

Korrektheit trotz Nicht-Determinismus

KI-Anforderungen in der Praxis



10. März 2026





Dr. Felix Kammerlander

Berater und Trainer für Softwarearchitektur



felix.kammerlander@embarc.de



linkedin.com/in/felix-kammerlander



fkammerlander.io



A close-up photograph of a brown dog's face. The dog has its mouth open, showing its teeth and pink tongue. On its forehead and around its eyes, there are several small, white, flower-shaped stickers. The background is a solid light purple color.

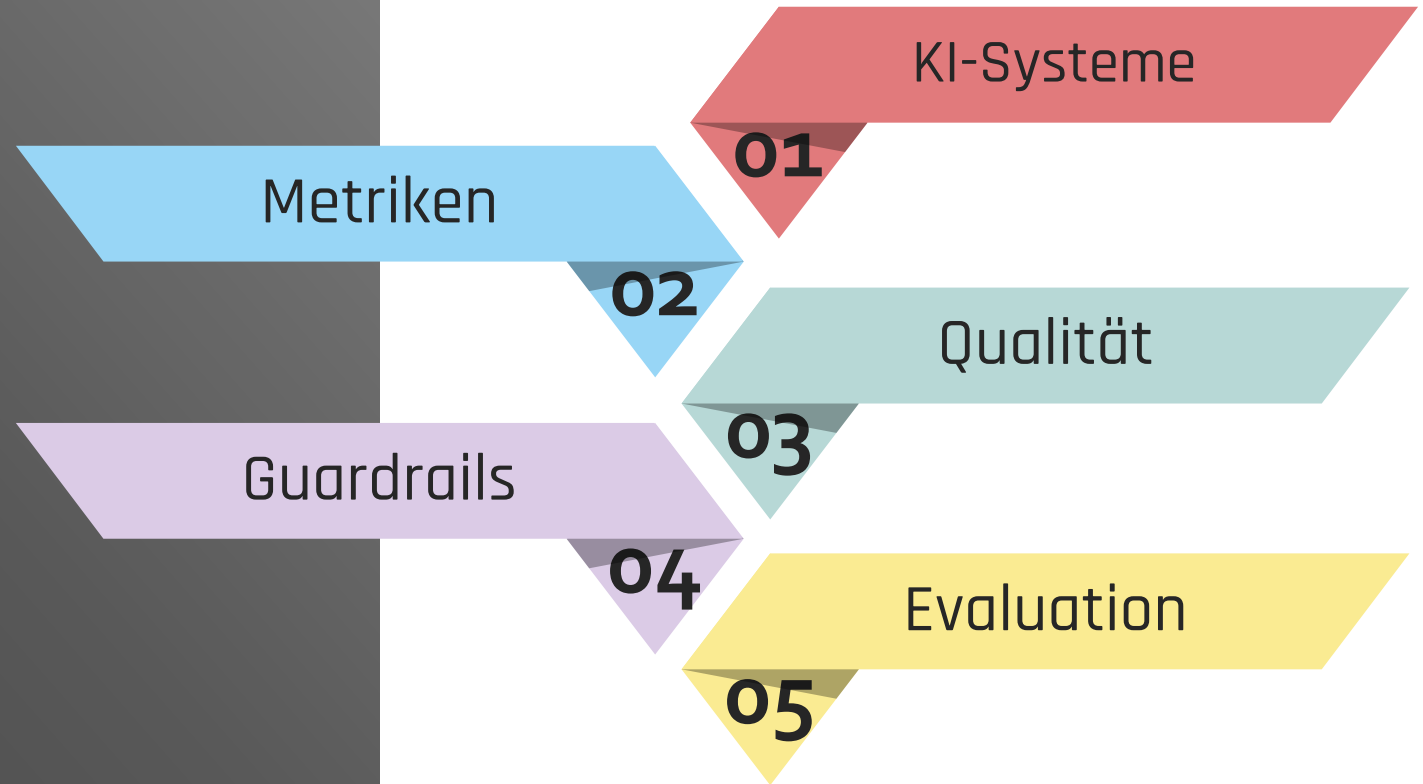
Wir müssen da jetzt mal was mit KI machen!



