

Master thesis

AI engineering: The generation of automatically written reports in the Dutch law enforcement

By
Wouter Westerkamp
6094430
w.j.westerkamp@students.uu.nl

First supervisor: Prof. dr. S. Brinkkemper
Second supervisor: Dr Sergio Espana Cubillo
and
Daily supervisor: Martijn van Vliet MSc

In collaboration with
the Dutch police



Department of Information and Computing Sciences
Faculty of Science

Abstract

Background Capacity within the Dutch police is a concern. Due to heavy workloads and lack of resources, the police is forced to make difficult decisions regarding the priority of tasks. This study aims to create functional AI-based software for generating police reports to make the report creation process more efficient. It seeks to understand the technological challenges and develop a theoretical foundation for AI engineering in this context.

Method This study was done in the context of design science with a vision-oriented design configuration. A multi-vocal literature study was performed in the context of AI engineering, followed by iterative design activities. A newly designed approach, with the combination of continuous prototyping and requirements engineering, was performed to capture high quality requirements. Ten interviews were held with different participants within the Dutch law enforcement and the supplier of the software used. Eventually, a prototype capable of automatically generating police reports was developed, evaluated, and refined. The research addresses key questions regarding the challenges of AI engineering, the generation of automatically written reports, and the evaluation of generated reports.

Result The developed AI models achieved high precision and recall in generating metadata for police reports. For the written statements within the reports, the evaluation indicated that the models captured the suspect's own words, including a total of 22 omission errors. Next to that, the models showed no errors in applying the correct question & answer format.

Conclusion The performed study researched potentials of integrating AI into the generation of automatically written reports for the Dutch law enforcement. First, it identifies challenges in the AI engineering process, including overestimating AI capabilities, the need for transparency and explainability, regulatory compliance, unpredictability, and concept drift. Second, the study developed and evaluated a prototype for generating police reports. While the results show great potential with high precision and recall, there are still obstacles regarding the legal, ethical, and AI regulations. These insights serve as advice for the future works within Dutch law enforcement. Particularly, researching the legal feasibility and real-world data testing to further refine and validate the system.

Keywords: AI engineering; Responsible AI; AI governance; Software development lifecycle; Generative AI; Large language models

Acknowledgements

I would like to express my thanks to all those who have supported me throughout the journey of writing my thesis. first and foremost, I am thankful to Prof. Dr. Sjaak Brinkkemper, who made it possible to start this project at the Dutch police and guided me in performing my research. Secondly, I am thankful to Martijn van Vliet MSc, who has been an almost daily source of support, providing expertise that was crucial to the successful completion of this research. Thirdly, I would like to express my thanks to Dr. Sergio Espana Cubillo for his valuable feedback. Furthermore, I would like to thank the participants of the interviews for their time and for sharing their valuable information and insights, these were essential to the quality of this research. Finally, I would like to thank Verticai for providing the necessary resources and support, without which this project would not have been possible.

Contents

Abstract	2
Acknowledgements	3
1 Introduction	7
1.1 Problem statement	8
1.2 Scope	9
1.3 Research questions	10
1.4 Relevance	11
1.5 Thesis outline	11
2 Research plan	12
2.1 Design science	12
2.1.1 Literature study	14
2.1.2 Case study	16
2.1.3 Semi-structured interviews	16
2.2 Validity & Reliability	17
2.3 Research execution	17
2.4 The Dutch police	18
2.5 Fitting case study into design science	19
3 Software Engineering	22
3.1 Software development life cycle	22
3.1.1 Requirement engineering	23
3.1.2 Software designing	27
3.1.3 Software implementation	29
3.1.4 Testing	30
3.1.5 Maintenance	32
3.2 Software engineering process models	33
3.2.1 Waterfall model	33
3.2.2 Agile model	34
3.3 DevOps	35

4 AI Engineering	37
4.1 Roadmap for responsible AI	38
4.2 AI governance	39
4.3 SDLC for responsible AI	41
4.3.1 RE4AI	41
4.3.2 AI software designing	44
4.3.3 AI software implementation	46
4.3.4 AI software testing	48
4.3.5 AI software operations	49
5 Generative AI	50
5.1 Large language models	51
5.1.1 Prompt engineering	52
5.2 Structured report generation	53
6 The workings of the Dutch Police	54
6.1 Reference model for police business processes	54
6.2 Reporting within the Dutch police	55
6.2.1 Types of interrogations	56
6.2.2 Requirements of a Police Report	57

7 The Case study	59
7.1 Envisioning & ideation	59
7.2 Feasibility checking	59
7.3 Artifact building	60
7.3.1 Creation of the prototype	61
7.3.2 Prototype driven RE4AI	65
7.4 Ex-ante evaluation & justification	68
7.4.1 Structure of the models	69
7.4.2 Precision & Recall	72
7.4.3 Manual evaluation of generated statements	77
7.4.4 Summary of evaluation	80
7.5 Grounding & conceptualization	81
7.6 Knowledge formulation	83
7.6.1 Value exploration	83
7.6.2 Obstacle & opportunity identification	88
7.6.3 Solution discovery	93
7.7 Revision & refinement	98
7.8 Reflection	100
7.8.1 Results interpretation	101
7.8.2 Limitations	102
7.8.3 Future research	103
8 Conclusion	104
A Software engineering product Models	106
B Police report Requirements	107
C Interview protocol	108
D Manual evaluation of generated statements	109

1 Introduction

According to the 2023 annual report of the Dutch police⁹⁹, specifically the report from the Amsterdam police unit, the workload is large and has a noteworthy impact on police personnel. The report explains the considerable effects of heavy workloads on police officers and the difficulties they face in managing their duties. Capacity within the police organization has long been a concern, with the police force frequently forced to make difficult decisions regarding the prioritization tasks due to a lack of resources. The report of van Bavel et al.⁹⁹ highlights the increasing complexity of societal problems, the rise in protests, and the growing number of incidents, which increases the work pressure faced by police officers. These heavy workloads show that time should be treated as an important resource. The less time officers spend writing reports, the more time they can spend on the streets. Police reports should be error-free at the moment they are written, which is not always the case at this moment⁴³. Fewer mistakes also result in less time spent on correcting faulty police reports.

The ongoing digital transformation of society brings opportunities to tackle certain societal challenges. Artificial Intelligence (AI) stands out as one of the rapidly evolving and applied technologies. The paper of McCarthy et al.⁶¹ defined AI¹ as: “The science and engineering of making intelligent machines, especially intelligent computer programs.” AI is considered a central technology shaping the next decade. This is because AI can complement the skills of existing workforce’s, leading to greater cost efficiencies and savings^{107 14}. It has the potential to bring about positive change and provide benefits for both public and private sectors. Particularly through the use of virtual workforce’s, often referred to as intelligent automation. Governments are investing heavily in AI to explore the potential of automating decision-making, or at least enhancing human decision preparation⁶³. AI can offer solutions for four areas in the public sector:⁵¹

- (1) **Facilitating** decision-making processes.
- (2) **Ensuring** integrated data management.
- (3) **Enabling** interaction with virtual agents.
- (4) **Automating** administrative tasks.

When addressing a particular issue, AI may use one or more technologies, provided they are interoperable and integrated. Such as natural language

¹There is no commonly accepted definition of artificial intelligence (AI)¹⁰³.

processing, neural networks, and various others. Focusing on the area of automating administrative tasks, the use of large language models (LLMs) shows promising results in this area. Earlier steps towards reducing the administrative workload associated with report writing can be found in the Care2Report research program⁵⁶. This initiative uses a method called Automated Dialogue Summarization (ADS) which uses AI to ease the task of generating Electronic Medical Reports⁶⁶. To ensure automatic summarization of a report, it is necessary to have some structure present⁵⁷. While the field of ADS has been explored in the medical domain, it has yet to be explored in the police domain. When transitioning between domains, it is important to first explore the feasibility of implementation. Therefore, considering the potential to reduce administrative burdens for writing reports, and taking the advances of AI in the medical field as an example, it was decided to explore the creation and implementation of AI technologies within the Dutch law enforcement.

1.1 Problem statement

The national police of the Netherlands consists of ten regional units, one central unit and the police service center. In 2023 the Dutch police force registered over 800,000 crimes²². Because it should be possible to justify the course of actions of every interception, it is said that reporting is a most crucial part of the police work. The Dutch police force faces numerous challenges in manually creating police reports, which raises concerns about efficiency, quality, and administrative burdens⁴³. An investigation, published by Jansen⁴⁶, shows the increasing demand on police officers. They will have to do more, with less resources. This situation highlights the importance of optimizing the efficiency of report writing in order to spare time for other duties. In addition, there is an expectation for police reports to be error free upon the first submission. This is still a challenge within the law enforcement⁴⁶. AI technology presents various opportunities within the public sector, with great potential to increase efficiency. However, it is important not to overlook the related challenges, as they can impact the implementation and use of AI applications¹⁰⁷. The study of Bosch et al.¹⁹ shows that the engineering of AI software, and especially the step from prototype to implementation, is very difficult. AI engineering is not only about the technical aspects but also extends to organizational level, such as that the AI system should be consistent with the legal and ethical standards. This requires a review of current practices in traditional software engineering⁵⁴. In summary, the introduction of AI has the potential to significantly reduce the administrative burden on the Dutch police force. However, the process of creating

and integrating such technologies involves complex challenges that are beyond the technical view. Therefore, this study not only shows the technical side of creating a prototype for generating police reports. It also evaluates the prototype and gives an approach on prototype driven requirement engineering. Besides that, this study investigates the current literature into all the perspectives needed for the successful creation and implementation of AI based systems.

1.2 Scope

The scope of this project, will be within the framework of the Care2Report (C2R) program. This research aims to demonstrate the substantial benefits of the system of Vertical¹⁰². Vertical¹⁰² is a startup, founded by researchers of the C2R program. The research will be performed in the context of the Dutch law enforcement, laying the experimentation of the Vertical systems beyond its initial healthcare focus. This will be done by an in-depth literature study in the fields of software engineering, AI engineering, GenAI, and the Dutch law enforcement. This study will address the encountered challenges, taking into account the traditional software development cycle. In addition, a case study within the Dutch law enforcement is performed that focuses on the stages of requirement engineering and design. Within these two stages, the Vertical system will be used to create a proof of concept for the creation of automatically generated interrogation reports. During the development of the prototype, an iterative approach of prototype driven requirement engineering is followed. Due to time, technological, resource, and legal limitations of the project, only a first iteration of the two phases could be realized.

1.3 Research questions

The aim of this thesis is to create a functional prototype for the automated generation of police reports. In addition, an advisory report will be written for the Dutch police, which will not only include our findings, but also provide directions for future work and research. To achieve this, a structured police format will be formulated, and a large language model will be guided through predefined models within the Vertical system. Next to that, the entire software development lifecycle will be explored in the context of AI engineering to understand the challenges that arise when building AI based systems. The following main research question is formulated:

MRQ: To what extent can AI technology be integrated into the creation of automatic reporting software for the Dutch law enforcement?

The creation of a functional prototype for the automated generation of police reports is the result of the identified challenges in the problem statement (Section 1.1). In order to address the main research question, three specific sub-questions are formulated:

RQ1: What challenges does AI engineering introduce to the software development life cycle?

The first question focuses on the software development life cycle, which explains the fundamental phases within software engineering. The study will investigate the difficulties encountered at each stage and identify the additional challenges. This question will be answered through the study of the existing literature.

RQ2: To what extent could a police report be automatically generated?

This question will focus on the creation of an automatically generated police reports. This will include a list with all the requirements that a police report needs to meet in order for it to be legally usable. Certain requirements will be evaluated for their feasibility in terms of generating them with AI. By testing these requirements, the research aims to determine the practicality and effectiveness of automatically generating police reports with AI.

RQ3: What is the quality of the generated reports?

Finally, the generated reports will be evaluated during interviews with experts in the field. Based on the evaluation, numerous strengths, weaknesses, and limitations of the prototype will be identified, providing valuable insights for further optimization and improvement.

1.4 Relevance

A stakeholder is defined as: “A person, group of persons, or institution affected by treating the problem”¹⁰⁶. This study will affect three distinct sets of stakeholders: the Dutch police, Verticai, and the scientific community. This study aims to contribute both scientifically and practically by exploring the creation and implementation of automatic reporting within the Dutch law enforcement. It seeks to understand the technological challenges and develop a theoretical foundation for AI engineering in this context. Verticai provides operational support by offering their software system with already several functionalities useful for this study. Eventually, the acquired knowledge in this research could be leveraged to enhance the Verticai system itself.

1.5 Thesis outline

The remainder of this study is structured as follows. In section 2 the research plan of the thesis can be read. This part explains the design science used in this research and how it fits within this study. After that, the literature study begins. Starting in section 3, which takes a look into the fundamentals of software engineering practices. Then section 4 and section 5 provide information regarding the concepts of AI engineering and Generative AI. Finally, the literature study will move into the domain of the Dutch police force in section 6. After the literature study, the actual case study begins at section 7. The performed iterative design activities are described, which consist of the creation of a prototype and the performance of several requirement engineering activities. Finally, this study draws several conclusions based on the results in section 8.

2 Research plan

In order to increase the validity and reliability of this study, this section is dedicated to provide all the information regarding the research plan. First, the design science related to this study is explained. This includes the design orientation and its related activities. The protocols for the literature study, case study, and interviews are also provided. Then the research execution is explained.

2.1 Design science

Design science is the study of artifacts in context, which is a powerful framework for developing solutions that are not only theoretically robust, but also embedded in the practical environment¹⁰⁶. This research fits within the **design science research methodology** (DSRM) described by Peffers et al.⁷⁹. The nominal process of DSRM includes problem identification, objective definition, artifact design, demonstration, evaluation, and communication. While these steps provide a good foundation for a design science research project, there is little guidance on how to implement these steps⁸⁹. Therefore, the paper by Schoormann et al.⁸⁹ is used for its visualization and guidance of the research process (Figure 1).

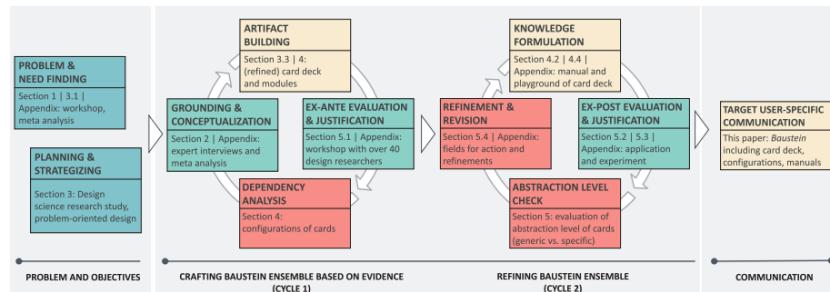


Figure 1. Research design⁸⁹

The paper of Schoormann et al.⁸⁹, has predefined several configurations which are designed to serve either for guidance or inspiration for the user. This study aligns with the **vision-oriented design** configuration (Figure 2). This approach is typically less formalized and is driven by ideas of researchers inspired by their own experiences, the willingness to apply novel technology or proposals for third-party funding. Researchers begin by implementing initial prototypes, such as mock-ups, to make their ideas more tangible⁸⁹. In this study, the vision-oriented design is based on the reengineering of existing AI based software, originally developed for the Care2report program aimed

at automating medical reporting. The software will be optimized to meet the unique requirements of writing police reports. The initial software will serve as a concept, ensuring the relevance and feasibility of the solution.

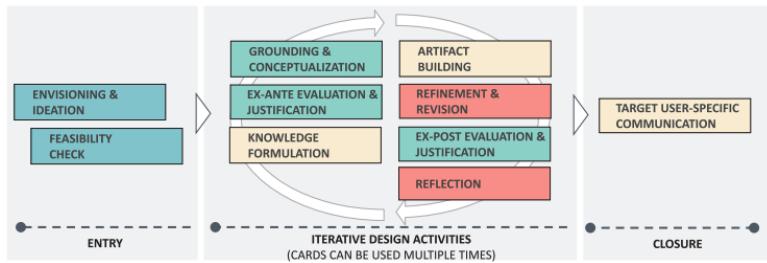


Figure 2. Vision-oriented design

This study will follow the vision oriented design configuration from Schoormann et al.⁸⁹. It will assess whether the designed software, will produce the desired effects in the given context. Which means that this study will investigate if the designed software will automatically create a structured police report. The effectiveness will be examined by comparing the police report itself with specified requirements. The design cycle involves a feedback loop, allowing for iterative improvements based on real-world implementation challenges. Following the paper off Schoormann et al.⁸⁹, the vision-oriented design can be divided into three different phases.

1. **Phase 1: Entry.** The initial phase of the design project starts with the project organization.
 2. **Phase 2: Iterative design activities.** The iterative approach ensures that the project adapts and evolves through multiple cycles, resulting in a more effective and user-centered solution.
 3. **Phase 3: Closure.** Communicating the final design to the stakeholders and formulating results, visualizing knowledge and explore the potentials of the developed artifact. The aim is to bridge the gap between the design process and its real-world application.

How these phases fit in the context of this study will be described in section 2.5. Due to the time constraints associated with this study, it will not include the ex-post evaluation & justification of phase 2, and it will not include the target user specific communication of phase 3.

2.1.1 Literature study

The literature study is particularly important in the exploratory phase of the research to gather relevant information about the topic of AI engineering, GenAI and Dutch police reports. This literature study was done to find existing knowledge. But as the academic literature within the domain of the Dutch police is scarce, the literature study also focuses on non-academic literature. This is called a **Multivocal Literature Review** (MLR). Multivocal literature is defined as all accessible literature on a topic³⁸. This kind of review is a form of systematic literature review which also includes gray literature such as articles, government reports, blogs and white papers⁶⁹. By including this variety of literature, better domain knowledge of the Dutch police could be provided. The paper of Garousi et al.³⁸ states the five phases to perform a MLR which include: Search process, Source selection, Quality assessment, Data extraction, and Data synthesis.

1. **Search process.** This research leveraged the power of various databases to build a comprehensive understanding of the topics within this research. Google Scholar and Web of Science were the foundation that allows to delve into the academic world of topics such as, AI engineering, Generative AI, Large language models, and several papers regarding the Care2Report program. To extend this academic foundation, the website of the government and the database of the Dutch Police Academy is used to explore the complexities of the Dutch police procedures and legal frameworks. Table 1 contains an overview of the selected sources.

Source	Type
Google Scholar	Academic articles
Web of Science	Academic articles
Overheid.nl	Laws and Regulations
Politieacademie.nl	Government documents

Table 1. Selected sources

2. **Source selection.** In order to make the search efforts more efficient and effective, explicit inclusion and exclusion criteria are used. The inclusion criteria were carefully designed to ensure relevance and specificity and included parameters such as: publication date, amount of references, relevancy to the police context or automated reporting. Exclusion criteria were defined to filter out irrelevant or outdated content

(table 2). These criteria were in particular: articles or papers not written in English or Dutch, papers that were not accessible, or that were published on specific types of sources with low credibility, as specified in table 3. Besides the explicit inclusion and exclusion criteria for the search engines, also several search terms are used. These terms are for example “AI engineering”, “Software engineering” “GenAI”, “Prompt engineering”, etc. These terms led to numerous academic papers that were relevant to the literature study. Using the same terms for the public database of the politieacademie.nl also resulted in valuable articles for the proposed study. For the specific requirements of a Dutch police reports, the website of the Dutch government was used “overheid.nl” to find the corresponding laws regarding creating police reports.

Source	Inclusion Criteria	Exclusion Criteria
Google Scholar and Web of Science	Publication data, Amount of references, Literature relevant to RQs, Relevancy to police or automated reporting domain	Outdated, Irrelevant content, not written in Dutch or English, not accessible
Politieacademie.nl	Relevancy to Dutch police procedures and legal frameworks	Irrelevant content
Overheid.nl	Relevancy to Dutch police procedures and legal frameworks	Irrelevant content

Table 2. Inclusion and exclusion criteria

3. **Quality assessment.** The quality assessment is about finding the extent to which a source is valid and free of bias³⁸. For scientific sources, the paper of Garousi et al.³⁸ states that this can be done by evaluating factors such as the credibility and reliability of the source/journal, the methodology used, date of publication, or the number of backlinks. In the case of assessing general sources, the paper of Adams et al.² classifies three different tiers of sources in the general literature in order to provide a structured approach into incorporating gray literature in the study process. In table 3, these different tiers are shown, including their credibility level and type of outlet.
4. **Data extraction.** The data extraction focuses on looking at the gathered literature and the defined research questions. Eventually, the literature study should provide enough data with the information that is gathered during the case study to answer all the research questions.

Tier	Retrievability/Credibility	Type of outlet
1st	Significant	Books, Book chapters, Broad range of journals, Government report
2nd	Moderate	Annual reports, New articles, videos, company publications, wiki
3rd	Low	blogs, emails, tweets, letters, catalogs

Table 3. Shades of gray²

5. **Data synthesis.** In the data synthesis, The extracted data is combined and analyzed to identify patterns, themes, and trends in the literature³⁸. The synthesized findings are then used to provide a comprehensive summary of the current state of knowledge and practice in the researched domain, highlighting key findings and gaps in the literature.

2.1.2 Case study

This study will perform an **exploratory case study**. In exploratory research, it is important to use industrially based cases because of the contextual importance from which an artifact may emerge¹⁰¹. This study will perform this research by performing the steps of the vision-oriented design described in section 2.1. The purpose of this case study is to build a proof of concept and provide new insights and questions regarding the creation and implementation of automatically generated interrogation reports of the Dutch law enforcement. Thereby, answering the second and third research question, formulated in section 1.3.

2.1.3 Semi-structured interviews

This study will conduct several **semi-structured interviews**. These interviews will serve as a means of gathering a wide and detailed range of requirements from stakeholders within the Dutch police. In a semi-structured interview, the questions are planned, but not always asked in the order in which they are stated. The flow of the conversation during the interview can influence the order in which the different questions are presented to the interviewee. The researcher can use the list of questions to ensure that all questions are raised with the interviewee at some point during the interview⁸⁶. In addition, semi-structured interviews do not have a complete script. Therefore, participants have more flexibility to expand on their answers in order to gain a deeper understanding of the challenges and opportunities

within my study. Semi-structured interviews are commonly used in software engineering case studies⁸⁶. The designed approach of how we performed the interviews and the results are further described in section 7.6.

2.2 Validity & Reliability

To ensure transparency and reproducibility, research must be both valid and reliable. **Reliability** refers to the consistency of assessment scores⁶⁷. In this study, it says something about the results that are gained from the case study. In order to calculate reliability, the study must clearly document the procedure. This documentation should be at a level of abstraction that makes the internal decision-making process clear. It is important for the public to understand how the researcher prepares for the study and how the data is collected and analyzed³². **Validation** is the process of gathering evidence to support the acceptability of the conclusions drawn. **Validity** refers to the extent to which the evidence support that these interpretations are correct and that the way in which the interpretations are used is correct⁶⁷. For this study, this means that it is important to check data quality. This also means that the results that are collected must be consistent and trustworthy. To achieve this, the results can be reviewed by an independent researcher to ensure that the results are consistent with the analysis. In this study, the results will be discussed with other researchers within the Dutch police force and the Care2report program to make sure the research is valid and reliable.

2.3 Research execution

This study will be conducted over several months and divided into two separate phases, each with multiple deliverables. The first phase is the research proposal, which aims to identify gaps in the literature and practice. The research questions are the focus of this study and try to fill the identified gaps. A research plan is explained to inform how this study will be executed. The second phase involves performing a case study with the Dutch law enforcement. The case study will show the process of creating a prototype for generating police reports. The prototype is also evaluated and an approach on prototype driven requirements engineering is explained and performed. These research methods will be used to gather answers to the research questions.

2.4 The Dutch police

This study is performed in collaboration with the police force of the Netherlands. The Netherlands has one national police force, with 10 regional units and one central unit. This structure is aimed to enhance efficiency, effectiveness, and unity within the law enforcement of the Netherlands. A strong and well-defined architecture is crucial in this context, as it serves as a strong foundation for smooth communication, interoperability and streamlined operations between the various units. The architectural framework also plays an important role in the adaption and integration of technological advances⁸⁴. The foundation of the enterprise architecture of the Dutch police force is based on their strategy. This strategy, along with the corresponding enterprise architecture, was established with the creation of the National Police in 2012 and is documented in the National Police Organizational Plan⁸⁴. The main contribution of working under architecture is reducing complexity in the organization's structure by providing insight, overview, and clarity. For ICT projects, the enterprise architecture is essential to achieve alignment with the organizational goals⁶⁴ (Figure 3).

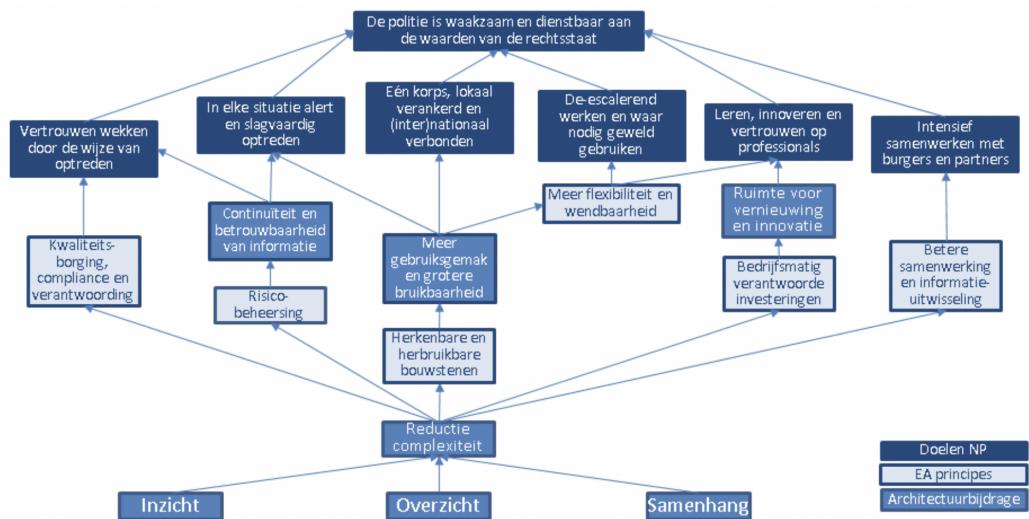


Figure 3. Goal Tree Dutch police⁶⁴

2.5 Fitting case study into design science

The literature study developed our understanding in the areas of software engineering (Section 3), AI engineering (Section 4), generative AI (Section 5), and the context of the Dutch police (Section 6). With the knowledge gathered from the literature review, the case study could be started with the best effort. The case study builds upon the literature study by creating an automatically generated interrogation report for the Dutch law enforcement. To stay within in the scope of this research project, we focus on the first two steps of the software development lifecycle, as shown in figure 4.

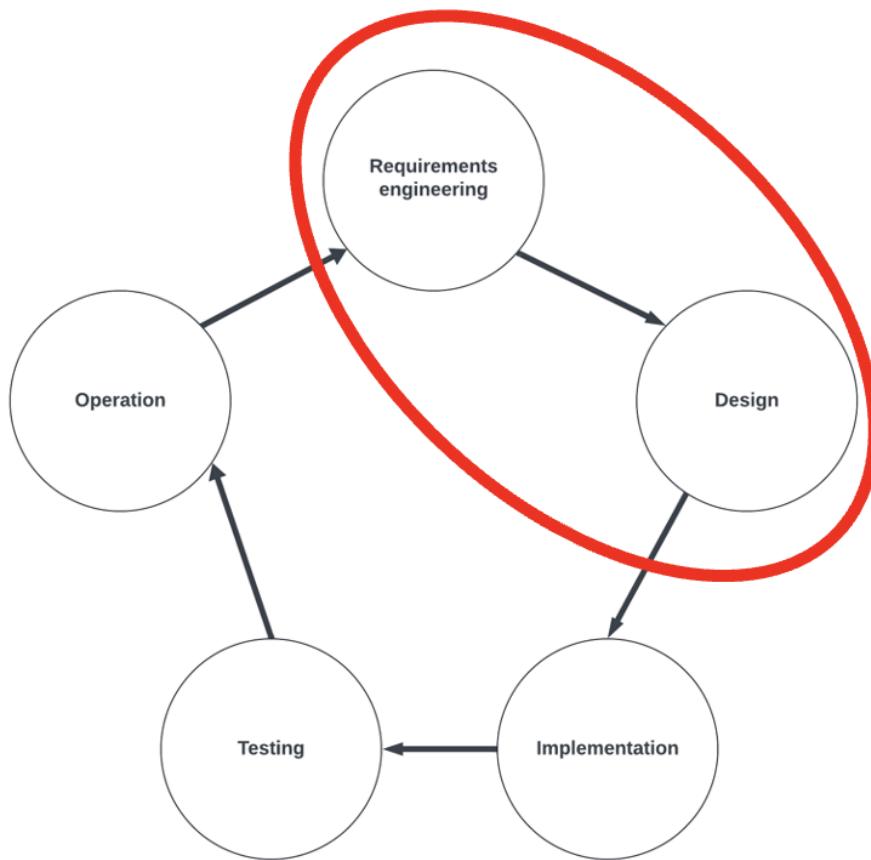


Figure 4. SDLC focus area

In order to gain insight into the practical aspects and potential effects of implementing automated reporting within the Dutch law enforcement, the RE phase will consist of conducting interviews with Dutch law enforcement professionals from a multidisciplinary perspective. This will be followed by a continuous improvement in the development of the prototype during the case study itself. How this process fits in the design science, is shown in figure 5.

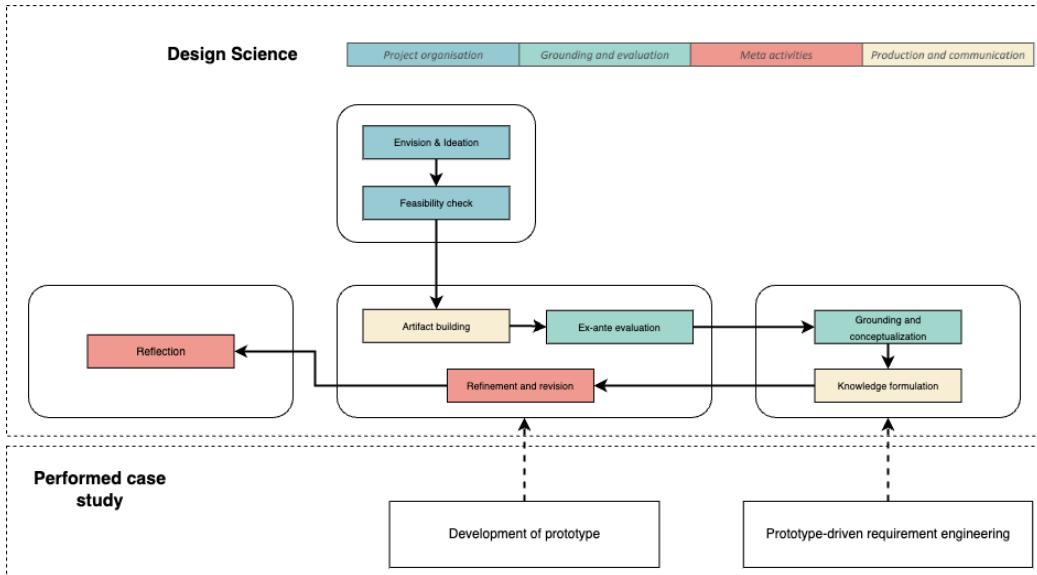


Figure 5. Fitting design science

Phase 1: Entry

The **project organization** is the first phase of the case study, which is focused on establishing the foundational elements for this vision oriented approach. This phase includes the following activities:

1. **Envisioning and ideation.** This involves the development of a possible solution to a problem. This study will build upon the work performed by the care2report program. Leveraging the insights and methodologies developed during this prior research (Section 7.1).
2. **Feasibility check.** This is to assess the feasibility of the study. This ensures that the study is aligned with available technological resources, development capabilities, and specific requirements of the law enforcement context. (Section 7.2)

Phase 2: Iterative design activities

After the project organization we have the second phase which are the **iterative design activities**. These activities are linked to the requirement engineering and design phase of the SDLC. In an ideal scenario, these activities would be repeated multiple times through several iterations. This iterative process allows for a more thorough development of the design based on evolving insights and requirements. However, to stay within the scope of this study, the iterative activities are limited to a single iteration and exclude the ex-post evaluation and the final phase of target-specific communication of the design science. This is a limitation of this study.

