Forschungsidee:
Closed-Loop-Validierung des EU AI Act im Requirements Engineering – Kurzbeschreibung

Warum jetzt der richtige Zeitpunkt ist
Die EU-Regeln für künstliche Intelligenz (EU AI Act) werden in den nächsten Jahren schrittweise umgesetzt, parallel entstehen Normen und Richtlinien. Teams, die für die Erfüllung dieser Vorgaben verantwortlich sind, müssen regulatorische Pflichten in konkrete, testbare Anforderungen überführen, diese kontinuierlich validieren und die Einhaltung nachvollziehbar dokumentieren.

In der Praxis fehlt dafür häufig ein durchgängiger Validierungsansatz. Verfügbare Daten sind begrenzt, Messgrößen sind nicht standardisiert, und die Überprüfung erfolgt überwiegend manuell, verteilt und zeitaufwändig. Das erhöht das Risiko für Lücken bei Abdeckung, Konsistenz und Traceability.

Hypothese und Vision
Die Hypothese ist, dass ein agentenbasierter Ansatz auf Basis von Foundation Models, Conversational Agents und Human-in-the-Loop (HITL) die Erfüllung der EU-AI-Act-Pflichten im Requirements Engineering wirksam unterstützen kann.

Die Vision ist ein transparenter, auditierbarer Closed-Loop, der:
- regulatorische Risikoanforderungen (HLR) systematisch in testbare Projektanforderungen (LLR) überführt,
- Abdeckung, Konsistenz und Traceability zwischen HLR und LLR bewertet,
- und den manuellen Prüfaufwand reduziert, ohne die fachliche und regulatorische Kontrolle aus der Hand zu geben.

Die Idee (kurz und konkret)
Kern des Ansatzes ist eine kuratierte EU-AI-Act-Wissensbasis, die in HLR-Checklisten und Prüfkriterien überführt und an ISO/IEC/IEEE 29148 ausgerichtet ist.

Auf dieser Basis:
- analysiert ein **Validierungsagent** verschiedene RE-Artefakte (z. B. Spezifikationen, User Stories, RTMs),
- prüft Abdeckung und Konsistenz zwischen HLR und LLR,
- und belegt seine Befunde mit Quellen (Citations) aus den zugrunde liegenden Dokumenten.

Zunächst werden Risikoanforderungen (HLR) klassifiziert. Diese Risikoanforderungen werden gegen Projektanforderungen (LLR) gemappt und validiert, um belegbare, testbare Zuordnungen mit hoher Abdeckung zu erhalten.

Conversational Agents klären Unschärfen im Dialog mit den Beteiligten und sammeln zusätzliche Evidenz (z. B. Kontext, Annahmen, Entscheidungen).

Der **HITL-Review** priorisiert Befunde, setzt Freigaben bzw. Policy-Gates und führt ein Audit-Log. Feedback aus den Reviews fließt in Wissensbasis, Prompts und Regeln zurück – so entsteht ein lernender, kontinuierlicher **Closed-Loop**.

Prozessfluss (vereinfacht)

Der angestrebte Prozessfluss lässt sich wie folgt skizzieren:

1. Anforderungs-Mapping (HLR ↔ LLR)
2. Query Router (Auswahl der passenden Agenten/Checks)
3. Evaluation Hub
    - Scorecards: Quelltreue, Vollständigkeit, Präzision/Testbarkeit
4. Scorecard-Aggregation
5. Kritiker/Scorer (Prüfung von Schwellwerten)
6. Entscheidungsgate
    - ggf. Human Review (HITL) bei Unsicherheit oder kritischen Findings
7. Reporter-LLM
    - Erstellung der Artefakte (Mapping-Liste, Scorecards, Audit-Log, Bericht)

