Aleksi Pakkala

Evaluation of AI Tools Suitability for Academic Writing Purposes for Internal Use

Metropolia University of Applied Sciences

Master's Degree

Degree Programme in Business Informatics

Master's Thesis

25 May 2025

I have been lucky to work with such an inspiring and relevant topic as generative AI evaluation. I want to thank Metropolia Business School, and especially the AI team at MBS. This work would not have been possible without the AI teams eager support and feedback for this Thesis.

I would also like to thank everyone involved with the Master's degree programme in Business Informatics at Metropolia University of Applied Sciences for providing me with the education that enabled me to conduct this research. Specifically, I want to thank my supervisor Dr. Minna Liikanen for the feedback and unrelenting support throughout the Thesis process, and Zinaida Grabovskaia, PhL and Misa Bakajic for valuable feedback and guidance.

Finally, I want to thank my family. I am grateful to my partner for supporting me throughout this long Thesis process, aiding me through the highs and lows.

Aleksi Pakkala
Nurmijärvi
May 25, 2025

Abstract

Author:	Aleksi Pakkala
Title:	Evaluation of AI Tools Suitability for Academic Writing Purposes for Internal Use
Number of Pages:	87 pages + 7 appendices
Date:	25 May 2025
Degree:	Master of Engineering
Degree Programme:	Business Informatics
Instructors:	Minna Liikanen, Dr. Sc. (Econ. and Business Admin.), Senior Lecturer; Misa Bakajic, Senior Lecturer; Zinaida Grabovskaia, PhL, Senior Lecturer

The rise of generative AI has had a profound impact on higher education, and Metropolia Business School wanted to provide guidance on which AI tools can be used reliably in research and development work by students. The objective of this Thesis was to evaluate the suitability of selected AI tools for academic writing purposes in Metropolia Business School. To achieve this, an AI tool Evaluation Framework was developed.

This research employed an applied action research strategy. Data was collected in three rounds through interviews to support current state analysis, initial proposal development, and validation of the proposal. The current state analysis was conducted first to understand the existing use of generative AI at Metropolia and the existing guidance on generative AI use at Metropolia. The theoretical framework focused on covering key elements of generative AI, accuracy and reliability of generative AI, and evaluation of generative AI output.

Based on the current state analysis, best available knowledge, and input from stakeholders, an AI tool evaluation framework with explicit guidelines for implementation was developed. The evaluation of the suitability of selected AI tools for academic writing purposes was done by utilizing this AI tool evaluation framework. The AI tool evaluation framework was further refined based on feedback from key stakeholder and insights gained during the initial implementation of the framework.

This Thesis presents evaluation results of the suitability of selected AI tools for academic writing purposes by students for year 2025, and an AI tool evaluation framework that can be used to evaluate the suitability of AI tools for a variety of tasks and workflows also in the future. The AI tool evaluation framework and guidelines were developed to be used with individuals without extensive experience with AI, with reasonable amount of work for Metropolia Business School.

Keywords: Generative AI, LLM-as-a-Judge, AI evaluation, Suitability assessment

Contents

List of Figures

List of Tables

1 Introduction

Easy-to-use generative AI was made available for the mass market when OpenAI released ChatGPT to the public in November 2022. This kickstarted the generative AI technology trend, which remains the biggest trend in technology over the last two years (Yee *et al.*, 2024).

The release of ChatGPT and other services utilizing generative AI with large language models (LLM) has had profound effects on various business areas, such as marketing and software development (Singla *et al.*, 2025, p. 24). School systems at various levels have struggled with generative AI, as it has become very easy to generate text content about a chosen topic in a matter of seconds. The text content can be quite convincing and factually accurate, and identifying AI-generated text can be a significant challenge for educational systems (Sharkey, 2025).

This Thesis focuses on developing recommendations for selecting AI tools for research and development projects for the case organization. The Thesis was completed for Metropolia UAS (University of Applied Sciences), and the main deliverable is a framework for AI tools evaluation, along with recommendations on how this evaluation can be used also in the future.

1.1 Business Context

This study was conducted for Metropolia UAS, specifically for the Master's Degree Programme in Business Informatics, serving as the case organization. Metropolia UAS is the largest university of applied sciences (UAS) in Finland. Metropolia Ammattikorkeakoulu Oy was founded in 2007 by the cities of Helsinki, Espoo, Vantaa, and Kauniainen, following the merger of the schools EVTEK UAS and Stadia UAS to form Metropolia. There are approximately 18 000 students in Metropolia, and around 1100 staff members (Metropolia, 2025a).

The Master's degree Programme in Business Informatics is offered by the Metropolia Business School, one of the 10 schools in Metropolia. Metropolia Business School (MBS) has 1600 enrolled students and 60 personnel, and approximately 400 professionals graduate each year. Some of the programs in MBS are conducted in English, and the

school heavily focuses on a multicultural learning environment, as stated on Metropolia's websites.

> We offer a wide range of courses in English, a multicultural learning and working environment and an extensive international partnership network. Our multicultural atmosphere is reinforced by our international degree students and numerous exchange students from around the world, as well as by our own multinational staff (Metropolia, 2025b).

Metropolia, along with other UASs in Finland, aims to equip its students with the best possible skills to succeed and excel in the workforce. Artificial Intelligence has come to stay and knowing how to utilize AI tools efficiently and responsibly, will become a mandatory skill for a variety of fields in the near future. The McKinsey & Company report "Superagency in the Workplace" indicates that '46 percent of leaders identify skill gaps in their workforces as a significant barrier to AI adoption.' (Mayer *et al.*, 2025, p. 39)

Metropolia needs to teach students how AI can be used responsibly. A core part is understanding what tools can be used for which tasks to get reliable results. Unethical exploiting of AI tools with course and Thesis work is already common in higher education. A survey done already in 2023 by KPMG in Canada revealed that almost 70 percent of students participating in the survey at least sometimes presented AI-generated content as their own (KPMG Canada, 2023).

1.2 Business Challenge, Objective, and Outcome

At its best, generative AI can help students learn complex topics with efficiency, offering more personalized teaching and even automating mundane tasks (Slimi, 2023, p. 1). Sadly, there are at least as many downsides to generative AI. Generative AI makes cheating and plagiarism easier for students than ever before, which can lead at its worst to the devaluation of degrees (Cotton, Cotton and Shipway, 2024, p. 3). Even though these AI tools can be problematic in education, utilizing AI is an important skill now, and especially in the future (Mayer *et al.*, 2025, p. 40). For Metropolia UAS, to ensure a high quality of education now and in the future, extensive knowledge of AI and available AI tools by staff and students is important. Students want to gain the skills needed in future work life, and Metropolia has a need to validate the students' learning.

Generative AI has the potential to be the next transformative leap in technology, increasing personal productivity and creativity (Mayer *et al.*, 2025, p. 6). Besides benefits

for the work life, there is potential with generative AI in the field of education as well. Personalized teaching can boost student learning (Vieriu and Petrea, 2025, p. 5), and teachers can offload administrative workload on generative AI by automating repetitive tasks, leaving more time to focus on education (Cardona, Rodríguez and Ishmael, 2023, p. 28).

One of the key problems with AI tools in education is that they make cheating and plagiarism easy for students (Bittle and El-Gayar, 2025, p. 6). Cotton et al. (Cotton, Cotton and Shipway, 2024, p. 4) states “one of the most effective ways to prevent plagiarism is to educate students on what it is and why it is wrong”. Metropolia, like any other school, needs tools to minimize the negative effects of AI. Importantly, it has already acted on this front, as a new course “Ethical practices of the Thesis work” will become mandatory for all students doing Thesis work.

The generative AI field is rapidly growing, and the selection of AI tools for research and development is growing each week. Already in the summer of 2024, there were over 2000 different generative AI tools available (ARTSMART AI, 2024). Metropolia wants to identify reliable AI tools, that could be utilized in different stages of research and development projects. To aid in this work, this Thesis will focus on evaluating which AI tools are suitable to be used in which academic writing related tasks and workflows.

The objective of this Thesis is *to evaluate a selected set of AI tools to determine their suitability for academic writing purposes by students.* To perform the evaluation, an AI tool evaluation framework needs to be developed. The outcome of this Thesis is *the AI tool evaluation framework with explicit guidelines for implementation, and the evaluation results for year 2025* about the suitability of AI tools for academic writing purposes by students for internal use.

1.3 Thesis Outline

This Thesis work is done especially for the Metropolia Business School (MBS), more specifically for the Master´s Degree Programme in Business Informatics as the case organization, even though the findings can benefit other schools as well. Therefore, the recommendations provided in this Thesis will also consider the existing AI usage guidelines set by Metropolia. As this school operates in Finland, the evaluation will be compliant with local laws.

This Thesis is based on seven sections. Section 1 introduces the Thesis topic and its objectives. Section 2 describes the method and materials used in conducting the research for this Thesis. Section 3, Current State Analysis, explores how AI tools are currently used at MBS and the existing guidance on AI tool use at MBS. Section 4 covers the literature review, where the focus concepts are “Elements of Generative AI”, “Accuracy and Reliability of Generative AI”, and “Evaluating Output of Generative AI tools”. A conceptual framework is developed based on the literature review. Section 5 focuses on the initial proposal building, which is based on the current state analysis, the conceptual framework, and interviews with the MBS lecturers. The outcome of Section 5 is an AI tool evaluation framework with explicit guidelines for implementation. In Section 6, the AI tool evaluation framework is implemented, and the evaluation results are presented. Further refinements are made to the AI tool evaluation framework based on feedback from key stakeholder and insights gained during implementation. The Thesis concludes with Section 7, with the conclusion and suggestions for next steps with AI tool evaluations.

1.4 Key Concepts

In this Thesis, the term “AI tool” is used to describe a general-purpose, generative AI application that utilizes a specific Large Language Model. For example, in the context of this Thesis, OpenAI ChatGPT GPT-4o, OpenAI ChatGPT o3, and Google Gemini 2.5 Pro (preview) are examples of three different AI tools. If we break down the name of one of these AI tools, “OpenAI ChatGPT GPT-4o”, we can identify its components. “OpenAI” is the developer of the AI tool, “ChatGPT” is the AI application, and “GPT-4o” is the Large Language Model that powers this AI tool. Together, these components form the AI tool entity.

2 Method and Material

This section looks at the selected research approach, research design, and data collection and analysis methods used in this Thesis. The first part of this section covers different common research approaches and my thought process on how I selected a research approach for this Thesis. The second part delves into the research design of this Thesis. The third part focuses on data collection and analysis methods, explaining how the data used in this Thesis was gathered and analysed in a reliable way.

2.1 Research Approach

By its nature, research aims to create new knowledge or expand existing knowledge. This knowledge is created by solving a research problem, by using research methods. Research approach is something that covers the broad approach to a problem, it does not explicitly state how exactly the research problem is going to be solved (Kananen, 2013).

All research has a specific purpose, and research work can be categorized into the following kinds of studies: exploratory, descriptive, explanatory, evaluative, or mixed studies. *Exploratory* studies are focused on exploring and clarifying understanding of an issue, problem, or phenomenon. *Descriptive* studies aim to accurately describe events, persons, or situations. *Explanatory* studies research relations between variables in a situation or a problem. *Evaluative* studies assess how well certain solutions work for certain problems, or how well a solution works. *Mixed* studies utilize elements from different studies (Saunders, Lewis and Thornhill, 2023, pp. 179–181).

Research can also be divided into big *research families*, such as basic and applied research. *Basic* research aims to generate new knowledge just for the sake of new knowledge, whereas *applied* research aims to solve real-world problems through research (Bentley, Gulbrandsen and Kyvik, 2015, p. 690).

From a data collection and analysis perspective, research can be divided into qualitative research and quantitative research. *Qualitative* research answers open questions and is exploratory by its nature. *Quantitative* research on the other hand focuses on measurable variables and metrics. All quantitative research is, however, based on previously done

qualitative research, and it's quite common to validate and further investigate the theories created in qualitative research by using quantitative research (Kananen, 2013, pp. 27–28). Yet, qualitative and quantitative research are not opposing forces, but they answer different kinds of questions, with different kinds of data. Qualitative research depends on unstructured data, originating as an example from interviews. Quantitative research requires the data with which calculations can be done, for example, sales numbers and structured survey responses. Quantitative research tests theories or models of a phenomenon that were produced by previously conducted qualitative research (Kananen, 2013, pp. 31–36).

In many cases with applied research, the *research strategy* cannot be strictly classified using only qualitative or quantitative research methods. Case research, action research, and design research (or one of its novel, smaller types called Applied action research) are the forms of research strategies with multiple techniques or methods. Case research closely examines a case or phenomenon, and it incorporates elements from both qualitative and quantitative research. In action research, the researcher is an active participant rather than an external observer. Applied action research, which shares similarities with both action research and design research, has a tighter focus on the practical results, instead of researching the change process (Kananen, 2013, pp. 45–48).

This Thesis represents applied research, focusing on providing recommendations to solve a real-world issue of how students can utilize generative AI reliably for academic writing tasks. In this Thesis, the applied action research strategy was selected. The Thesis researcher is also a student at Metropolia, acting as both an internal and external participant in this study. The Thesis researcher would see himself as more of an active actor in this research if he were a lecturer or staff member at Metropolia. However, as he is not, this affects his perspective on the topic, as the recommendations he provides as an outcome of this Thesis are aimed at lecturers and staff.

In this Thesis, there is no cyclic approach to the problem, but a one-iteration-based research design. A first version of the AI tool evaluation framework is created based on (a) the best currently available knowledge, (b) the current state analysis results, and (c) inputs from the stakeholders; and then it is refined in the final validation stage, based feedback from a key stakeholder interview and insights gained during the implementation of the initial proposal.

2.2 Research Design

In this section, the research design of this Thesis is presented, which explains how the Thesis is split and how the collected data supports different stages of research. Figure 1 shows the research design of this Thesis.

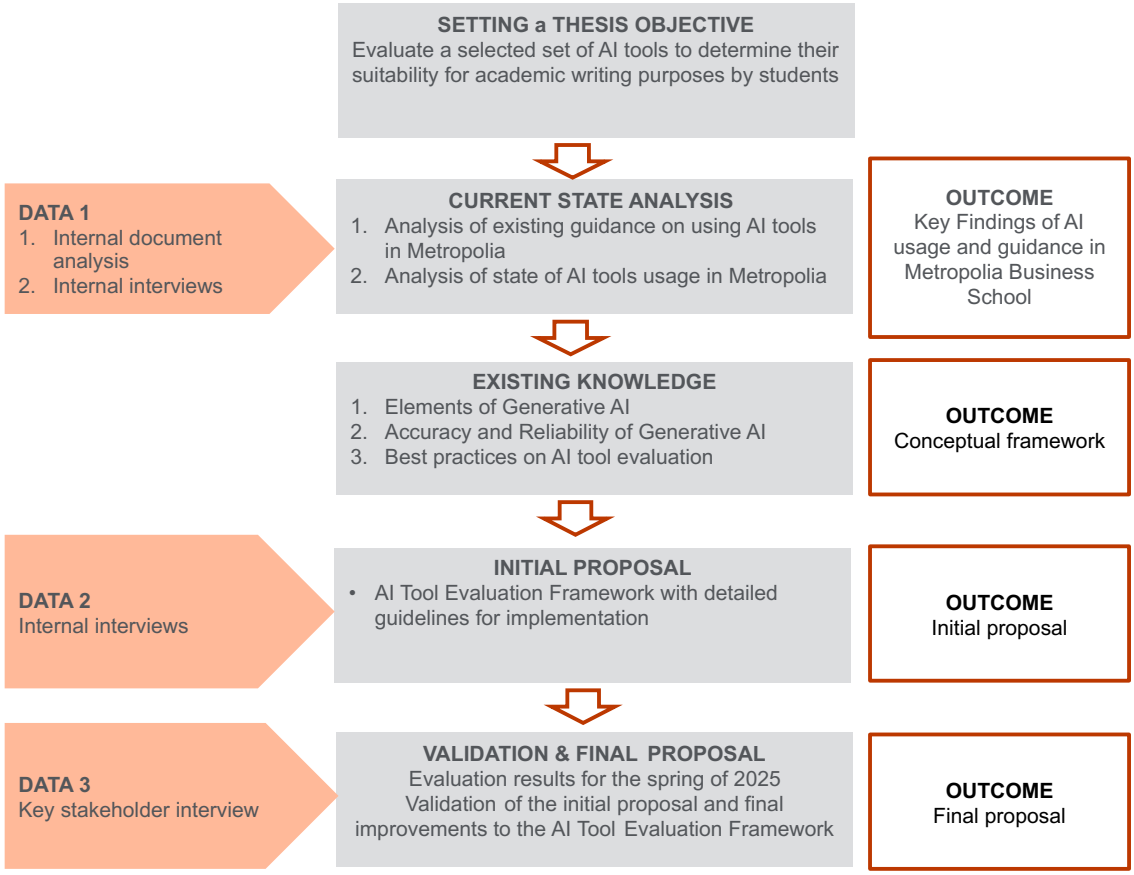


Figure 1: Thesis research design of this Thesis.

Figure 1 shows that the research process is split into five stages. The first stage is setting a Thesis objective. The second stage is the current state analysis, which is based on internal document analysis and the utilization of qualitative data from lecturer interviews. The outcome of the current state analysis is the key findings with the highest relevance for this Thesis.

The third stage in the Thesis process involves collecting existing knowledge and best practices relevant to the topic and formulating a conceptual framework for the Thesis. In the fourth stage, an initial proposal is built based on the previous steps and further interviews with the MBS lecturers. The outcome is an initial proposal for an AI tool

evaluation framework, which can be applied to assess the suitability of AI tools in MBS's specific workflows related to academic writing.

After the initial proposal is created, the AI tool evaluation framework is applied in the MBS context, and evaluation results are collected for the spring 2025. The AI tool evaluation framework and the results are reviewed with a key stakeholder from MBS, and the final proposal for the AI tool evaluation framework is built based on the feedback from the key stakeholder and insights gained during the implementation of the AI tool evaluation framework.

2.3 Data Collection and Analysis

The data used in this thesis is qualitative and consists of interviews and internal document analysis. This Thesis utilizes data from multiple sources, which were collected in multiple data collection rounds. Table 1 describes the various data sources used in this Thesis.

Table 1: Details of data collected in different stages of this research.

	Participants / role	**Data type**	**Topic, description**	**Date, length**	**Documented as**
	Data 1, for the Current state analysis (Section 3)				
1	Respondent 1: Metropolia business school personnel	Microsoft Teams Meeting	The state of AI usage and regulation at Metropolia UAS	May 2024, 1 hour	Recording
2	Respondent 2: Metropolia business school personnel	Microsoft Teams Meeting	The state of AI usage and regulation at Metropolia UAS	December 2024, 1 hour	Recording
3	Respondent 3: Metropolia business school personnel	Microsoft Teams Meeting	The state of AI usage and regulation at Metropolia UAS	December 2024, 1 hour	Recording
4	Respondent 4: Metropolia business school personnel	Microsoft Teams Meeting	The state of AI usage and regulation at Metropolia UAS	December 2024, 1 hour	Recording
5	Respondent 5: Metropolia business school personnel	Microsoft Teams Meeting	The state of AI usage and regulation at Metropolia UAS	January 2025, 1 hour	Recording

	Data 2, for Proposal building (Section 5)				
2	Respondent 6: Metropolia business school personnel	Microsoft Teams Meeting	Proposal building	March 2025, 1 hour	Recording
3	Respondent 7: Metropolia business school personnel	Microsoft Teams Meeting	Proposal building	March 2025, 1 hour	Recording
	Respondent 8: Metropolia personnel	Microsoft Teams Meeting	Proposal building	March 2025, 1 hour	Recording
	Respondent 9: Metropolia business school personnel	Microsoft Teams Meeting	Proposal building	March 2025, 1 hour	Recording
	Data 3, from Validation (Section 6)				
6	Metropolia business school key stakeholder	Microsoft Teams Meeting	Final proposal building	May 2025, 2 hours	Recording

In the first data collection round, data were collected for the current state analysis through personal interviews. The goal of the current state analysis was to understand how lecturers perceive the usage of AI tools in Metropolia UAS. The interviews focus on the lecturer´s experiences and challenges with using AI tools in research and development work. The interviewees were selected from the MBS AI team. The members of the AI team are lecturers with an interest in AI. All the interviews were conducted in a semi-structured format. For the first data collection round of interviews, four main themes were prepared that were to be discussed in each interview.

The second data collection round relies on stakeholders' inputs, but here the focus is on building the initial proposal. The interviews focused on the lecturers' needs and expectations regarding the Ai tool evaluation framework, as well as their views on the practical application of these proposed documents. Almost all the interviewees for the second data collection round were members of the MBS AI team, but this time, there were also AI experts from other parts of Metropolia.

The final data collection round involves validating the initial proposal through expert judgment. The initial proposal will be presented in an interview with key stakeholders from MBS for the final validation. The findings will be used to refine the initial proposal into the final proposal.