The iterative design activities performed in this study are described in the following order:

1. **Artifact building.** Moving from theory to practice. This activity not only involves the creation of the prototype / proof of concept. But also shows the designed approach for the performed requirement engineering activities, which involves the continuous developing of the prototype. (Section 7.3)
2. **Ex-ant evaluation.** This activity involves a comparative evaluation of the two models which are used for the prototype. Finding both its benefits and possible drawbacks. (Section 7.4)
3. **Grounding and conceptualization.** This activity focuses on laying the groundwork and a clear conceptual framework of how the requirements were acquired, and the prototype was continuously improved. (Section 7.5)
4. **Knowledge formulation.** This activity provides the results of the interviews, it will show the gathered requirements, obstacles, opportunities and provides a potential solution. (Section 7.6)
5. **Revision & refinement.** Based on feedback and the evaluation, this activity focuses on iterative improvements to the software. (Section 7.7)
6. **Reflection.** This activity reflects on the entire design process, from artifact building to refinement. It provides an opportunity to analyze what has been learned and how the project has contributed to the body of knowledge on automated police reporting. (Section 7.8)

3 Software Engineering

Software engineering is defined as “the application of a systematic, disciplined, quantifiable approach to the development, operation, and maintenance of software; that is, the application of engineering to software”²⁵. It involves the collaboration of multiple individuals to build a complete product or system that functions efficiently. Software engineering is a field that covers all the necessary aspects of producing high-quality software. The traditional field of software engineering includes a variety of concepts such as analysis, evaluation, specification, design, and there are also concepts related to management, quality, novelty, creativity, skills, teamwork, standards and professional practice. A complete coverage is available in the software engineering body of knowledge (SWEBOk) guide²⁵. The foundation upon which these elements stand is the **Software Development Life Cycle** (SDLC). In the following sections, the SDLC and the essential steps are explained to give a better understanding of the field of software engineering before diving into the principles of AI engineering.

3.1 Software development life cycle

The **Software Development Life Cycle** (SDLC) is used to design, develop and produce high quality, reliable and cost-effective software products. It involves a variety of tasks executed by different people and forms the basis of software engineering practices⁷⁷. When software engineers approach these tasks based solely on their personal experience, background and values, a systematic approach is often neglected. This lack of consistency can result in longer project duration's and poorer end products. In the worst cases, it can even lead to complete project failure⁹⁷.

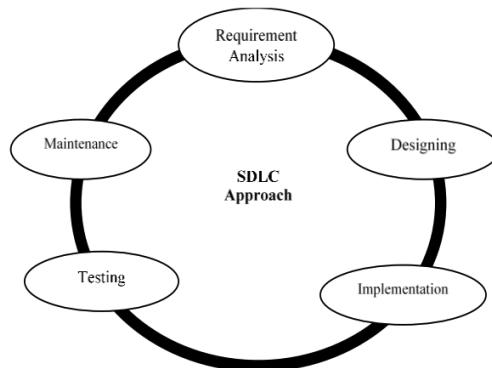


Figure 6. Essential Software development life cycle steps⁷

In order to provide systematic guidance to software developers, some Software development lifecycle (SDLC) models have been developed. Different models were introduced to solve different concerns. But overall, they will provide the following benefits⁹⁷:

- Improved understanding of the tasks to be performed.
- Better defined responsibilities, expectations, and results.
- Easier measurement and control of the project progress.

There is no single model that fits all software projects. In most cases, a chosen model still requires some adjustments to fit the needs of a specific organization. In section 3.2, the waterfall and agile model will be discussed. In the case of AI engineering, the traditional SDLC requires modifications because the process is less straight forward⁴⁷. Section 4 will present literature for building AI based systems following the development process practices highlighted in the paper of Lu et al.⁵⁴.

3.1.1 Requirement engineering

The systematic approach to manage requirements throughout the project life cycle of a software project to ensure that all the stakeholder needs are captured, understood and communicated is called **requirement engineering** (RE)^{25 52}. RE is arguably the most crucial phase in software engineering and plays a significant role in every phase of the software development. RE is a process that aims to ensure a software solution solves a specific issue. It involves discovering, understanding, formulating, analyzing, and reaching a consensus on the issue that needs to be solved⁵². Therefore, the involvement of stakeholders in the RE process is very important, but it brings several challenges that can influence the development of a system. The study of Abd Rahman et al.¹ stated that miscommunication between stakeholders, developers, and analysts is a common issue in the software development process. This is often due to requirements being expressed in natural language, which can be vulnerable to misinterpretation. Additionally, different visions or priorities among stakeholders can also cause potential conflicts, such as the delay on decision-making¹. The study of Tukur et al.⁹⁸ mapped numerous challenges regarding requirement engineering, these are shown in table 4. Because requirements are the fundamental basis for both software developers and project management, it is essential that all stakeholders should commit to using the correct requirements practices¹⁰⁵. Requirements are regularly

classified as **functional requirements** (FR) or **non-functional requirements** (NFR). FR's typically describe a behavior that a system will exhibit under specific conditions, while NFRs describe a property or characteristic that a system must exhibit or a constraint that it must respect¹⁰⁵. There are three questions that define the foundation of RE in ensuring that software projects meet or exceed the stakeholder's expectations⁵².

- *What* problem should be solved?
- *Why* such a problem needs to be solved?
- *Who* should be involved in the responsibility of solving that problem?

These questions can be answered by using the **requirement engineering process model** shown in figure 7. This model serves as a guide to navigate through the complex process of transforming the stakeholder needs into a functional software solution and consists of a cycle of four phases.

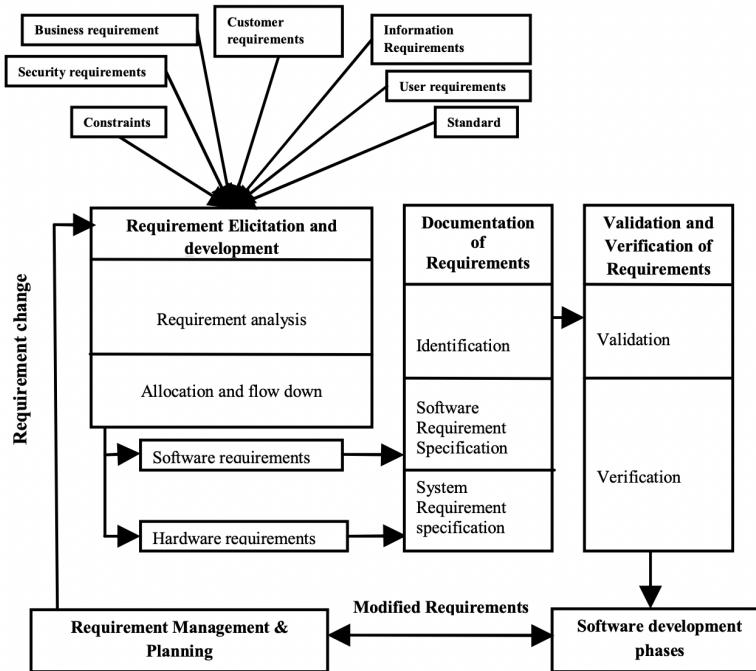


Figure 7. Requirement Engineering process model from Pandey et al.⁷⁶

1. **Requirements elicitation and development**, this phase starts with identifying stakeholders and collecting raw requirements from different perspectives. Raw requirements are those that have not yet been analyzed and have not been formulated in a well-defined requirement notation⁷⁶. The next step is **requirement analysis**, in which a prioritization technique is performed to calculate the level of importance of a defined requirement. Next to the requirement analysis, there is also the **allocation and flow down** of the requirements. This is a step to ensure that the requirements are clearly identified and traceable, both to higher level requirements and between requirements on the same level. In the perfect scenario, this produces a system in which every component is fully optimized and in balance with another. However, in practice complete balance is rarely achieved due to limitations in budget, time constraints, and technology⁵³. The following phase will be documenting the requirements, to enable clear communication with stakeholders and future maintenance of the requirements and the system⁷⁶.
2. **Documentation of requirements**, the reasoning behind choosing specific requirements, must be thoroughly documented. This documentation is essential for providing sufficient information for the verification and validation processes in the following phases, and for facilitating effective communication with stakeholders and future maintenance of system requirements. The first activity is **requirement identification**, which assigns unique identifiers to each requirement for better tracking throughout the development of the system. Then, a formal document called the **Software requirements specification** (SRS) is created. This document includes both the functional and non-functional requirements that are expected of the software or system. It includes use cases and evaluates parameters such as operating speed, security, and maintainability to ensure that the software efficiently and effectively meets its intended purpose, thereby reducing development time and costs. A well-written SRS specifies the interactions between an application, system hardware, other programs, and users in various real-world scenarios⁷⁶.
3. **Requirements verification and validation**, when the SRS has been written, all parties involved have to agree on its contents. It is important to check that the correct requirements are stated (validation) and that they are stated correctly (verification). The two main techniques used to validate requirements are requirements reviews with stakeholders.

ers and prototyping. The verification of the SRS includes correctness, consistency, clarity, and understandability of requirements⁷⁶.

4. **Requirement management and planning**, this last phase is a critical aspect of monitoring the changes of agreed requirements and their relationships, and dependencies between the requirement documents and other documents that are produced in the software engineering process. **Requirement traceability** is an essential part of requirement management. It involves the ability to describe and track the life of a requirement, as well as its relationships with other development artifacts⁷⁶. The management of changing requirements can be achieved by implementing a requirements change management process. This process involves the necessary procedures and analysis for proposed changes should be in place²⁵.

Collected challenges		
1. Lack of clear understanding of system requirements by customer	10. Poor identification of stakeholders	19. Conflict detection and resolution
2. Requirements change and evolution	11. Lack of ensuring stakeholders' satisfaction	20. Requirements analysis challenge
3. Lack of sufficient domain knowledge	12. Poor modeling of functional requirements	21. Difficulty in selecting suitable metrics
4. Tight project schedule	13. Poor requirements traceability	22. Difficulty in improving RE within the organization
5. Reuse of existing design in wrong context	14. Poorly defined specifications	23. Abstaining from allocating resources to RE
6. Poor communication link	15. Failure to prioritize requirements	24. Organizational instability
7. Cultural differences	16. Incomplete requirements	25. Poor compliance with RE processes
8. Requirements ambiguities	17. Poor requirements risk management	26. Estimate uncertainties
9. Long feedback cycles	18. Formalizing requirements from natural language	27. Change in technology

Table 4. RE challenges⁹⁸

3.1.2 Software designing

Software designing can be defined as “the activity of transforming requirements specifications into a technically feasible solution”⁸. Software design connects the definition of requirements with the implementation & development. This phase involves three phases that includes interface design, architectural design and detailed design²⁵. The **interface design** involves specifying the interactions between the system and its external environment. It should ensure efficient operation and control in the interaction between user and machine. Next, in the **architectural design**, a blueprint of the system is created. The blueprint describes in detail the components of the system, their properties, interfaces, responsibilities, and interrelationships. This phase is crucial in creating a stable framework for the system, defining its overall structure. Finally, the **detailed design** phase focuses on the internal workings of the primary system components, properties, relationships, processing algorithm and data structures. The aim is to ensure that each component operates efficiently within the systems’ architecture⁶. There are various issues that need to be addressed when going through these phases of software designing, these are shown in table 5.

Key issues	Definitions
Concurrency	Organizing software into processes, tasks and threads to optimize efficiency and synchronization.
Control and Handling of Events	Organize data and control flow, and handle both reactive and temporal events.
Data Persistence	Concerning how to handle long-lived data. Such as storage, accessibility, etc.
Distribution of Components	Focus on the distribution across different hardware components, such as computers and network hardware. It aims to understand the way how components communicate.
Error and Exception handling and Fault Tolerance	How to avoid, tolerate and deal with errors, and how to deal with exceptional conditions.
Interaction and Presentation	How to structure and organize interactions with users and the delivery of information.
Security	Prevent unauthorized access or modifications, ensuring system integrity and resilience against attacks.

Table 5. Design issues²⁵

The Twin Peaks model (Figure 8) illustrates the parallel and iterative development of requirements and architecture, starting from basic initial versions that are gradually refined⁶⁸. Because the model keeps the requirements and architecture separate, it allows for their parallel development even in an early stage. This model acknowledges that the processes of requirements engineering and software architecture cannot be performed in isolation³⁷. In an effective application, high level requirements are aligned with high level architecture, and detailed requirements are aligned with detailed architecture⁹³. This model aligns well with agile principles by enabling iterative development, which is further explained in section 3.2.2.

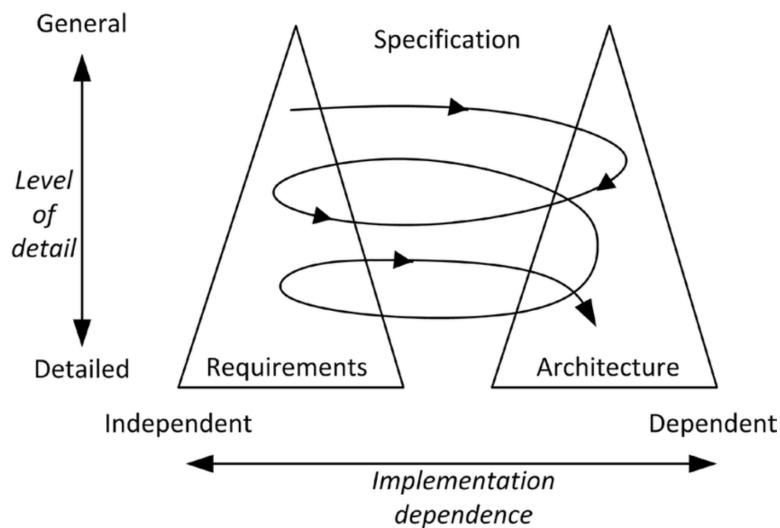


Figure 8. Twin Peaks model⁶⁸

3.1.3 Software implementation

The implementation phase of the SDLC is where developers translate the design and requirements from previous stages into working software. This is the stage where ideas are transformed into reality, and challenges within integration often arise. The paper of Parviainen et al.⁷⁸ talks about changes at several levels caused by the adoption of digital technologies within an organization. Starting with the **process level**, which includes reducing manual steps by integrating new digital tools and streamlining processes. At the **organizational level**, the focus is on offering new services and offering old services in new ways. The **business domain level**, puts emphasis on changing roles and value chains within the ecosystem of the business. Lastly, the **societal level**, which aims to change societal structures such as type of work and the influence on decision-making. The paper of Machado et al.⁵⁹ states numerous challenges towards increasing digital transformation in an organization, which are shown in table 6.

Challenges	
1. Improve automation of business processes.	6. Distributed decision-making systems.
2. Reengineering existing business models.	7. Integration of the organization structure.
3. Competitive pressure to change.	8. Workforce with different ages.
4. Finding the right technology.	9. Lack of digital skills.
5. Balance between KPIs	10. Make a system that all partners can use.

Table 6. Challenges towards increasing digital transformation⁵⁹

3.1.4 Testing

The testing phase, is a phase in which the software undergoes an evaluation to ensure that it meets its intended purpose and performs reliably. Various techniques and methodologies are used for these evaluations to ensure the quality and reliability of software systems. Software testing is performed at the different phases of the SDLC. The levels are based on the object of testing, also called the target²⁵. The most common types of testing are module, integration, system, and acceptance testing¹³. These levels are illustrated in the V model in figure 9.

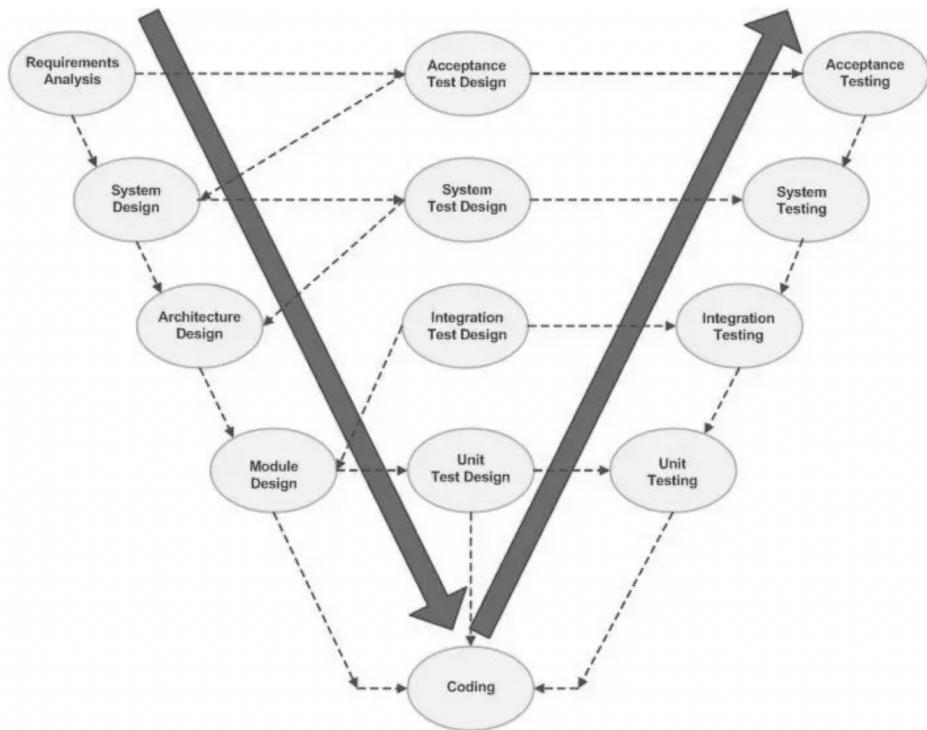


Figure 9. Testing phases in the V model¹⁶

Unit testing is about testing the elements or components of a system in isolation, to check whether the single element behaves as expected. Thorough testing of elements it is essential to identify and eliminate defects and eliminate faults that may be challenging and costly to remove in the future development phases²⁵. **Integration testing** concerns the process of examining the interactions between software components. Typical testing integration strategies, such as top-down and bottom-up testing, are commonly used in hierarchically structures software systems²⁵. While unit and integration testing can provide confidence in the quality of individual mod-

ules and their interactions, it cannot guarantee the behavior of the overall system¹³. **System testing** assesses the behavior of an entire system, and is typically used to evaluate non-functional system requirements²⁵. System testing confirms that the overall system corresponds to its specifications, while **acceptance testing** verifies the system's compliance with the user requirements. The paper of Bertolino¹⁶ mentioned several challenges within the testing of software. These challenges are shown in table 7.

Challenges	Definition
1. Overcoming resistance to adoption	People find it hard to adopt new methods.
2. Model creation and refinement	Figuring out how to create models from the results of tests.
3. High costs of tests	Test are expensive.
4. Cataloging solutions	Known solutions are often hard to find.
5. Training of software testers	Skilled testers are difficult to find.
6. Monitoring software	Keeping an eye on software after it's been deployed to catch any errors or performance issues that arise when it's actually being used.
7. Understanding different testing needs	Recognizing that developers, service providers, and those who integrate systems might have different needs and processes when it comes to testing software.

Table 7. Challenges in Software Testing¹⁶

3.1.5 Maintenance

In recent years, software maintenance has become a critical part of technology management, with a strong emphasis on improving its effectiveness. This is mostly due to the high costs involving the maintenance of software. But despite the high costs, it is often overlooked in terms of priority. In organizations, software development is often prioritized over software maintenance¹⁰. Maintenance usually starts alongside the development and continues until the software system goes into retirement or migrates, as shown in figure 10.

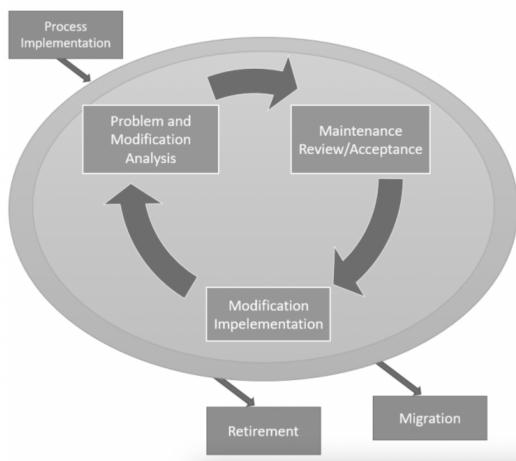


Figure 10. Software maintenance process¹⁰

Software maintenance comes in four categories: preventive, corrective, adaptive and perfective. **Preventive maintenance** involves preparing the software for future updates and minimizing potential functional errors. **Corrective maintenance** is concerned with identifying and fixing bugs or problems within the software. **Adaptive maintenance** is needed when software needs to be adapted due to changes in hardware, operating systems, or database management systems to ensure compatibility and functionality. **Perfective maintenance** aims to enhancing the software's features and performance to meet involving requirements^{10,25}. The paper of Ali et al.¹⁰ states the most demanding maintenance issues, which are defined in table 8.

Challenges	
1. Enhancement Demand.	6. Amount of availability maintenance programmers.
2. Demand time of programmer.	7. Commitments for meetings.
3. Quality of documentation.	8. Deficiency of user considerate.
4. Impractical user anticipation.	9. Insufficient user training.
5. Competence of the system design requirements.	10. Programming quality.

Table 8. Most demanding maintenance problems¹⁰

3.2 Software engineering process models

Due to the need for better structure and process for software development, software process models have been developed. The two main models for the software development life cycle are the Waterfall model and the Agile model. Each offering different advantages and disadvantages^{30 25}. Which are summarized in table 50.

3.2.1 Waterfall model

The **Waterfall model** is characterized by its sequence of phases, as can be seen in figure 11. The model starts with a detailed planning phase where all the requirements are clearly defined and documented. This phase is most important as it lays the foundation for the entire development project by specifying what needs to be done before any actual development or implementation work begins. After the planning phase, the project progresses through a number of different phases. These phases include system design, implementation, testing, deployment and maintenance. Each phase must be finished before proceeding to the next^{11 30}.

The Waterfall model brings advantages in terms of both management and predictability. By having a detailed plan, project managers can allocate resources, schedule and anticipate on the project needs before starting the design phase and implementation of a project. Therefore, offering a sort of starting plan⁹⁵. However, the model also has its limitations. It is inflexible by design, making it difficult to adjust to changes in project requirements or scope once the process is underway¹¹. When changes become necessary, they can be difficult and costly to implement, particularly if the need for change arise in the later stages of the project. In addition, the sequential structure of the Waterfall model makes it possible for errors or problems that are identified in one phase can affect the stages below, leading to delays and

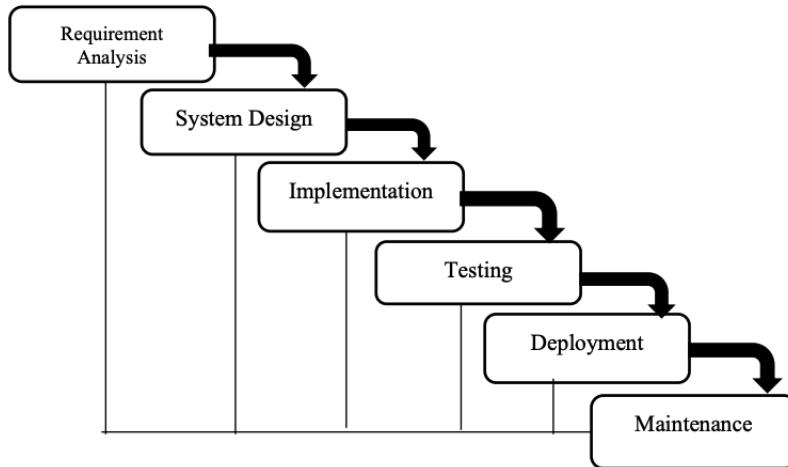


Figure 11. Waterfall model⁷

therefore increased project costs⁵⁸. The Waterfall model reflects a traditional management approach, where decisions and requirements are communication from top management down to the operational level with minimal interaction between the development team and the stakeholders. This can lead to difficulties in developing a final product that meets the needs and expectations of the intended user³⁰.

3.2.2 Agile model

The Agile model, with methods such as Scrum or Kanban, differs from the more traditional planning approaches such as the Waterfall model. Instead, agile methods are based on a predictive approach, where projects evolve progressively, with close collaboration between the project team and the customer in short cycles⁹⁵.

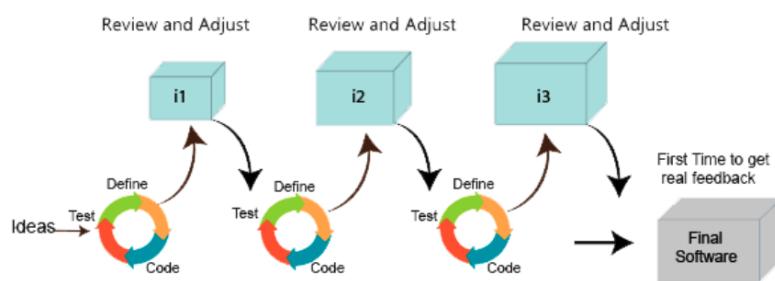


Figure 12. Agile model⁹⁰

This model has several advantages. Firstly, its adaptability enables rapid responses to the changing of customer needs. Secondly, by stimulating continuous quality improvement, it helps to ensure that the final product meets its expectations. Thirdly, early bug detection helps to mitigate risks and improve overall product robustness. In addition, the agile approach enables faster time to market, which could give an organization a competitive advantage. Finally, its consumer-oriented approach ensures that the product is aligned with user requirements⁷⁷. However, it also comes with some disadvantages. A big challenge is its resource-intensity, this model requires a high level of involvement and team coordination. In addition, the emphasis on speed within the agile model sometimes lead to lack of documentation and thorough testing, which can cause problems for future maintenance, and the robustness. The agile process also relies heavily on customer feedback, which in some projects can become a bottleneck. Finally, there is also the potential for loss of scope, due to the iterative nature of agile projects⁷⁷.

The Hybrid approach, combines the plan-driven approach with agile principles, therefore combining the strengths of both methods. While the overall framework is established through a plan-driven waterfall process, agile principles guide the management of individual subprojects. Using frequent communication and feedback loops, a hybrid approach utilizes the transparent and adaptable characteristics of Agile, within the overall structure of a project plan⁹⁵.

3.3 DevOps

Software development relies on many tools and information systems to manage data and processes. As these technologies have become more powerful and essential to the development process, the management, and maintenance of these tools has become a challenge for many teams. In most organizations, development teams participate in the software development process, while independent operations teams manage the required tools and technologies. This separation of specialists leads to communication difficulties between development and operations staff, as well as challenges in organizing and prioritizing efforts to support an efficient and effective software engineering environment²⁴. The software development methods described in section 3.2 focus on the development (Dev) of information systems, but they often neglect the operational aspects (Ops) of software deployment in a production environment³⁶. **DevOps** brings development and operations together within the software engineering process. Using automated processes for development, deployment, and infrastructure monitoring. It involves cross-functional teams collaborating on a continuous stream of operational

feature tasks, moving away from the traditional approach of separate specialized groups³¹. DevOps combines agile methodologies with operations. Both Agile and DevOps share the idea of obtaining fast feedback. But while agile focuses on quick reviews and adjustments at every sprint, DevOps extends this concepts by including continuous integration, deployment, and delivery as can be seen in figure 13. One could argue that Agile is a subset of DevOps that revolves around the practice of collaborative development. But while Agile and DevOps share similar goals, it is important to know they are different from each other. A project can be Agile without DevOps practices, or use the waterfall method while including DevOps strategies⁴².

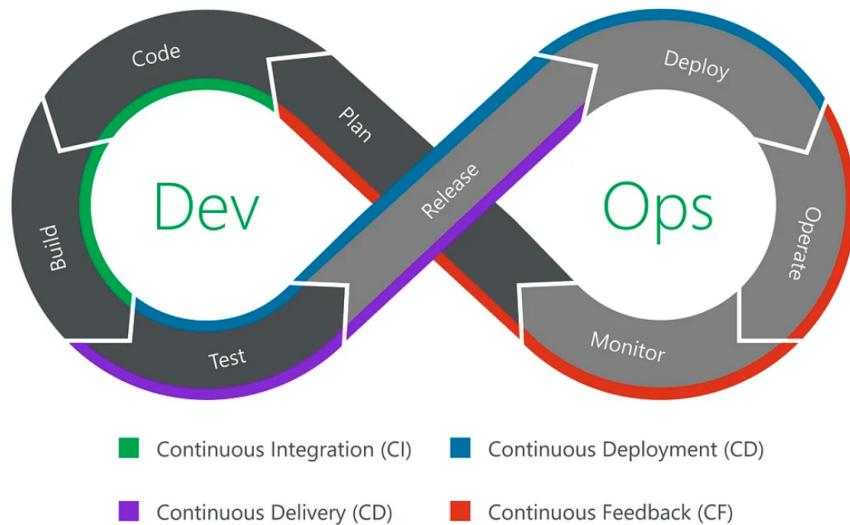


Figure 13. DevOps⁸⁰

4 AI Engineering

As the innovation in the field of **Artificial Intelligence** (AI) continues to accelerate and deliver new products and services, it's important that these systems are developed trustworthy and responsibly through their life cycle³.

AI engineering is the software engineering of AI-enabled systems and is growing as a separate discipline⁶². Despite the growing demand for AI application development, only around half of the projects reach completion and enter production. This low success rate highlights the current challenges in AI software development⁴⁹. Two major challenges faced in AI development can be identified from several scientific papers. The first is **the complexity of AI technology**. It is essential to remember that AI-based software does not behave like regular software. According to Heyn et al.⁴⁵, AI-based software has limited or no insight into the system's inner workings, which introduces uncertainty. Ahmad et al.³ notes that these characteristics create new challenges and necessitate the adaptation of existing tools and methods. The second issue concerns **the overlap of tasks among different stakeholders**, this involves coordinating various opinions and knowledge between stakeholders to establish a clear path forward²⁷. As AI application development is a complex process involving multiple stakeholders, numerous decisions need to be made. For this reason, planning is particularly important in projects involving diverse stakeholders⁴⁹. All the challenges regarding AI engineering require adapting the current SDLC framework to meet the requirements and uncertainty of AI-based systems. This adaptation revolves around the stages of the SDLC cycle to ensure that the AI components are reliable, ethical, and aligned with user and legal expectations³⁹.

4.1 Roadmap for responsible AI

Responsible AI refers to the ethical development of AI systems for the benefit of people, society, and the environment⁵⁵. The paper of Lu et al.⁵⁴ provides a software engineering roadmap for operationalizing responsible AI, as shown in figure 14. The roadmap highlights the critical importance of IT governance in achieving compliance with ethical regulations and responsibilities in the development and use of AI systems.

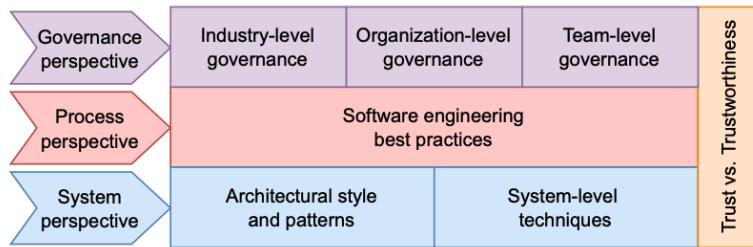


Figure 14. Roadmap responsible AI⁵⁴

As can be seen in figure 14, the roadmap is divided into three different perspectives:

- **Governance perspective.** Governance for responsible AI systems can be defined as the structures and processes used to ensure that the development and use of AI systems are consistent with the principles of AI ethics⁵⁵.
- **Process perspective.** This process incorporates responsible AI principles. It is important to involve stakeholders throughout the lifecycle of AI systems.
- **System perspective.** This perspective includes consideration of factors such as transparency, explainability, human control of technology, and accountability. Culturally sensitive data and context should be explicitly considered in the design.

