

DIPARTIMENTO DI INGEGNERIA E ARCHITETTURA

Corso di Laurea Magistrale Ingegneria Informatica

Enhancing Fault Isolation in Hardware Systems Using Large Language Models (LLM)

Miglioramento dell'isolamento dei guasti nei sistemi hardware utilizzando modelli linguistici di grandi dimensioni (LLM)

Relatore:

Prof. Gianfranco Lombardo

Correlatore:

Tutor aziendale Giuseppe Fiorelli

Tesi di Laurea di:

Matteo Gianvenuti Matricola: 364979

Table of Contents

[Introduction 2](#_Toc204329349)

[Working Environment 3](#_Toc204329350)

[1 State Of The Art (SOTA) 4](#_Toc204329351)

[1.1 Large Language Models (LLMs) 4](#_Toc204329352)

[1.2 Generative Pre-trained Transformer (GPT) 5](#_Toc204329353)

[1.3 Prompt Engineering 5](#_Toc204329354)

[1.4 Large Language Model Meta AI (Llama) 7](#_Toc204329355)

[1.5 Phi 7](#_Toc204329356)

[1.6 Hardware Diagnostics with AI 8](#_Toc204329357)

[1.7 Hardware Diagnostics Papers with AI 8](#_Toc204329358)

[2. Technological Analysis 10](#_Toc204329359)

[2.1 Models’ Comparison 10](#_Toc204329360)

[2.2 Final Comparison 10](#_Toc204329361)

[2.3 Quantization 11](#_Toc204329362)

[2.2.1 Bits and Bytes 8-bit Quantization 11](#_Toc204329363)

[2.4 Complexity of Implementation 12](#_Toc204329364)

[3. Method 12](#_Toc204329365)

[3.1 Preliminary Activities 12](#_Toc204329366)

[3.1.1 Questions Dataset 13](#_Toc204329367)

[3.1.2 Prompt Structure 13](#_Toc204329368)

[3.1.3 Textual Representation 15](#_Toc204329369)

[3.1.4 Json Representation 16](#_Toc204329370)

[3.1.5 Tabular Representation 17](#_Toc204329371)

[3.1.6 Model Dimension Limit 18](#_Toc204329372)

[3.1.7 Models Results 20](#_Toc204329373)

[3.1.8 PoC Conclusion 22](#_Toc204329374)

[3.2 Dataset 22](#_Toc204329375)

[3.2.1 Introduction 22](#_Toc204329376)

[3.2.2 Database Structure 22](#_Toc204329377)

[3.2.3 System Definition 24](#_Toc204329378)

[3.3 Scenarios 24](#_Toc204329379)

[3.4 ICL 25](#_Toc204329380)

[3.4.1 System Prompt without Few-shot 25](#_Toc204329381)

[3.4.2 System Prompt with Few-shot 26](#_Toc204329382)

[3.5 System Prompt Experiments 27](#_Toc204329383)

[3.6 Few-shot Experiments 27](#_Toc204329384)

[4. Results 27](#_Toc204329385)

[Bibliography 28](#_Toc204329386)

Lombardo’s structure:

Introduzione

- Ambito dell’attività di tesi

- Obiettivi dell’attività di tesi

Stato dell’arte

- Documentare tutto ciò che esisteva prima della tesi in termini di tecnologie, metodologie e modelli

2/3 capitoli di metodologia in cui spiega cosa ha effettivamente svolto, progettato e realizzato (architettura software, addestramento di modelli, dati utilizzati e loro composizione etc..)

- 1 capitolo su i risultati

- 1 breve capitolo di conclusioni in cui ricapitola il tutto traendo le conclusioni e proponendo degli sviluppi futuri.

# Introduction

In this thesis work we want to explore the use of LLMs in hardware diagnostic to analyse diagnostics test results and try to find their possible causes. LLMs have shown good memorization and "understanding" capabilities of natural language, even in reasoning tasks. Especially in the case in which the models know the environment in which they operate.

In our case, we initially want to make the LLMs understand how a hardware architecture is structured so that the model is able to solve the diagnostics at its best. Subsequently, we want to exploit the capabilities of the LLMs to identify any problems in the hardware by analysing system logs, telemetry data and hardware functionality test results. To carry out these activities it was necessary to create a specific dataset and a simulation environment. We also implemented the Open WebUI graphical interface.

So, we have two objectives, the first is to verify that the LLMs are able to understand the hardware architecture. While the second is to verify that the LLMs are able to understand and solve hardware problems. This work was based on In-Context Learning (ICL) and Prompt Engineering. In particular, we compared the performance of different models according to various metrics.

I developed this work thesis during an internship program in MBDA ITALIA SPA. MBDA, shortly called company in this work thesis, is a multinational corporation that works in defence.

The internship does not regard the development of weapons.

## Working Environment

Since I had to develop an AI project for this work thesis, the company assigned me a ZBOOK HP laptop with a small GPU (NVIDIA GA104GLM RTX A3000 Mobile GPU with 6GB of memory). That laptop has the CentOS 9 operative system. Some colleagues (Giuseppe and Alessandro) configured the Finx OS on a virtual machine through Virtual Box. It was necessary to use the diagnostics software tool HMI.

To use the GPU, I installed the necessary driver and the Miniconda software. I installed also Python. Then I created a virtual environment with all the libraries needed for this work. After the preliminary activities, I installed and configured Ollama and Open WebUI.

Miniconda is a reduced version of Anaconda. It includes conda, Python and other minor packages. Instead, Ollama is a software that allow you to download and run LLMs models locally. While Open WebUI is a web user interface like the ChatGPT web interface that can run also locally. With it you can add and configure a model. So, you can set a system prompt.

SOTA: we do this, what others do? Analyse connected technologies, test the tech did u like it? What about results?

