Informatica & Economie

Enhancing the Project Definition in Instrument Making Using Large Language Models

Ole Bulters

Supervisors:
Drs. J.B. Kruiswijk & Prof.dr. K.J. Batenburg

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl 18/07/2025

Abstract

This thesis explores how generative AI can support the project definition drafting process in the field of instrument making. The research combines a literature review on project definition best practices and AI applications with qualitative interviews conducted among Dutch academic instrument makers. Thematic analysis of the interviews revealed challenges such as inconsistent intake procedures, vague requirements, and communication barriers. Based on these insights, a project definition framework was developed and implemented as an UML activity diagram. Additionally, the possible options to utilize AI to improve the project definitions were analyzed. This led to the solution to fine-tune a large language model (LLM) using a parameter-efficient approach to assist in drafting project definitions by asking critical questions. The outcomes of this model were rated by experts and yielded an average helpfulness grade of 5,32. No significant difference between the baseline model and the finetuned model could be proven. Therefore, this finetuned model could not yet be considered an aid to instrument makers.

Contents

1 Introduction

Instrument making is a work field that is not commonly known. While the term instrument maker may lead to associations with the building of musical instruments, in this context it refers to someone involved in the design and development of technical or scientific instruments. This field of work plays a vital role in scientific research by designing and building, for example, custom tools, scientific test setups and medical equipment. Despite its importance, the internal processes within instrument-making departments are often informal and rely heavily on the expertise and experience of individual technicians. A particularly crucial step in these processes is the project definition phase. This starting phase is proven to be essential for product success (Xia et al., 2016). A clear project definition ensures that expectations, requirements, and objectives are understood by all stakeholders before production begins. However, defining these requirements precisely and consistently remains a challenge, especially when clients come from diverse academic or technical backgrounds.

This project definition often lacks standardization or involves incorrect requirements resulting in delays and other inefficiencies. This research attempts to close that gap by using the aid of AI and more specifically Large Language Models.

The idea for this research started by a research request of the Leidse Instrumentmakers School (LIS) to Leiden University. This instrument makers school wanted to do more with AI but did not know how. This marked the starting point of this thesis. In the initial orientation phase, the relevance of improving the quality of project definition was confirmed by the LIS, as well as by the FMD (Fine Mechanical Department), which is the internal instrument making branch of Leiden University. The possible use of AI to fulfill this goal was proposed, which is a specific topic that has not been researched yet, underlining the relevance of this research.

1.1 Research questions

The goal of this thesis is to explore how generative AI can improve the project definition drafting process for instrument makers. To accomplish this, the research is structured around one main research question and three subquestions:

Main question: How can generative AI improve the project definition drafting process in the field of instrument making?
Subquestion 1: What are best practices to construct a successful project definition?
Subquestion 2: How can AI be implemented in the workflow of instrument makers?
Subquestion 3: How can an AI model be constructed to assist instrument makers?

1.2 Thesis overview

This thesis is structured as follows:

- **Literature review** In the literature review, first some background about the project definition is explained after which the possible AI methods for a project definition are proposed and finishing with the ways of implementing, training and testing such a model.

- **Research method** In the research method the interviewing approach is explained after which the process of creating the framework is described and finishing with an explanation of the AI model training and testing.
- **Results** In the results section first the framework is created out of the interviews and after this the AI model is trained and tested following the approach out of the research method.
- **Discussion** In the discussion the results are considered on their relevance and importance.
- **Conclusion** In the conclusion the research questions are answered and the thesis is concluded.
- **Appendices** In the appendices there are the interview questions, the interview short answer transcription, the PEFT code, the statistical analysis code and an example output of the finetuned model.

2 Literature review

Before the research can be conducted, a thorough review of the relevant background literature is required. This review will begin by examining the project definition, as it forms the foundation of the research. Following this, the potential applications of artificial intelligence (AI) models for instrument makers will be explored. Finally, the review will consider various approaches to training such models.

2.1 Project definition

Since instrument making is a small and specialized field, there is limited literature available regarding the project definition within this context. To identify best practices, project definitions from other disciplines, such as healthcare and construction management, are examined and used as reference points.

The general definition of a project definition can be stated as the first phase of a project life cycle in which value creation is mainly done (Forgues et al., 2018). The usual case is structured into three stages: the planning phase, the programming phase and the schematic design phase, during which client needs are documented and preliminary design solutions are proposed (Chbaly et al., 2021). In prior research, the connection between a clear project definition and project success has been significantly proven. In the study by Xia, this has been done by setting up a questionnaire consisting of 120 executives in the construction industry measuring the connection between the project definition and project success (Xia et al., 2016). The proven significant importance of the project definition underlines and confirms the relevancy and correct scope of this research.

The next step involves analyzing the best practices and common pitfalls associated with the project definition phase. A key factor contributing to an unsatisfactory project definition is the inadequate conceptualization during this first stage. A lack of clarity of project goals, scope and expected outcome can be traced back to as a typical source for problems later on in the project (Xia et al., 2016). With the aim of developing a framework capable of enhancing the overall quality of project definitions, researchers have identified various methods to support this goal. An example of such a tool is the lean-led design approach. This method was developed by Grunden in 2012 and is based on the existing lean methodology. The lean method is a management philosophy that is build on two pillars: continuous process improvement and respect for people. Lean can be explained as a systematic approach to continuously expose and solve problems in order to eliminate waste within systems (Grunden & Hagood, 2012). These core fundamentals form the basis for the lean-led design approach. The initial application of lean-led design was in the field of hospital design. Building further on the basis of lean, lean-led design is focused on the participation of the actors involved, in which the design teams (in this research the instrument makers) work closely with the client. An interesting way to facilitate a Lean initiative can be through workshops, which are also called *Kaizen* in Japanese. This Japanese translation is relevant since it means ”change for the good” which represents the continuous improvement, which is a core aspect of the Lean-led design (Grunden & Hagood, 2012). The workshop setting is a rapid improvement event, used to resolve complex problems that flow across multiple departments. In this *Kaizen* session the root cause of the problem is analyzed, a future state is proposed and new workflows are experimented to develop a new method to resolve this problem over time. These core values of Lean-led design can be twisted to be applied to the specific case of a project definition for instrument makers. In a recent research by Chbaly, a Lean-led design step-by-step framework for a project definition is proposed. This framework is applied to healthcare related project definitions but can be easily converted to an instrument making version. The framework is described as follows (Chbaly & Brunet, 2022):

1. Definition of a common vision
2. Identification of client needs
3. Development of organizational chart
4. Evaluation of hypothesis
5. Definition of operation modes
6. Development of conceptual and definitive designs
7. Preparation of the transformation and the transition

Another tool to measure and improve project definition success is the PDRI, short for Project Definition Rating Index (ElZomor & Parrish, 2017). This framework was developed in the early 1990s by a research consortium from the University of Texas called the CII. The purpose of PDRI was to offer project teams a structured methodology to establish a well-defined project scope and to assess its completeness and clarity. The CII constructed multiple versions of the PDRI for different industries before making a specific version for small projects in 2013, which could be applicable to instrument making due to their small project nature. The general purpose of the PDRI consists of 3 parts. First, to facilitate a structured planning process to use during the front end planning phase of a project and therefore indicating the key processes involved. Second, to provide a quantitative measure (for example a score) to the level of scope definition of a project. Finally, to correlate the degree of scope definition to key project success factors, enabling stakeholders to make informed decisions about advancing the project to the detailed design and construction phases.