Oh boy...

00.

Agenda

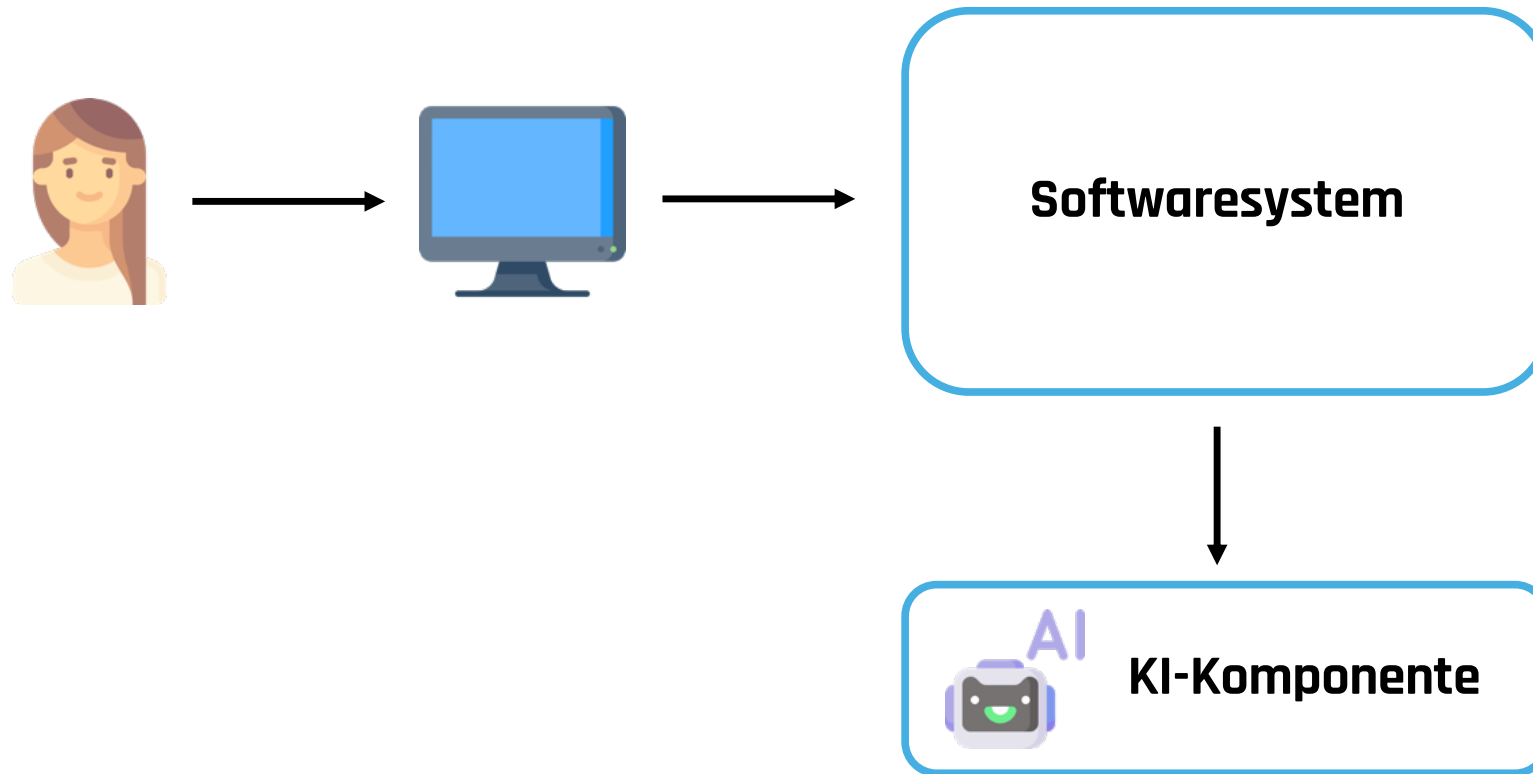


01.