# 1 State Of The Art (SOTA)

## 1.1 Large Language Models (LLMs)

LLMs are machine learning models with many parameters, designed for natural language processing (NLP) tasks, especially text generation. These models are typically built using deep neural networks, particularly transformer architectures, and are trained on massive corpora of textual data. They are trained with self-supervised learning on a vast amount of text. Their strength lies in the ability to capture complex syntactic, semantic, and contextual patterns, allowing them to perform a wide range of tasks such as translation, summarization, question answering, and reasoning.

LLMs can be fine-tuned for specific tasks or guided by prompt engineering, as done in this work. These models acquire predictive power regarding syntax, semantics, and ontologies inherent in human language corpora, but they also inherit inaccuracies and biases present in the data they are trained in.

The core idea behind LLMs is to learn a statistical model of language by predicting the next word (or token) in a sentence, given a context. This is achieved using self-supervised learning, where the model learns from raw text without the need for labelled data. During training, the model builds internal representations that encode meaning, structure, and dependencies across sequences of words.

The largest and most capable LLMs are generative pre-trained transformers (GPTs).

Modern LLMs like OpenAI’s GPT and DeepSeek’s DeepSeek series or Meta’s LLaMA models are characterized by having billions or even trillions of parameters. The performance of these models tends to improve significantly as the number of parameters and the size of the training data increase, a phenomenon often referred to as “scaling laws.”

In addition to linguistic capabilities, recent LLMs demonstrate emergent properties such as reasoning, commonsense inference, and even basic code generation. However, they also exhibit limitations, including sensitivity to prompt phrasing, hallucination of facts, and the presence of biases derived from the training data. So, in the context of training LLMs, datasets are typically cleaned by removing low-quality, duplicated, or toxic data. Cleaned datasets can increase training efficiency and lead to improved downstream performance.

As all the others machine learning (ML) algorithms, LLMs process numbers rather than text, the text must be converted to numbers through the tokenization process. Tokenization is a fundamental process in LLM pipelines. Before text is processed by a model, it is converted into tokens. In the first step, a vocabulary is decided upon, then integer indices are arbitrarily but uniquely assigned to each vocabulary entry, and finally, an embedding is associated to the integer index.

Most results previously achievable only by (costly) fine-tuning, can be achieved through prompt engineering, although limited to the scope of a single conversation (more precisely, limited to the scope of a context window). Due to the high training costs and the huge amount of data needed, we did not implement a fine-tune, instead we worked in the field of ICL. Also, because it is already sufficient to achieve our goals.

In our thesis work, LLMs serve a dual role: first, as systems capable of understanding the architecture and components of hardware environments through descriptive prompts; second, as tools capable of interpreting system logs and telemetry to support fault diagnosis.

## 1.2 Generative Pre-trained Transformer (GPT)

A GPT is a type of LLM and a prominent framework for generative AI. It is an artificial neural network that is used in NLP by machines. It is based on the transformer deep learning architecture, pre-trained on large datasets of unlabelled text, and able to generate novel human-like content. As of 2023, most LLMs had these characteristics and are sometimes referred to broadly as GPTs.

The first GPT was introduced in 2018 by OpenAI. OpenAI has released significant GPT foundation models (models trained on board data at scale such that they can be adapted to a wide range of downstream tasks) that have been sequentially numbered, to comprise it “GPT-n” series. Each of this was significantly more capable than the previous, due to increased size (number of trainable parameters) and training. Such models have been the basis for their more task specific GPT systems, including models fine-tuned for instruction following. GPT is based on the transformer architecture.

## 1.3 Prompt Engineering

Prompt engineering is the process of structuring or crafting an instruction in order to produce the best possible output from a generative AI model.

In case of text-to-text language models a *prompt* is natural language text describing the task that an AI should perform. It can be a query, a command, or a longer statement including context, instructions, and conversation history. Prompt engineering may involve phrasing a query, specifying a style, choice of words and grammar, providing relevant context, or describing a character for the AI to mimic.

There are many different types of prompt engineering:

* Chain-of-Thought (CoT): it is a reasoning technique that goes through several intermediate steps before giving a definitive answer to a question.

The CoT prompting improves reasoning ability by inducing the model to answer a multi-step problem with steps of reasoning that mimic a train of thought. These techniques were developed to aid common sense reasoning and multi-step reasoning.

The prompt specifies some examples of questions and answers (Q&A); this improved the results considerably. Even just writing "let's think step by step" improved the results a lot. Allowing you to generalize and avoid writing many specific examples that consume the context.

* In-Context Learning (ICL): it is the ability of models to learn information temporarily. The information of interest is specified in the prompt, typically in the initial part. This information is then used by the model in the session to answer questions by generating text, this allows the model to be trained without necessarily having to perform fine-tuning.

The weights of the network are not updated, so the information is temporary. Starting a new instance the model does not retain this knowledge.

* Prompting to estimate model sensitivity: research consistently demonstrates that LLMs are highly sensitive to subtle variations in prompt formatting, structure, and linguistic properties. Linguistic features significantly influence prompt effectiveness—such as morphology, syntax, and lexico-semantic changes—which meaningfully enhance task performance across a variety of tasks.

To address sensitivity of models and make them more robust, several methods have been proposed. Format Spread facilitates systematic analysis by evaluating a range of plausible prompt formats, offering a more comprehensive performance interval. Similarly, Prompt Eval estimates performance distributions across different prompts, enabling robust metrics such as performance quantiles and accurate evaluations under constrained budgets.

* Retrieval-augmented generation (RAG): it is a technique that enables generative AI models to retrieve and incorporate new information. It modifies interactions with a LLM so that the model responds to user queries with reference to a specified set of documents, using this information to supplement information from its pre-existing training data. This allows LLMs to use domain-specific and/or updated information.
* Few-shot: the model is already pre-trained and fine-tuned to become instruct. Few-shot means that in inference in the prompt are put from zero to 4/5 task examples before the request for an actual task.