For multiple different use cases the CII developed a different PDRI scoring scheme, typically summing up to 1000 points total. The PDRI system uses a cutoff score which varies per industry use case and measures the viability of a project. These quantitative insights can be utilized for overall project definition quality assurance and risk management inside the project.

2.2 AI possibilities for the project definition

As earlier described, a correct project definition is of great importance for project success. This phase is crucial for aligning stakeholder expectations and clarifying technical goals (Pohl, 2010). According to the FMD, which was the initial starting point of this research, the drafting process of this project definition does not always go as smoothly as intended. This fact makes it an ideal target for AI assistance. The following classification reflects common distinctions found in recent AI literature:

1. **Rule-Based Expert Systems**
   This approach is the oldest out of the options and is based on pre-programmed logic to suggest outputs based on structured inputs. The AI can be seen as a classic ”if-then” logical scheme. The model is mainly effective when the variability in the data is low and when the specifically needed data can be explicitly encoded. The method is highly predictable but still upfront knowledge engineering is needed (Ligêza, 2006). In a project definition context it could for example be used to:

   - Check whether all required fields or metadata are present in a project intake form
   - Validate that certain compliance or safety criteria are met

2. **Retrieval-Based Systems**
   A retrieval-based system works by transferring text data into vector representations (embeddings) and then comparing these vectors in a high-dimensional space to identify most relevant prior cases or documents (Karpukhin et al., 2020). In a project definition context this could for example be used to:

   - Suggest similar past projects based on early project descriptions
   - Help users reuse prior definitions or templates

3. **Traditional NLP Models**
   Natural language processing or NLP is a model that can extract structured data from text input. It serves as a foundational part of many document analysis systems. These NLP models are less flexible than the more advanced LLM models but are computationally lighter and can therefore be more efficient in some cases (Jurafsky & Martin, 2021). In a project definition context this could for example be used to:

   - Extract technical terms
   - Detect vague language in the project definition
   - Convert free-from text into structured summaries

4. **Large Language Models (LLMs)**
   LLMs are transformer-based neural networks that are trained on a massive amount of data to understand and generate natural language. This is a recent innovation which yields a big leap in the possibilities with AI. Examples of LLM base models are GPT, Claude, and PaLM. They can produce human-like responses, interpret unstructured input and can be used in diverse domains. The strength lies in the flexibility of the model: it can handle incomplete

inputs, unclear structure and still yield a correct end result (Radford et al., 2018). A downside is that due to the probabilistic nature of LLMs human-oversight is still needed as a check (Vogelsang & Fischbach, 2024). In a project definition context this could for example be used to:

- Generate clarification questions
- Validate requirements sets
- Rephrasing vague language

Due to the high ability of LLMs to process unstructured input and generate human-like language, LLMs are well-suited to assist in drafting project definitions. Since instrument making is a very specific and sensitive use case the complete drafting process of the project definition by AI is not a logical step. This theory is backed by several studies in related areas that do not give the full drafting authority to the LLM but merely use it as a check and validation tool. First of all, a study by Reinpuld finds that LLMs (e.g. GPT-4o, Claude) can yield high F1 scores by effectively verifying whether textual system specifications fulfill listed requirements. This while their methodology contrasts LLM outputs against rule-based benchmarks, underlining the validation role of the LLM instead of a drafting function (Reinpold et al., 2024). In another research by Vogelsang the caution is emphasized that while LLMs excel at supporting requirements tasks like parsing, classification, and prompting they are not standalone solution generators. They stress the importance of user oversight and task-specific model selection. This is called the human-in-the-loop approach. (Vogelsang & Fischbach, 2024). Therefore a LLM should be used in a supporting role to assist with the project definition while the human stays in charge.

2.3 LLM training and testing

A few factors are important to take into consideration when using and tuning a LLM. First of all, a suitable baseline model needs to be selected. Then this baseline model can be trained/interpreted in multiple ways by including data to learn from. After this training process this model can be tested to verify the relevance of an AI model.

To select a baseline model it is of course of great importance that the model is accurate. In order to test this accuracy of the different LLMs available, researchers recently invented a benchmarking strategy called DocBench to test the performance of the widely available LLMs. DocBench works by taking raw PDF files and accompanying questions as inputs with the aim of generating corresponding textual responses. The PDF documents are extracted from 5 domains, namely Academia, Finance, Government, Laws and News. For each domain the PDF files are downloaded from a respectable predetermined source. The construction of the testing dataset works in a three step proces. First, the documents from various domains are crawled from the predetermined sources. Next, corresponding QA pairs are generated using GPT-4 in combination with input from human annotators. Finally, an automatic filtering step is applied, followed by a manual review to validate the quality of the generated instances. After this test setup is complete, it is tested on 22 different LLM-based document reading systems including GPT, Llama, Mistral, Yi, InternLM, Phi-3, Gemma, ChatGLM3, and Command-R. This research led to the GPT-4 model being the most accurate model in all of the 5 domains with an overrall accuracy of 67,9 and leaving a second place for the Phi-3 algorithm with an overall accuracy of 57,4.(Zou et al., 2024).

After the baseline model selection, this model can be trained on the specific case of project definition validation for instrument makers. There are several methods to do this including prompt engineering, supervised fine tuning and parameter efficient fine tuning (PEFT). To start, prompt engineering is the process of formulating effective instructions for a pre-trained LLM without the necessity to change the underlying model weights. This is a very accessible method of tuning a LLM (Zhao et al., 2025). Prompt engineering is possible in multiple ways. The first is zero-shot prompting. This means that only a description of a task is given without examples of the desired output. With few shot prompting there are multiple inputs and corresponding example desired outputs given to train the model besides a description of a task (Brown et al., 2020). The advantages of prompt engineering are the easy implementation and no need for extensive training while the limitations are the risk of inconsistent responses. Another form of baseline model training is supervised fine tuning or also called instruction tuning. This involves training the model on domain-specific input-output pairs. In this way, the model is taught how to perform tasks by seeing many examples of correct behaviour. In practice, this can be done implementing for example the InstructGPT model. The advantages of instruction tuning are the strong domain adaptation and the ability to facilitate custom output behaviour. A possible limitation is the need for high-quality labeled training data which is labour expensive to make (Zhang et al., 2024). Finally, it is possible to use parameter efficient fine tuning (PEFT). This is a transfer learning method developed to adapt and tweak the parameters of large pre-trained models to fit new tasks and scenarios. By dynamically adjusting the model the effectiveness in performing certain tasks is enhanced while taking in account the important features and requirements of the target task. While updating the parameters, the efficiency increases which therefore reduces the computational cost of the algorithm. A condition is that the model used should be an open weight model, meaning that the parameters of the model can be tweaked. Different PEFT approach algorithms are possible, for example, LoRa, prefix tuning, prompt tuning and adapter tuning (L. Wang et al., 2025). In a research by Lialin, 50 methods of PEFT are compared, resulting in LoRa being one of the more efficient options for resource-limited environments (Lialin et al., 2024). This is relevant due to the limited computing abilities available for this research. The advantages of PEFT are the efficiency, the scalability and the ability to work with limited data supply. The limitations are the high level technical setup requirement and the need to still have task-specific examples.

