Korrektheit trotz Nicht-Determinismus

KI-Anforderungen in der Praxis

THE AGENTmakers

19. März 2026

embarc

Dr. Felix Kammerlander

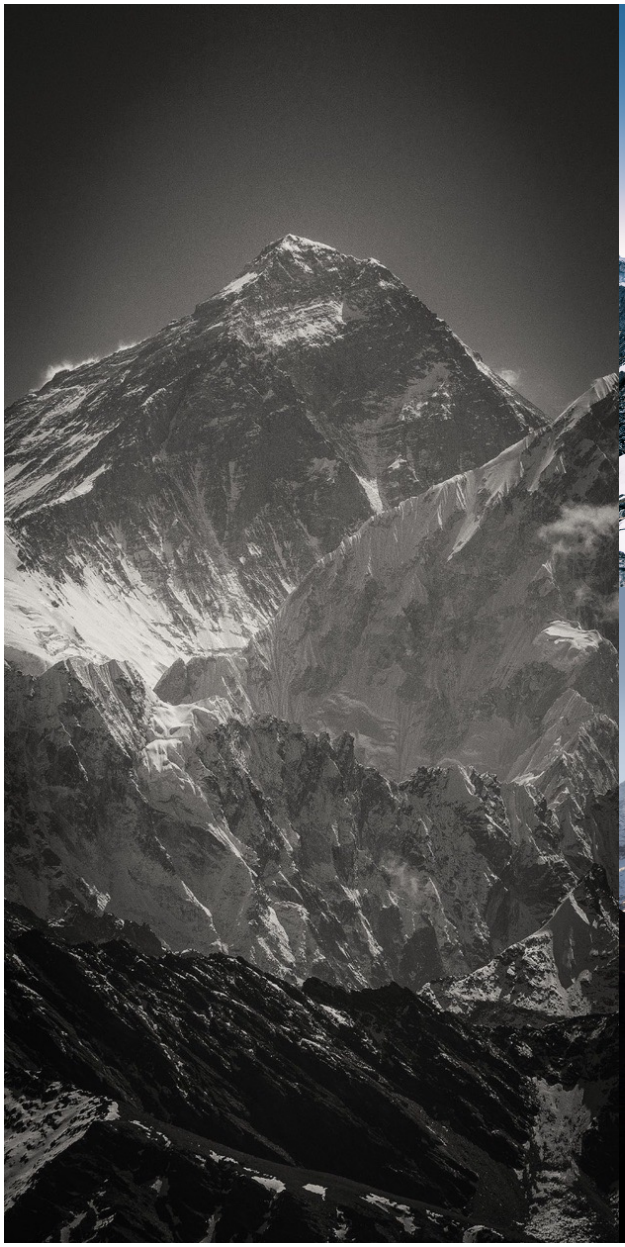
Berater und Trainer für Softwarearchitektur