There are many other techniques, but they are not relevant to this work. In this work we mainly focused on ICL and CoT, building a customized few-shot technique.

## 1.4 Large Language Model Meta AI (Llama)

Llama, also known as LLaMA, is a family of open-source LLMs released by Meta AI. In the last years Meta AI released many models with various parameters size. Furthermore, they released also fine-tuned versions (instruct). The Llama model demonstrated to be stronger than many other models in NLP tasks. Llama is able to generalize patterns and follow detailed instructions better than other models like Phi.

The Llama model architecture is autoregressive decoder-only transformer, similar to GPT-3 with minor differences in activation function, embeddings, and optimizer. So, Llama is based on the transformer architecture, then it is composed of many transformer layers. Transformers use self-attention mechanisms to process input sequences in parallel, allowing for efficient training and high-quality language modelling.

## 1.5 Phi

Phi is a series of small open-source AI models developed and published by Microsoft. There are two types of models: normal and reasoning. Phi models can be also base or instruct depending on the training. There are also reduced versions called “mini”.

The Phi-4 series are strong in reasoning and logics especially in math, due to a mix of unsupervised learning techniques such as multi-agent prompting and self-revision workflows that constitute the “synthetic data” approach.

Thanks to the innovative synthetic data generation methods, this small model can reach stronger performance than others size-equivalent and some bigger models. The synthetic data method is applied also in post training to refine the model output.

The development of Phi-4 is guided by three core activities:

1. Synthetic data (in pre-training and mid-training): thanks to appositely designed datasets of synthetic data focused on reasoning and problem solving, Phi can reach its strong results.
2. Data curation and filtering: the datasets contents are meticulously selected and filtered also for the material taken from the internet to promote educational values and obviously to have good datasets.
3. Post training: here are applied various techniques based on the pivotal token search.

Nevertheless, its strong performance, Phi has some weaknesses mainly due to the model size. Specifically in hallucinations regarding factual knowledge, rigorously following detailed instructions for a specific task and generate the output according to a given pattern.

## 1.6 Hardware Diagnostics Papers with AI

In this chapter, I analyse the SOTA papers and articles about the diagnostics and log analysis.

* “LLM4SecHW: Leveraging Domain-Specific Large Language Model for Hardware Debugging” (paper 2024): This paper realizes, a fine-tune of various models to help in identifying hardware bugs during the design phase. The fine-tune is based on a dataset built on a GitHub repository of machine code.

This system has two workflows, one to identify a bug and another to generate a possible solution. The solution is compared to the one in the dataset, through the Rouge-N F1 score. Since the Rouge metric cannot capture all the LLMs skills, the authors manually evaluated the models replies. The authors experimented fine-tuned and not fine-tuned models. The fine-tuned models have more specific information then they were able to complete the authors tasks in all cases while the not fine-tuned in some cases affirmed, they were unable to complete their tasks, in other cases they just identified the problem without proposing a solution.

In conclusion, with more data is possible to reach better results.

A diagram of a number of numbers

AI-generated content may be incorrect.

* “[Structuring Unstructured Data with LLMs | by Story\_Teller | Medium](https://medium.com/@vivekvjnk/structuring-unstructured-data-with-llms-13a19810174d)” (article 2024): LLMs thanks to their ability of understanding unstructured data can process raw input and transform it in structured output prompts.
* “FD-LLM: Large Language Model for Fault Diagnosis of Machines” (paper 2024): The method proposed in the paper, pre-process the raw signals data before analysing them with LLMs. In particular, focusing on the use of LLMs, this paper builds a multi-class classifier based on LLMs finetune.

After the signal pre-processing, prompt engineering technique is applied. For each signal there is in the dataset a corresponding prompt with information about the task requested, the context so information about the hardware machine and then its state. In this case the finetune is fundamental to teach to a model specific hardware information and fault mechanisms.

The optimization technique LoRA was applied to make the finetune efficient by the authors.

In conclusion, the authors finetuned Llama3-8B, Llama3-8B-instruct, Qwen1.5-7B, e Mistral-7B-v0.2. Llama demonstrated to be the best model between the others in that scenario. In fact, Llama showed a strong generalization ability.

A diagram of a diagram

AI-generated content may be incorrect.

* “LogGPT: Exploring ChatGPT for Log-Based Anomaly Detection” (paper 2023): This paper experiments with the log analysis on two supercomputers datasets (Spirit and BGL), to do a real time anomalies identification with ChatGPT.

Before applying the LLM, the data are pre-processed with the Drain method to collect the information in a structured way.

This paper is based on the prompt engineering and so ICL. A prompt with the task description, optional human knowledge and then the log information: context, events, sequence of events and their meaning.

Some important conclusions of the authors:

* The prompt structure and human knowledge are fundamental to improve the results
* Specifying the cause of an error helps the model in understanding the logic link from the effects rather than simply labelling the errors
* Adding some solved examples to the prompt (few-shot) lead to better performance than without examples (zero-shot)
* ChatGPT has a high false positive rate
* ChatGPT is too much sensible to small changes in the prompt

Overall, LLMs as the other AI methods can identify patterns better than rule-based systems.

A diagram of a log processing process

AI-generated content may be incorrect.A diagram of a process

AI-generated content may be incorrect.

* “LogLLM: Log-based Anomaly Detection Using Large Language Models” (paper 2024): This paper proposes a system that automatically collects logs in a structured way for a real time analysis to identify anomalies with Llama3-8B after a pre-processing based on BERT-base and a projector. BERT-base is used to extract semantic vectors, to be tokenized with the BERT tokenizer. While the projector is used to map the semantic vectors previously tokenized to another tokenized representation understandable by Llama. Finally, Llama is used to do the log analysis.

A custom prompt is used to introduce the log’s tokenized representation and to request the task to Llama.

