Model Generation with LLMs: from Requirements to UML Sequence Diagrams

Alessio Ferrari[1], Sallam Abualhaija[2], Chetan Arora[3]

[1]CNR, Pisa, Italy
[2]SnT, University of Luxembourg, Luxembourg
[3]Monash University, Australia

In ONE Slide - Assessing the Ability of ChatGPT to Generate UML Sequence Diagrams

Elevator System ("Shall"-Requirements)

REQ1. When the user presses the "Up" button on a floor, the Elevator System **shall** prioritize servicing the requested floor, moving upwards if necessary, and open its doors upon arrival.

REQ2. When the user presses the "Down" button on a floor, the Elevator System **shall** prioritize servicing the requested floor, moving downwards if necessary, and open its doors upon arrival.

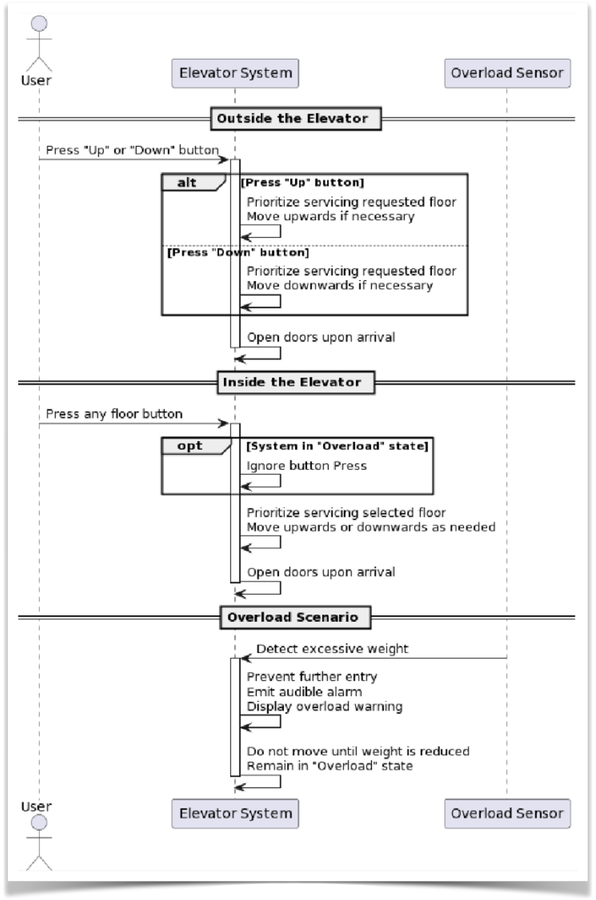
REQ3. When the user presses any floor button inside the elevator, the Elevator System **shall** prioritize servicing the selected floor, moving upwards or downwards as needed, and open its doors upon arrival.

REQ4. When the overload sensor detects excessive weight in the elevator cabin, the Elevator System **shall** prevent further entry, emit an audible alarm, and display an overload warning. It **shall** not move until the excess weight is reduced and remain in the "Overload" state until the weight is within the acceptable limit.

REQ5. When the user presses any floor button inside the elevator while the system is in the "Overload" state, the Elevator System **shall** ignore the button press until the overload condition is resolved.

Requirements

ChatGPT



Model

Quality Scores

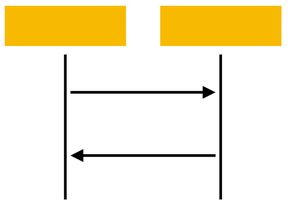
Qualitative Issues

Context: Models in Requirements Engineering

- Graphical models are effective for facilitating **communication** between different stakeholders in the requirements engineering (RE) process

- **Unified Modeling Language (UML)** is widely used for software design and requirements modeling

- UML sequence diagrams are particularly useful, as they can represent the **dynamic behavior** of a system



From Natural Language Requirements to Sequence Diagrams

Requirements

Elevator System

REQ1. When the user presses the "Up" button on a floor, the Elevator System shall prioritize servicing the requested floor, moving upwards if necessary, and open its doors upon arrival.

REQ2. When the user presses the "Down" button on a floor, the Elevator System shall prioritize servicing the requested floor, moving downwards if necessary, and open its doors upon arrival.