Thus, interviews are the form of data collection for this Thesis. All interviews in this thesis were conducted over Microsoft Teams, recorded, and later transcribed. For example, in total, five interviews were conducted for the current state analysis, each lasting approximately 45 minutes. The transcriptions of the interviews contained approximately 5500 words of text each and about 950 lines of text. The total content of transcribed texts was approximately 28,000 words. If we trim the time codes and line changes from the transcription files, the resulting text would be roughly 70-80 pages, formatted in Arial font, size 11pt, and line height 1.5.

All the interviews were semi-structured thematic interviews, and data collected from these were qualitative. The first of the interviews was conducted in the spring of 2024, and the later interviews were done at the end of 2024 and the beginning of 2025. The interviews were recorded using Microsoft Teams, and transcriptions were created with OpenAI Whisper. Whisper was run locally on my computer to ensure data safety and compliance, so no data was shared with OpenAI or any other third party. With this data, a thematic analysis was conducted by following an explanation of thematic analysis (Saunders et al., 2023).

The first step of the thematic analysis was to become more familiar with the collected data. Interview transcripts were checked to become field notes. After this, the data was “coded” using labels. The data set consisted of several transcribed interviews, so this process was quite straightforward. After coding my data, the common themes were identified by grouping the related codes together. Once the themes were identified, the most relevant themes were chosen for the analysis. Some of the identified codes that were not part of the most relevant themes still offered insights and were discussed in the text.

3 Current State Analysis of AI Tools Usage in Metropolia Business School

This section will showcase the results of the current state analysis of using Generative AI tools in Metropolia UAS. The first sub-section contains an overview of the current state analysis and covers the methodologies and techniques used to conduct this current state analysis.

The second sub-section is focused on the state of generative AI usage in MBS. This sub-section will cover topics such as how Metropolia supports the use of AI tools, what challenges lecturers face with using generative AI, and how generative AI already helps lecturers and other personnel do their jobs. This sub-section also explores how lecturers see generative AI already affecting the performance of students, what kind of trouble areas lecturers see with students using generative AI, and what kind of systems are already in place to mitigate these issues.

The third sub-section covers the strengths and weaknesses of the current state of generative AI usage in MBS. The fourth section then explains the key themes and the current state analysis is wrapped up with the fifth and final sub-section covering the Key Focus Areas.

3.1 Overview of the Current State Analysis

This current state analysis is based on five interviews with MBS personnel and publicly available documents on AI tools usage in Metropolia and higher education in Finland. This current state analysis aims to provide a deeper understanding and insights into the usage of AI tools in MBS, and what kind of existing guidelines are in place for AI tool use in MBS.

This current state analysis consisted of three phases. In the first phase, knowledge about generative AI usage in Metropolia was gathered and analysed from publicly available online materials, such as Metropolia strategy and available courses. This first step gave a preconception of the current state of generative AI usage in MBS, but the main purpose of this first step was to gain a bit of knowledge about the subject.

The second step was thematic interviews with selected lecturers. For this Thesis, several lecturers from MBS were interviewed who had some experience with generative AI. All the interviewed lecturers were part of MBS AI team. For this end, an explorative type of interview was needed. The chosen people have vastly different experiences with generative AI, so a structured interview wouldn't have been a good fit for this analysis.

The third step of the current state analysis was performing a thematic analysis of the data collected in interviews, and identifying the results of the current state analysis based on that thematic analysis.

3.2 Description of AI Guidance in MBS

Metropolia guides the use of AI in education for both students and personnel. The guidance focuses on responsible, ethical, and transparent use of AI tools (Koirikivi *et al.,* 2024, p. 1). Although Metropolia provides general guidelines for using AI tools that apply to all schools in Metropolia, still, degree programs can have more specific guidelines, as is the case with Metropolia Business School. The essential thing is to follow good study practices and report the use of AI in accordance with the guidance (Koirikivi *et al.,* 2024, p. 5).

For students, the Metropolia general guidelines for AI tools usage suggest that AI should be used as a support tool. Some examples of use cases provided in those guidelines include ideation with AI, using the AI as a tutor and sparring companion. AI can also be used to correct, summarize, and translate the text that a student has produced. Here the MBS specific guidelines differ quite from the general Metropolia guidelines, as the written submissions are checked for AI usage by Turnitin, as stated in the MBS guidelines.

> When a written submission includes checking for plagiarism, the detection tool that is currently used by Metropolia is Turnitin.com (note, only available in English). It also includes detection of AI-generated text. When using it, it would be important to have the acceptable levels set by the course instructor in advance (for example, for plagiarism it is typically below 25% for copy-pasting, which still includes the quality check for specific violations)." (Metropolia Business School, 2023, p. 2)

These guidelines about using Turnitin in AI detection mean that generative AI cannot be used in MBS to correct, summarize, or translate the text that a student has produced. This means that the written assignments shouldn't contain text generated by AI at all.

The general guidelines for AI usage in Metropolia, that were published one year later, state the following:

> "AI detection tools can be used cautiously as one tool among others when reviewing English language works. However, it should be noted that AI detection tools can give false results and can be deceived. Therefore, percentage scores from detection tools should not be given threshold values, and their results should always be interpreted as part of the overall assessment." (Koirikivi *et al.,* 2024, p. 6)

Thus, in this particular case, the guidelines provide somewhat contradictory guidance on the use of AI detection programs. It may be due to the fact that in 2023 the all-school guidelines were not yet produced, and MBS pioneered this area at Metropolia. By now, the AI guidelines for MBS, released in autumn 2023, feel somewhat outdated and are not in sync with the latest guidelines from Arene and the general Metropolia guidance (2024). Another example is that MBS guidance does not specifically mention that AI tools cannot help in research-based tasks, such as literature search (Metropolia Business School, 2023, p. 1), whereas the general guidance for AI tools usage stress the information retrieval as one of the main ways for use of AI in theses (Koirikivi *et al.,* 2024, p. 4).

This points to the need to update the MBS guidelines. Even though the MBS guidance on AI tool usage is based on recommendations from Arene and closely follows the "Guidelines for the use of AI in teaching at the University of Helsinki" (and some other leading universities), many documents have undergone significant updates since August 2023.

3.3 Analysis of Current Use of Generative AI in MBS

Generative AI as a concept in the context of this Thesis means tools and technologies that utilize modern transformer-based large language models to generate textual content, images, audio, or video. This sub-section covers how generative AI is currently being used in Metropolia both by lecturers and students, how artificial intelligence is present in the current Metropolia strategy, and how Metropolia supports the adoption of generative AI in education.

According to the lecturers interviewed, no official generative AI tools are procured for Metropolia (Interviewee 2, Interviewee 3, Interviewee 4, 2024). The lecturers interviewed

are all part of an internal group at MBS that focuses on driving AI usage forward in Metropolia. However, they were participating in a trial where they were given access to a paid version of ChatGPT (Interviewee 4, 2024).

Presently, Metropolia uses cloud-based services from Microsoft and Google as part of its IT infrastructure, but there are technical and data-protection reasons why Microsoft Copilot integrations haven't been taken into use with the Microsoft 365 products that Metropolia IT offers for personnel and students. Google Gemini features are not enabled for students or personnel either (Interviewee 4, 2024).

At the same time, Metropolia has noticed the need to offer studies about AI and generative AI. In their strategy, Metropolia states the following under the "Forerunning digital transformation" heading:

> Rapid technological development and digital transformation bring new opportunities, but at the same time challenge us to continuously renew our operations. Our goal is to be a forerunner in this transformation.
>
> We use data, artificial intelligence, and other digital solutions to support Metropolia's strategic management and decision-making (Metropolia, 2024).

At the time of writing, MBS does not offer studies directly related to artificial intelligence. But there are 3 artificial intelligence-related courses in the School of ICT offered (and their number is constantly growing); they are available for and aimed at industry experts (Master´s level).

3.3.1 Generative AI usage by MBS lecturers

Based on the interviews, there is no widespread usage of generative AI amongst the MBS personnel. The lecturers interviewed are all part of MBS AI team, so they are all quite familiar with the most popular generative AI tools, and almost all of them use generative AI tools daily. Everyone interviewed mentioned that they had used ChatGPT and Microsoft Copilot, and some of them had used Google Gemini, Elicit & Perplexity. All of the interviewed lecturers mentioned using ChatGPT the most.

Amongst the interviewed lecturers, the most common task where generative AI was used is the creation of structures for textual content. There were diverging opinions on producing the final text with generative AI. Interviewee 2 thought it was only natural to

translate textual content with generative AI and publish or use that content as part of their work. In contrast, Interviewee 4 didn't want to use anything created by generative AI in the final results of their work and only utilized generative AI for the ideation and planning phase.

The interviewed lecturers wish for improvements in the tooling that allows the identification of generative AI usage in the works produced by students (Interviewee 2, Interviewee 3, Interviewee 4, 2024).

> Then there's also the fact that many people are still hoping for these kinds of AI detection tools. Tools that would catch students using AI. But even among colleagues, there's already been a change, with some saying, "Well, there's nothing really wrong with (using generative AI responsibly)." That, of course, we now simply have to accept. (Interviewee 2, 2024)

Metropolia is organizing the training for on the usage of generative AI, but there have been varying levels of experience with generative AI. Some resistance was observed during the initial widespread adoption of ChatGPT, but the opinions on generative AI quickly shifted to a more open attitude towards generative AI (Interviewee 2, Interviewee 5, 2024, 2025).

3.3.2 Generative AI usage by MBS students

A common theme that was raised during the interviews with the lecturers about generative AI usage by MBS students was that it is surprisingly rare amongst the students. Interviewee 3 mentioned that in a course that they teach, only about 20% of students claimed to have any previous experience with generative AI tools when asked. This was in September 2024.

During 2024, there were a couple of separate cases where significant generative AI usage was identified in Thesis works produced in Metropolia. This was noticed during the Thesis work, well before the Thesis was released to the public or graded. Lecturers have also noticed that some students use generative AI to produce textual content for coursework. In some rare cases, most of the work was created by generative AI, which was discovered by detected with a routine check with turnitin.com done by some lecturers as a screen check for all submitted papers (Interviewee 1, Interviewee 4, 2024)

A threat that multiple interviewees identified was the fact that as the generative AI models are improving, the output of these models looks better and more convincing. The accuracy of the generated content might not increase at the same pace, however, and this situation might lead to students placing too much trust in the output of generative AI models.

Another issue identified by the lecturers was that the generative AI often produces text on a high level, using broad terms, and if the students use this output just for ideation, they might be led to produce texts themselves that don't delve into details enough. This was a pattern that Interviewee 4 identified in some of their courses (Interviewee 3, Interviewee 4, 2024).

> Yes, well, in my opinion, the downside is that, in some cases, I've noticed it steers students to think that these "consultant texts," these high-level descriptions—like, for example, talking about their organization's data environment in a general way ("we collect data widely, then bring it into centralized databases, where we use master data methods, etc.") actually amount to analysis, when they don't really mean anything. I suspect much of it comes from AI. Either it's text produced by an AI or based on AI-generated answers, and then people assume that this is an acceptable level of content. (Interviewee 4, 2024)

3.4 Key Themes of Generative AI Usage in MBS

Generative AI in its current form, as a tool utilized by hundreds of millions of people, is a new phenomenon, and this is reflected in the status of generative AI usage in MBS as well. Generative AI usage and knowledge are limited among both lecturers and students.

There are several challenges that arose from the interviews. The process of getting new tools to be approved to use in Metropolia is slow and difficult for the lecturers to handle. This then leads to the lecturers using generative AI tools that are not approved by Metropolia.

Another core theme is how Metropolia can verify in the future that the students have learned and gained the skills that the courses aim to give. For the future success of the school, it's critical that the degrees earned are valued in work life and academia. If the situation gets to the point where a student can get through the courses and gain a degree with minimal effort utilizing generative AI, the value of any degrees could be undermined.

3.4.1 AI’s Perceived Impact on Students

The most alarming theme for the lecturers was that generative AI is already affecting students' thinking. The interviewees demonstrated a concern that students may not understand the principles of how generative AI systems work, and therefore, have misconceptions about their reliability (Interviewee 1, Interviewee 4, 2024). Interviewees anticipated that the situation would only get worse as generative AI models become more widespread in the coming years, but with this current transformer-based architecture, they can never become fully reliable (Interviewee 1, Interviewee 4,2024).

Even in situations where generative AI is only used at the ideation stage, it’s already visible that the students' answers for tasks are becoming more homogenous, which indicates that the answers lack original thinking and analysis (Interviewee 4, 2024). Often, the answers generated by AI are way too high-level in their analysis, which can be misleading for students if they think that the answers AI gives are examples of good answers (Interviewee 3, 2024).

On the other hand, overall, perceived AI usage is low among students. For example, in one course for first-year students, only 20% reported having any experience with AI tools (Interviewee 3, 2024). These students had experience only with ChatGPT and Microsoft Copilot, and usage of other AI tools was reported to be surprisingly minimal.

3.4.2 Organizational Challenges

As the generative AI scene has moved quickly, it’s difficult for many organizations to keep up. However, even with this in mind, the interviewed personnel saw that the change processes with IT tools are too slow and complicated.

The main thing that the interviewees were struggling with was that they lacked access and licenses to new IT tools (Interviewee 3, Interviewee 4, 2024). The process of gaining access to new IT tools is perceived as too complex and requiring unrealistic effort; and they feel, there doesn’t seem to be any process for how new tools could be acquired fast (Interviewee 3, Interviewee 4, 2024). Based on the interviewees, as an example, the lecturers would be required to perform data privacy and data-protection analyses on these tools, and this is something that they often don’t have the time for, nor the skills, and the process of getting new IT tools easily gets stuck.

This leads to the lecture using generative AI tools that are not approved and managed by internal IT. The user is, of course, responsible for their actions, but this can lead to data protection and data privacy-related issues for the whole organization. There have been trials aimed at mitigating this issue. As an example, the AI team at MBS has been granted access to a paid version of ChatGPT.

3.4.3 Pedagogical Adjustments Concerning AI

In the interviews, there were mentions of students misusing generative AI both in coursework and theses. The cases of misuse with Thesis work were rare and isolated, but still, this proves that the misuse of generative AI is a real issue. (Interviewee 1, Interviewee 4, 2024).

Some teachers hope for generative AI detection tools, such as TurnItIn (which is an officially procured tool at the organization), to become more advanced and accurate in identifying generative AI in student work. The hope is that these tools will serve to deter students from using generative AI in unethical ways. The guidelines from Metropolia state the following:

> AI detection tools can be used cautiously as one tool among others when reviewing English language works. However, it should be noted that AI detection tools can give false results and can be deceived. Therefore, percentage scores from detection tools should not be given threshold values, and their results should always be interpreted as part of the overall assessment (Koirikivi et al., 2024, p. 6).

Some interviewees saw that traditional methods of validating student learning can still be effective, such as discussing the subject with the student. Interviewee 2 gave an example: As part of the project presentation, the interviewee likes to ask clarifying questions about the subject, and it's often painfully clear if the student is not familiar with the subject and the content has been produced by generative AI.

3.5 Key Findings from Current State Analysis

There were multiple findings of varying importance regarding the use of generative AI in MBS, but not all of them can be addressed within the scope of this Thesis work. Importantly, many findings related to the significance of *reliable* use of AI tools, which

points to the need to evaluate the reliability of these AI tools used, especially for the purposes of academic writing, such as coursework and theses.

The three key findings from the current state analysis within the scope of the Thesis were as follows: (a) the organization does not yet have AI tools to offer student & adopting new tools requires a time-consuming procurement process (that should by default be preceded by *a thorough evaluation process*), and (b) there is a lack of specific, detailed instructions for using AI tools for academic writing purposes in a reliable way.

As for the first challenges, the organization does not yet offer a general-use AI tool for students. There are some LLM-powered tools available for more specific tasks, such as mikro-Mikko, an AI chatbot assistant designed to help with IT-related questions (since 2025), and Keenious in the library (since 2025). There is also an AI-based AINI-chatbot for screening the well-being and thesis problems among students (since 2025) and a few smaller traditional chatbots, such as for Admissions and Library customers (non-AI based). Yet, multiple higher education facilities in Finland already offer general-purpose AI tools for their students. For example, Aalto University offers an “Aalto AI Assistant” to its students, and the University of Helsinki provides a “CurreChat” service to its students. Offering a unified solution as an AI tool in education ensures that students have equal access to high-quality AI resources. If the school doesn’t provide a general-purpose AI tool for its students, the only option for students is to use commercial AI tools, which raises the risk for data privacy violations.

Moreover, the slow IT tool procurement process hinders the agile experimentation of generative AI tools. This issue is already being somewhat mitigated with the work of MBS AI team, but it is still an issue for the wider adaptation of AI tools. The slow IT tool procurement process leads to the same issues as with the students: using non-compliant generative AI tools increase the likelihood of problems, from data breaches up to various misuse.

Secondly, the guidelines offered by MBS for students are presented at a high level and do not provide exact details on how students could utilize AI tools in a more ethical and reliable way. This shortcoming relates to the slow IT tool selection process discussed above.

These key findings point to the need for MBS to start a systematic process for *evaluating the suitability of AI tools for specific tasks,* but especially for academic writing. There is a clear need for a framework that would enable the evaluation of selected AI tools, in order to be better equipped when deciding on the tool's suitability for students, and also for procuring them.

4 Available Knowledge and Best Practice on Evaluating Generative AI Output in Higher Education Context

This section examines the currently available knowledge on the key concepts of evaluation of generative AI tools and their output. The first subsection focuses on generative AI, briefly touching on history of AI, what generative AI is, its differences from other forms of artificial intelligence, and the components required for generative AI solutions. The second subsection discusses the reliability and accuracy of generative AI. This section covers potential issues with generative AI, what it means for a generative AI service to be accurate, and how different services or models can be compared with each other. The third subsection is about evaluating generative AI output. This subsection discusses different methods and metrics for generative AI evaluation.

4.1 Generative Artificial Intelligence

Modern generative AI is a highly complex topic. To understand Generative Artificial Intelligence, it is best to first discuss Artificial Intelligence. Artificial Intelligence (AI) is a broad umbrella term, and under it, there are numerous fields of science, like machine learning, probability mathematics, and natural language processing, to mention a few (Morandín-Ahuerma, 2022, p. 1). On a high level, artificial intelligence is the ability to perform tasks on computers that mimic human intelligence (Russell, 2021, p. 18). These tasks can vary from simple ones, like sorting algorithms, to more complicated ones, like image generation. One popular method of defining AI is the Turing test, proposed by Alan Turing in 1950. A computer passes the test if a human operator cannot identify whether the answers are received from a person or a computer. To succeed in this, the computer needs four capabilities: natural language processing, knowledge representation, automated reasoning, and machine learning. These capabilities are at the core of generative AI as well (Russell, 2021, pp. 19–20). In the following subsections, we will look at the capabilities most relevant for generative artificial intelligence in greater detail.

The current artificial intelligence “hype” revolves around generative artificial intelligence. Generative AI systems can produce text, images, and even music. Generative AI systems can also learn natural language, programming languages, fields of science, or almost any complex topic (Amazon Web Services, 2024c).

The introduction of ChatGPT sparked the discussion about the possibility of artificial general intelligence (AGI), but we are still far away from achieving that. Artificial general intelligence is a term used for AI that can learn and apply knowledge across domains and perform any intellectual task a human can do (Amazon Web Services, 2024b). This is still in the realm of science fiction, and our solutions now are referred to as narrow AI (Amazon Web Services, 2024b). Narrow AI, as the name suggests, excels at solving specific domain problems, like playing chess or interpreting if an email is spam or not. Even complex tasks utilizing AI, like self-driving cars, are still considered narrow AI, even though they use multiple features that use different kinds of AI solutions, like traffic-aware navigation, computer vision, and decision-making. (Amazon Web Services, 2024b)

4.1.1 Transformer Models

Modern generative AI is enabled by the relatively recent advances in transformer models, which allow natural language inputs and natural language outputs. Transformer is a deep-learning architecture, first introduced by the “Attention Is All You Need” research paper from 2017, created by researchers from Google. The breakthrough in this paper was the introduction of self-attention mechanisms. Self-attention mechanisms allow the model to define the importance of words in a sentence, taking the sentence context into consideration. In addition to understanding the context of a sentence, transformer models understand the relationship between sentence components. This is a simplification, and to be more exact, the model weights the importance and relationship of tokens (words) in a sequence (sentence) (Vaswani *et al.*, 2023, p. 10; Amazon Web Services, 2024a).

Before transformer models, Recurrent Neural Networks (RNN) are used to understand dependencies and relationships in text content, very similarly what transformers do. RNNs were used by services like Google Assistant for the speech recognition (Beaufays, 2024). The use cases and features of Recurrent Neural Networks and Transformer Models are similar, but the main difference is that RNNs process data sequentially, while transformer models process data using parallelization. Utilizing parallelization makes transformers much faster and more efficient compared to RNNs, which enables shorter training times and the ability to handle much longer sequences. With the self-attention mechanism, transformer models can take into account the whole data of a sequence simultaneously, compare to RNNs where the data handling process is cyclic. RNNs

process the input sequence one token per time, storing the processed data in a hidden state vector which is updated on each cycle (Amazon Web Services, 2024a).