The whole model system is finetuned in three steps. One to teach to Llama to reply by saying if a log is anomalous or normal. A second to train BERT-base and the projector to find the best token representation for Llama. And finally, a third step in which BERT-base, projector and Llama are finetuned as a single system. The finetunes are done with the QLoRA optimization technique.

In conclusion, this system maintains a low rate of false positives but there are also some disadvantages it has a high computational cost due to the high number of parameters. As this system limits the power of an LLM to a binary classification. The model cannot handle unbalanced scenarios.

The authors experimented also a finetune on the pure logs, but they had a worsening in performance.

A diagram of a projector

AI-generated content may be incorrect.

* “An Automated Machine Learning Approach for Real-Time Fault Detection and Diagnosis” (article 2022): This paper proposes a system for real-time fault detection on industrial machines based on ML. This system uses only commonly available data in industrial systems. The method was tested in a 3D machine simulation system in faulty and non-faulty conditions. The system consists in a classifier trained in supervised learning to identify known faulty situations. The system is able to do re-generation with new situations.

The authors experimented with sixteen classifiers to find the best. They selected Extra trees and Random Forest as final classifiers with an F1 score of 100% and 85% respectively for the two “Furnace” and “Pick And Place” simulated industrial machines.

The authors observed that increasing the examples lead to a good improvement in results. The F1 score reached more than 90% with more samples. Also, thanks to an ablation study on the features that allowed to increase the performance up to 6%.

* “Machine Learning for Anomaly Detection: A Systematic Review” (paper 2021): This paper analyses a collection of other publications in years 2000-2020 about anomaly detection with ML. The authors systematically analysed all the publications with the “Kitchenham and Charters methodology”.

According to the authors, the ML papers can be divided in the following techniques: classification, ensemble, rule system, clustering and regression. Those techniques are applied in two forms as standalone or hybrid (their combination) models. Unsupervised anomaly detection has been adopted more than classification techniques.

In conclusion, the article consists in a comparison of the ML papers in that period. Practically is a comparison of the other methods, it shows for each period which technique was mostly used and their performances. The most used technique in anomaly detection is based on unsupervised learning while the second most used is the classification even if, according to the authors of this analysis, the papers’ authors did not mention the classification type applied in many cases. While the semi-supervised learning technique is rarely used in anomaly detection. Many papers did not mention also the extraction/selection type for the features.

This publication limits its analysis to strictly scientific papers/articles only.

## 1.7 Metrics

The Recall-Oriented Understudy for Gisting Evaluation (ROUGE) evaluation metric measures the recall. It’s typically used for evaluating the quality of generated text and in machine translation and automatic summarization tasks in NLP. This metric and its variations range between 0 and 1, with higher scores indicating higher similarity between the automatically produced summary and the reference.

While the ROUGE-N measures the number of matchings 'n-grams' between a reference (a) and test (b) strings.

# 2. Technological Analysis

Since I had at disposition a GPU with only 6GB of RAM the technological analysis is mainly driven by hardware resources. Consequently, in this work thesis I experimented only models with less than seven billion parameters.

## 2.1 Models’ Comparison

Initially, I compared various models such as Llama, Gemma, Phi, Mistral, DeepSeek and Qwen. I compared them in terms of performances according to different metrics and size due to hardware limitations.

In this comparison, I directly discard ChatGPT due the monetary cost of the API unwanted by the company and the fact that the software is covered by license and not open source.

During the experiments on the GPU done in the [*preliminary activities*](#_3.1_Preliminary_Activities), I noticed the available hardware did not support the model Phi-4-mini-Instruct even with 4-bit quantization. While Gemma-3-4B-it was not compatible with the virtual environment due to some vision libraries dependency. So, I practically experimented with Llama-3.2-3B-Instruct, Phi-4-mini-reasoning and Mistral-7B-Instruct-V0.3.

During the experiments I noticed the model Llama is the best in understanding the natural language — compared to the others cited. In fact, this model is able to easily generalize a sentence.

Instead, Phi-4-mini-reasoning is stronger than the others in logics. This model reached the best results in the preliminary activities. Even if sometimes the hardware did not support it, or the model did not answer to all the questions.

While the last model Mistral-7B-Instruct-V0.3 has the worst performances, probably due to the 4-bit quantization.

For these reasons, I decided to use the Llama model for all phases after the preliminary activities.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Parameters** | **Input-output modalities** | **Context length** | **Provenience** | **Author** |
| Llama-3.2-3B-Instruct | 3B | Text-to-text | 128k | Hugging Face | Meta AI |
| gemma-3-4B-it | 4B | Text/Image-to-text | 128k | Hugging Face | Google DeepMind |
| Mistral-7B-Instruct-V0.3 | 7B | Text-to-text | 32,768 | Hugging Face | Mistral AI |
| Phi-4-mini-reasoning | 3.84B | Text-to-text | 128k | Hugging Face | Microsoft |
| Phi-4-mini-Instruct | 3.84B | Text-to-text | 128k | Hugging Face | Microsoft |

## 2.3 Quantization

It is the process of reducing the memory space used by a number. It reduces the precision by approximating the model weights and then reducing the model size making faster the model usage.

There are two main quantization techniques with various implementations.

* Post- training quantization: it reduces the precision after the model training for the inference usage.
* Quantization aware training: it quantizes the model already during the training and reduces the performance degradation

I used the post-training quantization in this work thesis with the BitsAndBytesConfig class by the transformers library.

### 2.2.1 Bits and Bytes 8-bit Quantization

Starting from float16 matrices X for the input state and W for the weights state. To quantize, values are extracted column-wise from X forming the matrix Cx. While from W are extracted row-wise forming the Cw. Then with the following formula the two matrices can be quantized in 8-bit.