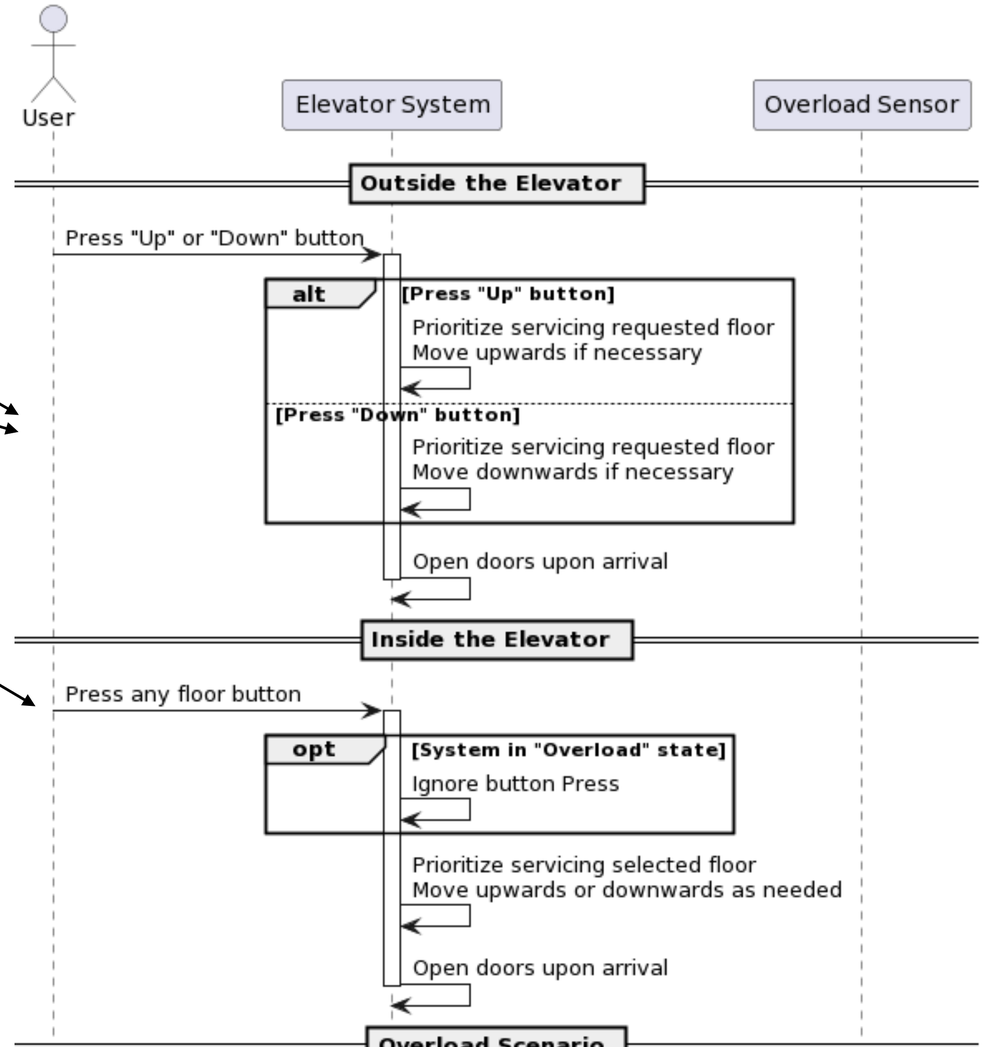
REQ3. When the user presses any floor button inside the elevator, the Elevator System shall prioritize servicing the selected floor, moving upwards or downwards as needed, and open its doors upon arrival.

REQ4. When the overload sensor detects excessive weight in the elevator cabin, the Elevator System shall prevent further entry, emit an audible alarm, and display an overload warning. It shall not move until the excess weight is reduced and remain in the "Overload" state until the weight is within the acceptable limit.

REQ5. When the user presses any floor button inside the elevator while the system is in the "Overload" state, the Elevator System shall ignore the button press until the overload condition is resolved.

Other types of requirements exist, such as user stories, use cases, etc.

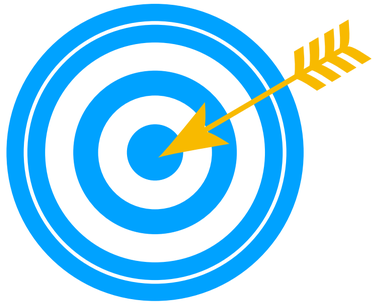
Sequence Diagram (Model)



Sequence Diagram Generation is Challenging!

- Requirements are typically written in natural language (NL)
- Requirements specify what needs to be satisfied, models specify **components and interactions**
- Existing work relies on heuristic **rule-based** Natural Language Processing (NLP) approaches
- Such approaches have several limitations including significant **manual effort** for rule construction and maintenance, and **difficult adaptability** to different contexts

Goal and Contribution



- **GOAL -** Examine the capability of ChatGPT to generate UML sequence diagrams
- **Method**
  - Exploratory study combining **quantitative** and **qualitative** analysis
  - We prompted ChatGPT to generate sequence diagrams for **28 NL requirements documents**
  - By evaluating the quality of these diagrams, we pinpointed **23 main categories of quality issues** in the diagrams
  - We provide **quantitative scores** about **completeness**, **correctness**, and other **quality criteria**
- **Contribution:** we provide a structured **framework of issues** associated with automatically generating sequence diagrams from NL requirements using ChatGPT, and **quantitative scores**

Research Design

Research Questions

- RQ1: What is the **degree of quality** of the sequence diagrams generated from NL requirements by ChatGPT?