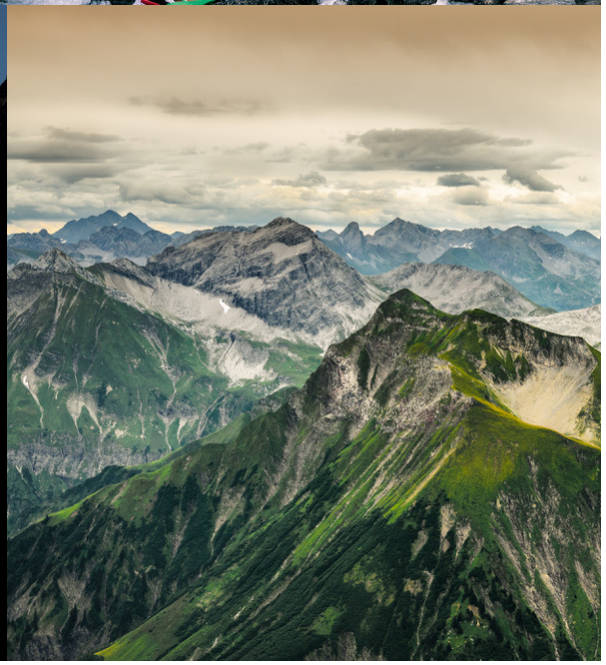
felix.kammerlander@embarc.de

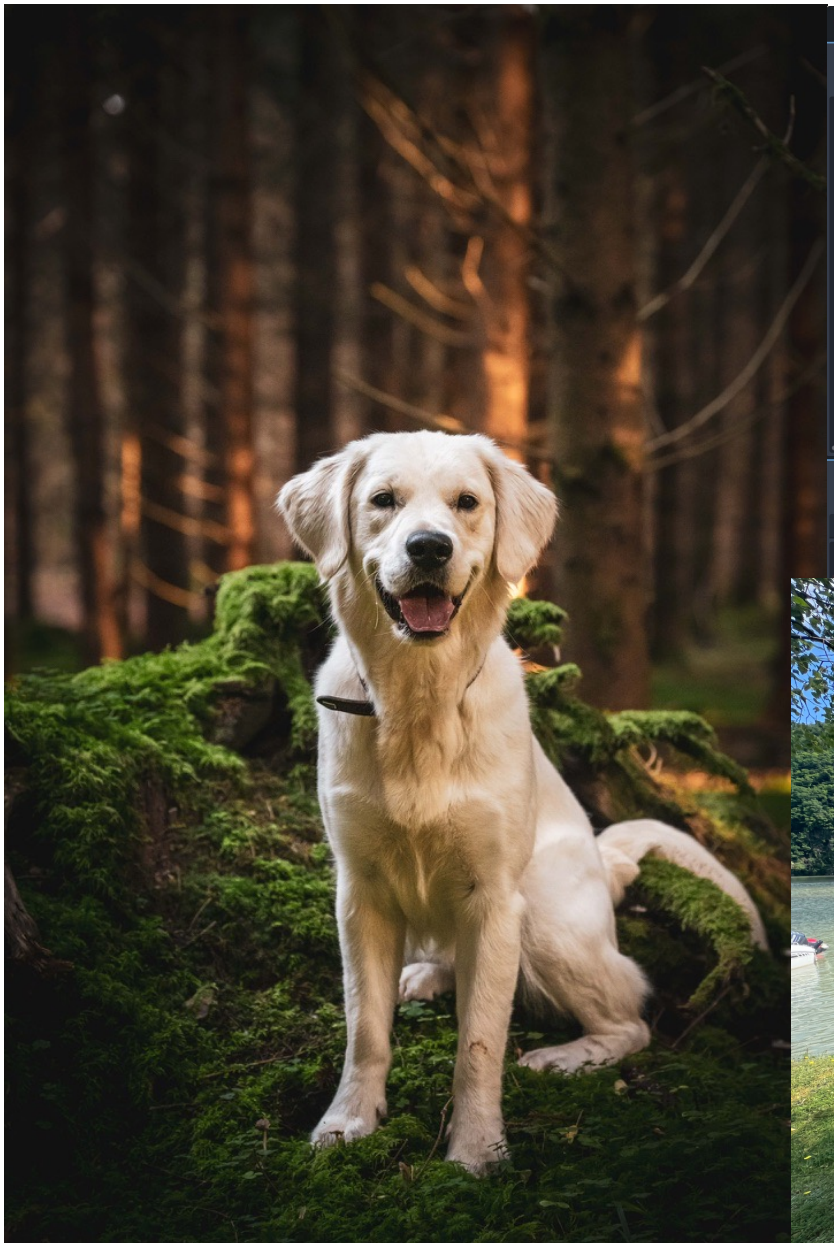
linkedin.com/in/felix-kammerlander

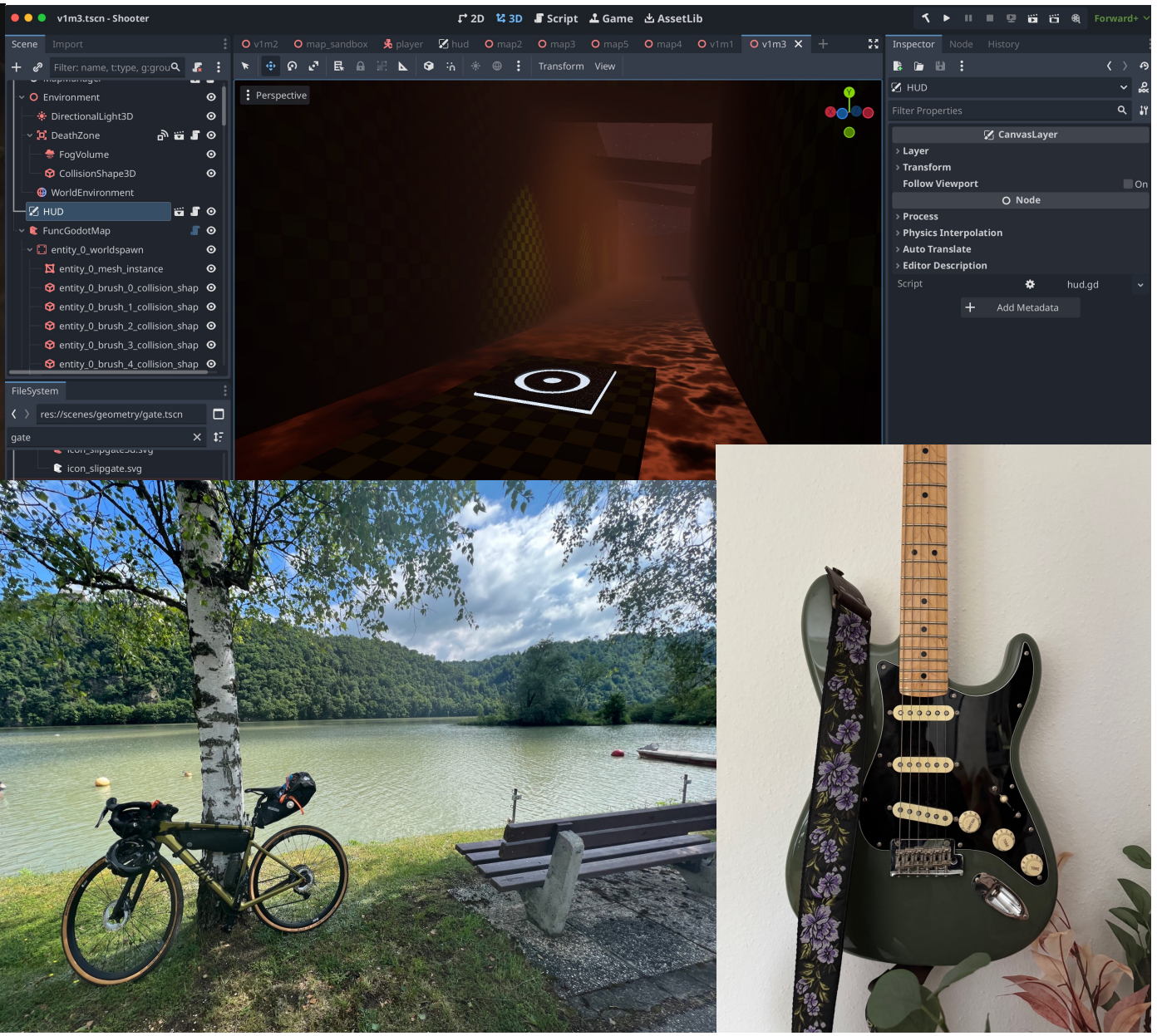
fkammerlander.io

embarc








v1m3.tscn - Shooter
2D
3D
Script
Game
AssetLib
Forward+
Scene
Import
Filter: name, t:type, g:grou
Environment
DirectionalLight3D
DeathZone
FogVolume
CollisionShape3D
WorldEnvironment
HUD
FuncGodotMap
entity_0_worldspawn
entity_0_mesh_instance
entity_0_brush_0_collision_shap
entity_0_brush_1_collision_shap
entity_0_brush_2_collision_shap
entity_0_brush_3_collision_shap
entity_0_brush_4_collision_shap
FileSystem
res://scenes/geometry/gate.tscn
gate
icon_slipgate.svg
v1m2
map_sandbox
player
hud
map2
map3
map5
map4
v1m1
v1m3
Transform
View
Perspective
Inspector
Node
History
HUD
Filter Properties
CanvasLayer
Layer
Transform
Follow Viewport
On
Node
Process
Physics Interpolation
Auto Translate
Editor Description
Script
hud.gd
Add Metadata


Wir müssen da jetzt mal was mit KI machen!


Oh boy...

00.

