

Development and Application of a Robustness Evaluation Framework of LLM-based C-to-Rust Translations

Referent:

Martin Weiss

Masterstudent der Otto-von-Guericke Universität Magdeburg, FIN

Betreuer:

Prof. Dr. Andreas Schmietendorf, Uni Magdeburg/HWR Berlin

Dr. Jesko Hecking-Harbusch, Robert Bosch GmbH

Abstrakt zur Masterarbeit:

Legacy-Codebasen in C sind bekanntermaßen anfällig für Speichersicherheitsprobleme, weshalb eine Migration in sicherere Sprachen wie Rust sehr wünschenswert ist. Das manuelle Übertragen der immensen Codemengen ist jedoch unrealistisch, und klassische, regelbasierte Transpiler erzeugen weder idiomatischen noch sicheren Rust-Code. Die Programmierfähigkeiten großer Sprachmodelle (LLMs) erscheinen hier als vielversprechende Alternative für eine idiomatische Übersetzung. Ihre praktische Einsetzbarkeit erfordert jedoch eine umfassende Untersuchung, insbesondere im Hinblick auf ihre Robustheit bei der Übersetzung der vielfältigen, realen C-Codebasen. Diese Arbeit schließt diese Lücke, indem sie ein umfassendes Robustheits-Evaluations-Framework für LLM-basierte C-zu-Rust-Übersetzung entwickelt und anwendet.

Das Framework bewertet die Anfälligkeit von LLMs gegenüber Prompt-Variationen mithilfe eines breiten Spektrums codefokussierter Perturbationen. Es nutzt ein aktuelles Codeübersetzungssystem, das funktionale Äquivalenz direkt prüft und bei Fehlschlägen automatische Feedback-Schleifen einsetzt. Durch die Quantifizierung der Übersetzungsleistungsabweichungen zwischen perturbierten und unveränderten Eingaben liefern die Metriken des Frameworks einen detaillierten Blick auf die Robustheit. Ergänzend wird erfasst, ob die perturbationsbedingten Leistungsschwankungen über die erwarteten Variationen von LLMs hinausgehen, welchen Einfluss Feedback-Schleifen auf die Robustheit haben, und ob semantische Ähnlichkeiten zwischen Eingabevarianten die Robustheit der LLMs vorhersagen können.

Die Anwendung des Frameworks zeigt, dass moderne Modelle gegenüber den meisten Eingabevariationen robust sind, insbesondere, wenn Feedback-Schleifen eingesetzt werden. Diese Schleifen verbessern sowohl den Gesamterfolg der Übersetzung als auch die Robustheit und sind somit ein unverzichtbarer Bestandteil LLM-basierter Übersetzungssysteme. Trotzdem offenbart die Evaluation Schwachstellen bei bestimmten Perturbationen, von denen eine modellunabhängig und mehrere modellspezifisch sind. Während moderne Modelle überwiegend robustes Verhalten zeigen, besitzen sie individuelle Stärken und Schwächen, die die Wahl des optimalen Modells für ein Übersetzungssystem beeinflussen. Da sich die Robustheit gegenüber bestimmten Variationen nicht auf alle Modelle verallgemeinern lässt, schließt die Arbeit, dass eine kontinuierliche Bewertung neuer Systeme notwendig ist, um deren spezifische Stärken und Schwächen zu verstehen. Nur eine umfassende Evaluation ermöglicht eine feingranulare Entscheidung darüber, welches LLM am besten zu den Anforderungen einer C-zu-Rust-Migration passt.