

## Automatische Identifikation und Kategorisierung von Erklärungsbedarf in App-Reviews mittels LLMs

**Kommentiert [M1]:** Erhöhter Reflexionsbedarf, auch über die Threats to Validity

### Hintergrund

Im Rahmen unseres Projekts [softXplain](#) haben wir Erklärungsbedarf von Nutzern in App-Reviews identifizieren können [1]. Dieser Bedarf tritt in der Regel auf, wenn Nutzer Fragen zur Software haben, insbesondere bei unerwartetem Systemverhalten. Um diesen Bedarf besser zu strukturieren, haben wir eine Liste von Aspekten erstellt, die Erklärungsbedarf in verschiedene Kategorien unterteilt [2], wie z.B. zu Software-Features oder Interaktionen mit der Software. Diese Kategorien dienen als Werkzeug, um neuen Erklärungsbedarf automatisch zu identifizieren und an die zuständigen Personen weiterzuleiten. Es fehlt jedoch ein multidimensionaler Goldstandarddatensatz, der basierend auf diesen Aspekten gelabelt und anschließend für das Training von LLMs (Large Language Models) verwendet werden soll.

### Aufgabe

Im Rahmen dieser Arbeit sollen Sie basierend auf den identifizierten Aspekten einen multidimensionalen Goldstandarddatensatz erstellen (wie z.B. [3]). Anschließend sollen Sie (generative) LLMs wie BERT, Setfit, ChatGPT, Lama und Gemini trainieren und verwenden, um die App-Reviews hinsichtlich des Erklärungsbedarfs zu klassifizieren. Darüber hinaus sollen Sie die LLMs einsetzen, um die relevanten Textstellen in den Reviews zu markieren, die den Erklärungsbedarf sowie die zugehörigen Aspekte enthalten.

Diese Arbeit gliedert sich in die folgenden Schritte:

1. Einarbeiten in die Grundlagen von Erklärbarkeit und die Erstellung von Goldstandarddatensätzen
2. Entwicklung eines Konzepts für die Erstellung des Goldstandarddatensatzes und das Training der LLMs
3. Datensatz bestehend aus App-Reviews scrapen (mindestens 7.500 Reviews)
  - a. Vorfilterung auf Erklärungsbedarf
4. Labeling des Datensatzes (2-3 Rater)
  - a. Einteilung der Reviews in verschiedene Aspekte und Markierung der relevanten Stelle
  - b. Optional: Einteilung in expliziten und impliziten Erklärungsbedarf
5. Training und Evaluation von LLMs wie BERT, Setfit, ChatGPT und Gemini
6. Integration der Ergebnisse in eine vorhandene Weboberfläche
  - a. Einarbeiten in den vorhandenen Code
  - b. Implementierung des besten LLMs zur Identifikation und Kategorisierung von Erklärungsbedarf
7. Evaluation und Diskussion der Ergebnisse

### Organisatorisches

**Betreuer:** Martin Obaidi  
**Beginn:** ab sofort möglich

**Prüfer:** Prof. Dr. Kurt Schneider

### ***Literaturempfehlung***

- [1] [Automatisierte Detektion von Erklärungsbedarf in Nutzerfeedback zu Software](#)
- [2] Identification of Criteria for Classifying the Explanation Needs in Apps
- [3] [A Four-Dimension Gold Standard Dataset for Opinion Mining in Software Engineering](#)
- [4] [Entwicklung einer Software zur Extrahierung und Analyse von Reviews aus App Stores](#)
- [5] [Automatische Erstellung von Erklärbarkeitsanforderungen und Erklärungen](#)