To summarize:

Strategy	**Training Required**	**Data Requirement**	**Cost**	**Customization Level**
Prompt Engineering	No	Low	Low	Low–Medium
Instruction Tuning	Yes (Full Fine-Tuning)	High	High	High
PEFT	Yes (Partial)	Medium	Medium	High

Now if the model is trained and tweaked, it is important to test and evaluate it to confirm the working of the model. This testing could be done in multiple ways. To start a performance evaluation could be performed which is the intrinsic evaluation. This evaluates how the model performs the targeted task. This can for example be done by F1 score for classification tasks and the BLEU / ROUGE / METEOR metrics for text generation quality (Celikyilmaz et al., 2021). Next, an extrinsic evaluation can be done as well. This evaluates how well the model supports the real-world task. There are multiple ways to perform this, one is to test the quality with the opinion of experts (instrument makers). Another option is to use a scoring framework such as the PDRI as defined in section 2.1 (Xia et al., 2016). And finally, to compare the effect of AI assistance, a test is needed to find whether the differences in test scores are statistically significant. This can be done with an independent samples t-test (data must be normally distributed) in order to test, for example, mean PDRI scores. If improvements are categorized (for example "correct" or "incorrect), the significance can be tested using the chi-square test (Lakens, 2013).

3 Research method

In order to utilize AI to improve the project definition for instrument makers, and therefore answer the research question, several steps are needed to reach a working end model. Before building and training this AI model an activity diagram will be created consisting of the best practices in the project definition drafting process for instrument makers. This framework will be based on the prior knowledge as described in the literature review and customized to the case of instrument makers. There is very limited information available about instrument makers so further investigation is necessary. This has been done by conducting interviews with instrument makers which is described in more detail in section 3.1. Out of these interviews, the common thread will be identified and the insights drawn from it will be analyzed and incorporated into the research. This will than be incorporated into an activity diagram, which is described in section 3.2. This diagram will be an important aspect since it serves as the bridge between the qualitative interview data and the functional AI model.

After this drafting process is complete, the AI model will be trained using PEFT, which is further described in section 3.3. This trained model will spit out critical questions about a given project definition. The quality of this model will be tested by using an adjusted version of the PDRI which will be further explained in section 3.3.

3.1 Interviews

In order to gain a better insight into the somewhat unknown and unresearched field of instrument making, the first step in the research was to investigate how the FMD (Fine Mechanical Department) of Leiden University operates. The FMD is an internal instrument making division of the university, where researchers can have their test setups and scientific instruments fabricated. This initial insight gave a strong impression into the instrument making world. However, a broader view is needed to get a good grasp of the best practices for the project definition in the instrument making world. Given that most universities in the Netherlands maintain such an in-house instrument-making facility as the FMD, they form a logical and accessible sampling group for this research. The academic instrument making branches in the Netherlands are united in a partnership called the UPTO (Upto, n.d.). To get a diverse view on the instrument making workfield, all the 19 members of the UPTO have been invited for an interview. In this standardized interview the participants will be asked 10 pre-determined questions that can be found in Appendix A. These questions have been formulated based on the information given by the FMD. The questions range from which project management techniques are used, to the perceived importance of a well-defined project definition, and how they see the potential of an AI-based tool for assisting in the project definition drafting process. From these interviews the shortened essential answers will be recorded in a table added in Appendix B.

3.2 Framework creation

In order to transfer the qualitative interview data into a structured workflow in the form of an activity diagram, a recognized analysis method should be chosen. These qualitative analytic methods can be roughly split into two options. The first are tied to or are derived from a specific theoretical position. In these models there is limited room for change in how this method is applied. Examples of these models are grounded theory and discourse analysis. The second camp are methods that are in essence independent from theory and can therefore be applied for a broader range of approaches. A model from this second division is thematic analysis. This method is known for its high flexibility due to the theoretical freedom of it (Braun & Clarke, 2006). The flexibility and methodological independence of thematic analysis is an important aspect for this research since the goal (creating a practical of framework for the project definition) is not directly theory related. In contrast, the grounded theory approach is focused on theory construction which is not the goal of this research. Research by Nowell also underlines the high ability of the thematic analysis to identify patterns correctly (Nowell et al., 2017). Due to this flexibility and pattern recognition ability the thematic analysis will be the method of choice.

The framework creation with the thematic analysis approach is a multi-step process. It starts with the transcription process. In this case it is done with the table in Appendix B. Next the meaningful insights need to be coded using open coding to highlight the meaningful units. After this similar ”codes” are grouped into broader themes. These themes will be translated into steps or checkpoints in the workflow. These steps will be compared and if relevant be combined with the project definition framework from the literature review (Braun & Clarke, 2006). This plan will then be translated into an activity diagram by using the Unified Modelling Language (UML) where each node represents a key step or decision point and each transition is a flow of actions (Booch et al., 2005).

3.3 AI model training

As described in the literature review section 2.3 there are multiple options to finetune a baseline model. If these models are compared, the main pillars of comparison are the requirement of training, the data requirement, the cost and the customization level. For the case of this research for project definitions, training is required to get a good level of understanding of the structure in the specific field of instrument making. This makes prompt engineering an insufficient option since no training is involved. Secondly, the data input requirement cannot be immense due to the small nature of the instrument making world and therefore the limited supply of project definitions available. The cost of computation should be preferably not be too high for this use case since there is no access to high-level computing facilities for this research and this is also not viable for small organizations like instrument making branches. Lastly, the customization level should be as high as possible since instrument making is a very specific field and therefore needs a highly customized model to compete in this technical and specific world. Adding these factors up, the best choice for this research seems to be the PEFT (Parameter Efficient Fine Tuning). Although this choice for PEFT may simplify a complex variety of methods, it reflects the practical realities of this research context.

First of all a baseline model needs to be chosen to be finetuned by PEFT. This model should be an open weight model due to the need for parameter tuning. In literature review section 2.3 the different models are compared through the DocBench approach. The most accurate model is the GPT-4 model. However, this is not an open weight model. The second most accurate is the Phi-3 model, which is open weight and will therefore be the model used for this research.

Due to the computational limitations, the LoRa approach will be used, which is accessed through the Huggingface transformers library. This can be combined with the Huggingface PEFT library to train the model in Python (Mangrulkar et al., 2022). Training data can be uploaded in the form of CSV or JSON files and need to be structured in the form of an instruction, input and desired output. For this research, the instruction will be to ask critical questions about the input project definition while using the framework that results out of the interviews. This instruction variable is provided in the results section since the framework is needed for this. This variable will be the same for all training cases. Both the input cases (an example project definition) and the output cases (critical questions) are supplied by the interview participants and the FMD. Due to the relatively low number of input options self-instruct is used to enlarge this dataset. This method uses LLMs to reproduce similar input and output data to train the model. Recent research by Wang yielded a 33% increase in performance over the original model (Y. Wang et al., 2023). For this research, the GPT-4 model is used to generate this dataset due to the results out of DocBench where GPT-4 is proposed as the most accurate model. With this self-instruct method, the orginal data is enlarged to 80 input/output pairs for training and 20 input project definitions for testing afterwards. An example of a self-instructed input/output pair is provided below:

Input: Project: Dual-Beam Spectroscopic Ellipsometer with Real-Time Feedback Loop
Project code: 3000000008
Client: Dr. M. de Vries (LIC-Instrumentation)
Projectmanager: P. Jansen
Author: M. Novak
Date/version: 03 July 2025
Date last mod.: 03 July 2025
Status: in progress

Content:
Brief introduction:

This project aims to design and develop the ”Dual-Beam Spectroscopic Ellipsometer with Real-Time Feedback Loop”, a critical component for advanced physical and chemical experimentation. The current limitations in precision, environmental control, or integration compatibility present barriers for several ongoing research efforts at LIC. The deliverable must bridge engineering constraints with experimental goals and adhere to institutional safety and standardization protocols.

