

University of Stuttgart
Institute of Industrial Automation
and Software Engineering



Automating Safety and Risk Management with Large Language Models Agents

Belal Abulabn
Supervisor: Yuchen Xia
Study Program: Elektromobilität



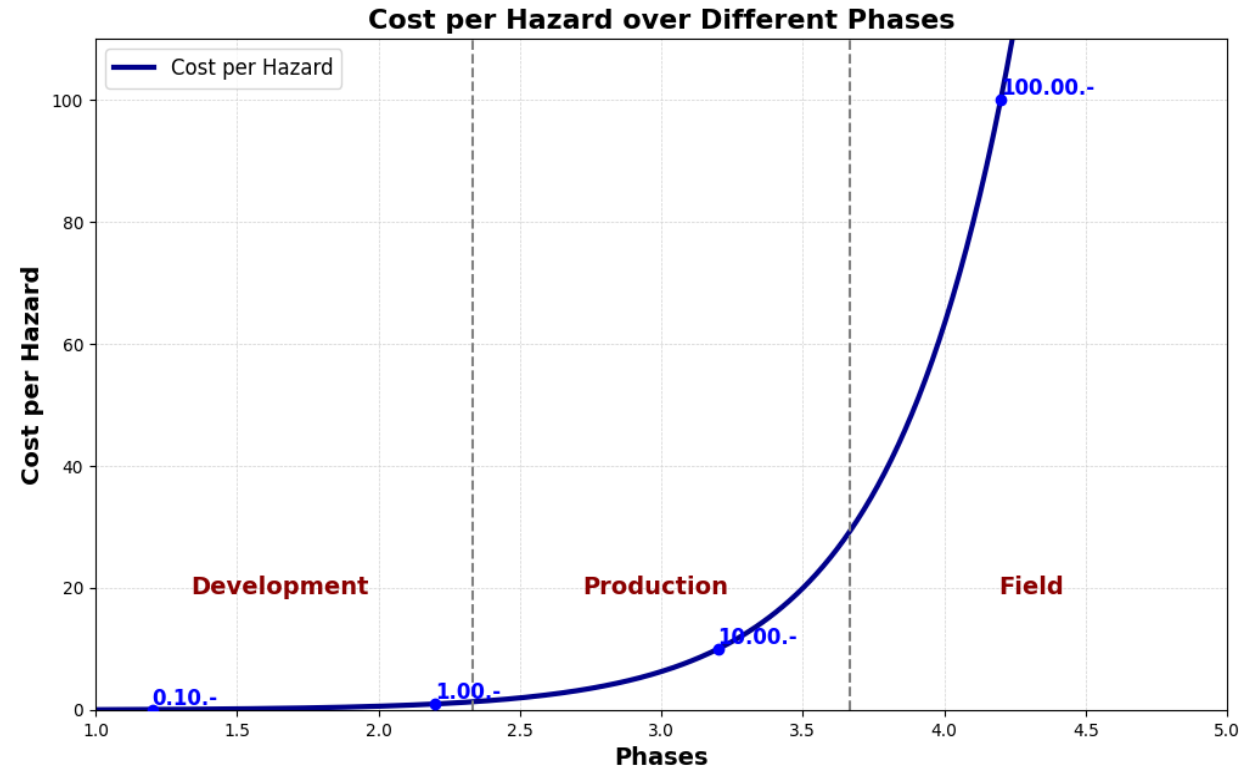
Agenda

- Einführung
- Forschungsfragen und Ziele
- Konzept Entwurf
- Implementierung
- Auswertung und Ergebnisanalyse
- Fazit

Motivation

Risikoeinschätzung

- Je früher die **Risikoeinschätzung** erfolgt, desto **kosteneffizienter** ist es im Produktlebenszyklus.
- Methods: **Failure Modes and Effects Analysis (FMEA)**, **Fault Tree Analysis (FTA)**
- Standards wie **ISO 14971** (Risikobewertung im Produktlebenszyklus), **ISO 15026** (System- und Softwarelebenszyklus-Prozesse – Risikomanagement) und **ISO 31000** (Risikomanagement).



Hintergrund

LLM

Problem

Manuell, arbeitsintensiv, könnten kritische Probleme übersehen und erfordern Fachkräfte in unterschiedlichen Bereichen.

- Warum LLM nutzen?
 - Internalisiertes menschliches Wissen
 - Verschiedene Domänen
 - Automatisierte Informationsverarbeitung
 - Anpassungsfähigkeit
 - Skalierbarkeit durch Agent-design

Forschungsfrage

Wie können LLMs für Risikomanagementprozesse eingesetzt werden?

Forschungsfragen und -schritte

Einsatz von LLMs zur Unterstützung von Ingenieuren im Risikomanagement.

**Befolgung von ISO-Normen:
ISO 14971, ISO 15026, ISO 31000**

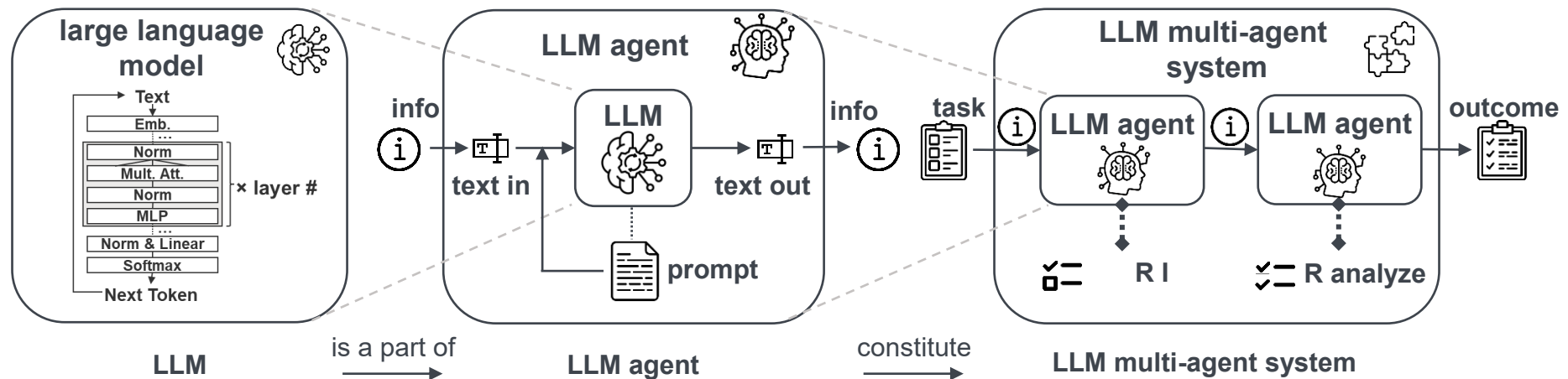
**Entwicklung eines automatisierten
Workflows**