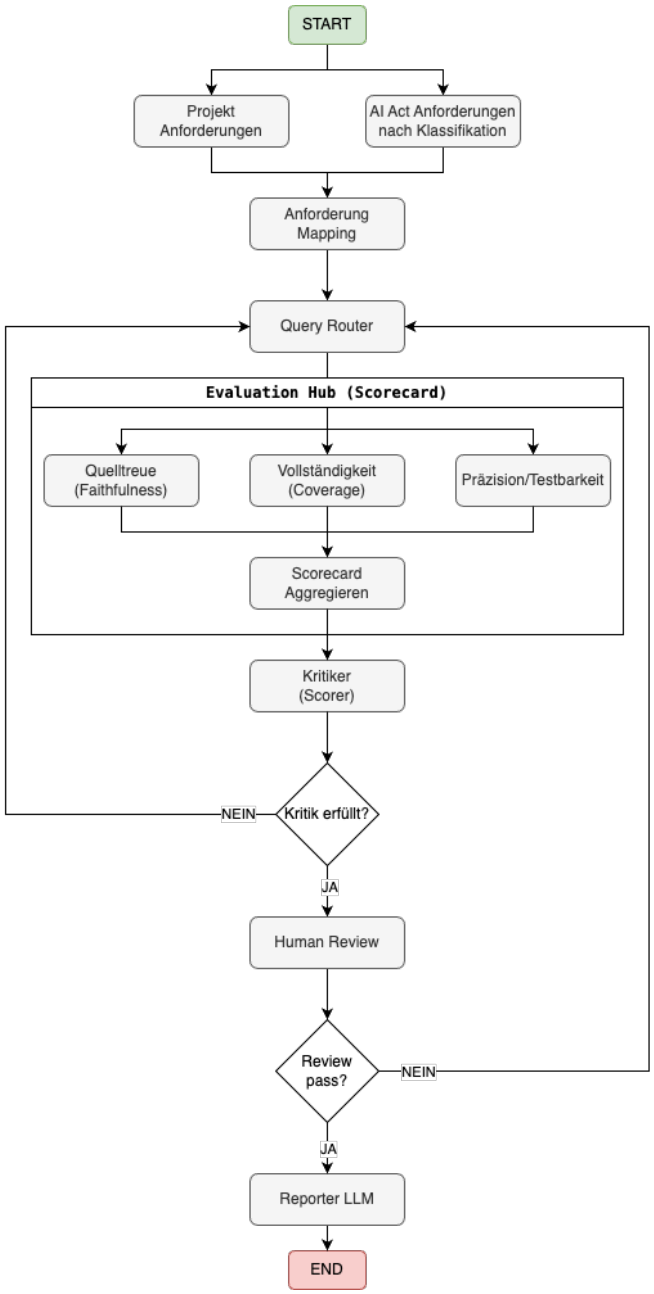

Abbildung 1: Übersicht des Closed-Loop-Prozessflusses (Anforderungs-Mapping → Query Router → Evaluation Hub → … → Reporter-LLM).

Ground-Truth / Goldset für die Evaluation

Zur Bewertung des Closed-Loops wird ein Ground-Truth-Datensatz aufgebaut. Dieser besteht aus zwei Bausteinen:

1. einem Goldset für HLR-↔-LLR-Mappings
2. annotierten RE-Artefakten (inkl. Leistungsbeschreibungen, sofern vorhanden)

1. Goldset für HLR↔LLR-Mappings

Manuell geprüfte Zuordnungen zwischen HLR und LLR werden mit einem Relations Label versehen (z. B. „korrekt", „teilweise korrekt", „nicht zutreffend"). Die Agenten-Ergebnisse werden mit diesem Goldset verglichen (z. B. bezüglich Coverage, Präzision und Zitat-Treue).

Beispiel (vereinfacht):

HLR-ID	LLR-ID	Relation-Label	Kommentar Reviewer	Quelle / Artefakt
HLR-9.1	LLR-9.1a	korrekt	Deckt alle geforderten Risiko-Felder ab	RiskRegister_v1.3
HLR-9.1	LLR-9.1b	nicht zutreffend	Betrifft nur interne Reporting-Logik	Reporting-Konzept v2
HLR-9.2	LLR-9.2a	teilweise korrekt	Lifecycle-Aspekt nur teilweise adressiert	CI-Konfiguration, Policy Gate P2

2. Ground-Truth auf Artefakt-Ebene (Annotationen)

Zusätzlich werden ausgewählte RE-Artefakte (z. B. Spezifikationen, Leistungsbeschreibungen, User Stories) abschnittsweise annotiert. Ziel ist zu bewerten, ob die Agenten:

- die relevanten Dokumentbereiche korrekt identifizieren,
- die richtigen Abschnitte zitieren,
- und regulatorisch relevante Inhalte zuverlässig erkennen.