Popular models built on the transformer model architecture include BERT, developed by Google, GPT-n, developed by OpenAI, and DALL-E, developed by OpenAI as well. BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer) are transformer models that focus on textual input and output but have different focus areas. BERT excels in understanding natural language (Devlin *et al.*, 2019, p. 9), and GPT focuses more on text generation (Brown *et al.*, 2020, p. 34). DALL-E, on the other hand, is a transformer architecture-based model built for text-to-image generation, enabling services to generate images based on textual input (OpenAI, 2024a).

4.1.2 Large Language Models

The terms “Transformer Model” and “Large Language Model” are closely related, but they describe different concepts. The term “Transformer Model” describes the neural network architecture, and large language models are applications that use the transformer architecture. Large Language Models are deep learning models that have been trained with very large amounts of data and utilize transformer architecture (Amazon Web Services, 2025). The primary purpose of large language models is to understand and generate human-like text (Zhao *et al.*, 2024, p. 56).

The best-known LLM applications are generative AI applications with chat interfaces, like OpenAI ChatGPT, Perplexity, or Google Gemini. Large Language Models can be used for more narrow tasks as well, such as text translation, code autocompletion, or text classification (Amazon Web Services, 2025). A key benefit of transformer-based LLMs is the model's ability to self-learn. This ability is what enables the models to learn basic grammar, languages, and knowledge. Because of this, one model can perform various kinds of more complex tasks, like question answering and document summarization, to name a few (Amazon Web Services, 2025).

Large Language Models are very big and require vast amounts of data. When discussing model size, we refer to the number of tunable parameters the model has. Large Language Models have multiple kinds of parameters, mostly weights and biases. Parameters are tunable variables learned in the model training process, and they define

the model's ability to produce accurate output based on the given input (Bronsdon, 2025). Generally, models with more parameters perform better (Kaplan *et al.*, 2020, p. 8), but the latest advancements in large language model architecture find alternative ways to improve model performance besides increasing the parameter count (Wang, Li and Xu, 2025, p. 4).

With modern large language models, the parameter sizes are counted in hundreds of billions. Most companies no longer release their exact parameter counts, but the OpenAI GPT-3 model already had 175 billion parameters in 2020 (Brown *et al.*, 2020, p. 1). There isn't publicly available data about the recent models from OpenAI like the o1-family, or Google's Gemini models, but the recent R1 model from DeepSeek that caused quite a buzz in the generative AI scene has 671 billion parameters, with 37 billion activated parameters (Hugging Face, 2025). From this, we can conclude that modern large language models can have hundreds of billions of parameters.

With the latest advancements in generative AI, there are now large language models with over a trillion parameters. These models utilize AI sparsity, a mixture-of-experts architecture. These models might have hundreds of billions, or even trillions of parameters, but only a much smaller subset of those parameters is used at any given time (Merritt, 2022). Examples of such models are the DeepSeek R1 (Hugging Face, 2025), and the Switch-C Transformer by Google (Fedus, Zoph and Shazeer, 2022, p. 1)

Huge models are not always the solution, as large models also have downsides (L. Zhou *et al.*, 2024, p. 1). With a bigger model size comes increased computational costs for training and using the model in an application. With current large language model architectures, the computational workload and memory requirements grow linearly with the number of parameters (Fernandez *et al.*, 2024, p. 1). This is why smaller models could be used for more mundane tasks, such as translation, or in use cases with huge scale, for example, as phone AI companions, such as Apple Siri (Parthasarathy *et al.*, 2024, p. 69).

There are multiple methods of reducing the required computational workload for both AI training and model usage in an application. One of these methods is fine-tuning a model for a more specific task. Fine-tuning is the process of taking a pre-trained model like GPT-4 from OpenAI and training it with a smaller, domain-specific dataset. Not only does

fine-tuning reduce the computation costs, it also improves the domain-specific performance (Parthasarathy *et al.*, 2024, p. 70).

It's important to differentiate the term “model size” from training dataset size. As mentioned previously, model size just refers to the number of parameters a model has. Dataset size, however, describes the data the model was trained on. Dataset sizes are measured in tokens, where the token is a piece of text, most commonly a single word (Sharma, 2025).

4.1.3 Overview of Generative AI Application Architecture

As almost all software, generative AI services consist of multiple software layers (Bitloops, 2025). There’s the client frontend, often a web or mobile application, which handles the logic for the user interface (Bitloops, 2025; ELITEX, 2025a). The frontend client communicates with an API (Application Programming Interface) or a collection of APIs, often referred to as the backend (ELITEX, 2025b). The next layer in a generative AI system is the model layer, which houses the large language model (Epical, 2024; Bucaioni *et al.*, 2025, p. 3). The model layer receives input from the API layer (Epical, 2024; Bucaioni *et al.*, 2025, p. 3). There are more layers to generative AI services (Epical, 2024), but focusing on these three is sufficient for this Thesis context.

When consumers use services like ChatGPT, they communicate with the application layer (Bucaioni *et al.*, 2025, p. 1). Their prompts are sent in an HTTP request to the API layer with relevant context (Bucaioni *et al.*, 2025, p. 1). The API layer handles multiple things, like request validation, rate limiting, authentication, and analytics, to name a few (Bhavandla, 2025, pp. 1–3; Bucaioni *et al.*, 2025, p. 1). With generative AI systems, the API layer can also handle things like censoring the large language model output (Bucaioni *et al.*, 2025, p. 3), to ensure security and compliance. The internal logic of an API is not visible to the end users (Bratslavsky, 2025). After processing, the API layer sends the user input to the Model layer and receives the LLM output from the model layer (Bucaioni *et al.*, 2025, p. 2)

Companies offering Generative AI services, such as OpenAI, Google, and Anthropic, to name a few, usually provide an LLM API that can be integrated into another piece of software, such as a chatbot (Bucaioni *et al.*, 2025, p. 1). These APIs can differ from one another from a software architectural perspective. For example, some Google Gemini

APIs handle the web search-based retrieval augmented generation internally (Google AI, 2025c), and with OpenAI APIs, the web-based retrieval augmented generation needs to be configured explicitly (OpenAI, 2025f). Anthropic Claude APIs didn't include Search functionality until recently (Antrophic, 2025). From the evaluation perspective, this can cause issues. Suppose we try to measure the performance differences of LLM APIs with differences in the web search architecture. In that case, an LLM API with search functionality can offer superior performance in tasks that benefit from web-based retrieval augmented generation (Lakatos *et al.*, 2024, p. 1). This can cause misleading results that don't correctly portray the capabilities of the LLM APIs.

When evaluating the performance of generative AI solutions, communication happens with the application layer, the API layer, or, in some cases, the Model layer (Epical, 2024; Yucong, Zhendong and Fuliang, 2025, p. 1). It's crucial to understand what part we want to measure. The application layer is the easiest to start the evaluation on, compared to testing the API layer. Assessment through the API layer requires signing up with a developer account, and making requests to the API is a bit more involved than using the web interfaces (OpenAI, 2025c). Evaluating the model layer is rarely possible for external evaluators (Yucong, Zhendong and Fuliang, 2025, p. 1). This evaluation methodology is possible only with large language models published publicly, such as models from DeepSeek (Yucong, Zhendong and Fuliang, 2025, p. 1).

Whichever layer of generative AI architecture is used for evaluation, the choice should be intentional and based on the evaluation needs. For example, If the evaluation purpose is to gain understanding on which LLM API would be the optimal choice for certain tasks, the evaluation should be conducted utilizing the LLM API layer.

4.2 Accuracy and Reliability of Generative AI

ChatGPT, developed by OpenAI, brought an easily accessible, easy-to-use user interface for the masses, and almost overnight, millions of people had access to generative AI utilizing large language models (Chow, 2023). There has been much discussion that ChatGPT cannot be considered reliable, as generative AI models are prone to "hallucination," meaning that the text they generate contains false information (University of Arizona Libraries, 2024; Mayer *et al.*, 2025, p. 44).

When discussing the reliability of generative AI tools, it is important to understand, at least on a high level, the process of how generative AI generates textual output. For textual content, generative AI solutions use Large Language Models to analyse the input and generate the output. As an example, when given a prompt, “Write me one paragraph of text about best practices of web development with the React Framework, but use sources only written by Dan Abramov and mark the citations to the text”, one would think that the model under the hood would start searching for blog posts written by Dan Abramov, and then giving answers based on those, that is not at all what is happening. The model will analyze the text and try to understand its context and intent as well as it can, and then the model will try to guess what the first word (or token) for the output is, that would best fit the given prompt. Then, it will try to guess the second word, and so on. The model does not know what is true and what is false, and it doesn’t understand what is written by Dan Abramov, but it will predict a sequence of words that seems very plausible for the given prompt (Amazon Web Services, 2024a, 2025; Elastic, 2024).

Developers from OpenAI state that the problem-solving abilities of o1-preview are superior to previous models because the LLM can now “think”. OpenAI states this about their new o1 family of large language models: “o1 thinks before it answers—it can produce a long chain of thought before responding to the user.” (OpenAI, 2024c). This thinking process is called reasoning, and it’s implemented with chain-of-thought prompting (Wei *et al.*, 2023, pp. 2–3). With automatic chain-of-thought prompting, the system first clusters the questions if the prompt contains multiple questions and then creates a reasoning process for the prompt, by feeding the LLM with a prompt “Let’s think step by step”. This reasoning process, together with the original prompt, works as the enhanced prompt for the LLM (Zhang *et al.*, 2022, pp. 2–3).

4.2.1 Challenges in AI reliability

As AI solutions are being used in various sectors, there is much discussion about the reliability of AI services (Mortaji and Sadeghi, 2024, p. 1). ChatGPT hallucinates rather often (Cardona, Rodríguez and Ishmael, 2023, p. 7; Gimpel *et al.*, 2023, p. 40), the data that powers the LLMs is often biased (Guan *et al.*, 2024, p. 1), and most of the AI systems’ internal logic is a “black box”, where the systems offer little to no transparency on the decision-making processes (Mortaji and Sadeghi, 2024, p. 5).

One should never take the output of a generative AI system as an absolute truth (Yuxia Wang *et al.*, 2024, p. 1). Even though there are systems in place for some services where the AI system checks the factuality of its output, utilizing retrieval augmented generation (Y. Zhou *et al.*, 2024, p. 1), the core idea of large language models is based on probability (Y. Zhou *et al.*, 2024, p. 4). Large language models produce output that matches the given prompt (input) with the highest probability. This is important to keep in mind when discussing how reliable the current generative AI services can be (Zhao *et al.*, 2024, p. 4).

With the current way transformer-based large language models work, they can never be considered fully reliable (Li *et al.*, 2024, p. 44; Mirzadeh *et al.*, 2024, p. 12; Y. Zhou *et al.*, 2024, p. 4). This is due to their probabilistic nature and the fact that LLMs don't have a grounded understanding of what is true and what is not (Y. Zhou *et al.*, 2024, p. 4). Increasing the size of the model improves the accuracy and reliability of the model, but only to a point. Even though an LLM with a larger training dataset and a larger number of parameters is more accurate and reliable, no increase in size will make the current transformer-based large language model output completely reliable (Kaplan *et al.*, 2020; Bender *et al.*, 2021).

When discussing the reliability of generative AI, the suitable level of reliability for the given use case should be considered. For example, Generative AI can offer tremendous help in the Thesis research process (Alasadi and Baiz, 2023, p. 6), even though it might make factual mistakes or provide subpar sources (Vieriu and Petrea, 2025, p. 8). Ultimately, it is the researcher's responsibility to validate their sources and the factuality of their statements, preferably from multiple sources.

4.2.2 Benchmarks for evaluating Generative AI performance and accuracy

There is a need to measure LLMs' accuracy and have quantitative metrics with what the LLMs can be compared against each other. One way to measure the accuracy of LLMs is with benchmarks based on question-answer datasets, for example, HotPotQA, StrategyQA, and MMLU (HotpotQA, 2024; Yubo Wang *et al.*, 2024, p. 1). New benchmarks are released at quite a rapid pace, and OpenAI released their open-source SimpleQA benchmark at the end of October 2024 (OpenAI, 2024b; Wei *et al.*, 2024, p. 1) HotPotQA, StrategyQA, and SimpleQA are all question-answering benchmarks, where the LLM is given a set of questions, which it needs to answer. Previously

mentioned benchmarks are quite similar, with varying focus areas. For example, HotPotQA has questions, that require complex reasoning and the ability to switch between multiple sources of information to produce a correct answer (HotpotQA, 2024). SimpleQA, on the other hand, focuses on simpler question-answer pairs, and is designed to measure how accurate the model is with facts, and how much the model "hallucinates" (Wei *et al.*, 2024, p. 1).

These benchmarks are suitable for standardization and LLM comparison but don't paint a clear picture of the actual accuracy of the LLM. As an example, even though the o1-preview scores 90,8 points out of 100 on the MMLU benchmark (OpenAI, 2024c), it doesn't mean that the model would always be accurate over 90 % of the time – it simply means that it got 90,8 % of answers correct in that specific benchmark.

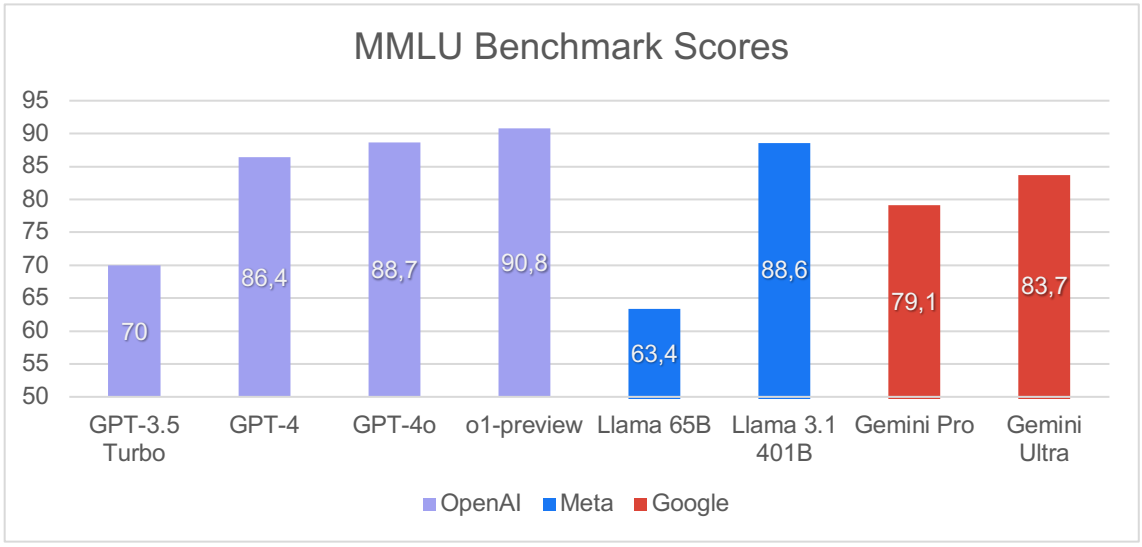


Figure 2: Benchmark scores from the MMLU benchmark (Artificial Analysis, 2025a).

When looking at the performance improvements between models, we can start to paint a picture of how the field is improving. In 2022, GPT-3.5 Turbo was the best model OpenAI had on offer, and it scored 69,8 on the MMLU benchmark as shown in Figure 2. With a quick look, that improvement doesn't look that impressive, with an increase in the relative accuracy of "just" 30 percent. If we consider that the o1-preview made a mistake in the test 9,2 % of the time and the GPT-3.5 Turbo 31,2 % of the time, that's an improvement of 69,5% in relative error reduction rate. In this section, we have discussed many of the AI products developed by OpenAI. Still, many of these findings apply to models created by other companies, such as Llama, developed by Meta, or Gemini,

developed by Google, as seen from Figure 2. OpenAI boldly states that, in selected benchmarks, their reasoning models perform at the level of PhD students in specific fields (OpenAI, 2024c).

Very few tools utilizing generative AI that are aimed at research disclose what large language models they use. When a tool such as Elicit uses a large language model from one of these providers, it pays per token for the usage. In most cases, the pricing is informed in dollars per million tokens, as that is a scale that makes sense. The pricing between the models differs wildly, for example, Google Gemini 1.5 Flash costs $0,60 / 1M tokens (output), and the more capable Google Gemini 1.5 Pro costs $10 / 1M tokens (output) (Google AI, 2024). A price increase of 1566 % between the Gemini models would indicate that tools utilizing these large language models are probably not using these state-of-the-art versions of the models, at least not with the free versions. Perplexity.ai (Perplexity AI, 2024) claims that their free version uses the “Standard Perplexity AI Model optimized for speed and quality”, and with a paid version, the user can choose their preferred AI model.

4.3 Evaluating Output of Generative AI tools

Looking at the performance metrics of different large language models, as shown in the previous chapter, gives us some idea of how the various models stack up against each other. Still, a great-performing model in standardized benchmarks is not a guarantee that the application using that model gives good output (Cohen-Inger *et al.*, 2025, p. 1). The output quality depends on multiple things, besides the model’s capabilities. The quality of the output is affected by the quality of the prompt, context window size, and additional features in the generative AI application, such as retrieval augmented generation (Gimpel *et al.*, 2023, p. 15; Santu and Feng, 2023, p. 1; Y. Zhou *et al.*, 2024, p. 1).

Next, the process of evaluating Generative AI applications is discussed from quantitative and qualitative perspectives. Quantitative evaluation presents the performance of a General AI application in numbers, making surface-level comparisons between applications effortless. According to Fragiadakis *et al.* (2025, p. 7), qualitative evaluation offers more insights into why specific applications perform better than others.

4.3.1 Quantitative Evaluation

Quantitative evaluation best serves the following purposes in assessing generative AI applications: comparing different tools, tracking progress between tool versions, and automated assessments. Quantitative metrics alone don't paint a complete picture of a generative AI application's performance, and in most cases, qualitative evaluation is required to offer meaningful analysis of generative AI application performance (Fragiadakis *et al.*, 2025, p. 7). Numeric values make comparing different generative AI tools efficient and easy. The quality of the comparison is highly dependent on the metrics being measured.

New versions of large language models are published rapidly. For example, Google released 11 different versions of Gemini large language models in four months, from January 2025 to the end of April 2025 (Google AI, 2025d). Quantitative metrics enable the tracking of progress in the development of large language models in this rapidly evolving industry.

Traditionally, quantitative metrics can be divided into automated metrics and human-based metrics (Chan *et al.*, 2023, p. 1). In the context of generative AI, automated metrics include the BLEU score and the ROUGE score, which can be determined through algorithms (Zhao *et al.*, 2024, p. 56). Human-based metrics are defined based on human judgment, such as pairwise comparison results or numeric ratings based on a pre-defined scale. These lines become slightly blurred with the use of generative AI, as these systems can also be employed in the evaluation process. Traditionally human-based metrics can now be measured using large language models, employing the LLM-as-a-Judge methodology (Li *et al.*, 2025, pp. 7, 10–18). This methodology will be covered in more detail in the forthcoming section.

Quantitative evaluation of generative AI is well-developed when the objective is to assess large language models. It is relatively easy to build datasets and create scripts that provide a large number of prompts for large language models, and then compare the output to predefined ground truths (Wang *et al.*, 2019, p. 1; Brown *et al.*, 2020, p. 17). These kinds of evaluation tools and metrics exist already, and these were covered in more detail in the previous chapter, "4.2.2 Benchmarks for Evaluating Generative AI Performance and Accuracy".

According to Zhao *et al.* (2024, p. 58), the performance of the large language model should be one of the metrics considered in the tool selection process, but in the case of selecting AI tools for research and development work in higher education, the benchmark numbers should not be the only method of evaluation. Benchmark numbers from standardized benchmarks do not indicate how well the tool suits the specific domain in which it will be used

4.3.2 Qualitative Evaluation

When assessing the usability of generative AI tools and comparing them against each other, qualitative evaluation done by humans remains the gold standard. In human-led evaluation, the relevancy, accuracy, coherence, and depth of the generative AI output are examples of qualities that can be evaluated. Compared to the benchmark results of different AI models available online, evaluation done by humans can provide domain-specific insights into the evaluation (Bronsdon, 2024).

Human-led qualitative evaluation also introduces some issues. The person conducting the evaluation should be an expert in the domain in which the generative AI model is being evaluated (Bommasani *et al.*, 2022, p. 125). When evaluating the readability or understandability of generated text, the evaluator does not require domain-specific expertise (Bronsdon, 2024).

> While human evaluation is prevalent and indispensable in NLP, it is notoriously unstable (Gillick and Liu, 2010; Clark et al., 2021). Karpinska et al. (2021) has shown that low-quality workforces in human evaluation can have a detrimental effect on the evaluation result, making it impossible to compare the performance among different systems (Gillick and Liu, 2010). (Chiang and Lee, 2023, p. 1)

Humans tend to have strong biases, and research shows that when people are aware that an AI has written something, they tend to be more critical of it (Zhu *et al.*, 2024, p. 1). This is especially problematic with few human evaluators, as human evaluation is highly susceptible to biases (Chan *et al.*, 2023, p. 2). Human evaluation is also heavily dependent on the expertise of the evaluators, and a low-quality workforce can do more harm than good in the evaluation of AI-generated content (Chiang and Lee, 2023, p. 1).