Project definition:

The objective is to create a working prototype within the defined research constraints:

- Mechanical and thermal design within laboratory scale tolerances

- Interface with existing DAQ or microscope infrastructure

- Integration of safety interlocks and modular parts for future upgrades

- Documentation for manufacturing reproducibility and operational SOPs

A feasibility phase will validate material selection, interface calibration, and environmental stability. A detailed design phase will then proceed using SolidWorks and LabVIEW integration diagrams. Fabrication and test phases will follow pending stakeholder review.

Output:

- **Project Intake & Vision Alignment:**
  - Is the scientific problem and associated measurement challenge clearly described?
  - How does this instrument contribute to existing or planned research efforts?
- **Clarify Requirements & Scope:**
  - Are precision and sensitivity targets defined quantitatively?
  - Are there constraints from existing lab infrastructure (mounting, wiring, software)?
  - Are calibration procedures and performance benchmarks anticipated?
- **Feasibility Check / Initial Design Draft:**
  - Have materials and manufacturing feasibility (thermal, mechanical, cost) been considered?
  - Are system-level interactions (thermal drift, EMI, noise sources) modeled?
- **Identify Key Constraints (Time, Budget, Safety):**
  - Is the prototyping budget sufficient to meet spec requirements?
  - Are safety protocols for cryogenics, lasers, or high voltages integrated?
- **Stakeholder Review & Clarification:**
  - Have end-users (e.g. PhD researchers, technicians) reviewed the interface layout?
  - Are stakeholder expectations for longevity, maintenance, and data handling aligned?
- **AI-Supported Review (LLM critical questions):**

  - Is there ambiguity in design tolerances or test conditions?
  - Are any dependencies not under project control (e.g. vendor parts, licenses)?

- **Iterate on Project Definition if Needed:**
  - What parts of the current scope require risk-based revision?

- **Final Approval & Transition to Design Phase:**
  - What experimental test plan or validation is needed for approval?
  - Are roles and responsibilities during fabrication and integration phase clearly assigned?

Now the code can be constructed using the Huggingface PEFT documentation (Mangrulkar et al., 2022). The base will be a LoRa algorithm with the standard values of parameters. The training process will be optimized for using it on Apple Silicon chips since it will be trained on a Apple computer. Now the ”phi-3-mini-4k-instruct” model will be loaded in. After the model is loaded, it is possible to see how much of the model can be trained and altered by the finetuning process by running the `model.print_trainable_parameters()` command. This results in the fact that the model contains 3.825.798.144 parameters of which 4.718.592 are trainable, translating to 0,12% of the model. After this the dataset is formatted and loaded in. Now the training arguments are determined, which will be based on the standard settings provided by the PEFT documentation and slightly down scaled to perform best on a laptop. The training process can now begin and the model can be saved locally if finished. The full code can be found in appendix C.

3.4 AI model testing

After the AI model is trained it can be used by entering a project definition in the model which will then spit out critical questions about this specific definition. However, empirical evidence is needed to prove the relevance of this model. This will be both for the relevance of the finetuned model in comparison to the baseline model as for the quality of the produced questions in general. For this research the choice will be to do this using expert opinions.

This will be done by entering the 20 test project definitions gathered from the self-instruction into both the trained model and the untrained regular Phi-3 model. The untrained model will be given the same instruction that the trained model is given during training. After this, the proposed questions are given to an expert, which in this case will be a group of multiple instrument makers. The questions will be randomly given to a participant, who will rate the quality of these critical questions on a scale from 1-10. After this, the average score for both the trained model and the baseline model can be calculated. On this data a paired t-test will be done if the data is normally distributed and a Wilcoxon signed-rank test if the data is not normally distributed. The normally distribution test will be done using Shapiro. This test for significance will be done with a p-value of 0,05. All statistical tests will be performed using Python. In this way there can be tested if the newly trained LLM yields a significant better result than the baseline model (Agresti, 2018).

4 Results

4.1 Interviews

For the interviews with the members of the UPTO 6 of the 19 organizations participated in an online meeting while answering the questions provided in appendix A. The shortened answers of these questions can be found in appendix B. In order to check for consensus, the general opinion is visualized where possible in the following table:

Q#	Topic	Consensus	Summary
1	Intake process	Partial	Most begin with walk-ins or email followed by a meeting; smaller projects vary.
2	Standardization	No	Some use flowcharts or templates; others have no standardization at all.
3	PM techniques	No	Few use formal project management methods; approaches vary widely.
4	Deliverables	No	Wide variation; some use requirement sets, others deliver feasibility studies or prototypes.
5	Miscommunication	Partial	Most acknowledge occasional miscommunication, especially on assumptions or expectations.
6	Impact of poor definitions	Strong	Clear agreement that poor definitions lead to project issues.
7	Common mistakes	Partial	Incorrect assumptions and vague requirements are recurring problems.
8	Conservativeness	Strong	General agreement that the field is conservative and resists standardization.
9	AI use	Strong	AI is not currently in use at any of the interviewed organizations.
10	AI potential	Strong	Almost all see value in AI if it's user-friendly and supportive (not replacing human judgment).

Now we can start the thematic analysis process as described by Braun (Braun & Clarke, 2006). This starts with the transcription process which already have been done. Now the meaningful insights need to be coded using open coding to spot important themes. This process can be seen as marking all the the key aspects with a marker. The next step is to group these markings or ”codes” in a overarching theme. Now the most important themes, thus the most common, can be chosen which will then serve as the core of the framework. These themes are: lack of standardization, communication barriers, estimation errors, high impact of poor definitions, resistance to change and AI as a support tool.

These found themes, which serve as the common thread out of the interviews, can now be combined with the frameworks and literature proposed in the literature review to construct an activity diagram. We start with the same 7 steps out of the lean-led design approach and alter them where needed. Below the changes with explanation:

1. **Project Intake & Vision Alignment**
   Based on lean-led design step 1: "Definition of a common vision" is focused more towards the actual intake process due to the lack of standardization.

2. **Clarify Requirements & Scope**
   Based on lean-led design step 2: "Identification of client needs" is leaned more towards hard requirements and scopes since the deliverables are often vague and not well set.

3. **Feasibility Check / Initial Design Draft**
   Based on lean-led design step 6: "Development of conceptual and definitive designs" is altered more towards the feasibility check as this is described as an essential process by DEMO Delft.

4. **Identify Key Constraints (Time, Budget, Safety)**
   This step is not based on lean-led design and comes straight out of the interview analysis since the underestimation of time and cost is a common problem which should be controlled more in the starting phase.

5. **Stakeholder Review & Clarification**
   This is to counter the medium level of miscommunications experienced. In this phase the stakeholders are identified and the project is rehearsed and clarified.

6. **AI-Supported Review (LLM critical questions)**
   In this step the proposed AI model is used to serve as a review in the form of critical questions.

7. **Iterate on Project Definition if Needed**
   Allows for revisions based on stakeholder or AI feedback, similar to lean workshops or Kaizen cycles.

8. **Final Approval & Transition to Design Phase**
   Based on lean-led design step 6: "Preparation of the transformation and the transition" but altered to the instrument making case due to the design nature of instrument making.