Agenda

01 KI-Systeme

02 Metriken

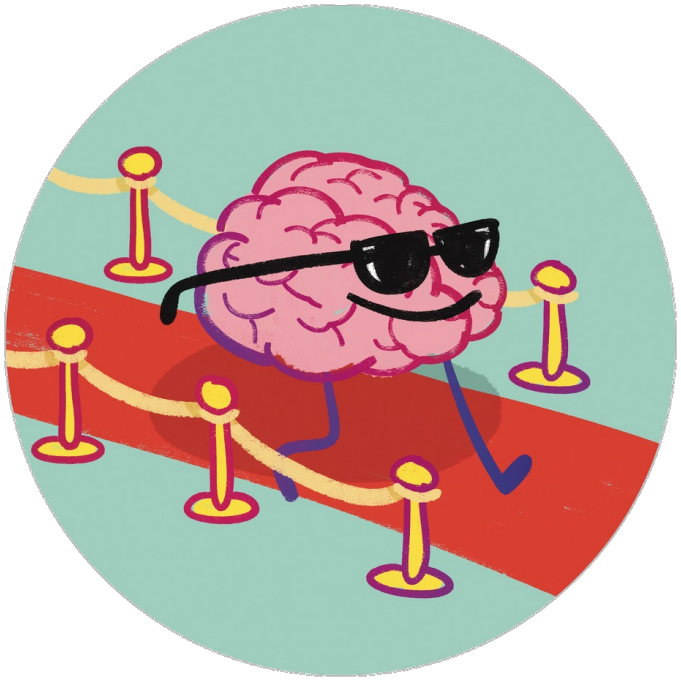
03 Qualität

04 Guardrails

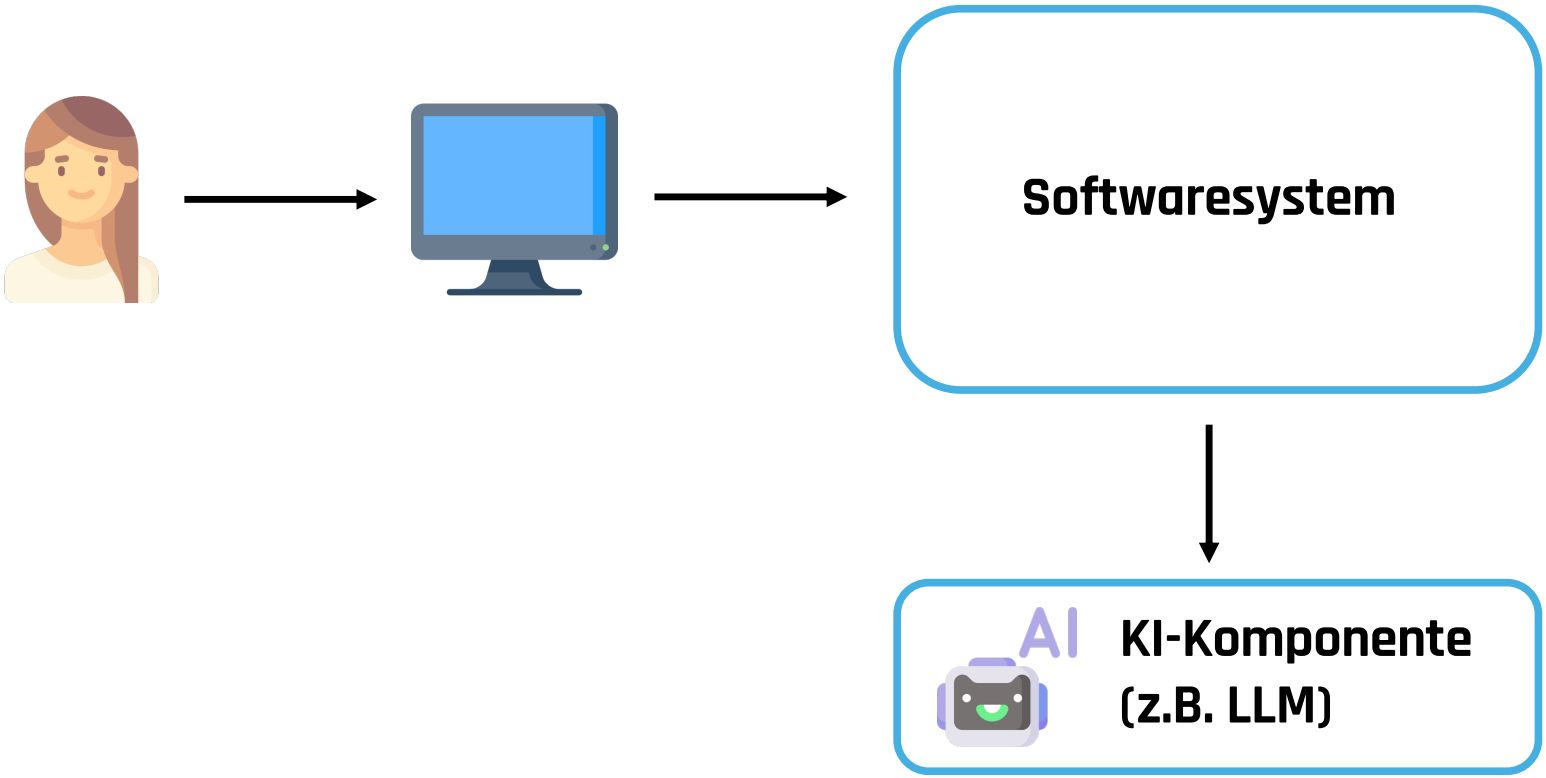
05 Evaluation

01.

KI-Systeme sind anders



KI-Systeme



Unterschiede zu klassischen Softwaresystemen

Induktiv vs. deduktiv

Daten vs. Code

Zufall vs. Determinismus

Blackbox vs. Whitebox

Präzise vs. datengetriebene Spezifikation

```python
def is_toxic(post: str) -> bool:
    if contains_insult_word(post):
        return True
    elif matches_pattern(post, "du bist [negativ]"):
        return True
    elif contains_silencing_or_exclusion_phrase(post):
        return True
    elif devalues_person_instead_of_argument(post):
        return True
    else:
        return False
```

```
Post: „Danke für die Erklärung, das hat mir geholfen.“
Label: nicht toxisch

Post: „Du bist echt zu dumm, um das zu verstehen.“
Label: toxisch

Post: „Ich sehe das anders, weil deine Argumente wichtige Punkte auslassen.“
Label: nicht toxisch

Post: „Halt einfach die Klappe, niemand will deinen Müll lesen.“
Label: toxisch

Post: „Dein Beitrag ist unklar formuliert, kannst du ihn präzisieren?“
Label: nicht toxisch

Post: „Was für ein kompletter Idiot schreibt so etwas?“
Label: toxisch
```