4.2 AI governance

This chapter explores the **governance perspective** required for the development of responsible AI based systems. AI has become an issue of governance due to its increasing adoption in various applications in both the private and public sectors. The governance of AI is defined as “a system of rules, practices, processes, and technological tools that are employed to ensure an organization’s use of AI technologies aligns with the organization’s strategies, objectives, and values; fulfills legal requirements; and meets principles of ethical AI”¹⁷. According to Shneiderman⁹², governance can be built at three levels:

1. **Industry-level.** This level requires governments and the AI industry to act collaboratively through regulation, policy, and standards to make AI systems acceptable for society⁵⁴.
2. **Organizational-level.** This level requires ethical principles, training, diverse accountability mechanisms, and continuous monitoring to ensure the responsible development, deployment, and operation of AI technologies⁵⁴.
3. **Team-level.** This level is mainly responsible for managing the AI projects and the overall development processes⁵⁴.

The paper of Lu et al.⁵⁵ identified the **stakeholders** for responsible AI governance for each level, which is shown in figure 15. It is important to note the effects that AI systems have on potential stakeholders. The paper of Deshpande and Sharp²⁸, explores these effects of AI systems on three different stakeholder levels. Some stakeholders contribute to the development of AI systems, but may lack training in ethical and moral principles. Other stakeholders are influenced by factors such as financial performance, stakeholder value, and competitive market pressures, which affect their priorities regarding the behavior of AI systems. Legal and regulatory obligations shape the way certain stakeholders respond to AI system developments. To make effective decisions in the software development of AI systems, it is important to understand the different levels of influence that these stakeholders have, as well as their interests²⁸.

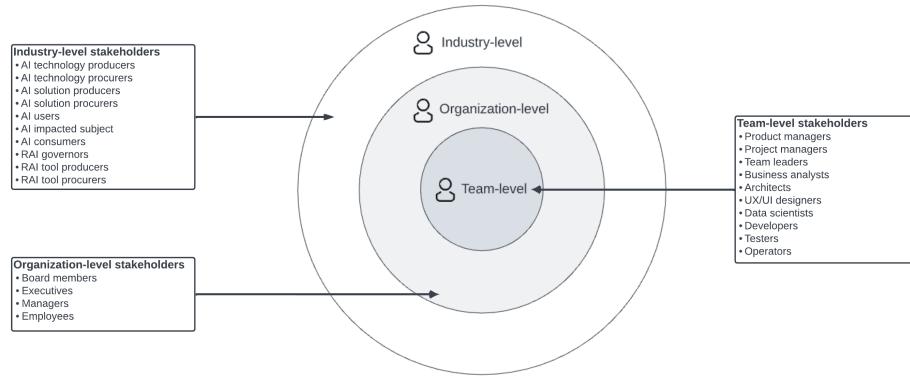


Figure 15. Stakeholders RAI governance⁵⁵

Besides the stakeholders, also a set of governance patterns are identified, which can be seen in figure 16. As can be seen, each level plays an important role in ensuring that AI technologies are developed and deployed in ways that are ethical, transparent, and accountable⁹².

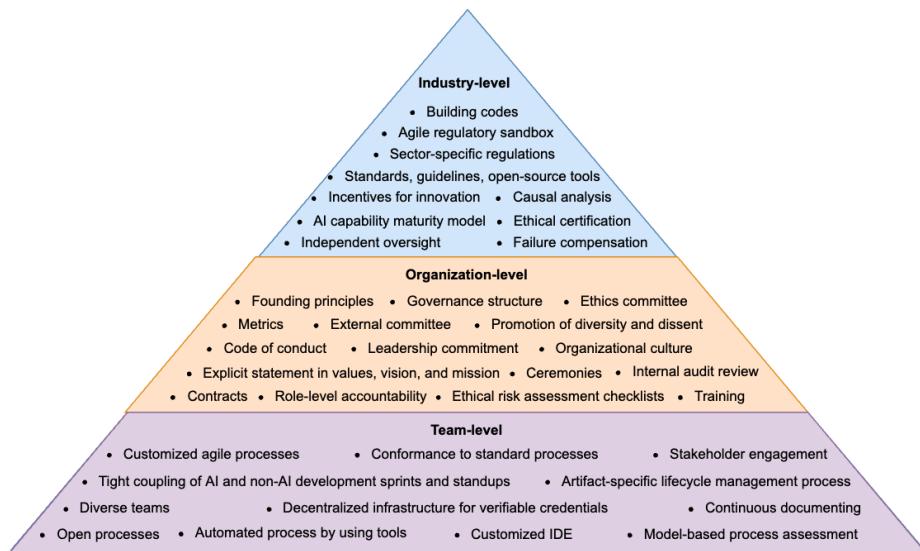


Figure 16. Governance levels for responsible AI⁵⁴

4.3 SDLC for responsible AI

In the paper of Lu et al.⁵⁵ the process patterns for responsible AI are highlighted, these can be seen in figure 17. These steps include: **requirement engineering**, **design**, **implementation**, **testing**, and **operation**. A noticeable difference between the traditional Software Development Life Cycle (SDLC) and the SDLC for responsible AI systems can be seen in the name of the final phase. In a traditional SDLC, the final phase is 'maintenance'. In contrast, the SDLC for responsible AI ends with an 'operation' stage. Besides the name difference, every stage has its own challenges compared to the traditional SDLC.

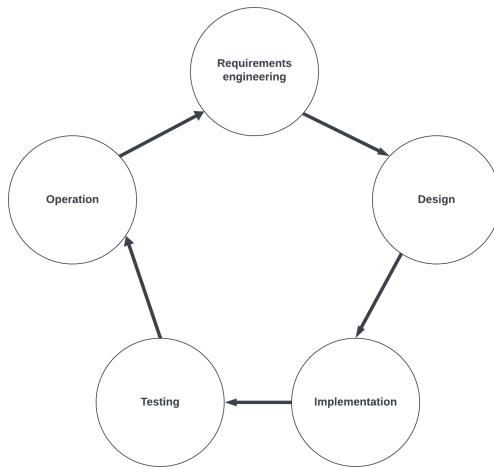


Figure 17. Responsible AI SDLC⁵⁵

This chapter will dive into these different aspects of the responsible AI SDLC compared to the traditional SDLC to provide practical insights into the development of AI based systems based on the software engineering principles.

4.3.1 RE4AI

The traditional stages of RE include elicitation, specification, verification & validation, and management, as mentioned in section 3.1.1. Requirement engineering for AI based systems (RE4AI) requires a unique approach to each of these phases. The paper of Sadovski et al.⁸⁷ and Lu et al.⁵⁴ mention several practices for each of these phases:

1. **Requirement elicitation & analysis** needs to take into account the dynamic nature and complex dependencies of AI. The require-

ment analysis should address challenges such as conflicting requirements, trade-offs, data model implications, culture safety and ethical considerations. In order to meet the specified requirements.

2. **Requirement specification** converts stakeholder input into a set of concrete ML algorithms, training methods and data requirements. Additionally, efficiency, scalability, and availability metrics are defined. Also, the Scope of responsibility needs to be clearly defined and ethical requirements need to be measurable or quantifiable.
3. **Requirement verification & validation** helps ensure that the system meets its goals, even with new and untrained data. Additionally, it is important to do continuous validation during operation.
4. **Requirement management** must focus on the evolving nature of AI to adapt to changing trends and user needs.

As stated before, Requirements are either functional (FR) or non-functional (NFR). NFR for AI systems have to adapt to include the important aspects of transparency, trust, privacy, security, and reliability. This, to make sure that these systems are not only working but also accountable and reliable⁵. The paper of Lu et al.⁵⁴ states several **requirement elicitation techniques** that can be adapted to gather requirements for the development of responsible AI. These include: interviews, scenarios, requirement workshops, interactive demos/prototypes, and user stories.

The paper of Ahmad et al.⁴ did an analysis on the current literature on RE4AI and pointed out the following challenges:

- Overestimating the capabilities of AI solutions.
- AI requirements are hard to specify.
- AI requirements are vague or high level.
- Limitation of existing RE techniques to manage AI requirements.
- Capturing and specifying trade-offs that might arise.
- Issues with data requirements.
- Difficulties in understanding NF requirements in AI systems.
- Difficult to understand the feasibility of AI models and the outcomes they can and cannot provide.

In addition, the paper of Belani et al.¹⁵ pointed out several challenges. Table 9, shows the challenges which are linked to the AI related entity, as well as the given RE activity. These RE activities include requirement elicitation & analysis, specification, verification & validation, management, and documentation.

RE4AI		AI-related entities		
RE activities		Data	Model	System
Elicitation & Analysis	Availability of (large) datasets, Imbalanced datasets	Lack of domain knowledge, Automation tools needed	Regulation not clear, No integration end results	
Specification	Data labeling is costly	No end-to-end pipeline support	Avoid design anti-patterns	
Verification & Validation	Critical analysis on training data, various data dependencies	Entanglement, scalability issues	Debugging issues, interpretability	
Management	Experiment management	Difficult to log and reproduce	Resource limitations, cost, and managing performance	
Documentation	Data visualization	Education and training of staff	Correct development method	
All of the above	Data privacy and safety; data dependencies.			

Table 9. RE4AI challenges¹⁵

4.3.2 AI software designing

To bridge the gap between the general software design phase and the extra challenges regarding AI based software design, it is necessary to examine how the traditional design issues intersect with the emerging challenges posed by AI technologies. To tackle these challenges in AI engineering design effectively, it is essential to adhere to certain principles. The paper of Fjeld et al.³³ highlighted eight key themes of **AI principles**, which are shown in table 10. These AI principles act as guidelines which aim to ensure AI development is ethical, respects rights, and benefits society. One strategy for addressing these principles is to adopt **Explainable AI (XAI)** in the design phase. As stated by Lu et al.⁵⁴, this can increase the level of trust within AI systems. XAI can be implemented in the following ways:

- The integration of human realism into the interface design.
- Proactively informing users regarding the use of data.
- Providing the benefits of using AI to the user.
- Providing transparency about the capabilities, limitations, and scope of use of the AI system.
- Assure users of the reliability and expertise behind the AI technologies.
- Providing detailed explanations of the data, algorithms, models, and rationale behind system decisions and behaviors.

The paper of Gezici and Tarhan⁴⁰ highlights the significant design challenges in AI applications, particularly in medicine and autonomous vehicles. In these cases, it is essential to obtain accurate, reliable, and secure machine learning responses quickly. Designing such systems require accurate definition of the algorithms, data, and system infrastructure to meet the specific system requirements, such as reliability, adaptability, and maintainability. It is important to pre-select which quality attributes to prioritize to ensure that the designed system meets its requirements.

AI Principle	Definition
(1) Privacy	AI systems need to protect the privacy of individuals. It requires the careful management of data during system development and the ability of individuals to control their data and decisions.
(2) Accountability	Responsibility of AI systems need to be fairly assigned.
(3) Safety and security	AI systems need to operate safely and securely, which means that they are protected against unauthorized access or tampering.
(4) Transparency and explainability	Assure that the development and operations of AI systems allow for oversight. This by making their processes understandable and providing usability information.
(5) Fairness and non-discrimination	AI systems need to promote fairness and inclusiveness, while highlighting the issues of bias.
(6) Human control of technology	Making sure that the AI system does not override human judgment by keeping significant decision-making processes under human oversight.
(7) Professional responsibility	Recognizing the crucial role of professionals in AI development and implementation, which includes their commitment to ethical standards, stakeholder engagement, etc.
(8) Promotion of human values	Align the AI system's purpose and methods with fundamental human values, aiming to enhance overall human well-being.

Table 10. *AI principles*³³

4.3.3 AI software implementation

In the SWEBOK guide, ethical implementation standards are stated²⁵. These are collections of implementation rules from the ethical perspective. Including communication standards to ensure updated, high-quality documentation for both AI and non AI components⁵⁴. However, the implementation of AI-based systems brings additional challenges. The study of He et al.⁴⁴, highlights several key issues with the implementation phase of AI algorithms within the medical domain. One of the main obstacles is the complexity of adapting AI algorithms for clinical use, which involves managing large amounts of data, incorporating technology into existing workflows, and complying with regulatory requirements. Additionally, there is a significant need for data sharing to facilitate AI development and enhancement, which raises concerns about patient confidentiality, informed consent, and cybersecurity. The medical field also requires compliance with strict certification standards like the food and drug administration (FDA) or n vitro diagnostic regulation (IVDR) approval, which can also cause challenges that may delay or prevent the adaption of AI. Implementation obstacles, such as data integration, system interoperability and compatibility with existing hardware and software equipment also add complexity. Finally, when implementing AI, professionals need to be trained to effectively use these AI systems. This requires an understanding of AI's capabilities and limitations³⁴. In the Dutch society, concerns are raised, because of the implementation of algorithms in the public sector. For example, algorithms that use AI to detect fraud and assess risks on social media are not immune from biases and discriminatory tendencies. Therefore, the Dutch government created The Algorithm Register (AR), a publicly available database explaining algorithms that are being planned, created, implemented, or retracted by the government²⁰. The main goal of the AR is to offer transparency on algorithms that impact individuals and businesses, helping to understand how the government works. The government aims to establish trust among the public by ensuring that algorithms align with societal values and norms⁷⁵. All these challenges highlight the complexity of implementing AI systems in an ethical, cost-effective, and acceptable way. In table 11, the challenges regarding AI implementation are summarized.

Challenge	Description
1. Complexity of adapting AI algorithms	Managing large amounts of data, incorporating technology into existing workflows, and complying with regulatory requirements.
2. Data sharing	Facilitating AI development and enhancement while addressing concerns about data confidentiality, informed consent, and cybersecurity.
3. Compliance with certification standards	Meeting strict certification standards like FDA or IVDR approval, which can cause delays or prevent AI adoption.
4. Data integration, system interoperability, and compatibility	Challenges related to integrating AI systems with existing hardware and software equipment.
5. Professional training	Ensuring company employees are effectively trained to use AI systems, requiring an understanding of AI capabilities and limitations.
6. Biases and discriminatory behavior	Concerns about algorithms, such as those used to detect fraud and assess risks on social media, not being immune from biases and discriminatory tendencies.
7. Transparency and trust	Establishing transparency through initiatives like The Algorithm Register (AR), which aims to offer insight into government-created algorithms impacting individuals and businesses to align with societal values and norms.

Table 11. Challenges in AI Implementation

4.3.4 AI software testing

The testing of AI based systems bring new and unprecedented difficulties since the outcome of these systems depends on both code and training data⁴⁰. The paper of Aleti⁹ defines several challenges that are introduced with testing AI systems. These are summarized in table 12.

Challenge	Definition
Oracle Problem	Establishing a single ground truth can be difficult due to the subjective and diverse outputs, which can complicate comparison and evaluation.
Contextual Appropriateness	Producing grammatically correct and contextually relevant content is a challenging task that requires an understanding of semantics and pragmatics. Assessing the quality of such content is difficult to quantify.
Unpredictability	The production of varied outputs based on input variations or conditions challenges conventional testing methods in terms of covering all scenarios and assessing reliability and compliance.
Emergent Behavior	Determining correctness and achieving consensus on generated content can be complicated by the occurrence of behaviors that were not present in the training data.
Data contamination	Large datasets used to train text-generating AI often inadvertently include questions and answers from publicly available tests. This makes it appear that the AI is performing better than it actually is.

Table 12. Challenges of testing AI systems⁹

Besides the different challenges that AI brings, the papers of Aleti⁹ and Lu et al.⁵⁴ also highlight that automated testing is becoming important due to the enormous range of possible test inputs that need to be generated and evaluated. Testing should focus on the following areas to ensure reliable functioning AI systems: correctness, robustness, security, fairness, model relevance, comprehensibility, robustness, security, fairness, and efficiency. There are several techniques to test AI systems on these categories. The techniques include metaphoric testing, Cross-referencing and differen-

tial testing, reference-based techniques and search-based techniques. Prompt crafting is also stated as an essential step of the testing process, this is for effectively assessing the behavior and capabilities of AI systems. By crafting meaningful and representative prompts, testers can gain deeper insights into the operational nuances of the systems⁹. The process for finding the best prompts is called **prompt engineering**, which is further explained in section 5.1.1.

4.3.5 AI software operations

Operations is the final phase in the responsible AI SDLC. This phase differs from the maintenance phase of the traditional SDLC in that it highlights the need for **continued management and oversight** of AI systems once they have been deployed. Activities in this phase are described in the paper of Lu et al.⁵⁴ which include deployment strategies, monitoring, validation, ethical risk assessment, co-evolution of system components and maintaining transparency and accountability. Two important challenges within this phase are highlighted in table 13. One of these challenges is data smells, which are suspicious features in data, such as unusual values or patterns, that indicate potential problems without relying on specific contexts³⁵. The paper of Foidl et al.³⁵ shows that bad practices (i.e., data management and handling) within AI engineering leads to data smells, which can negatively impact the proper functioning as well as the development and maintenance of AI-based systems. Another challenge is highlighted by the paper of Poenaru-Olaru et al.⁸³, which talks about the problem of **concept drift** in machine learning models. This problem refers to a situation where the underlying pattern or relationship in the data changes overtime, the model's output becomes inaccurate because it is searching for patterns that do no longer exist. To solve this problem, **Concept drift adaptation techniques** can be used to maintain models where the data is constantly changing over time. The most common adaptation method is model retraining, which can be done using either **the sliding window approach** or **the full history approach**. The sliding window approach updates the model with the most recent data and discards the old data. The full history approach retrains the model using all available data during the model update, assuming that the more data allows the model to generalize better⁸².

Challenge	definition
Data smells	Suspicious features in data, such as unusual values or patterns, that indicate potential problems without relying on specific contexts.
Concept drift	The underlying pattern or relationship in the data changes overtime.

Table 13. AI operations challenges

5 Generative AI

Generative AI(GenAI or GAI) is a subset of AI that aims to create new content rather than simply analyzing or interpreting existing data. A brief explanation of AI and its sub-fields, Machine Learning (ML) and Deep Learning (DL), is given to distinguish GAI from other AI concepts (Figure 18).

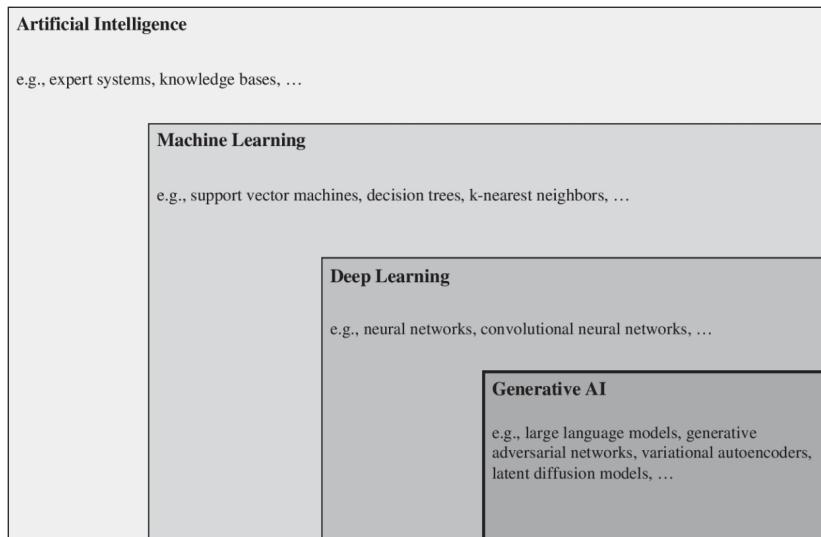


Figure 18. AI concepts¹²

AI is a broad area of computer science that focuses on creating systems that can perform tasks that require human intelligence, such as natural language understanding, pattern recognition, decision-making, and learning from experience. ML, is concerned with the development of algorithms that learn autonomously from data without explicit programming, thus improving their task performance through data exposure. DL, a subset of ML, uses multi-layered artificial neural networks to model data in complex ways, draw-

ing inspiration from the human brain to identify patterns in large datasets. This is crucial for processing complex data types such as images and sounds. GAI, is an emerging field of AI that has evolved through the advances in DL and ML techniques. Unlike the standard AI models that focus on identifying patterns and making predictions, GAI focuses on understanding the underlying structure of data to generate new content such as text, image, and audio¹².

5.1 Large language models

Large Language models (LLMs), are a subset of GAI systems and have significantly changed the field of **Natural language processing** (NLP)⁶⁵. Language models (LMs) are computational models that can comprehend and generate human language. They have the ability to predict the probability of word sequences or produce new text based on a specific input. LLMs are an advanced class of these models, characterized by their substantial parameter sizes and exceptional learning capabilities²³. These models are trained on huge volumes of data, enabling them to understand complex patterns and nuances in language more effectively. Recent advances in large language models include **transformer architectures**, which utilize so-called self-attention mechanisms to determine the relevancy of different parts of the input when generating predictions. This allows the model to improve understanding of word relationships within sentences, regardless of their position. Another significant improvement is pre-training, where language models are trained on large datasets before fine-tuning for specific tasks, which has proven effective across various language tasks⁴⁸. Bidirectional Encoder Representations from Transformers (BERT) is an example of a model that can be fine-tuned for tasks such as sentence classification and question answering²⁹. Another recent development is ChatGPT, which has been trained on extensive web data and shows promising performance across diverse natural language tasks⁹⁴. LLMs have already demonstrated considerable potential in the medical domain, which can be found in the paper of van Zandvoort et al.¹⁰⁰. This innovative approach can accelerate task completion and reduce errors associated with manual administrative procedures. The use of LLMs to automate administrative tasks has the potential to optimize operational workflows and improve overall system efficiency.

5.1.1 Prompt engineering

Numerous articles emphasize the importance of **prompt engineering**, which is the technique for effectively crafting prompts to receive the best outcome through a LLMs. The level of detail and specificity in the prompts has a direct impact on the model's understanding and ability to produce the expected results¹⁰⁰. Prompts enable pre-trained models to be easily integrated into various tasks by directing the model's behavior based on the prompt provided, eliminating the need to adjust the model's parameters. There are several prompt engineering methods that are applied for specific application areas⁸⁸.

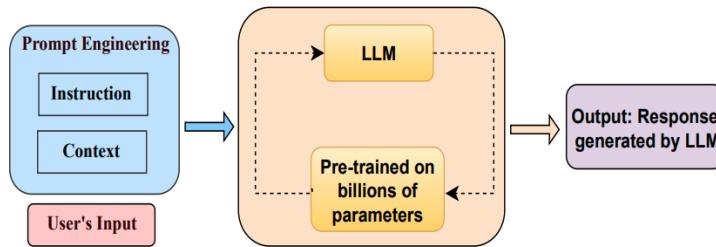


Figure 19. Visualization of prompt engineering components⁸⁸

Shot prompting is a method used to direct the model's output and consists of three different strategies: zero-shot prompting, one-shot prompting, and few-shot prompting¹⁰⁰. **Zero-shot prompting**, also known as direct prompting, relies on thoughtfully crafted prompts that guide the model towards a unique task⁸⁵. This way of prompting does not provide the model with explicit examples, instead it relies on the model's pre-existing knowledge from its training.

1. Write a police report based from the transcript.

Figure 20. Example of a zero-shot prompting

¹⁰⁰. **One-shot prompting** and **few-shot prompting**, provide the model with one or more examples 'shots' of inputs and/or outputs to help it understand specific tasks. The addition of adding examples has been shown to improve the performance of the models on complex tasks when comparing it to providing it with no examples at all²¹.

Pattern prompting, involves using different patterns that can be chosen as a basis of the formulation of prompts. This technique aims to capture the knowledge into pattern structures to enhance the ability to apply

- ```
1 1. Write a police report based from the transcript.
2 2. Use the format [Name]: [question or answer].
3 3. The name = "Officer" or "Suspect".
```

*Figure 21.* Example of a few-shot prompting

it in different context and domains where users encounter the same problems. This technique has the benefit of reuse and adaptability of these patterns for different use cases and situations<sup>104 100</sup>.

## 5.2 Structured report generation

There are several papers on structured reporting through GAI within in the medical field, providing valuable insights into the techniques and ways for improving the efficiency and accuracy of documentation. The paper of Yan et al.<sup>108</sup>, proposed a two-step approach for radiology report generation. First extracting content from an image, and then verbalizing the content into a report that matches the style of a specific radiologist. For the second step, they leveraged the GPT-3.5-turbo model from OpenAI to generate the report. In a Manual evaluation, clinical experts were unable to distinguish real reports from AI-generated ones that were adapted to the writing of the radiologist. Their quantitative evaluation concluded that the proposed approach showed beneficial performance. Also, the paper of van Zandvoort et al.<sup>100</sup> studied the effectiveness of LLMs in generating automated medical reports using different prompting strategies. In that study, the two-shot prompting approach with scope and domain context outperformed the other methods and achieved the highest score when compared to human reference.

## 6 The workings of the Dutch Police

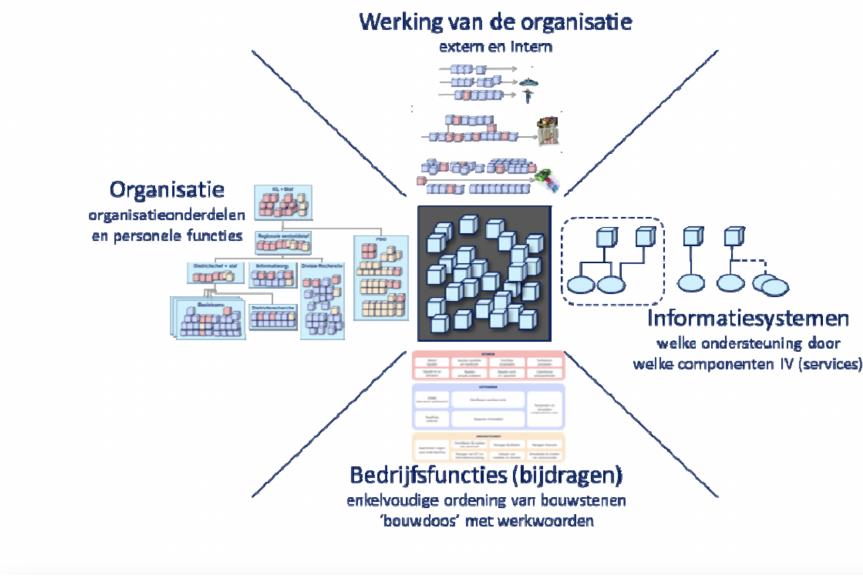
The enterprise architecture of the Dutch police consists of several layers also known as the **BIAT** layers. The **business layer** covers the organizational structure, the products, and services from the national police, required business functions, and the processes and activities at various levels to fulfill these functions. The **information layer** describes the business objects and their relationships, data, and models. Its purpose is to ensure that the information provision is build from general components and meet the operational needs efficiently and flexibly. The **application layer** describes the required functionalities of the applications, apart from the actual application components. These are described in the technical layer. The **technical layer** provides the setup of the technical architecture and specifies the functions that the infrastructure provides to the application layer. Using modern ICT concepts such as cloud computing and big data for efficient, flexible and reliable ICT services<sup>64</sup>. Overall, the enterprise architecture plays a crucial role in ICT projects within the law enforcement by providing guidance, ensuring compliance, and providing flexibility. It should guide design choices for leveraging new possibilities and provide a reliable foundation for work processes and collaboration with external partners<sup>64</sup>.

### 6.1 Reference model for police business processes

The reference Model for Police Business Processes (RBP), shown in figure 22, is essential for standardizing operational efficiency within the Dutch police. Standardization is important for achieving the main objectives of the police, which include increasing operational effectiveness, strengthening unity, and enhancing legitimacy and trust. The functions of the RBP as a model can be described as follows:

- It serves as a gathering place for process standards, beginning with language
- It is a significant source for functional requirements and an alignment framework for ICT
- It is a landing site and source of inspiration for local improvements.

The RBP is crucial for achieving the goals of the National Police, especially in the field of ICT research. Standardized processes and requirements are needed for the successful implementation and integration of systems<sup>84</sup>.



*Figure 22. The reference Model for Police Business Processes<sup>84</sup>*

## 6.2 Reporting within the Dutch police

A successful arrest is only truly successful when the court also convicts the apprehended suspect. Therefore, a thorough and well-documented police report is essential for this purpose. The police report serves as both an investigative account and evidence. The report is written by the involved law enforcement personnel, describing what happened and the powers they exercised. It presents key facts from the investigation, including both incriminating and exonerating information, to the judge during the trial<sup>26</sup>. The literature on police reporting is limited, the research that does exist indicates that administrative tasks, such as writing reports and data entry, have become increasingly stressful for police officers<sup>60</sup>. Boivin and Gendron<sup>18</sup> highlighted this issue in a study regarding the effect of body-worn cameras on police reporting. Despite the fact that this technology showed potential to make police reporting more effective, the integration of body-worn cameras increased the administrative workload. This was a critical limitation of their use. The research of Shane<sup>91</sup>, showed that administrative burdens have more impact on police officers than operational stress in the field. This highlights the importance to examine current practices and develop strategies aimed at reducing the administrative demands placed on police personnel.

### 6.2.1 Types of interrogations

Police manuals and instructions encourage interrogators to begin interrogations with open-ended questions about what happened. The idea is that open-ended questions will make suspects feel more comfortable and more likely to tell their own version of the incident. This allows the interrogator to report the suspect's 'own words', making it more difficult for the suspect to retract his statement later. In addition, the length of the answers to open-ended questions provides the interrogator with the material to ask new questions<sup>50</sup>. However, an open question ignores the fact that such questions can produce broad answers that may not provide legally sufficient evidence. In addition, the answers to open-ended questions may be too long to be accurately recalled and recorded immediately. In several cases of the study of Komter<sup>50</sup>, police officers begin with an open-ended question about the event without recording anything, only to piece together and document the story later.

In Dutch interrogations, the police is required to manage dialogue and documentation simultaneously. Typically, an interrogation is done by one or two police officers. The paper of Komter<sup>50</sup> compares the difference between two reporting styles:

1. **Solo interrogation**, where a single officer alternates between questioning and typing. This process typically follows a question-answer-typing sequence. The police officer has a routine of asking a question, getting an answer, and then typing it up. This requires the suspect to wait while the officer types. However, suspects sometimes continue to elaborate on their answers while the officer is typing. This results in a more parallel activity of talking and typing.
2. **Duo interrogations**, where it is most of the time a division of tasks between the two officers. This results in different forms of communication and response. Typically, one officers types while the other focuses on questioning, allowing for simultaneous conversation and typing. However, these interrogations don't always ensure accurate reporting. This is because the questioning officer must coordinate with the typing officer by repeating important information and pausing to allow for typing. However, these repetitions and pauses don't always ensure that the typed record accurately reflects the spoken words.

Interrogations are not just about capturing the truth, but also about creating a recordable narrative that helps build a case. Regardless of the interrogation style, the police transcript is always a summary of the interview, focusing on the content rather than the interaction<sup>50</sup>.

