Corporate Execution Intelligence Benchmark (CEIB)

Ein architektur-agnostisches Framework zur Evaluation operativer KI-Systeme

---

Was ist das CEIB?

Das CEIB bewertet KI-Systeme nach einem simplen Prinzip: Es zählt nur, was hinten rauskommt. Egal ob klassisches LLM, Chain of Thought, Mixture of Expert, Multi-Agent-System oder experimenteller Ansatz - entscheidend ist die Qualität der generierten Handlungsanweisungen für reale Unternehmensaufgaben.

Die Performance Assessment Base (PAB) ist dabei keine theoretische Spielerei, sondern ein praktisches Werkzeug zur Quantifizierung operativer Leistung.

Warum noch ein Benchmark?

Traditionelle KI-Benchmarks messen linguistische Fähigkeiten, Faktenwissen und Reasoning. Das CEIB dagegen fragt: Kann dieser Output korrekt darstellen, was ein menschlicher Experte tun würde und wie er mit anderen menschlichen Experten zur Problemlösung zusammenarbeitet.

Typische Anwendungsfälle:

- Evaluation von Prozessautomatisierung
- Vergleich verschiedener Architekturen in operativen Szenarien
- Qualitätssicherung vor Produktiveinsatz

Vendor Selection im Enterprise-Bereich

Die Methodik

Performance Assessment Base (PAB)

Die Bewertung erfolgt über eine gewichtete Summation normalisierter Einzelscores:

$PAB = \Sigma(i=1 \text{ to } n)\ wi \times si$

wobei:

- $wi \in [0,1]$ das normierte Gewicht der Kategorie i
- $si \in [0,100]$ den erzielten Score in Kategorie i
- $n = 5$ die Anzahl der Bewertungskategorien
- $\Sigma(i=1 \text{ to } n)\ wi = 1$ (Gewichte summieren sich zu 1)

Die fünf Bewertungskategorien

Kategorie	Gewicht wi	Was wird gemessen?
Execution Granularity Index (EGI)	0,25	Wie konkret sind die Handlungsschritte?
Aufgabenerfüllung	0,25	Wird das Ziel erreicht?
Reaktivität/Praktikabilität	0,20	Funktioniert es im echten Leben?
Vollständigkeit	0,15	Fehlt etwas Wichtiges?
Handlungslogik	0,15	Macht die Abfolge Sinn?

Verifikation: 0,25 + 0,25 + 0,20 + 0,15 + 0,15 = 1,00 ✓

Die Gewichtung ist kein Zufall. EGI und Aufgabenerfüllung sind mit je 25% am höchsten gewichtet, weil sie direkt über Brauchbarkeit entscheiden. Reaktivität folgt mit 20%. Vollständigkeit und Logik sind mit je 15% gewichtet - wichtig, aber oft durch Nacharbeit kompensierbar.

Das CEIB evaluiert **Execution Intelligence** – die Fähigkeit, operative Probleme durch konkrete Handlungsschritte zu lösen. Der Fokus liegt auf den **ausführenden und koordinierenden Aktivitäten**, die zur Problemlösung führen.

Abgrenzung: Was gehört nicht zur Execution Intelligence?

Nachgelagerte Prozesse (außerhalb des Bewertungsumfangs):

Buchhaltung & Finanzverwaltung:

- Buchungsanweisungen, Kontierungen, Bilanzierungsschritte
- Diese Aktivitäten erfolgen nach der operativen Problemlösung

Dokumentation & Archivierung:

- Speicherung in Systemen, Archivierung von Vorgängen
- Diese Aktivitäten dienen der Nachvollziehbarkeit, nicht der Problemlösung selbst

---

Die Kategorien im Detail

1. Execution Granularity Index (EGI) - Gewicht: 0,25

Der EGI misst, ob das Modell in der Lage ist Einzel-Schritte konzeptionell so aufzubauen, dass es ohne weiteres Nachfragen und Halluzination damit das Problem lösen kann. Es kommt auf Feingliedrigkeit und textuellen Umfang an, wobei der Text stets auch relevant für die Problemlösung und ohne unnötige Wiederholung auskommen muss.

Bewertungsskala:

- 9-10: Sofort umsetzbar, alle Infos vorhanden (Namen, Kontakte, Zeiten, Referenznummern)
- 7-8: Gut detailliert, kleinere Details fehlen
- 5-6: Grundstruktur da, aber vage Formulierungen ("baldmöglichst", "zuständige Person")
- 3-4: Zu abstrakt für direkte Umsetzung
- 0-2: Praktisch unbrauchbar

2. Aufgabenerfüllung - Gewicht: 0,25

Simpler Check: Wurde die Aufgabe gelöst oder nicht?

