Gliederung Bachelorthesis – Copyright-Scanner

1. **Einleitung**
   - 1.1. Einführung in die Relevanz des License Compliance Managements und die Rolle von Copyright-Statements.
   - 1.2. Darstellung der Problemstellung bei der automatisierten Extraktion von Copyright-Informationen.
   - 1.3. Zielsetzung der Arbeit: Entwicklung und Bewertung eines AI-gestützten Copyright-Scanners.
   - 1.4. Abgrenzung des Projektumfangs und Erläuterung der methodischen Vorgehensweise.
2. **Stand der Technik**
   - 2.1. Das ScanCode Toolkit
   - 2.2. Der ScanCode Service
   - 2.3. Analyse bestehender Lösungen und deren Schwächen.
   - 2.4. Erläuterung der Policy und Einordnung der vorhandenen Lösungen
   - 2.5. Vorstellung von Large Language Models und deren Potenzial für die Extraktion unstrukturierter Daten.
   - 2.6. Verwandte Arbeiten
3. **Anforderungen**
   - 3.1. Funktionale Anforderungen und ihre Priorisierung
   - 3.2. Nicht-funktionale Anforderungen und ihre Priorisierung
4. **Daten**
   - **4.1.** Wahl der Datenquelle
   - **4.2.** Erzeugung des Ausgangsdatensatzes
   - **4.3.** Analyse des Datensatzes und Kategorisierung der Daten
   - **4.4.** Herausforderungen bei der Datenaggregation
   - **4.5.** Qualität der Daten
5. **Benchmark und Modellauswahl**
   - 5.1. Definition der Auswahlkriterien für geeignete Sprachmodelle.
   - 5.2. Konzeption des Benchmarks anhand der genannten Auswahlkriterien.
   - 5.3. Erstellung eines Testdatensatzes für die Durchführung eines Benchmarks.
   - 5.4. Durchführung des Benchmarks und Auswertung der Ergebnisse.
   - 5.5. Begründete Auswahl des Modells für die Implementierung des Copyright-Scanners.
6. **Experimente zur LLM-gestützten Extraktion**
   - 6.1. Beschreibung der Extraktionsexperimente und Formulierung einer Erwartungshaltung.
   - 6.2. Identifikation problematischer Copyrights und Lizenztexte.
   - 6.3. Durchführung von Extraktionsexperimenten mittels Prompt Engineering und Evaluierung der Ergebnisse.
   - 6.4. Nutzung eines weiteren LLMs zur Validierung der Ergebnisse.
   - 6.5. Konzeption einer Umsetzung mit Hilfe von Fine-Tuning.
7. **Implementierung des Copyright-Scanners**
   - 7.1. Funktionale & nicht-funktionale Anforderungen.
   - 7.2. Konzeption des Copyright-Scanners.
   - 7.3. Beschreiben der Schnittstellen und Integration in bestehende Systeme.
   - 7.4. Dokumentation des Copyright-Scanners und seiner Komponenten.
8. **Evaluation und Bewertung der Ergebnisse**
   - 8.1. Definition der Evaluationskriterien.

8.2. Analyse der Ergebnisse anhand der Evaluationskriterien.
8.3. Vergleich der Evaluierungsergebnisse mit bestehenden Lösungen.

9. Diskussion

9.1. Interpretation der Ergebnisse im Kontext der ursprünglichen Problemstellung.
9.2. Reflexion über Herausforderungen und Limitationen bei der Umsetzung und Evaluierung.
9.3. Betrachtung der Integrationsfähigkeit der entwickelten Lösungen in bestehende Abläufe und Kundenlandschaften.
9.4. Kritische Würdigung der Vorgehensweise in Hinsicht auf Erfolge & Versäumnisse.

10. Fazit und Ausblick

10.1. Zusammenfassung der wichtigsten Erkenntnisse und Ergebnisse der Arbeit.
10.2. Bewertung des entwickelten Prototyps hinsichtlich seines Potenzials für den praktischen Einsatz.
10.3. Ausblick auf mögliche Weiterentwicklungen, Optimierungen und Forschungsperspektiven im Bereich der automatisierten Lizenz-Compliance.

11. Anhang

12. Literaturverzeichnis