4.3.3 LLM-as-a-Judge based evaluation

As discussed in the previous sections, meaningful evaluation of generative AI applications requires quantitative and qualitative evaluation. Generative AI technologies are evolving so rapidly that manual human-based evaluation is not a sustainable approach for measuring the performance of generative AI applications for specific use cases (Tan *et al.*, 2025, p. 1). For this purpose, an LLM-as-a-Judge methodology was developed.

On a very high level, LLM-as-a-Judge methodology describes a process where generative AI output is evaluated using a large language model based on predefined criteria (Evidently AI, 2025; Li *et al.*, 2025, p. 1). The definition of LLM-as-a-Judge methodology is quite abstract by design, and it can be fitted to suit multiple kinds of evaluation situations. In the overview described in Figure 3, the LLM-as-a-Judge process has been split into 3 phases: Input, System, and Output.

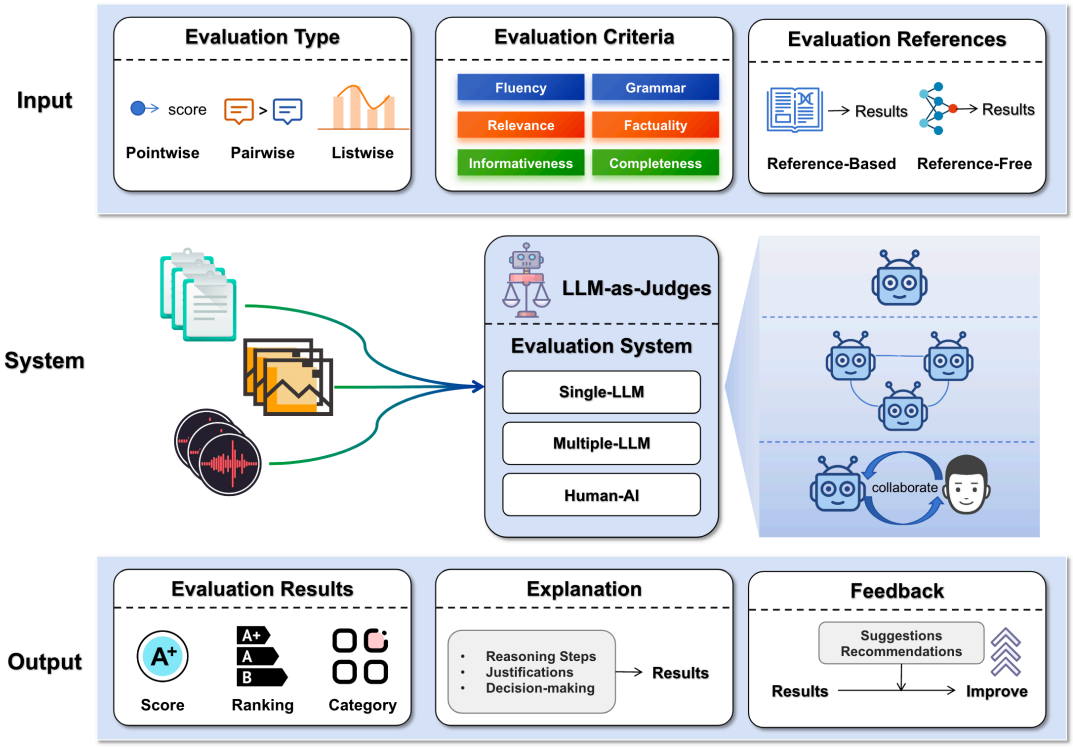


Figure 3: Overview of an LLM-as-a-Judge system (Li et al., 2024).

As shown in Figure 3, in the Input phase, three decisions need to be made. The evaluator chooses what kind of evaluation type will be used: a scoring system, pairwise evaluation, or listwise evaluation. The evaluation criteria are defined, which sets the metrics to be measured in the evaluation and how these metrics are measured. The last stage of the Input phase is selecting whether the evaluation will be reference-based or reference-free. Once these decisions have been made, the evaluation prompt can be built.

4.3.3.1 Creating AI responses for evaluation

Figure 3 only includes the evaluation flow and omits the part where the evaluated generative AI output is generated. The purpose of LLM-as-Judge-based evaluation is to assess the generative AI performance in context-specific tasks (Li *et al.*, 2025, p. 1). The evaluated generative AI applications require prompts that reflect context-specific tasks to achieve this.

The first step in the evaluation process is identifying these context-specific tasks (Peng *et al.*, 2024, p. 2; Zhao *et al.*, 2024, p. 66). These can be simple tasks, such as translating context-specific text, or more complex and abstract, such as brainstorming sessions about context-specific topics (Chan *et al.*, 2023, p. 4; Li *et al.*, 2024, p. 23).

The same prompts are given to each evaluated LLM or AI application, and the outputs are collected (Chiang and Lee, 2023, pp. 2–3). These are later used in the evaluation process.

4.3.3.2 Evaluation types

The selection of the evaluation type affects how the evaluation prompt is built. In a pointwise evaluation, only one generative AI answer is given in the evaluation prompt for the LLM to evaluate. The Judge LLM will evaluate the answer individually based on the defined evaluation criteria and possible reference materials (Verga *et al.*, 2024, p. 2). This process is then repeated with all the answers. Table 2 shows the pros, cons and situations when to use each of the evaluation types.

Table 2: Table describing the pros, cons, and situations when to use each of the evaluation types

	Pros	Cons	When to use
Pointwise	Assessing specific aspects of output (Li *et al.*, 2025, p. 4)	Lacking the relative performance to other answers	Assessing if a generative AI system is suitable for a specific task (Evidently AI, 2025)
Pairwise	Best choice for comparative assessment (Verga *et al.*, 2024, p. 2)	Positional bias (Li *et al.*, 2025, p. 8)	Defining which of the two systems is better in a specific task (Verga *et al.*, 2024, p. 2)
	More accurate for comparison than pointwise (Evidently AI, 2025)	Time-consuming when evaluating more than two systems (Li *et al.*, 2025, p. 10)	
Listwise	Ranking a collection of generative AI outputs (Li *et al.*, 2025, p. 3)	Positional bias (Li *et al.*, 2025, p. 8)	Ranking multiple generative AI systems in order of performance in a specific task (Li *et al.*, 2025, p. 3)
	Less time-consuming compared to pairwise evaluation (Li *et al.*, 2025, p. 10)	Listwise evaluation leads to largest evaluation prompts, causing potential issues with Judge LLM context window size (Tan *et al.*, 2025, p. 12)	

As shown in Table 2, Pointwise scoring is a good choice when the goal is to assess specific aspects of the generative AI output, such as helpfulness, clarity, or tone (Evidently AI, 2025). In a pairwise evaluation, two generative AI outputs are given for evaluation in a single evaluation prompt. Pairwise evaluation evaluates the relative quality between two generative AI outputs. In listwise evaluation, more than two generative AI outputs are given in a single evaluation prompt for the LLM Judge for evaluation (Verga *et al.*, 2024, p. 2).

Pairwise and listwise evaluation might provide more granular scoring between the generative AI outputs. A pointwise system might evaluate two outputs as “4,” but a pairwise comparison will say which answers were better (Evidently AI, 2025). Pairwise and listwise evaluations are best suited for situations where the generative AI outputs are ranked, and pointwise evaluation can be a better choice when determining if a model is suitable for a specific task (Li *et al.*, 2025, p. 3).

A pointwise scoring system might not be as consistent in scoring as pairwise or listwise evaluation, as the answers are evaluated in isolation (Evidently AI, 2025). Then again, Pairwise and listwise evaluations are known to suffer from positional bias, and to acquire as accurate results as possible, the evaluations should be repeated by altering the order of the generative AI outputs (Li *et al.*, 2025, p. 8).

Aspects of the different evaluation types can also be used together (Li *et al.*, 2024, p. 7). For example, in a listwise comparison, in addition to ranking the answers, the LLM Judge can be prompted to include numeric scoring based on the evaluation criteria, just like in pointwise evaluation (Li *et al.*, 2024, p. 6). Here, it is good to keep in mind that the probability of mistakes increases with the complexity of the evaluation prompt (Wang *et al.*, 2023, p. 2; Li *et al.*, 2024, p. 1; Zhao *et al.*, 2024, p. 47).

4.3.3.3 Evaluation Criteria

Evaluation criteria consist of a set of rules, based on which the generative AI output is evaluated. The evaluation criteria should be task-specific and measure relevant metrics for the evaluation context (Li *et al.*, 2024, pp. 3, 7). Good evaluation criteria are critical in an LLM-as-a-Judge evaluation (Li *et al.*, 2024, p. 7; Ragolane, Patel and Salikram, 2024, p. 18).

When using a pointwise evaluation type, the criteria should define how the answers are scored. Binary or low-precision scoring is preferred here (Evidently AI, 2025). Binary scoring systems are more reliable, where the answers are labelled, for example, as “polite” or “impolite” (Evidently AI, 2025). For some evaluation needs, binary scoring might not be accurate enough, and in these cases, a low-precision numeric scoring system should be used.

When using a scoring system, the meaning of each score should be clearly defined in a scoring rubric (Evidently AI, 2025; Li *et al.*, 2025, p. 9). If multiple aspects of the answer are being scored, for example, “politeness,” “clarity,” and “helpfulness,” the score meanings for each of the different aspects should be defined explicitly (Fu *et al.*, 2023, pp. 3–4). A 1-5 Likert scale is a good fit for a scoring system in many cases, where more granular scoring is needed instead of binary scoring (Chiang and Lee, 2023, p. 2; Evidently AI, 2025).

When defining what metrics to evaluate from the generative AI answer, the learning on how to perform human evaluation can also be applied in the LLM-based assessment (Chiang and Lee, 2023, p. 1). Based on the “Best practices for the human evaluation of automatically generated text” by Lee et al. (2019), good evaluation criteria are task-relevant, specific, and focused, and acknowledge Context and Potential Overlaps (van der Lee *et al.*, 2019, pp. 4–9)

Google has recently published a Generative AI evaluation service as part of the Vertex AI suite. It lets developers benchmark any generative model or AI application with their evaluation criteria (Google Cloud, 2025b). The Google Generative AI evaluation service documentation about the evaluation metric definition includes the following code piece, which gives an example on how custom metrics can be defined with rating rubrics for task-specific evaluations (Google Cloud, 2025a).

```
# Define a pointwise metric with two criteria: Fluency and
Entertaining.
custom_text_quality = PointwiseMetric(
    metric="custom_text_quality",
    metric_prompt_template=PointwiseMetricPromptTemplate(
        criteria={
            "fluency": (
                "Sentences flow smoothly and are easy to read,
avoiding awkward"
                " phrasing or run-on sentences. Ideas and
sentences connect"
                " logically, using transitions effectively where
needed."
            ),
            "entertaining": (
                "Short, amusing text that incorporates emojis,
exclamations and"
                " questions to convey quick and spontaneous
communication and"
                " diversion."
            ),
        },
        rating_rubric={
            "1": "The response performs well on both criteria.",
            "0": "The response is somewhat aligned with both
criteria",
            "-1": "The response falls short on both criteria",
        },
    ),
)
```
(Google Cloud, 2025a)

4.3.3.4 Evaluation references

In an LLM-as-a-Judge process, the generative AI outputs can be evaluated based on references, or the evaluation can be reference-free as well (Li *et al.*, 2024, p. 8). In

reference-based evaluation, reference material is added to the evaluation prompt. The generative AI output can be directly compared against this reference material, or the reference material can guide the LLM in the evaluation process (Li *et al.*, 2024, p. 8). The reference material can be an example of a desired answer or the original prompt given to the generative AI system being evaluated (Li *et al.*, 2024, p. 8; Verga *et al.*, 2024, p. 2). The reference material can be obtained using Retrieval-Augmented Generation, where the Judge AI retrieves additional context to aid the evaluation (Li *et al.*, 2025, p. 15). The additional context can be retrieved from the internet or a more specific dataset containing context-specific information, such as a curated collection of research papers (Li *et al.*, 2025, p. 15).

Whether or not to base the evaluation on references should be based on the evaluation task. Reference-based evolutions are suitable for tasks where the quality of the AI answer can be objectively measured by comparing it to an established reference. Translation tasks are a good example where reference-based evaluations can be used (Li *et al.*, 2024, p. 8).

In a reference-free evaluation, the AI answer is judged based on the evaluation criteria and the Judge LLM's knowledge. For example, when evaluating the coherence of a text content generated by AI, the judge LLM applies its knowledge of grammar and semantic understanding (Li *et al.*, 2024, p. 8).

4.3.3.5 Evaluation Systems

Once the generative AI outputs have been collected, the evaluation type has been selected, and the evaluation criteria have been defined, the evaluation prompt can be built. The evaluation prompts should be carefully crafted, following prompt engineering best practices, as the quality of prompts will drastically effect the evaluation quality (Wang *et al.*, 2023, p. 7). They should be clear and precise, the instructions should be exact, and the judge LLM should be given a clear role (Zhao *et al.*, 2024, p. 47). If human-based evaluation results are available, these evaluation results can be used as few-shot examples (Li *et al.*, 2025, p. 10). The Google Gen AI evaluation service documentation contains an example of an evaluation prompt used in LLM-as-a-Judge evaluation (Google Cloud, 2025c).

```
# Instruction
```

```
You are an expert evaluator. Your task is to evaluate the quality
of the responses generated by AI models.
We will provide you with the user input and an AI-generated
response.
You should first read the user input carefully for analyzing the
task, and then evaluate the quality of the responses based on the
Criteria provided in the Evaluation section below.
You will assign the response a rating following the Rating Rubric
and Evaluation Steps. Give step-by-step explanations for your
rating, and only choose ratings from the Rating Rubric.

# Evaluation
## Metric Definition
You will be assessing summarization quality, which measures the
overall ability to summarize text.

## Criteria
Instruction following: The response demonstrates a clear
understanding of the summarization task instructions, satisfying
all of the instruction's requirements.
Groundedness: The response contains information included only in
the context. The response does not reference any outside
information.
Conciseness: The response summarizes the relevant details in the
original text without a significant loss in key information
without being too verbose or terse.
Fluency: The response is well-organized and easy to read.
Reference alignment: The response is consistent and aligned with
the reference response.

## Rating Rubric
5: (Very good). The summary follows instructions, is grounded,
concise, fluent and aligned with reference summary.
4: (Good). The summary follows instructions, is grounded, concise,
and fluent but not aligned with reference summary.
3: (Ok). The summary mostly follows instructions, is grounded, but
is not very concise and is not fluent and is not aligned with
reference summary.
2: (Bad). The summary is grounded, but does not follow the
instructions.
1: (Very bad). The summary is not grounded.

## Evaluation Steps
STEP 1: Assess the response in aspects of instruction following,
groundedness, conciseness, fluency and reference alignment
according to the criteria.
STEP 2: Score based on the rubric.

# User Inputs and AI-generated Response
## User Inputs
### Reference
{reference}

### Prompt
{prompt}

## AI-generated Response
{response}
```
(Google Cloud, 2025c).

The evaluation can be based on a single LLM Judge or multiple Judges (Li *et al.*, 2024, p. 6; Verga *et al.*, 2024, p. 2). Single LLM evaluation is simpler to set up (Li *et al.*, 2024, p. 46), but large language models suffer from intra-model bias in evaluation

(Panickssery, Bowman and Feng, 2024, p. 8), which can be mitigated using multiple LLM judges (Li *et al.*, 2024, p. 6).

Judge selection can also be an involved process. Generally, bigger, better-performing models perform better as judges (Verga *et al.*, 2024, p. 1). More competent models tend to be more consistent in their evaluations and can pick up nuances that less capable models might miss (Sreekar *et al.*, 2024, p. 8). Using a large, highly performing model in the evaluation process can also become costly when the scale is large, and the evaluation is automated and frequent (Verga *et al.*, 2024, p. 2). Recent research shows that a jury of multiple smaller, large language models can be a more performant and cost-efficient solution than a single large model (Li *et al.*, 2024, p. 46; Verga *et al.*, 2024, pp. 2, 6).

Evaluation can also be performed in a human-AI collaboration model, where humans and LLMs work with human evaluators. The model aims to mitigate potential issues with judge model biases and improve evaluation quality with nuanced human expert knowledge (Li *et al.*, 2024, p. 6).

4.3.3.6 Evaluation Output

In a typical LLM-as-a-Judge process, the system generates three kinds of outputs: the evaluation result, the explanation, and feedback. The evaluation results can be numeric based on an evaluation rubric, ranking order of AI outputs, a categorical label, or a qualitative assessment. The explanation provides justification for the evaluation and contains insights on why the AI outputs were scored a certain way. The feedback includes actionable suggestions on improving the AI output (Li *et al.*, 2024, p. 8).

These parts are not all mandatory in the evaluation (Li *et al.*, 2024, p. 8), and they need to be explicitly requested in the evaluation prompt if they are required as part of the evaluation (Verga *et al.*, 2024, p. 2). Including the explanation and feedback sections as part of the evaluation output improves the consistency and accuracy of the LLM-based evaluation (Evidently AI, 2025).

The LLM-as-a-Judge methodology is a critical part of my Thesis, as the conceptual framework is built around this methodology. Utilizing LLM-as-a-Judge methodology as

part of AI tool evaluation process can bring big efficiency improvements to the evaluation process, compared to solely relying on human evaluation.

4.4 Conceptual Framework of This Thesis

This chapter introduces the conceptual framework of this Thesis. Here, the best available knowledge about generative AI tool evaluation is gathered in a visualized format. The process shown in the conceptual framework is utilized in the section 5.

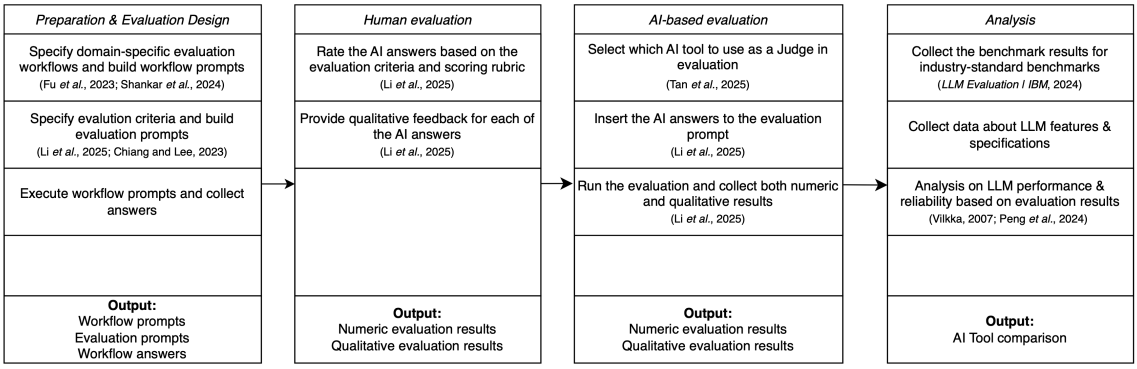


Figure 4: Conceptual framework for the generative AI application evaluation process.

Figure 4 illustrates the conceptual framework for the generative AI service evaluation process. My goal with this Thesis is to evaluate a selection of AI tools for academic writing purposes. To achieve this goal, a process was needed to evaluate and compare different generative AI tools against each other, as well as their suitability for specific tasks. This process is based on the LLM-as-a-Judge methodology. This conceptual framework builds upon the LLM-as-a-Judge methodology, demonstrating how human evaluation can be integrated with the LLM-as-a-Judge evaluation model.

The first step in the evaluation process is *the preparation and evaluation design*. In this step, context-specific evaluation workflows are defined, and workflow prompts are created. The evaluation criteria and scoring rubrics are defined, and the evaluation prompts are built. The final step of the preparation and evaluation design is to run the workflow prompts with all the evaluated AI tools and collect the answers.

The second step of the evaluation process is *the human evaluation*. All answers should be graded by a domain expert, using the same evaluation criteria and scoring rubrics as

used in the AI-based evaluation. In addition to the numeric scoring, quantitative feedback should also be provided.

The third step in the process is *the AI-based evaluation*. The first phase is deciding on which AI tool or LLM to use as the Judge. Multiple aspects should be considered in selecting the LLM Judge, and the amount of work required for the selection process also depends on the evaluation type. For manual evaluations that are run infrequently, using a large and expensive model is usually the best choice. For more frequent automated evaluation, further consideration of judge selection is needed from both cost and performance reasons. In the AI-based evaluation, the AI's answers to the workflow prompts are given to the judge for evaluation based on the evaluation criteria and the scoring rubrics, and the results are collected.

The final step in the evaluation process is *the Analysis.* As part of the Analysis step, results from industry-standard benchmarks should be collected for the selected AI applications from online resources (IBM, 2024). These are available for free from services like Artificial Analysis. The selection of benchmark results is extensive in Artificial Analysis, so collecting benchmark scores should be straightforward and requires a minimal amount of work. There are multiple types of benchmarks for measuring the performance of generative AI, and the benchmarks that best suit the domain and use case should be selected. As an example, if evaluating what AI application would be the best tool for aiding students in learning coding, more coding-related benchmarks like LiveCodeBench and SciScode should be added to the benchmark score dataset. At this stage, the features and specifications of the evaluated AI tools or large language models (LLMs) should be collected to aid in the analysis. The final phase is analysing the gen AI tool or LLM performance and reliability based on the evaluation results, qualitative feedback, industry standard benchmark results, and the features and specifications.