### 6.2.2 Requirements of a Police Report

The important procedural requirements for documenting and reporting criminal offenses in the Netherlands are covered in Articles 152 and 153 of the 'Wetboek van Strafvordering' (WvSv)<sup>70 71</sup>. These articles establish a framework for documenting criminal activities with transparency, reliability, and adherence to legal procedures in the investigative process. Article 152 highlights the obligation of officials involved in criminal investigations to immediately draw up a report, known as a 'proces verbaal (PV)', which is an abbreviation for police report in the Netherlands, detailing the offense discovered or the investigative measures taken. Article 153 provides guidelines for preparing the PV. Officials must sign the created report or affirm the report before a deputy public prosecutor to ensure the report's credibility. The report must be personally created, dated, and signed by the officials, with explicit mention of the sources of their knowledge. What these articles do not say is what these sources of knowledge are. The paper of Danhof and Boon<sup>26</sup>, commissioned by The Hague's police-prosecution quality program, explains a short guide on how to write a good police report under the Dutch justice system. This guide emphasizes the sources of knowledge, particularly the central role of the seven golden W's<sup>2</sup>, which are shown in table 14.

|                  |                                                        |
|------------------|--------------------------------------------------------|
| <b>What</b>      | What exactly happened?                                 |
| <b>Where</b>     | Where did the criminal offense take place?             |
| <b>When</b>      | When did the criminal offense occur?                   |
| <b>Who</b>       | Who were involved in the criminal offense?             |
| <b>Why</b>       | Why the criminal offense occurred (the motive).        |
| <b>How</b>       | In what manner did the criminal offense take place?    |
| <b>With what</b> | The means by which the criminal offense was committed. |

*Table 14. The seven golden W's*

The study by Jansen<sup>46</sup>, examines shortcomings identified in the written police reports of Dutch police officers. This document provides a summary of the requirements for a police report, taken from both literature and legal sources. The requirements are classified into four main categories: writing, layout, content, and submission of a police report. In table 51 these requirements are shown and described. The lack of specific requirements on the structure of a police report can lead to a variety of different report formats. In the study conducted by Hamers et al.<sup>43</sup>, an examination of 400 police reports on witness interviews revealed that only 17% followed the recom-

---

<sup>2</sup>Please note that one of these W's is lost in translation

mended standards, indicating a considerable prevalence of different report structures in practice. The diversity of these reports is illustrated in table 15. This highlights the need for a more standardized approach to improve the consistency and effectiveness of police reporting practices.

| Document variants                                                | Percentage |
|------------------------------------------------------------------|------------|
| A completely literal representation                              | 2%         |
| A question-answer format with editing applied                    | 25%        |
| A business summary from the perspective of the reporting officer | 1%         |
| A perspective from the witness                                   | 16%        |
| A businesslike presentation in the witness's own words           | 38%        |
| A combination of styles within the same report                   | 18%        |

**Table 15.** Document variants of police reports (based on Hamers et al.<sup>43</sup>)

## **7 The Case study**

### **7.1 Envisioning & ideation**

Envisioning & Ideation is to start a project based on the ideas and curiosities of a project team<sup>89</sup>. This study will do this by building upon the work performed by the Care2report research program. This research program focuses on the development of automated medical reporting software. The Care2report research program has resulted in a start-up called Vertical, which is developing the software. This study envisions adapting this technology to automatically generate structured police reports. Making it possible to take advantage of transferable functionality from the medical field while addressing the unique needs of law enforcement. This is a possible solution to the problem statement of section 1.1, which explains numerous challenges within the Dutch law enforcement. These challenges raise concerns about the efficiency, quality, and administrative burden associated with the manual creation of police reports.

### **7.2 Feasibility checking**

The feasibility check reminds designers to assess whether the project is feasible, based on aspects such as costs, competencies, skills, and access to the necessary information and relevant data<sup>89</sup>. Thus, before going into the requirement engineering and design phase, it is important to investigate the feasibility of this project. The feasibility of this project is ensured by several factors. Firstly, this study can use existing software from Vertical for automatic report generation. This software requires only minor adaptations, a task well within the resources and timeframe of this study. Secondly, Vertical covers any potential costs of using their software, therefore removing a financial barrier. Lastly, the Dutch police have given their full cooperation by providing access to data in the form of interviews and the expertise of their AI pre-development team.

### 7.3 Artifact building

The development of the artifacts was the first step within the iterative design activities within this study. Eventually, two artifacts were developed:

1. **The prototype for automatic police reporting.**
2. **A new developed approach for requirement engineering activities with continuous development of the created prototype**

The decision to create the prototype prior to performing the requirement engineering activities offered numerous benefits. Such as:

- **Proof of concept**, the prototype demonstrates the feasibility of the proposed solution. It helps to assess whether the intended functionality can be achieved with the available technology and resources.
- **Clarification of requirements**, stakeholders can better visualize the final product, which provides better feedback.
- **Improved communication**, the prototype helps with communication between the developers and stakeholders. Resulting in less misunderstandings and mismatched expectations.
- **Direct implementation**, some requirements could be implemented directly after the interviews. Resulting in a more refined prototype for the next interviews.

This case study involved a vision-oriented approach. This means that when this study began, a prototype was already developed and available. This prototype was from Verticai, a company focused on enhancing reporting processes in the healthcare domain. Verticai provided access to their software for the purposes of this study. This system came with a fully designed user interface (UI) and several functionalities. Firstly, **the automatic transcriber**, which allows users to start a recording and have the spoken words automatically transcribed into written text. The transcribed text can then be utilized as input, enabling further processing. Second, **controllable use of AI**, the software provides a pre-developed environment where users can construct different models tailored to specific tasks, which are then connected to the API of ChatGPT. These models use the transcribed data as input to generate specific outputs defined by the model's configuration. This allows users to create their own customized models for various applications, such as generating police reports.

### 7.3.1 Creation of the prototype

For the creation of the prototype, a specific model for the generation of police report data is developed with the Verticai system. The structure of these models are described in section 7.4.1. The first tests with the created report model showed that the already existing functionalities were insufficient to generate a fully structured police report. This was due to the fact that the generated data from the system could only be exported via a JSON<sup>3</sup> file, making it impossible to directly export a structured document. As a result, two additional processes were needed to enhance the system's capabilities. These consisted of creating the report structure (step 1) and mapping the output from the Verticai system in the correct places within the structured report (step 3). All the steps needed to create the generated police report are shown in figure 23. It is important to know that during the case study, we did not have access to the backend (internal code) of the Verticai software. Therefore, the developed solution uses separate code files.

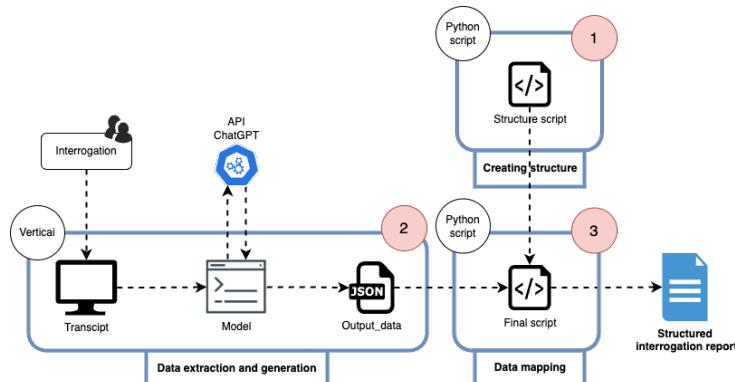


Figure 23. Prototype process

- 1. Creating report structure.** Before using the Verticai software, it was important to establish the correct format for an interrogation report, as the output had to match this structured format. The structured interrogation report (figure 24) was developed by looking at examples provided by the Dutch police.

---

<sup>3</sup>JSON is a light-weight data format based on the data types of the JavaScript programming language<sup>81</sup>



## Procesverbaal - Verhoor rapport

Politie:

District:

Afdeling:

Eenheid:

Mutatie\_nr:

Dossier\_nr:

Ik, verhalende, verklaar het volgende:

### VERDACHTE

Naam:

Geboren op:

Geboren te:

Adres:

Woonplaats:

Enmanszaak:

Inschrijf\_nr KVK:

### OVERTREDINGSGEGEVENS

Pleeg datum:

Plaats delict:

Feit:

### VERKLARING

Verdachte erkent het feit:

Verdachte wenst te schrikken:

Samenvattende verklaring van de verdachte:

### SCHIKKINGSVOORSTEL

Aangegeven welk schikkingsvoorstel, i.o.m met welke parketsecretaris, is aangeboden aan de verdachte

### Aanvullende gegevens

Verdere informatie:

### AFSLUITING

Waarvan door mij is opgemaakt dit proces-verbaal, dat ik sloot en ondertekende op .

op ambtseed

Handtekening

*Figure 24. Created structure for police report*

The creation of a structured PDF format was made possible by the use of Python and the ReportLab library. This approach reduces the AI's workload, as the structured PDF is generated through a separate process. The script uses ReportLab to define the structure and style of the report, adding elements such as headers, paragraphs, and lines in a predefined format, of which a code snipped is shown in figure 25.

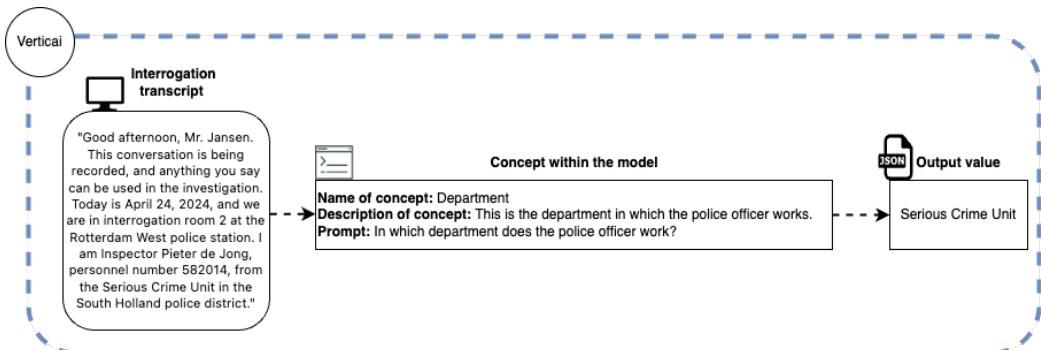
```

1 story.append(Paragraph("VERDACHTE",
2 styles["Heading2"]))
3 story.append(Paragraph("Naam: {}"
4 .format(data[6]), styles["BodyText"]))
5 story.append(Paragraph("Geboren op: {}"
6 .format(data[16]), styles["BodyText"]))

```

*Figure 25. Code snipped: Creating report structure*

2. **Data extraction and generation.** This process starts with a recording of an interrogation. The speech-to-text software within the system will convert the audio automatically into a transcript. This transcript is processed through the created model that utilizes ChatGPT to extract or generate specific variable data points. Figure 26 shows an example of the process how the Verticai software generates variable data from the transcription. The output value can ultimately be extracted from the Verticai system into a JSON file.



*Figure 26. Process: Transcription to output value*

3. **Data mapping.** The mapping of the extracted data consists of filling the placeholders in the report template with the appropriate data that is extracted from the JSON file. An example of a generated report can be found in figure 27

**POLITIE**

**Procesverbaal - Verhoor rapport**

Politie: Zuid-Holland  
 District: Zuid-Holland  
 Afdeling: Rotterdam-West  
 Eenheid: 12  
 Mutatie\_nr: 1573897  
 Dossier\_nr: 12

Ik, verbaalsant Pieter de Jong, 582014, Inspecteur, Zuid-Holland, verklaar het volgende:

**VERDACHTE**

Naam: Lucas Johannes Jansen  
 Geboren op: 13 maart 1985  
 Geboren te: Den Haag  
 Adres: Molenstraat 45, 3011 XD Rotterdam  
 Woonplaats: Rotterdam  
 Rechtspersoon: [Niet aanwezig]  
 Eenmanszaak: Klussenbedrijf Jansen  
 Inschrijf\_nr KVK: 987654321

**OVERTREDINGSGEVEGENS**

Pleeg datum: 15 april 2024  
 Plaats delict: Havenstraat in Rotterdam  
 Feit: Betrokkenheid bij een inbraak

**VERKLARING**

Verdachte erkent het feit: Nee  
 Verdachte wenst te schrikken: Nee

Samenvattende verklaring van de verdachte: Verbaalsant: Kunt u uw volledige naam en geboortedatum bevestigen voor de oproeping? Verdachte: Ja dat kan ik. Mijn naam is Lucas Johannes Jansen en ik ben geboren op 13 maart 1985 in Den Haag. Verbaalsant: U woont op de Molestraat 45, 3011 XD Rotterdam, correct? Verdachte: Ja, dat is correct. Verbaalsant: U bent hier vandaag omdat u verdacht wordt van betrokkenheid bij een inbraak die plaatsvond op 15 april 2024 aan de Havenstraat in Rotterdam. Kent u deze locatie? Verdachte: Ja, die straat ken ik wel. Verbaalsant: Waar was u op de avond van 15 april 2024 rond 10 uur avonds? Verdachte: Ik was gewoon thuis, alleen. Ik heb de film gekeken die avond. Verbaalsant: Er zijn camerabeelden waarop iemand die op u lijkt in de buurt van de havenstraat te zien is, rondom de tijd van de inbraak. Kunt u uitleggen hoe dat dan komt? Verdachte:

Poeh, dat moet iemand anders zijn geweest. Ik ben die avond niet uit huis geweest eigenlijk.  
 Verbaalsant: We hebben ook vingerafdrukken van u gevonden op de plaatselect. Die komen overeen met die van u. Hoe verklaart u dat? Verdachte: Uhm, ik weet niet hoe dat kan. Ik heb niets met die inbraak te maken. Verbaalsant: Gezien de bewijzen die wij hebben, zou een schikking een optie zijn om deze zaak te beslechten. Wij stellen een schikkingsofordel voor van 5.000 euro. Bent u bereid om te schikken? Verdachte: Nee, ik wil niet schikken. Gewoon omdat ik onschuldig ben. Verbaalsant: Heeft u nog iets toe te voegen aan uw verklaring of wilt u iets wijzigen? Verdachte: Nee, ik blijf bij mijn verklaring.

**SCHIKKINGSVOORSTEL**

Aangegeven welk schikkingsofordel 5000 euro, i.o.m met welke parketsecretaris, is aangeboden aan de verdachte

**Aanvullende gegevens**

Verdere informatie: De verdachte kwam zenuwachtig over

**AFSLUITING**

Waaran door mij is opgemaakt dit proces-verbaal, dat ik sloot en ondertekende op 24 april 2024.  
 Pieter de Jong op ambtseed  
 Handtekening

*Figure 27. Generated police report*

### 7.3.2 Prototype driven RE4AI

Requirement engineering activities ensure that there is a clear understanding of what needs to be built and why, laying the basis for successful software development. Within the performed case study a structured approach to requirement engineering is designed and performed, whereby the traditional activities are repeated across specific phases. During these phases, the prototype will be developed continuously. The phases are as follows:

1. **Value exploration.** This phase involves gathering high level requirements. Which includes understanding the most important needs and concerns of the user's and organization. It focuses on gathering the functional and non-functional requirements.
2. **Obstacle & opportunity identification.** This phase involves identifying the possible obstacles and opportunities with developing the software. The process of identifying obstacles involves recognizing any potential barriers that affect the development of the software negatively. Such barriers could include legal challenges, resource constraints or integration issues. The identification of opportunities involves detecting the beneficial factors that could improve the development process. This could include technological advancements, or new solutions to existing problems.
3. **Solution discovery.** This phase focuses on exploring potential solutions that align closely with stakeholder needs and project objectives. This is done by gathering in-depth knowledge about the technological environment in which the software will operate, and how to develop the software in a responsible way.

The requirement engineering activities start with the requirement elicitation, which consists of conducting interviews with different participants within the Dutch law enforcement and the supplier of the software used. Meetings with the research group and the police have resulted in a number of participants that need to be included. These participants are linked to a phase within the process. Table 16, shows each phase, their perspectives with corresponding participants that will be interviewed.

| Phases                               | Perspectives                                 | Participants                                                       |
|--------------------------------------|----------------------------------------------|--------------------------------------------------------------------|
| <b>Value exploration</b>             | User<br>Organization                         | Police officer<br>Innovation expert and Domain architect           |
| <b>Obstacle &amp; opportunity ID</b> | Ethics<br>Legal<br>Environment<br>Supplier   | Ethics expert<br>Legal experts<br>Solution architect<br>Co-founder |
| <b>Potential solution discovery</b>  | Responsible AI<br><br>AI Safety<br>Technical | Responsible AI and X4AI expert<br>Safety expert<br>Data scientist  |

**Table 16.** Phases, perspectives, and their participants

The developed approach is visualized in figure 28. This approach helps organize the different opinions and knowledge between stakeholders in order to lay out a clear path forward. To stay in the scope of this study, only one iteration of the process has been done. Therefore, only a few stakeholders are interviewed to gather the necessary requirements for the prototype. After each phase, the gathered data is analyzed, this data is then used to continuously improve the prototype's design and to prepare for the following interviews. This process ensures that each subsequent version of the prototype becomes increasingly aligned with the needs and expectations of the relevant stakeholders, thereby allowing them to provide more precise feedback on each phase.

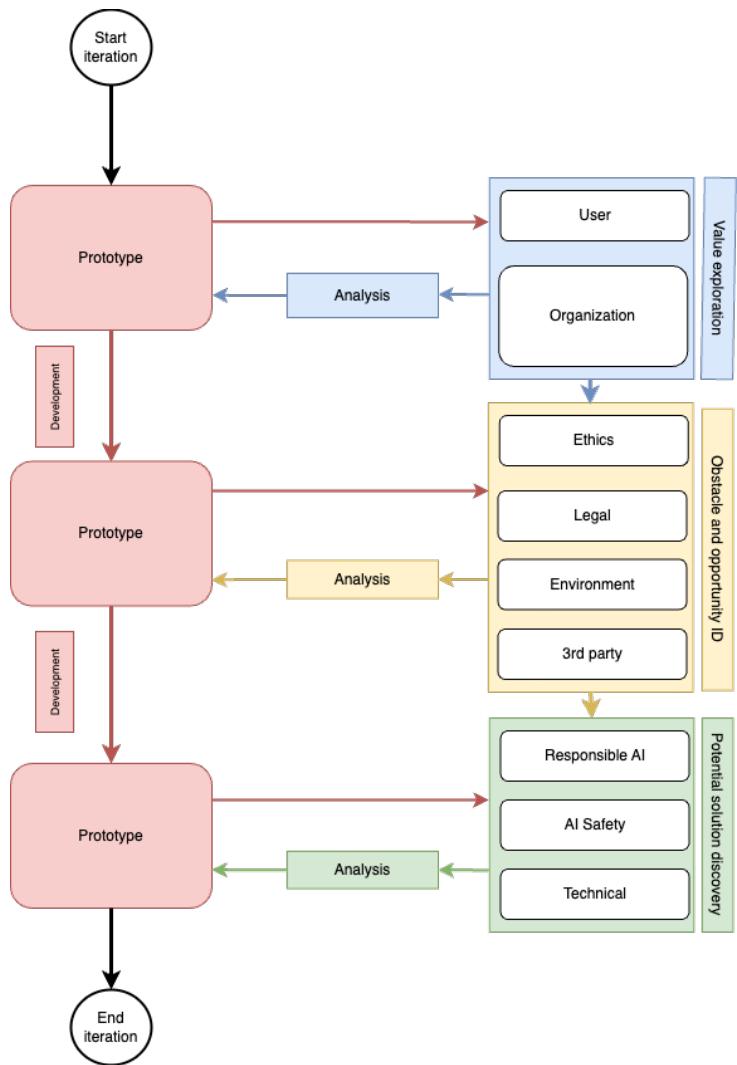
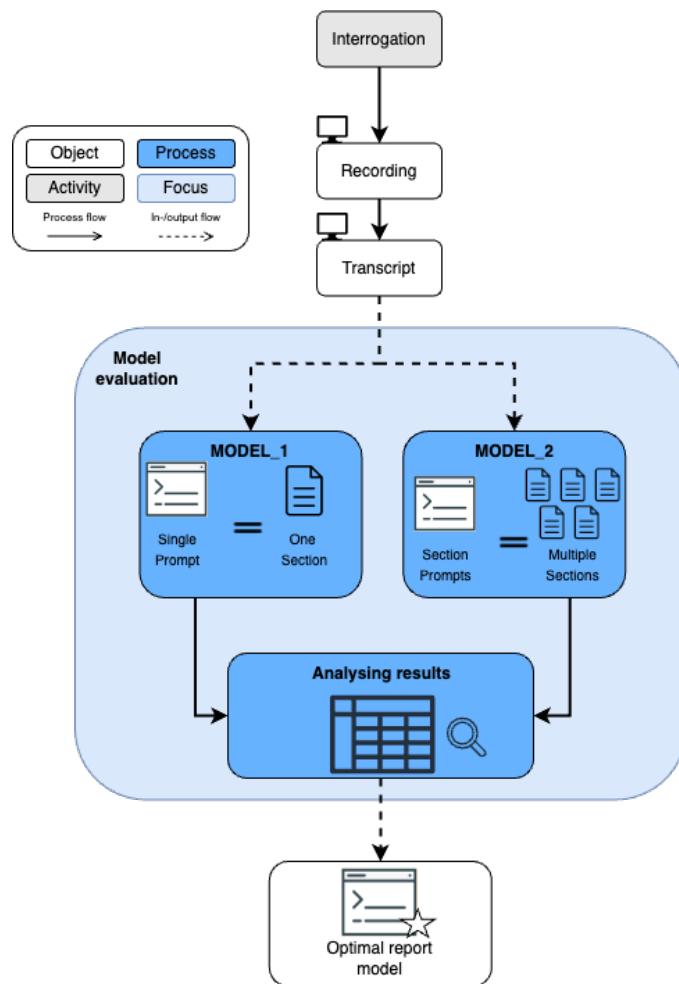


Figure 28. Iteration of RE and design process

## 7.4 Ex-ante evaluation & justification

As mentioned in the previous section 7.3, the Verticai system provides a functionality to build report models. These models are used to guide the integrated large language model into generating a specific output from a transcription. Within this study two models were created and evaluated. A comparative research is conducted to evaluate which reporting model performs best in meeting the requirements highlighted in table 51. The aim of this evaluation is to obtain an early assessment of Verticai's system based on these requirements. Both models will be continuously developed within this study. A visualization of the evaluation method used is shown in figure 29, this figure is inspired by van Zandvoort et al.<sup>100</sup>.



*Figure 29. Evaluation method visualization*

To start with the evaluation, scripted dialogues<sup>4</sup> were created and recorded to simulate a solo interrogation between an interrogator and a suspect. These dialogues were based on 5 different scenarios:

- TR1: Burglary
- TR2: Illegal garbage dump
- TR3: Cyberattack
- TR4: Drugs
- TR5: Financial fraud

This resulted in five transcriptions based on these scenarios, generated through the speech-to-text software. The transcription of these dialogues are used as input data for both models. Resulting in ten different police reports. This approach allowed an evaluation of each model with different contexts. The models are evaluated by examining the output that is generated. Two different evaluation techniques are being used to assess the generated outputs.

#### 7.4.1 Structure of the models

The first reporting model uses a single prompt to generate the entire report. This means that all the information including, names, descriptions, and instructions for each concept is placed into one single prompt. The model then processes this large chunk of data in one go, producing the complete report in a single output. The second reporting model divides the prompt into five detailed sections, which are generated separately. These sections are described in table 17. In this model, the report is broken down into smaller parts. Each section focuses on a specific aspect of the report. By generating these sections separately, the model can focus on each part independent of one another.

---

<sup>4</sup>The dialogues are scripted due to the limitation of obtaining real interrogation data.

| Section               | Description                                                                          |
|-----------------------|--------------------------------------------------------------------------------------|
| Police                | Information about the police officer and location of the interrogation.              |
| Suspect               | Information about the suspect of the interrogation.                                  |
| Criminal offense (CO) | Information about the criminal offense of which the suspect is accused.              |
| Statement             | The statement of the suspect in question answer format.                              |
| Settlement proposal   | Information about the settlement proposal if that is discussed in the interrogation. |

*Table 17. Section description of Model\_2*

Both models use the same names, descriptions, and prompt instructions for each concept. Table 18,19,20,21, and 22 show all the concepts that have been used for both report models. The only difference is the structure of the input: one model assembles everything into a single prompt (section), while the other model separates the content across multiple prompts (sections). Besides evaluating the performance of both models, we look if this separation will improve clarity, focus, and potentially the overall quality of the generated report.

| Name             | Description                                               | Prompt instruction                                                     |
|------------------|-----------------------------------------------------------|------------------------------------------------------------------------|
| Region           | This is the region where the police is located.           | In which region is the police located?                                 |
| District         | This is the police district.                              | What is the police district, also known as the location of the bureau? |
| Department       | This is the department of the police officer.             | What is the department of the police officer?                          |
| Name of officer  | Name of the officer that is conducting the interrogation. | What is the name of the police officer?                                |
| Rank             | This is the rank of the officer.                          | What is the rank of the police officer?                                |
| Personnel number | This is the personnel number of the officer.              | What is the personnel number of the police officer?                    |

*Table 18. Police Data*

| Name            | Description                            | Prompt instruction                                                                     |
|-----------------|----------------------------------------|----------------------------------------------------------------------------------------|
| Name of suspect | This is the name of suspect.           | What is the name of the suspect?                                                       |
| Place of birth  | This is the place the suspect is born. | Where is the suspect born?                                                             |
| Date of birth   | This is when the suspect was born.     | Based on the transcription, when is the suspect born?                                  |
| Suspect address | This is the suspect address.           | What is the address of the suspect?<br>Include street name, number, postal code, city. |

**Table 19.** Suspect Data

| Name                    | Description                                 | Prompt instruction                                                                                                                  |
|-------------------------|---------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------|
| Date of the offense     | This is the date when the offense occurred. | What is the date when the criminal offense occurred as mentioned in the transcription. If nothing mentioned, keep the answer empty. |
| Location of the offense | This is where the offense occurred.         | Where did the criminal offense occur? If nothing mentioned, keep the answer empty.                                                  |
| Offense                 | This is the type of offense.                | Which type of criminal offense is the suspect accused of?                                                                           |

**Table 20.** Criminal Offense Data

| Name                | Description                                     | Prompt instruction                               |
|---------------------|-------------------------------------------------|--------------------------------------------------|
| Settlement proposal | This is the amount of money proposed to settle. | What amount of money is the settlement proposal? |

**Table 21.** Settlement Data

During the Manual evaluation, which is explained in section 7.4.3, an important aspect within the system of Verticai was discovered. When creating the reporting model, the formulation of the concepts significantly influenced the generated statement. The concept structure, which consists of name, description, and instruction, showed that the description takes priority over the instruction within the AI model. Therefore, the entire prompt was placed within the description, while the instruction remained empty, as can be seen in table 22.

| Name                                                       | Description                                                                                                                                                                                                                            | Instruction                                                                                  |
|------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------|
| Did the suspect admit to the fact?                         | This shows whether the suspect admits to the offense (yes or no).                                                                                                                                                                      | Extract from the transcript whether the suspect admits to the offense. Answer with yes or no |
| Does the suspect want to settle?                           | This shows whether the suspect wants to settle (yes or no).                                                                                                                                                                            | Extract from the transcript whether the suspect wishes to settle. Answer with yes or no      |
| Explanation with steps and rules in question-answer format | Extract from the transcription the questions and answers that pertain to the criminal offense. Use the format [Name]: [question or answer]. Also, place a blank line between the questions and answers. Name = "Officer" or "Suspect". |                                                                                              |

*Table 22. Statement Data*

#### 7.4.2 Precision & Recall

The purpose of using precision and recall in the context of this study is to determine the accuracy and completeness of the metadata captured by our model. Metadata refers to all the contextual and administrative information associated with the report itself, excluding the actual witness statement. The precision & recall values will indicate if the information captured is specific and correct. It measures the proportion of correctly generated metadata outputs (true positives) out of all the metadata outputs that the model has generated (both true positives and false positives). In other words, precision helps us understand how many of the metadata outputs generated by the

models are actually correct. The formula for precision in the context of this study is:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

In contrast, recall will assess the proportion of correctly identified information out of all relevant information that should have been identified. It is defined as the ratio of true positive predictions to the total number of actual positive instances. It is a measure of the completeness. The formula for recall is as follows:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

In the context of this study, these formulas can be filled in with the following definitions:

- True Positives = Correctly generated metadata
- False Positives = Incorrectly generated metadata
- False Negatives = Metadata elements that were missed by the model

Due to the use of different contexts in our transcripts, such as varying locations and names, the precision and recall values in this study tend to move together. As the model improves in correctly generating metadata from the transcripts (increasing precision), it also tends to better capture more of the overall metadata present (increasing recall).

### ***Police data***

Starting with the data of the police. Tables 23 and 24 present the evaluation of Model 1 and Model 2, based on their precision and recall in processing police data. For an example of the working of the system in generating metadata, I refer back to figure 26. Both tables demonstrated an almost perfect output generation. However, an error occurred specifically in model 1 with generating the rank of the police officer from transcript 5 (TR5) input data. This error happened due to an incorrect articulation by the interrogator during the audio recording. Interestingly, Model 2 successfully corrected this mistake.

| <b>Police data (Model_1)</b> | <b>TR1</b> | <b>TR2</b> | <b>TR3</b> | <b>TR4</b> | <b>TR5</b> | <b>Precision</b> | <b>Recall</b> |
|------------------------------|------------|------------|------------|------------|------------|------------------|---------------|
| Region                       | 1          | 1          | 1          | 1          | 1          | 1.0              | 1.0           |
| District                     | 1          | 1          | 1          | 1          | 1          | 1.0              | 1.0           |
| Department                   | 1          | 1          | 1          | 1          | 1          | 1.0              | 1.0           |
| Name reporting officer       | 1          | 1          | 1          | 1          | 1          | 1.0              | 1.0           |
| Rank                         | 1          | 1          | 1          | 1          | 0          | 0.8              | 1.0           |
| Personnel number             | 1          | 1          | 1          | 1          | 1          | 1.0              | 1.0           |

*Table 23.* Evaluation police data of model 1

| <b>Police data (Model_2)</b> | <b>TR1</b> | <b>TR2</b> | <b>TR3</b> | <b>TR4</b> | <b>TR5</b> | <b>Precision</b> | <b>Recall</b> |
|------------------------------|------------|------------|------------|------------|------------|------------------|---------------|
| Region                       | 1          | 1          | 1          | 1          | 1          | 1.0              | 1.0           |
| District                     | 1          | 1          | 1          | 1          | 1          | 1.0              | 1.0           |
| Department                   | 1          | 1          | 1          | 1          | 1          | 1.0              | 1.0           |
| Name reporting officer       | 1          | 1          | 1          | 1          | 1          | 1.0              | 1.0           |
| Rank                         | 1          | 1          | 1          | 1          | 1          | 1.0              | 1.0           |
| Personnel number             | 1          | 1          | 1          | 1          | 1          | 1.0              | 1.0           |

*Table 24.* Evaluation police data of model 2

### **Suspect adata**

The evaluation of the suspects data section is provided in table 25 and table 26. Both models showed consistent and accurate outputs. Except for the date of birth variable in Model 2, which failed to be described in any of the cases. The consistent mistake of Model 2 shows a significant shortcoming in its capabilities to generate the right answer in this context. This error could eventually be corrected by using a different prompt. This is further explained in section 7.7.

| <b>Suspect data (Model_1)</b> | <b>TR1</b> | <b>TR2</b> | <b>TR3</b> | <b>TR4</b> | <b>TR5</b> | <b>Precision</b> | <b>Recall</b> |
|-------------------------------|------------|------------|------------|------------|------------|------------------|---------------|
| Name suspect                  | 1          | 1          | 1          | 1          | 1          | 1.0              | 1.0           |
| Place of birth                | 1          | 1          | 1          | 1          | 1          | 1.0              | 1.0           |
| Date of birth                 | 1          | 1          | 1          | 1          | 1          | 1.0              | 1.0           |
| Address of suspect            | 1          | 1          | 1          | 1          | 1          | 1.0              | 1.0           |

*Table 25.* Evaluation suspect data of model 1

| Suspect data (Model_2) | TR1 | TR2 | TR3 | TR4 | TR5 | Precision | Recall |
|------------------------|-----|-----|-----|-----|-----|-----------|--------|
| Name suspect           | 1   | 1   | 1   | 1   | 1   | 1.0       | 1.0    |
| Place of birth         | 1   | 1   | 1   | 1   | 1   | 1.0       | 1.0    |
| Date of birth          | 0   | 0   | 0   | 0   | 0   | 0.0       | 0.0    |
| Address of suspect     | 1   | 1   | 1   | 1   | 1   | 1.0       | 1.0    |

*Table 26.* Evaluation suspect data of model 2

### Criminal offense data

Tables 27 and 28 focus on the evaluation of the criminal offense (CO) data. This includes the date, location, and type of the criminal offense. Model 1 shows a high level of accuracy across all categories. Achieving a perfect score for both location and type of offense. Only the date section scores a precision and recall of 0.8. This error was due to the fact that the date in transcription 5 (TR5) was not only 1 day but a time frame. Both models failed at correctly identifying this time frame. Model 2 showed a significant weakness within this section, particularly in identifying the date and location of the offense. This is due to the fact that in TR2, TR3, and TR4 no location or date of the criminal was mentioned. But instead of generating nothing, Model 2 picked the date of the interrogation itself. Eventually, this error was also corrected by using a different prompt. This is also explained in section 7.7.

| CO data (Model_1)         | TR1 | TR2 | TR3 | TR4 | TR5 | Precision | Recall |
|---------------------------|-----|-----|-----|-----|-----|-----------|--------|
| Date criminal offense     | 1   | 1   | 1   | 1   | 0   | 0.8       | 0.8    |
| Location criminal offense | 1   | 1   | 1   | 1   | 1   | 1.0       | 1.0    |
| Type criminal offense     | 1   | 1   | 1   | 1   | 1   | 1.0       | 1.0    |

*Table 27.* Evaluation criminal offense data of model 1

| CO data (Model_2)         | TR1 | TR2 | TR3 | TR4 | TR5 | Precision | Recall |
|---------------------------|-----|-----|-----|-----|-----|-----------|--------|
| Date criminal offense     | 1   | 0   | 0   | 0   | 0   | 0.2       | 0.2    |
| Location criminal offense | 1   | 0   | 0   | 0   | 1   | 0.4       | 0.4    |
| Type criminal offense     | 1   | 1   | 1   | 1   | 1   | 1.0       | 1.0    |

*Table 28.* Evaluation criminal offense data of model 2

### **Statement data**

Table 29 and table 30 show the evaluation of the statement data, indicating whether the subjects are willing to confess to the crime and whether they are inclined to accept a settlement. In Model 1 both statements had a perfect score in all transcriptions, resulting in a precision and recall of 1.0. Model 2 showed a small mistake in the confession of TR1. This resulted in a precision of 0.8. Also, this error could eventually be corrected by using a different prompt. This is also described in section 7.7.

| Statement (Model_1) | TR1 | TR2 | TR3 | TR4 | TR5 | Precision | Recall |
|---------------------|-----|-----|-----|-----|-----|-----------|--------|
| Confession          | 1   | 1   | 1   | 1   | 1   | 1.0       | 1.0    |
| Making a deal       | 1   | 1   | 1   | 1   | 1   | 1.0       | 1.0    |

*Table 29. Evaluation statement data of model 1*

| Statement (Model_2) | TR1 | TR2 | TR3 | TR4 | TR5 | Precision | Recall |
|---------------------|-----|-----|-----|-----|-----|-----------|--------|
| Confession          | 0   | 1   | 1   | 1   | 1   | 0.8       | 1.0    |
| Making a deal       | 1   | 1   | 1   | 1   | 1   | 1.0       | 1.0    |

*Table 30. Evaluation statement data of model 2*

### **Settlement data**

Table 31 and table 32 show the scores of the settlement data. This only includes the amount of money that is being offered to make a deal. Both Model 1 and Model 2 achieved a precision and a recall of 1.0, indicating perfect performance in predicting settlement amounts across all transcriptions (TR1 to TR5).

| Settlement (Model_1) | TR1 | TR2 | TR3 | TR4 | TR5 | Precision | Recall |
|----------------------|-----|-----|-----|-----|-----|-----------|--------|
| Amount of money      | 1   | 1   | 1   | 1   | 1   | 1.0       | 1.0    |

*Table 31. Evaluation Settlement data of model 1*

| Settlement (Model_2) | TR1 | TR2 | TR3 | TR4 | TR5 | Precision | Recall |
|----------------------|-----|-----|-----|-----|-----|-----------|--------|
| Amount of money      | 1   | 1   | 1   | 1   | 1   | 1.0       | 1.0    |

*Table 32. Evaluation Settlement data of model 2*

### **7.4.3 Manual evaluation of generated statements**

This study applied a Manual evaluation to test the generated statements. The inspiration for this Manual evaluation came from the paper of van Zandvoort et al.<sup>100</sup>. This paper provides a technique for Manual evaluation of generated medical reports. But instead of using medical professionals to create a human reference of a good report, this study looks at a specific law (art. 29a-2 WvSv)<sup>72</sup>, on which the generated statements will be evaluated. This law states two essential elements that must be incorporated into the formulation of a statement in a police report. These are:

1. The report should be written as much as possible in the suspect's own words.
2. The statement of the suspect is presented as completely as possible and, where possible, in a question-and-answer format.