Diese Annotationen sind Teil des Ground-Truth und ermöglichen:

- eine Bewertung der Retrieval- und Zitationsqualität der Agenten,
- reproduzierbare Benchmarks für unterschiedliche Modell- und Prompt-Varianten.

Erwartete Ausgaben

- **Mapping-Liste:** HLR mit LLR-Subliste und belastbaren Zitationen.
- **Scorecards:** pro Pflicht und aggregiert (Coverage, Zitat-Treue, semantische Ähnlichkeit, Testbarkeit).
- **Review-Protokoll:** HITL-Entscheidungen und Freigaben als Audit-Log.
- **Bericht / Trace:** konsolidierter Validierungsbericht mit Traceability-Verweisen.
- **Zusätzlich:** Grundlage für synthetische bzw. augmentierte Testdaten und Audits.

Nutzen

- Transparente, nachvollziehbare Zuordnung von Risikoanforderungen zu Projektanforderungen – mit Citations.
- Höhere Abdeckung kritischer Risiken bei reduziertem manuellen Prüfaufwand.
- Auditfähige Dokumentation und belastbare Basis für PMM/Audits und die Augmentation von Testdaten.

Kontakt & Teilnahme

Fragebogen:

- Dauer: ca. 20–30 Minuten
- Teilnahme: freiwillig, anonym
- Thema: Closed-Loop-Validierung des EU AI Act im Requirements Engineering (RE)

Kontakt: emmy.lai@iais.fraunhofer.de

Glossar

- **High-Level Requirements (HLR):** regulatorische Pflichten bzw. Schutzziele auf höherer Abstraktionsebene.
- **Low-Level Requirements (LLR):** konkrete, testbare Projektanforderungen mit klaren Akzeptanzkriterien.
- **Traceability:** bidirektionale Nachvollziehbarkeit zwischen HLR und LLR inkl. Audit-Log.
- **Human-in-the-Loop (HITL):** qualitätssichernde menschliche Reviews und Freigaben.
- **Foundation Models (FM):** vortrainierte KI-Grundmodelle.
- **Large Language Models (LLM):** große Sprachmodelle für Textaufgaben.
- **Requirements Traceability Matrix (RTM):** Zuordnungstabelle (z. B. HLR ↔ LLR).
- **Ground-Truth / Goldset:** manuell geprüfter Referenzdatensatz (z. B. HLR-↔-LLR-Mappings und annotierte RE-Artefakte), der zur Bewertung und zum Vergleich der Agenten-Ergebnisse dient.

Beispielausgabe (vereinfacht)

HLR-ID	AI-Act Bezug	Zitat (sinngemäß)	LLR-ID	LLR (testbar)	Akzeptanzkriterien	Evidenz	Traceability
HLR-9.1	Art. 9(1)	Anbieter richten ein dokumentiertes Risikomanagementsystem ein und pflegen es.	LLR-9.1a	Risikoregister in Tool X mit Feldern: Hazard, Harm, Severity, Likelihood, Residual Risk, Mitigation, Owner.	100% Hazards haben vollständige Felder; wöchentlicher Auto-Report ohne Lücken.	Link: RiskRegister v1.3; CI-Report #224	HLR-9.1 ↔ LLR-9.1a (1:n)
HLR-9.2	Art. 9(2)	Kontinuierlich und iterativ über den gesamten Lebenszyklus.	LLR-9.2a	CI-Pipeline triggert Risk-Review bei jeder Modell-/Prompthub-Änderung.	≥95% relevanter Commits erzeugen Risk-Review-Job; Job-Log vorhanden.	CI Logs run-8742; Policy Gate P2	HLR-9.2 ↔ LLR-9.2a