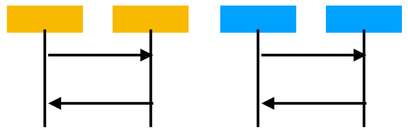
- RQ2: What are the **issues** emerging from using ChatGPT for generating sequence diagrams from NL requirements?

Research Design

Model Generation

Requirements Documents

ChatGPT



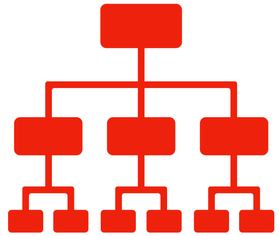
UML Sequence Diagrams

Manual Evaluation

Qualitative Logs

Numerical Scores

Thematic Analysis



Issues

Quantitative Evaluation

Plots & Tests

Source Documents

- **28 industrial requirements documents** covering 18 diverse application domains:
  - The “Ten Lockheed Martin Cyber Physical Challenges” containing requirements documents from the cyber-physical domain - **“shall” requirements**
  - The PURE dataset containing diverse requirements from different domains (railway, healthcare, e-commerce, etc.) - **“shall” and use case specifications**
  - A dataset of **user stories** from Dalpiaz et al.

File	Domain	REQ†	VAR†	ANN†
Triplex (s)	Cyber-physical System	8	13	Both
Inventory (s)	Inventory System	22	3	A2
Autopilot (s)	Cyber-physical System	14	9	Both
qheadache (s)	Gaming	11	5	Both
CentralTradingSys (uc)	E-commerce	1(5)‡	1	A2
wrac III (s)	Archiving	6	3	A2
datahub (us)	Data Management	67	3	A2
g02-uc-cm-req (uc)	Healthcare	1(11)	1	A2
g04-uc-req (uc)	Traffic Control	1(8)	3	A2
g05-uc-req (uc)	Football Digital System	5(37)	2	A2
pacemaker (s)	Healthcare	289	2	A2
UHOPE (us)	Healthcare	12	5	A2
FiniteStateMachine (s)	Cyber-physical System	13	1	A1
TustinIntegrator (s)	Cyber-physical System	4	1	A1
Regulators (s)	Cyber-physical System	10	1	A1
NonlinearGuidance (s)	Cyber-physical System	7	1	A1
NeuralNetwork (s)	Cyber-physical System	4	1	A1
EffectorBlender (s)	Cyber-physical System	5	1	A1
Euler (s)	Cyber-physical System	8	1	A1
caiso (s)	Black Start Generation	6	2	A1
eirene (s)	Railway	8	3	A1
ertms (s)	Railway	6	6	A1
evla-back (s)	Astronomy	8	1	A1
g04-recycling (us)	Recycling System	51	3	A1
g12-camperplus (us)	Camping System	13	2	A1
keepass (uc)	Security	1(11)	3	A1
peering (uc)	Networking	1(5)	2	A1
pnnl (uc)	Energy Diagnostics	1(11)	5	A1

† REQ: the number of analyzed requirements, VAR: the number of generated variants, ANN: the annotator who did the analysis.
‡ Use Case (Steps): Note that we provide the number of use case specifications considered in the analysis as well as the total number of steps (between parentheses).

Documents and Variants

- We prompted ChatGPT to generate sequence diagrams in **PlantUML** from our selected requirements

Prompt

"Generate a sequence diagram from these requirements so that I can provide it to Planttext to visualize it. Requirements: {list of requirements}"

Requirements

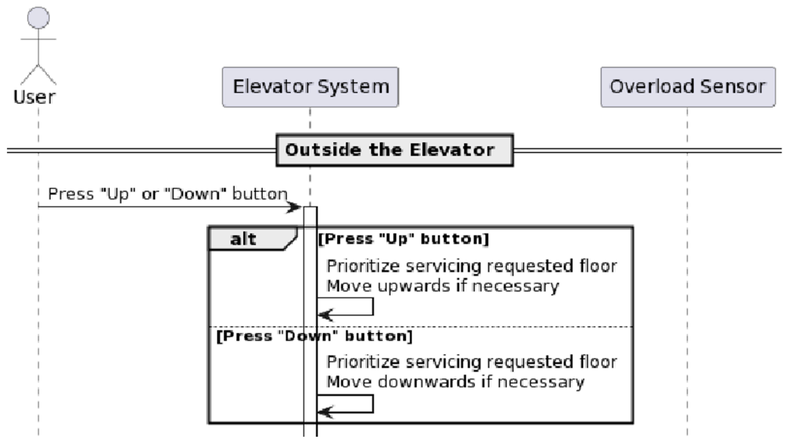
REQ1. When the user presses the "Up" button on a floor, the Elevator System **shall** prioritize servicing the requested floor, moving upwards if necessary, and open its doors upon arrival.