Xi8 = X \* (127/Cx)

Wi8 = W \* (127/Cw)

Then, the two matrices Xi8 and Wi8 are multiplied resulting in Outi32. To obtain a partial output for the inference is applied the following formula.

OutF16 = (Outi32 \* (Cx ⊗ Cw)) / 1272

The remaining values from X and W build two new matrices that are multiplied to obtain a second partial output OutFP16.

The final output matrix is the xor between the two partial outputs.

Out = OutF16 ⊕ OutFP16

## 2.4 Complexity of Implementation

All the models required to install the necessary Nvidia drivers, Miniconda and a Python virtual environment.

All the models can be used through a Python script with their setup and the necessary libraries. During the preliminary activities I used the Python “input” to grasp the user input from a terminal. In particular, for the preliminary activities I experimented the models with the library transformers by Hugging Face.

After, for all the successive phases, I implemented the Open WebUI to grasp the user input. It required the installation of Ollama for the backend to handle models. Then for all the successive phases I used Llama from Ollama instead of Llama by Hugging Face used for the first phase.

# 3. Method

## 3.1 Preliminary Activities

The purpose of these firsts activities is to evaluate the ability of LLMs in understanding the hardware structure of a sample system and its interconnections, and to determine whether representation of the system architecture is more effective in making the model understand the system architecture. This is therefore an initial proof of concept (PoC).

I have experimented three system architecture representations: textual, JSON and tabular representation of the network architecture. All these representations are always in text format since LLMs only understand text.

We have experimented with various models according to the available hardware capacity for this phase, especially meta-llama/llama3.2-3B-Instruct and microsoft/Phi-4-mini-reasoning.

For this work thesis I had a small GPU. Consequently, I had to perform the tests with small models and the use of 8-bit and 4-bit quantization.

To test the models on the understanding of the connection topology (the architecture of the System) I created a dataset of 94 questions, in which it is asked what is connected from port/interface A to B, and vice versa. From this dataset I extract a subsample of questions to ask the model to verify if it has "understood" and remembers the architecture. Since the GPU used in this phase has only 6GB of memory, it is not possible to ask all the questions together.

Since the LLMs answer open questions with generally always different and long text, I taught the model (specifying rules in the system prompt) to answer by indicating only the name of the connected component. In this way it is possible to automatically parse the answers to evaluate the accuracy.

The system prompt consists of a short description of the task that the model must perform and its role, a set of rules on the representation of the architecture and the names of the devices, a list of the devices, the representation itself in its format and finally some examples of questions and answers. This PoC is therefore based on a combination of the prompt engineering techniques shown in the previous chapter.



### 3.1.1 Questions Dataset

I manually built a dataset with 94 questions and corresponding answers about the system interconnections topology. The dataset contains questions like “What is connected to A?” for all ports and interfaces of all devices. So, if A is connected to B there is the dual question (the answer is B when asking about A and vice versa). In the dataset there are also few questions about components that do not exists. The responses are of the type "The C is connected to D", "Nothing" if there is nothing connected or "The interface/port D does not exist for device G".

To calculate the accuracy, a subset of questions is randomly extracted and asked to the model. Due to GPU memory limitations, I could not ask the models all questions at once.

I calculate the accuracy in reporting the asked questions and in correctly responding to questions. All the models can report the questions (the accuracy was always 100%), so they know the request but in answering/reasoning they had more difficulties.

### 3.1.2 Prompt Structure

The prompt is specific for each type of representation. At the beginning of the prompt is defined the model task. In Json and tabular representations, the task description varies slightly as I specify what type of representation is shown.



In the successive part initially, I put the components definitions



After some experiments I removed the definitions because it was not helpful in interconnections understanding and the size of the prompt matter.

In the successive part there are the naming conventions and other rulers depending on the representation and the representation itself.

In the last part of the prompt, I put some Q&A examples. This teaches the model to respond in the specified format. At this point of the work, it is mainly useful to parse the answers and automatically evaluate the accuracy.

Not all models clearly understand that they must respond in the specified way. So, I added a parenthetical note to the prompt to clarify it, but models like mistralai/Mistral-7B-Instruct-v0.3 sometimes still do not follow it, while meta-llama/llama3.2-3B-Instruct already understood it correctly from the start.

Here the last part of the prompt, notice that the last generalized sentence allows you to narrow down the space of possible output values ​​and therefore narrow the field to just your domain of interest. This is powerful.

A screenshot of a computer program

AI-generated content may be incorrect.

### 3.1.3 Textual Representation

In this representation the system interconnections are described in natural language. Before the description of the system, I put some naming rules and a list of all devices.



Here the system representation:



### 3.1.4 Json Representation

In this representation I described the network and its components as Json objects. I introduced this representation in the prompt as there:

A screenshot of a computer program

AI-generated content may be incorrect.

Here an example of one component in Json format, the switch-1:

A screen shot of a computer

AI-generated content may be incorrect.

### 3.1.5 Tabular Representation

As in all the other representation there is a part with rules and the components list:



After the system architecture topology is in a tabular format. The table is manually hardcoded, since it must be text:

A screenshot of a computer program

AI-generated content may be incorrect.

### 3.1.6 Model Dimension Limit

The tests for this part have been done with meta-llama/llama3.2-3B-Instruct model and the textual representation.

In the first experiments I noticed most errors were caused by a similar nomenclature in multi-device components (e.g., Switch-1, Switch-2, Swtich-3). In questions like “What is connected to the Workstation-1-eth0?” the model answered more than one time like “The Switch-3-port-2” while the correct answer was “The Switch-2-port-2”. Then to better highlight the number that makes the devices different I tried a different representation by separating the device name and number from the port/interface with the colon symbol (‘:’).