KI-Systeme sind anders



KI-Systeme

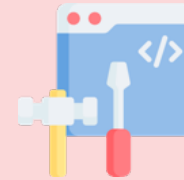


Einsatz von KI-Systemen vs. Klassische Softwareentwicklung



KI-Systeme

- Unklare Regeln
- Induktiv, Datengetrieben
- Sich ändernde Muster



Klassische Softwareentwicklung

- Eindeutige oder einfache Regeln, deduktiv
- Geringe Fehlertoleranz
- Mangelnde Datenbasis
- Hohe Erklärbarkeit gefordert



Unterschiede zu klassischen Softwaresystemen



Induktiv vs. deduktiv



Daten vs. Code



Zufall vs. Determinismus



Blackbox vs. Whitebox



Präzise vs. datengetriebene Spezifikation

```

1 def is_toxic(post: str) -> bool:
2     if contains_insult_word(post):
3         return True
4     elif matches_pattern(post, "du bist [negativ]"):
5         return True
6     elif contains_silencing_or_exclusion_phrase(post):
7         return True
8     elif devalues_person_instead_of_argument(post):
9         return True
10    else:
11        return False

```

```

1 Post: „Danke für die Erklärung, das hat mir geholfen.“
2 Label: nicht toxisch
3
4 Post: „Du bist echt zu dumm, um das zu verstehen.“
5 Label: toxisch
6
7 Post: „Ich sehe das anders, weil deine Argumente wichtige Punkte auslassen.“
8 Label: nicht toxisch
9
10 Post: „Halt einfach die Klappe, niemand will deinen Müll lesen.“
11 Label: toxisch
12
13 Post: „Dein Beitrag ist unklar formuliert, kannst du ihn präzisieren?“
14 Label: nicht toxisch
15
16 Post: „Was für ein kompletter Idiot schreibt so etwas?“
17 Label: toxisch

```

Herausforderungen



Spezifikation nicht vollständig möglich



Neue Qualitätsmerkmale, enge Verknüpfung mit funktionalen Anforderungen



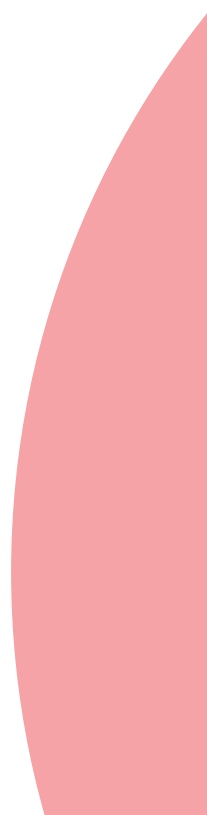
Komplexe Trade-offs



Distanz zwischen Technik und Business

Wann KI-Systeme einsetzen?

	Deduktive Probleme	Induktive Probleme
Niedrige Fehlertoleranz	Klassische Software	„KI-Assistent“
Hohe Fehlertoleranz	Heuristiken / einfache Skripte	„KI-System“



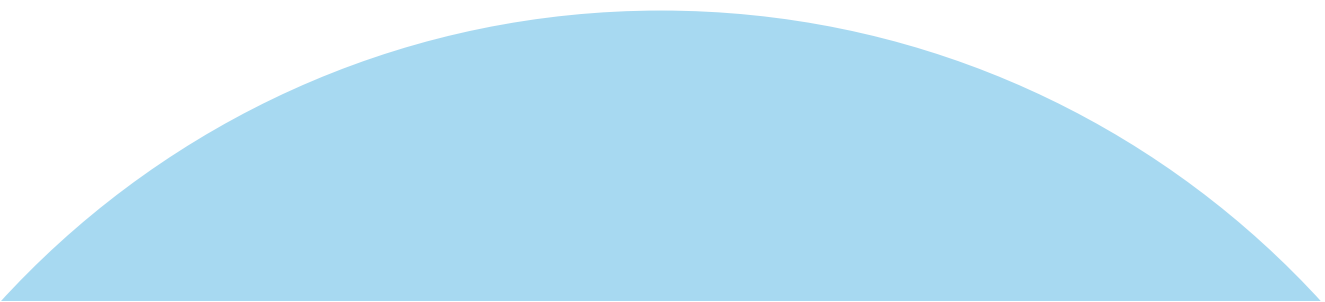
02.