z.B. bei Few-Shot Prompting

Herausforderungen

Spezifikation nicht vollständig möglich

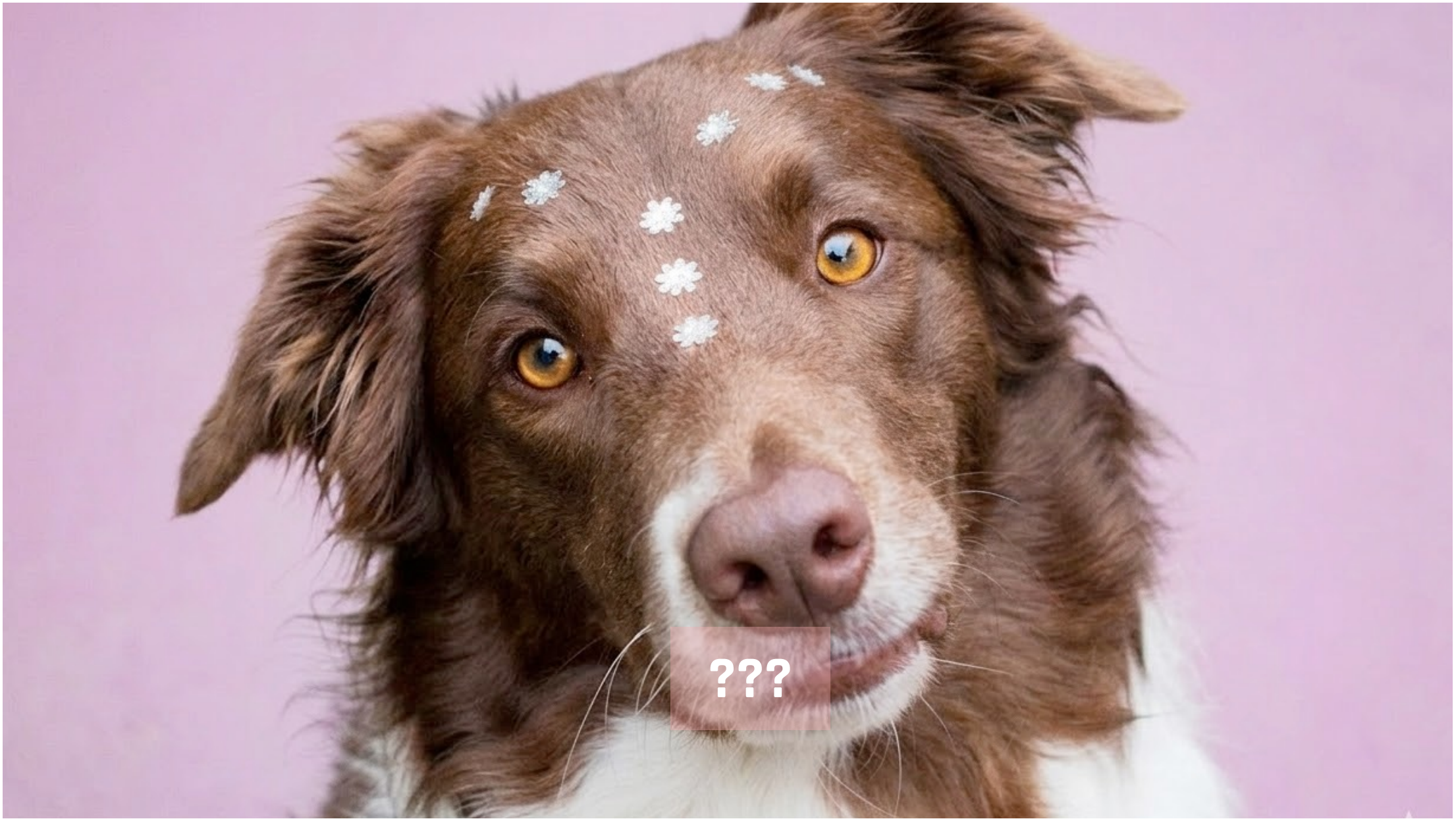
Neue Qualitätsmerkmale, enge Verknüpfung mit funktionalen Anforderungen

Komplexe Trade-offs

Zufall und Unsicherheit


Aha. und wie macht man solche Systeme dann korrekt?


That‘s the neat thing. You don‘t.




Deal with it!

02.