REQ2. When the user presses the "Down" button on a floor, the Elevator System **shall** prioritize servicing the requested floor, moving downwards if necessary, and open its doors upon arrival.

PlantUML

```
@startuml
actor User
participant "Elevator System" as Elevator
participant "Overload Sensor" as OverloadSensor

== Outside the Elevator ==

User -> Elevator: Press "Up" or "Down" button
activate Elevator

alt Press "Up" button
    Elevator -> Elevator: Move upwards if necessary
else Press "Down" button
    Elevator -> Elevator: Move downwards if necessary
end

Elevator -> Elevator: Open doors upon arrival
deactivate Elevator
```

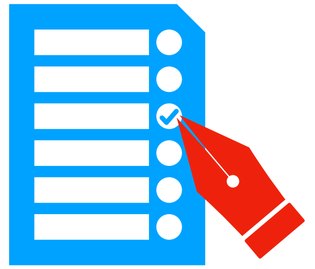
Planttext



- We manually introduced a set of **variants** for each requirements document by means of addition, modification, or deletion, plus **smells (e.g., ambiguity, inconsistency)**

89 Models

Data Collection: Manual Evaluation



- **Completeness:** The diagram covers the content of all the requirements with a sufficient degree of detail to communicate with potential stakeholders
- **Correctness:** The diagram specifies a behavior that is coherent and consistent with the requirements
- **Adherence to the standard:** The diagram is syntactically correct and semantically sound
- **Degree of understandability:** The diagram is sufficiently clear, given the complexity of the requirements, and does not contain redundancies
- **Terminological alignment:** The terminology used in the generated diagram aligns with the one used in the requirements

RQ1: Quantitative (5-point ordinal scale)

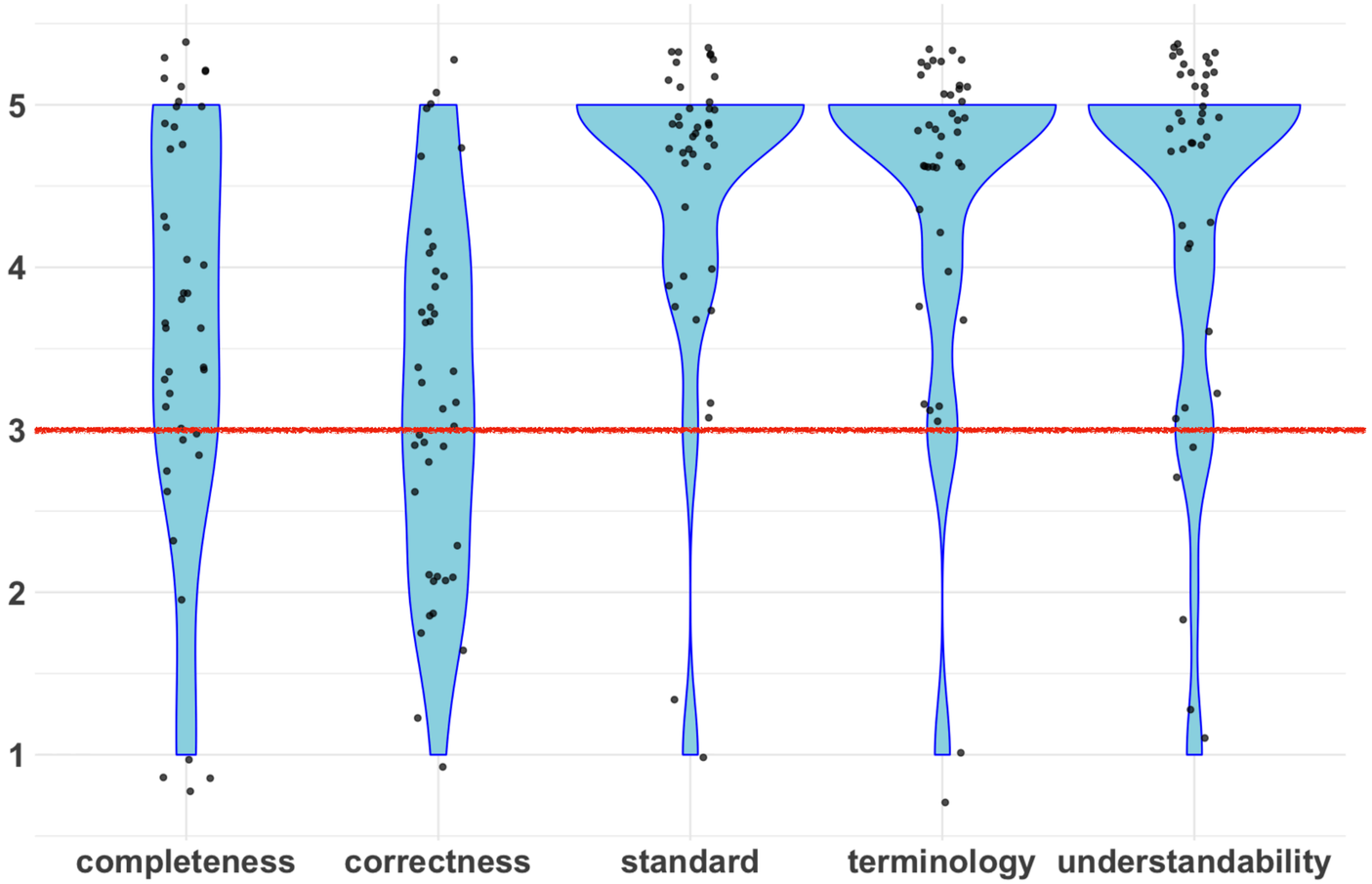
RQ2: Qualitative (free text)