Metriken



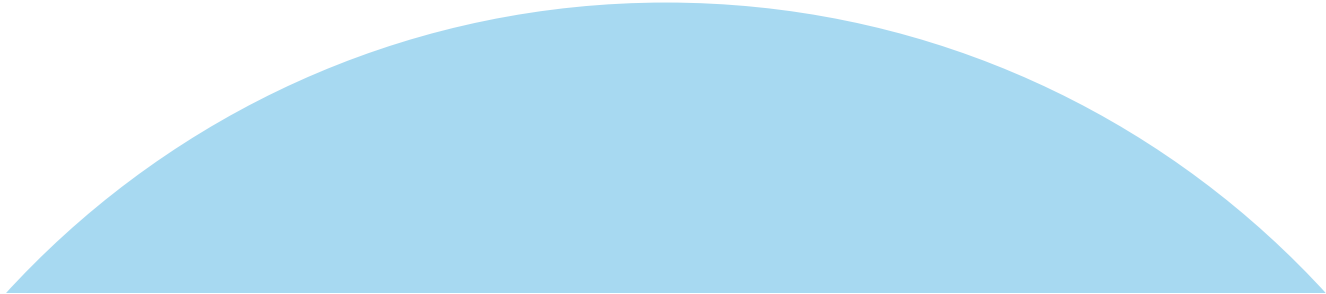
Was ist das Ziel? Ein Beispiel

Wir wollen Kreditkartenbetrug mit KI erkennen.



Was ist das Ziel? Ein Beispiel

Wir brauchen ein System, das finanzielle Verluste durch Kreditkartenbetrug in Echtzeit verhindert.



Was ist das Ziel? Ein Beispiel

„Wann rechnet sich das?“

Wir brauchen ein System, das finanzielle Verluste durch Kreditkartenbetrug in Echtzeit verhindert.

„Wie gut erkennen wir das?“

„Wie schnell muss das sein?“



Metriken



Domänenspezifische Fähigkeiten

Functional Correctness:

pass@k

Faktenwissen und Logik:

Accuracy, Precision, Recall



Generierungsqualität

Lokale / Globale Konsistenz:

Entailment vs. Contradiction

Safety:

Violation Rate, False Refusal Rate



Instruction-Following

Formate:

% Korrekt befolgte Anweisungen

Role-Playing:

% passende Antworten, AI-Judge Score



Kosten und Latenz

Latenz:

TTFT, TPOT, Total Latency = $TTFT + TPOT * \#Tokens$

Durchsatz:

TPS, RPM, Goodput

Kosten:

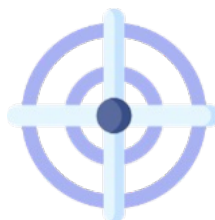
Cost per Request, Cost per Token

Beispiel: Kreditkartenbetrug



Latenz

Total Latency < 50ms, sonst
Zahlung akzeptieren



Korrektheit

Präzision > 95% (max. 5
Fehlalarme / 100 Sperrungen
Wichtiger als Recall, da
Kunden verärgern teurer ist



Kosten

Erwarteter Verlust durch Betrug
X Euro, 10.000 Überprüfungen
am Tag
 $CPR \ll X / 10.000$

03.

Qualität



Klassische vs. KI-Anforderungen

	Klassische Anforderungen (Traditional RE)	Anforderungen an KI-Systeme (RE4AI)
Natur	Deterministisch. Gleicher Input führt immer zum gleichen Output	Probabilistisch & Nicht-Deterministisch. Ergebnisse sind Wahrscheinlichkeiten, Verhalten kann variieren Benötigt Schwellenwerte
Fokus	Funktionale Korrektheit und definierte Geschäftsregeln	Datenqualität, Trainingsziele, Performanz-Metriken (Genauigkeit) und ethische Leitplanken
Spezifizierbarkeit	Vollständig und präzise vorab spezifizierbar ("Das System muss...")	Oft schwer abstrakt zu definieren (z.B. "Erkenne einen Fußgänger"). Spezifikation erfolgt oft implizit durch Daten
Qualitätsziele (NFRs)	Wartbarkeit, Sicherheit, Performance, Usability, ...	+ Erklärbarkeit (Explainability), Fairness, Vertrauen, Robustheit gegen neue Daten, Vermeidung von Bias, ...

