1 Projektziel

Das Ziel dieses Forschungsprojekts ist es, die Fähigkeit von Large Language Models (LLMs), insbesondere GPT-4o mini (aktuell beste frei verfügbare Variante), zur Identifikation von Bugs in Programmcode zu untersuchen. Ein besonderer Fokus liegt auf dem Unterschied zwischen der Nutzung eines festgelegten Vokabulars zur Fehlerbeschreibung und einer offenen Suche nach Fehlern. Die zentrale Frage ist, ob es hilfreich ist, das LLM gezielt auf bestimmte Bug-Kategorien hinzuweisen oder ob dies zu Halluzinationen und einer einseitigen Fehlerbewertung führt.

2 Forschungsfragen

Die Untersuchung soll folgende Forschungsfragen beantworten:

- **Leistungsfähigkeit der Fehlersuche:** Wie gut erkennt GPT-4o mini Bugs in Programmcode?
- **Einfluss von Vorgaben:** Führt die explizite Vorgabe bestimmter Bug-Typen zu besseren oder verzerrten Ergebnissen?
- **Vergleich zwischen freier und gezielter Suche:** Welche Unterschiede bestehen zwischen einer offenen Fehlersuche und einer spezifischen Fehlersuche nach bestimmten Bug-Kategorien?
- **Halluzinationen und Relevanz der Fehler:** Wie häufig erkennt das Modell Bugs, die nicht existieren? Gibt es Muster in den Fehldiagnosen?

3 Methodik und Relevanz

Zur Beantwortung der Forschungsfragen wird ein empirischer Ansatz verfolgt:

- Testfälle mit synthetischem und realem Code werden verwendet.
- Das LLM wird mit zwei Methoden getestet:
  - Offene Bug-Suche: Das Modell wird ohne Vorgaben aufgefordert, Fehler zu finden.
  - Gezielte Bug-Suche: Das Modell erhält Vorgaben, nach bestimmten Bug-Typen zu suchen.
- Die gefundenen Bugs werden klassifiziert und mit tatsächlichen Fehlern abgeglichen.
- Eine Analyse der Halluzinationen erfolgt durch Überprüfung falsch erkannter Fehler.