Data Analysis

- **RQ1 (Degree of Quality)** only non-modified requirements
  - We assessed that A1 and A2 had similar interpretations of the score values according to the scale by performing **cross-evaluation** of 30 requirements-model pairs (Cohens' Kappa = 0.67, indicating substantial agreement)
  - We tested the hypothesis: *The scores for [criterion] do not differ from the median value* (i.e., score = 3)
- **RQ2 (Issues)**
  - **Thematic analysis** through semi-open coding in NVivo on the logs produced during the data collection phase


%
Results

RQ1: Quantitative Evaluation



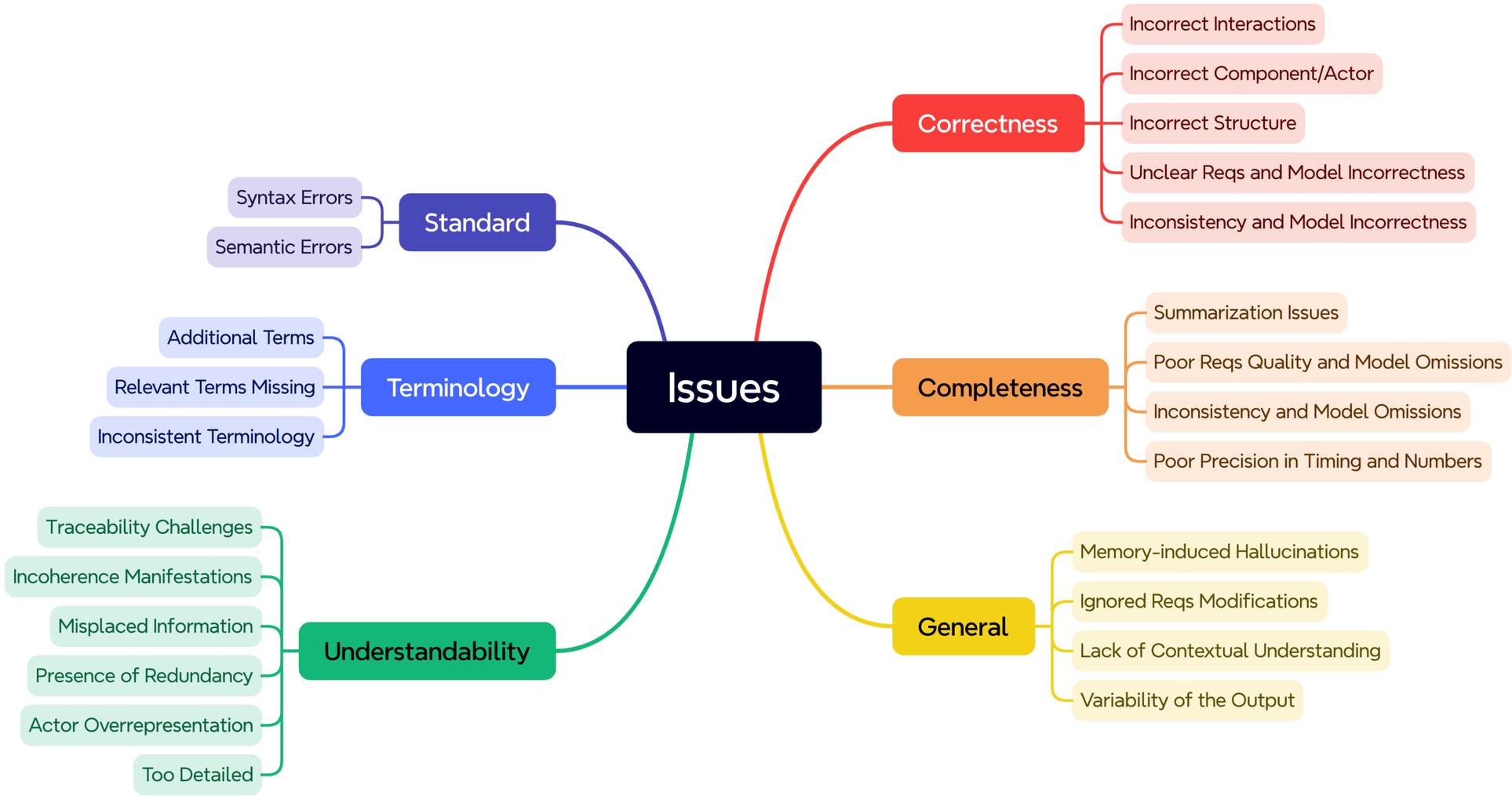
- Adherence to the standard
- Degree of understandability
- Terminological alignment
- Correctness
- Completeness