Modelle bewerten und Fein tuning

Grundlagen LLM-Agent

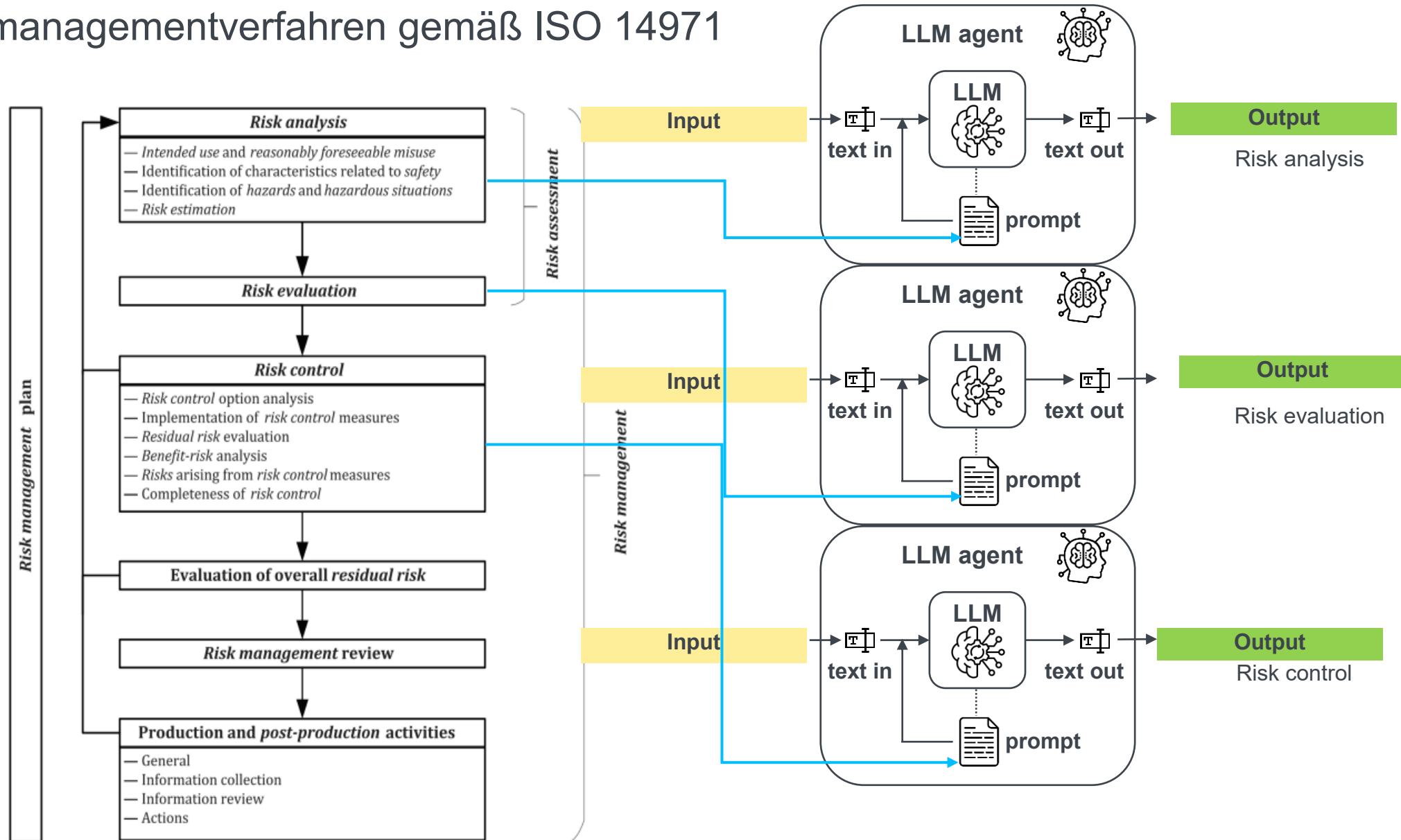
Allgemeine Methodik für den Entwurf von LLM-Systemen

Vom LLM-Modell zum LLM-Agentensystem



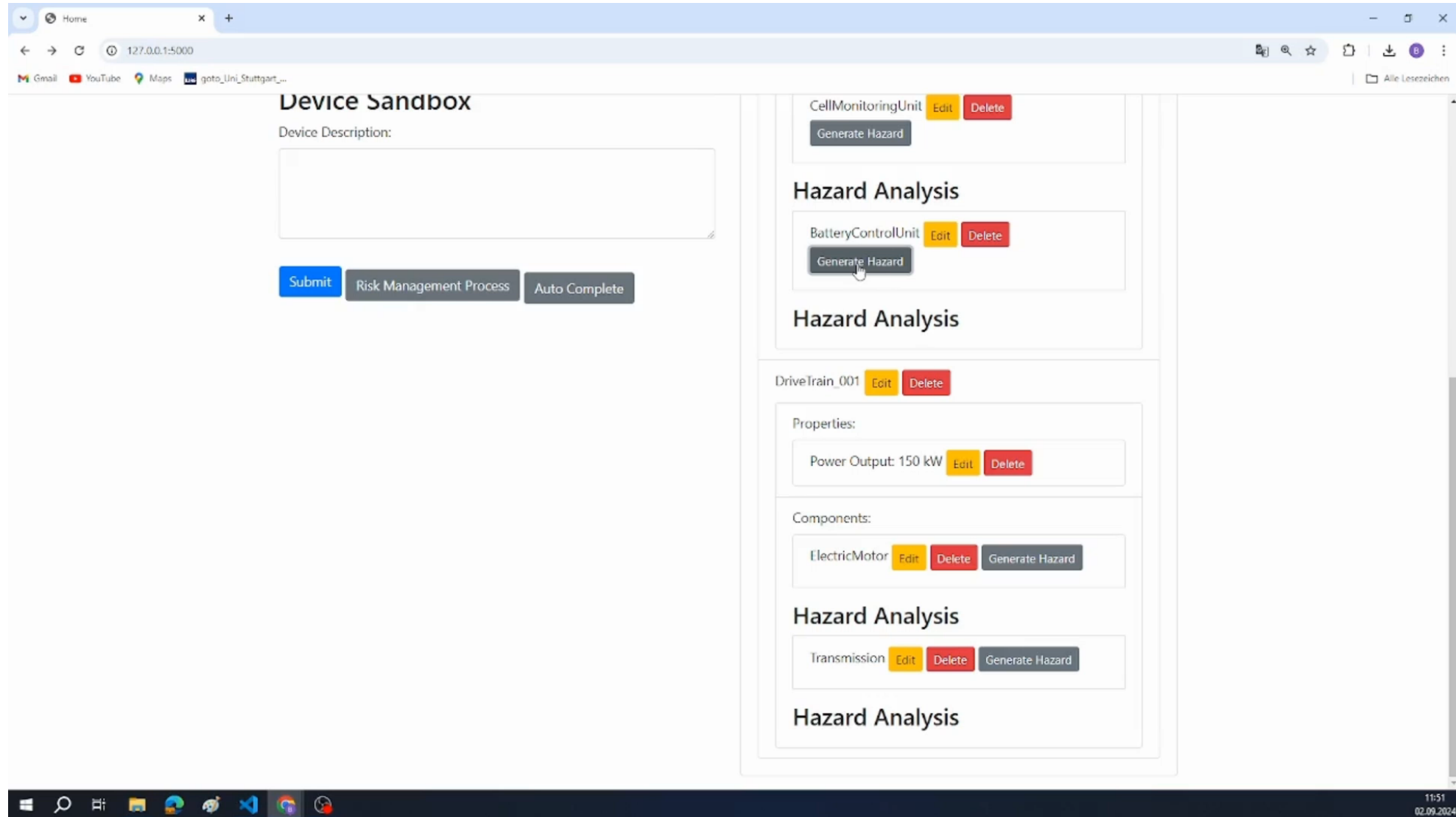
Verfahren der Risikoanalyse

Risikomanagementverfahren gemäß ISO 14971



Ergebnis-Vorschau

Prozess Ablauf



Autocomplete-Textfunktion

The screenshot shows a web browser window with the title 'Risk Management Process' and the URL '127.0.0.1:5000/risk_management_process'. The browser's address bar and tabs are visible at the top. The main content area displays a list of text blocks, each with a bolded header and a descriptive sentence. The blocks are:

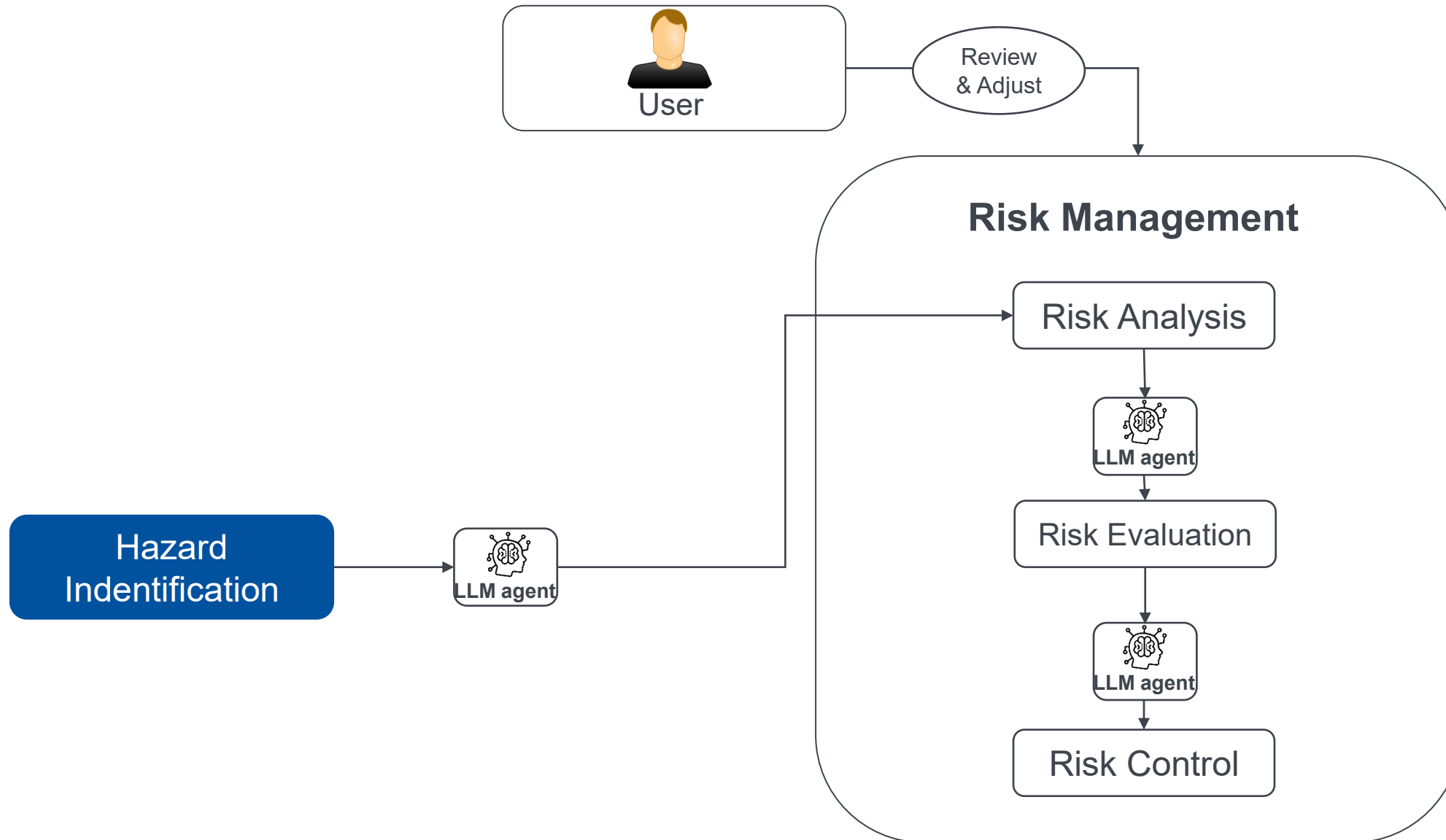
- Measure:** Error handling protocols to mitigate incorrect commands.
Practicability: Practical and necessary but requires robust protocol development and testing.
- Effectiveness of Measures:** The control measures are effective in reducing the risk levels.
- Performance Verification:** Measures are performing as expected and meeting intended safety objectives.
- Residual Risk Acceptable:** true
- New Hazards or Risks:** No new hazards or risks introduced as a result of the control measures.
- Risk Assessment Update:** Risk assessment updated based on the effectiveness of implemented measures.
- Advantages:** Control measures reduce risk levels and enhance safety by preventing failures and maintaining system integrity.
- Drawbacks:** Some measures require regular maintenance and investment in additional infrastructure, which may incur additional costs.

At the bottom of the content area, there are three buttons: 'Text Edit' (yellow), 'Save Risk Control' (green), and 'Generate Residual Risk' (blue). The Windows taskbar is visible at the bottom of the screen, showing the time as 12:41 on 02.09.2024.