As proposed this can be translated in an activity diagram which results in the following:

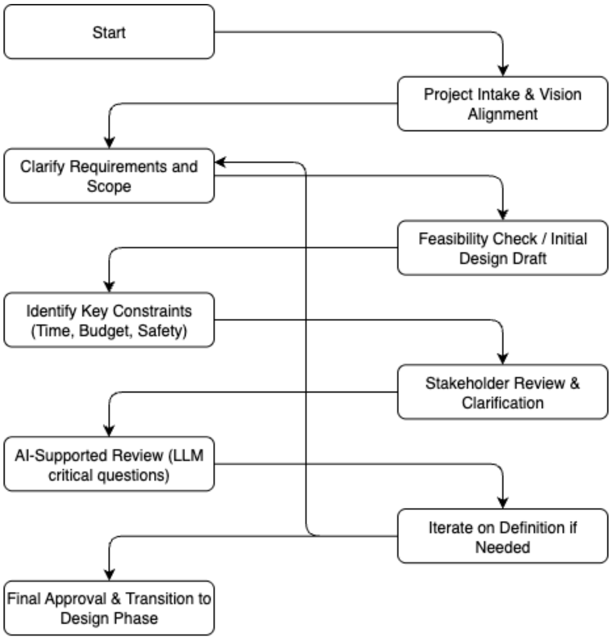
Start
Project Intake & Vision Alignment
Clarify Requirements and Scope
Feasibility Check / Initial Design Draft
Identify Key Constraints (Time, Budget, Safety)
Stakeholder Review & Clarification
AI-Supported Review (LLM critical questions)
Iterate on Definition if Needed
Final Approval & Transition to Design Phase

Figure 1: Instrument making project definition framework

4.2 AI model

In order to train the baseline model, which is picked using the DocBench measurement described in the literature review, the dataset is build up out of 3 variables per instance. To start, the instruction variable, which will be the same for all instances. It gives an instruction to the model what to do with the input and output variables. For this research this is:

```
"Ask critical questions about the following project definition. Use the following framework
1. Project Intake & Vision Alignment ->2. Clarify Requirements & Scope ->3. Feasibility
Check / Initial Design Draft ->4. Identify Key Constraints (Time, Budget, Safety ->5.
Stakeholder Review & Clarification ->6. AI-Supported Review (LLM critical questions)
->7. Iterate on Project Definition if Needed ->8. Final Approval & Transition to Design
Phase"
```

Now a multi step approach is initialized:

- **Step 1:** Combine the provided project definitions (input) and the provided critical questions (output) with the above instruction variable into a JSONL file. This data is provided by the FMD and the interviewees and consisted of 5 input/output pairs. There were no more inputs and outputs available due to the low number of standardized work in instrument making.
- **Step 2:** Enlarge this dataset using self-instruct to 80 instances with the GPT-4 model. The prompt used is ”Given this testdata, expand it using selfinstruct but keep the instruction variable the same. Make sure the style and subject of the project definitions are similar and it yields somewhere the same length. Supply the file in JSONL”.

- **Step 3:** Construct a test dataset by using project definitions provided without output pairs by the FMD and the interviewees. It does include the instruction variable. These definitions are different from those used in step 2. Also enlarge this to 20 instances using self instruct. There were 4 project definitions given.

- **Step 4:** Load this the training JSONL file out of step 2 into the training code provided in appendix C. Now a trained model is available.

- **Step 5:** Now iteratively perform the test project definitions out of the test dataset from step 3. Also test these on the baseline model. This leads to 40 outputs. In this case 2 of them were unusable due to a model error leading to 38 outputs from which 19 are basemodel and 19 finetuned. An example input/output result pair is enclosed in appendix E.

- **Step 6:** Add the input/output pairs to a survey tool and send out to the FMD and interviewees. This resulted into 3 answers. The results will be analyzed below.

First, the difference in mean scores can be measured. This is done with Python (code in Appendix D). It leads to the difference between the finetuned model and the baseline model being -0,1053 point on the scoring of the models. The mean score on the baseline model is 5,35 and the mean score on the finetuned model is 5,32.

In order to see if the difference between the baseline model and the finetuned model are significant, a significance test will be performed. This significance will be tested using a paired t-test if the test data is normally distributed or a Wilcoxon signed-rank if not normally distributed. This distribution will be tested using the Shapiro test. With this test a p-value >0,05 will be considered normal distributed and p-value <0,05 will be considered non-normal distributed. This test is performed using the statsmodels package in Python and can be seen in appendix D. This yields a p-value of 0,6369 and can therefore be seen as normally distributed, so a paired t-test will be performed.

This significance test is done in multiple steps. It starts with the assumptions. It can be assumed normally distributed due to the Shapiro value. Now the hypotheses can be made, which are as follows:

Null Hypothesis ($\mathbf{H_0}$): There is no significant difference in helpfulness ratings between outputs generated by the fine-tuned model and those generated by the baseline model.

Alternative Hypothesis ($\mathbf{H_1}$): Outputs generated by the fine-tuned model will receive significantly higher helpfulness ratings than those generated by the baseline model.

The test will be performed on a p-value threshold of 0,05.

The implementation of the paired t-test leads to an p-value of 0,742 after which there can be stated that the test failed to reject the null hypotheses and there are therefore no significant differences between the finetuned model and the baseline model in the test data.

5 Discussion

This research aimed to explore how generative AI can support the project definition process in the field of instrument making, using a combination of interviews, thematic analysis, and AI model development. The interviews reached a strong consensus on a few topics, namely the fact that poor definitions lead to project issues, the fact that the instrument makers world is a conservative field, the lack of AI use and the anticipated added value of an assisting AI model. This confirms the statement out of the literature review that the project definition is of great importance for project success (Xia et al., 2016). The other results coming out of the interviews are not findings that have been researched yet and can therefore be considered new results.

These results have been combined with the literature review components into the proposed framework. Each step in the framework, was directly informed by recurring issues in the interviews, such as vague requirements, unstructured communication, and a lack of standard documentation. Compared to frameworks like Lean-Led Design (Grunden & Hagood, 2012), the proposed approach is significantly more lightweight and customized to the needs of small, technically focused teams. Its sequential yet flexible design allows for iterative adjustments and human oversight, matching well with the conservative nature of the instrument-making field. The inclusion of AI as a support element, rather than a decision-maker, also reflects both the potential and limitations of technology and the adaption willingness of instrument makers in this context.

This framework has been implemented in a proof of concept in the form of a finetuned AI model. This model failed to yield a significant difference between the baseline model and the finetuned model. In fact, the baseline resulted in a slightly better average helpfulness rating. Besides the absence of significance, both the models led to unsatisfactory results, namely an average grade of 5,35 for the baseline model and a 5,32 for the finetuned model. These failed results can be addressed to multiple possible causes. To start, the choice of the Phi-3 model has now been done based on the DocBench approach, while more extensive research may result in a better fit. Next, the minimal input provided due to the low number of available project definitions is not optimal for training a model. Furthermore, the efficiency of the self-instruct method may have been overestimated. Lastly, the impact of the instruction prompt provided may not have been optimal and could therefore have directed critical questions in the wrong direction. On the testing site, it would have yielded more precise results if the survey could have yielded higher response rates.

Furthermore, the choice for PEFT has been done based on a few key characteristics of the available techniques. In future research this can be researched more thoroughly to make a more grounded choice for a specific method. Some of the options such as, for example, prompt engineering are now discarded quite easily while it may yield better results than expected beforehand. An approach with prompt engineering skips the finetuning aspect which may have been an incorrectly assigned strong requirement for this research.

In future research, the proposed framework could be tested in a real world scenario to validate the working of it. Secondly, the AI model introduced could be trained with a higher number of project definition to reach a more accurate result. Besides this, more AI implementations for instrument makers can be made, for example a tool to assist in the administrative processes. This was mentioned in the interviews as a great addition to an AI product. It is also a possibility to broaden the research to commercial instrument making companies and implement interviews in that work field.