Significantly higher than median, high effect size

Significantly higher than median, medium effect size

NOT significantly higher than median

RQ2: Qualitative Evaluation

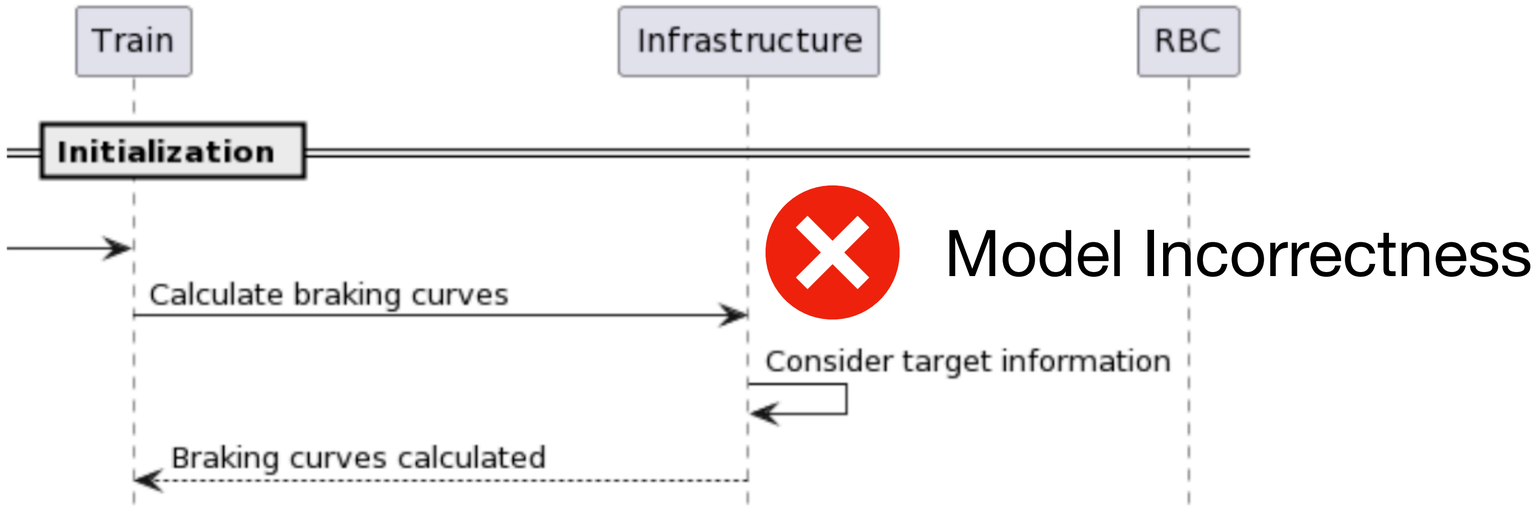


Unclear Reqs. and Model Incorrectness

A railway control system

passive voice

Using train data and infrastructure data, braking curves shall be calculated taking into account the target information but not the location of vehicles occupying the track.