Qualitätsmerkmale (Begriffe nach ISO 25010:2023)



Funktionale Eignung (Functional Suitability)

Sind die berechneten Ergebnisse genau genug / exakt, ist die Funktionalität angemessen? ...



Effizienz (Performance Efficiency)

Antwortet die Software schnell, hat sie einen hohen Durchsatz, einen geringen Ressourcenverbrauch? ...



Kompatibilität (Compatibility)

Ist die Software konform zu Standards, arbeitet sie gut mit anderen zusammen? ...



Benutzbarkeit (Interaction Capability)

Ist die Software intuitiv zu bedienen, wiedererkennbar, leicht zu erlernen, attraktiv? ...



Zuverlässigkeit (Reliability)

Ist das System verfügbar, tolerant gegenüber Fehlern, nach Abstürzen schnell wieder hergestellt? ...



Sicherheit (Security)

Ist das System sicher vor Angriffen? Sind Daten und Funktion vor unberechtigt Zugriff geschützt? ...



Wartbarkeit (Maintainability)

Ist die Software leicht zu ändern, erweitern, testen, verstehen? Lassen sich Teile wiederverwenden? ...



Übertragbarkeit (Flexibility)

Ist die Software leicht auf andere Situationen oder Zielumgebungen (z.B. anderes OS) übertragbar? ...



Betriebssicherheit (Safety)

Sind Personen, Tiere, Sachen und Umwelt vor Schäden durch die Software geschützt? ...

Reproduzierbarkeit und Prüfbarkeit



Reproduzierbarkeit

Gleiche Eingaben führen zu gleichen Ergebnissen



Auditierbarkeit

Nachvollziehbare Beweiskette für jede Vorhersage eines KI-Systems. Ermöglicht externe Überprüfung

Erklärbarkeit und Interpretierbarkeit



Erklärbarkeit

„The ability to describe the behavior of a system in understandable language to humans.

Explainability helps us understand what caused an AI system to reach a prediction“



Interpretierbarkeit

„Refers to the visibility and understanding of the inner logic and mechanics of the AI model.

An AI model with high interpretability allows us to understand how the components of the AI model (nodes and weights in deep neural network models) produce a mapping between a system input and output“

Beispiel: Kreditkartenbetrug



Erklärbarkeit

Ablehnung von Transaktionen muss für rechtliche Prüfungen begründet werden können



Fairness

Keine Ungleichbehandlung, etwa von ethnischen Gruppen

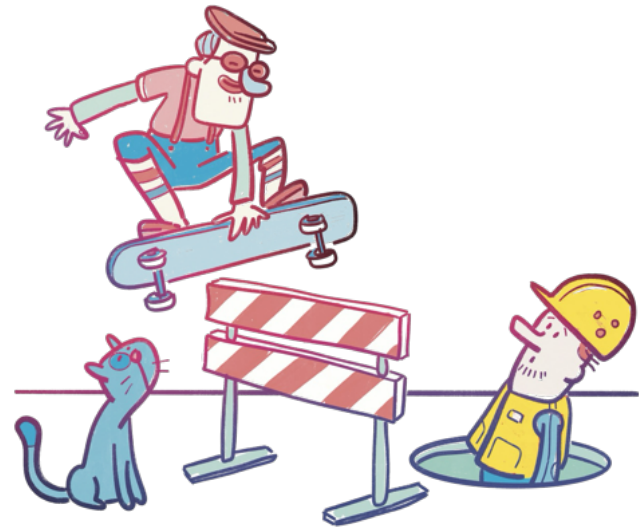


Reproduzierbarkeit

Wiederholung einer Prüfung mit gleichen Daten gelangt zum selben Ergebnis

D4.