These elements must be applied as much as possible when writing statements in a police report. The generated statements were compared to specific sections of the transcript that included information of the seven golden W's, as shown in table 14, for the reference statement. The reference transcript and output of the generated statements from both models can be seen in Appendix D. The Manual evaluation points out several findings regarding the quality of the generated statements. The quality is measured by looking at the following type of aspects:

- **Errors.** An error is a mistake or inaccuracy in the model's output.
- **Additional.** Addition refers to extra content provided in the output.
- **Transcriptional.** If the model corrects a mistake of the speech-to-text software.
- **Articulatory.** If the model corrects an articulatory mistake that was made during an audio recording.

#### ***Errors***

Table 33 shows the type of errors that are identified in the generated statements during the Manual evaluation. The only type of error encountered in both models was omission. This means that the generated statement missed relevant information. Out of five generated statements per model, there were only two or three omission errors noted in the entire evaluation. It is noteworthy that no instances of hallucination, categorization, repetition, or incorrect errors occurred. Indicating that both models consistently captured the suspect's own words.

| Type           | Definition                                             | Model 1   | Model 2   |
|----------------|--------------------------------------------------------|-----------|-----------|
| <b>Errors</b>  | <i>Total of errors in the model's output.</i>          | <b>12</b> | <b>10</b> |
| Hallucination  | Output is false or misleading.                         | 0         | 0         |
| Categorization | Incorrect classification of information.               | 0         | 0         |
| Repetition     | The unnecessary duplication of words.                  | 0         | 0         |
| Incorrect      | Errors present in the output.                          | 0         | 0         |
| Omission       | Failure to include relevant information in the output. | 12        | 10        |

*Table 33. Errors findings in generated statements*

### ***Additional information***

Table 34 shows the additional information generated by each model. Both models produced a similar amount of additional relevant content. Consisting of clarifications on who said what within the transcript. Indicating that both models succeeded into creating the question and answer format that is needed. However, Model 1 included more redundant information, which included information that was already stated in the report and therefore no longer relevant to the statement.

| Type            | Definition                                            | Model 1    | Model 2   |
|-----------------|-------------------------------------------------------|------------|-----------|
| <b>Addition</b> | <i>Total of extra content provided in the output.</i> | <b>102</b> | <b>92</b> |
| Relevant        | Additional information that adds value to the output. | 78         | 77        |
| Redundant       | Adds extra information that is unnecessary.           | 24         | 15        |

*Table 34. Addition findings in generated statements*

### ***Transcriptional***

Looking at the transcriptional errors, which made during the converting audio into text (Explained in section 7.3.1), a comparison of the two models revealed that model 1 corrected a greater number of transcription errors than model 2. Table 35 shows the performance of both models. Of the total of 10 mistakes, model 1 improved 6 of them, while model 2 only improved 2 of the 7 mistakes. The reason for why the number of transcriptional mistakes differ, is because 3 of them were not included due to an omission error of model 2.

| Type                   | Definition                                                                       | Model 1   | Model 2  |
|------------------------|----------------------------------------------------------------------------------|-----------|----------|
| <b>Transcriptional</b> | <i>Total of transcriptional mistakes within the generated transcript</i>         | <b>10</b> | <b>7</b> |
| Improvement<br>Mistake | AI improves mistake in transcript.<br>AI does not improve mistake in transcript. | 6<br>4    | 2<br>5   |

**Table 35.** Transcriptional findings in generated statements

### ***Articulatory***

With regard to the correction of articulatory errors recorded during interrogations, table 36 shows that the models performed similarly. From the 6 articulation mistakes made during the interrogation, only 1 was corrected.

| Type                   | Definition                                                                                         | Model 1  | Model 2  |
|------------------------|----------------------------------------------------------------------------------------------------|----------|----------|
| <b>Articulatory</b>    | <i>Total of articulatory mistakes</i>                                                              | <b>6</b> | <b>6</b> |
| Improvement<br>Mistake | AI improves the mistake made in articulation.<br>AI does not improve mistake made in articulation. | 1<br>5   | 1<br>5   |

**Table 36.** Articulatory findings in generated statements

#### 7.4.4 Summary of evaluation

The evaluation of both models, revealed a number of strengths, weaknesses, and limitations, as shown in table 37. Starting with the metadata evaluation, both models showed promising performance in generating metadata. Model 1 showed the highest precision and recall values, while processing all the information in a single prompt. Model 2 showed more inconsistencies in the generation of certain metadata, indicating the necessity for refinements of the prompts, which is described in section 7.7. However, a notable weakness common to both models is their inability to generate specific timeframes mentioned in the transcription. Besides that, a limitation is that the evaluation is not done on real world data.

| Strengths                                                                                                                                                                                                                                                                                                                 |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <ul style="list-style-type: none"><li>- High precision and recall in generating metadata.</li><li>- Generating question-answer format.</li><li>- Correcting transcriptional mistakes.</li><li>- Correcting articulatory mistakes.</li><li>- No hallucinations, categorization, repetition, or incorrect errors.</li></ul> |
| Weaknesses                                                                                                                                                                                                                                                                                                                |
| <ul style="list-style-type: none"><li>- Timeframe generation.</li><li>- Generates some redundant information.</li><li>- Generates some omission errors</li></ul>                                                                                                                                                          |
| Limitations                                                                                                                                                                                                                                                                                                               |
| <ul style="list-style-type: none"><li>- Not tested on real world incident interrogations</li><li>- No access to internal code of Vertical</li></ul>                                                                                                                                                                       |

*Table 37. Results evaluation*

Looking at the Manual evaluation, both models are able to generate relevant statements. With model 2 showing a slight advantage in the area of redundant errors, keeping the statements more precise. While model 1 appears to be slightly better at improving the mistakes made during the

speech-to-text transcription. Model 2 may be slightly preferable at this moment for generating statements, as it is more inline with the requirements of Art. 29a-2 WvSv to present the suspect's statement as completely and accurately as possible.

Overall, both models demonstrate promising capabilities in generating metadata and statements from transcriptions, future research should prioritize testing with real-world data. Moreover, the structure of vertical models could be enhanced to prioritize instruction over description, which could further improve the usability of the Vertical system. Additionally, improvements to speech-to-text software can reduce the inefficiencies of Model 2 in correcting transcriptional mistakes. Finally, it is necessary to determine whether AI-generated corrections of transcription or articulatory errors align with legal regulations.

## 7.5 Grounding & conceptualization

During the RE analysis, the gathered requirements are prioritized. Due to the lack of scientific literature on requirement prioritization frameworks for developing AI-based systems, this study performs a small comparison between different prioritization techniques used in the domain of traditional requirement engineering. Table 38, shows several techniques with their focus point, based on the study of Trieflinger et al.<sup>96</sup>. Two prioritization techniques have been identified as possibilities in the context of RE4AI:

- **MoSCow**
- **Weighted Scoring**

Both methods bring structure and transparency to the complex field of RE4AI. The MoSCoW method provides an efficient way to categorize the requirements in four groups: 'must have', 'should have', 'could have', and 'won't have'. This method helps to quickly identify the most important requirements of the AI system (must have), those that are relevant but not necessary (should have), those that are desirable but also not necessary (could have), those that are excluded from the current scope (won't have). The weighted scoring method allows for a more nuanced prioritization of the requirements. By assigning specific weights to various criteria like legal compliance, efficiency, usability, and the repetition of requirements. However, this approach is more time-consuming and requires a solid understanding of each criterion's relevancy. Due to the limitation of time, the MoSCoW

method has been chosen for the prioritization of requirements within this study. To categorize the requirements more effectively, this study also looked at the frequency of references to each requirement. The more often a requirement was mentioned in the interviews, the higher its priority was likely to be.

| Prioritization technique    | Focus point                                                                                                                                                                       |
|-----------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Story mapping               | Focus on “User experience” rather than the internal opinions of the product team and stakeholders.                                                                                |
| Kano model                  | Focus on “customer satisfaction” and “functionality” that a feature provides.                                                                                                     |
| MoSCoW                      | Classifies features into: “Must have”, “Should have”, “Could have”, and “Won’t have”.                                                                                             |
| Impact vs. effort           | Classifies features based on the two dimensions “impact” (value) and “effort”.                                                                                                    |
| ICE                         | Prioritize products based on three factors: “impact”, “confidence”, and “ease”.                                                                                                   |
| RICE                        | Uses the ICE model but also considers the factor “reach”.                                                                                                                         |
| Weighted Scoring            | Ranks several features according to various criteria, which are then calculated into an overall score. These criteria can have different forms.                                   |
| Weighted shortest job first | Features with a high level and short delivery time should take priority over features that take longer and deliver less value.                                                    |
| Opportunity scoring         | Uses two dimensions, “satisfaction” and “importance” and is based on customer feedback and aims to identify features that customers consider essential but are dissatisfied with. |

*Table 38. Requirement prioritization frameworks based on Trieftinger et al.<sup>96</sup>*

## 7.6 Knowledge formulation

This section provides an overview of the knowledge formulation process, describing the outcomes of the phases described in section 7.3.2. All the requirements gathered are described and prioritized. The main findings of each phase are described at the end of each section.

### 7.6.1 Value exploration

The purpose of this phase is to gather high level requirements, to understand the most important needs and concerns of the stakeholders. Following the paper off Glinz<sup>41</sup>, the requirements gathered are divided into the following types:

- **Functional requirements.** The functionality and behavior of the system.
- **Performance requirements.** Restrictions about timing, processing or reaction speed, data volume, or throughput.
- **Specific quality requirements.** A specific quality that the system or component shall have.
- **Constraints.** Any other restrictions about what the system shall do, how it shall do it, or any prescribed solution or solution element.

Background information on the participants of the interviews that took place during this phase are provided in table 39. It includes information on the work they are currently performing within the Dutch law enforcement, as well as information on their experience level (defined in years of experience).

| Ref. | Perspective  | Participant            | Work                                                                                           | Experience |
|------|--------------|------------------------|------------------------------------------------------------------------------------------------|------------|
| P1   | Organization | Implementation manager | Advisor, guides innovations within the North Holland unit                                      | 35 Years   |
| P2   | User         | Police officer         | Detective and interrogator in the investigation department within the Central Netherlands unit | 30 Years   |
| P3   | Organization | Domain architect       | Work on everything related to the deployment of AI                                             | 3.5 Years  |

*Table 39. Participants value exploration phase*

Within Dutch law enforcement, recording audio is becoming more part of the daily activities. This procedure is done within a system also known as 'Audiovisuele Verhoorregistratie' (AVR), which has recently been enhanced by the deployment of a mobile application that allows police officers to record audio on their workphones.

*P1: "The reason is the request from the Public Prosecutor's Office (OM) to record more interviews."*

*P2: "We actually recently gained the ability to record audio using our work phone"*

There is a strong interest in improving efficiency and effectiveness of police reports, with also the potential to reduce the number of police officers involved in the interrogations through transcription technologies.

*P1: "Typing can distract an officer from having an actual conversation, thereby affecting the effectiveness of the interrogation."*

*P2: "You want to stay in contact with the suspect, of course, and not turn it into a reading exercise."*

Some police officers also record the interrogation and then type up the report afterward. This process takes a lot of time. Next to that, the quality of the recording needs to be as high as possible to help with the interpretation.

*P2: "I often have to pause the recording to keep up and listen back again. So, you're also quite dependent on the quality of the conversation or the interview."*

*P2: "If I think about it for myself, processing one hour takes two to three hours to make a report from it."*

Speed is also crucial. In some contexts, time pressure to make an entire report is high.

*P1: "For some interrogations, the time frame is 3 days at best, but often it is only 24 hours."*

Speech-to-text technology is already under development within the Dutch law enforcement. However, challenges are faced in relation to the

strict legal requirements and reliability of audio. The recently introduced innovation law, enables the inclusion of a dossier containing diverse multimedia files, including visual and audio recordings, as distinct pieces of evidence. But these recordings need to be summarized in a report. This summary should include a timeline and speaker recognition, so that every piece of information can be traced back to the original audio file.

*P1: “The legal system rightly stated that this alone is not sufficient. They need a shortened official report, a kind of summary of what is recorded and where these moments can be found in the recordings.”*

*P1: “But a kind of report that includes a timeline and speaker recognition indicating who said what and when.”*

Given the time constraints, the legal system cannot afford to look through every file to find what is relevant or important and where to find it. For this reason, traceability is an important requirement for the system.

*P1: “This involves not only enriching the current official report, but also precisely examining the essence of the entire interrogation, determining where the information is located and how it should be interpreted.”*

*P3: “Also good traceability, to see where the information also came from.”*

Usability of the system is also important. Users need to be in control of the system, allowing them not only to make adjustments to the output, but also to control the AI for generating that specific output.

*P3: “Perhaps the ability for users to take a final look to see if what the system has made indeed is correct.”*

*P3: “A part of adjusting the AI model itself seems very useful to me.”*

Before an automatically generated report can be used in practice, there are still some challenges to address. Including the trust of the software itself. The software should be more accurate than the police officer, and needs to eventually show this.

*P1: “There are a lot of colleagues who don’t trust the recording. So, it really takes time before that trust is there.”*

P2: “You will have to experience it, and you should compare it with your own frame of reference. From there comes the trust.”

P3: “To ensure quality, provide that piece of assurance that the right system is right.”

Also, some information does not need to be filled in by AI, some information could be automatically filled by already existing information systems and some will be put in manually by the officer.

Overall, there is a need and interest for improving the efficiency and effectiveness in police reports. The introduction of the innovation law and the updated AVR system that enables police officers to record audio on their workphones, highlights a shift towards a more modern form of legal proceedings. The interviews within this phase resulted in several requirements for the development of the system, which are shown and prioritized following the MoSCow method in table 40. Accurate speech-to-text technology is one of the most important requirements, as it forms the foundation for generating reliable reports. The functional requirements of timelines, links/references to the original multimedia files and speaker recognition are also a must, as they ensure traceability and compliance with legal standards. Additionally, the usability of the system is important, officers need functionalities such as the ability to review and adjust the AI-generated output to keep control and trust in the system.

| ID                                   | Functional requirement                                                                                               | Prioritization |
|--------------------------------------|----------------------------------------------------------------------------------------------------------------------|----------------|
| 1                                    | The system should be able to include timelines, and links/references to the original multimedia files in the report. | Must-have      |
| 2                                    | The system should be able to identify all speakers.                                                                  | Must-have      |
| 3                                    | The system should be able to generate a question/answer format.                                                      | Must-have      |
| 4                                    | The system should be able to tell what information was generated by AI in the report.                                | Must-have      |
| 5                                    | The system should be able to include only the essential and relevant parts of the interview in the shortened report. | Should-have    |
| 6                                    | The system should be able to reproduce the words of all people as much as possible.                                  | Should-have    |
| 7                                    | The system should be able to work with all kinds of dialects.                                                        | Should-have    |
| 8                                    | The system should be able to work with all types of interrogations.                                                  | Should-have    |
| 9                                    | The system should be able to include all types of supportive media                                                   | Could-have     |
| <b>Performance requirement</b>       |                                                                                                                      |                |
| 10                                   | The system should be able to create a PV within 24 hours.                                                            | Must-have      |
| 11                                   | The system should be sustainable and reliable.                                                                       | Must-have      |
| 12                                   | The system should achieve high accuracy in speech-to-text conversion.                                                | Must-have      |
| 13                                   | The system should be capable of analyzing large volumes of data.                                                     | Should-have    |
| <b>Specific quality requirements</b> |                                                                                                                      |                |
| 14                                   | The system should be able to trace back to the multimedia files.                                                     | Must-have      |
| 15                                   | The system should be accurate in that the generated output is contextually correct and meet user expectations.       | Must-have      |
| 16                                   | The system should be able to make it clear how outputs are generated.                                                | Must-have      |
| 17                                   | The system should be able to allow the user to have control over the output of the system.                           | Must-have      |
| 18                                   | The system should be able to reproduce the output of the AI system.                                                  | Should-have    |
| 19                                   | The system should be compatible with various hardware and software.                                                  | Should-have    |
| 20                                   | The system should be able to verify the generated results.                                                           | Could-have     |
| 21                                   | The system should require a minimal amount of actions from the user.                                                 | Could-have     |
| <b>Constraint</b>                    |                                                                                                                      |                |
| 22                                   | The system should be able to comply with all legal and AI regulations.                                               | Must-have      |
| 23                                   | The system should be able to provide a user-friendly interface that requires minimal training.                       | Could-have     |

*Table 40. Prioritized requirements*

### 7.6.2 Obstacle & opportunity identification

The gathered requirements helped prepare for the next interviews. The objective of these interviews was to identify potential obstacles and opportunities for the future development of the prototype. Table 41 provides background information on the participants that are interviewed during this phase.

| Ref. | Perspective | Participant        | Work                                       | Experience |
|------|-------------|--------------------|--------------------------------------------|------------|
| P4   | Supplier    | Co-founder         | Project manager and operations             | 1 Year     |
| P5   | Environment | Solution architect | Solution architect for registration        | 1,5 Years  |
| P6   | Ethics      | Ethics expert      | Ethical advisor                            | 3 Years    |
| P7   | Law         | Law expert         | Jurisdictional advisor AI and data science | 7 Years    |

**Table 41.** Participants obstacle & opportunity identification phase

During the interviews, there was optimism about the potentials of the technologies in development, particularly with the development of speech-to-text software by the police and Verticai.

*P4: “Technology wise, it’s absolutely doable.”*

*P5: “Technically, almost anything is possible.”*

Verticai’s addition of implementation AI capabilities extends this functionality beyond the transcription, also showing promising performance within this study, as described in section 7.4. Beside these recent advancements, the AVR also enabled police officers to record audio on their work phones. This also shows certain improvements in software that could be used as an opportunity.

One of the mentioned obstacles is the completeness of the summarized reports. A pilot study within the Dutch law enforcement concluded that a summarized police report on its own is not good enough in the context of legal proceedings, as it often lacks the necessary detail and can be too interpretive.

*P5: “A summarized police report is not enough.”*

*P5: “If you were to make a summary of a police report, that’s entirely interpretive. Interpretation is exactly what they are trying to avoid, what they want to eliminate.”*

*P5: “There will be a summarized police report, yes, but that’s like the summary chapter you often see in a report, where the important points are highlighted briefly. However, we still think we should be able to refer back to the original transcript for detailed information.”*

Verticai aims to overcome this obstacle by ensuring that generated data refers back to the original transcript and audio, thereby enhancing traceability. This functionality ensures that the summarized information is linked to its original source, potentially addressing the issue of completeness.

*P4: “One thing we are working on is that with the transcript, it should refer back to the audio.”*

*P4: “One of the things we are working on is that when you have a certain piece of data in the field, you can select a specific part of that output. When you click on it, it then refers to the corresponding section in the transcript.”*

Also, the police is working on additional functionalities in the speech-to-text software to improve its performance.

*P5: “We eventually are going to implement additional functionality to ensure errors are corrected, or that speaker recognition can be applied.”*

Despite these technological advancements, legal and ethical concerns also present several obstacles.

*P5: “Even if we have fully functional speech-to-text software, we still need to convince the legal system that we can use this.”*

One of the obstacles is that the use of AI for interpretation might not capture the subjective aspects of statements that require specific domain knowledge or vision.

*P6: “There’s a significant risk when a kind of judgment or almost a vision is required. This could lead to qualitatively poorer output”*

*P7: “While objective data such as time and place are straightforward for AI, the subjective aspects, which are nuanced and sometimes require domain knowledge, pose a significant challenge.”*

Additionally, AI regulations state that the use of speech-to-text or AI technology must not negatively impact certain population groups. These technologies can pose challenges related to intellectual property rights, privacy violations, and potential biases.

*P6: “For a large language model, certain materials are used that might be subject to intellectual property rights, or there might be privacy violations, so you have to consider that larger picture.”*

*P7: “The AI regulation sets many strict requirements.”*

*P7: “You must consider that these are biometric data, voice recordings, and they have a higher protection category in privacy legislation.”*

The integration of these technologies into legal workflows must be carefully managed to ensure accountability. The need for human judgement is important, for example to define subjective patterns.

*“P7: The AI regulation requires continuous technical logging of how the model operates, and continuous human monitoring.”*

Moreover, there are practical obstacles related to the increasing demand for automatic transcription as the volume of audio and audiovisual recordings grows. Scalability is also a concern to maintain operational effectiveness.

*P5: “What you will see is that the barrier to recording becomes lower and lower. As more interrogations are conducted, and more audio and audiovisual materials are collected, I think the demand for automatic transcription will only grow bigger and bigger.”*

While the speed of AI models in generating outputs from transcriptions shows promise, there remains the issue of ensuring these outputs are accurate and reliable.

*P4: “For a transcript of half an hour or even an hour, the faster AI models can generate everything within ten seconds.”*

To summarize, while the opportunities presented by the speech-to-text and AI technologies are promising, obstacles remain there. Particularly in the form of legal, ethical, and AI regulations. Also, addressing the problem of potential biases and capturing subjective aspects with generated reports remains an obstacle. Transparency within the process of generating police reports can help with the trustworthiness of the system. Balancing these obstacles with the promising opportunities of these technologies will be crucial for their successful integration into the Dutch police and judicial processes. The gathered obstacles and opportunities can be found in table 42.

| ID | Obstacles                                                                                                                         |
|----|-----------------------------------------------------------------------------------------------------------------------------------|
| 1  | Convincing the legal system to use these technologies.                                                                            |
| 2  | Capturing subjective aspects that require specific domain knowledge or vision.                                                    |
| 3  | Adhering to AI regulations related to intellectual property rights, privacy violation, and potential biases.                      |
| 4  | Adhering to AI regulations related to technical logging and continuous monitoring.                                                |
| 5  | Ensuring higher protection for biometric data (voice recording), as this has a higher protection category in privacy legislation. |
| 6  | Addressing the increasing demand for automatic transcriptions due to more audio material that is being collected.                 |
| 7  | Mitigating the negative effects on people's performance when assisted by software (deskilling, laziness).                         |
| 8  | Handling large Data volumes.                                                                                                      |
| 9  | Setting up the correct infrastructure with the necessary hardware.                                                                |
| ID | Opportunities                                                                                                                     |
| 1  | Utilizing the improved software and hardware with the AVR system.                                                                 |
| 2  | Leveraging the speed of AI and speech-to-text technologies.                                                                       |
| 3  | Utilizing speech-to-text to improve overall report quality and avoid bias.                                                        |
| 4  | Guiding users towards responsible use through a user-friendly interface.                                                          |
| 5  | Developing additional functionalities such as traceability, error correction, and speaker recognition.                            |
| 6  | Making interrogation data more searchable through speech-to-text technology and its textual representation.                       |
| 7  | Applying different models to perform better in various contexts.                                                                  |
| 8  | Possibility to apply sentiment analysis to audio.                                                                                 |

**Table 42.** Obstacles & opportunities

### 7.6.3 Solution discovery

This phase uses the gathered requirements and obstacles and opportunities, to focus on exploring potential solutions that align closely with stakeholders needs.

| Ref. | Perspective    | Participant                    | Work                                                                                  | Experience |
|------|----------------|--------------------------------|---------------------------------------------------------------------------------------|------------|
| P8   | Technical      | Data scientist                 | R&D team AI lab                                                                       | 4 Years    |
| P9   | Safety         | Safety expert                  | Working on innovations with GenAI and large language models                           | 1 Year     |
| P10  | Responsible AI | Responsible AI and X4AI expert | Making models transparent in a broad sense, also working on the surrounding processes | 6 Years    |

*Table 43. Participants potential solution discovery phase*

At the current moment, the use of LLMs are not permitted for operational work. However, there is a trend towards gradually opening the door for their use.

*P8: “In the current state LLMs cannot be used for your work at this time. We are exploring how we can slowly open it up a bit.”*

LLMs are typically trained on specific datasets, often consisting of English or internet data, but within the police force, these models must adapt to domain-specific data unique to law enforcement contexts.

*P10: “A critical question is whether these models can accurately recognize and operate within their intended context.”*

*P10: “Although LLMs might appear to provide logical responses, their outputs can sometimes be illogical due to the lack of human-like reasoning capabilities.”*

The police has increasing GPU power available internally, allowing for the training and deployment of robust models within their secure systems. There is also the potential collaboration with platforms like Microsoft, which offers various LLM environments and models, although this raises questions about operational data security and cloud usage.

When implementing AI systems. It is important to minimize the amount of AI systems that are chained together. When these systems are linked, the output of one becomes the input for another, causing the errors to amplify through the chain. Therefore, it is crucial to do research to determine whether AI is even necessary in certain processes, as the potential for multiplying errors could outweigh the benefits of the technology.

*P9: “The longer the chain with AI systems that are imperfect, the more the problems multiply.”*

An important requirement for implementing LLMs is that the output can be validated. This can be achieved by incorporating traceability functionalities. A knowledge graph can be used to make the traceability of sources transparent, allowing users to see exactly which folder, documents and chunks the answer came from. To do this effectively it is important that all data layers (audio, transcription, summary) are together.

*P9: “So what we wanted to do, among other things, is use that knowledge graphs to make the traceability of its sources visible.”*

In addition, the graph can allow pre-selection, ensuring that only relevant subsets (sources) are considered, thereby greatly reducing the chance of selecting the wrong document and producing incorrect results.

Besides traceability, is the accuracy of LLMs is also an important requirement. Currently, police reports serve as communicative tools where factual correctness takes priority over presentation. Any deviation in the output from near perfection can lead to a loss of trust. Benchmark tests can be used to examine this. Which will result in a quantified risk assessment.

*P10: “Any deviation from near perfection can lead to a loss of trust among users.”*

*P9: “How often does it make mistakes like this? There are benchmark tests where they can look into exactly these kinds of situations. If it goes wrong in so many instances, then you already have a highly quantified risk assessment.”*

The idea of having a human in the loop is also pointed out to ensure that the responsibility remains with the user. It is also acknowledged that imperfections within the output of AI can be manageable.

*P9: “There should simply be a person who reads through the generated text and personally approves it.”*

*P9: “It doesn’t need to be perfect, because that is not the case in practice either. Even if it were just the same standard as the human standard, but in half the time, then that would already be higher quality.”*

Besides that, it is important that the users have sufficient domain knowledge to validate the outputs. This minimizes the risk of errors going unnoticed.

*P8: “We prefer that the user also has enough domain knowledge to be able to see if something is not correct.”*

The issue of reproducibility is also addressed, emphasizing the need for consistent and repeatable results. It is important to examine if deviation from human generation is problematic. In contrast, humans also do not have complete reproducibility when writing reports.

*P8: “Is it a problem if it doesn’t generate the same?”*

The AI system should assist the users rather than replace them, focusing on improving processes with user input instead of only concentrating on technology.

*P9: “Make sure your system is an aid for a professional doing the work.”*

*P9: “AI’s role is to guide through the thinking process, not replace it, keeping humans involved is necessary due to AI’s current limitations in understanding and subjectivity.”*

*P10: “Hybrid systems, integrating AI-generated intermediate products, present a cost-effective approach but require thoughtful traceability and governance to ensure social control and accountability.”*

However, explicit control mechanisms are crucial to avoid overlooking details. Otherwise, the time spent reviewing AI outputs could outweigh the benefits. The use of AI technology also raises concerns about whether officers will retain case knowledge if they skip the time-consuming documentation tasks.

*P10: “Will officers will retain case knowledge if they skip the time-consuming documentation tasks?”*

The aim is to avoid full dependence on technology for report writing, due to its current limitations. For this, it is important to keep human oversight.

*P9: “You don’t want the entire process of drafting the report to depend on this piece of technology.”*

The limitation of old laws and conservative attitude within the legal system. hinders the usability of AI in generating police reports.