Next, this conceptual framework is used to guide the proposal building for the case organization.

5 Building Proposal for AI Tool Evaluation Framework

This section presents an initial proposal for the AI tool evaluation framework, accompanied by detailed guidelines. This initial proposal is based on findings from the current state analysis, the literature review, and stakeholder feedback.

5.1 Overview of the Proposal Building Stage

This section outlines the development of the initial proposal for a framework to evaluate a selected set of AI tools in order to determine their suitability for academic writing purposes by students. To achieve this, the AI tools are systematically evaluated within the MBS context.

The initial proposal was built in 2 steps. First, four interviews were conducted with stakeholders from MBS and the key common findings from the interviews, CSA, and conceptual framework were identified. Second, the initial proposal was built based on these findings.

Findings from Data 2

This section presents the findings from Data 2, related to the key focus areas identified in the current state analysis and the best available knowledge from the literature review. The key findings from the current state analysis affecting this Thesis work were that MBS doesn't yet have a centralized solution for general-use AI to offer to students, and implementing any new IT tool can be a lengthy and challenging process due to missing skills in evaluation of AI tools, and there are also no detailed AI tools instructions for academic writing purposes.

Table 3: Key stakeholder suggestions (findings of Data 2) for Proposal building in relation to findings from the CSA (Data 1) and the Conceptual framework.

	*Key findings from **CSA** (from **Data 1**)*	*Inputs from literature (**CF**)*	*Findings from stakeholders for the Proposal, summary (from **Data 2**)*	*Descriptions of their suggestions (in detail)*
1	No general-use AI solution to	Large Language Model	The evaluation should focus on	MBS is investigating an LLM agnostic solution for a general-use AI application.

	offer for students or personnel	Evaluation utilizing LLM-as-a-Judge methodology (Li *et al.*, 2024, 2025; Evidently AI, 2025)	general-use AI tools	
	Taking new tools into use requires a thorough and time-consuming evaluation process		Evaluation should be based on domain-specific tasks, and focus should be on qualitative analysis of the AI's answers	The standardized benchmarks are insufficient in measuring the LLM capabilities in the MBS context. *"Perhaps the biggest problem is that people evaluate mainly just quantitatively because it's easy, and that leads to the problem that, well, these benchmarks don't necessarily describe anything at all."*
	No detailed AI tools instructions for academic writing purposes		The evaluation framework should be reusable, and the evaluation workload should be so that it can be done i.e. every 6 months	*"That is precisely the concrete timeframe in this case, that this is thought of on a six-month timeframe, not over the next 5 years."*

As seen from Table 3, the main insights from interviews with key stakeholders were that the evaluation should focus on general-use AI tools, the evaluation should be based on domain-specific tasks, and the evaluation framework should be easy enough to implement, so that the evaluation can be quite constant, if needed.

Focusing on the evaluation of general-use AI tools, directly assesses the situation that MBS doesn't yet have a general-use AI tool to offer to students and lecturers, which was identified in the current state analysis. In the key stakeholder interviews, a proposal was made to focus on general-use AI tools based on the findings from the CSA, and each of the stakeholders agreed with this. The focus on general-use AI tools was also backed by the findings from CSA, which stated that taking new tools into use requires a thorough and time-consuming validation process.

Based on the stakeholder insights, the evaluation should be done with domain-specific tasks, emphasizing qualitative analysis of the AI answers. The industry standard benchmarks offer little value when trying to depict whether a tool is suitable for use in the MBS Context, as they only portray the LLM reasoning capabilities on a general level.

Other issues with the benchmarks, such as model fitting for the benchmarks and data contamination, are discussed in more detail in Section 4.

As the generative AI industry is developing rapidly, the evaluation framework should be relatively lightweight to implement. Based on the stakeholder interviews, the evaluation frequency should be approximately 6 months to maintain an up-to-date understanding of developments with AI tools.

5.2 Initial Proposal: A Framework & Guidelines for AI Tool Evaluation

The proposed AI tools evaluation framework has four phases. To facilitate the adoption of this AI tool evaluation framework, explicit guidelines have been developed to guide the evaluator through the evaluation process.

5.2.1 Phase 1: Preparation and Evaluation Design

The first phase in this evaluation process is the preparation and design phase for evaluation. In this phase, the first step is to create evaluation workflows based on the best available knowledge.

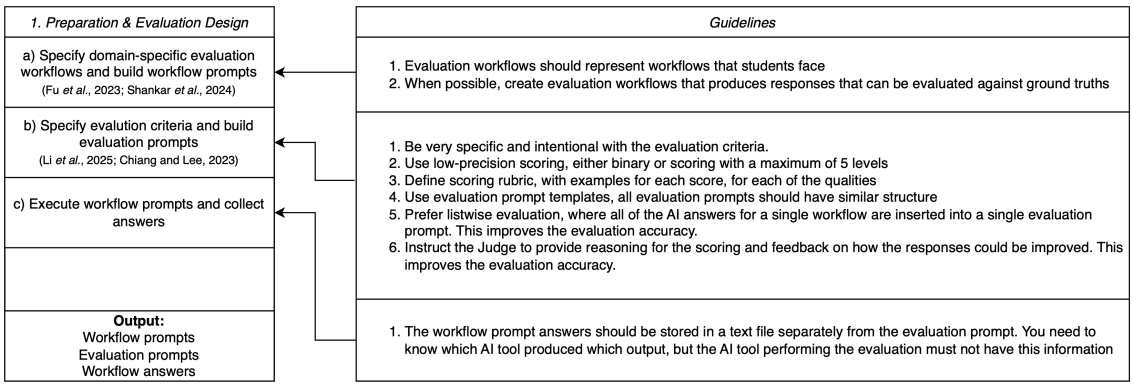


Figure 5: Preparation and evaluation design of the AI tool evaluation framework with guidelines. The preparation and evaluation design consists of three distinct steps: specifying the evaluation workflows and criteria, executing the workflow prompts, and collect.

The first step (a) in the preparation and evaluation design is **crafting the evaluation workflow prompts**. The workflow prompts describe the task or set of tasks that the

evaluated AI should perform. The AI answers to these workflow prompts are used to evaluate the AI tools.

Benchmarks like MMLU score the LLMs based on a set of questions and multiple-choice answers. This makes it easy to automatically validate the answers in these benchmarks, as there is a single correct answer to each question and no room for interpretation. The domain-specific tasks for this evaluation should supplement these benchmarks and focus on more open-ended answers.

In specific tasks, evaluation accuracy can be improved by having high-quality, exhaustive ground truths that are used in the evaluation process. For example, when an AI tool was prompted to identify knowledge gaps in a set of research sources, the ground-truth materials should contain an exhaustive list of sources for the specific topic being tested, so that during validation, it can be identified if the AI tool has missed some critical sources or added sources that are not relevant to the topic. To avoid any copyright issues while testing the AI tools, it's crucial to use only openly available research papers. Evaluation containing ground truths is called reference-based evaluation, and evaluation without ground truths is called reference-free evaluation.

The next step (b) in the Preparation and Evaluation Design is to **create evaluation criteria and develop the evaluation prompts**. The evaluation prompt should include at least the following sections: General Instructions and role setting, overall metric definition, evaluation steps, criteria for evaluation, instructions for output format, original workflow prompt, and the AI responses.

General Instructions and role setting provide the AI Judge a role and high-level instructions on how the evaluation should be conducted. Setting a role for the AI Judge is a standard prompt engineering method that has proven to improve the quality of AI responses. In the overall metric definition section, the qualities on which the AI responses are evaluated should be briefly reviewed.

The next part of the evaluation prompt is the evaluation steps. This section guides the AI Judge through the reasoning process, explaining the evaluation steps in detail for the AI Judge. Having the steps implicitly defined improves the AI tools' accuracy and reliability

The evaluation criteria should be precise and intentionally crafted. The evaluation criteria describe how the AI responses should be evaluated and list qualities based on which the answer should be graded. This evaluation criterion is then inserted into the evaluation prompt, together with instructions and the AI response or responses. For numeric scoring, a low-definition system should be used, either boolean results or a scale from 1 to 5. Each score should have a clear definition in the scoring rubric.

The evaluation prompt should also include specific instructions for the output format. Here, the most critical area is instructing the AI Judge to provide detailed justification for the scoring and ranking of responses, as well as actionable feedback for each AI response. Research has shown that this is an effective way to improve the accuracy of the evaluation (Li *et al.*, 2024, p. 8). The instructions for output should also include guidelines for the structure of the AI Judge output.

To perform the evaluation, the evaluation prompt contains both the original workflow prompt given to the AI tools and the AI responses being evaluated. The most important thing in this section is not to include the information on which AI tool produced which response. Having this information in the evaluation prompt would drastically affect the evaluation. The evaluation prompt can consist of just one AI tool response for evaluation or multiple. Providing multiple AI tool answers in a single evaluation prompt is preferable, as enables the AI Judge to reference different AI tool responses in the evaluation.

The last step (c) of the Preparation and Evaluation Design is **executing the workflow prompts and collecting the AI responses**. The AI responses should be stored in a separate text file from the evaluation prompt, as the evaluator needs to know which response was produced by which AI tool. However, this information must be kept hidden from the AI judge.

5.2.2 Phase 2: Human evaluation

With the workflow prompts defined and executed, and the evaluation prompt defined, the next phase in the AI tool evaluation framework is human evaluation.

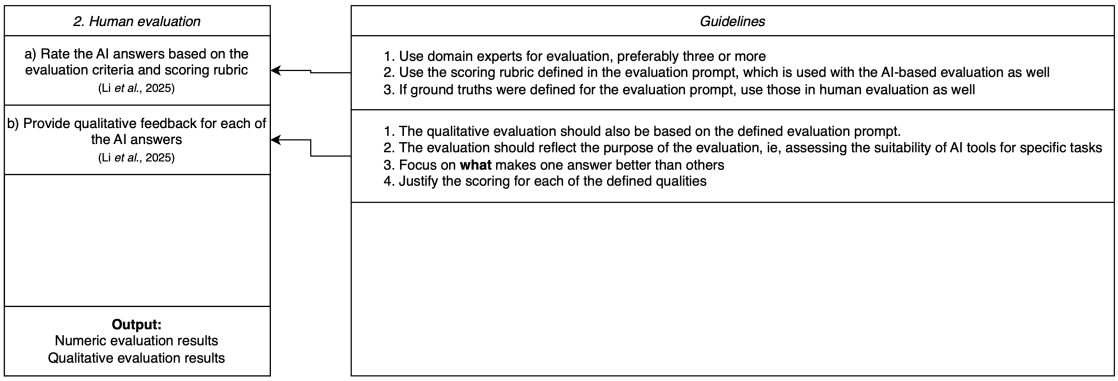


Figure 6: The human evaluation phase of the AI tool evaluation framework with guidelines. This phase consists of two steps: rating the answers based on the evaluation criteria and scoring rubric and providing qualitative feedback for the answers.

In this phase, domain experts assess the AI responses and provide feedback and scoring based on the specified evaluation criteria. A crucial aspect of the human evaluation is that it follows the same evaluation criteria defined for the evaluation prompt. This ensures that human evaluators assess the AI responses based on the same qualities as the AI judge.

In the AI tool evaluation framework, the human evaluation phase is divided into two steps: (a) **Providing numeric scoring for the AI responses** and (b) **providing qualitative feedback for the AI responses**. In practice, however, these steps occur simultaneously and are documented as separate steps to enhance clarity.

For numeric scoring, the same scoring rubric defined in the evaluation prompt should be used in human evaluation. If ground-truths have been described for the evaluation, these should be referenced in the human evaluation. The qualitative evaluation of AI tools should focus on why specific answers are more effective than others. The qualitative assessment should regard each of the qualities defined in the evaluation criteria.

The human evaluation results should be stored to a spreadsheet file for further analysis, in the last phase of the evaluation framework.

5.2.3 Phase 3: AI-based evaluation

The third phase is the AI-based evaluation. In this phase, the selection of which AI tools act as the AI judge is made, and the AI responses are evaluated by the AI judge or

judges.

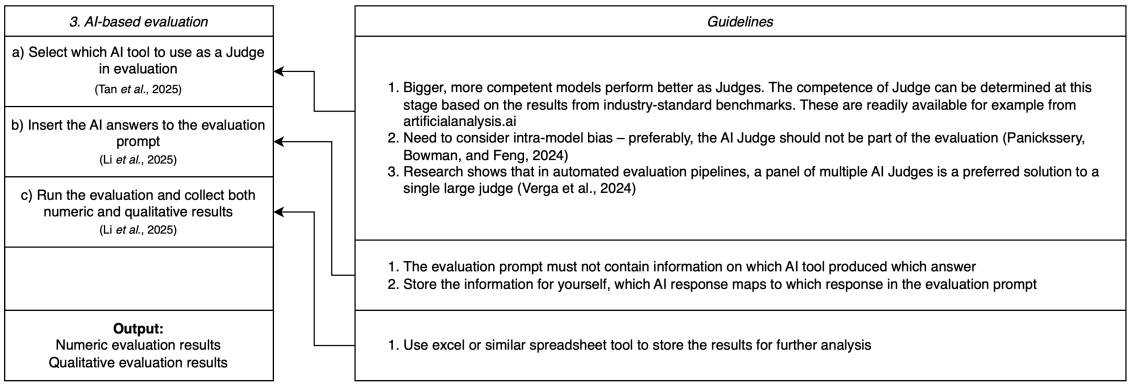


Figure 7: The AI-based evaluation phase. AI-based evaluation is split into 3 distinct steps: selecting the LLM Judge, inserting the evaluation workflow answers to the evaluation prompt, and executing the evaluation.

The first step (a) in the AI-based evaluation is **selecting the AI Judge**. Based on research (Tan *et al.*, 2025, p. 8) bigger, more competent AI tools perform better as Judges. To determine which AI tools are more competent, results from industry-standard benchmarks can be used to guide the selection. These results can be found, for example, from artificialanalysis.ai. All AI Judges are susceptible to multiple kinds of bias, one of which is the intra-model bias. Intra-model bias means that the AI tool prefers responses in evaluation, which have been created with the same model that is evaluating the responses (Zhao *et al.*, 2024, p. 73). Using an AI tool as the AI judge, which is part of the evaluation, should be avoided due to intra-model bias. Basing the AI judge selection solely on the AI tool's competence is a valid method for manually performed evaluations, where the number of prompt calls and token usage is relatively limited, as this generates little to no cost. However, in an automated setup where evaluations run daily, using a highly competent AI tool for evaluation can result in unreasonable costs. To mitigate this, a panel of smaller models should be used for the evaluation (Verga *et al.*, 2024, p. 1).

The next step (b) in AI-based evaluation is **inserting the AI responses to the evaluation prompt**. Again, it should be remembered that the evaluation prompt should not contain the information about which AI tool produced which response. The responses should be labelled with alphabets, such as “Response A, Response B…” and so on. The evaluator should keep track of this, however, and here it’s beneficial to store that information, for example, in a spreadsheet, where the response alphabets are mapped to the correct AI tools.

The final step (c) in the AI-based evaluation phase is **executing the evaluation prompts and collecting results**. After the evaluation prompt has been executed with the AI judge, the evaluation results should be stored in a spreadsheet file for further analysis.

5.2.4 Phase 4: Analysis of Evaluation Results

The last phase in the AI tool evaluation framework is the analysis of evaluation results. In this phase, the results from the various evaluation methods are brought together, along with the features and specifications of the evaluated AI tools and the outcomes from industry-standard benchmarks.

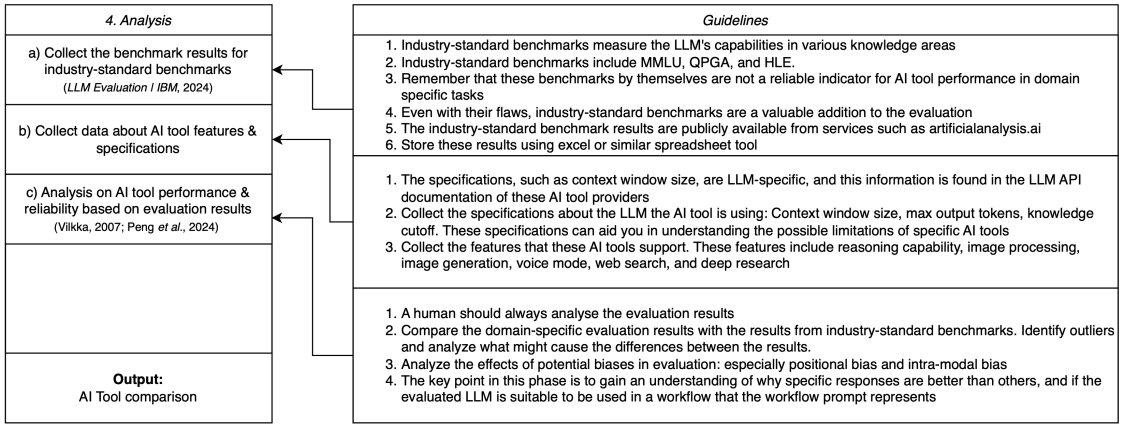


Figure 8: The Analysis of Results phase. Analysis phase consist of three steps: Collecting industry-standard benchmark results from online sources, collecting data about AI tool features and specifications, and analysis on LLM performance.

The first step (a) in the analysis phase is to **collect results from industry-standard benchmarks for the AI tools being evaluated**. These results can be found from artificialanalysis.ai. To enable analysis of these results, the results should be stored in a spreadsheet file.

The second step (b) in the analysis phase involves **collecting features and specifications for the evaluated AI tools**. This step can be a bit tricky, as not all the specifications are published by the AI tool developers. Specifications for the corresponding LLM APIs are available, and these should be used to define the specifications of the specific AI tool, such as context window size, knowledge cutoff date, and token limits. If the AI tool developer has published these specifications for the AI tool, they should be used instead; however, this isn't the case in most es. The AI tool

features are often better documented, but depending on the AI tool developer, this documentation can also be lacking. The feature collection is much easier to assess by manually testing the available features in the AI tools.

The final step (c) in the analysis phase is the **analysis** itself, which should be performed by humans (Cardona, Rodríguez and Ishmael, 2023, p. 44; Ragolane, Patel and Salikram, 2024, p. 18). The numeric domain-specific evaluation results should be compared with those from industry-standard benchmarks. Often, these results should correlate strongly. The correlation should be measured by applying the Pearson correlation coefficient (Schober, Boer and Schwarte, 2018, p. 3). From this correlation analysis, outliers in the pattern can be identified, and a deeper analysis should be conducted to understand why specific AI tools performed better or worse than expected based on the results from industry-standard benchmarks. At this stage, the potential biases affecting the evaluation should be considered. This includes, for example, positional bias and intra-model bias (Zhao *et al.*, 2024, p. 73).

The key point to assess at this stage is why specific responses are better than others, and which AI tools produce satisfactory responses for the workflow prompt, indicating that the AI tool would be suitable for use in the corresponding workflow by students.

5.3 Summary of the Initial Proposal

The initial proposal of this Thesis is the AI Tool Evaluation Framework, accompanied by specific guidelines on its implementation. Guidelines are created for each step of each evaluation framework phase, so that someone not familiar with the subject can implement the framework accurately with a reasonable amount of effort.