System Design

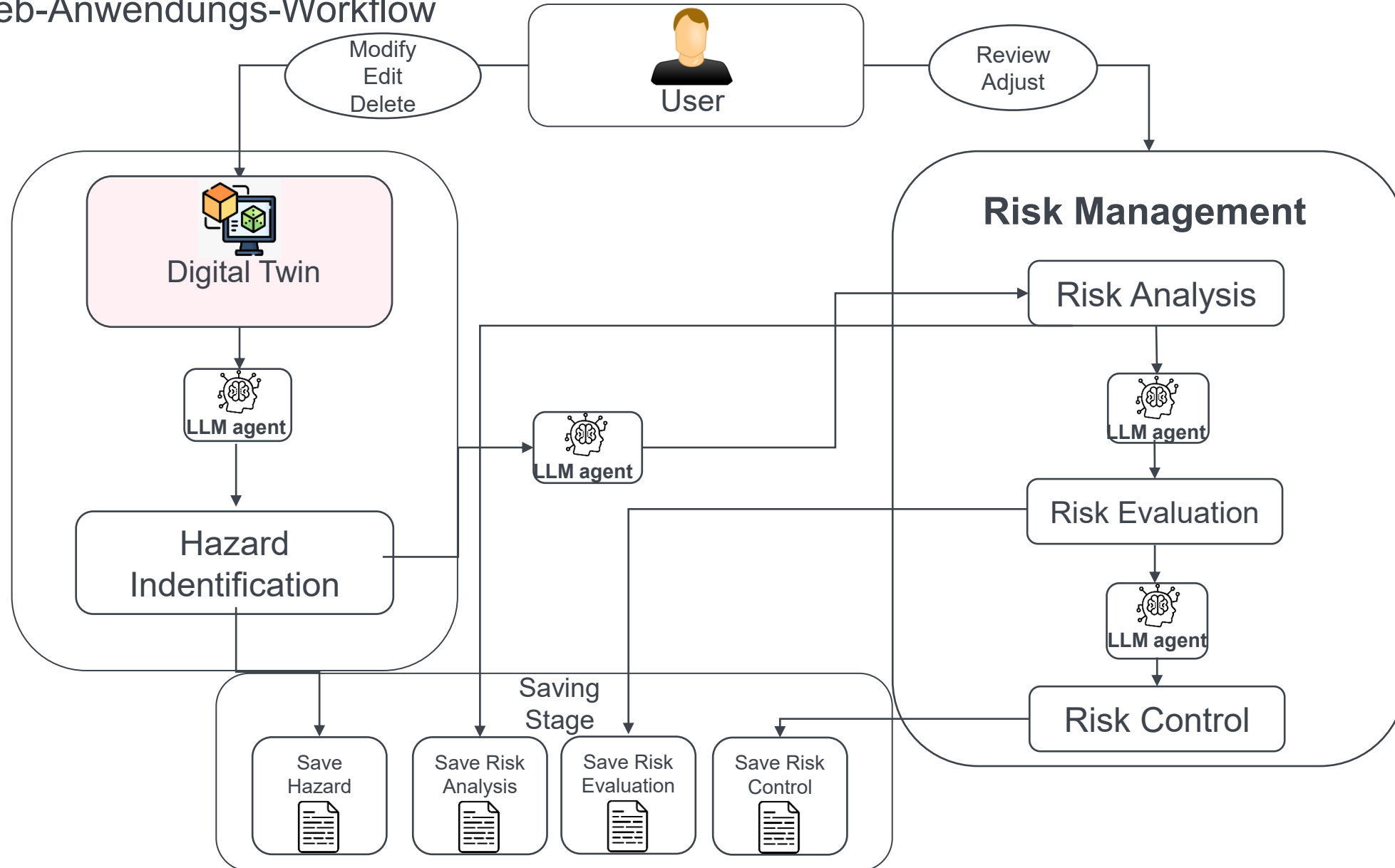
System Design

Web-Anwendungs-Workflow



System Design

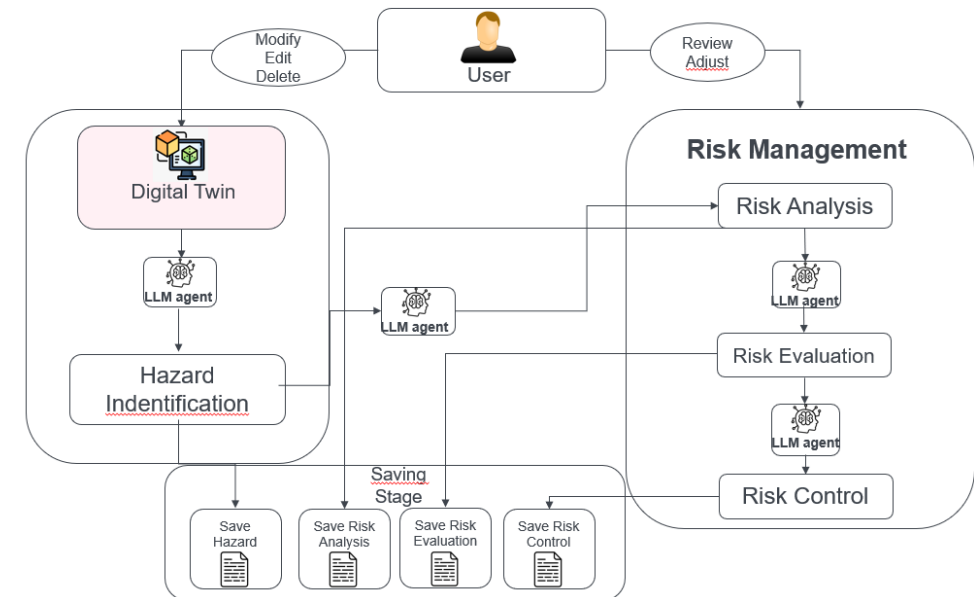
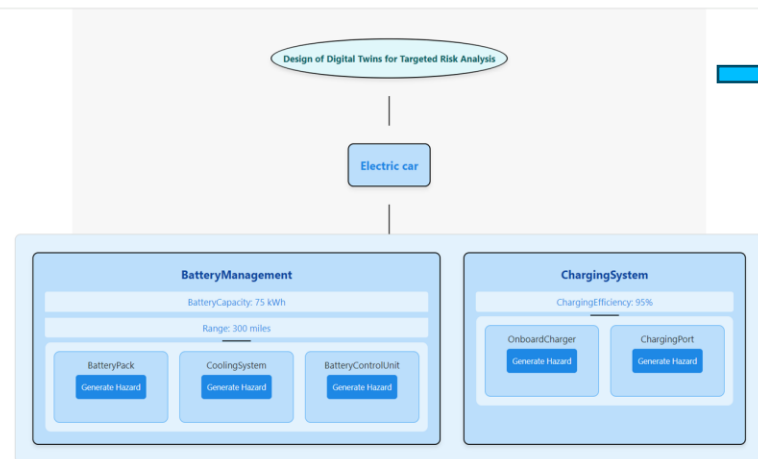
Web-Anwendungs-Workflow



Integriere das Informationsmodell in den digitalen Zwilling

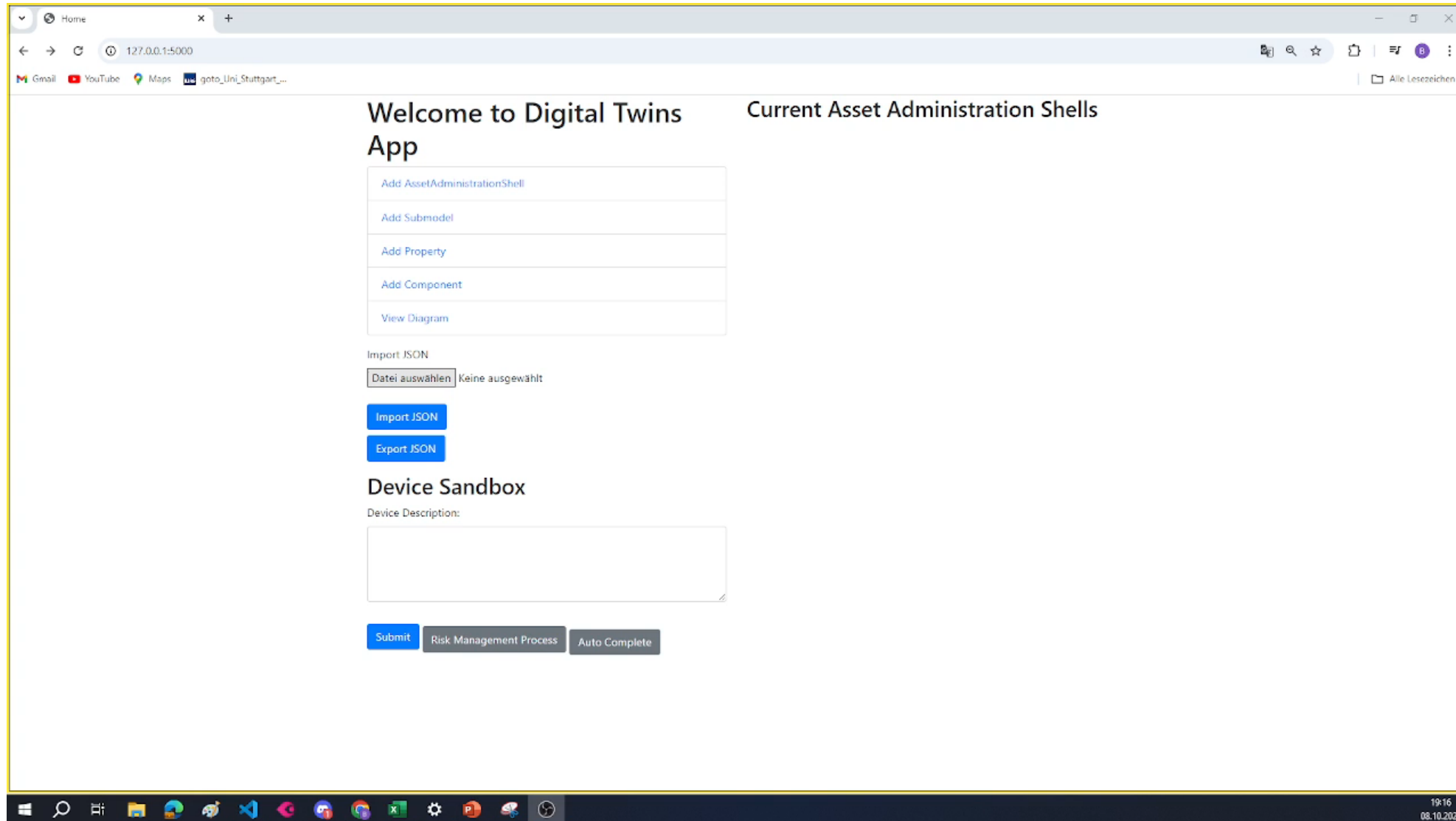
Wir benötigen einen digitalen Zwilling des Analyseobjekts, um die notwendigen Daten für die Risikobewertung bereitzustellen.