Beispiel - Aufgabe: "Erstelle einen Workflow für die Bearbeitung einer Mieterschadenmeldung."

Score 9-10/10:

- End-to-End-Prozess von Meldung bis Behebung

Score 2-3/10:

- Nur Schadensmeldung behandelt

3. Reaktivität/Praktikabilität - Gewicht: 0,20

Hier trennt sich Theorie von Praxis. Funktioniert die Lösung im echten Unternehmensalltag?

Gut:

- "während der Bürozeiten 9-17 Uhr"
- "innerhalb von 2 Werktagen"
- Eskalationswege bei Nichterreichbarkeit
- Alternative Kontaktmöglichkeiten

Schlecht:

- "sofort"
- "innerhalb einer Stunde" (für nicht-kritische Prozesse)
- Ignorieren von Hierarchien
- Unklare Verantwortlichkeiten

4. Vollständigkeit - Gewicht: 0,15

Checkliste:

- Vorbereitung abgedeckt?
- Hauptprozess vollständig?
- Ausnahmen behandelt?
- Kommunikationswege definiert?
- Dokumentation berücksichtigt?
- Nachbereitung integriert?
- Eskalationspfade vorhanden?

5. Handlungslogik - Gewicht: 0,15

Die innere Konsistenz der Lösung. Bauen die Schritte aufeinander auf? Gibt es Widersprüche? Sind Abhängigkeiten beachtet?

Warnsignale:

- Zirkuläre Abhängigkeiten
- Schritte setzen nicht vorhandene Informationen voraus
- Inkonsistente Detailtiefe

- Widersprüchliche Anweisungen
- Unnötige Wiederholungen

Score-Berechnung

Normalisierung

Jeder Rohscore ri wird auf [0,100] normalisiert:

si = (ri / ri_max) × 100

Beispiel: Rohscore 8,5 von 10 → si = (8,5/10) × 100 = 85

Gesamtformel

PAB = 0,25 × sEGI + 0,25 × sAufgabe + 0,20 × sReaktivität + 0,15 × sVollständigkeit + 0,15 × sLogik

Wertebereich: PAB ∈ [0, 100]

Corporate Utility Score (CUS)

Für den Vergleich zweier Systeme A und B:

CUS(A,B) = (PAB_A / (PAB_A + PAB_B)) × 100

Der CUS gibt den prozentualen Leistungsanteil von System A an.

Interpretation:

- CUS = 50: Gleichwertige Performance
- CUS > 50: System A überlegen
- CUS < 50: System B überlegen

Symmetrie: CUS(A,B) + CUS(B,A) = 100

Alternative Darstellung über Differenzen: CUS = 50 + Σ(i=1 to n) wi × Δsi / 2 mit Δsi = si(A) - si(B)

Praxisbeispiel

Zwei Systeme werden für einen Hausverwaltungsprozess evaluiert.

Aufgabe: "Erstelle einen detaillierten Workflow für die Bearbeitung einer Schadensmeldung eines Mieters, inklusive Koordination mit der Hausverwaltung, Beauftragung eines Handwerkers."

System A - Rohscores:

- EGI: 8,5/10
- Aufgabenerfüllung: 9,0/10
- Reaktivität: 8,0/10
- Vollständigkeit: 7,5/10
- Handlungslogik: 8,8/10

System B - Rohscores:

- EGI: 7,0/10
- Aufgabenerfüllung: 8,5/10
- Reaktivität: 9,0/10
- Vollständigkeit: 8,0/10
- Handlungslogik: 7,2/10

Nach Normalisierung und Gewichtung:

- PAB_A = 84,20
- PAB_B = 79,55

CUS(A,B) = (84,20 / 163,75) × 100 = 51,4% CUS(B,A) = 48,6%

Ergebnis: System A zeigt mit 2,8% Vorsprung eine marginale Überlegenheit. Tr*otz des knappen Gesamtergebnisses ist zu beachten, dass System A in der für Workflows kritischen Kategorie 'Handlungslogik' signifikant besser abschneidet.*