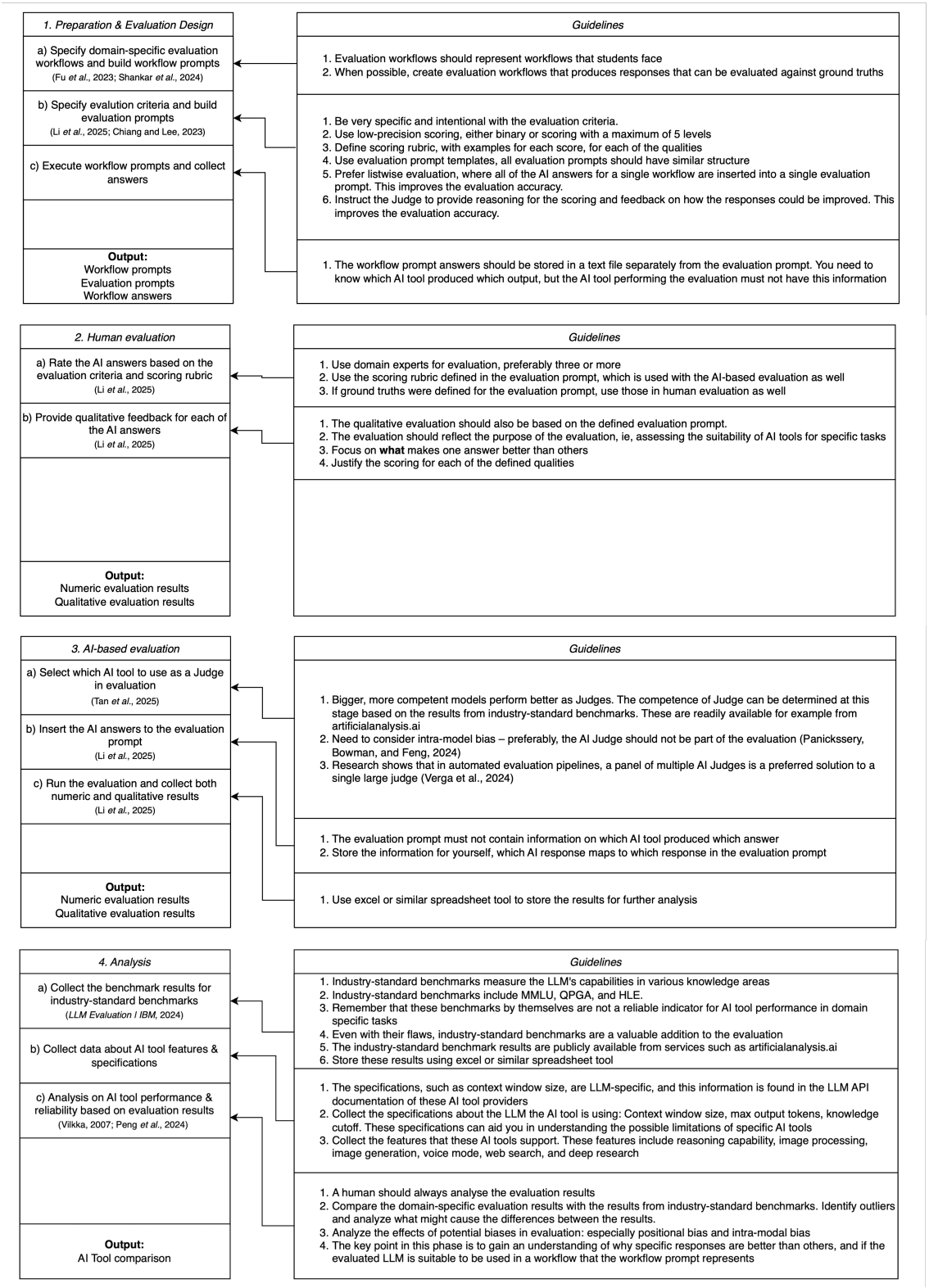


Figure 9: Initial proposal of AI tool evaluation framework with explicit guidelines for implementation.

This framework focuses on evaluating AI tools, which represent the “application layer” in a generative AI service architecture. As discussed in Section 4.1.3 “Overview of Generative AI Application Architecture”, a Generative AI service consists of multiple layers. Evaluating the application layer provides the best picture of how the specific AI application performs for end-users. Manually conducting the evaluation is straightforward, and this approach yields the most significant level of feature parity between the AI tools.

In the future, evaluating the LLM API layer could be beneficial for the development work of MBS general-use generative AI tool. This evaluation framework can be applied closely as-is for the evaluation of the LLM API layer as well. Evaluating the LLM API layer would require some amount of software development for each of the LLM provider services; separate implementations would need to be created for OpenAI APIs, Gemini APIs, and so on. For more information on testing the LLM API layer, refer to the service provider’s API documentation (Google AI, 2025b; OpenAI, 2025c).

The next part of the Thesis work involves applying the AI Tool Evaluation Framework to MBS-specific workflows.

6 Implementation and Validation of the Proposal

This section reports the implementation of the AI tool evaluation framework by presenting the evaluation results, and introduces the developments made to the initial proposal based on key stakeholder feedback and insights gained during the implementation phase. At the end of this section, the final proposal for the AI tool evaluation framework is presented.

6.1 Overview of the Validation Stage

This section reports on the validation results of the proposal developed in Section 5. The goal of this section is to present the results from the AI tool evaluation and how the AI tool evaluation framework was further developed based on the key stakeholder feedback and insights gained during this implementation.

The validation was done in three steps. The first step focused on the practical implementation of the AI tool evaluation framework and the evaluation results. This part included obtaining the results for each evaluation workflow, along with an analysis of the evaluation outcomes.

The second step focused on gathering the feedback and developments ideas to the proposal, by discussing with the key stakeholders and showcasing how the insights gained were used to refine the AI tool evaluation framework. The results were presented to the MBS AI team, and feedback was gathered for further development. This feedback, together with insights gained during the implementation of the evaluation was used to refine each of the evaluation phases.

The last step was pulling together the final proposal and another round of validation discussion with a lkey stakeholder, which led to the final AI tool evaluation framework along with refined guidelines.

6.2 Implementation of the AI Tool Evaluation Framework

This section provides an in-depth look into the key steps on how the AI tool evaluation framework was implemented to evaluate a selected set of AI tools to determine their suitability for academic writing purposes by students (spring 2025).

This sub-section starts by describing the creation of evaluation workflows. Secondly, it presents the selection process for the evaluated AI tools, followed by a section on how the AI judge for the AI-based evaluation was selected. After that, the results from industry-standard benchmarks are presented, followed by those from domain-specific evaluation workflows. This sub-section ends with the analysis of the evaluation results.

6.2.1 Creation of Evaluation Workflows

For this evaluation, three evaluation workflows were selected. The goal was to create the workflows with varying levels of difficulty, so that to test the effects of prompt quality. For testing the prompt quality, the TELeR Taxonomy was utilized (Santu and Feng, 2023, pp. 3–5) to create prompts with varying levels of quality. The first workflow represents a level 1 prompt from the TELeR taxonomy, and the second and third workflows are level 3 prompts.

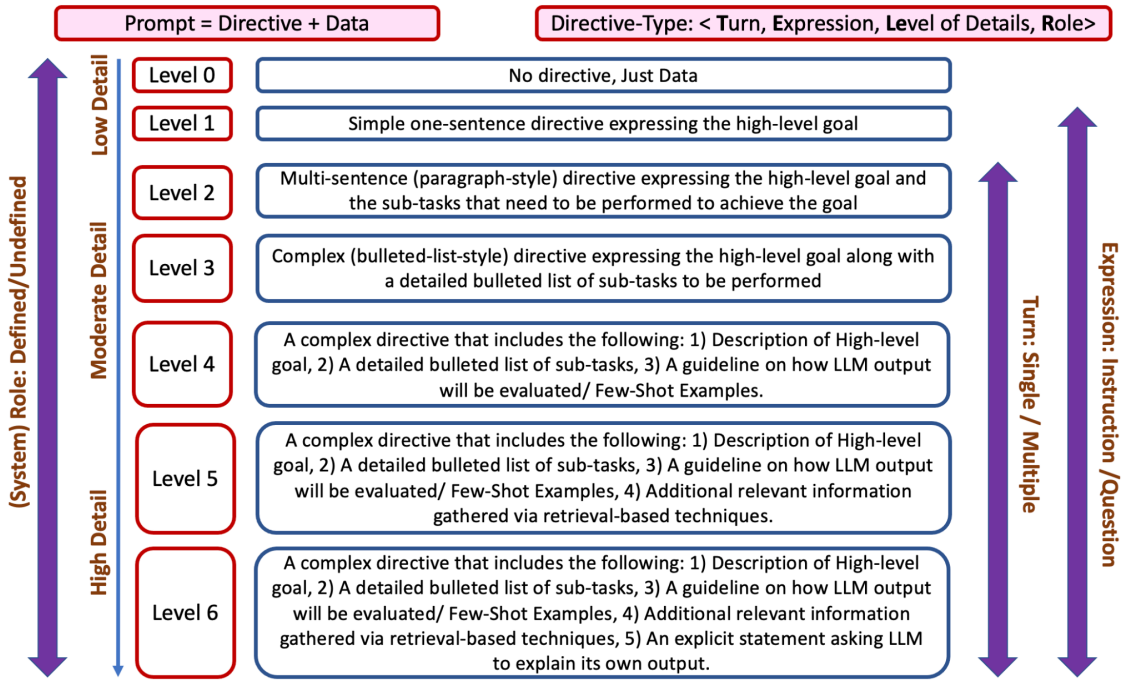


Figure 10: TELeR taxonomy, describing six different levels of prompt details (Santu and Feng, 2023, p. 3).

The first evaluation workflow is “Information gathering with a sub-optimal prompt”, and the purpose of this workflow is to tests the AI tools capability to coherently explain topics with factual accuracy and reasonable depth, when provided with limited context. The second workflow is “Academic editing & proofreading”, which tests the AI tools capability to handle large amounts of textual content, follow specific instructions, and perform academic editing and proofreading. The third evaluation workflow was “Identifying knowledge gaps & academic text generation”, and the goal with this workflow was to tests the AI tools capability to identify gaps in research from a relatively large amount of text, follow specific instructions, and the ability to generate high-quality academic text content. The results sub-section contains more detailed descriptions for each evaluation workflow.

The evaluation workflows were designed to represent workflows that students could use in the Thesis process. I had a couple of goals in mind when designing these evaluation workflows. The workflows had to be complex enough to challenge the AIs, and there would be some variation in the answers.

6.2.2 How the Generative AI tools were selected

There are tens, if not hundreds, generative AI tools available that can be used in a higher education context. Ithaka S+R offers a “Generative AI Product Tracker” service, where the available tools are split into the following categories: General Purpose Tools, Discovery Tools, Teaching and Learning Tools, Research Workflow Tools, Writing Tools, Coding Tools, Image Generation Tools, and Other Tools (Ithaka S+R, 2025). There are many ways to split these tools into categories, and I have chosen this categorization to streamline the communication of my choices.

This Thesis focuses on tools that would belong in the “General Purpose Tools” category. This category focuses on tools that can be used in a wide variety of tasks in the Thesis workflow. Tools belonging to categories like “Discovery”, “Research Workflow” or “Writing” are often hyper-focused on certain tasks, and their overall usability is limited to certain tasks. Often these same tasks can be completed with a general-purpose AI tool if the user is experienced with creating prompts.

For the students of a university of applied sciences, there is a clear need for a general-use generative AI tool. When discussing the selected tools for evaluation with lecturers, they all supported focusing on general-use generative AI tools. Focusing on the general-use generative AI tools in this Thesis would help that work.

The following tools were tested in this Thesis, as listed in Table 4 below.

Table 4: AI tools evaluated in spring 2025.

1. OpenAI ChatGPT 4o
2. OpenAI ChatGPT 4.5 (preview)
3. OpenAI ChatGPT o3
4. OpenAI ChatGPT o4-mini (high)
5. xAI Grok 3
6. DeepSeek R1
7. Mistral Pixtral Large
8. Google Gemini 2.5 Pro (preview)
9. Google Gemini 2.5 Flash (preview)

As seen from Table X, a wide range of large language models tested from OpenAI were selected to showcase the difference in model performance moving from more basic models (4o) to more advanced thinking models (o3, o4-mini (high)). GPT-4.5 (preview) is the latest model from OpenAI, and it should excel especially in writing tasks; however, it lacks the reasoning, also known as a chain-of-thought, that models like o3 and o4-mini possess (OpenAI, 2025e). The GPT-4.5 is also wildly expensive to use, but it paints a picture of how well models without reasoning capabilities can perform in complex tasks.

The other models represent a selection of the best models from each company. The xAI Grok 3, DeepSeek R1, and Google Gemini 2.5 Pro (preview) are all extremely well-performing models in the industry-standard benchmarks. Additionally, the thesis researcher wanted to include at least one large language model from a European company, and Mistral Pixtral Large represents Europe in this comparison.

Yet, this is not an exhaustive evaluation and it was limited due to both workload and resource reasons. Also, some of the models were not accessible without a paid subscription. Most notably, this evaluation is missing Antrophic's Claude Sonnet 3.7, Meta Llama 3.3, and AWS Nova Pro. To further expand this evaluation, it would be beneficial to include cheaper models in the evaluation besides OpenAI o3-mini. These models would include for example Google Gemini 2.0 Flash Thinking and DeepSeek V3.

6.2.3 Selection of the AI Judge

The Generative AI tool used for evaluation was selected based on AI tool features and competence in industry-standard benchmarks.

There were three generative AI tools found suitable for use as a Judge in evaluation: Google Gemini 2.5 Flash (preview), Gemini 2.5 Pro (preview), and xAI Grok 3. They could all be used as the evaluation tool, as a large token limit was required to perform the evaluation. Most AI tools have limited token amounts in their web applications. Based on my testing, the OpenAI GPT-4.0 and GPT-4.5 models have a token limit of approximately 4000 tokens in the ChatGPT web application. I first noticed this limitation during my testing and verified this by asking these models to generate text output in English with a length of 10,000 tokens. These outputs terminate abruptly at approximately 4000 tokens. I calculated the number of tokens by dividing the number of words by 0,75, which gives me a rough estimate of the token count (OpenAI, 2025g).

GPT-4o and GPT-4.5 (preview) models have higher token limits when used through the OpenAI API. The output token limit for both GPT-4o and GPT-4.5 (preview) APIs is 16,384 tokens (OpenAI, 2025b, 2025a).

For this evaluation, Google Gemini 2.5 Pro (preview) was chosen as the tool to perform the Generative AI-based evaluation. The choice was based on exceptional performance in benchmarks (MMLU-Pro, GPQA Diamond, Humanity's Last Exam, and MATH-500). At the time of writing, Gemini 2.5 Pro (preview) has the highest scores out of all tested large language models in artificialanalysis.ai (Artificial Analysis, 2025b). Gemini 2.5 Pro is in a "preview" phase. Still, it performed excellently when testing the consistency, and it was by a large margin the most critical of the models in the evaluation. The Gemini 2.5 Pro has exceptionally high token limits, making it ideal for handling large amounts of text, which is necessary for evaluation. Gemini 2.5 Pro (preview) supports 1 million tokens of input and 64,000 tokens of output (Google AI, 2025a).

The Gemini 2.5 Pro (preview) was not an optimal choice to serve as the judge in this evaluation, as it contains both Gemini 2.5 Pro (preview) and Gemini 2.5 Flash (preview) AI tools. This will have an impact on the AI-based evaluation results, as AI tools exhibit intra-model bias when evaluating answers. AI judges prefer responses created with the same model (Verga *et al.*, 2024, p. 5). The other choices, xAI Grok 3 and Gemini 2.5 Flash (preview), were not any better choices, as those AI tools were part of the evaluation as well. In the end I decided not to remove Gemini 2.5 Pro (preview) from the evaluation, but the intra-model bias needs to be taken into account in the result analysis phase.

6.2.4 Results from Evaluation Workflow 1: Information Gathering

The first evaluation workflow was straightforward. In the first workflow, the AI was tasked with a suboptimal prompt to explain the meaning of "critical literature review" in the context of working on a master's-level Thesis, with a fictitious business informatics-related title. The full prompt for the evaluation workflow 1 was:

> "I'm doing my master's Thesis on "Big Data Analytics for Customer Churn Prediction" and starting my literature review. Please explain in detail what critical literature review means in this context."

The goal of this evaluation workflow was to assess how well different AI tools can handle short, suboptimal prompts and provide valuable, easy-to-understand explanations of

research-related topics. The categories for evaluation were the explanation of the term “critical” in this context, the purpose of critical review, the depth of explanation, structure and key elements of the answer, contextualization, coherence and cohesion, and factual accuracy of the answer.

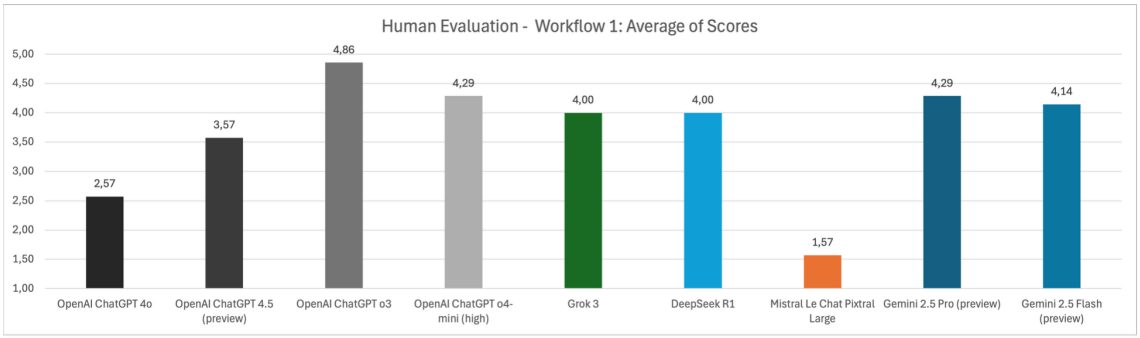


Figure 11: Workflow 1 - Numeric results of the human evaluation, showing the average score of different qualities for each AI tool. A higher number is better.

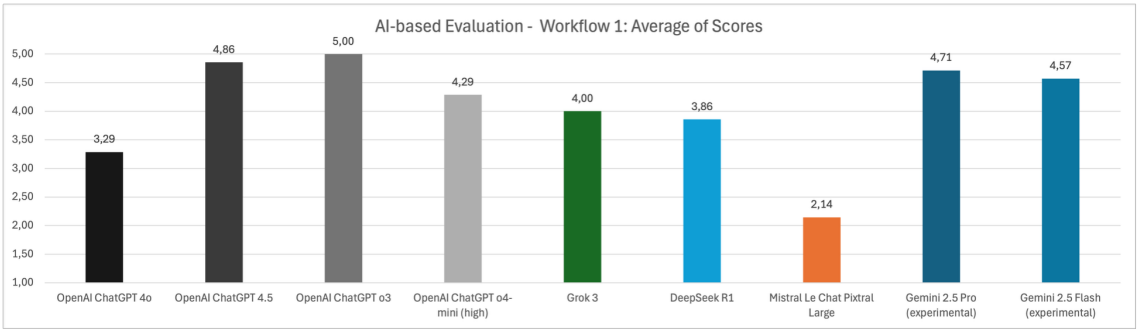


Figure 12: Workflow 1 - Numeric results of the AI-based evaluation, showing the average score of different qualities for each AI tool. A higher number is better.

The answers for the evaluation workflow were provided to the AI performing the evaluation within an evaluation prompt. The evaluation prompt includes instructions to provide both written feedback and a score of 1-5 for each of the specified qualities, with explanation of the scoring. The evaluation prompt for workflow 1 can be found in the appendix of this Thesis.

Looking at the total scores for this task, Pixtral Large and GPT-4o are again performing worse than the reasoning-powered models in both human and AI-based scoring. GPT-4.5, however, performs surprisingly well in AI-based evaluation in this workflow, even without the reasoning capabilities. OpenAI ChatGPT o3 performed the best in both human-based scoring and AI-based scoring.

The responses for Evaluation Workflow 1 were scored based on seven qualities: explanation of “critical”, Purpose of the critical literature review, depth of explanation, structure and key elements, contextualization, coherence and cohesion, and factual accuracy.

In the “Explanation of critical” and ”Depth of the explanation” categories GPT-4o and Pixtral Large struggled with similar issues. They were able to describe literature reviews on a surface level, but both tools failed to explain what makes a literature review “critical”, i.e. questioning assumptions, comparing different perspectives, and identifying potential biases (as described by: Saunders, Lewis and Thornhill, 2023, pp. 75–77). GPT-4o and Pixtral Large describe the evaluative part of the literature review but fail to explain the critical thinking part.

Overall, GPT-4o and Pixtral Large models were only capable of providing surface-level answers, and a student could get the same amount of information on this subject by skimming the titles of a related textbook. These models also generated output that’s borderline misleading, for example, this part of Pixtral Large output:

> The **primary purpose** of a critical literature review is to:
>
> 1. **Identify Key Studies**: Locate and evaluate the most relevant and influential studies in the field of big data analytics and customer churn prediction.
>
> 2. **Analyze Methodologies**: Examine the various methodologies, algorithms, and tools used in big data analytics for predicting customer churn.
>
> 3. **Evaluate Findings**: Assess the findings and conclusions of these studies to understand their contributions and limitations.
>
> 4. **Identify Gaps**: Highlight gaps in the existing literature that your research can address.
>
> 5. **Synthesize Knowledge**: Integrate the findings from different studies to provide a coherent overview of the current state of knowledge.

The mentioned tasks are parts of a critical literature review, and the explanations are factually correct (if based on: Saunders, Lewis and Thornhill, 2023, p. 78) those are not the primary purpose of a critical literature review.

The rest of the answers were similar, all providing good, valuable information in a concise package. They describe the key concepts of a critical literature review and its purpose, and the answers are easy to understand. Compared to GPT-4o and Pixtral Large, the rest of the models can provide much more in-depth answers, with actionable advice. GPT-4.5 (preview) deserves honourable mention for providing a good example of a critical evaluation:

> Example of Critical Perspective:
>
> Instead of simply writing: "Smith et al. (2020) found that XGBoost outperformed logistic regression in customer churn prediction."
>
> A critical review would state: "Although Smith et al. (2020) demonstrated the superior predictive accuracy of XGBoost models compared to logistic regression in churn prediction, their research was conducted only in a telecom setting, potentially limiting the generalizability of results. Furthermore, their study did not adequately address issues regarding model interpretability, which is crucial in real-world business applications where managers must understand the reasons behind churn."

Even though this example was not perfect, as the counterargument is missing a reference, it was still an effective way of explaining the concept of a critical literature review. DeepSeek R1 explained the concepts of a critical literature review with examples as well, but those examples were somewhat simpler.