6 Conclusion

To conclude, this research had the goal to answer the research question: "How can generative AI improve the project definition drafting process in the field of instrument making?" This has been done by first incorporating a literature review which served as a background on project definition best practices and the possible AI approaches. This therefore answered subquestion 1. After this, interviews with instrument makers have been conducted which were analyzed using thematic analysis. This resulted in a framework specifically for instrument makers which could be used as a bridge between this qualitative interview data and the functional AI model.

The interviews revealed several findings in the current project definition practices of instrument makers. A key finding was the lack of standardization in how projects are initiated and documented. Most organizations rely on informal processes, often dependent on the individual experience of the instrument maker rather than a structured approach. Additionally, communication gaps between scientists and technicians were frequently mentioned, with unclear requirements and unspoken assumptions leading to misunderstandings or delays. Despite these challenges, there was a strong consensus that a well-defined project at the outset leads to smoother workflows and fewer iterations. Furthermore, while AI tools are not currently in use within this domain, nearly all interviewees saw potential value in AI assistance, especially if it could support, rather than replace, the existing human-centered workflow and was easy to use within their established practices.

Continued on these findings, a custom project definition framework was made to address the specific needs of instrument makers. The framework integrates both literature-based best practices and themes identified through the interviews, resulting in a structured yet lightweight process. It emphasizes vision alignment, clear requirement definition, stakeholder communication, and iterative refinement. Notably, it includes a step for AI-supported review, ensuring that LLMs are used to assist, rather than automate, the drafting process, aligning with the field's conservative and human-driven nature.

This framework was used in combination with example project definitions and example critical questions to train a finetuned AI model. This working of this model has been tested by feeding both a baseline model and the finetuned model example project definitions and letting these outputs be rated by experts (instrument makers). After this, subquestion 2 and 3 have been answered. However, the finetuned model did not yield significant better results. In fact, the baseline model was slightly better rated than the finetuned model. The average helpfulness ratings are also unsatisfactory. The mean grade for the finetuned model is namely 5,32. With these findings it can be concluded that the current model is not yet helpful to instrument makers and should be implemented differently to serve as an aid for instrument makers. Therefore it is not yet possible with this research to succesfully answer the research question: How can generative AI improve the project definition drafting process in the field of instrument making?

References

Agresti, A. (2018). *Statistical Methods for the Social Sciences 5th edition.*

Booch, G., Rumbaugh, J., & Jacobson, I. (2005). *Unified Modeling Language User Guide, The (2nd Edition).*

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners.

Celikyilmaz, A., Clark, E., & Gao, J. (2021). Evaluation of Text Generation: A Survey.

Chbaly, H., & Brunet, M. (2022). Enhancing Healthcare Project Definition with Lean-Led Design. *Sustainability*, *14*(3), 1588. https://doi.org/10.3390/su14031588

Chbaly, H., Forgues, D., & Ben Rajeb, S. (2021). Towards a Framework for Promoting Communication during Project Definition. *Sustainability*, *13*(17), 9861. https://doi.org/10.3390/su13179861

ElZomor, M. A., & Parrish, K. (2017). *Development of the Project Definition Rating Index (PDRI) for Small Infrastructure Projects* [Doctoral dissertation]. Arizona State University. https://www.proquest.com/dissertations-theses/development-project-definition-rating-index-pdri/docview/1947583748/se-2?accountid=13598%20https://media.proquest.com/media/hms/ORIG/2/baz5K?_a=ChgyMDI1MDUwNjEzMTc1NDk1NDo4NDExNjESBTk4MjY0GgpPTkVfU0VBU
2BgIBToIDA1dlYooDHENJRDoyMDI1MDUwNjEzMTc1NDk1NDoxNzI1OTA%3D&_s=Oni3S3RVUgg%2BrK%2BRohCn%2BJcyLto%3D%20http://eur.on.worldcat.org/atoztitles/link?sid=ProQ:&issn=&volume=&issue=&title=Development+of+the+Project+Definition+Rating+Index+%28PDRI%29+for+Small+Infrastructure+Projects&spage=&date=2017-01-01&atitle=Development+of+the+Project+Definition+Rating+Index+%28PDRI%29+for+Small+Infrastructure+Projects&au=ElZomor%2C+Mohamed+A.&id=doi:

Forgues, D., Brunet, M., & Chbaly, H. (2018, September). Lean-Led, Evidence-Based and Integrated Design: Toward a Collaborative Briefing Process. https://doi.org/10.1007/978-3-030-00560-3{_}11

Grunden, N., & Hagood, C. (2012, March). *Lean-Led Hospital Design*. Productivity Press. https://doi.org/10.1201/b11766

Jurafsky, D., & Martin, J. (2021). Speech and Language Processing (3rd ed. draft). https://web.stanford.edu/~jurafsky/slp3/

Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W.-t. (2020). Dense Passage Retrieval for Open-Domain Question Answering.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00863

Lialin, V., Deshpande, V., Yao, X., & Rumshisky, A. (2024). Scaling Down to Scale Up: A Guide to Parameter-Efficient Fine-Tuning.

Ligêza, A. (2006). *Logical Foundations for Rule-Based Systems* (Vol. 11). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-32446-1

Mangrulkar, S., Gugger, S., Debut, L., Belkada, Y., Paul, S., & Bossan, B. (2022). PEFT: State-of-the-art Parameter-Efficient Fine-Tuning methods.

Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic Analysis. *International Journal of Qualitative Methods*, *16*(1). https://doi.org/10.1177/1609406917733847

Pohl, K. (2010). *Requirements Engineering*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-12578-2

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving Language Understanding by Generative Pre-Training* (tech. rep.).

Reinpold, L. M., Schieseck, M., Wagner, L. P., Gehlhoff, F., & Fay, A. (2024). Exploring LLMs for Verifying Technical System Specifications Against Requirements. http://arxiv.org/abs/2411.11582

Upto. (n.d.). Upto.

Vogelsang, A., & Fischbach, J. (2024). Using Large Language Models for Natural Language Processing Tasks in Requirements Engineering: A Systematic Guideline.

Wang, L., Chen, S., Jiang, L., Pan, S., Cai, R., Yang, S., & Yang, F. (2025). Parameter-efficient fine-tuning in large language models: a survey of methodologies. *Artificial Intelligence Review*, *58*(8), 227. https://doi.org/10.1007/s10462-025-11236-4

Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., & Hajishirzi, H. (2023). Self-Instruct: Aligning Language Models with Self-Generated Instructions.

Xia, B., Xiong, B., Skitmore, M., Wu, P., & Hu, F. (2016). Investigating the Impact of Project Definition Clarity on Project Performance: Structural Equation Modeling Study. *Journal of Management in Engineering*, *32*(1). https://doi.org/10.1061/(ASCE)ME.1943-5479.0000386

Zhang, S., Dong, L., Li, X., Zhang, S., Sun, X., Wang, S., Li, J., Hu, R., Zhang, T., Wu, F., & Wang, G. (2024). Instruction Tuning for Large Language Models: A Survey.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., ... Wen, J.-R. (2025). A Survey of Large Language Models.

Zou, A., Yu, W., Zhang, H., Ma, K., Cai, D., Zhang, Z., Zhao, H., & Yu, D. (2024). DOCBENCH: A Benchmark for Evaluating LLM-based Document Reading Systems.