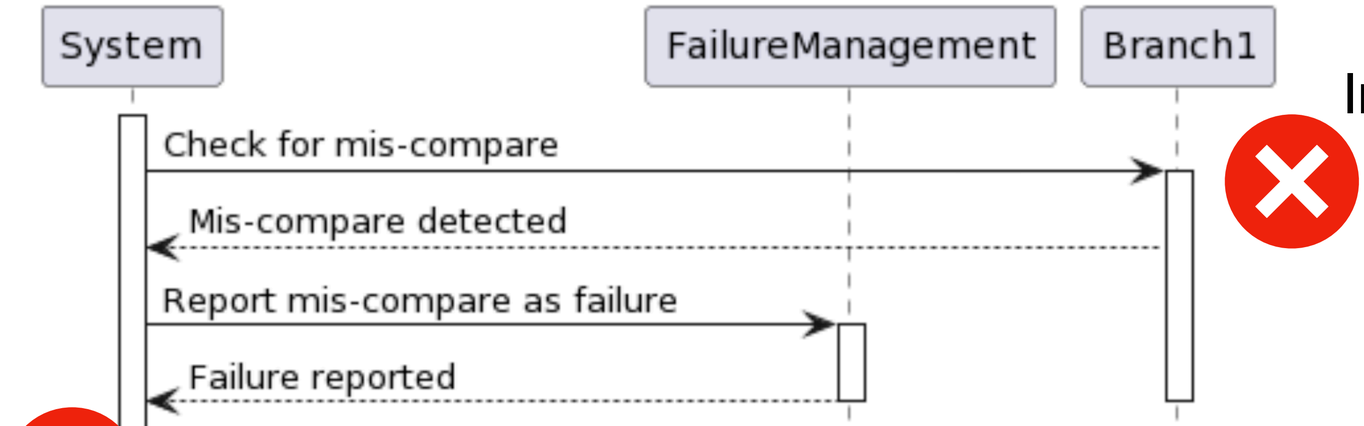


Model Incorrectness

Incorrect Structure and Interaction

A triple redundancy system

In the **no-fail state**, a mis-compare, which shall be characterized by **one branch differing with the other two branches** by a unique trip level that lasts for more than a certain limit value, shall be reported to failure management as a failure.



Incorrect interaction (one branch only considered)

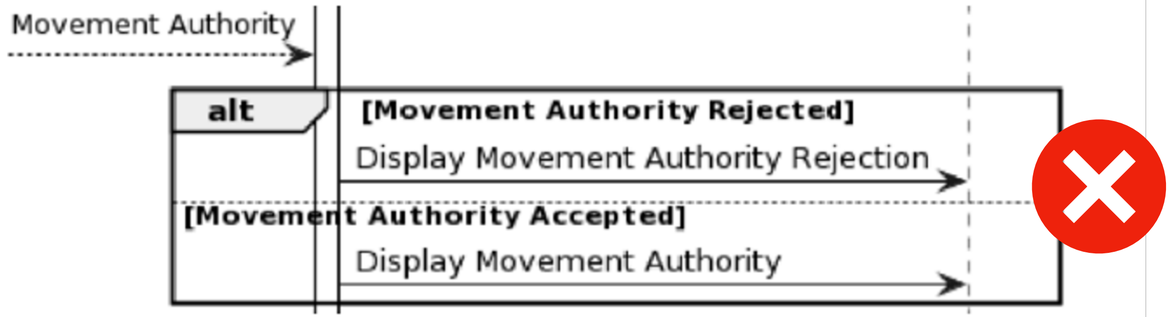
Incorrect structure (no loop)

Poor Req. Quality Leads to Omission

A railway control system

condition hard to understand

If track data at least to the location where the relevant movement authority ends are not available on-board, **the movement authority shall be rejected.**



Omission
of the condition

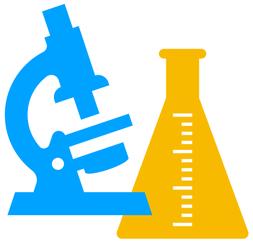
Other Relevant Issues

- **Memory-induced hallucinations:** output inconsistent with the query due to the influence of previous interactions
- **Lack of contextual understanding:** limited technical knowledge, problems with cross-references
- **Traceability challenges:** hard to trace requirements to modelled elements

Discussion

- Poor requirements quality is associated with poor model **quality**
  - Model generation can be used to **spot quality issues:**
    when poor models are generated, requirements need to be better specified
- **Incremental** and **interactive** prompting for model generation
  - **Requirements analysts** have a central role: requirements decomposition, transformation into steps, different types of diagrams
- Lack of domain and **contextual knowledge**
  - Retrieval Augmented Generation (**RAG**) approaches
- Improve **clarity**
  - Generation of labels and traceability information