Guardrails



Einschränkungen & Guardrails

Input Guardrails

PII-Erkennung
Toxicity
Prompt-Injection



Output Guardrails

Leak-Prevention
Faktencheck
Redaktion



Tool / Action Guardrails

Allowlist
Parameter-Schema
Action Scope



Context Guardrails

Rollenrechte
Freigaben
Audit trail

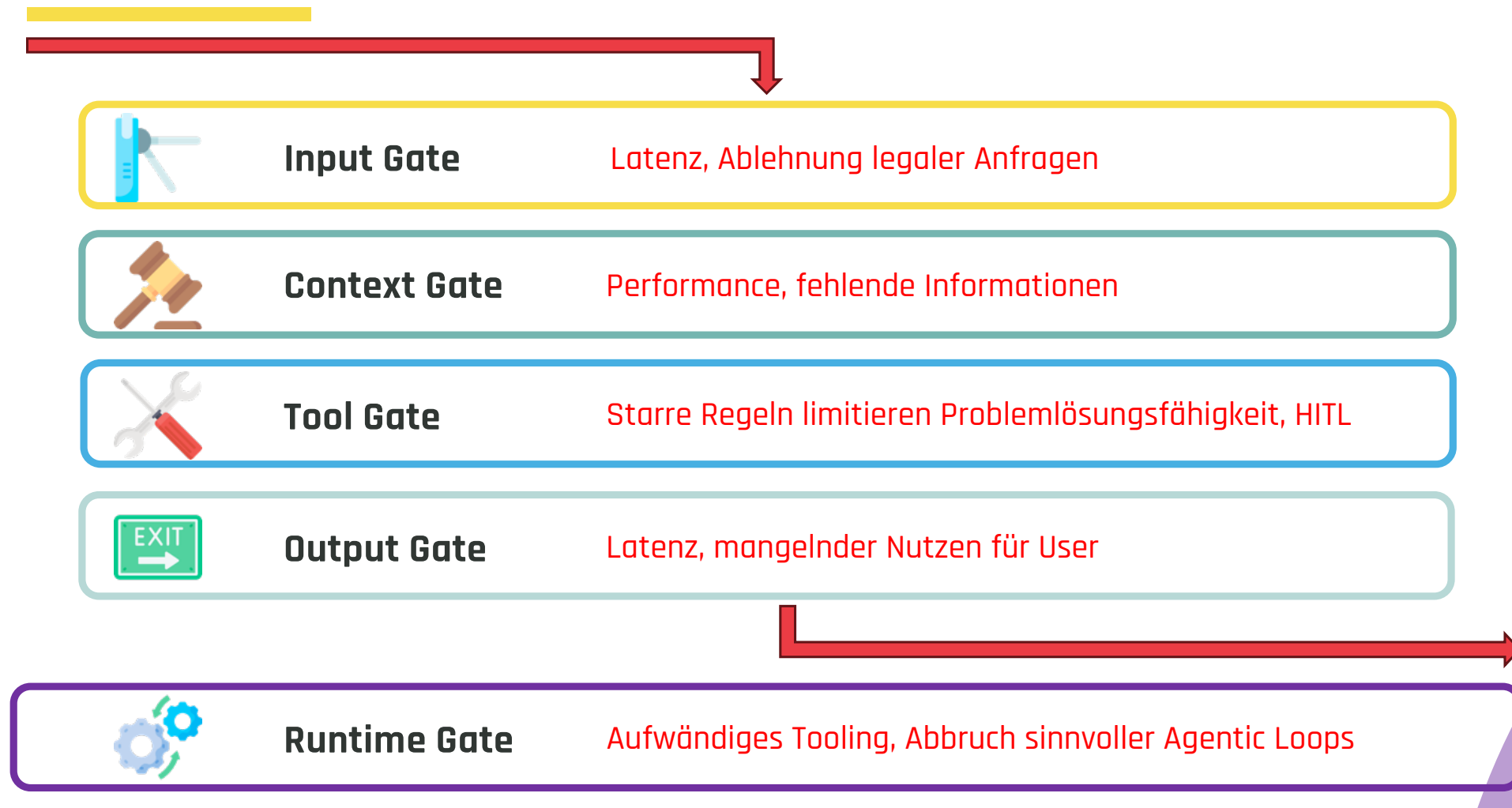


Runtime Guardrails

Timeouts
Budgets
Fallbacks



Einschränkungen & Guardrails: Tradeoffs



Beispiel: Kreditkartenbetrug

Runtime Guardrails

Timeout: x ms, danach
Transaktion durchlassen
und Nachprüfung
veranlassen



Input Guardrails

Entfernung geschützter
Merkmale (Alter,
Geschlecht,
Herkunft)