Metriken

embarc
Software Consulting GmbH

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

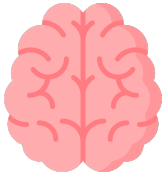
„Wann rechnet sich das?“

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

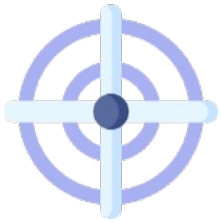
Kosten und Latenz

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

- Betrugsversuch benötigt 98%+ Konfidenz + lokale Konsistenz, **sonst Zahlung akzeptieren**
- 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 << X / 10.000

03.

Qualität

embarc
Software Consulting GmbH

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 unberechtigtem Zugriff geschützt? ...

Wartbarkeit
(Maintainability)

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

Übertragbarkeit
(Flexibility)

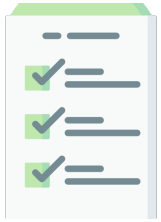
Ist die Software leicht auf andere Situationen oder Zielumgebungen (z.B. anderes OS) übertragbar? ...

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.

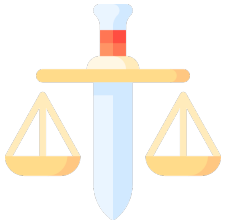
An AI model with high interpretability allows us to understand how the components of the AI model (nodes and weights in deep neural network models) produce a mapping between a system input and output“

https://www.splunk.com/en_us/blog/learn/explainability-vs-interpretability.html

Beispiel: Kreditkartenbetrug

Erklärbarkeit

Ablehnung von Transaktionen muss für rechtliche Prüfungen begründet werden können. **Dazu müssen folgende Informationen geliefert und nachvollzogen werden können: …**



Fairness

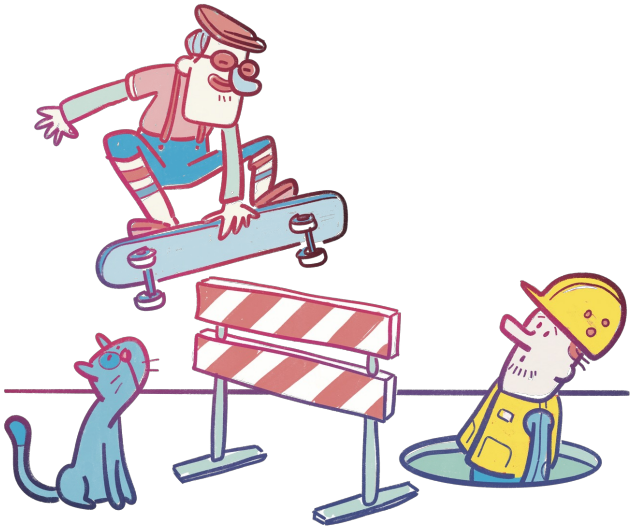
Keine Ungleichbehandlung, etwa von ethnischen Gruppen, **insbesondere keine Berücksichtigung von Geschlecht, Herkunft, Alter, …**

Reproduzierbarkeit

Wiederholung einer Prüfung mit gleichen Daten gelangt zum selben Ergebnis

04.

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

Input Gate — Latenz, Ablehnung legaler Anfragen

Context Gate — Performance, fehlende Informationen

Tool Gate — Starre Regeln limitieren Problemlösungsfähigkeit, HITL

Output Gate — Latenz, Überprüfung bei Streaming schwierig, mangelnder Nutzen für User