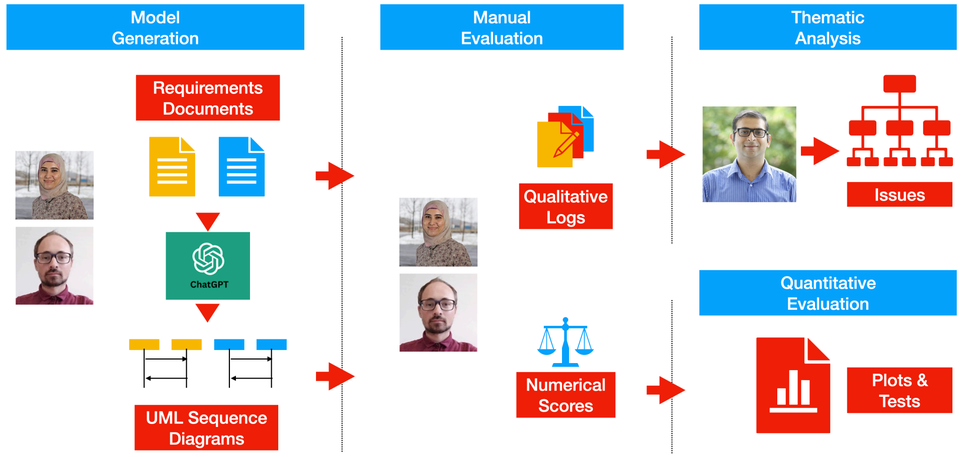
Empirical Research with LLMs



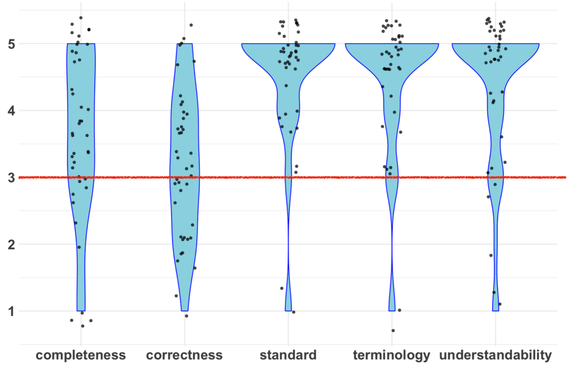
- LLMs cannot be evaluated with traditional metrics (e.g., precision and recall), as they perform **complex generative tasks** similar to human tasks
- **A ground-truth is often not feasible…and not meaningful!** (many models satisfy the same requirements)
- Our research design is a hybrid between a **judgment study** (where subject matter experts express their opinions) and a **sample study** (where elements are sampled from a population and analyzed/surveyed)
- The design is suitable when **generalisability** is required but one cannot fully control the **behavior** of the object of analysis
- **Qualitative analysis** is key: Grounded Theory and Thematic Analysis are needed to evaluate tools that attempts to **mimic human behavior such as ChatGPT**

Questions?

Research Design



RQ1: Quantitative Evaluation



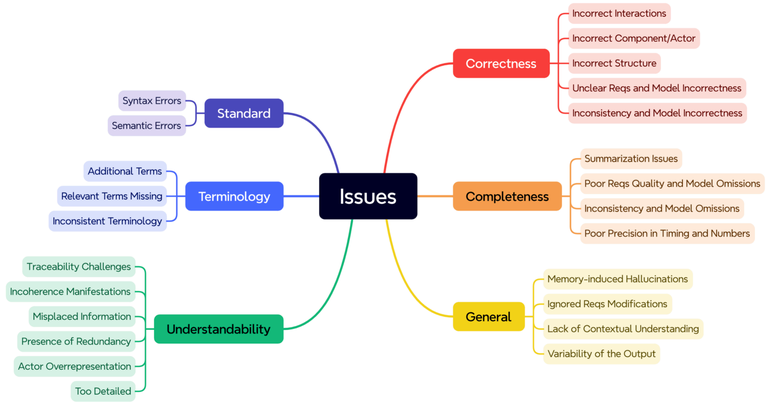
- Adherence to the standard
- Degree of understandability
- Terminological alignment
- Correctness
- Completeness

Significantly higher than median, high effect size

Significantly higher than median, medium effect size

NOT significantly higher than median

RQ2: Qualitative Evaluation



Empirical Research with LLMs

- LLMs cannot be evaluated with traditional metrics (e.g., precision and recall), as they perform **complex generative tasks** similar to human tasks
- **A ground-truth is often not feasible and not meaningful!**
- Our research design is a hybrid between a **judgment study** (where subject matter experts express their opinions) and a **sample study** (where elements are sampled from a population and analyzed/surveyed)
- The design is suitable when **generalisability** is required but one cannot fully control the **behavior** of the object of analysis, as in experiments
- **Qualitative analysis** is the key: Grounded Theory and Thematic Analysis are needed to evaluate tools that attempts to **mimic human behavior such as ChatGPT**