Asset Administration Shell (AAS)



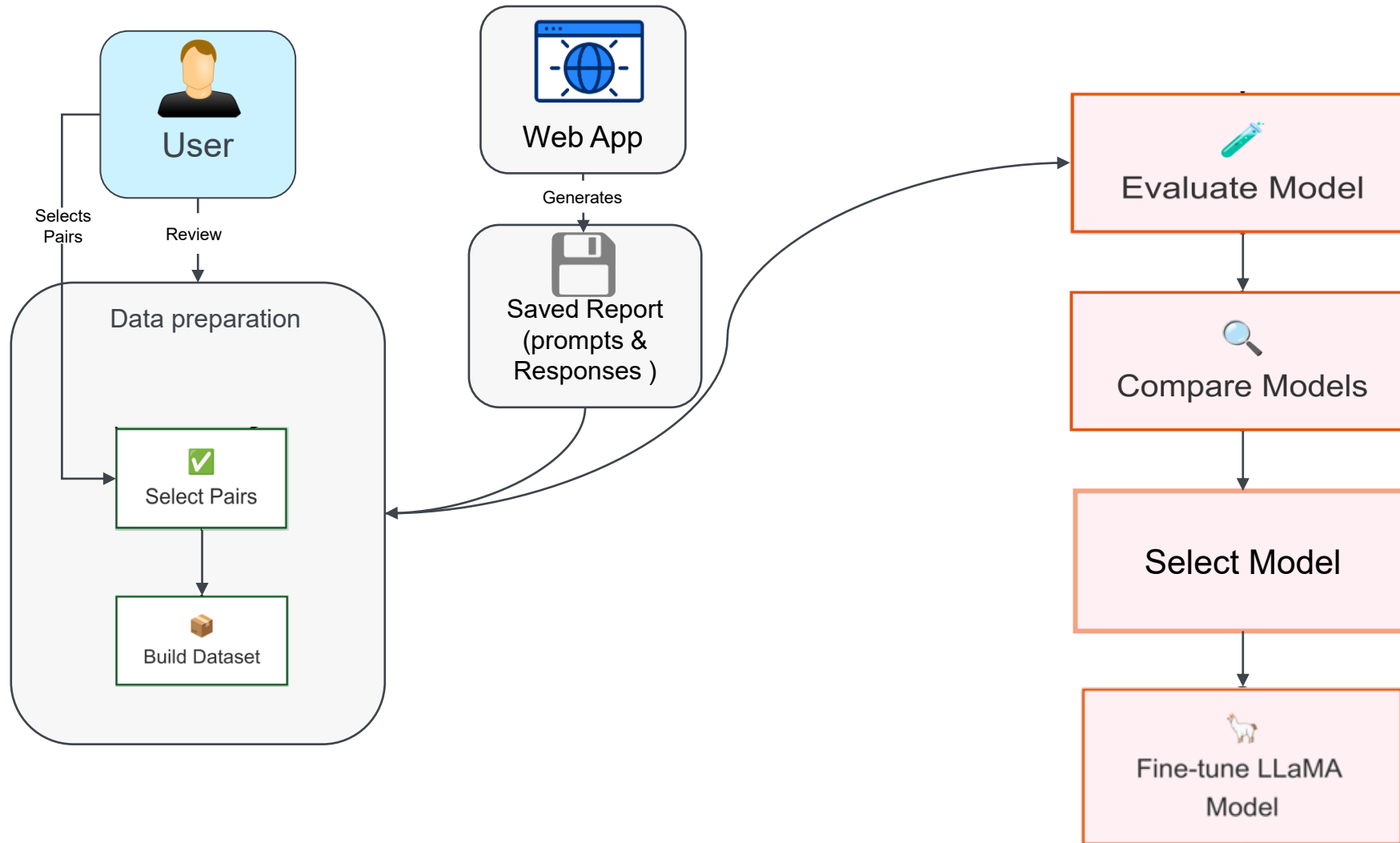
- **AAS (Asset Administration Shell)**
- **Informationsmodell**
- **Vorteile:**
Ermöglicht Datenanalyse, Automatisierung und Interoperabilität.