Overall, the best response was produced by OpenAI ChatGPT o3, with Gemini 2.5 Pro (preview) offering a similar high-quality response. All of the tested AI tools were able to explain the meaning of a critical literature review, but the less-capable models struggled with providing a depth of explanation and contextualization. Especially models with reasoning capabilities were able to produce answers that could be valuable for students learning about critical literature review. No hallucinations were identified in the responses for this evaluation workflow.

6.2.5 Results from Evaluation Workflow 2: Academic Editing & Proofreading

The second evaluation workflow was developed to test the AI tools capability of performing academic editing and proofreading to a section of a literature review. For this evaluation workflow, a literature review was prepared by the thesis research, which contained obvious errors. This was achieved by gathering a piece of literature review from Theseus, feeding it to an AI tool, and asking it to introduce errors to the text.

Here is the full prompt for this evaluation workflow. The “literature review” part in this prompt is generated using Gemini 2.5 Pro (preview):

> I’m working on my master’s Thesis with the topic ”A Plan for New Customer Acquisition”, and I’m studying in Metropolia University of Applied Sciences, more specifically at Metropolia Business School.
>
> Your task is to perform academic editing and proofreading for a part of my literature review
>
> ”Literature contain several defintions for outsourcing. Outsourcing can be considered as discontinuation of companys internal activities, which are substituted by external provision. Outsourcing can be also defined as an activity, where company produces products or services out-side the companys boundaries in search of business competitivnes. (Van Weele 2014; Insinga and Werle 2000, Gilley and Rasheed 2000)
>
> As a notion outsourcing contains several activitys within it. Outsourcing contains an aspect of decision-making, and has been known as “make-or-buy” decision or as a decision of companys level of vertical integration. (Arnold 2000; Quinn and Hilmer 1994; Holcomb & Hitt 2007, Slack, Lewis 2008). In outsourcing, by set of complex choices companys decide what to perform internally and what to source from the marketplace, and by doing so determine the boundaries and scope of the company. Outsourcing also includes a process of transision. In this transision functions that was previously kept in-house are transferred to be performed outside the company (Kakabadse and Kakabadse 2000; Ellram, Billington 2001). These functions include specific assets that can be equipment knowledge people or anyother required factors of production that is transferred to the outside supplier. Also, decision rights to manage these factors are transferred (Greaver 1999: Van Weele 2014). However Gilley and Rasheed (2000) state that outsourcing does not necessarily require transfer of assets, but is defined by the capabilities of the outsourcing company. In cases where a company have capabilities to produce activitys by itself but decides to buy them from the markets, company is outsourcing. When company buys from the markets because it does not possess required capabilities, it is conducting procurement
>
> At least two party's exist in any outsourcing arangemant. Client is the party that contracts out selected activities and provider is the party that provides agreed activities for agreed amount of time to the client (Kern & Willcocks, 2000). Outsourcing creates a relationship between the outsourcing parties, where both parties seek for rewarding exchange (Moore 1998). Relationships are unique to each arrangement, and in them benefits risks and information are shared in different quantities depending on the nature of the relationship. Nature of the relationship is defined by the scope and criticality of the outsourced functions, as well as objectives of the outsourcing initiative (Gardner et. al 1999; Sanders et al, 2007). In management perspective outsourcing requires management of the client - supplier relationship that aims for adding value in cooperation (Hätönen and Eriksson, 2009).
>
> Outsourcing relationships create interdependancies between the parties. In addition to transfer of specific assets, client transfers decision rights over these assets to the supplier and thereby lose direct control over them. Provider takes over assets and depends on receiving compensation from the provided services, Therefore outsourcing can change both the client’s and the supplier’s cost and risk

profiles drastically. To avoid risks, outsourcing relationships are commonly governed by contracts, by forming a contract, both parties mitigate for external and internal uncertainties, as well as risks of opportunistic behavoir of the other party (Ellram and Billington 2001, Van Weele 2014).

Several possible benefits exist that companys seek to achieve by outsourcing. Also, several possible risks prevail in outsourcing activities. Drivers for outsourcing can be divided to two catagories financial and strategic (Vagadia 2012; Sanders et al. 2007). Financial reasons relate to lowering operational costs of a company. transaction cost theory is commonly used to explain economic benefits of outsourcing (Hätönen and Eriksson 2009; Ellram and Billington 2001; McIvor 2000). transaction cost theory considers the economic impacts of performing activities internally or buying them from the market. Transaction costs are affected by asset specificity, internal and external uncertainty and infrequency of transactions that lead to opportunistic behavior and increased transaction costs. Asset specificity relate to assets used in the transaction and there uniqueness to that specific transaction. High asset specificity increases the dependency between the contracting parties as well as increases the possibility of opportunistic behavior. According to Transaction cost theory, result of the transaction costs analysis should define the boundaries of the company. Other economic reasons for outsourcing include reduced fixed costs in internal functions, induced cash flow from sold assets and the need of balance sheet restructuring (Kakabadse & Kakabadse 2005; Vagadia, 2012).

Using economic benefits as a basis of outsourcing decision contains specific risks. Risk of hidden costs relates to those costs that company is unable to include to the cost calculations when making the decision. These costs can induce ex-post transaction cost that hinder the benefits of sought cost benefits Typically hidden cost relate to provider search and contracting cost as well as costs related to managing the relationship (Barthelemy 2003; Barthelemy and Quelin 2006). Solely concentrating on economic perspectives in outsourcing also risk avoiding the possible strategic level benefits of outsourcing. Economic benefits from cost related outsourcing can be short term and when focusing purely on costs, risk of outsourcing activities that have strategic value to the company in the long run can occur. Also economic measures can indicate symptoms of underperforming functions but reactive decisions based on these factors can lead to neglecting other possible development and solution methods that could be more beneficial (Barthelemy 2003; Sanders et al. 2007; McIvor 2000).

Strategic reasons to outsource integrate outsourcing activities to companys overall strategy process (Kakabadse and Kakabadse 2000) Instead of aiming for short term cost benefits outsourcing is used to align company’s scope of internal activities to its overall strategic objectives. In literature Core competence theory is widely used to explain rationale behind strategic outsourcing. According to core competence theory, to create true competitive edge, companys need to develop and maintain core competencies (Prahalad & Hamel, 1990). Core competencies are non-physical assets that are unique to a specific company and through which the company is able to create unique value to end customers. Core competencies allow companies to adapt and to re-align themselves to changing market requirements and sustain there competitive edge in the long run. According to core competence theory, all internal resources should be focused on those competencies while strategically outsourcing other non core activities. Outsourcing non core activities allow better managerial focus and resource allocation to most important activities (Quinn & Hilmer 1994; Gilley and Rasheed 2000).”

Required output:

1. Clean Edited Version: Provide the full text with corrections integrated (spelling, grammar, punctuation, typos).

2. Tracked Changes/Suggestions (Conceptual): If possible, indicate suggested rephrasing for clarity, flow, structure, and conciseness directly within the text (like track changes).

3. Numbered Comments List: Provide a separate, numbered list of comments that:

1. Reference the specific sentence or paragraph in the original text.

2. Explain the issue clearly (e.g., lack of clarity, weak transition, potential argument gap, tone issue).

3. Provide concrete suggestions for how to fix it.

4. Distinguish between mandatory corrections (grammar) and stylistic suggestions.

In this workflow, the low token limit of GPT-4o and GPT-4.5 (preview) proved to be problematic. When using these models through the ChatGPT web application, they have a token limit of approximately 4000 tokens, as previously discussed. This meant that GPT-4o and GPT-4.5 (preview) were not able to produce meaningful responses, and they were dropped out of this evaluation workflow. I could have split the workflow prompt into multiple pieces, but that might have given these tools an unfair advantage, and the results would not have been comparable to those of others.

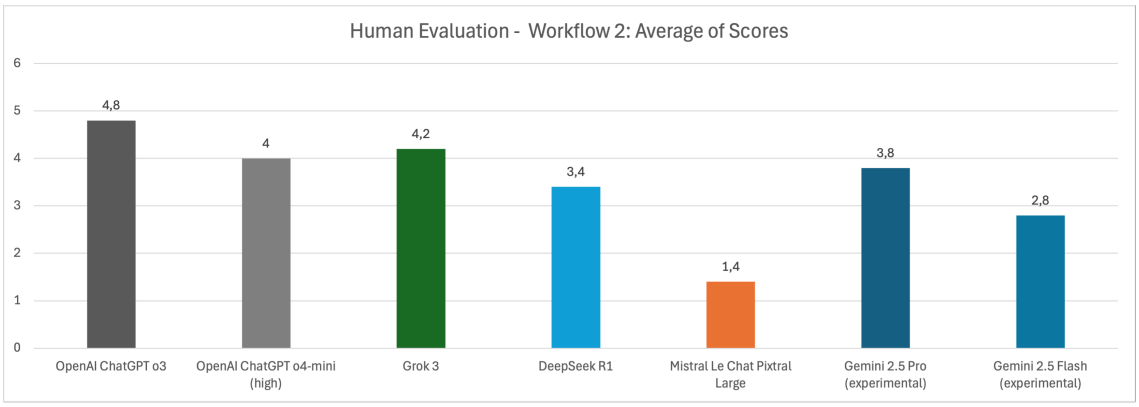


Figure 13: Workflow 2 - Numeric results of the human evaluation, showing the average score of different qualities for each AI tool. A higher number is better.

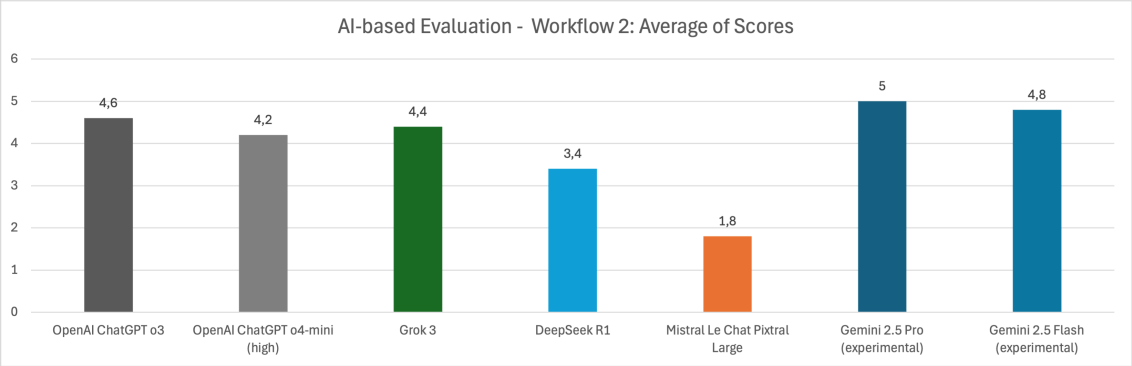


Figure 14: Workflow 2 - Numeric results of the AI-based evaluation, showing the average score of different qualities for each AI tool. A higher number is better.

In Evaluation Workflow 1, my interpretation of the quality of AI answers aligned quite well with the results from the AI-based evaluation, but that is not the case with Evaluation Workflow 2. All models with reasoning capabilities scored very well; however, in my opinion, the quality of the answers varied significantly.

The biggest difference between the answers was the text section with tracked changes. Grok 3 and DeepSeek R1 stated that they are unable to output the tracked changes in the requested format. However, these tools attempted to produce the tracked changes in an alternative format, which wasn't particularly usable. The formats of tracked changes varied significantly, and ChatGPT o3 presented the tracked changes in the most effective format, as determined by human evaluation.

Workflow 2 was especially challenging for human evaluation. The sheer amount of content was overwhelming, as the evaluation prompt was over 33 000 words, or 79 pages, in length. It was not feasible to manually evaluate the quality of all the grammar suggestions. Instead, the human evaluation focused on the improvement suggestions provided and the quality of the suggestions.

Almost all editing suggestions provided by the AI tools were at the sentence level. Most of the answers lacked suggestions for paragraph-level coherence and overall text structure. Suggestions for paragraph-level coherence and overall text structure would likely be included in the answers if these were explicitly requested in the prompt.

Gemini 2.5 Pro (preview), ChatGPT o3, and ChatGPT o4-mini produced easy-to-read responses that included actionable conceptual suggestions and tracked changes.

6.2.6 Results from Evaluation Workflow 3: Identifying Knowledge Gaps and Academic Text Generation

The third evaluation workflow was created to measure the AI tools' capability to identify possible academic knowledge gaps in academic text and fill these gaps by generating an academic text with references. For this workflow, a section from a published Master's Thesis literature review was selected and paragraphs were removed from that section to create artificial knowledge gaps in the remaining text. The original complete literature review section served as the ground truth in the evaluation.

Here is the full prompt for this evaluation workflow:

> I'm working on my master's Thesis: A Plan for New Customer Acquisition at Metropolia University of Applied Sciences. I have a short excerpt of my literature review below. You are my research assistant, and here is your step-by-step task:
>
> 1. Identify any knowledge gaps (without modifying the text).
>
> 2. Draft new content to fill in these gaps, using relevant, peer-reviewed scientific references (in APA style or another standard academic format)
>
> 3. Insert this new content into the text exactly where needed, keeping the original text intact.
>
> 4. Return the entire updated text with inserted content.
>
> Important:
>
> * Do not remove or modify existing text.
>
> * If you do not return the entire original text with your inserted content, your answer is invalid.
>
> * Provide the final updated text as follows:
>
> ### FULL UPDATED TEXT BEGINS ###
>
> <Paste original text with new content inserted>
>
> ### FULL UPDATED TEXT ENDS ###
>
> Original Text: "Knowing who are and who could be customers is crucial for any business. Understanding the target group's needs and desires helps the company to fulfil them. This gives the best knowledge of tailoring marketing strategies and

messaging accordingly. Next, the company should define a compelling value proposition which is far apart from competitors' propositions. It should attract customers and be visible and situated in spectacular location in all used media like the company's own landing page." (HubSpot, 2023.) According to HubSpot (2023), all marketing messages should be designed using these effective lead generation strategies. Every used Social Media Channel and the website and landing pages should be optimized for the pleasant and successful customer journey. (HubSpot, 2023.)

According to the classics of value creation (Grönroos and Voima 2012; Vargo and Lusch 2004, etc.), salespeople who practice value-based selling help their customers attain their fundamental exchange goal, by working together with the customer to facilitate creation of their value-in-use and business profitability. Therefore, value-based selling facilitates superior customer goal attainment and long-term satisfaction (Terho et al. 2012). In addition, value-based selling enables salespeople to communicate what the supplier's offerings are actually worth in customers' usage situations in monetary terms, which should help the customer buying center make its purchase decision". (Anderson, Narus, and Narayandas 2009; Ulaga and Eggert 2006; Terho, Ulaga, Eggert, Haas, Boehm, 2017.) The success or failure of a company's business model depends largely on how it interacts with those of the other players in the industry. Almost any business model will perform brilliantly if a company is lucky enough to be the only one in the market. (Casadesus-Masanell, Ricard, 2011.) Moreover, Osterwalder et al. 2010 believe a business model can best be described through nine basic building blocks that show the logic of how a company intends to make money. The nine blocks cover the four main areas of a business: customers, offer, infrastructure, and financial viability. The business model is like a blueprint for a strategy to be implemented through organizational structures, processes, and systems. The Value Propositions Building Block describes the bundle of products and services that create value for a specific Customer Segment. (Osterwalder et al. 2010.) Figure 20 shows the Business Model Canvas with the Value Proposition Block. 56 Figure 20. Business Model Canvas (Osterwalder, 2010). Figure 20 represents the Business Model Canvas and its Value Propositions Building Block as one out of nine blocks. According to Osterwalder et al. 2010, the Value Proposition is the reason why customers turn to one company over another. It solves a customer problem or satisfies a customer's need. Each Value Proposition consists of a selected bundle of products and/or services that caters to the requirements of a specific Customer segment. In this sense, the Value Proposition is an aggregation, or bundle, of benefits that a company offers to its customers. (Osterwalder et al. 2010.) "

Please confirm you understand by restating these instructions in your own words, then proceed with steps 1, 2, 3, and 4.

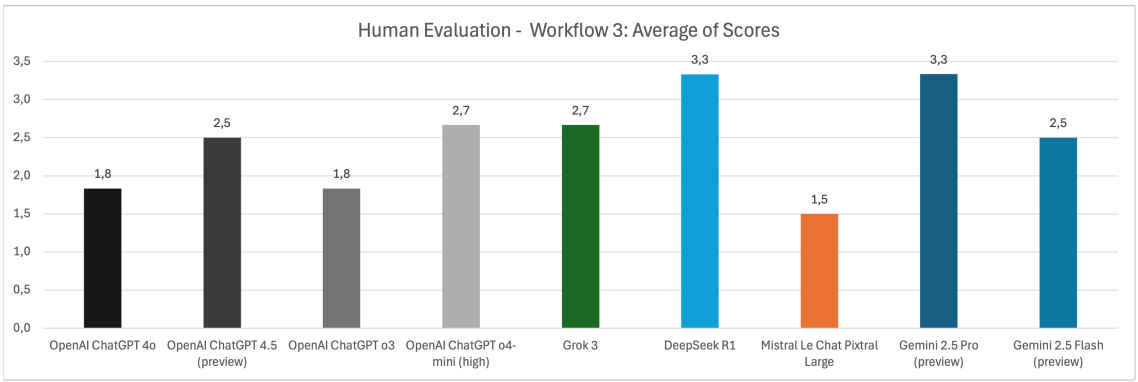


Figure 15: Workflow 3 - Numeric results of the human evaluation, showing the average score of different qualities for each AI tool. A higher number is better.

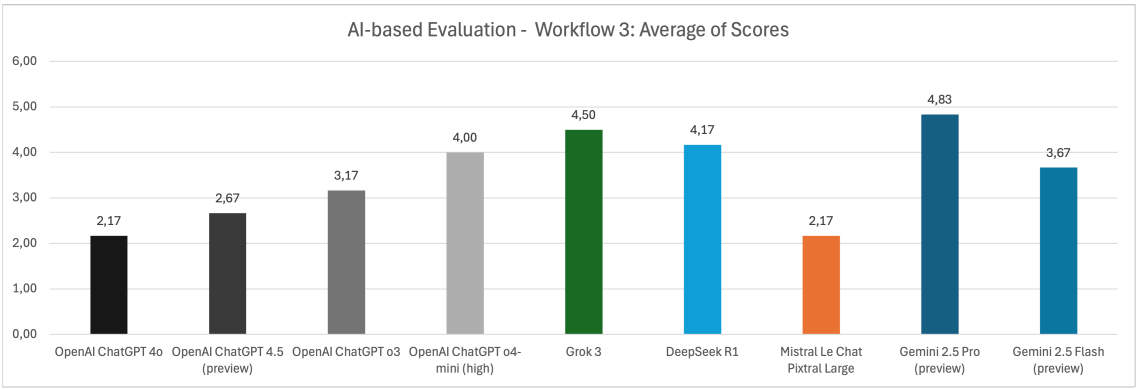


Figure 16: Workflow 3 - Numeric results of the AI-based evaluation, showing the average score of different qualities for each AI tool. A higher number is better.

Workflow 3 yielded interesting results, which differed significantly from those of the previous evaluation. ChatGPT o3, which had been performing quite well, produced poor results in this evaluation workflow. From a human evaluation perspective, excluding the issue with references, the biggest problem with the o3 answer was that, although the concepts introduced in the added text were relevant in the context of the original text, the style of the text was jarringly different from the original. In human evaluation, DeepSeek R1 and Gemini 2.5 Pro gained the best results, and the responses were quite similar. With the DeepSeek R1 and Gemini 2.5 Pro responses, the inserted text pieces cover relevant topics to the original text, match the style of the original text, and the flow between the original and inserted text is natural.

The evaluation workflow 3 was evaluated based on 6 qualities: instruction following, factual accuracy, scientific rigor, references, coherence and cohesion, and stylistic consistency.

The biggest problem identified in workflow 3 was the hallucinations with references. Mistral Pixtral Large did not include any references at all, even though those were specifically requested. All tools in this evaluation that inserted references into the text made mistakes with the references. Most commonly, errors occurred in the author lists, and the AI tools referenced papers that didn't exist.

Sadly, but not surprisingly, the AI judge was unable to identify these errors with the references. Even though references was one of the evaluated qualities, the AI judge was giving high scores to the answers with hallucinated references. In the same manner as the AI tools are not able to generate references accurately, the AI judge does not have a method to verify these references accurately.

Ultimately, none of the evaluated AI tools produced satisfactory responses for this workflow, even though there was significant variance in the quality of the responses.

6.2.7 Analysing the Evaluation Results

Analysing the evaluation results started by collecting the results from industry-standard benchmarks from artificialanalysis.ai. The benchmarks in this collection are MMLU-Pro, GPQA Diamond, Humanity's Last Exam, and MATH-500. These benchmarks were selected from a larger set of AI benchmarks, as they are the most relevant for MBS Primarily, benchmarks related to solving coding problems were excluded.

