

ARBEITSBLATT · VERSION 1.0

Agent Workforce Eval Scorecard

Bewerten Sie in 20 Minuten, ob ein KI-Agent nur überzeugend wirkt oder bereits messbar, sicher und wirtschaftlich betrieben werden kann.

Unternehmen**Agent / Use Case****Fachlicher Owner****Datum / Version**

Bewertungsprinzip: Vergeben Sie pro Aussage 0 Punkte (nicht vorhanden), 1 Punkt (teilweise/ad hoc) oder 2 Punkte (dokumentiert und im Betrieb nachweisbar). Übertragen Sie anschließend den Dimensionswert in die Zusammenfassung.

Die Scorecard ersetzt kein Security-, Datenschutz- oder Rechtsaudit. Sie schafft eine gemeinsame Gesprächsgrundlage für Fachbereich, IT, Datenschutz und Management.

DIMENSIONEN 1–3

Auftrag, Kontext und Qualität

1. Zweck und Verantwortung

× 1,5

Maximal 15 Punkte

- | | |
|---|---|
| ☐ Der geschäftliche Zweck ist in einem Satz definiert. | ☐ Ein fachlicher Owner verantwortet Ergebnis und Änderungen. |
| ☐ Nutzergruppe und erlaubte Aufgaben sind abgegrenzt. | ☐ Nicht erlaubte Aufgaben sind ausdrücklich dokumentiert. |
| ☐ Ein messbares Geschäftsergebnis ist festgelegt. | |

Rohwert (0–10)

× 1,5 =

2. Kontext und Daten

× 1,5

Maximal 15 Punkte

- | | |
|--|---|
| ☐ Freigegebene Wissens- und Datenquellen sind benannt. | ☐ Aktualität und verantwortliche Pflege sind geregelt. |
| ☐ Datenklassen und Übertragungsgrenzen sind dokumentiert. | ☐ Der Agent erhält nur aufgabenrelevanten Kontext. |
| ☐ Quellen können im Ergebnis nachvollzogen werden. | |

Rohwert (0–10)

× 1,5 =

3. Evals und Qualität

× 2

Maximal 20 Punkte

- | | |
|---|--|
| ☐ Mindestens 20 reale Testfälle bilden typische Arbeit ab. | ☐ Erwartete Ergebnisse und Akzeptanzkriterien sind definiert. |
| ☐ Grenz- und Fehlerfälle sind Teil des Testsets. | ☐ Relevante Produktionsfehler werden zu Regressionstests. |
| ☐ Vor Änderungen gibt es ein verbindliches Qualitätsgate. | |

Rohwert (0–10)

× 2 =

DIMENSIONEN 4–7

Handeln, Governance und Betrieb

4. Tools und Berechtigungen

× 1,5

Maximal 15 Punkte

- | | |
|--|---|
| ☐ Werkzeuge und Rechte folgen dem Minimalprinzip. | ☐ Schreibende Aktionen sind von lesenden getrennt. |
| ☐ Kritische Aktionen benötigen Freigabe. | ☐ Secrets liegen nicht in Prompts oder Skills. |
| ☐ Aktionen werden nachvollziehbar protokolliert. | |

Rohwert (0–10) × 1,5 =

5. Governance und Übergaben

× 1,5

Maximal 15 Punkte

- | | |
|---|---|
| ☐ Unsicherheit und Eskalationsregeln sind festgelegt. | ☐ Menschen können übernehmen, stoppen und korrigieren. |
| ☐ Freigaben richten sich nach Risiko und Datenklasse. | ☐ Vorfälle und Änderungen werden dokumentiert. |
| ☐ Datenschutz und Informationssicherheit sind eingebunden. | |

Rohwert (0–10) × 1,5 =

6. Kosten und Leistung

× 1

Maximal 10 Punkte

- | | |
|---|---|
| ☐ Kosten pro akzeptiertem Ergebnis werden gemessen. | ☐ Wiederholungen und menschliche Nacharbeit fließen ein. |
| ☐ Budgets und Warnschwellen sind technisch hinterlegt. | ☐ Modellwahl folgt Qualität, Risiko, Latenz und Kosten. |
| ☐ Eine manuelle Prozessbaseline ist dokumentiert. | |

Rohwert (0–10) =

7. Monitoring und Betrieb

x 1

Maximal 10 Punkte

- | | |
|---|---|
| ☐ Qualität, Kosten, Latenz und Eskalationen sind sichtbar. | ☐ Es gibt Alarmierung für kritische Abweichungen. |
| ☐ Modelle, Prompts, Skills und Quellen sind versioniert. | ☐ Ein regelmäßiger Review-Rhythmus ist festgelegt. |
| ☐ Backup, Wiederanlauf und Abschaltweg sind getestet. | |

Rohwert (0–10)

=

Agentifizierung · Agent Workforce Eval Scorecard · 3/4

AUSWERTUNG UND NÄCHSTE SCHRITTE

Vom Score zur Entscheidung

Dimension	Ergebnis	Größte Lücke / nächste Maßnahme
1. Zweck und Verantwortung		
2. Kontext und Daten		
3. Evals und Qualität		
4. Tools und Berechtigungen		
5. Governance und Übergaben		
6. Kosten und Leistung		
7. Monitoring und Betrieb		

Gesamtergebnis von 100 Punkten

Punkte	Reifestufe	Empfehlung
0–39	Nicht produktionsreif	Zweck, Datenzugriff und Verantwortung zuerst klären.
40–59	Pilot mit Leitplanken	Begrenzter Use Case, menschliche Freigabe und Basis-Evals.
60–79	Kontrollierter Betrieb	Produktiver Einsatz mit Monitoring und festen Qualitätsgates.
80–100	Skalierbar	Betriebsmodell auf weitere Agenten übertragen.

Die drei wichtigsten Maßnahmen für die nächsten 30 Tage