Informationsmodell



- **Testing und Evaluierung von Modelle**
- **Training**

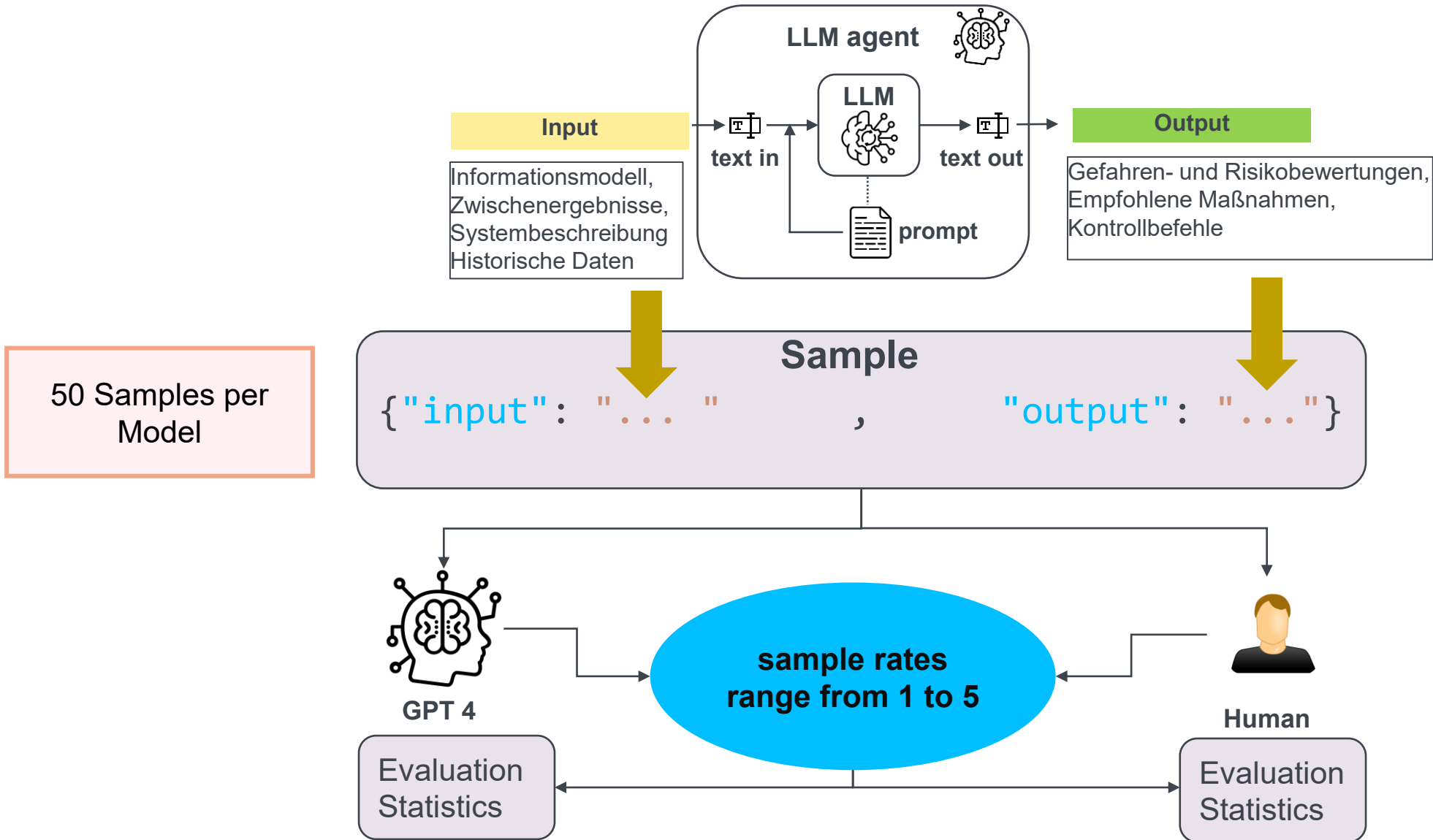
Gesamtkonzept Design



Evaluierungskonzept für jedes Modell

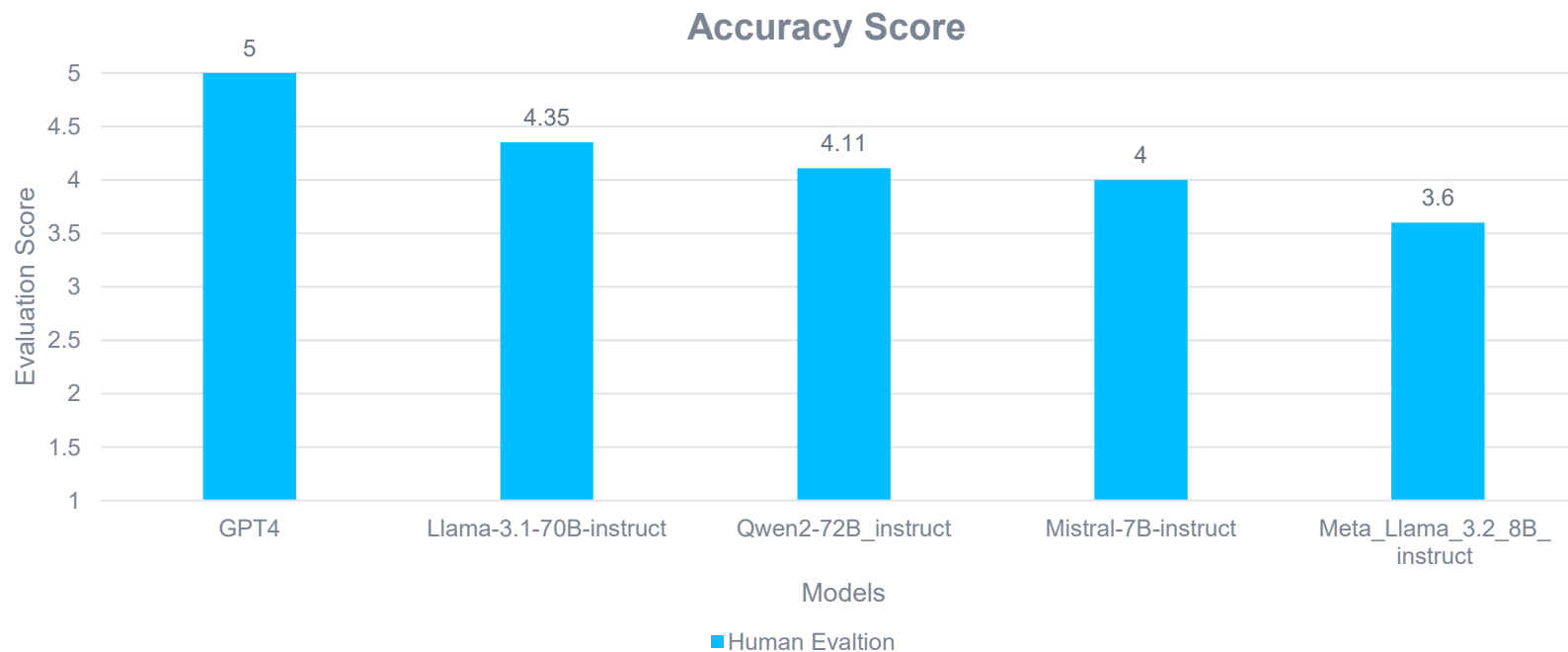


Evaluate Model



Evaluierung verschiedener Modelle für Risikomanagementaufgaben

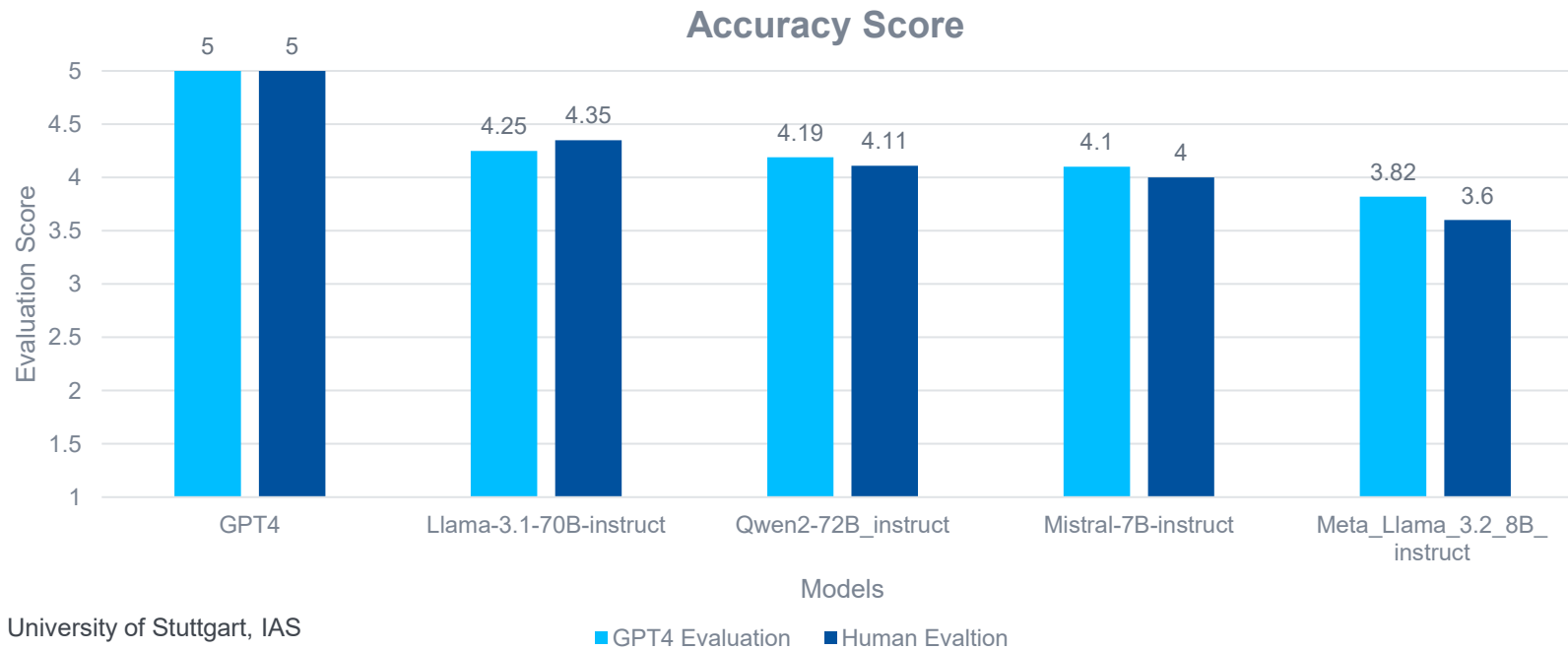
Model	Human Evaluation (%)
GPT-4	100.0%
Llama-3.1-70B-instruct	87.0%
Qwen2-72B_instruct	82.2%
Mistral-7B-instruct	80.0%
Meta_Llama_3.2_8B_instruct	72.0%



