

FM & AI ?

Prof. Dr. Holger Schlingloff ^{1,2}

Sunday, June 15th, 2025

(1) Institut für Informatik der Humboldt Universität(2) Fraunhofer Institut für offene Kommunikationssysteme FOKUS

H. Schlingloff, IFIP-WG 1.3, 15.6.2025

Slide 1





- AI is here to stay (we can't pretend it isn't)
- It already now massively affects both teaching and industrial practice
- We have to position ourselves or we risk becoming insignificant

Two research directions

- AI for FM
 - (How) can we use NN to find proofs, invariants, examples, counterexamples, ...?
 - (How) can we use LLMs to support education in FM?
- FM for AI
 - (How) can we give correctness guarantees for AI generated results?
 - If this is not possible, what is the second best we can do?

H. Schlingloff, IFIP-WG 1.3, 15.6.2025

A slide from my last year's talk



o Du

A so-called "Chris-Rupp-Sentence" consists of the following components: First a precondition, then the words "the system", the one of the words "shall", "should" or "will", then a process, and finally an object or process detail. A process is a word describing an activity, or the words "be able to" followed by a word describing an activity, or the word "provide" followed by an object followed by the words "the ability to" followed by a process. Is the following a Chris-Rupp-Sentence? "If the signal is on, the system should provide the student the ability to submit the thesis."

ChatGPT

Yes, the given sentence fits the structure of a Chris-Rupp-Sentence. Let's break it down:

- Precondition: "If the signal is on,"
- "the system"
- "should" (indicating a requirement)
- Process: "provide"
- Object: "the student"
- Process Detail: "the ability to submit the thesis."

So, the sentence follows the specified format for a Chris-Rupp-Sentence.

Me:

Can you re-formulate the following as a Chris-Rupp-Sentence? "Let us assume that the signal is set to on. Then the intended system must be programmed such that it offers students a possibility to hand in their work for evaluation as a thesis."

ChatGPT

Certainly! Here is the re-formulated sentence in the Chris-Rupp-Sentence format:

"If the signal is set to on, the system shall be programmed to offer students the ability to hand in their work for evaluation as a thesis."

H. Schlingloff, IFIP-WG 1.3, 15.6.2025

New paper

Evaluation of the idea with different LLMs and prompting techniques Results in a nutshell:

- Few-shot learning (by example) quite successful
- Formal layer essential for correctness
- The prompt does matter
- Mostly 100% matches, rare hallucinations
- Responses from DeepSeek and GPT-4o were identical
- Small models may also be ok

LESS is more: Guiding LLMs for Formal Requirement and Test Case Generation

Abhishek Shrestha
 $^{1[0009-0007-9627-8355]}$ and Bernd-Holger Schlingloff
 $^{1[0000-0001-9601-157X]}$

Fraunhofer-Institut für Offene Kommunikationssysteme FOKUS, Berlin 10589, Germany {abhishek.shrestha, holger.schlingloff}@fokus.fraunhofer.de

Abstract. Large Language Models (LLMs) demonstrate impressive incontext reasoning capabilities; however, generating structured outputs remains challenging. In this paper, we investigate prompt-based techniques to guide LLMs in producing outputs compliant with a domainspecific language called Language for Embedded Safety and Security (LESS). Additionally, we evaluate the effectiveness of LLMs in automating test case generation. Our results show that structured prompt engineering significantly enhances the accuracy and consistency of generated requirements, and that using controlled language formats such as LESS as an intermediate representation substantially improves test case generation accuracy.

Keywords: Large language model \cdot Domain-specific language \cdot Controlled language generation \cdot Test case generation

1 Introduction

Natural language processing techniques, and, in particular, Large Language Models (LLMs) from Artificial Intelligence (AI), have a significant potential in automating software engineering tasks [6]. This includes the generation of requirements, program code, and test cases. However, generating accurate and structured output, often crucial in this domain, presents a significant challenge



AI in Verification and Education



A small experiment, conducted yesterday:

Me: Can you help me to derive a loop invariant for Dafny?

ChatGPT: Absolutely, I can help with that. To start, I'll need a few things from you: ...

- **Me:** The program is while (i>0) {i--; x+=y}; precondition is (x=0 && i=n); postcondition is x=n*y.
- **ChatGPT:** Thank you for sharing the loop! Let's walk through this step by step and figure out the invariant. ...

```
method LoopExample(n: int, y: int) returns (x: int)
requires n >= 0
ensures x == n * y

{
    var i := n;
    x := 0;
    }
    while (i > 0)
    invariant 0 <= i <= n;
    invariant x == (n - i) * y;
    {
        i := i - 1;
        x := x + y;
    }
}</pre>
```

Yes, I know, this is susceptible to hallucinations, but still ...

H. Schlingloff, IFIP-WG 1.3, 15.6.2025

ANN Verification



- NN = directed graph, where each node computes a simple mathematical function (e.g., relu(x)=x, if x>0, 0 otherwise)
- Easy to encode in SMT: $F_v \triangleq (\underbrace{v^{\text{in},1} > 0}_{x>0} \Rightarrow v^{\text{o}} = v^{\text{in},1}) \land (\underbrace{v^{\text{in},1} \leqslant 0}_{x\leqslant 0} \Rightarrow v^{\text{o}} = 0)$
- Typical properties
 - robustness: similar inputs lead to similar results ٠ e.g. pictures distorted w.r.t. brightness, color, positioning of objects, etc.
 - properties modulo semantic similarity, i.e., word embeddings
- SAT-approach currently works for ANNs of up to 10⁶ nodes
 - "symbolic" representation for large ANNs?

Discussion points



- Can AI be a breakthrough in theorem proving?
- How should we change FM education, is it still needed after all?
- Can we specify a picture of a cat?
- Can we ever verify ANNs of size 10¹² nodes?