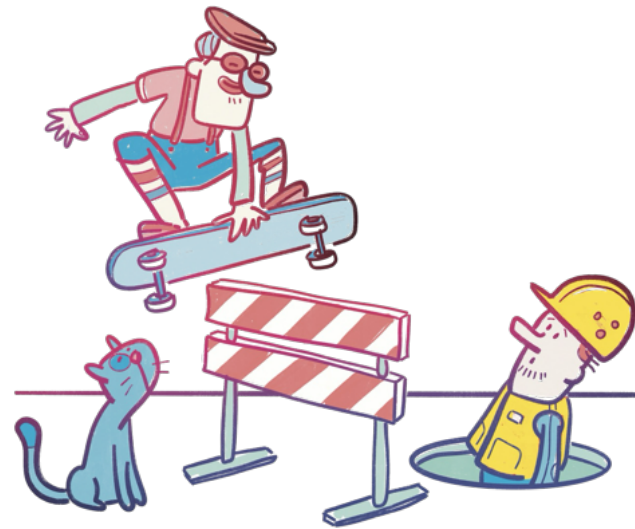
Output Guardrails

Bei Ablehnung:
Begründung,
Quellenverweise,
Konfidenz-Score



05.

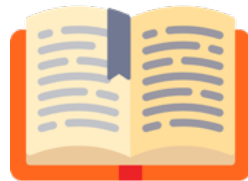
Evaluation



Evaluation: Herausforderungen



Zufall / Nicht-Determinismus



Offene Aufgabenstellung, riesiger Raum an Möglichkeiten



Hohe Komplexität der Outputs



Stilbewertungen



Biases & Fairness

Exact Evaluation & Similarity

Vergleich des Outputs mit einer Referenzlösung;
Überprüfung mit “Golden Sets”



Exakte
Überprüfung

- Einfache Evaluation bei gut überprüfbaren Ergebnissen
- Reproduzierbarkeit
- Automatisierbar, Hohe Skalierbarkeit

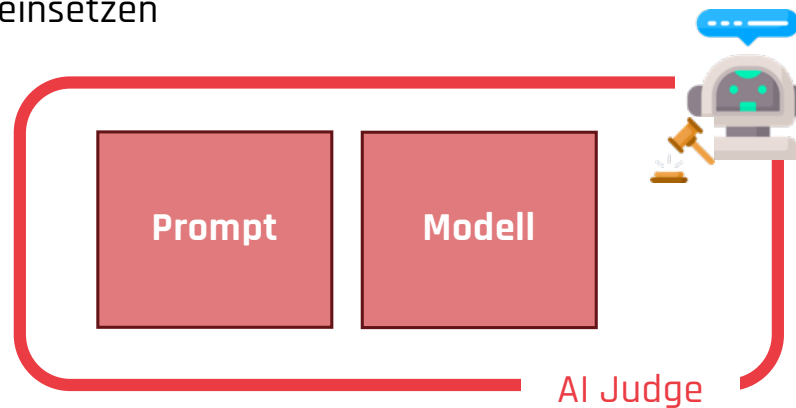


Semantische Ähnlichkeit überprüfen,
etwa über Distanzen von
Embeddings

- Nicht für alle Aufgaben geeignet
- Schwierig für Outputs, die nicht Text sind
- Anfällig für oberflächliche Übereinstimmung
- Stil oder Argumentation wird nicht erfasst
- Testfälle müssen gut entworfen und kuratiert werden