**Nachteil
Zeitaufwand
Fachwissen**

Evaluierung verschiedener Modelle für Risikomanagementaufgaben

Model	Human Evaluation (%)	GPT-4 Evaluation (%)
GPT-4	100.0%	100.0%
Llama-3.1-70B-instruct	87.0%	85.0%
Qwen2-72B_instruct	82.2%	83.8%
Mistral-7B-instruct	80.0%	82.0%
Meta_Llama_3.2_8B_instruct	72.0%	76.4%

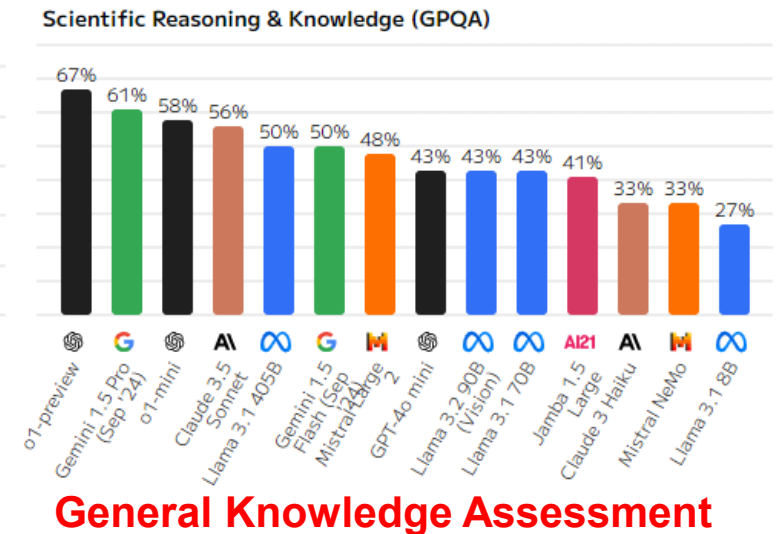
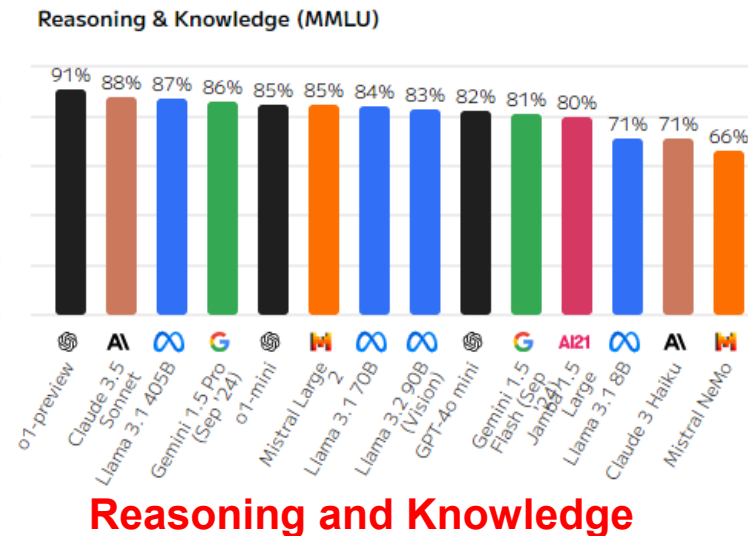
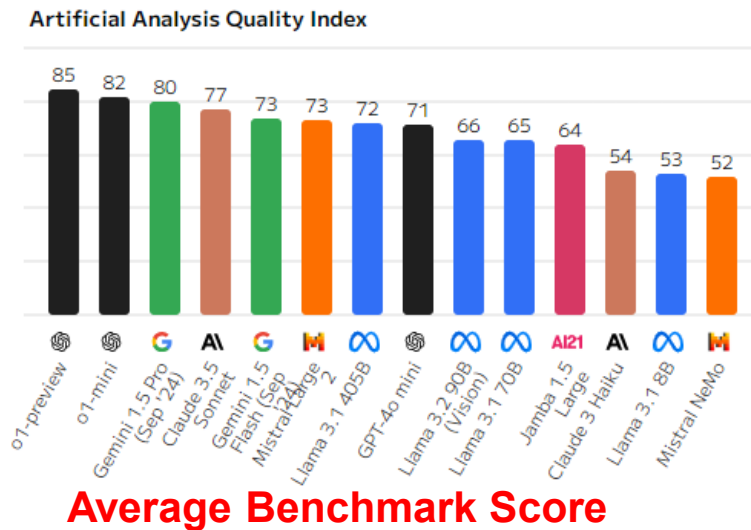


Übereinstimmung
Korrelationskoeffizient (r) = 0,97

Nachteil
Teure

Wie wählt man in Zukunft ein Modell für die Risikoanalyse aus?

- **Problem:** Es kostet Zeit und Geld, um ein Modell für die Risikoanalyse zu testen.
 - **Token-Kosten:** 200 Euro
 - **Zeitaufwand:** 5 Stunden für jede Modell
- **Vorschlag:** Korrelation zwischen der Leistung in der Risikoanalyse und allgemeinen Benchmarks finden.



Benchmark- und Korrelationsanalyse von LLMs für das Risikomanagement

Top 3 Pearson-Korrelationen

valuation Pair	Correlation Coefficient (r)	P-Value
Human vs. GPQA	0.948	0.014
Human vs. AAQI	0.850	0.068
Human vs. MMLU	0.802	0.102

- **Zweck dieser Benchmarks:**

- Messen die Fähigkeit, verschiedene Problemstellungen über **mehrere Domänen** hinweg zu lösen.
- Bewerten das **logische Denken** und die **Analysefähigkeiten** der Modelle.

- **Bedeutung der Fähigkeiten:**

- Entscheidend zur Identifikation von Modellen, die effizient Risiken analysieren und bewerten können.

