Projektüberblick

Ziel des Projektes ist es, einen KI-Router zu entwickeln, d.h. ein automatisches System, das zu einer neuartigen Nutzeranfrage diejenige KI auswählt, die die Anfrage gemäß individueller und situationsabhängiger Kriterien am besten bearbeiten kann. Diese Vorhersage beruht auf Evaluationsdaten der KIs auf bestehenden, annotierten Datensätzen. Neben reiner Aufgabenerfüllung ermöglichen wir durch dedizierte Datensätze und Modellkonfigurationen, dass in diese Beurteilung auch individuell bedeutende Aspekte wie Compliance mit verschiedenen europäischen Normen einfließen können. Im Gegensatz zu reinen Auswahlroutern ermöglichen wir für jede einzelne KI eine quantitative, nach Aspekten aufgeschlüsselte Qualitätsvorhersage.

Detaillierte Leistungsbeschreibung

Arbeitspaket 1 - Predictive Evaluations:

Im ersten Arbeitspaket werden wir den Prototypen unseres KI-Routers “MIMIR.PREDICTOR” konstruieren und erproben.

Dabei wollen wir folgende Meilensteine erreichen:

- Meilenstein 1 **(Predictor Benchmarking)**: Entwicklung einer Testumgebung, die Erfolgsmetriken für Entwicklungsversionen des Predictors definiert und einen Vergleich mit anderen Ansätzen als Baselines erlaubt.
- Meilenstein 2 **(Predictor Implementation)**: Implementierung und Training eines transformer-basierten Predictors mit grundlegender Funktionalität für einzelne und mehrere Agentenkandidaten, der verwertbare Vorhersagen messbar erfolgreich trifft.
- Meilenstein 3 **(Predictor Optimization)**: Optimierung des Predictors für Anwendungen durch Anpassung des Trainings (inkl. verwendeter Evaluationsdaten), Ausweitung der Vorhersagekategorien (correctness, compliance, costs etc.), Effizienzsteigerung bei Vielzahl von Agenten und Datenpunkten, Unsicherheitsquantifizierung sowie Umgang mit lückenhaften Agentendaten.

Arbeitspaket 2 - Router Interface and Infrastructure:

Im zweiten Arbeitspaket werden wir eine Infrastruktur aufbauen, um effizient Evaluationsdaten verschiedener KI-Agenten auf vorliegenden annotierten Daten zu generieren. Außerdem entwickeln wir Methoden zur Aufbereitung der Ergebnisse sowohl dieser rohen Auswertungsdaten als auch des Predictor-Outputs, damit diese vom menschlichen Benutzer oder bei automatisierter Weiterverarbeitung (z.B. multi-agent routing) effektiv nutzbar sind.

Dabei wollen wir folgende Meilensteine erreichen:

- Meilenstein 1 **(Server Set-up)**: Aufsetzen einer den Anforderungen entsprechenden und anpassbaren Compute-Infrastruktur für Evaluationsdaten, Predictor-Training und Anwendungen wie Routing. Inkl. Dokumentation und Reproduzierbarkeit z.B. bei Dienstleisterwechsel.
- Meilenstein 2 **(LLM Evaluation Pipeline)**: Erstellung einer Evaluationspipeline, die auf reproduzierbare Weise Auswertungsdatensätze lädt, die Ergebnisse von lokalen und per API angezapften KI-Agenten ausgeben lässt, diese einheitlich formatiert und vorauswertet, und zur direkten weiteren Nutzung bereitstellt.

- Meilenstein 3 **(Router Reporting)**: Entwicklung und Implementierung von Prozessen zur kundenorientierten Aufbereitung von Evaluations-, Predictor- und Routingergebnissen für Betrachtung und Interpretierbarkeit. Vergleich unseres Routers mit Standard Baselines durch das Sammeln und Aufsetzen von existierenden Router-Lösungen und Router Benchmarks.

Arbeitspaket 3 - Evaluation Data for Performance, Democratization and Legal Risk Management of AI-Agents:

Im Im dritten Arbeitspaket werden wir eine strukturierte Sammlung von LLM-Evaluationsmethoden und -datensätzen (Evaluationsressourcen) aufbauen, die insbesondere sowohl die Anforderungen europäischer Normen als auch des unternehmerischen Risikomanagements abbilden. Neben deren Zusammenstellung in einer Evaluationsplattform wollen wir Lücken in der vollständigen Abdeckung dieser Aspekte identifizieren und durch eigens entwickelte Evaluationsressourcen schließen.

Dabei wollen wir folgende Meilensteine erreichen:

- Meilenstein 1 **(Systematization of Evaluation Resources)**: Systematische Erfassung und Kartierung vorhandener Evaluationsressourcen (v.a. Benchmarks) hinsichtlich Performance-Aspekten (wie Nützlichkeit zur Aufgabenbewältigung, Effizienz, Genauigkeit) sowie Compliance-Anforderungen (wie Recht u. Haftung, weitere Unternehmensrisiken).
- Meilenstein 2 **(Identifying and Closing Gaps in the Evaluation Landscape)**: Kundenorientierte Analyse unzureichend abgedeckter Risikoaspekte inkl. Darstellung des Kundennutzens (Risikoscore/Risikomatrix). Methoden zum Sichtbarmachen unserer Daten und ihrer Abdeckung und Qualität. Gezielte Entwicklung neuer Evaluationsressourcen zur Schließung festgestellter Lücken.
- Meilenstein 3 **(Evaluation Data Platform)**: Aufbau eines strukturierten Vergleichs- und Auswahlverfahrens hilfreicher Evaluationsressourcen. Zusammenstellung in einer hochwertigen und dynamischen Datenbank, die die Verfügbarkeit, Nachvollziehbarkeit und Aktualität aller Evaluationsressourcen sicherstellt und sowohl dem Training unseres Routers als auch der Demokratisierung der Auswahl von AI-Agenten dient. Implementierung und Bereitstellung einer dynamischen Evaluationsplattform.

Zeitplan

Die Projektlaufzeit beträgt 5 Monate. Wir bearbeiten die drei einzelnen Arbeitspakete parallel. Die Meilensteine bauen jeweils aufeinander auf und werden, wie in folgender Tabelle dargestellt, bearbeitet. Beispielsweise sollen die Meilensteine 1 jedes Arbeitspakets nach drei Monaten erfolgreich zur Verfügung stehen und danach weiter ergänzt werden.

Monate / **Arbeitspakt**	**1**	**2**	**3**	**4**	**5**
Meilenstein 1					
Meilenstein 2					
Meilenstein 3					