Runtime Gate — Aufwändiges Tooling, Abbruch sinnvoller Agentic Loops

Beispiel: Kreditkartenbetrug

Runtime Guardrails

Timeout: x ms, danach Transaktion durchlassen und Nachprüfung veranlassen

Input Guardrails

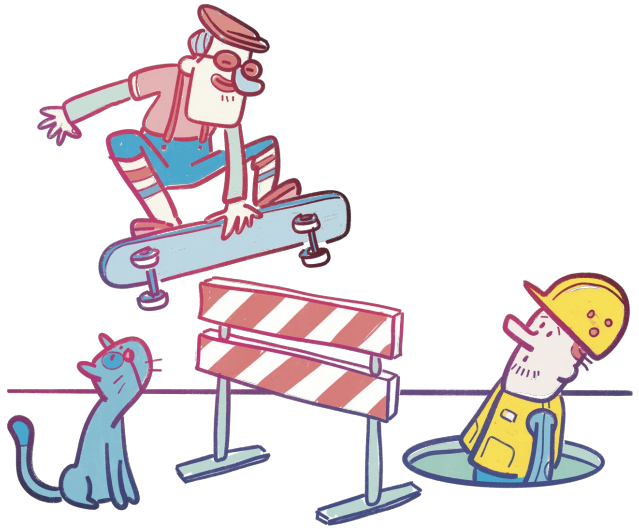
Entfernung geschützter Merkmale (Alter, Geschlecht, Herkunft)

Output Guardrails

Bei Ablehnung: Begründung, Quellenverweise, Konfidenz-Score

05.

Evaluation

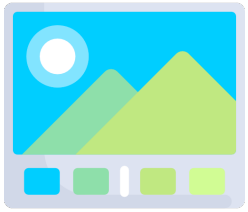


embarc
Software Consulting GmbH

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

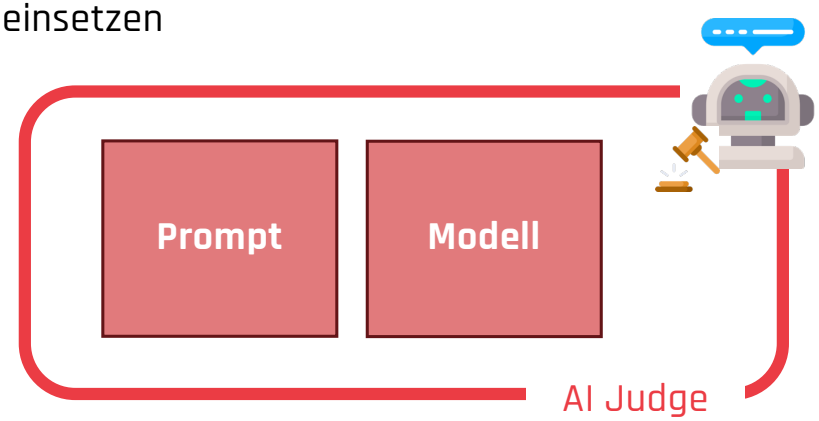
- Einfache Evaluation bei gut überprüfbaren Ergebnissen
- Reproduzierbarkeit
- Automatisierbar, Hohe Skalierbarkeit

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

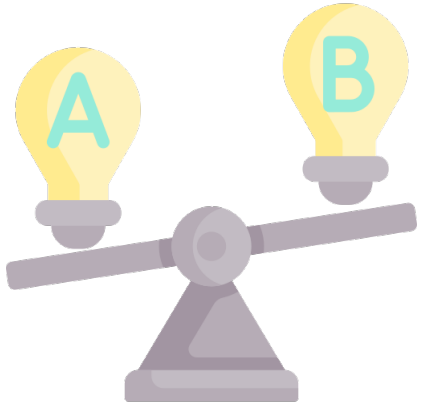


- Für jeden Use-Case und jedes Output-Format einsetzbar
- Breite Bewertungskriterien
- Skalierbar, automatisierbar

- 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%**)
- Ground Truth wird wie folgt ermittelt: ...
- Auf folgenden Datensätzen darf (nicht) evaluiert werden: ...
- Überprüfung auf Golden Set mit 3% Betrugsversuchen, 90% legitime Transaktionen, 7% Grenzfälle: ... (**Precision > 95%**)

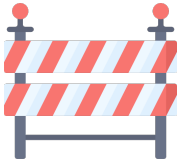
KI-Anforderungen auf einen Blick

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?

Vielen Dank!

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

fk@embarc.de

linkedin.com/in/felix-kammerlander/

fkammerlander.io