With that names representation I did the following experimental test with twenty questions:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Device name format | Run 1 | Run 2 | Run 3 | Mean |
| Colon | 85.0% (17/20) | 70.0% (14/20) | 100.0% (20/20) | 85.0% |
| Dash | 90.0% (18/20) | 87.5% (17.5/20) | 100.0% (20/20) | 92.5% |

After this test I did the tests in the section [2.1.7](#_2.1.6_Models_results) (I expanded the context, so more questions, resulting in a worsening of accuracy). With the [2.1.7](#_2.1.6_Models_results) tests it became clear that the problem was due to the size of the model also because the model did not answer correctly to other questions (like “What is connected to ...?”) where the answer was clearly defined in the prompt where the topology is defined. Indeed, a smaller size model has less storage/reasoning capacity especially with a “large” context. To empirically demonstrate this, I did another experiment: I reduced the size of the system (maintaining more questions) resulting in an improvement.

Information that is already available from scientific literature (papers) on diagnostics. However, the company wanted to experiment with it. I also feed the system architecture to larger models like GPT-4 and DeepSeek to verify it again. Obviously, these models correctly understand and answer questions.

Here the smaller system architecture:

A diagram of a computer system

AI-generated content may be incorrect.

To calculate the accuracy for the experiment two I defined a smaller dataset with 34 questions (to test all the interconnections of a smaller system). I feed to the model all the 34 questions for five runs. For simplicity I used only the “dash” representation for device names, also because as I said before it does not really matter. Here the results of the experiment:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Run 1 | Run 2 | Run 3 | Run 4 | Run 5 | Mean |
| 88.2% (30/34) | 94.1% (32/34) | 91.2% (31/34) | 97.1% (33/34) | 94.1% (32/34) | 92.94% |

### 3.1.7 Models Results

#### 3.1.7.1 meta-llama/llama3.2-3B-Instruct

I tested this model with 8-bit double quantization.

This model has been tested on all the representation. In few runs the model did not answer all the questions, so I ran it again in those cases.

In the following table the accuracy in answering the 30 questions with the textual representation:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Device name format | Run 1 | Run 2 | Run 3 | Run 4 | Run 5 | Mean |
| Colon | 86.7% (26/30) | 83.3% (25/30) | 73.3% (22/30) | 73.3% (22/30) | 86.7% (26/30) | 80.66% |
| Dash | 80.0% (24/30) | 83.3% (25/30) | 73.3% (22/30) | 83.3% (25/30) | 90.0% (27/30) | 81.98% |

In the following table the accuracy in answering the 30 questions with the JSON representation:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Device name format | Run 1 | Run 2 | Run 3 | Run 4 | Run 5 | Mean |
| Colon | 80.0% (24/30) | 93.3% (28/30) | 83.3 (25/30) | 80.0% (24/30) | 76.7% (23/30) | 82.66% |
| Dash | 63.3% (19/30) | 83.3% (25/30) | 83.3% (25/30) | 73.3% (22/30) | 90.0% (27/30) | 78.64% |

In the following table the accuracy in answering the 30 questions with the tabular representation:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Device name format | Run 1 | Run 2 | Run 3 | Run 4 | Run 5 | Mean |
| Colon | 63.3% (19/30) | 70.0% (21/30) | 66.7% (20/30) | 80.0% (24/30) | 93.3% (28/30) | 74.6% |
| Dash | 86.7% (26/30) | 80.0% (24/30) | 50.0% (15/30) | 53.3% (16/30) | 63.3% (19/30) | 66.66% |

The tabular representation has the worst performances.

#### 3.1.7.2 microsoft/Phi-4-mini-reasoning

I tested this model with 8-bit double quantization.

This model is focused on reasoning, it has the best performances, but it is a bit bigger (3.84B parameters) than the used llama so after two test with 30 and one with 25 questions the hardware stopped working (always out of memory errors).

Since it is focused on reasoning was better able to understand logical relationships and answer correctly.

In this table I report the results of the few tests (dash textual representation):

|  |  |  |  |
| --- | --- | --- | --- |
| Run 1 | Run 2 | Run 3 | Mean 30 quests |
| 96.7%  (29/30) | 90.0%  (27/30) | 84.0%  (21/25) | 93.35% |

This model is 2x or 3x slower than the llama used due to the reasoning phase. In most other cases it answered but only to 15-20 of 25-30 questions.

#### 3.1.7.3 mistralai/Mistral-7B-Instruct-v0.3

I tested this model with 4-bit double quantization, the hardware did not support the 8-bit quantization. I reduced also the number of questions to 25.

The model struggles to follow the Q&A output format, often ignoring the prompt and answering however it wants. I had to run it multiple times before getting the responses in the desired format. Its poor performance is likely due to quantization or hardware limitations.

In the following table the answers accuracy with the dash textual representation, the hardware didn't hold up with the other representations:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Run 1 | Run 2 | Run 3 | Run 4 | Run 5 | Mean |
| 76.0% (19/25) | 64.0% (19/25) | 72.0% (18/25) | 72.0% (18/25) | 92.0% (23/25) | 75.2% |

### 3.1.8 PoC Conclusion

The hardware available for this phase allowed only these experiments to be performed, however quantizing below 16-bit generally makes neural networks decidedly unstable and extensively imprecise. Despite this, the results obtained have an average accuracy of around 80-90 percent. So, the PoC was successful even if carried out in very precarious conditions. Surely, as expected, having more resources it is possible to make the model “understand” enough the system architecture to then carry out the diagnostics.

## 3.2 Dataset

### 3.2.1 Introduction

In the company the diagnostics activity is almost completely based on the HMI software. That software allows you to load a system architecture by its system definition file that contains a list of the system components and tests to perform. With the HMI you can connect to the agent tester on the device to be tested and request the tests execution. The HMI also allow you to view the results and save them on a PDF or HTML file for a single or all tests.

The HMI application save the tests results on a database. It allows also to save the data in pdf or html format.