Interpretationsleitfaden

PAB-Skala

PAB 90-100: Produktiveinsatz empfohlen PAB 75-89: Produktiveinsatz nach Review PAB 60-74: Optimierung vor Produktiveinsatz nötig PAB 45-59: Wesentliche Überarbeitung erforderlich PAB 0-44: Nicht produktionsreif

CUS-Differenzen

ΔCUS < 3%: Praktisch gleichwertig ΔCUS 3-5%: Marginaler Vorteil ΔCUS 5-10%: Relevanter Unterschied ΔCUS 10-15%: Deutliche Überlegenheit ΔCUS 15-20%: Starke Überlegenheit ΔCUS > 20%: Dominanz

---

Qualitätssicherung

Inter-Rater-Reliabilität

Zur Objektivitätssicherung sollte die Übereinstimmung zwischen Evaluatoren gemessen werden:

Übereinstimmung = 1 - (|Score_Evaluator1 - Score_Evaluator2| / Score_max)

Zielwert: > 0,80

Häufige Fehlerquellen

1. Halo-Effekt: Gute Bewertung in einer Kategorie färbt auf andere ab
2. Reihenfolge-Bias: Erstes System wird bevorzugt/benachteiligt
3. Erwartungs-Bias: Vorwissen über Systemarchitektur beeinflusst Bewertung

Gegenmaßnahmen: Randomisierung, Anonymisierung der Systeme, strukturierte Bewertungsbögen.

Anpassung an spezifische Kontexte

Das Framework lässt sich an verschiedene Problemstellungen anpassen. Die Summe der Gewichte muss dabei stets 1 ergeben.

Beispiel Healthcare (Vollständigkeit kritisch):

- EGI: 0,20
- Aufgabe: 0,25
- Reaktivität: 0,15
- Vollständigkeit: 0,25 (erhöht)
- Logik: 0,15

Beispiel Logistik (Pragmatismus und Routenoptimierung zählt):

- EGI: 0,20
- Aufgabe: 0,30 (erhöht)
- Reaktivität: 0,30 (erhöht)
- Vollständigkeit: 0,10 (reduziert)
- Logik: 0,10 (reduziert)

Bei Bedarf können zusätzliche Kategorien ergänzt werden (z.B. Compliance, Kreativität), wobei die Gewichte entsprechend angepasst werden müssen.

Limitationen

Das CEIB bewertet nur die Outputqualität. Nicht erfasst werden:

- Technische Effizienz (Inferenzzeit, Ressourcenverbrauch)
- Kosten pro Anfrage
- Skalierbarkeit unter Last
- Robustheit bei fehlerhaften Eingaben
- Konsistenz bei wiederholten Anfragen
- Aufgaben der administrativen Nacherfassung wie Dokumentation und Rechnungslegung

Diese Aspekte müssen separat evaluiert werden.

Trotz strukturierter Kriterien bleibt eine gewisse Subjektivität in der Bewertung. Multiple Evaluatoren und klare Kriterien helfen, diese zu minimieren.

Durchführung einer Evaluation

Vorbereitung:

1. Aufgabe klar definieren
2. Systeme unter identischen Bedingungen testen
3. Evaluatoren kalibrieren

Bewertung:

1. Unabhängige Bewertung durch alle Evaluatoren
2. Dokumentation aller Scores mit Begründung
3. Konsensbildung bei Abweichungen > 2 Punkte

Dokumentation:

1. Detaillierte Kategorienbewertungen
2. PAB und CUS Berechnungen
3. Interpretation und Empfehlungen

Zusammenfassung

Das CEIB bietet ein praktisches Framework zur Bewertung operativer KI-Systeme. Es fokussiert auf die tatsächliche Brauchbarkeit der Outputs statt auf technische Details. Die mathematische Fundierung ermöglicht objektive Vergleiche, während die Flexibilität Anpassungen an spezifische Kontexte erlaubt.

Der Kern: Fünf gewichtete Kategorien ergeben einen Gesamtscore (PAB), der absolute Leistung quantifiziert. Der CUS ermöglicht relativen Systemvergleich.

Version: CEIB 1.1
Stand: August 2025

Das CEIB (Corporate Executive Intelligence Benchmarking) wird als Open-Source-Framework unter der MIT-Lizenz veröffentlicht, um eine breite Adoption in Forschung und Praxis zu fördern und die transparente Bewertung von Execution Intelligence europaweit zu standardisieren.This is a living document.

Urheber ist die European AI Watch Association. Wir freuen uns über jede freundliche namentliche Erwähnung.