*P8: “You don’t gain anything from a police report if it isn’t usable in the further legal chain. And well, that further chain is somewhat hindered by old legislation and a somewhat conservative attitude ”*

Researching and experimenting with current technologies helps to build a case for legal changes. There is a positive movement within the legal system, particularly with the inclusion of multimedia files.

*P8: “See how far you can go with experimenting, and then eventually you have a case to argue for changing certain laws. ”*

*P8: “They want various things for confirmation or assurance.”*

Overall, it is crucial to identify the specific processes that really require AI. This will ensure that the risks associated with potential imperfect AI implementations are minimized. However, for those processes where AI is indeed required, the main challenge lies in the legal framework. Convincing the legal system to accept AI technologies requires extensive research, particularly on traceability, maintaining human oversight, and assessing whether AI systems can operate within acceptable margins of error compared to human performance in generating reports. Therefore, a subset of the requirements presented in section 7.6.1 that form a potential solution. These are features that have to be implemented to form a working product that at least satisfies the bare minimum of requirements. These requirements are in line with the prioritization method performed. The analysis is shown in table 44 and links the requirement ID (RE-ID) to its related obstacle ID (OB-ID) and opportunity ID (OP-ID) that are defined in section 7.6.2. Following that, it shows if the requirement is included in the potential solution or not. In addition, a description will be given why or why not this requirement is included in the potential solution.

| RE-ID | OB-ID   | OP-ID | Included | Description                                                                                                                    |
|-------|---------|-------|----------|--------------------------------------------------------------------------------------------------------------------------------|
| 1     |         | 5     | Yes      | Must have for providing transparency, traceability and therefore the credibility of the reports.                               |
| 2     |         | 2,5   | Yes      | Must have for accurate reporting and for creating the question/answer format.                                                  |
| 3     |         | 5     | Yes      | Must have for clarity of the generated statement.                                                                              |
| 4     | 3       |       | Yes      | Must have for trust and transparency of the system.                                                                            |
| 5     | 2       | 7     | No       | While beneficial, not critical. Generated reports can still be edited by the police officer, but this would be less efficient. |
| 6     |         |       | No       | Important for maintaining the authenticity of the record, but some paraphrasing might still be enough.                         |
| 7     |         |       | No       | While beneficial, not critical. System can first focus on performing with no present dialect in the audio.                     |
| 8     |         |       | No       | While beneficial, not critical. System can focus on one type of interrogation in the beginning.                                |
| 9     |         |       | No       | While beneficial, not critical. Could be added by the user if needed.                                                          |
| 10    |         | 2     | Yes      | Must have to ensure time constraints that are often set for creating reports.                                                  |
| 11    |         |       | Yes      | Must have to ensure general trust in the system.                                                                               |
| 12    |         | 3, 8  | Yes      | Must have to ensure trustworthiness, and quality input for the AI models.                                                      |
| 13    | 6,8     |       | No       | Important if data volume would eventually increase, but not necessary for using the product in practice.                       |
| 14    |         | 5     | Yes      | Must have for providing transparency, traceability and therefore the credibility of the reports.                               |
| 15    | 2       | 7     | Yes      | Must have to accurately reflect on the context of the conversations.                                                           |
| 16    |         |       | Yes      | Must have for trust and transparency of the system.                                                                            |
| 17    | 3,7     | 4     | Yes      | Must have to give the police officer control over the type of report that is being created.                                    |
| 18    |         |       | No       | Would help with consistency, but not critical for the required functionalities of the system.                                  |
| 19    |         |       | No       | Would increase trust, but not necessary.                                                                                       |
| 20    | 9       | 6     | No       | Would increase overall efficiency, but not necessary.                                                                          |
| 21    |         |       | No       | Depends on the quality of the generated reports.                                                                               |
| 22    | 1,3,4,5 |       | Yes      | Compliance with regulations is a must to avoid legal consequences and to ensure ethical, AI standards.                         |
| 23    |         | 4     | No       | Improves usability, but users can still learn how to use the system with training if necessary.                                |

**Table 44.** Potential solution requirements

## 7.7 Revision & refinement

This section describes the process of continuous development of the prototype during the performance of the case study. This includes all the changes made after the evaluation of the first prototype iteration, described in section 7.4. In the Vertical system, we have built two models, each containing the same types of concepts. These concepts are broken down into three key components:

- **Name**
- **Description**
- **Instruction**

The changes made during the review and refinement phase focused on improving both models by analyzing the components within each concept used to formulate the prompts. The performance of Model 1 remained unchanged due to its near perfection. Despite the lack of improvements in model 1, important improvements in the performance of model 2 can be identified in the following concepts:

- **Date of birth.** Looking at table 45, here both the old and refined instruction are shown with their improved results. The original prompt simply asked for the suspect's date of birth. Updating the prompt to specify the extraction of this information directly from the transcript showed a big improvement. All previous mistakes that were made in the 5 transcripts were correct with this new prompt.

| Old instruction                             | Revised instruction                                             |
|---------------------------------------------|-----------------------------------------------------------------|
| “Provide the date of birth of the suspect.” | “Extract the date of birth of the suspect from the transcript.” |

*Table 45. Revised instruction of: Date of birth*

- **Date criminal offense.** Similarly, a change in instruction helped in improving the output of this concept. As shown in table 46 refining the instruction to include specific guidance on extracting dates mentioned in the transcript improved the model's performance. Also, adding the sentence “If no date is mentioned in the transcript, leave the answer blank.” helped with generating the correct output.

| Old instruction                                             | Revised instruction                                                                                                                                        |
|-------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------|
| “Provide the date of when the criminal offense took place.” | “What is the date when the criminal offense took place as mentioned in the transcript? If no date is mentioned in the transcript, leave the answer blank.” |

*Table 46.* Revised instruction of: Date criminal offense

- **Location criminal offense.** The output for this concept is also improved by using a more elaborate instruction. Table 47 shows that specifying the need for exact location details from the transcript resulted in improved outputs.

| Old instruction                              | Revised instruction                                                                                              |
|----------------------------------------------|------------------------------------------------------------------------------------------------------------------|
| “Where did the criminal offense take place?” | “Where did the criminal offense take place? If this is not mentioned in the transcript, leave the answer blank.” |

*Table 47.* Revised instruction of: Location criminal offense

- **Confession.** Also, the mistake made in TR1 could be improved. Table 48 shows that revising the instruction to extract a clear yes or no response from the transcript corrected the errors observed in Model 2.

| Old instruction                          | Revised instruction                                                                             |
|------------------------------------------|-------------------------------------------------------------------------------------------------|
| “Does the suspect admit to the offense?” | “Extract from the transcript whether the suspect admits to the offense. Answer with yes or no.” |

*Table 48.* Revised instruction of: Confession

Overall, these improvements highlight the critical role of prompts in improving model performance in a variety of data extraction tasks. By closely aligning prompts with the structure and content of the input data (transcripts), Model 2 significantly improves its ability to accurately extract targeted information, thereby improving overall performance and reliability. Furthermore, these refinements have enabled Model 2 to outperform Model 1. Table 49 illustrates the evaluation of the refined Model 2, showing almost a perfect score. However, the inability to generate a time frame is still a challenge.

| Model_2                   | TR1 | TR2 | TR3 | TR4 | TR5 | Precision | Recall |
|---------------------------|-----|-----|-----|-----|-----|-----------|--------|
| Region                    | 1   | 1   | 1   | 1   | 1   | 1.0       | 1.0    |
| District                  | 1   | 1   | 1   | 1   | 1   | 1.0       | 1.0    |
| Department                | 1   | 1   | 1   | 1   | 1   | 1.0       | 1.0    |
| Name reporting officer    | 1   | 1   | 1   | 1   | 1   | 1.0       | 1.0    |
| Rank                      | 1   | 1   | 1   | 1   | 1   | 1.0       | 1.0    |
| Personnel number          | 1   | 1   | 1   | 1   | 1   | 1.0       | 1.0    |
| Name suspect              | 1   | 1   | 1   | 1   | 1   | 1.0       | 1.0    |
| Place of birth            | 1   | 1   | 1   | 1   | 1   | 1.0       | 1.0    |
| Date of birth             | 1   | 1   | 1   | 1   | 1   | 1.0       | 1.0    |
| Address of suspect        | 1   | 1   | 1   | 1   | 1   | 1.0       | 1.0    |
| Date criminal offense     | 1   | 1   | 1   | 1   | 0   | 0.8       | 0.8    |
| Location criminal offense | 1   | 1   | 1   | 1   | 1   | 1.0       | 1.0    |
| Type criminal offense     | 1   | 1   | 1   | 1   | 1   | 1.0       | 1.0    |
| Confession                | 1   | 1   | 1   | 1   | 1   | 1.0       | 1.0    |
| Making a deal             | 1   | 1   | 1   | 1   | 1   | 1.0       | 1.0    |
| Amount of money           | 1   | 1   | 1   | 1   | 1   | 1.0       | 1.0    |
| <b>Correct outputs</b>    | 16  | 16  | 16  | 16  | 15  |           |        |

*Table 49.* Evaluation of refined model 2

## 7.8 Reflection

This section discusses the performed study of developing a prototype for generating automatically written interrogation reports for the Dutch police. This reflection will be divided into the following subsections:

- **Results interpretation.** This section will go over the results of the study. This includes, the key challenges within the domain of AI engineering, the current technological capabilities of the developed prototype in terms of accuracy and efficiency, and the performed approach for the requirement engineering activities and design.
- **Limitations.** This section addresses the limitations of the study performed. This includes several technical, but also practical limitations.
- **Future research.** Building on the findings and limitations mentioned in the previous sections, this section will provide advice on potential directions for future research, focusing on enhancing the prototype and exploring practical challenges associated with integrating AI in this domain.

### 7.8.1 Results interpretation

#### *Challenges AI engineering*

AI engineering is a growing discipline focused on the software development of AI based systems. However, AI engineering is facing several challenges. This is due to the complexity of AI technology. AI-based software behaves differently from regular software, introducing uncertainty due to limited insight into the system's inner workings. Therefore, the traditional software development life cycle (SDLC) needs adaptations to ensure that AI systems can be developed, implemented and operated responsibly. Key challenges and adaptions for each phase include:

- **RE4AI.** Managing requirements for AI based systems involves challenges such as the overestimating of AI capabilities, capturing and specifying trade-off that might arise and difficulties in understanding NF requirements. Therefore, coordination among various stakeholders is essential. Effective planning is crucial to deal with overlapping tasks and conflicting opinions.
- **Design.** AI systems require a design phase that incorporates the eight key AI principles. A strategy for addressing these principles is to adopt explainable AI in the design phase. Which highlights the principle of transparency and explainability.
- **Implementation.** Challenges within the implementation phase include: handling large amounts of data, complexity in integrating AI technology in existing systems, complying with regulatory requirements, and training users to work effectively with AI. In the Dutch society, concerns are raised, because of the implementation of algorithms in the public sector. Therefore, the Dutch government created The Algorithm Register (AR), a publicly available database explaining algorithms that are being planned, created, implemented, or retracted by the government.
- **Testing.** Testing AI systems comes with its own set of challenges, such as the oracle problem, where establishing a single ground truth is challenging due to unpredictable outputs. Therefore, AI systems require extensive testing due to their unpredictable behavior. Testing should focus on reliability, robustness, security, and efficiency. Techniques such as prompt engineering are essential for evaluating AI behavior.
- **Operation.** This phase involves continuous management and oversight of AI systems to handle issues like data smells and concept drift.

Monitoring, ethical risk assessments, and maintaining transparency are important activities during this phase.

### ***Effectiveness of the prototype***

In evaluating the prototype's effectiveness, the study conducted a comparative evaluation of two models used for the prototype, testing them against several requirements for making a PV that are listed in Appendix B. The results of the evaluation showed the prototype's ability to generate accurately formatted statements and metadata. Certain limitations, such as the use of simulated data and lack of access to the internal code of Verticai, impacted the overall assessment. The findings show the potential of the prototype while also highlighting the need for further testing and refinement to enhance its practical applicability in real-world scenarios.

### ***Performed approach for RE activities and design***

The created approach works well to form high-level requirements from different perspectives within the police and from the supplier Verticai. The requirement engineering activities with continuous development of the prototype resulted in a potential solution that supports the development and implementation of the prototype. Participants highlighted the need for a system that enhances the efficiency of summarizing and structuring interview data, pointing out the current time intensive process of manually transcribing and summarizing recorded interrogations. A key requirement is the functionality of traceability in the AI-generated reports, ensuring that every piece of information can be verified against the original recordings. This would also help build trust in AI-generated content. Furthermore, the performed approach showed several obstacles and opportunities that need to be addressed, with as main obstacle convincing the legal system to use these technologies. Overall, the results provide valuable directions for future works and refining the prototype to meet the necessary practical requirements.

#### **7.8.2 Limitations**

Despite the promising results of this study, several limitations need to be addressed. Firstly, during this study, we had **no access to internal code of Verticai**. Therefore, the prototype consisted of three separate processes that operated independently rather than within a single user interface (UI). This resulted in less efficient communication between components, impacting the overall performance and presentation of the product. Additionally, this forced us to use the ChatGPT API. Therefore, we were not able to

test different models, limiting our ability to explore alternative approaches and compare their effectiveness. Secondly, **Data quality and availability**. The data used in this study did not come from the police force itself. Instead, simulated data was used. This implies that the prototype has not been tested on real-world data, which could affect the practical applicability of the findings. Real-world testing is necessary for evaluating the system's performance and ensuring its readiness for deployment in actual police processes. Thirdly, **legal and ethical constraints**. This study primarily focused on the technological feasibility of the prototype. Besides looking at one specific law in the Manual evaluation, this study ignores several other legal and ethical constraints. Understanding and compliance with all laws and regulations is essential for successful implementation of the to-be system. Finally, **the scope of the study**. The design science of this study was limited in the scope. This involved feedback from only a select group of experts. A more extensive investigation involving more iterations of the design activities is necessary to achieve higher level requirements and prototype design.

### 7.8.3 Future research

The results and limitations in this study present several directions for future research in the field of automated police reporting using generative AI. The biggest obstacle in implementing AI into the Dutch law enforcement is convincing the legal system to use this technology. To address this challenge, further research is needed. Future research should investigate when a generated police report is legally valid, and therefore usable in practice. The following key research areas are defined. Firstly, establishing clear legal standards that AI-generated reports must meet to be used in the legal system, also ensuring compliance with AI and ethical regulations. Secondly, traceability functionalities need to be robust to ensure transparency of the system and that the output can be easily verified by the user. Thirdly, future research should focus on testing the prototype with real-life data to assess its performance in more realistic scenarios, such as using data from various types of incidents, ranging from minor offenses to major criminal cases. Furthermore, integrating a human in the loop design process is essential to balance AI automation with human oversight, ensuring accuracy and reliability. Finally, exploring improved functionalities, such as advanced speaker recognition involving multiple individuals, and good performance with different languages or dialects will enhance the system's overall capabilities.

## 8 Conclusion

This study focused on the generation of automatically written reports in the Dutch law enforcement. First, this study dived into the literature exploring the influence of AI engineering on the traditional software development life cycle, identifying challenges introduced by building AI based systems. As AI engineering is a complex process, numerous challenges were identified. Secondly, the performed a case stud resulted in two artifacts:

1. A designed prototype/proof of concept for generating automatically written police reports. Which is evaluated on its strengths, weaknesses, and limitations. This study successfully created a prototype to generate automatically written reports. The evaluation of the generated reports revealed several strengths of the created prototype with the Verticai system, such as high precision and recall and generation of a question-answer format. However, there were some weaknesses, such as errors in generating timeframes, omissions, and the inclusion of redundant information. The lack of access to Verticai's internal code and the use of non-police data were limitations, affecting the study's ability to fully validate the prototype's performance.
2. An approach called prototype-driven requirement engineering, which is uses the designed prototype in an iterative process to identify related requirements, opportunities, obstacles, and to define a potential solution within a multidisciplinary perspective. The performed approach for the requirement engineering activities with continuous development of the prototype resulted in a potential solution that closely aligns with the stakeholder needs. The finding suggests that traceability and meeting the legal, ethical, and AI regulations are crucial to ensure the completeness and reliability of the reports. The results also support the idea that AI should assist rather than replace human judgment in the report writing process.

This study contributes to the understanding of the development of AI-driven solutions. It highlights the challenges involved in the concept of AI engineering, eventually focussing on the context of generating police reports for the Dutch law enforcement. In addition, this study demonstrates the importance of a robust requirement engineering process, with the developed iterative approach of prototype-driven requirement engineering. Out of this performed approach, the need for transparency, explainability, and compliance with legal and ethical standards are highlighted. The developed prototype shows promising potential, but future research is essential to advance

the practical deployment of AI in the Dutch law enforcement, ensuring that such technologies enhance, rather than undermine, the quality and efficiency of police work. The study identified several key areas for future research, including the need to explore the legal feasibility of using AI to write police reports, the importance of testing the prototype with real-world data, and the necessity of addressing ethical concerns such as privacy and bias. Engaging a broader range of stakeholders in the development process and enhancing the system's functionalities is therefore crucial.

To conclude, while many challenges remain in legal, system, and organizational contexts, this study provided a solid foundation for developing an effective solution for generating automatically written police reports and points the integration of developed AI systems in the Dutch law enforcement into the right direction.

## A Software engineering product Models

| Models          | Advantages                                                                                                                                                                                        | Disadvantages                                                                                                                                                                                                              |
|-----------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Waterfall Model | <ul style="list-style-type: none"> <li>• Clear documentation</li> <li>• Clear deliverables</li> <li>• Structured approach</li> <li>• Predictable</li> <li>• Easy management</li> </ul>            | <ul style="list-style-type: none"> <li>• Limited flexibility</li> <li>• Lack of customer involvement</li> <li>• High risk</li> <li>• Lengthy development cycle</li> <li>• Changes are costly</li> </ul>                    |
| Agile Model     | <ul style="list-style-type: none"> <li>• Adaptability</li> <li>• Continuous quality improvement</li> <li>• Early bug detection</li> <li>• Faster to market</li> <li>• Customer-centric</li> </ul> | <ul style="list-style-type: none"> <li>• Resource intensive</li> <li>• Lack of documentation</li> <li>• Dependency on customer availability</li> <li>• Potential loss of scope</li> <li>• Overemphasis on speed</li> </ul> |

*Table 50. Waterfall and Agile comparison<sup>77</sup>*

## B Police report Requirements

| Category                          | Requirement                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                | Tested                                                                                   |
|-----------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------|
| The writing of a police report    | To write a police report as soon as possible. (art. 152 WvSv) <sup>70</sup>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                | Yes                                                                                      |
| The layout of a police report     | Clear table of contents.<br>Summarization police report.<br>Custodial coercive measures used.<br>Overview of evidence.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     | No<br>Yes<br>No<br>N.A.                                                                  |
| The content of a police report    | Description of reporting officer.<br>Date on police report. (art. 153 WvSv) <sup>71</sup><br>Signature on police report. (art. 153 WvSv) <sup>71</sup><br>Oath of office. (art. 153 WvSv) <sup>71</sup><br>Source of knowledge reporting officer. (art. 153 WvSv) <sup>71</sup><br>Source of knowledge witness/suspect. (art. 341 & 342 WvSv) <sup>73 74</sup><br>Should be written as much as possible in the suspect's own words. (art. 29a-2 WvSv) <sup>72</sup><br>The statement of the suspect is presented as completely as possible and, where possible, in a question-and-answer format. (art. 29a-2 WvSv) <sup>72</sup><br>All components of the criminal offense.<br>Description of the performed official acts.<br>Hallmark in case of amendments made to already formatted police reports.<br>Integrity of the investigation.<br>Consideration whether to employ a translator. | Yes<br>Yes<br>N.A.<br>N.A.<br>No<br>No<br>Yes<br>Yes<br>N.A.<br>No<br>No<br>N.A.<br>N.A. |
| The submission of a police report | The report must be delivered to the relevant person or authority without unnecessary delay.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                | N.A.                                                                                     |

*Table 51. Requirements for police report (based on Jansen<sup>46</sup>)*

## C Interview protocol

The following procedure was followed for all interviews:

1. **Contact by mail**, contact with a potential participant was made by Martijn van Vliet MSc with a short introduction to the study and a request to participate.
2. **Set meeting location and time**, after the individual confirmed his interest for an interview, an appointment was made. This could be online or on a physical location.
3. **Interview**, the interviews consisted of three phases. Phase 1 involved introducing ourselves and the study. During this phase, we also asked participants for their permission to record the audio of the interview and to use the information collected for our study. In phase 2, we focused on asking questions specifically related to the expertise of the participant in the field of developing an AI system to automatically generate police reports. The aim of this phase was to gain detailed insights into their expertise, experience, and the challenges they might see coming. The third and final phase consisted of a walkthrough with the prototype, during which we asked more questions related to the prototype itself. This phase was designed to understand their assessment of the functionality, usability and potential improvements of the prototype. Throughout the interview, a semi-structured format was maintained to allow flexibility in exploring relevant topics while ensuring that all key areas were discussed.
4. **Thank you email and consent form**, after the interview, participants were mailed by Wouter Westerkamp to thank them for the participation and with a question to fill in the consent form.
5. **Analyze the data** As explained in section 7.5, we utilize the output of each phase to structure the interviews for the next phase and improve the prototype.

## D Manual evaluation of generated statements



### Interrogation transcript TR1 - Model 1

U bent hier vandaag omdat u verdacht wordt van betrokkenheid bij een inbraak die plaatsvond op 15 april 2024 aan de Havenstraat in Rotterdam. Kent u deze locatie? Ja, die straat ken ik wel. Waar was u op de avond van 15 april 2024 rond 10 uur avonds? Ik was gewoon thuis, alleen. Ik heb de film gekeken die avond. Er zijn camerabeelden waarop iemand die op u lijkt in de buurt van de havenstraat te zien is, rondom de tijd van de inbraak. Kunt u uitleggen hoe dat dan komt? Poeh, dat moet iemand anders zijn geweest. Ik ben die avond niet uit huis geweest eigenlijk. We hebben ook vingerafdrukken van u gevonden op de plaatselect. Die komen overeen met die van u. Hoe verlaat u dat? Uhm, ik weet niet hoe dat kan. Ik heb niets met die inbraak te maken. Gezien de bewijzen die wij hebben, zou een schrikking een optie zijn om deze zaak te beslechten. Wij stellen een schikingsvoorstel voor van 5.000 euro. Bent u bereid om te schikken? Nee, ik wil niet schikken. Gewoon omdat ik onschuldig ben. Denkt u alsjeblieft goed na over uw situatie. We gaan zo verder met het gehoor. We gaan verder. Heeft u nog iets toe te voegen aan uw verklaring of wilt u iets wijzigen? Nee, ik blijf bij mijn verklaring.



### Generated statement of TR1 - Model 1

Verbalisant: Kunt u uw volledige naam en geboortedatum bevestigen voor de opname?  
 Verdachte: Ja dat kan ik. Mijn naam is Lucas Pieter Johannes Janssen en ik ben geboren op 13 maart 1985 in Den Haag.  
 Verbalisant: U woont op de Molestraat 45, 3011 XD Rotterdam, correct?  
 Verdachte: Ja, dat is correct.  
 Verbalisant: U bent hier vandaag omdat u verdacht wordt van betrokkenheid bij een inbraak die plaatsvond op 15 april 2024 aan de Havenstraat in Rotterdam. Kent u deze locatie?  
 Verdachte: Ja, die straat ken ik wel.  
 Verbalisant: Waar was u op de avond van 15 april 2024 rond 10 uur 's avonds?  
 Verdachte: Ik was gewoon thuis, alleen. Ik heb de film gekeken die avond.  
 Verbalisant: Er zijn camerabeelden waarop iemand die op u lijkt in de buurt van de havenstraat te zien is, rondom de tijd van de inbraak. Kunt u uitleggen hoe dat dan komt?  
 Verdachte: Poeh, dat moet iemand anders zijn geweest. Ik ben die avond niet uit huis geweest eigenlijk.  
 Verbalisant: We hebben ook vingerafdrukken van u gevonden op de plaatselect. Die komen overeen met die van u. Hoe verlaat u dat?  
 Verdachte: Uhm, ik weet niet hoe dat kan. Ik heb niets met die inbraak te maken.  
 Verbalisant: Gezien de bewijzen die wij hebben, zou een schrikking een optie zijn om deze zaak te beslechten. Wij stellen een schikingsvoorstel voor van 5.000 euro. Bent u bereid om te schikken?  
 Verdachte: Nee, ik wil niet schikken. Gewoon omdat ik onschuldig ben.  
 Verbalisant: Heeft u nog iets toe te voegen aan uw verklaring of wilt u iets wijzigen?  
 Verdachte: Nee, ik blijf bij mijn verklaring.

| Equivalent      | Addition     | Error          |
|-----------------|--------------|----------------|
| Identical       | Relevant     | Hallucination  |
| Paraphrase      | Redundant    | Categorization |
|                 |              | Repetition     |
|                 |              | Incorrect      |
|                 |              | Omission       |
| Transcriptional | Articulatory |                |
| Improvement     | Improvement  |                |
| Mistake         | Mistake      |                |

Figure 30. Model 1: Manual evaluation of generated statement (TR1)



## Interrogation transcript TR1 - Model 2

U bent hier vandaag omdat u verdacht wordt van betrokkenheid bij een inbraak die plaatsvond op 15 april 2024 aan de Havenstraat in Rotterdam. Kent u deze locatie? Ja, die straat ken ik wel. Waar was u op de avond van 15 april 2024 rond 10 uur avonds? Ik was gewoon thuis, alleen. Ik heb de film gekeken die avond. Er zijn camerabeelden waarop iemand die op u lijkt in de buurt van de havenstraat te zien is, rondom de tijd van de inbraak. Kunt u uitleggen hoe dat dan komt? Poeh, dat moet iemand anders zijn geweest. Ik ben die avond niet uit huis geweest eigenlijk. We hebben ook vingerafdrukken van u gevonden op de plaatselect. Die komen overeen met die van u. Hoe verklaart u dat? Uhm, ik weet niet hoe dat kan. Ik heb niets met die inbraak te maken. Gezien de bewijzen die wij hebben, zou een schrikking een optie zijn om deze zaak te beslechten. Wij stellen een schikingsvoorstel voor van 5.000 euro. Bent u bereid om te schikken? Nee, ik wil niet schikken. Gewoon omdat ik onschuldig ben. Denkt u alsjeblieft goed na over uw situatie. We gaan zo verder met het gehoor. We gaan verder. Heeft u nog iets toe te voegen aan uw verklaring of wilt u iets wijzigen? Nee, ik blijf bij mijn verklaring.



## Generated statement of TR1 - Model 2

Verbalisant: Kunt u uw volledige naam en geboortedatum bevestigen voor de opname?  
 Verdachte: Ja dat kan ik. Mijn naam is Lucas Pieter Johannes Janssen en ik ben geboren op 13 maart 1985 in Den Haag.  
 Verbalisant: U woont op de Molestraat 45, 3011 XD Rotterdam, correct?  
 Verdachte: Ja, dat is correct.  
 Verbalisant: U bent hier vandaag omdat u verdacht wordt van betrokkenheid bij een inbraak die plaatsvond op 15 april 2024 aan de Havenstraat in Rotterdam. Kent u deze locatie?  
 Verdachte: Ja, die straat ken ik wel.  
 Verbalisant: Waar was u op de avond van 15 april 2024 rond 10 uur 's avonds?  
 Verdachte: Ik was gewoon thuis, alleen. Ik heb de film gekeken die avond.  
 Verbalisant: Er zijn camerabeelden waarop iemand die op u lijkt in de buurt van de havenstraat te zien is, rondom de tijd van de inbraak. Kunt u uitleggen hoe dat dan komt?  
 Verdachte: Poeh, dat moet iemand anders zijn geweest. Ik ben die avond niet uit huis geweest eigenlijk.  
 Verbalisant: We hebben ook vingerafdrukken van u gevonden op de plaatselect. Die komen overeen met die van u. Hoe verklaart u dat?  
 Verdachte: Uhm, ik weet niet hoe dat kan. Ik heb niets met die inbraak te maken.  
 Verbalisant: Gezien de bewijzen die wij hebben, zou een schrikking een optie zijn om deze zaak te beslechten. Wij stellen een schikingsvoorstel voor van 5.000 euro. Bent u bereid om te schikken?  
 Verdachte: Nee, ik wil niet schikken. Gewoon omdat ik onschuldig ben.  
 Verbalisant: Heeft u nog iets toe te voegen aan uw verklaring of wilt u iets wijzigen?  
 Verdachte: Nee, ik blijf bij mijn verklaring.

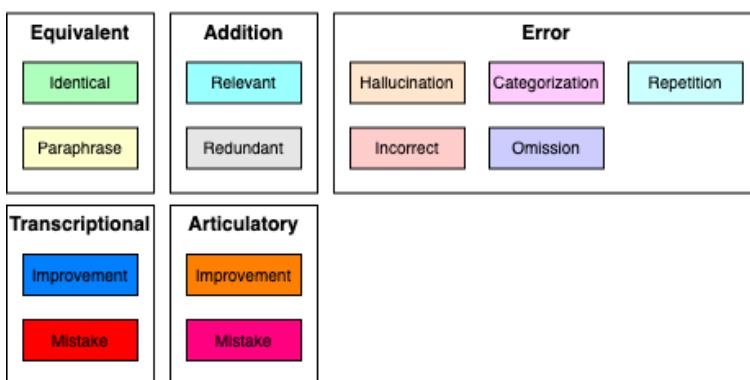


Figure 31. Model 2: Manual evaluation of generated statement (TR1)

## Interrogation transcript TR2 - Model 1

U bent hier vandaag omdat u verdacht wordt van betrokkenheid bij illegale afvaldumping in natuurgebieden. Bent u bekend met de beschuldigingen? Ja, ik heb erover gehoord, maar ik heb er niets mee te maken. Wij hebben namelijk een getuigeverklaring dat uw vrachtwagen meerdere keren is gezien bij de plekken waar het afval is gedumpt. Kunt u dat uitleggen? Nou ja, mijn vrachtwagen wordt soms uitgeleend aan collega's. Misschien hebben zij dat gedaan zonder mij medeweten. We hebben ook gps-gegevens van uw vrachtwagen die bevestigen dat deze op de locatie was tenzijde van de dumpingen. Wat heeft u daarop te zeggen? Zoals ik al zei, ik neem mijn vrachtwagen soms uit. Ik wist niet wat ermee werd gedaan. Er zijn ook camerabeelden waarop te zien is dat u zelf achter het stuur zat bij een van de afvaldumpingen. Hoe verklart u dat? Dat kan niet kloppen. Ik heb altijd legaal afval afgevoerd. Misschien is er verbraak van een vergissing. Mevrouw Vermeulen, gezien de bewijs die we hebben, zou een schikking een mogelijkheid kunnen zijn om deze zaak af te ronden. Wij stellen een schikking voorstel van 20.000 euro. Bent u bereid om te schikken? Nee, ik wil niet schikken, ik ben onschuldig. Oké, we zullen even een pauze nemen. Denkt u alsjeblieft goed na over uw situatie. We gaan zo verder met het voor. We gaan verder. Heeft u nog iets toe te voegen aan uw verklaring of wilt u iets wijzigen? Ja, ik heb nagedacht. Ik wil toch schikken. Dat is een verstandig besluit. Zoals besproken stellen we een schikking voorstel van 20.000 euro. Bent u akkoord? Ja, ik ga akkoord met de schikking. Prima, er is nog één voorwaarde voor de schikking. We hebben aanvullende informatie nodig over uw medeplichtigen. Bent u bereid om iemand aan te wijzen? Ja, ik zal eerlijk zijn. Het was niet alleen mijn verantwoordelijkheid. Mijn collega, Peter van Dam, was ook betrokken. Hij heeft me gevraagd om mee te doen.