[HMI, diagnostics, test]

### 3.2.2 Database Structure

The database of the HMI is composed by six tables with the following scheme. This database structure is maintained also in the simulation environment used for this work:

* acceptanceTest (id, begin, end, open)
* generalInformations (id, infoLabel, infoData): this table contains information about the operator
* note (id, session, testId, infoData): this table contains only notes
* notifications (id, type, testID, session, timestamp, testStatus, testResult, error, updateType, updateValue, report):

this is the most important table, because it contains the notifications exchanged between the agents that perform the tests and the tested system components. The most important fields of that table are these:

* id: integer that represents notification arrival order. It is the primary key with autoincrement
* type: integer that indicates the notification type. It can be 0 = Unknown, 1 = Init, 2 = Running, 3 = Update, 4 = Terminated, 5 = Stopped, 6 = Deleted and 7 Error
* testID: string that uniquely identifies the test of the notification. It is the same of testCode filed of the table testIdVocabulary where is located the corresponding label, it is also the test identifier in the system definition
* seission: integer that represents the ID of the test session or process. So, the number of test executions
* timestamp: hour and date of the notification
* testStatus: integer that represents the actual test state. It can be 0 = Default, 1 = Init, 2 = Running, 3 = Terminated, 4 = Stopped, 5 = Deleted and 6 Error
* testResult: Integer that represents the final test result. It can be 0 = Default, 1 = NotExecuted, 2 = Successful and 3 = Fail
* error: description of an eventual error. It can be empty (the default value is the dash ‘-‘)
* report: This is the most important field, it contains the results of the test and all its subtests
* sqlite\_sequence (name, seq): this table contains the number of instances for each other table
* testIdVocabulary (id, testCode, testLabel, subsystemId, subsystemLabel): this table contains the corresponding names/labels for the system codes/tags.

The most important table is “notifications”, because it contains the notifications exchanged between the agents that perform the tests and the tested system components. The most important fields of that table are these:

* id: integer that represents notification arrival order. It is the primary key with autoincrement
* type: integer that indicates the notification type. It can be 0 = Unknown, 1 = Init, 2 = Running, 3 = Update, 4 = Terminated, 5 = Stopped, 6 = Deleted and 7 Error
* testID: string that uniquely identifies the test of the notification. It is the same of testCode filed of the table testIdVocabulary where is located the corresponding label, it is also the test identifier in the system definition
* seission: integer that represents the ID of the test session or process. So, the number of test executions
* timestamp: hour and date of the notification
* testStatus: integer that represents the actual test state. It can be 0 = Default, 1 = Init, 2 = Running, 3 = Terminated, 4 = Stopped, 5 = Deleted and 6 Error
* testResult: Integer that represents the final test result. It can be 0 = Default, 1 = NotExecuted, 2 = Successful and 3 = Fail
* error: description of an eventual error. It can be empty (the default value is the dash ‘-‘)
* report: This is the most important field, it contains the results of the test and all its subtests

In this work thesis I focused on the last field for the notifications with type 4 (a successfully completed test with result pass/fail) and type 7 (error case).

I modified the base database structure by adding two fields reportLabel and errorLabele. These fields contain a reduced version of the same data of the corresponding fields report and error to reduce the context for the ICL activity.

### 3.2.3 System Definition

The system definition XML file is composed by a tree structure. Inside the “Diagnostics” root element there are the following elements:

* System: defines the system components
* Agents: define the agent testers which perform the tests
* TestCategories, MonitorCategories, StressCategoires: contains a set of test categories for the following elements
* testList: contains all the tests to be performed interactively (a human operator must manually start/stop them)
* MonitorList: contains a set of tests that are iteratively repeated if started by a human operator. In this case it is empty
* StressList: contains a set of tests that can be iteratively executed without the need of a human operator, they can be monitored by MonitorList. In this case it is empty

## 3.3 Scenarios

For the sample system we defined three scenarios:

* In the first scenario the system is correctly working, then all the tests are successful
* In the second scenarios, the switch 2 is completely broken, so all the tests on it and on the linked devices fail
* In the third scenario, the SBC2 is completely broken and so all the tests on it fails

For all the three scenarios we have defined a configuration file, a database with the described structure, a pdf report with a summary of all the executed tests and their results, finally various system prompt with and without few-shot.

The configuration file allows to run the tests in a simulation environment without the need of physical hardware.

## 3.4 ICL

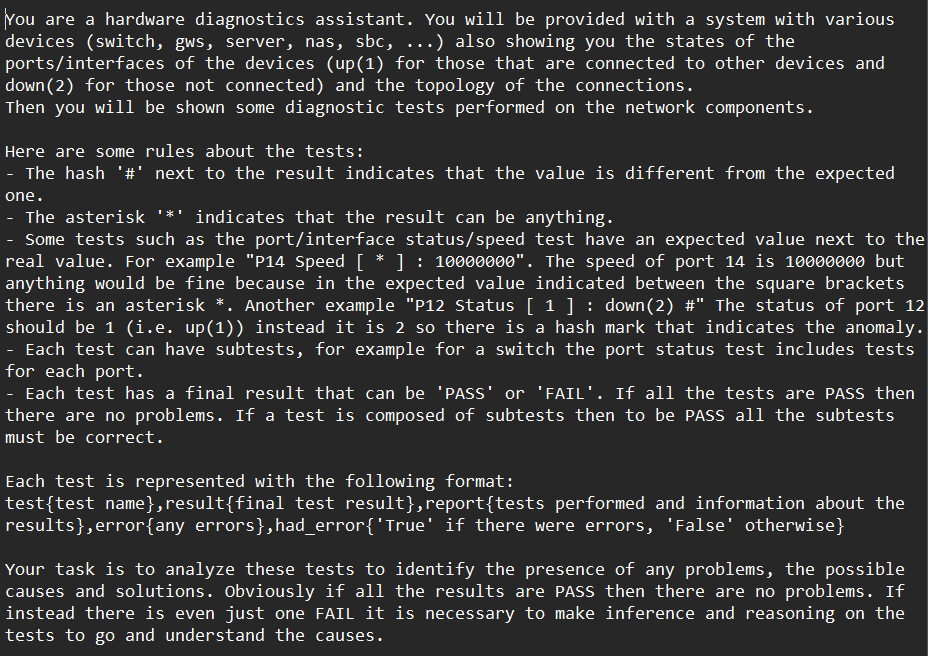
This part of the thesis work is the most important, it consists in teaching to the model how to do a diagnostics task. So, how to analyse the test results and how to infer to identify the causes of any possible issue. For this part of the work, I experimented various system prompt with and without few-shot.

