systems	all + retired-system archive	mandatory (NL public)
DPIA coverage (% of high-risk systems with current DPIA)	≥ 80%	100%	100% + change-triggered	mandatory
Audit-trail completeness (% of decisions with verifiable trail)	core flows	≥ 95%	100%	mandatory
Conformity assessment turnaround	n/a	≤ 90 days	≤ 30 days	indicative

KA11 - AI Ethics & Responsible AI

Requirement	L2	L3	L4	Tag
Bias-audit coverage (% of decision-affecting systems audited)	≥ 50%	≥ 90%	100% + change-triggered	indicative

Requirement	L2	L3	L4	Tag
Ethics committee cadence	annual	quarterly	as-needed + quarterly	mandatory
Stakeholder participation cycles per year	informal	annual	continuous	indicative

KA12 - Knowledge & Context Management

Requirement	L2	L3	L4	Tag
Trusted-source register coverage	priority sources	all RAG sources	all knowledge interactions	mandatory
Grounding rate (RAG systems: % responses with cited source)	≥ 70%	≥ 90%	≥ 95%	indicative
Source freshness audit cadence	annual	quarterly	continuous	indicative

KA13 - AI Literacy & Capability Development

Requirement	L2	L3	L4	Tag
Competency profiles (% of role x system intersections defined)	priority roles	≥ 90%	100%	mandatory
Literacy register currency (% of individuals current on required modules)	≥ 70%	≥ 90%	≥ 95%	mandatory (Art. 4)
Refresh adherence (% of modules within refresh window)	≥ 70%	≥ 90%	≥ 95%	mandatory
Drill cadence (high-risk-system supervisors)	n/a	annual	semi-annual	indicative
Article 4 evidence file readiness	started	maintained + auditable	continuously verified	mandatory

3. Threshold tuning

Indicative thresholds may be tuned by adopters with documentation. Mandatory thresholds may not be relaxed without invalidating the level claim. A documented exception path (cf. rubric aggregation rule) exists, with the following structure:

```

exception_id: EX-2026-Q2-01
ka: KA10
requirement: Audit-trail completeness ≥ 95% at L3
realised: 88%
reason: legacy classifier without backfill; planned retirement Q3 2026
mitigation: out-of-band trace via case-management system; auditor accepts compensating control
approved_by: AI Governance Lead + AI Auditor
approved_on: 2026-04-15
expires: 2026-09-30

```

Exceptions are reviewed at the cadence of the rubric's inter-assessor protocol. An exception lasting beyond two review cycles loses status (no longer an exception; becomes either a closed item or a downgrade).

4. Calibration kit - three worked assessments

Three fictitious organisations, three different maturity profiles. An assessor reading these can anchor their own scoring against the worked exemplars.

4.1 Assessment A - "Gemeente Veenwoorden" (≈120k inhabitants, public sector)

Profile: just past MVG, working toward L3 on the seven MVG items.

KA	Score	Evidence summary
KA1	L3	Charter approved March 2026; annual review scheduled; inventory at 96%
KA2	L2	Portfolio register exists; quarterly reviews informal; no value-realisation dashboard yet
KA3	L2	Reference architecture v1.1; cognition plane positioned for agentic pilot; conformity checks manual
KA4	L2	Gates documented; applied to 85% of changes (just below L3 90% threshold)
KA5	L2	Data-source policy; one SKOS framework; data-quality KPIs for top sources
KA6	L2	Model registry; model cards on 78% of production (below L3 90%)
KA7	L3	WCAG AA across production; prompt-pattern library v0.3 versioned
KA8	L2	Monitoring on 4/7 high-volume systems; MTTD ~1h; manual deploys
KA9	L3	Risk classification all systems; quarterly register; one red-team done; FM-coverage 75% (just below L3 80%) → flagged , accept as L2
KA10	L3	Algorithmeregister 100%; DPIAs current; audit-trail >95% on flagship system

KA	Score	Evidence summary
KA11	L2	Framework; bias audits on 3/3 classifiers; ethics committee quarterly
KA12	L2	Trusted-source register; RAG grounding rate 82% (below L3 90%)
KA13	L2	Profiles for 5 role groups; register currency 78%; refresh adherence 81%

Aggregate maturity: Minimum across KAs = **L2**, with KA1, KA7, KA10 at L3. The dominant constraint is the KA9 FM-coverage gap (75% vs. 80%); closing it lifts an L3-eligible KA without affecting overall score. **Aggregate stays L2** until the lowest KAs lift.

4.2 Assessment B - "FinanCo NV" (mid-sized financial services firm, ~5,000 FTE)

Profile: mature on risk and compliance; behind on agentic-AI integration.