Appendix A: Interview questions

1. How does the application process for an instrument-making project work?
2. How is the intake and project definition process standardized?
3. Are specific project management techniques used in this drafting process?
4. What are the final deliverables/results of the project definition phase?
5. Is miscommunication between scientists and instrument makers considered a common issue?
6. To what extent does a poorly defined project lead to a worse end result or a more difficult process?
7. What are common mistakes made during the project definition phase?
8. To what extent is the instrument-making field perceived as rigid and conservative, making it resistant to change and less standardized?
9. Is AI currently being used in the instrument-making process?
10. Do you see added value in an AI-based product for the project definition phase?

Appendix B: Interviews

Table 1: Part 1 (Q1–Q5)

Continued on next page ⟹

Organisation	Question 1	Question 2	Question 3	Question 4	Question 5
Wageningen Technical Solutions	By walk-in or email followed up by introduction meeting, from which a proposal is made	There is a flowchart which is used as a step by step plan. This is mainly used for big projects, for smaller cases it may be skipped	Most projects are done individually and therefore the project management is an individual responsibility. So no specific project management techniques.	This differs, party due to the fact that people have a different educational level and work method. In practice this leads to everyone implementing their own method with no standardized deliverable but a set of requirements are almost always present.	No strong miscomminication but the researcher may push his own vision true while neglecting the input from the instrument makers. Slight details may be overseen due to too many assumptions.
Nederlands Herseninstituut	Due to the small organisation the intake process is fairly straightforward and directly onto the person in charge.	No the process is not standardized. There used to be a form that would be filled in but that is not relevant anymore.	Project management techniques are barely used. SMART working is introducted sometimes but no constant technique.	Some small requirements are made, but no official project definition phase is used. The approach is described as informal.	Yes this is a problem. Both in the field of level difference between the researcher and instrument maker and in the field of a difference in expectations.
Ontwikkeling Medische Technologie UMC Utrecht	By walk-in or email followed up by introduction meeting, from which a proposal and quote is made	Specifically the instrumentmakers only take on projects with a maximum duration of 16 hours in which it is not standardized and it is the responsibility of the instrumentmaker . For longer projects a team is made in a more standardized workflow.	The waterfall method is used. The iterations are short cyclic. The goal is to move towards an agile workflow	If it is a project a hourly plan which is connected to the cost is made. The global requirements are made, not to specific. The requirements can both be implicit and explicit. Within these requirements there is not a lot of standardization	Miscommunications often lay in the quality of the intake. The definition of time is often vague. It rarely happens that the end product is not to desire.
DEMO TU Delft	This is diverse. It always starts with a introduction meeting with a client. The process starts with the formulation of the hard requirements. A set of requirements is also included.	There is a standard workflow but not in the form of a standard template. The expectation management is seen as an important aspect.	Yes we use a specific project management system based on the system setup by Roel Grit. The project leaders are also engineers and will lead this process. Scrum is used by some, but not organisation wide.	The result can be a pre investigation of the project. It can be seen as a feasibility research. Sometimes a market investigation is done to see what options there are already available. A protoype is often used as a proof of concept.	It is not an often occuring problem, but it does happen from time to time. The essence is to make sure every decision is recorded in writing.
Precision Mechanics and Engineering Group VU	There are 3 working streams, quickservice, jobs, and projects. The system specifications are the key variable of the project document which can be seen as the project definition.	Yes this is standardized in the form of a project document that is used for every project.	A standard project management technique which is personally custumized is used. No scrum or agile approach is used. A 3 week communication cycle is used.	The system specifications are a key part of the deliverables. A planning is also included in it.	The miscommunication is not seen as an issue and does happen often. The distance between the researchers and instrument makers is low.
LUMC	There is the distinction between the clinical jobs and the "regular" instrument making jobs. The jobs flow in by diverse ways, for example by email or by walk-in.	If a job is medical than it is strongly standardized due to the rules involved with it. Other projects may be less standardized.	Scrum has been tried before but was not a good fit. No specifc project management technique is used now.	A program of requirements is a deliverable. Apart from this the design framework and material choices are also included.	Miscommunications do not happen regularly. The instrument makers can level with the researchers involved. It did not happen yet that an end product is presented that did not serve it's purpose due to a miscommunication.

Table 3: Part 2 (Q6–Q10)

⟸ **Continued from previous page**

Organisation	Question 6	Question 7	Question 8	Question 9	Question 10
Wageningen Technical Solutions	Yes this is connected, mainly with bigger projects. A good project definition lowers the iterations but overall the end result stays high quality.	An incorrect intrepretation of requirements is a problem. The risks involved with this, both financially and safety wise, can be a sensitive point as well.	Yes this is defenitely true. The old-school workflow is present but not everybody sees this lack in standardized work as a problem.	No, AI is not being used yet.	Yes this would be useful. However it should be a system that is user friendly due to the presence of old-school intstrument makers
Nederlands Herseninstituut	Due to the small-size of the organisation and therefore the low entrance bar this is minimal. But in essence the statement is true, it just does not happen often.	The assumptions made for the working of a product may not be useful for the actual use case.	This is true. But due to the small nature of the institution this does not lead to problems	No, AI is not being used yet.	It is hard to say, it is essential that it is easy to use since part of the team is not proficient in computer usage.
Ontwikkeling Medische Technologie UMC Utrecht	This is definitely true. Sometimes there are projects that, for example, become more expensive than expected and therefore almost get cancelled. It is crucial to implement a correct projectdefinition but it does not often lead to problems in this organisation.	There are not a lot of mistakes that lead to qualitive issues. An overestimation of hours in order to reach a safe space can be a problem. Apart from these quote related issues not a lot of mistakes are made.	Yes everybody works in there own way. But due to the small setup of the team it does not lead to issues.	The instrumentmakers do not use AI. Due to the analog nature this is also not easily implemented.	Yes AI can definitely assist in this. It would be great to implement this with the historical projects as well to serve as an aid to predict planning or cost for a project.
DEMO TU Delft	An incorrect project definition always leads to problems later on in the project. This is one on one connected.	The importance of certain aspects is sometimes wrongly interpreted. This can be led back to an incorrect project definition. A scope greed is a risk in the project. If the feedback loop is than incorrect, mistakes are unavoidable.	Since we also have project managers this is limited. The chance of tunnelvision is therefore less.	No, AI is not being used yet.	Yes this would be useful. A generic system for project management would also be great to assist.
Precision Mechanics and Engineering Group VU	An incorrect project definition leads to problems in almost all the cases and leads to a setback to the complete beginning again. It is one on one connected.	The incorrect formulation of the system specifications leads to the most problems. The specification of the circumstances under what an instrument must be working needs to be specified correctly.	This is not seen as a problem since sometimes it is most efficient to immidiately start working on the project instead of the project definition drafting process.	No, AI is not being used yet.	This can definitifely have added value, however, there is doubt that the specific instrument making jobs can be assisted by AI.
LUMC	An insufficient project definition does lead to a harder process where a lot of adjustments are needed during the process. The end result does not suffer from this, only a longer process can be a problem.	A possible problem can be to copy the requirements from the client too quickly and therefore not foreseeing the faults in these requirements. A shortage in the drafting of requirements can be a problem.	This problem has become less over the years since the digital design work has increased which leaves more room for standardization. Due to this faults are also seen easier.	No, only some small ChatGPT help.	Yes this would definitely be helpful. It would also be perfect if it could assist in the administrative tasks of filling in the forms for the medical part.