### 3.4.1 System Prompt without Few-shot

In this prompt I described:

* At the beginning, the model role and briefly its task
* In the successive part I explained the rules to interpretate the test results. In particular, the hash ‘#’ near to a test result mean that the test failed (the result is different from the expected), the star ‘\*’ mean that the test can have any result. I put also some examples to let the model understand the pattern. At the end of that section, I put an explanation about the result for a test which is ‘PASS’ if all the subtests are successful and ‘FAIL’ if even just one fail
* In the next part there is the format I decided to use to show each test and its information. I used a Python script to retrieve the tests data from the database and format them
* At the end of the prompt, there is an explanation about its task that is analyse the test results and find the causes of any possible issue

The following are the short “working” and the longer system prompts



A computer screen with text on it

AI-generated content may be incorrect.

Before arriving to the working system prompt version, I did many other system prompts with more details, but they led to a worsening in results. The model was unable to understand its task and correctly answer the questions.

Instead, with this prompt the examples are feed to the model as user input (few-shot). The working prompt had the best results in replying to questions.

### 3.4.2 System Prompt with Few-shot

For each scenario, I realized a system prompt with one-shot. The prompt contains:

* A little description of the model task
* The sample system components with their ports/interfaces status. The status can be ‘up(1)’, ‘down(2)’ or ‘Unknown’
* A textual description of the network topology (the system architecture)
* The tests interpretation rules.
* A description of the tests format
* All the tests with their results in that format
* Finally, the most important part, an example of answer. It is a logical deduction and a results analysis

This prompt is widely longer than the others without examples. All this information overhead the model and lead it to unrelated replies. With two-shot the results are worse, increasing the prompt size results in a performance degradation.

These experiments and colleague discussions lead to define a system prompt without few-shot included.

## 3.5 System Prompt Experiments

I experimented with various system prompts. At the beginning, I tested the system prompt with few-shot included with one example. With this system prompt the model was unable to correctly answer the questions. In particular, the model did not identify the failed tests in many cases, consequently the model did not find the problems and their possible causes.

Subsequently, I experimented with a short system prompt without few-shot. This prompt has the best performance in replying to questions. In many cases, the model correctly identified the issue and their possible causes. Since the model did not identify the correct solution in all cases, I tried to improve the prompt.

Then, I created another longer system prompt with an explanation of how to logically identify any issue. But this prompt led to worse performance. In fact, the model was again unable to identify the failed tests. So, I decided to maintain the previous prompt.

## 3.6 Few-shot Experiments

I experimented two ways to do few-shot. One where I put the example and its solution directly in the system prompt and another where I put the example with its solution in the user prompt. I composed this prompt, with a first part where I briefly repeated the model and the network components and their ports/interfaces status, and a second part where I put all the tests with their results. For this method I used the short system prompt, that led to better performance, as I said before.

# 4. Results

Initially, I “manually” evaluated the model replies. The answers obtained with the longer prompts and with the few-shot prompts was not satisfactory. So, with the enterprise tutor we decided to analyse less tests with the model with a smaller sample system.



Then, I experimented with a shorter prompt having better results but not always correct.

# Bibliography

[] NLP: <https://en.wikipedia.org/wiki/Natural_language_processing>

[] LLMs: <https://en.wikipedia.org/wiki/Large_language_model> and

[https://web.stanford.edu/~jurafsky/slp3/](https://eur01.safelinks.protection.outlook.com/?url=https%3A%2F%2Fweb.stanford.edu%2F~jurafsky%2Fslp3%2F&data=05%7C02%7Cmatteo.gianvenuti%40studenti.unipr.it%7C3eb6fbcc055f4e8a0aff08dd8bd29bee%7Cbb064bc5b7a841ecbabed7beb3faeb1c%7C0%7C0%7C638820458860462928%7CUnknown%7CTWFpbGZsb3d8eyJFbXB0eU1hcGkiOnRydWUsIlYiOiIwLjAuMDAwMCIsIlAiOiJXaW4zMiIsIkFOIjoiTWFpbCIsIldUIjoyfQ%3D%3D%7C0%7C%7C%7C&sdata=R76CP%2Fvpd1I40a8MReCrxBrrjUE6ColECYnknFGpswY%3D&reserved=0)

[] Prompt engineering: <https://en.wikipedia.org/wiki/Prompt_engineering>.

[] Quantization: <https://huggingface.co/blog/merve/quantization>

[] Phi-4: <https://www.microsoft.com/en-us/research/wp-content/uploads/2024/12/P4TechReport.pdf>

[] Kitchenham and Charters methodology: B. Kitchenham and S. Charters, ‘‘Guidelines for performing systematic literature reviews in software engineering version 2.3,’’ *Engineering*, vol. 45, no. 4, p. 1051, 2007, doi: <https://dl.acm.org/doi/10.1145/1134285.1134500>

[] ROUGE, ROUGE-N: <https://learn.microsoft.com/en-us/ai/playbook/technology-guidance/generative-ai/working-with-llms/evaluation/list-of-eval-metrics> and <https://en.wikipedia.org/wiki/ROUGE_(metric)>