KA	Score	Notes
KA1	L4	Annual self-assessment; KPIs in board dashboard; policy-as-code piloting
KA2	L3	Portfolio review quarterly; value measured per system
KA3	L3	Reference architecture quarterly-reviewed; cognition plane positioned but not yet operational for any system
KA4	L3	Gates applied to 94%; CI/CD automated for ML
KA5	L3	Data-quality KPIs continuous; conceptual frameworks for core domain
KA6	L3	Model cards 92%; automated bias detection in CI/CD
KA7	L3	A/B testing + HITL protocols per risk level
KA8	L4	Full observability; MTTD <5 min; deploy 100% automated
KA9	L4	Quarterly red-team; supply-chain risks assessed; FM-coverage 100% but no agentic system yet
KA10	L4	ISO 42001 certified; automated compliance monitoring
KA11	L3	Bias audits in lifecycle; fairness monitoring nascent
KA12	L3	Trusted-source register; grounding rate 91%
KA13	L2	Competency profiles for board + key roles; register currency 76%; refresh adherence 78%; lowest KA

Aggregate maturity: **L2** (constrained by KA13). Even with multiple L4 KAs, the aggregation rule (minimum across KAs) holds the headline at L2 until KA13 lifts. **This is the right answer** under the

rubric: literacy is structurally critical and cannot be hidden behind strength elsewhere. A documented exception is possible (Art. 4 readiness gradual rollout), but should not become permanent.

4.3 Assessment C - "AgenticStartUp" (50-FTE startup, agentic AI as core product)

Profile: deep on architecture and risk for the agentic flagship; thin governance and literacy.

KA	Score	Notes
KA1	L2	Charter drafted; AI Board = co-founders
KA2	L2	Portfolio = 1 product + 3 internal tools
KA3	L4	Cognition plane operational; authority register live; signed decision artefacts; modelling-conventions clean variant
KA4	L3	Gates applied; CI/CD automated
KA5	L2	Provenance on all RAG sources; no conceptual framework
KA6	L3	Model cards 100%; foundation-model relayed
KA7	L3	UX patterns versioned; HITL for medium-risk
KA8	L3	Observability and CI/CD; cost monitoring active
KA9	L4	FM-coverage 100%; per-FM controls operational; red-team monthly
KA10	L2	Algoritmeregister N/A (not Dutch public); EU AI Act conformity in progress
KA11	L2	Ethical framework v0.5; ethics committee = co-founder + advisor
KA12	L3	Trusted-source register; grounding rate 93%
KA13	L1	No competency profiles; no literacy register; ad-hoc training; mandatory threshold not met

Aggregate maturity: L1 (KA13 below L2 mandatory threshold). Even though KA3, KA9 are L4, the framework's aggregation rule treats this as L1. **This is the correct verdict** despite feeling "wrong" - an organisation with zero literacy structure is *not* defensibly mature, however strong the architecture. The result is meant to be uncomfortable: lift KA13 to L2 (competency profiles for the 4 key roles, register started, 70% currency) and the aggregate jumps to L2. The growth path is short and high-leverage.

5. How to use the calibration kit

1. **Read all three exemplars first** before scoring your own organisation.
2. **Pick the exemplar closest to your portfolio** in size and sector.
3. **Per KA, compare evidence shape and thresholds.** Use the table in §2 to check whether your metrics clear the threshold.
4. **Flag every borderline score** explicitly. Borderline = within 10% of a mandatory threshold or missing one of the kinds (Document / Metric / Log).
5. **Run inter-assessor calibration** above L3 (rubric): two assessors independently, peer review on disagreement.
6. **Apply the aggregation rule** rigorously. Resist the temptation to average - the rule is *minimum* across KAs unless a documented exception covers the gap.
7. **Document exceptions** as in §3; review at the rubric's cadence.

6. Inter-assessor variance - known traps

- **Overscoring KA1.** Approved charter ≠ working governance. L3 requires evidence the governance *operates*.
- **Underscoring KA8.** Operations may be more mature than it looks because telemetry runs invisibly. Ask for the metrics rather than the runbook.
- **Overscoring KA10.** Algorithm register entry without a current DPIA is L2, not L3.
- **Overscoring KA11.** Ethics committee exists ≠ ethics committee functions. L3 needs cycle of decisions.
- **Overscoring KA13.** Articles 4 evidence file exists ≠ register currency at threshold. Check the currency metric.
- **Underscoring KA3 in non-agentic estates.** The cognition-plane mandatory items apply only when agentic systems are present. Pre-agentic, KA3 is scored against the classical architecture criteria.

7. Relationship with rubric

This kit is *additive* on the rubric. The rubric still defines what kinds of evidence are required; the kit adds the thresholds and the worked anchors. An organisation that prefers the rubric without numerical thresholds (e.g., because their portfolio is too small to support meaningful percentages) may use only the rubric - but they then take on the variance the panel reviewers flagged (P8).