Appendix C: PEFT training code

```python
from transformers import AutoTokenizer, AutoModelForCausalLM, TrainingArguments, Trainer, DataCollatorForLanguageModeling
from datasets import load_dataset
from peft import get_peft_model, LoraConfig, TaskType
import torch

# Check if MPS is available (Apple GPU support)
device = "mps" if torch.backends.mps.is_available() else "cpu"
print(f"Using device: {device}")

# Load tokenizer and base model (no quantization)
model_name = "microsoft/phi-3-mini-4k-instruct"
tokenizer = AutoTokenizer.from_pretrained(model_name, trust_remote_code=True)

tokenizer.pad_token = tokenizer.eos_token  # Add padding token if needed

model = AutoModelForCausalLM.from_pretrained(
    model_name,
    trust_remote_code=True,
    torch_dtype=torch.float16,
).to(device)

# Apply LoRA adapters
peft_config = LoraConfig(
    r=8,
    lora_alpha=16,
    lora_dropout=0.05,
    bias="none",
    task_type=TaskType.CAUSAL_LM,
    target_modules=["qkv_proj", "o_proj"]
)

model = get_peft_model(model, peft_config)

# Load and format dataset
def format_example(example):
    prompt = f"<|user|>\nInstruction: {example['instruction']}\n\nInput: {example['input']}\n<|assistant|>\n{example['output']}"
    return {"text": prompt}

dataset = load_dataset("json", data_files="data/data.jsonl")["train"]
dataset = dataset.map(format_example)

# Tokenize
```

```python
def tokenize(example):
    return tokenizer(
        example["text"],
        padding="max_length",
        truncation=True,
        max_length=1024
    )

tokenized_dataset = dataset.map(tokenize, batched=True)
data_collator = DataCollatorForLanguageModeling(tokenizer=tokenizer, mlm=False)

# Training arguments
training_args = TrainingArguments(
    output_dir="./phi3-lora-output",
    per_device_train_batch_size=1,
    gradient_accumulation_steps=4,
    learning_rate=1e-4,
    num_train_epochs=3,
    fp16=True,
    logging_dir="./logs",
    logging_steps=10,
    save_strategy="epoch",
    save_total_limit=1
)

# Trainer
trainer = Trainer(
    model=model,
    args=training_args,
    train_dataset=tokenized_dataset,
    data_collator=data_collator
)

trainer.train()

model.save_pretrained("./phi3-lora-trained")
tokenizer.save_pretrained("./phi3-lora-trained")
```

Appendix D: Statistical analysis code

```python
import pandas as pd
import matplotlib.pyplot as plt
from scipy import stats
from statsmodels.stats.multitest import multipletests
import numpy as np


df = pd.read_csv("results.csv", sep=";")

# Check which are the baseline columns and which are the finetuned columns
base_cols   = [c for c in df.columns if c.startswith("BASELINE")]
fine_cols   = [c for c in df.columns if c.startswith("FINETUNED")]

assert len(base_cols) == len(fine_cols) == 19, "column count mismatch" # check for incorrectness

# Compute avarage scores
baseline_mean   = df[base_cols].mean(axis=1)
finetuned_mean  = df[fine_cols].mean(axis=1)
diff = finetuned_mean - baseline_mean


#Total baseline mean
total_base = 0
for base in baseline_mean:
    total_base += base
total_base = total_base/len(baseline_mean)
print(f"Total mean score baseline: {total_base}")

#Total finetuned mean
total_fine = 0
for fine in finetuned_mean:
    total_fine += fine
total_fine = total_fine/len(finetuned_mean)
print(f"Total mean score finetuned: {total_fine}")

#Total difference
total_diff = 0
for difference in diff:
    total_diff += difference
print(f"Total difference in score finetuned model : {total_diff}")

# Shapiro test
w, p_norm = stats.shapiro(diff)
```

```python
print(f"Shapiro: p = {p_norm:.4f}")

# T-test
t_stat, p_value = stats.ttest_rel(finetuned_mean, baseline_mean)
print(f"T-value: {t_stat} p-value = {p_value}")
```

Appendix E: Example output finetuned model

Project: CryoLab
Project code:
Group: Low Temperature Physics Lab
Project leader: Lars Verboom
Author: Nina Chung and Lars Verboom
Date last mod.:
Status:
Source:

Content:

1. Brief introduction
2. Project definition
3. Phasing and activities of the project
4. Achieving results
5. Preconditions
6. Planning/roadmap
7. Budget
8. Information

1. Brief introduction:
This project is a central effort within the Low Temperature Physics Lab, focused on developing the CryoLab, a precision scientific instrument designed for frontier experimental research. The instrument aims to enable measurements at the atomic scale under extreme conditions such as ultra-low temperature or high magnetic fields. The project will integrate mechanical engineering, electronics, and material science. Leadership will be provided by Lars Verboom, with technical design led by Nina Chung and their engineering...

2. Project definition:
The CryoLab will combine scanning probe technology with integrated cryogenic control, allowing sub-nanometer resolution on heterogeneous quantum materials. It will support multiple probe modes (STM/AFM) and operate below 1K temperature using a dilution refrigerator system. Vibration isolation will be embedded using custom air-damped mounts.
The device will include a closed-loop nanopositioner stack with sub-picometer accuracy and automated feedback alignment. Custom-designed low-noise preamplifiers and temperature sensors will be implemented.

3. Phasing and activities of the project:

- Phase 1: Literature review and functional specification
- Phase 2: CAD modeling and finite element vibration analysis
- Phase 3: Assembly of structural and thermal subcomponents
- Phase 4: Integration of scanning stage and control electronics
- Phase 5: Testing and system calibration in cryo environment
- Phase 6: Performance validation and publication of benchmarks

4. Achieving results:
Expected outcomes:

- Operational platform for atomic-scale imaging
- Stable operation at mK temperature and UHV conditions
- Integration with existing lab infrastructure
- Enablement of novel quantum material experiments

5. Preconditions:

- Acquisition of cryo-ready materials
- Technical support for vacuum and electronics subsystems
- Coordination with cleanroom scheduling
- Securing safety approval for high voltage and cryogenic setup

6. Planning/roadmap:

- Q1: Design & review
- Q2: Component sourcing and construction
- Q3: Assembly and testing
- Q4: Demonstration and user training

7. Budget:

- Estimated €35,000 for dilution refrigerator and nanopositioners
- €12,000 for electronics and sensors
- €8,000 contingency and transport costs

8. Information:

This project start document initiates the formal construction phase and aligns the multidisciplinary team toward clear scientific and technical objectives.

Questions

1. Project Intake & Vision Alignment:

- Is the scientific problem and experimental goal clearly defined?
- How does the CryoLab fit within the broader research program or facility capabilities?

2. Clarify Requirements & Scope:

- Are the functional specifications derived from existing literature or preliminary experiments?
- How will the instrument interface with other lab components or data acquisition systems?

3. Feasibility:

- Are there existing design templates or similar instrumentation that can be adapted?
- What are the critical engineering challenges (thermal isolation, mechanical stability, vacuum sealing)?

Check / Initial Design Draft:

- Is there a preliminary CAD model or schematic available?
- How are safety and environmental constraints addressed in the design?

4. Identify Key Constraints:

- Time: Is the project timeline realistic given the complexity and integration steps?
- Budget: Are material and fabrication costs within the allocated funding?
- Safety: Are high voltage or cryogenic hazards adequately mitigated?

5. Stakeholder Review & Clarification:

- Is the project definition aligned with lab director's research goals?
- Are there any technical or logistical dependencies not yet considered?