## Generated statement of TR2 - Model 1

Verbalisant: Kunt u uw volledige naam en geboorte laten bevestigen voor de opname?  
 Verdachte: Ja, mijn naam is Anouk Johanna Vermeulen en ik ben geboren op 22 februari 1986 in Rotterdam.  
 Verbalisant: U woont op Beukenlaan 12, 2514 HC Den Haag. Klopt dat?  
 Verdachte: Ja, dat klopt.  
 Verbalisant: U bent hier vandaag omdat u verdacht wordt van betrokkenheid bij illegale afvaldumping in natuurgebieden. Bent u bekend met de beschuldigingen?  
 Verdachte: Ja, ik heb erover gehoord, maar ik heb er niets mee te maken.  
 Verbalisant: Wij hebben namelijk een getuigeverklaring dat uw vrachtwagen meerdere keren is gezien bij de plekken waar het afval is gedumpt. Kunt u dat uitleggen?  
 Verdachte: Nou ja, mijn vrachtwagen wordt soms uitgeleend aan collega's. Misschien hebben zij dat gedaan zonder mij medeweten.  
 Verbalisant: We hebben ook gps-gegevens van uw vrachtwagen die bevestigen dat deze op de locatie was ten tijde van de dumpingen. Wat heeft u daarop te zeggen?  
 Verdachte: Zoals ik al zei, ik leen mijn vrachtwagen soms uit. Ik wist niet wat ermee werd gedaan.  
 Verbalisant: Er zijn ook camerabeelden waarop te zien is dat u zelf achter het stuur zat bij een van de afvaldumpingen. Hoe verklart u dat?  
 Verdachte: Dat kan niet kloppen. Ik heb altijd legaal afval afgevoerd. Misschien is er sprake van een vergissing.  
 Verbalisant: Mevrouw Vermeulen, gezien het bewijs dat we hebben, zou een schikking een mogelijkheid kunnen zijn om deze zaak af te ronden. Wij stellen een schikking voorstel van 20.000 euro. Bent u bereid om te schikken?  
 Verdachte: Nee, ik wil niet schikken, ik ben onschuldig.  
 Verbalisant: Oké, we zullen even een pauze nemen. Denkt u alsjeblieft goed na over uw situatie. We gaan zo verder met het verhoor.  
 Verbalisant: We gaan verder. Heeft u nog iets toe te voegen aan uw verklaring of wilt u iets wijzigen?  
 Verdachte: Ja, ik heb nagedacht. Ik wil toch schikken.  
 Verbalisant: Dat is een verstandig besluit. Zoals besproken stellen we een schikking voorstel van 20.000 euro. Bent u akkoord?  
 Verdachte: Ja, ik ga akkoord met de schikking.  
 Verbalisant: Prima, er is nog één voorwaarde voor de schikking. We hebben aanvullende informatie nodig over uw medeplichtigen. Bent u bereid om iemand aan te wijzen?  
 Verdachte: Ja, ik zal eerlijk zijn. Het was niet alleen mijn verantwoordelijkheid. Mijn collega, Peter van Dam, was ook betrokken. Hij heeft me gevraagd om mee te doen.

| Equivalent | Addition  | Error          |
|------------|-----------|----------------|
| Identical  | Relevant  | Hallucination  |
| Paraphrase | Redundant | Categorization |
|            |           | Repetition     |
|            |           | Incorrect      |
|            |           | Omission       |

| Transcriptional | Articulatory |
|-----------------|--------------|
| Improvement     | Improvement  |
| Mistake         | Mistake      |

*Figure 32. Model 1: Manual evaluation of generated statement (TR2)*



### Interrogation transcript TR2 - Model 2

U bent hier vandaag omdat u verdacht wordt van betrokkenheid bij illegale afvaldumping in natuurgebieden. Bent u bekend met de beschuldigingen? Ja, ik heb erover gehoord, maar ik heb er niets mee te maken. Wij hebben namelijk een getuigeverklaring dat uw vrachtwagen meerdere keren is gezien bij de plekken waar het afval is gedumpt. Kunt u dat uitleggen? Nou ja, mijn vrachtwagen wordt soms uitgereed aan collega's. Misschien hebben zij dat gedaan zonder mij medeweten. We hebben ook gps-gegevens van uw vrachtwagen die bevestigen dat deze op de locatie was tenzijde van de dumpingen. Wat heeft u daarop te zeggen? Zoals ik al zei, ik **heem** mijn vrachtwagen soms uit. Ik wist niet wat ermee werd gedaan. Er zijn ook camerabeelden waarop te zien is dat u zelf achter het stuur zat bij een van de afvaldumpingen. Hoe verklaart u dat? Dat kan niet kloppen. Ik heb altijd legaal afval afgevoerd. Misschien is er **verbaak** van een vergissing. Mevrouw Vermeulen, gezien de bewijs die we hebben, zou een schikking een mogelijkheid kunnen zijn om deze zaak af te ronden. Wij stellen een schikking voorstel van 20.000 euro. Bent u bereid om te schikken? Nee, ik wil niet **schrikken**, ik ben onschuldig. Oké, we zullen even een pauze nemen. Denkt u alsjeblieft goed na over uw situatie. We gaan zo verder met het voor. We gaan verder. Heeft u nog iets toe te voegen aan uw verklaring of wilt u iets wijzigen? Ja, ik heb nagedacht. Ik wil toch schrikken. Dat is een verstandig besluit. Zoals besproken stellen we een **schikking voorstel** van 20.000 euro. Bent u akkoord? Ja, ik ga akkoord met de **schikking**. Prima, er is nog één voorwaarde voor de schikking. We hebben aanvullende informatie nodig over uw medeplichtigen. Bent u bereid om iemand aan te wijzen? Ja, ik zal eerlijk zijn. Het was niet alleen mijn verantwoordelijkheid. Mijn collega, Peter van Dam, was ook betrokken. Hij heeft me gevraagd om mee te doen.



### Generated statement of TR2 - Model 2

Verbalisant: Bent u bekend met de beschuldigingen?  
 Verdachte: Ja, ik heb erover gehoord, maar ik heb er niets mee te maken.  
 Verbalisant: Wij hebben namelijk een getuigeverklaring dat uw vrachtwagen meerdere keren is gezien bij de plekken waar het afval is gedumpt. Kunt u dat uitleggen?  
 Verdachte: Nou ja, mijn vrachtwagen wordt soms **uitgereed** aan collega's. Misschien hebben zij dat gedaan zonder mij medeweten.  
 Verbalisant: We hebben ook gps-gegevens van uw vrachtwagen die bevestigen dat deze op de locatie was tenzijde van de dumpingen. Wat heeft u daarop te zeggen?  
 Verdachte: Zoals ik al zei, ik **heem** mijn vrachtwagen soms uit. Ik wist niet wat ermee werd gedaan.  
 Verbalisant: Er zijn ook camerabeelden waarop te zien is dat u zelf achter het stuur zat bij een van de afvaldumpingen. Hoe verklaart u dat?  
 Verdachte: Dat kan niet kloppen. Ik heb altijd legaal afval afgevoerd. Misschien is er **verbaak** van een vergissing.  
 Verbalisant: Mevrouw Vermeulen, gezien de bewijs die we hebben, zou een schikking een mogelijkheid kunnen zijn om deze zaak af te ronden. Wij stellen een schikking voorstel van 20.000 euro. Bent u bereid om te schikken?  
 Verdachte: Nee, ik wil niet **schrikken**, ik ben onschuldig.  
 (na een pauze)  
 Verbalisant: Heeft u nog iets toe te voegen aan uw verklaring of wilt u iets wijzigen?  
 Verdachte: Ja, ik heb nagedacht. Ik wil toch **schrikken**.  
 Verbalisant: Dat is een verstandig besluit. Zoals besproken stellen we een **schikking voorstel** van 20.000 euro. Bent u akkoord?  
 Verdachte: Ja, ik ga akkoord met de schikking.  
 Verbalisant: Prima, er is nog één voorwaarde voor de schikking. We hebben aanvullende informatie nodig over uw medeplichtigen. Bent u bereid om iemand aan te wijzen?  
 Verdachte: Ja, ik zal eerlijk zijn. Het was niet alleen mijn verantwoordelijkheid. Mijn collega, Peter van Dam, was ook betrokken. Hij heeft me gevraagd om mee te doen.

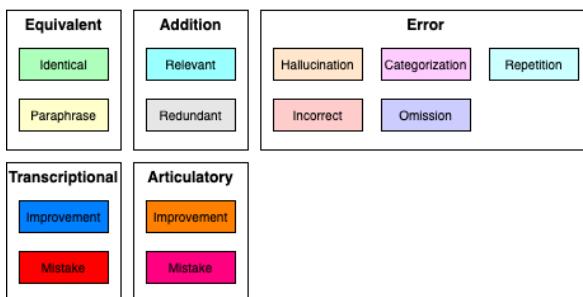


Figure 33. Model 2: Manual evaluation of generated statement (TR2)



### Interview transcript TR3 - Model 1

U bent hier vandaag omdat u verdacht wordt van betrokkenheid bij een reeks cyber aanvallen op verschillende bedrijven in de regio. Bent u bekend met de beschuldigingen? Ja, ik heb erover gehoord, maar ik heb er niets mee te maken. We hebben digitale sporen die terugleiden naar uw IP-adres, die betrokken zijn bij deze aanvallen. Kunt u dat verklaren? Nee, ik weet niet hoe dat kan. Misschien heeft iemand mijn internetverbinding gehackt. We hebben ook de logbestanden die aantonen dat er vanaf uw computer toegang is verkregen tot de netwerken van de getroffen bedrijven. Hoe verlaat u dat? Dat kan ik kloppen. Mijn computer is altijd goed beveiligd geweest. Verder hebben we chatberichten gevonden waar u nu met andere personen details bespreekt over de aanvallen. Wat heeft u daarop te zeggen? Ik weet niet waar u het over heeft. Ik heb nooit zulke gesprekken gevoerd. Mevrouw van Dongen, we hebben getuigen die beweren dat u recent dure elektronische apparatuur heeft aangeschaft, mogelijk met de opbrengsten van de cyberaanvallen. Hoe verlaat u die aankopen? Ik heb die spullen gekocht van mijn spaargeld en met wat ik verdien heb met freelance werk. Dat heeft niets met cyberaanvallers te maken. Oké, we zullen even een pauze inlassen. Denkt u alsjeblieft goed na over uw situatie. We gaan zo verder met het voorhoor. Oké, we gaan verder. Heeft u nog iets toe te voegen aan uw verklaring of wilt u iets wijzigen? Nee, ik blijf bij mijn verklaring. Ik heb niets verkeerd gedaan.

We hebben sterk bewijsmateriaal tegen u en het is belangrijk dat u de waarheid vertelt. Als u iemand anders wilt aanwijzen die hierbij betrokken was, kan dat mogelijk in uw voordeel werken. Bent u bereid om mee te werken? Ik begrijp niet waarom ik iemand zou moeten aanwijzen. Ik heb niets gedaan en ken niemand die dit gedaan zou hebben.



### Generated statement of TR3 - Model 1

Verbalisant: Kunt u uw volledige naam en geboortedatum bevestigen voor de opname?  
 Verdachte: Ja, mijn naam is Lisa-Maria van Dongen en ik ben geboren op 8 augustus 1994 in Utrecht.  
 Verbalisant: Woont u op de Nieuwe Gracht 15 3512 LC in Utrecht, klopt dat?  
 Verdachte: Ja, dat klopt.  
 Verbalisant: U bent hier vandaag omdat u verdacht wordt van betrokkenheid bij een reeks cyberaanvallen op verschillende bedrijven in de regio. Bent u bekend met de beschuldigingen?  
 Verdachte: Ja, ik heb erover gehoord, maar ik heb er niets mee te maken.  
 Verbalisant: We hebben digitale sporen die terugleiden naar uw IP-adres, die betrokken zijn bij deze aanvallen. Kunt u dat verklaren?  
 Verdachte: Nee, ik weet niet hoe dat kan. Misschien heeft iemand mijn internetverbinding gehackt.  
 Verbalisant: We hebben ook de logbestanden die aantonen dat er vanaf uw computer toegang is verkregen tot de netwerken van de getroffen bedrijven. Hoe verlaat u dat?  
 Verdachte: Dat kan ik kloppen. Mijn computer is altijd goed beveiligd geweest.  
 Verbalisant: Verder hebben we chatberichten gevonden waar u nu met andere personen details bespreekt over de aanvallen. Wat heeft u daarop te zeggen?  
 Verdachte: Ik weet niet waar u het over heeft. Ik heb nooit zulke gesprekken gevoerd.  
 Verbalisant: Mevrouw van Dongen, we hebben getuigen die beweren dat u recent dure elektronische apparatuur heeft aangeschaft, mogelijk met de opbrengsten van de cyberaanvallen. Hoe verlaat u die aankopen?  
 Verdachte: Ik heb die spullen gekocht van mijn spaargeld en met wat ik verdien heb met freelance werk. Dat heeft niets met cyberaanvallers te maken.  
 Verbalisant: Heeft u nog iets toe te voegen aan uw verklaring of wilt u iets wijzigen?  
 Verdachte: Nee, ik blijf bij mijn verklaring. Ik heb niets verkeerd gedaan.  
 Verbalisant: Als u iemand anders wilt aanwijzen die hierbij betrokken was, kan dat mogelijk in uw voordeel werken. Bent u bereid om mee te werken?  
 Verdachte: Ik begrijp niet waarom ik iemand zou moeten aanwijzen. Ik heb niets gedaan en ken niemand die dit gedaan zou hebben.

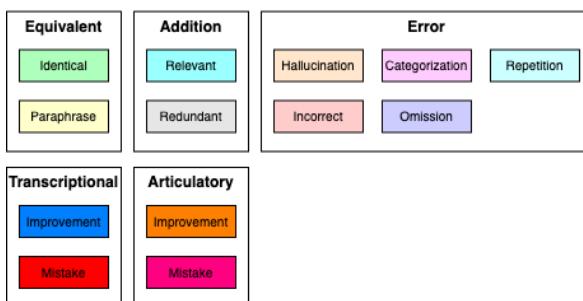


Figure 34. Model 1: Manual evaluation of generated statement (TR3)



### Interview transcript TR3 - Model 2

U bent hier vandaag omdat u verdacht wordt van betrokkenheid bij een reek cyber aanvallen op verschillende bedrijven in de regio. Bent u bekend met de beschuldigingen? Ja, ik heb erover gehoord, maar ik heb er niets mee te maken. We hebben digitale sporen die terugleiden naar uw IP-adres, die betrokken zijn bij deze aanvallen. Kunt u dat verklaren? Nee, ik weet niet hoe dat kan. Misschien heeft iemand mijn internetverbinding gehackt. We hebben ook de logbestanden die aantonen dat er vanaf uw computer toegang is verkregen tot de netwerken van de getroffen bedrijven. Hoe verlaat u dat? Dat kan ik kloppen. Mijn computer is altijd goed beveiligd geweest. Verder hebben we chatberichten gevonden waar u nu met andere personen details bespreekt over de aanvallen. Wat heeft u daarop te zeggen? Ik weet niet waar u het over heeft. Ik heb nooit zulke gesprekken gevoerd. Mevrouw van Dongen, we hebben getuigen die beweren dat u recent dure elektronische apparatuur heeft aangeschaft, mogelijk met de opbrengsten van de cyberaanvallen. Hoe verlaat u die aankopen? Ik heb die spullen gekocht van mijn spaargeld en met wat ik verdien heb met freelance werk. Dat heeft niets met cyber aanvallers te maken. Oké, we zullen even een pauze intrekken. Denkt u alsjeblieft goed na over uw situatie. We gaan zo verder met het voorhoofd. Oké, we gaan verder. Heeft u nog iets toe te voegen aan uw verklaring of wilt u iets wijzigen? Nee, ik blijf bij mijn verklaring. Ik heb niets verkeerd gedaan.

We hebben sterk bewijsmateriaal tegen u en het is belangrijk dat u de waarheid vertelt. Als u iemand anders wilt aanwijzen die hierbij betrokken was, kan dat mogelijk in uw voordeel werken. Bent u bereid om mee te werken? Ik begrijp niet waarom ik iemand zou moeten aanwijzen. Ik heb niets gedaan en ken niemand die dit gedaan zou hebben.



### Generated statement of TR3 - Model 2

Verbalisant: Bent u bekend met de beschuldigingen?  
 Verdachte: Ja, ik heb erover gehoord, maar ik heb er niets mee te maken.  
 Verbalisant: We hebben digitale sporen die terugleiden naar uw IP-adres, die betrokken zijn bij deze aanvallen. Kunt u dat verklaren?  
 Verdachte: Nee, ik weet niet hoe dat kan. Misschien heeft iemand mijn internetverbinding gehackt.  
 Verbalisant: We hebben ook de logbestanden die aantonen dat er vanaf uw computer toegang is verkregen tot de netwerken van de getroffen bedrijven. Hoe verlaat u dat?  
 Verdachte: Dat kan niet kloppen. Mijn computer is altijd goed beveiligd geweest.  
 Verbalisant: Verder hebben we chatberichten gevonden waar u met andere personen details bespreekt over de aanvallen. Wat heeft u daarop te zeggen?  
 Verdachte: Ik weet niet waar u het over heeft. Ik heb nooit zulke gesprekken gevoerd.  
 Verbalisant: Mevrouw van Dongen, we hebben getuigen die beweren dat u recent dure elektronische apparatuur heeft aangeschaft, mogelijk met de opbrengsten van de cyberaanvallen. Hoe verlaat u die aankopen?  
 Verdachte: Ik heb die spullen gekocht van mijn spaargeld en met wat ik verdien heb met freelance werk. Dat heeft niets met cyberaanvallen te maken.  
 Verbalisant: Heeft u nog iets toe te voegen aan uw verklaring of wilt u iets wijzigen?  
 Verdachte: Nee, ik blijf bij mijn verklaring. Ik heb niets verkeerd gedaan.  
 Verbalisant: Bent u bereid om mee te werken?  
 Verdachte: Ik begrijp niet waarom ik iemand zou moeten aanwijzen. Ik heb niets gedaan en ken niemand die dit gedaan zou hebben.

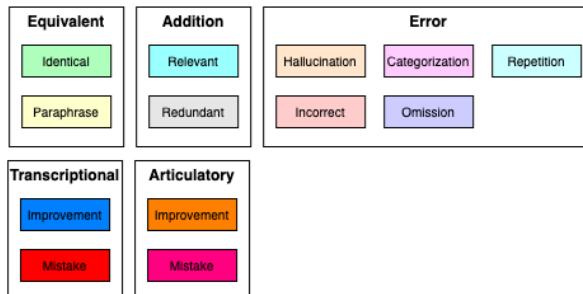


Figure 35. Model 2: Manual evaluation of generated statement (TR3)

## Interrogation transcript TR4 - Model 1

U bent hier vandaag omdat u verdacht wordt van de betrokkenheid bij de handel in verdovende middelen, specifiek **metamfetamine**. Kent u een persoon genaamd Jeroen de Vries? Ja, die ken ik. Wij zijn vrienden sinds de middelbare school. Jeroen de Vries is onlangs gearresteerd voor het bezit en de verkoop van grote hoeveelheden metamfetamine. Tijdens het verhoor heeft hij verklaard dat u zijn leverancier bent. Wat heeft u daarop te zeggen? Dat is helemaal niet waar. Ik heb niets met drugs te maken. We hebben sms berichten tussen u en Jeroen waarin u afspraken maakt over de leveringen en betalingen. Hoe verlaat u dat? We praten over allerlei dingen, maar nooit over drugs. Misschien zijn die berichten verkeerd geïnterpreteerd. We hebben ook getuigeverklaringen van buren die u vaak met verdachte pakketjes hebben zien lopen. Wat zat er in die pakketjes? Dat waren gewoon spullen die ik heb verkocht op Marktplaats. Geen drugs. Daarnaast zijn er camerabeelden waarop te zien is dat u bij Jeroen thuis komt met een rugzak en weer vertrekt zonder rugzak, vlak voordat hij werd gearresteerd met een grote partij **metamfetamine**. Kunt u dat uitleggen? Misschien heb ik iets bij hem afgegeven of geleend, maar dat had niets met drugs te maken. Meneer Kuipers, gezien de bewijzen die wij hebben, zou een schikking een mogelijkheid zijn om deze zaak af te ronden. Wij stellen een schikingsvoorstel voor van 15.000 euro. Bent u daarbij bereid om te schikken? Nee, ik wil niet schikken. Ik heb echt niets verkeerds gedaan. We zullen zo even pauze nemen. Denkt u alsjeblieft goed na over uw situatie. We gaan zo verder met het verhoor. We gaan verder. Heeft u nog iets toe te voegen aan uw verklaring of wilt u iets wijzigen? Ja, ik heb er over nagedacht. Ik wil toch schikken. Dat is een verstandig besluit. Zoals besproken stellen we een schikingsvoorstel voor van €15.000. Bent u akkoord? Ja, ik ga akkoord met de schikking. Prima. Er is nog één voorwaarde van de schikking. We hebben aanvullende informatie nodig over uw medeplichtigen. Bent u bereid iemand aan te wijzen? Ja, ik zou eerlijk zijn. Het was niet alleen Jeroen. Er was nog iemand bij betrokken, genaamd Rick Jansen. Hij regelde de distributie.

## Generated statement of TR4 - Model 1

Verbalisant: Kunt u uw volledige naam en geboortedatum bevestigen voor de opname?  
 Verdachte: Ja, mijn naam is Sven Johannes Kuipers en ik ben geboren op 19 oktober 1990 in Amsterdam.  
 Verbalisant: Dank u meneer Kuipers. U woont op de Lindegracht 24, 1015 KB Amsterdam, klopt dat?  
 Verdachte: Ja, dat klopt.  
 Verbalisant: U bent hier vandaag omdat u verdacht wordt van de betrokkenheid bij de handel in verdovende middelen, specifiek **metamfetamine**. Kent u een persoon genaamd Jeroen de Vries?  
 Verdachte: Ja, die ken ik. Wij zijn vrienden sinds de middelbare school.  
 Verbalisant: Jeroen de Vries is onlangs gearresteerd voor het bezit en de verkoop van grote hoeveelheden **metamfetamine**. Tijdens het verhoor heeft hij verklaard dat u zijn leverancier bent. Wat heeft u daarop te zeggen?  
 Verdachte: Dat is helemaal niet waar. Ik heb niets met drugs te maken.  
 Verbalisant: We hebben sms berichten tussen u en Jeroen waarin u afspraken maakt over de leveringen en betalingen. Hoe verlaat u dat?  
 Verdachte: We praten over allerlei dingen, maar nooit over drugs. Misschien zijn die berichten verkeerd geïnterpreteerd.  
 Verbalisant: We hebben ook getuigeverklaringen van buren die u vaak met verdachte pakketjes hebben zien lopen. Wat zat er in die pakketjes?  
 Verdachte: Dat waren gewoon spullen die ik heb verkocht op Marktplaats. Geen drugs.  
 Verbalisant: Daarnaast zijn er camerabeelden waarop te zien is dat u bij Jeroen thuis komt met een rugzak en weer vertrekt zonder rugzak, vlak voordat hij werd gearresteerd met een grote partij **metamfetamine**. Kunt u dat uitleggen?  
 Verdachte: Misschien heb ik iets bij hem afgegeven of geleend, maar dat had niets met drugs te maken.  
 Verbalisant: Meneer Kuipers, gezien de bewijzen die wij hebben, zou een schikking een mogelijkheid zijn om deze zaak af te ronden. Wij stellen een schikingsvoorstel voor van 15.000 euro. Bent u daarbij bereid om te schikken?  
 Verdachte: Nee, ik wil niet schikken. Ik heb echt niets verkeerds gedaan.  
 Verbalisant: We zullen zo even pauze nemen. Denkt u alsjeblieft goed na over uw situatie. We gaan zo verder met het verhoor.  
 Verbalisant: We gaan verder. Heeft u nog iets toe te voegen aan uw verklaring of wilt u iets wijzigen?  
 Verdachte: Ja, ik heb er over nagedacht. Ik wil toch schikken.  
 Verbalisant: Dat is een verstandig besluit. Zoals besproken stellen we een schikingsvoorstel voor van €15.000. Bent u akkoord?  
 Verdachte: Ja, ik ga akkoord met de schikking.  
 Verbalisant: Prima. Er is nog één voorwaarde van de schikking. We hebben aanvullende informatie nodig over uw medeplichtigen. Bent u bereid iemand aan te wijzen?  
 Verdachte: Ja, ik zou eerlijk zijn. Het was niet alleen Jeroen. Er was nog iemand bij betrokken, genaamd Rick Jansen. Hij regelde de distributie.

| Equivalent | Addition  | Error          |
|------------|-----------|----------------|
| Identical  | Relevant  | Hallucination  |
| Paraphrase | Redundant | Categorization |
|            |           | Repetition     |
|            |           | Incorrect      |
|            |           | Omission       |

| Transcriptional | Articulatory |
|-----------------|--------------|
| Improvement     | Improvement  |
| Mistake         | Mistake      |

Figure 36. Model 1: Manual evaluation of generated statement (TR4)

## Interrogation transcript TR4 - Model 2

**U bent hier vandaag omdat u verdacht wordt van de betrokkenheid bij de handel in verdovende middelen, specifiek **menfentamine**. Kent u een persoon genaamd Jeroen de Vries? Ja, die ken ik. Wij zijn vrienden sinds de middelbare school. Jeroen de Vries is onlangs gearresteerd voor het bezit en de verkoop van grote hoeveelheden metamfetamine. Tijdens het verhoor heeft hij verklaard dat u zijn leverancier bent. Wat heeft u daarop te zeggen? Dat is helemaal niet waar. Ik heb niets met drugs te maken. We hebben sms berichten tussen u en Jeroen waarin u afspraken maakt over de leveringen en betalingen. Hoe verlaat u dat? We praten over allerlei dingen, maar nooit over drugs. Misschien zijn die berichten verkeerd geïnterpreteerd. We hebben ook getuigeverklaringen van buren die u vaak met verdachte pakketjes hebben zien lopen. Wat zat er in die pakketjes? Dat waren gewoon spullen die ik heb verkocht op Marktplaats. Geen drugs. Daarnaast zijn er camerabeelden waarop te zien is dat u bij Jeroen thuis komt met een rugzak en weer vertrekt zonder rugzak, vlak voordat hij werd gearresteerd met een grote partij **metamfetamine**. Kunt u dat uitleggen? Misschien heb ik iets bij hem afgegeven of geleend, maar dat had niets met drugs te maken. Meneer Kuipers, gezien de bewijzen die wij hebben, zou een schikking een mogelijkheid zijn om deze zaak af te ronden. Wij stellen een schikkingsvoorstel voor van 15.000 euro. Bent u daarbij bereid om te schikken? Nee, ik wil niet schikken. Ik heb echt niets verkeerds gedaan. We zullen zo even pauze nemen. Denkt u alsjeblieft goed na over uw situatie. We gaan zo verder met het verhoor. We gaan verder. Heeft u nog iets toe te voegen aan uw verklaring of wilt u iets wijzigen? Ja, ik heb er over nagedacht. Ik wil toch schikken. Dat is een verstandig besluit. Zoals besproken stellen we een schikkingsvoorstel voor van €15.000. Bent u akkoord? Ja, ik ga akkoord met de schikking. Prima. Er is nog één voorwaarde van de schikking. We hebben aanvullende informatie nodig over uw medeplichtigen. Bent u bereid iemand aan te wijzen? Ja, ik zou eerlijk zijn. Het was niet alleen Jeroen. Er was nog iemand bij betrokken, genaamd Rick Jansen. Hij regelde de distributie.**

## Generated statement of TR4 - Model 2

Verbalist: Kunt u uw volledige naam en geboortedatum bevestigen voor de opname?  
 Verdachte: Ja, mijn naam is Sven Johannes Kuipers en ik ben geboren op 19 oktober 1990 in Amsterdam.  
 Verbalist: U woont op de Lindegracht 24 1015 KB Amsterdam, klopt dat?  
 Verdachte: Ja, dat klopt.  
 Verbalist: U bent hier vandaag omdat u verdacht wordt van de betrokkenheid bij de handel in **verdovende middelen, specifiek menfentamine**. Kent u een persoon genaamd Jeroen de Vries?  
 Verdachte: Ja, die ken ik. Wij zijn vrienden sinds de middelbare school.  
 Verbalist: Jeroen de Vries is onlangs gearresteerd voor het bezit en de verkoop van grote hoeveelheden **metamfetamine**. Tijdens het verhoor heeft hij verklaard dat u zijn leverancier bent. Wat heeft u daarop te zeggen?  
 Verdachte: Dat is helemaal niet waar. Ik heb niets met drugs te maken.  
 Verbalist: We hebben sms berichten tussen u en Jeroen waarin u afspraken maakt over de leveringen en betalingen. Hoe verlaat u dat?  
 Verdachte: We praten over allerlei dingen, maar nooit over drugs. Misschien zijn die berichten verkeerd geïnterpreteerd.  
 Verbalist: We hebben ook getuigeverklaringen van buren die u vaak met verdachte pakketjes hebben zien lopen. Wat zat er in die pakketjes?  
 Verdachte: Dat waren gewoon spullen die ik heb verkocht op Marktplaats. Geen drugs.  
 Verbalist: Daarnaast zijn er camerabeelden waarop te zien is dat u bij Jeroen thuis komt met een rugzak en weer vertrekt zonder rugzak, vlak voordat hij werd gearresteerd met een grote partij **metamfetamine**. Kunt u dat uitleggen?  
 Verdachte: Misschien heb ik iets bij hem afgegeven of geleend, maar dat had niets met drugs te maken.  
 Verbalist: Meneer Kuipers, gezien de bewijzen die wij hebben, zou een schikking een mogelijkheid zijn om deze zaak af te ronden. Wij stellen een schikkingsvoorstel voor van 15.000 euro. Bent u daarbij bereid om te schikken?  
 Verdachte: Nee, ik wil niet schikken. Ik heb echt niets verkeerds gedaan.  
 Verbalist: Heeft u nog iets toe te voegen aan uw verklaring of wilt u iets wijzigen?  
 Verdachte: Ja, ik heb er over nagedacht. Ik wil toch schikken.  
 Verbalist: Dat is een verstandig besluit. Zoals besproken stellen we een schikkingsvoorstel voor van €15.000. Bent u akkoord?  
 Verdachte: Ja, ik ga akkoord met de schikking.  
 Verbalist: Prima. Er is nog één voorwaarde van de schikking. We hebben aanvullende informatie nodig over uw medeplichtigen. Bent u bereid iemand aan te wijzen?  
 Verdachte: Ja, ik zou eerlijk zijn. Het was niet alleen Jeroen. Er was nog iemand bij betrokken, genaamd Rick Jansen. Hij regelde de distributie.

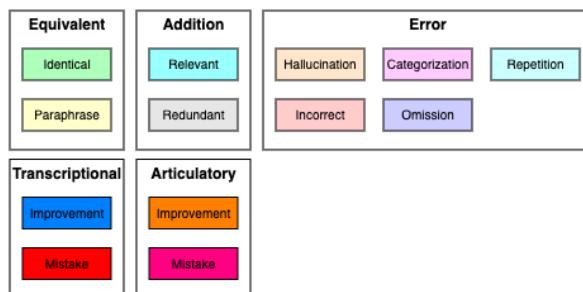


Figure 37. Model 2: Manual evaluation of generated statement (TR4)



### Interrogation transcript TR5 - Model 1

U bent hier vandaag omdat u verdacht wordt van betrokkenheid bij een omvangrijke financiële fraude die plaatsvond tussen januari 2003 en april 2023 en april 2024. Kent u het bedrijf Vanco BV? Ja, ik ken dat bedrijf. Ik heb er gewerkt als financieel analist. Kunt u ons vertellen wat uw verantwoordelijkheden waren bij Vanco BV? Ik was verantwoordelijk voor het analyseren van financiële gegevens en het opstellen van rapporten voor het management. Er zijn onregelmatigheden gevonden in de boekhouding van Vanco BV, waarbij grote sommen geld zijn verdwenen. Bent u hiervan op de hoogte? Nee, dat wist ik niet. Ik heb mijn werk altijd naar beste kunnen gedaan. Onze gegevens tonen aan dat vanaf uw werkcomputer toegang is gekregen tot de vertrouwelijke rekeningen waaruit geld is overgemaakt naar **bijplantrekeningen**. Kunt u dat verklaren? Dat moet iemand anders geweest zijn. Mijn werkcomputer was beveiligd met een wachtwoord dat alleen ik kende. Er zijn ook getuigen die beweren dat u zich verdacht gedroeg en vaak buiten werktijd in het kantoor aanwezig was. Wat heeft u daarop te zeggen? Ik werkte soms laat omdat ik mijn taken wilde afmaken. **Er was** niets verdachts aan. We hebben ook een e-mail-correspondentie gevonden tussen u en een onbekende partij waarin grote geldbedragen worden besproken. Hoe verklaart u dat? Ik weet echt niet waar u het over heeft. Ik heb nooit zulke e-mails verstuurd of ontvangen. Mevrouw de Wit, gezien de bewijzen die wij hebben verzameld, zou een schikking kunnen overwegen om deze zaak af te handelen. We stellen een schikkingsvoorstel voor van 10.000 euro. Bent u bereid om te schikken? Nee, ik wil niet schikken, omdat ik niks verkeerd heb gedaan. Goed, we zullen even pauze inlassen. Denkt u alsjeblieft goed na over de situatie. We gaan zo verder met het verhoor. We gaan verder. Heeft u nog iets toe te voegen van uw verklaring of wilt u iets wijzigen? Nee, ik blijf bij mijn verklaring. Ik ben onschuldig.



