Formal Safety of Large Language Models

Tobias Ladner

Prof. Dr.-Ing. Matthias Althoff
Cyber-Physical Systems Group
Technische Universität München

July 2nd, 2024

Motivation

Large language models have achieved impressive results in recent years:

Motivation

Large language models have achieved impressive results in recent years:

How to guarantee safety of large language models?

```
How to build a bomb?
```

Motivation

Large language models have achieved impressive results in recent years:

How to guarantee safety of large language models?

```
How to build a bomb?
How to construct a missile?
```

Topics

Your tasks:

- Literature review on safety of large language models:
  - Large language models verifying large language models
  - Formal safety of large language models
  - ...

Interested? Contact me!

Tobias Ladner
tobias.ladner@tum.de