- **Schlussfolgerung:**

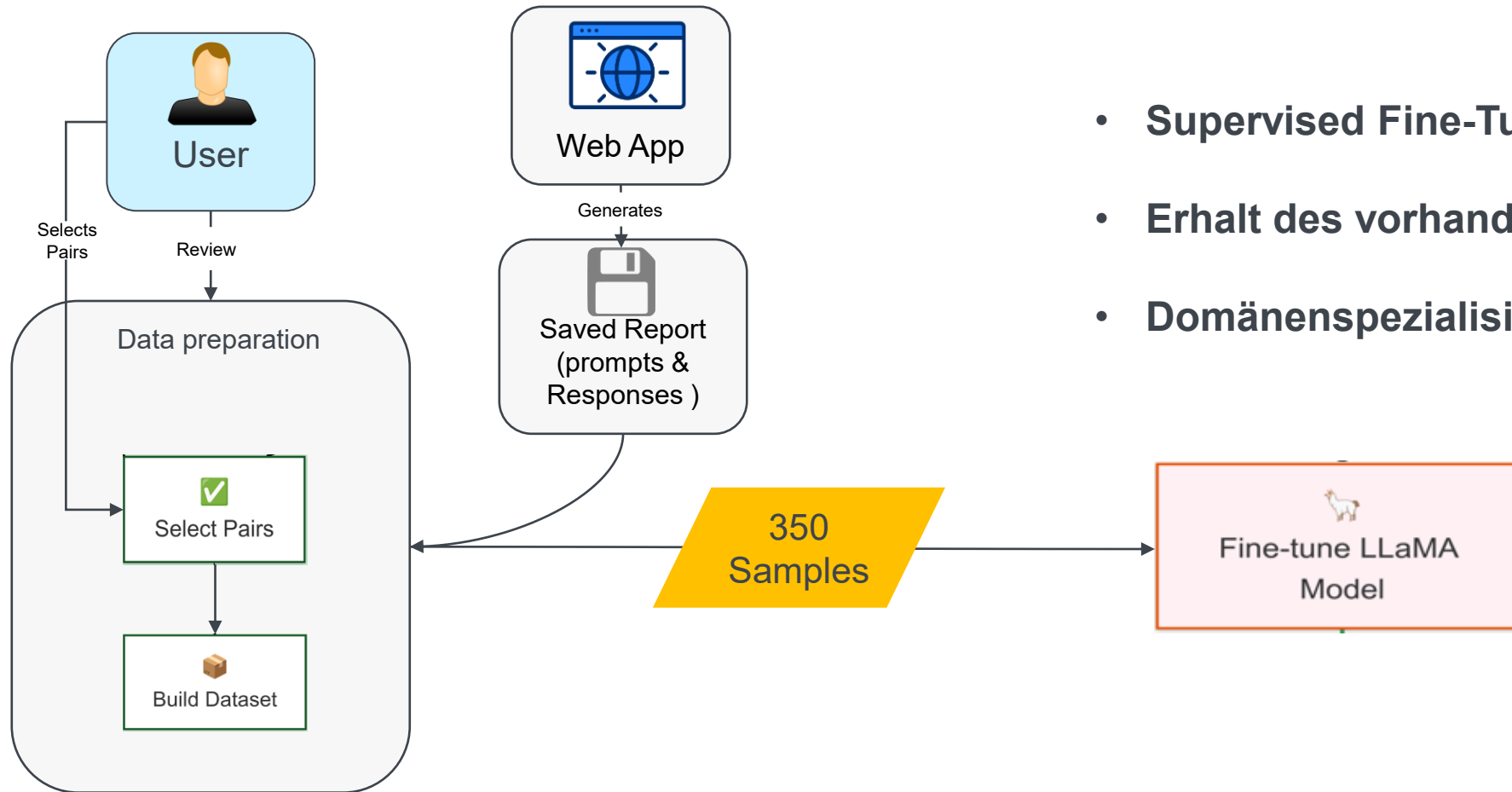
- Mit diesen Benchmarks können in Zukunft auch andere Modelle für die Risikoanalyse ausgewählt werden.

Erstellung von Daten

- Testing und Evaluierung
- **Training (mit ausgewählter GPT4-Generation)**

Overall Concept Design

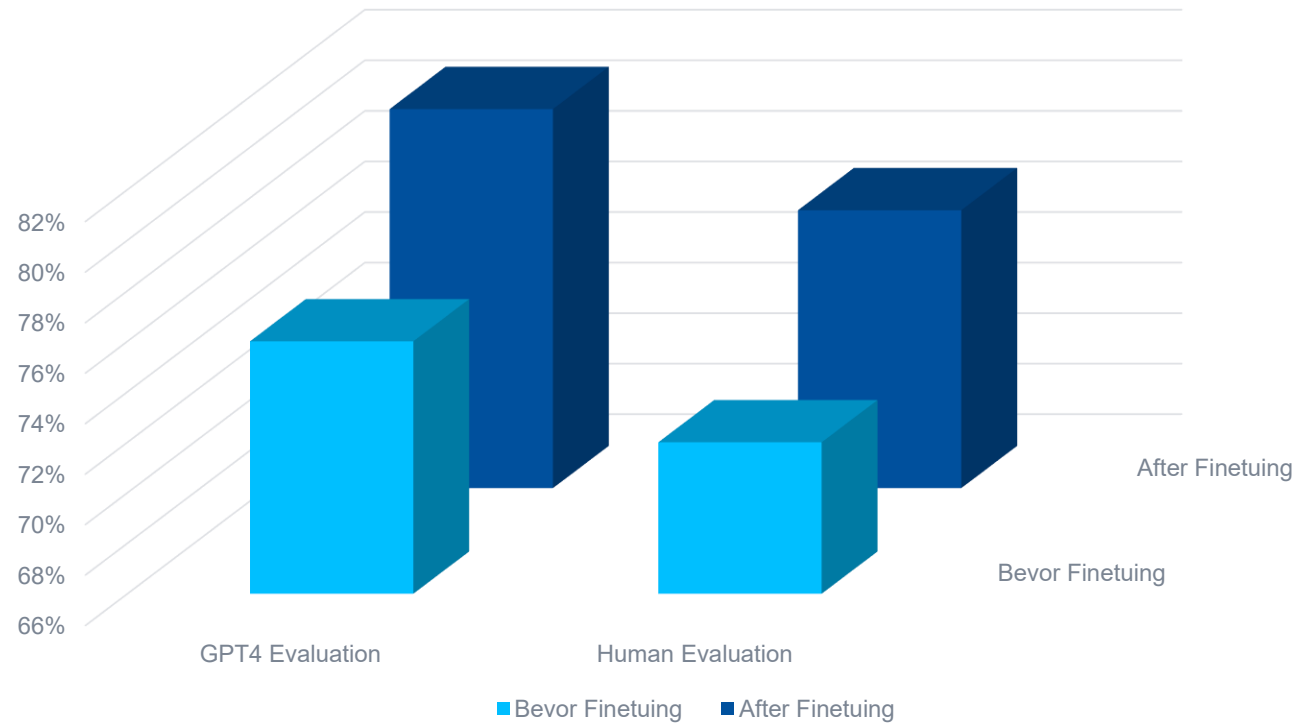
Fine tuning



- **Supervised Fine-Tuning mit LoRa**
- **Erhalt des vorhandenen Wissens**
- **Domänenspezialisierung**

Fine-Tuning the Llama 8B Model

Evaluation Metric	Before Fine-Tuning (%)	After Fine-Tuning (%)	Improvement (%)
GPT-4 Evaluation	76.4%	81.0%	+6.0%
Human Evaluation	72.0%	77.0%	+6.9%



Fazit

Fazit

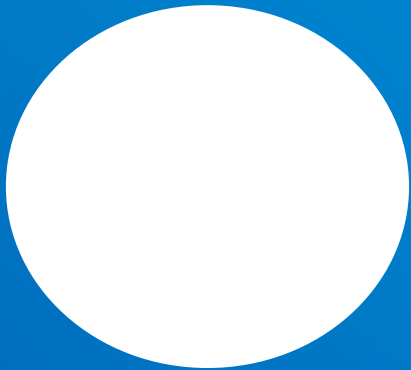
- **Effiziente Automatisierung durch LLM Agents**
- **Gezielte Modellauswahl mithilfe von Benchmarks**
- **Leistungssteigerung durch Feinabstimmung**
- **Kosteneffizienz durch angepasste Modelle**
 - **Ausblick auf zukünftige Verbesserungen**
 - **Erstellung qualitativ hochwertiger Datensätze**
 - **Vergleich RAG vs. Fine-Tuning**



University of Stuttgart

Institut of Industrial Automation
and Software Engineering

Thank you!



Belal Abulabn

e-mail st18211@stud.uni-stuttgart.de

phone +49 (0) 711 685-

fax +49 (0) 711 685-

University of Stuttgart