AI-as-a-Judge

KI zur Bewertung von Outputs einsetzen

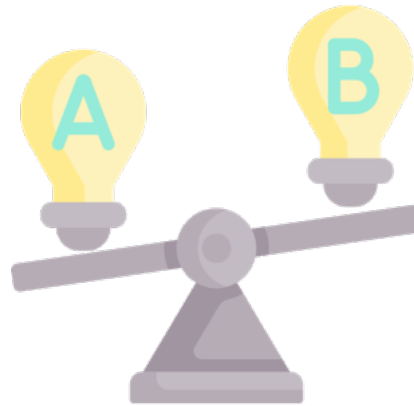


- Für jeden Use-Case und jedes Output-Format einsetzbar
- Breite Bewertungskriterien
- Skalierbar, automatisierbar

- Bias, etwa Bevorzugung gleicher Modelle
- Inkonsistente Bewertung
- Halluzination
- Erfasst ggf. formale Kriterien nicht zuverlässig
- Veränderungen über Zeit
- Teuer

Comparing Evaluation

Vergleich zweier Outputs und Ranking von Optionen (etwa durch Menschen oder Heuristiken)



- Robust gegen Skalenprobleme
- Stil, Klarheit, Nützlichkeit etc. fließen ein
- Einfacher als absolute Bewertung
- Multimodal möglich

- Viele Vergleiche nötig
- Reihenfolge präsentierter Optionen beeinflussen das Urteil
- Inkonsistente Beurteilungen
- Unklare Bewertungsmaßstäbe
- Nicht für jede Fragestellung sinnvoll

Beispiel: Kreditkartenbetrug



Exact Evaluation

Überprüfung auf fortlaufendem Satz an
Transaktionsdaten der letzten 30 Tage (Data Drift)
Precision > 95%

KI-Anforderungen auf einen Blick



Business Value & Metriken

Was ist der Zweck für den KI-Einsatz? Welche Metriken sollen verbessert werden?



Qualitative Aspekte

Welche neuen Qualitätsmerkmale spielen eine Rolle? Welche Trade-offs müssen ausgehandelt werden?



Einschränkungen

Welche Grenzen dürfen nicht überschritten werden? Wo muss KI überwacht oder reglementiert werden?



Evaluation

Wie überprüfen wir das System? Wann sind die Anforderungen erfüllt?



Lust auf mehr? Besucht unser 3-Tage iSAQB CPISA-A Modul SWARC4AI!



SOCREATORY
The Software Creators' Academy

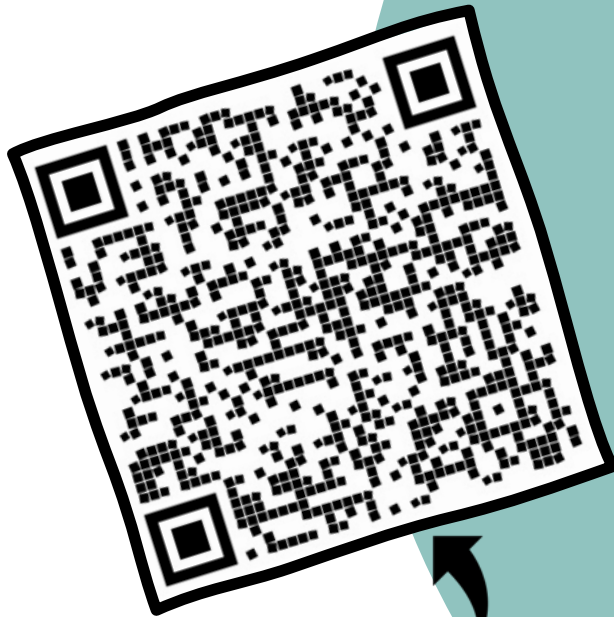
d ML-Systeme entwerfen

els), Agentic Systems und GenAI:
nächste Level und lernen Sie, wie Sie
KI-Systeme entwickeln.



Vielen Dank!

An welche Anforderungen mit KI-Bezug müsst ihr “zu Hause” nochmal ran?



Folien
downloaden



fk@embarc.de



linkedin.com/in/felix-kammerlander/



fkammerlander.io