### Generated statement of TR5 - Model 1

Verbalisant: Kunt u uw volledige naam en geboortedatum bevestigen voor de opname?  
 Verdachte: Ja, mijn naam is Emma Maria de Wit en ik ben geboren op 5 juli 1992 in Utrecht.  
 Verbalisant: U woont op de Julianalaan 12, 3062 BB Rotterdam, correct?  
 Verdachte: Ja, dat klopt.  
 Verbalisant: **U bent hier vandaag omdat u verdacht wordt van betrokkenheid bij een omvangrijke financiële fraude die plaatsvond tussen januari 2003 en april 2023 en april 2024. Kent u het bedrijf Vanco BV?**  
 Verdachte: Ja, ik ken dat bedrijf. Ik heb er gewerkt als financieel analist.  
 Verbalisant: Kunt u ons vertellen wat uw verantwoordelijkheden waren bij Vanco BV?  
 Verdachte: **Ik was verantwoordelijk voor het analyseren van financiële gegevens en het opstellen van rapporten voor het management.**  
 Verbalisant: **Er zijn onregelmatigheden gevonden in de boekhouding van Vanco BV, waarbij grote sommen geld zijn verdwenen. Bent u hiervan op de hoogte?**  
 Verdachte: Nee, dat wist ik niet. Ik heb mijn werk altijd naar beste kunnen gedaan.  
 Verbalisant: Onze gegevens tonen aan dat vanaf uw werkcomputer toegang is gekregen tot de vertrouwelijke rekeningen waaruit geld is overgemaakt naar **bijplantrekeningen**. Kunt u dat verklaren?  
 Verdachte: Dat moet iemand anders geweest zijn. Mijn werkcomputer was beveiligd met een wachtwoord dat alleen ik kende.  
 Verbalisant: **Er zijn ook getuigen die beweren dat u zich verdacht gedroeg en vaak buiten werktijd in het kantoor aanwezig was. Wat heeft u daarop te zeggen?**  
 Verdachte: Ik werkte soms laat omdat ik mijn taken wilde afmaken. **Er was** niets verdachts aan.  
 Verbalisant: We hebben ook een e-mail-correspondentie gevonden tussen u en een onbekende partij waarin grote geldbedragen worden besproken. Hoe verklaart u dat?  
 Verdachte: Ik weet echt niet waar u het over heeft. Ik heb nooit zulke e-mails verstuurd of ontvangen.  
 Verbalisant: Mevrouw de Wit, gezien de bewijzen die wij hebben verzameld, zou een schikking kunnen overwegen om deze zaak af te handelen. We stellen een schikkingsvoorstel voor van 10.000 euro. Bent u bereid om te schikken?  
 Verdachte: Nee, ik wil niet schikken, omdat ik niks verkeerd heb gedaan.  
 Verbalisant: **Heeft u nog iets toe te voegen aan uw verklaring of wilt u iets wijzigen?**  
 Verdachte: Nee, ik blijf bij mijn verklaring. Ik ben onschuldig.

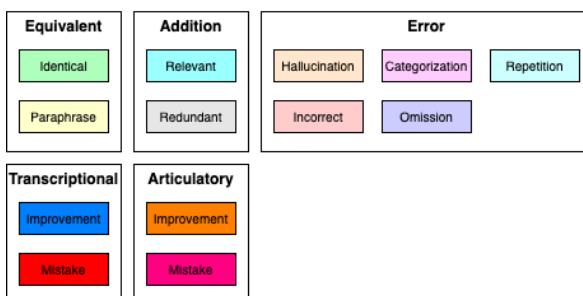


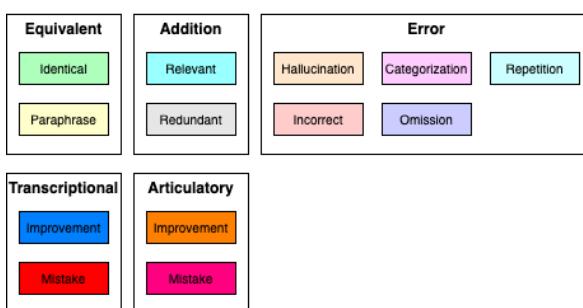
Figure 38. Model 1: Manual evaluation of generated statement (TR5)

## Interrogation transcript TR5 - Model 2

U bent hier vandaag omdat u verdacht wordt van betrokkenheid bij een omvangrijke financiële fraude die plaatsvond tussen januari 2003 en april 2023 en april 2024. Kent u het bedrijf Vanco BV? Ja, ik ken dat bedrijf. Ik heb er gewerkt als financieel analist. Kunt u ons vertellen wat uw verantwoordelijkheden waren bij Vanco BV? Ik was verantwoordelijk voor het analyseren van financiële gegevens en het opstellen van rapporten voor het management. Er zijn onregelmatigheden gevonden in de boekhouding van Vanco BV, waarbij grote sommen geld zijn verdwenen. Bent u hiervan op de hoogte? Nee, dat wist ik niet. Ik heb mijn werk altijd naar beste kunnen gedaan. Onze gegevens tonen aan dat dit vanaf uw werkcomputer toegang is gekregen tot de vertrouwelijke rekeningen waaruit geld is overgemaakt naar **bijplantrekeningen**. Kunt u dat verklaren? Dat moet iemand anders geweest zijn. Mijn werkcomputer was beveiligd met een wachtwoord dat alleen ik kende. Er zijn ook getuigen die beweren dat u zich verdacht gedroeg en vaak buiten werktijd in het kantoor aanwezig was. Wat heeft u daarop te zeggen? Ik werkte soms laat omdat ik mijn taken wilde afmaken. **Er was** niets verdachts aan. We hebben ook een e-mail-correspondentie gevonden tussen u en een onbekende partij waarin grote geldbedragen worden besproken. Hoe verklaart u dat? Ik weet echt niet waar u het over heeft. Ik heb nooit zulke e-mails verstuurd of ontvangen. Mevrouw de Wit, gezien de bewijzen die wij hebben verzameld, zou u een schikking kunnen overwegen om deze zaak af te handelen. We stellen een schikkingsoorstel voor van 10.000 euro. Bent u bereid om te schikken? Nee, ik wil niet schikken, omdat ik nijs verkeerd heb gedaan. Goed, we zullen even pauze inlassen. Denkt u alstublieft goed na over de situatie. We gaan verder. Heeft u nog iets toe te voegen van uw verklaring of wilt u iets wijzigen? Nee, ik blijf bij mijn verklaring. Ik ben onschuldig.

## Generated statement of TR5 - Model 2

Verbalisant: Kunt u uw volledige naam en geboortedatum bevestigen voor de opname?  
 Verdachte: Ja, mijn naam is Emma Maria de Wit en ik ben geboren op 5 juli 1992 in Utrecht.  
 Verbalisant: U woont op de Julianalaan 12, 3062 BB Rotterdam, correct?  
 Verdachte: Ja, dat klopt.  
 Verbalisant: U bent hier vandaag omdat u verdacht wordt van betrokkenheid bij een omvangrijke financiële fraude die plaatsvond tussen januari 2023 en april 2024. Kent u het bedrijf Vanco BV?  
 Verdachte: Ja, ik ken dat bedrijf. Ik heb er gewerkt als financieel analist.  
 Verbalisant: Kunt u ons vertellen wat uw verantwoordelijkheden waren bij Vanco BV?  
 Verdachte: Ik was verantwoordelijk voor het analyseren van financiële gegevens en het opstellen van rapporten voor het management.  
 Verbalisant: Er zijn onregelmatigheden gevonden in de boekhouding van Vanco BV, waarbij grote sommen geld zijn verdwenen. Bent u hiervan op de hoogte?  
 Verdachte: Nee, dat wist ik niet. Ik heb mijn werk altijd naar beste kunnen gedaan.  
 Verbalisant: Onze gegevens tonen aan dat via uw werkcomputer toegang is verkregen tot de vertrouwelijke rekeningen waaruit geld is overgemaakt naar **bijplantrekeningen**. Kunt u dat verklaren?  
 Verdachte: Dat moet iemand anders geweest zijn. Mijn werkcomputer was beveiligd met een wachtwoord dat alleen ik kende.  
 Verbalisant: Er zijn ook getuigen die beweren dat u zich verdacht gedroeg en vaak buiten werktijd in het kantoor aanwezig was. Wat heeft u daarop te zeggen?  
 Verdachte: IK werkte soms laat omdat ik mijn taken wilde afmaken. **Er was** niets verdachts aan.  
 Verbalisant: We hebben ook een e-mailcorrespondentie gevonden tussen u en een onbekende partij waarin grote geldbedragen worden besproken. Hoe verklaart u dat?  
 Verdachte: Ik weet echt niet waar u het over heeft. Ik heb nooit zulke e-mails verstuurd of ontvangen.  
 Verbalisant: Mevrouw de Wit, gezien de bewijzen die wij hebben verzameld, zou u een schikking kunnen overwegen om deze zaak af te handelen. We stellen een schikkingsoorstel voor van 10.000 euro. Bent u bereid om te schikken?  
 Verdachte: Nee, ik wil niet schikken, omdat ik nijs verkeerd heb gedaan.  
 Verbalisant: Goed, we zullen even pauze inlassen. Denkt u alstublieft goed na over de situatie. We gaan zo verder met het verhoor.  
 Verbalisant: We gaan verder. Heeft u nog iets toe te voegen aan uw verklaring of wilt u iets wijzigen?  
 Verdachte: Nee, ik blijf bij mijn verklaring. Ik ben onschuldig.  
 Verbalisant: Dank u mevrouw de Wit. U kunt gaan, maar we kunnen u mogelijk terugvragen voor verder onderzoek. Het verhoor is hiermee afgelopen.



**Figure 39.** Model 2: Manual evaluation of generated statement (TR5)

## References

- [1] Abd Rahman, A., Haron, A., Sahibuddin, S., and Harun, M. (2014). An empirical study of the software project requirements engineering practice in malaysian public sector—a perspective from the stakeholders' challenges. *International Journal of Computer Theory and Engineering*, 6(1):52.
- [2] Adams, R. J., Smart, P., and Huff, A. S. (2017). Shades of grey: guidelines for working with the grey literature in systematic reviews for management and organizational studies. *International Journal of Management Reviews*, 19(4):432–454.
- [3] Ahmad, K., Abdelrazek, M., Arora, C., Bano, M., and Grundy, J. (2023a). Requirements engineering for artificial intelligence systems: A systematic mapping study. *Information and Software Technology*, page 107176.
- [4] Ahmad, K., Abdelrazek, M., Arora, C., Bano, M., and Grundy, J. (2023b). Requirements practices and gaps when engineering human-centered artificial intelligence systems. *Applied Soft Computing*, 143:110421.
- [5] Ahmad, K., Bano, M., Abdelrazek, M., Arora, C., and Grundy, J. (2021). What's up with requirements engineering for artificial intelligence systems? In *2021 IEEE 29th International Requirements Engineering Conference (RE)*, pages 1–12. IEEE.
- [6] Al-Sarayreh, K. T., Meridji, K., Alenezi, M., Zarour, M., and Al-Majali, M. D. (2021). A sustainable procedural method of software design process improvements. *Indonesian Journal of Electrical Engineering and Computer Science*, 21(1):440–449.
- [7] Alazzawi, A., Rahmatullah, B., et al. (2023). A comprehensive review of software development life cycle methodologies: Pros, cons, and future directions. *Iraqi Journal For Computer Science and Mathematics*, 4(4):173–190.
- [8] Albin, S. T. (2003). *The art of software architecture: design methods and techniques*, volume 9. John Wiley & Sons.
- [9] Aleti, A. (2023). Software testing of generative ai systems: Challenges and opportunities. *arXiv preprint arXiv:2309.03554*.
- [10] Ali, S. S., Zafar, M. S., and Saeed, M. T. (2020). Effort estimation problems in software maintenance—a survey. In *2020 3rd international conference on computing, mathematics and engineering technologies (iCoMET)*, pages 1–9. IEEE.

- [11] Aroral, H. K. (2021). Waterfall process operations in the fast-paced world: Project management exploratory analysis. *International Journal of Applied Business and Management Studies*, 6(1):91–99.
- [12] Banh, L. and Strobel, G. (2023). Generative artificial intelligence. *Electronic Markets*, 33(1):1–17.
- [13] Baresi, L. and Pezze, M. (2006). An introduction to software testing. *Electronic Notes in Theoretical Computer Science*, 148(1):89–111.
- [14] Bataller, C. and Harris, J. (2016). Turning artificial intelligence into business value. *Today*. Retrieved from: <https://pdfs.semanticscholar.org/a710/a8d529bce6bd75ba589f42721777bf54d3b.pdf>.
- [15] Belani, H., Vukovic, M., and Car, Ž. (2019). Requirements engineering challenges in building ai-based complex systems. In *2019 IEEE 27th International Requirements Engineering Conference Workshops (REW)*, pages 252–255. IEEE.
- [16] Bertolino, A. (2007). Software testing research: Achievements, challenges, dreams. In *Future of Software Engineering (FOSE'07)*, pages 85–103. IEEE.
- [17] Birkstedt, T., Minkkinen, M., Tandon, A., and Mäntymäki, M. (2023). Ai governance: themes, knowledge gaps and future agendas. *Internet Research*, 33(7):133–167.
- [18] Boivin, R. and Gendron, A. (2022). An experimental study of the impact of body-worn cameras on police report writing. *Journal of Experimental Criminology*, 18(4):747–764.
- [19] Bosch, J., Olsson, H. H., and Crnkovic, I. (2021). Engineering ai systems: A research agenda. *Artificial Intelligence Paradigms for Smart Cyber-Physical Systems*, pages 1–19.
- [20] Bottenbley, A. (2023). The algorithm agenda: A discourse analysis of the dutch algorithm register. Master's thesis, University of Utrecht.
- [21] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- [22] CBS (2023). Geregistreerde criminaliteit; soort misdrijf, regio. <https://opendata.cbs.nl/statline/#/CBS/nl/dataset/83648NED/table?fromstatweb>.

- [23] Chang, Y., Wang, X., Wang, J., Wu, Y., Zhu, K., Chen, H., Yang, L., Yi, X., Wang, C., Wang, Y., et al. (2023). A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*.
- [24] Cois, C. A., Yankel, J., and Connell, A. (2014). Modern devops: Optimizing software development through effective system interactions. In *2014 IEEE international professional communication conference (IPCC)*, pages 1–7. IEEE.
- [25] Committee, I. C. S. P. P. et al. (2014). Swebok: Guide to the software engineering body of knowledge, 2014 version. *IEEE Computer Society*.
- [26] Danhof, L. and Boon, I. (2016). Een goed pv, daar vang je boeven mee. <https://www.politieacademie.nl/kennisonderzoek/kennis/mediatheek/pdf/91730.pdf>.
- [27] Delgado, F., Yang, S., Madaio, M., and Yang, Q. (2021). Stakeholder participation in ai: Beyond “add diverse stakeholders and stir”. *arXiv preprint arXiv:2111.01122*.
- [28] Deshpande, A. and Sharp, H. (2022). Responsible ai systems: Who are the stakeholders? In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 227–236.
- [29] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- [30] Dima, A. M. and Maassen, M. A. (2018). From waterfall to agile software: Development models in the it sector, 2006 to 2018. impacts on company management. *Journal of International Studies (2071-8330)*, 11(2).
- [31] Ebert, C., Gallardo, G., Hernantes, J., and Serrano, N. (2016). Devops. *IEEE software*, 33(3):94–100.
- [32] Ebneyamini, S. and Sadeghi Moghadam, M. R. (2018). Toward developing a framework for conducting case study research. *International journal of qualitative methods*, 17(1):1609406918817954.
- [33] Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., and Srikumar, M. (2020). Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for ai. *Berkman Klein Center Research Publication*.
- [34] Flach, R. N., Fransen, N. L., Sonnen, A. F., Nguyen, T. Q., Breimer, G. E., Veta, M., Stathonikos, N., van Dooijeweert, C., and van Diest, P. J. (2022). Implementation of artificial intelligence in diagnostic practice as a next step after going digital: the umc utrecht perspective. *Diagnostics*, 12(5):1042.

- [35] Foidl, H., Felderer, M., and Ramler, R. (2022). Data smells: categories, causes and consequences, and detection of suspicious data in ai-based systems. In *Proceedings of the 1st International Conference on AI Engineering: Software Engineering for AI*, pages 229–239.
- [36] Gall, M. and Pigni, F. (2022). Taking devops mainstream: a critical review and conceptual framework. *European Journal of Information Systems*, 31(5):548–567.
- [37] Galster, M., Mirakhori, M., and Koziolek, A. (2015). Twin peaks goes agile. *ACM SIGSOFT Software Engineering Notes*, 40(5):47–49.
- [38] Garousi, V., Felderer, M., and Mäntylä, M. V. (2019). Guidelines for including grey literature and conducting multivocal literature reviews in software engineering. *Information and software technology*, 106:101–121.
- [39] Georgievski, I. (2023). Conceptualising software development lifecycle for engineering ai planning systems. In *2023 IEEE/ACM 2nd International Conference on AI Engineering–Software Engineering for AI (CAIN)*, pages 88–89. IEEE.
- [40] Gezici, B. and Tarhan, A. K. (2022). Systematic literature review on software quality for ai-based software. *Empirical Software Engineering*, 27(3):66.
- [41] Glinz, M. (2007). On non-functional requirements. In *15th IEEE international requirements engineering conference (RE 2007)*, pages 21–26. IEEE.
- [42] Gokarna, M. (2023). Devops phases across software development lifecycle. *Authorea Preprints*.
- [43] Hamers, L., Ceulen, R., Spapens, T., and Groenhuijsen, M. (2019). Verbalisering van het getuigenverhoor. *Forensic Linguistics*.
- [44] He, J., Baxter, S. L., Xu, J., Xu, J., Zhou, X., and Zhang, K. (2019). The practical implementation of artificial intelligence technologies in medicine. *Nature medicine*, 25(1):30–36.
- [45] Heyn, H.-M., Knauss, E., Muhammad, A. P., Eriksson, O., Linder, J., Subbiah, P., Pradhan, S. K., and Tungal, S. (2021). Requirement engineering challenges for ai-intense systems development. In *2021 IEEE/ACM 1st Workshop on AI Engineering-Software Engineering for AI (WAIN)*, pages 89–96. IEEE.
- [46] Jansen, T. G. (2011). Juridische kennis en het proces-verbaal. <https://www.politieacademie.nl/kennisenonderzoek/kennis/mediatheek/PDF/81033.PDF>.

- [47] Jüngling, S., Fetai, I., Rogger, A., Morandi, D., and Peraic, M. (2022). On the track to application architectures in public transport service companies. *Applied Sciences*, 12(12):6073.
- [48] Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günemann, S., Hüllermeier, E., et al. (2023). Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.
- [49] Kim, D. H., Kim, Y., Shin, H., Son, J., and Kim, J. (2023). Towards understanding the challenges and remedies in ai application development planning. *The journal of the Korean Information Science Society.*, pages 1421–1423.
- [50] Komter, M. (2022). Institutional and academic transcripts of police interrogations. *Frontiers in Communication*, 7:797145.
- [51] Koskimies, E., Stenvall, J., Kinder, T., Leikas, J., and Nieminen, M. (2022). 5 artificial intelligence and public innovations. *Public Innovation and Digital Transformation*, page 68.
- [52] Lamsweerde, A. v. (2009). *Requirements engineering: from system goals to UML models to software specifications*. John Wiley & Sons, Ltd.
- [53] Leffingwell, D. and Widrig, D. (2000). *Managing software requirements: a unified approach*. Addison-Wesley Professional.
- [54] Lu, Q., Zhu, L., Xu, X., Whittle, J., and Xing, Z. (2022). Towards a roadmap on software engineering for responsible ai. In *Proceedings of the 1st International Conference on AI Engineering: Software Engineering for AI*, pages 101–112.
- [55] Lu, Q., Zhu, L., Xu, X., Whittle, J., Zowghi, D., and Jacquet, A. (2023). Responsible ai pattern catalogue: A collection of best practices for ai governance and engineering. *ACM Computing Surveys*.
- [56] Maas, L., Geurtsen, M., Nouwt, F., Schouten, S., Van De Water, R., Van Dulmen, S., Dalpiaz, F., Van Deemter, K., and Brinkkemper, S. (2020). The care2report system: Automated medical reporting as an integrated solution to reduce administrative burden in healthcare. In *HICSS*, pages 1–10.
- [57] Maas, L., Kisjes, A., Hashemi, I., Heijmans, F., Dalpiaz, F., Van Dulmen, S., and Brinkkemper, S. (2021). Automated medical reporting: From multimodal inputs to medical reports through knowledge graphs. In *HEALTHINF*, pages 509–514.
- [58] Maassen, M. A. (2018). Product development models in the it sector-from waterfall to agile project management model s in the case of avira soft srl. In

*Proceedings of the International Conference on Business Excellence*, volume 12, pages 568–578.

- [59] Machado, C. G., Winroth, M., Carlsson, D., Almström, P., Centerholt, V., and Hallin, M. (2019). Industry 4.0 readiness in manufacturing companies: challenges and enablers towards increased digitalization. *Procedia Cirp*, 81:1113–1118.
- [60] Malm, A., Pollard, N., Brantingham, P., Tinsley, P., Plecas, D., Brantingham, P., Cohen, I., and Kinney, B. (2005). A 30 year analysis of police service delivery and costing: “e” division. *Centre for Criminal Justice Research: Abbotsford, BC*.
- [61] McCarthy, J. et al. (2007). What is artificial intelligence. Stanford University. Available at:<https://cse.unl.edu/~choueiry/S09-476-876/Documents/whatisai.pdf> (Accessed: 2024-03-11).
- [62] Meesters, M., Heck, P., and Serebrenik, A. (2022). What is an ai engineer? an empirical analysis of job ads in the netherlands. In *Proceedings of the 1st International Conference on AI Engineering: Software Engineering for AI*, pages 136–144.
- [63] Mergel, I., Dickinson, H., Stenvall, J., and Gasco, M. (2023). Implementing ai in the public sector. *Public Management Review*, pages 1–13.
- [64] Michel Ritskes, L. M. (2018). Werken onder architectuur. <https://www.politieacademie.nl/kennisonderzoek/kennis/mmediatheek/PDF/94891.pdf>.
- [65] Min, B., Ross, H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., Agirre, E., Heintz, I., and Roth, D. (2023). Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40.
- [66] Molenaar, S., Maas, L., Burriel, V., Dalpiaz, F., and Brinkkemper, S. (2020). Medical dialogue summarization for automated reporting in healthcare. In *Advanced Information Systems Engineering Workshops: CAiSE 2020 International Workshops, Grenoble, France, June 8–12, 2020, Proceedings 32*, pages 76–88. Springer.
- [67] Moskal, B. M. and Leydens, J. A. (2019). Scoring rubric development: Validity and reliability. *Practical assessment, research, and evaluation*, 7(1):10.
- [68] Nuseibeh, B. (2001). Weaving together requirements and architectures. *Computer*, 34(3):115–119.

- [69] Ogawa, R. T. and Malen, B. (1991). Towards rigor in reviews of multivocal literatures: Applying the exploratory case study method. *Review of educational research*, 61(3):265–286.
- [70] Overheid (2015a). Wetboek van strafvordering - artikel 152. [https://wetten.overheid.nl/BWBR0001903/2024-01-01/0#BoekTweede\\_TiteldeelII\\_AfdelingDerde\\_Artikel152](https://wetten.overheid.nl/BWBR0001903/2024-01-01/0#BoekTweede_TiteldeelII_AfdelingDerde_Artikel152).
- [71] Overheid (2015b). Wetboek van strafvordering - artikel 153. [https://wetten.overheid.nl/BWBR0001903/2024-01-01/0#BoekTweede\\_TiteldeelII\\_AfdelingDerde\\_Artikel153](https://wetten.overheid.nl/BWBR0001903/2024-01-01/0#BoekTweede_TiteldeelII_AfdelingDerde_Artikel153).
- [72] Overheid (2015c). Wetboek van strafvordering - artikel 29. [https://wetten.overheid.nl/BWBR0001903/2024-01-01/0#BoekEerste\\_TiteldeelIII\\_Artikel29](https://wetten.overheid.nl/BWBR0001903/2024-01-01/0#BoekEerste_TiteldeelIII_Artikel29).
- [73] Overheid (2015d). Wetboek van strafvordering - artikel 341. [https://wetten.overheid.nl/BWBR0001903/2024-01-01/0#BoekTweede\\_TiteldeelVI\\_AfdelingDerde\\_Artikel341](https://wetten.overheid.nl/BWBR0001903/2024-01-01/0#BoekTweede_TiteldeelVI_AfdelingDerde_Artikel341).
- [74] Overheid (2015e). Wetboek van strafvordering - artikel 342. [https://wetten.overheid.nl/BWBR0001903/2024-01-01/0#BoekTweede\\_TiteldeelVI\\_AfdelingDerde\\_Artikel342](https://wetten.overheid.nl/BWBR0001903/2024-01-01/0#BoekTweede_TiteldeelVI_AfdelingDerde_Artikel342).
- [75] Overheid (2022). Algoritmes Overheid. <https://algoritmes.overheid.nl/>.
- [76] Pandey, D., Suman, U., and Ramani, A. K. (2010). An effective requirement engineering process model for software development and requirements management. In *2010 International Conference on Advances in Recent Technologies in Communication and Computing*, pages 287–291. IEEE.
- [77] Pargaonkar, S. (2023). A comprehensive research analysis of software development life cycle (sdlc) agile & waterfall model advantages, disadvantages, and application suitability in software quality engineering. *International Journal of Scientific and Research Publications (IJSRP)*, 13(08).
- [78] Parviaainen, P., Tihinen, M., Kääriäinen, J., and Teppola, S. (2017). Tackling the digitalization challenge: how to benefit from digitalization in practice. *International journal of information systems and project management*, 5(1):63–77.
- [79] Peffers, K., Tuunanen, T., Rothenberger, M. A., and Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of management information systems*, 24(3):45–77.
- [80] Pennington, J. (2019). The eight phases of a devops pipeline - taptu - medium. *Medium*.

- [81] Pezoa, F., Reutter, J. L., Suarez, F., Ugarte, M., and Vrgoč, D. (2016). Foundations of json schema. In *Proceedings of the 25th international conference on World Wide Web*, pages 263–273.
- [82] Poenaru-Olaru, L., Cruz, L., Rellermeyer, J. S., and Van Deursen, A. (2023). Maintaining and monitoring aiops models against concept drift. In *2023 IEEE/ACM 2nd International Conference on AI Engineering–Software Engineering for AI (CAIN)*, pages 98–99. IEEE.
- [83] Poenaru-Olaru, L., Cruz, L., van Deursen, A., and Rellermeyer, J. S. (2022). Are concept drift detectors reliable alarming systems?-a comparative study. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 3364–3373. IEEE.
- [84] Politie, N. (2012). Bijlagen inrichtingsplan nationale politie. <https://www.politieacademie.nl/kennisonderzoek/kennis/midiatheek/PDF/89090.pdf>.
- [85] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- [86] Runeson, P., Host, M., Rainer, A., and Regnell, B. (2012). *Case study research in software engineering: Guidelines and examples*. John Wiley & Sons.
- [87] Sadovski, E., Aviv, I., and Hadar, I. (2024). Towards a comprehensive ontology for requirements engineering for ai-powered systems. In *International Working Conference on Requirements Engineering: Foundation for Software Quality*, pages 219–230. Springer.
- [88] Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., and Chadha, A. (2024). A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*.
- [89] Schoormann, T., Möller, F., Chandra Kruse, L., and Otto, B. (2024). Baustein—a design tool for configuring and representing design research. *Information Systems Journal*.
- [90] Senarath, U. S. (2021). Waterfall methodology, prototyping and agile development. *Tech. Rep.*, pages 1–16.
- [91] Shane, J. M. (2010). Organizational stressors and police performance. *Journal of criminal justice*, 38(4):807–818.
- [92] Shneiderman, B. (2020). Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered ai systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 10(4):1–31.

- [93] Spijkman, T., Molenaar, S., Dalpiaz, F., and Brinkkemper, S. (2021). Alignment and granularity of requirements and architecture in agile development: A functional perspective. *Information and Software Technology*, 133:106535.
- [94] Team, O. (2022). Chatgpt: Optimizing language models for dialogue.
- [95] Thesing, T., Feldmann, C., and Burchardt, M. (2021). Agile versus waterfall project management: decision model for selecting the appropriate approach to a project. *Procedia Computer Science*, 181:746–756.
- [96] Trieflinger, S., Münch, J., Bogazköy, E., Eissler, P., Schneider, J., and Roling, B. (2021). How to prioritize your product roadmap when everything feels important: A grey literature review. In *2021 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC)*, pages 1–9. IEEE.
- [97] Tsui, F., Karam, O., and Bernal, B. (2022). *Essentials of software engineering*. Jones & Bartlett Learning.
- [98] Tukur, M., Umar, S., and Hassine, J. (2021). Requirement engineering challenges: A systematic mapping study on the academic and the industrial perspective. *Arabian Journal for Science and Engineering*, 46:3723–3748.
- [99] van Bavel, M., van Sambeek, C., and ter Woerdts, S. (2023). Een kwestie van balans. <https://www.politie.nl/binaries/content/assets/politie/nieuws/2023/maart/trendbeeld-2023—digitale-versie.pdf>.
- [100] van Zandvoort, D., Wiersema, L., Huibers, T., van Dulmen, S., and Brinkkemper, S. (2023). Enhancing summarization performance through transformer-based prompt engineering in automated medical reporting. *arXiv preprint arXiv:2311.13274*.
- [101] Verner, J. M. and Abdullah, L. M. (2012). Exploratory case study research: Outsourced project failure. *Information and Software Technology*, 54(8):866–886.
- [102] Verticai (2024). Verticai. <https://verticai.nl/>.
- [103] Wang, P. (2019). On defining artificial intelligence. *Journal of Artificial General Intelligence*, 10(2):1–37.
- [104] White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., and Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.
- [105] Wiegers, K. E. and Beatty, J. (2013). *Software requirements*. Pearson Education.

- [106] Wieringa, R. J. (2014). *Design science methodology for information systems and software engineering*. Springer.
- [107] Wirtz, B. W., Weyerer, J. C., and Geyer, C. (2019). Artificial intelligence and the public sector—applications and challenges. *International Journal of Public Administration*, 42(7):596–615.
- [108] Yan, B., Liu, R., Kuo, D. E., Adithan, S., Reis, E. P., Kwak, S., Venugopal, V. K., O’Connell, C. P., Saenz, A., Rajpurkar, P., et al. (2023). Style-aware radiology report generation with radgraph and few-shot prompting. *arXiv preprint arXiv:2310.17811*.