These industry-standard benchmarks are a suitable way to assess how different AI tools compare with each other quickly, but they are not without their issues and are only suitable for surface-level evaluation. As previously noted in research (Yubo Wang et al., 2024; Cohen-Inger et al., 2025), and based on data collected from two interviews, AI tools, such as ChatGPT, have been optimized and fine-tuned to perform well in these industry-standard benchmarks. These benchmarks may not accurately reflect how well these large language models will perform in specific workflows (Liang et al., 2023, pp. 156–157). Even with their issues, these benchmarks do offer value, and they work as a

quick and easy way to evaluate one model against another (Yubo Wang et al., 2024, p. 2)

MMLU-Pro is a more complex version of the original Massive Multitask Language Understanding benchmark, focusing on assessing reasoning across diverse academic and professional domains (Yubo Wang *et al.*, 2024, p. 1). GPQA Diamond is a benchmark that focuses on graduate-level scientific “Google-proof” questions, which don’t have straightforward answers, from the fields of biology, physics, and chemistry (Rein *et al.*, 2023, p. 1).

Many of the benchmarks designed to test the capabilities of large language models are quickly saturated, meaning that almost all models score very highly in these tests, rendering them redundant as a result. Humanity’s Last Exam is the exception here, as it was explicitly designed to mitigate this issue (Phan et al., 2025, p. 1).

> High accuracy on HLE (Humanity’s Last Exam) would demonstrate expert-level performance on closed-ended, verifiable questions and cutting-edge scientific knowledge, but it would not alone suggest autonomous research capabilities or “artificial general intelligence.” HLE tests structured academic problems rather than open-ended research or creative problem-solving abilities, making it a focused measure of technical knowledge and reasoning. HLE may be the last academic exam we need to give to models, but it is far from the last benchmark for AI. (Phan et al., 2025, p. 8)

The final benchmark in this collection is MATH-500. The questions in the MATH-500 benchmark are designed to test an AI tool’s performance in solving varying complex mathematical problems. The benchmark uses a smaller subset of questions from the MATH dataset, which OpenAI has developed, described in their ”Let’s Verify Step by Step” research paper (Lightman et al., 2023, p. 3)

This section will analyse the absolute results from these benchmarks and examine the normalized results.

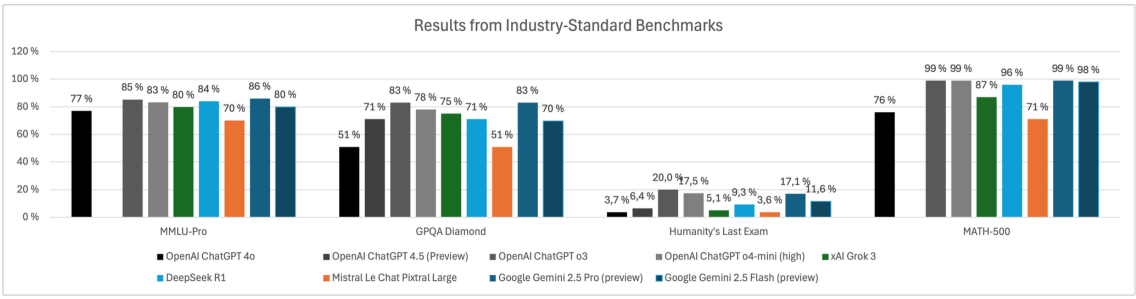


Figure 17: Absolute Results from Industry-Standard Benchmarks. A higher value is better.

Examining the absolute benchmark results reveals several key insights. GPT-4o and Pixtral Large perform quite poorly in all benchmarks, with a noticeable performance gap observed when compared to the reasoning models. The GPT-4.5 (preview) shows promising results in all the benchmarks it's included in, even surpassing xAI Grok 3 in HLE. This is astonishing, considering that GPT-4.5 (preview) lacks reasoning capability.

OpenAI GPT-4.5 is missing MMLU-Pro and MATH-500 results, most likely because the model was recently replaced by GPT-4.1 (OpenAI, 2025d). Overall, these industry-standard benchmarks shouldn't be the only metric when considering the suitability of an AI tool for any task.

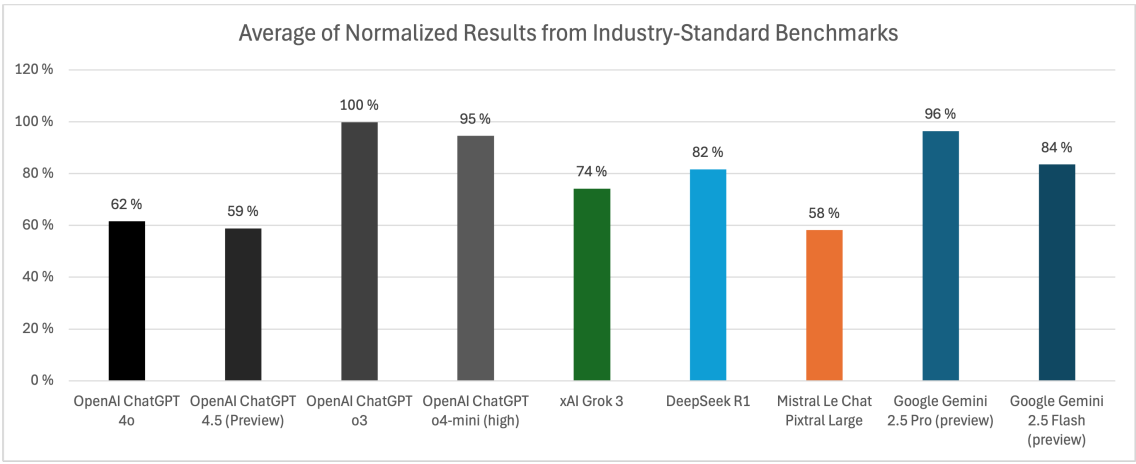


Figure 18: Average of Normalized Results from Industry-Standard benchmarks. Higher value is better.

As shown in Figure 17, the average scores from HLE were significantly lower than those from the other results. The average of the scores from HLE was 9,0 %, and the average of all other results was 79,9 %. Because of this, the results had to be normalized first, allowing us to compare the numbers with the normalized results.

Even with the normalized results, OpenAI GPT-4.5 (preview) results are not comparable with the other results, as GPT-4.5 is missing results from MMLU-Pro and MATH-500.

Figure 18 shows that Pixtral Large and GPT-4o were the outliers amongst the competition, scoring significantly lower in the benchmarks. OpenAI ChatGPT o3 scored a perfect 100% in the normalized benchmarks, indicating that it achieved the highest scores in all the included benchmarks.

Next step in the analysis of results was collecting the features as specifications for the AI tools. These are presented in the following tables.

Table 4: AI tool features.

AI tool	Image Processing	Image Generation	Voice Mode	Reasoning capabilities	Web Search	Deep Research
ChatGPT GPT-4o	Supported	Supported	Supported	Not supported	Supported	Supported
ChatGPT GPT-4.5 (preview)	Supported	Supported	Not supported	Not supported	Supported	Supported
ChatGPT o3	Supported	Not supported	Supported	Supported	Supported	Supported
ChatGPT o4-mini (high)	Supported	Not supported	Supported	Supported	Supported	Supported
Grok 3	Supported	Supported	Supported	Supported	Supported	Supported
DeepSeek R1	Supported	Supported	Not supported	Supported	Supported	Not supported
Mistral Pixtral Large	Supported	Not supported	Not supported	Not supported	Not supported	Not supported
Gemini 2.5 Pro (preview)	Supported	Supported	Not supported	Supported	Supported	Supported
Gemini 2.5 Flash (preview)	Supported	Supported	Supported	Supported	Supported	Supported

The above table presents the which evaluated AI tools support which key features. The main differentiation factor for this evaluation is the lack of reasoning capabilities on ChatGPT GPT-4o, GPT-4.5 (preview) and Mistral Pixtral Large.

Table 5: AI tool specifications.

AI tool	Token limit (AI tool)	Output token limit (API)	Context Window Size	Knowledge cutoff
ChatGPT GPT-4o	~4000	16 384	128 000	October 2023

ChatGPT GPT-4.5 (preview)	~4000	16 384	128 000	October 2023
ChatGPT o3	N/A	100 000	200 000	June 2024
ChatGPT o4-mini (high)	N/A	100 000	200 000	June 2024
Grok 3	N/A	N/A	131 072	November 2024
DeepSeek R1	N/A	8000	64 000	N/A
Mistral Pixtral Large	~4000	4096	128 000	Not supported
Gemini 2.5 Pro (preview)	1 000 000	65,536	1 000 000	January 2025
Gemini 2.5 Flash (preview)	1 000 000	65,536	1 000 000	January 2025

The technical specifications are challenging to discover for the AI tools. In most cases, AI service developers provide documentation only for the LLM APIs. However, these specifications may differ from what is available in the AI tool utilizing the same LLM. These specifications are not particularly valuable on their own, but they can offer insights into why certain models perform poorly in specific tasks. As an example, the GPT-4o and GPT-4.5 (preview) AI tools were unable to complete evaluation workflow 2 due to their low token limits. Some of the specifications are not available, and the token limits for ChatGPT GPT-4o, GPT-4.5 (preview) and Mistral Pixtral Large are estimation based on testing.

With the results from industry-standard benchmarks, collected features and specifications, and both human evaluation and AI-based evaluation results, the suitability of AI tools for academic writing purposes can be determined, as described by the three evaluation workflows in this evaluation.

Almost all the evaluated AI tools produced suitable results in the Workflow 1: “Information gathering with sub-optimal prompt”, except for Mistral Pixtral Large. This suggests that AI tools can be utilized to gain an initial understanding of topics relevant to academic writing in the MBS context.

The evaluation workflow 2: “Academic editing and proofreading” showcased that not all AI tools are suitable for all workflows. GPT-4o, GPT-4.5 (preview) and Pixtral Large were unable to produce responses for this workflow due to their low token limits. DeepSeek R1 and Grok 3 struggled with the formatting of tracked changes, and overall, OpenAI o3 and Gemini 2.5 Pro (preview) produced the best answers in workflow 2.

In workflow 3: “Identifying knowledge gaps & academic text generation” the limits of the current AI tools were identified. None of the AI tools were able to produce suitable results, as all the responses contained hallucinations with references. However, most of the AI tools were able to identify relevant knowledge gaps and produce sound textual content. The best AI tools stood out with their ability to match the original tone of the provided text and seamlessly integrate the added paragraphs with the original text. DeepSeek R1 and Gemini 2.5 Pro (preview) produced the best answers in workflow 3.

In this evaluation, Gemini 2.5 Pro and ChatGPT o3 are the top performers. Especially when examining the numerical results, the intra-modal bias with Gemini 2.5 Pro must be taken into consideration.

When considering recommending these tools for students, pricing is a crucial factor. Google offers limited use with Gemini 2.5 Pro for free, and ChatGPT o3 is part of their paid “plus” membership. Out of the tools in this evaluation, only ChatGPT o3 and o4-mini require a paid subscription. Based on the results of this evaluation, I would suggest that students avoid using ChatGPT GPT -4, as there are many better free alternatives available.

Table 6: Correlation between the industry-standard benchmark results, numeric human evaluation results and numeric AI-based evaluation results.

Correlation	Average of normalized industry-standard benchmark results	Average of normalized domain-specific AI-based evaluation results	Average of normalized domain-specific human evaluation results
Average of normalized industry-standard benchmark results	1	0,824	0,823
Average of normalized domain-specific AI-based evaluation results	0,824	1	0,929
Average of normalized domain-specific human evaluation results	0,823	0,929	1

To assess the alignment of the results from domain-specific evaluations with industry-standard benchmarks, the correlation between these results was calculated. To do this, the average of all industry-standard benchmarks for each AI tool was used, as well as the average of each of the three evaluation workflows for each AI tool. As the industry-

standard benchmarks report their scores on a 0-100 scale, and the 1-5 scale was used in this evaluation. First, the results were normalized to a 0-1 scale.

Correlation measures the strength of a linear relationship between two sets of scores. In this case, high correlation means that AI tools that received poor scores from industry-standard benchmarks also received poor scores from domain-specific evaluations. While there isn't an official categorization for correlation coefficient, a correlation above 0,9 is considered to indicate a very strong relationship between values (Schober, Boer and Schwarte, 2018, p. 3) Table 7 shows that the human evaluation results correlate strongly with AI-based results (0,929 correlation coefficient), and the industry-standard benchmarks results correlate relatively strongly as well with the results from domain specific evaluation.

6.3 Developments to the Proposal (based on Data Collection 3)

The initial proposal was presented to a key stakeholder from the MBS AI team, and feedback was gathered for further development. This feedback, together with insights gained during the implementation of the evaluation, using the AI tool evaluation framework, is used to refine each of the evaluation phases.

6.3.1 Refined Guidelines for Phase 1: Preparation & Evaluation Design

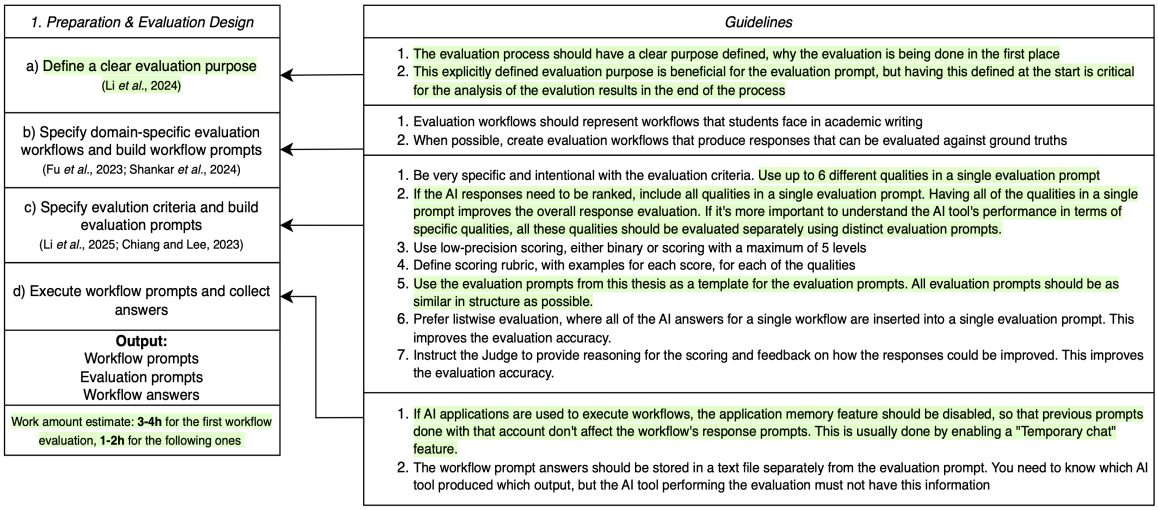


Figure 19: Refined preparation and evaluation design phase of the AI tool evaluation framework with guidelines.

One critical improvement suggestion gained from key stakeholder was that each of the evaluation phases should have work amount estimates.

> A workload should make a sort of projection of how much time you think each stage can take from your perspective, and based on your own evaluation, which was done of course for the first time, but now you created the guidelines how much time it takes so that we can resource this work (Stakeholder H)

Work amount estimates were added to the AI tool evaluation framework guidelines for each step. These are estimates, reflecting the time it took me to complete the evaluation process.

During the implementation of the evaluation framework, it became clear that the evaluation purpose should be defined explicitly and clearly as the first step in the evaluation process. The evaluation purpose should clearly explain why the evaluation is being conducted, which aids in building the evaluation prompt and informs the analysis of evaluation results.

When building and testing the evaluation prompts, the most consistent results were obtained when all the AI responses were inserted into a single evaluation prompt, facilitating a listwise evaluation. With listwise evaluation, the AI judge had multiple responses to compare against each other in the evaluation, and this enabled the AI judge to pick up nuances in the AI responses that it missed when the AI responses were inserted to evaluation prompts separately. The evaluation prompts used in this evaluation are crafted with care, utilizing prompt engineering best practises, such as role setting, and these evaluation prompts produce consistent evaluations. These evaluation prompts can be used as templates in future evaluations.

When evaluating AI tools, a “Temporary Chat” feature should be used, when available. With most AI tools, this blocks the AI tool from accessing previous discussions with the tool, which could influence the evaluation.

6.3.2 Refined Guidelines for Phase 2: Human Evaluation

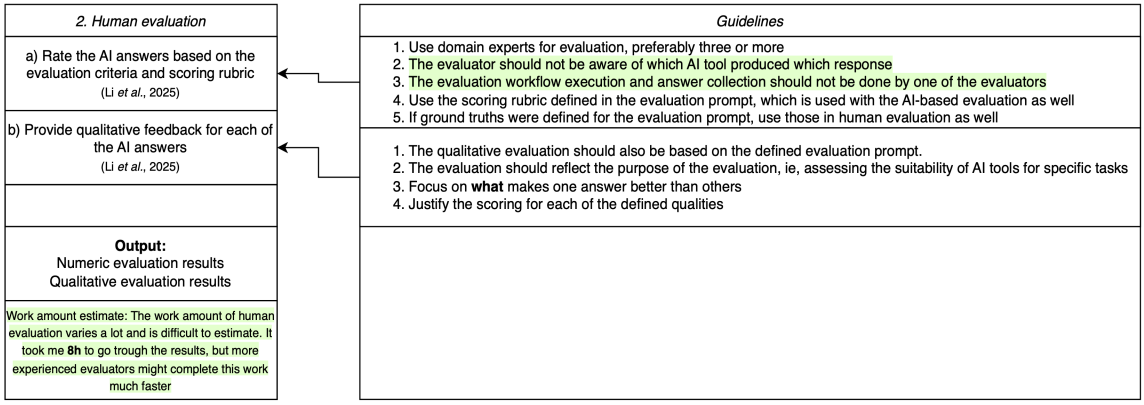


Figure 20: Refined human evaluation phase of the AI tool evaluation framework with guidelines.

As a single person (i.e. the thesis researcher) implemented the AI tools evaluation, this caused issues, particularly in the human evaluation phase. As the thesis researcher knew which AI tool had produced which answer, he had a strong bias when evaluating the responses. Noticeably, the thesis researcher preferred the responses created by higher-competence AI tools. Even though he hadn't analysed the industry-standard benchmark results at this point, the thesis researcher was still fully aware, for example, that Gemini 2.5 Pro should produce better responses than Gemini 2.5 Flash.

To mitigate this bias in the future, the human evaluators should not be aware of which tool produced which result. The AI responses should be collected by someone who is not evaluating the AI responses.

The key stakeholder from Metropolia had high confidence that Metropolia personnel can perform the human evaluation phase with high efficiency.

> We understand the challenges of human evaluation, but it doesn't mean that it's not possible to do it. To do it compactly, this is what we're doing all the time. (Stakeholder H)

6.3.3 Refined Guidelines for Phase 3: AI-based Evaluation

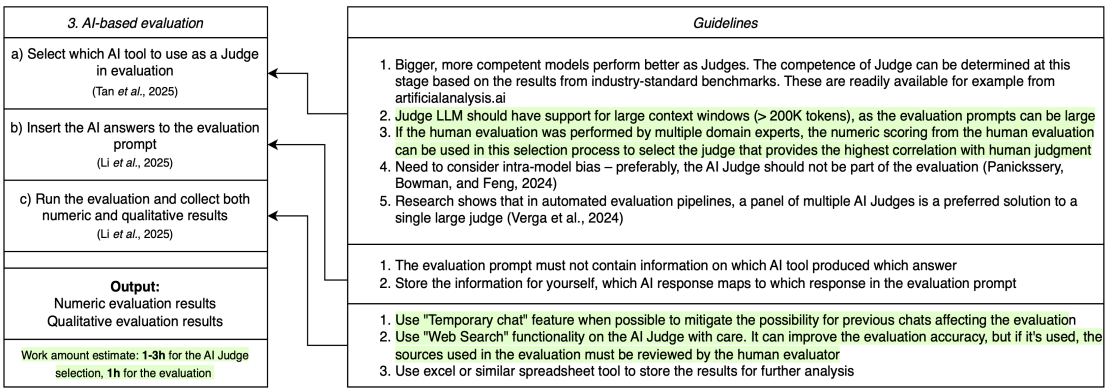


Figure 21: Refined AI-based phase of the AI tool evaluation framework with guidelines.

While implementing phase 3: AI-based evaluation, the significance of using an AI judge with large context windows became clear. Workflow 2: Academic editing and proofreading had an evaluation prompt that exceeded 50 000 tokens in length. When we consider that the reasoning process creates tokens stored in the same context window, AI tools with context windows smaller than 200 000 tokens are not suitable choices to act as judges. The evaluation prompt itself could easily exceed the 200 000 token limit if the evaluation workflow responses contained longer paragraphs of text.

If the human evaluation was conducted with a panel of domain experts, the numeric scoring from the human evaluation can be used in the AI judge selection process to select an AI tool that best aligns with human judgment.

If an AI tool is used to perform the evaluation, instead of calling a LLM API directly, a “Temporary Chat” feature should be used to prevent previous discussions with the AI tool from affecting the evaluation results. Most AI tools include a “Web Search” feature, which enables them to retrieve up-to-date information from search engines to improve their responses. A “Web Search” feature could be beneficial in an evaluation that requires up-to-date knowledge, but it should be used with care, and a human should verify the sources that it applies to the evaluation.

6.3.4 Refined Guidelines for Phase 4: Analysis of Evaluation Results

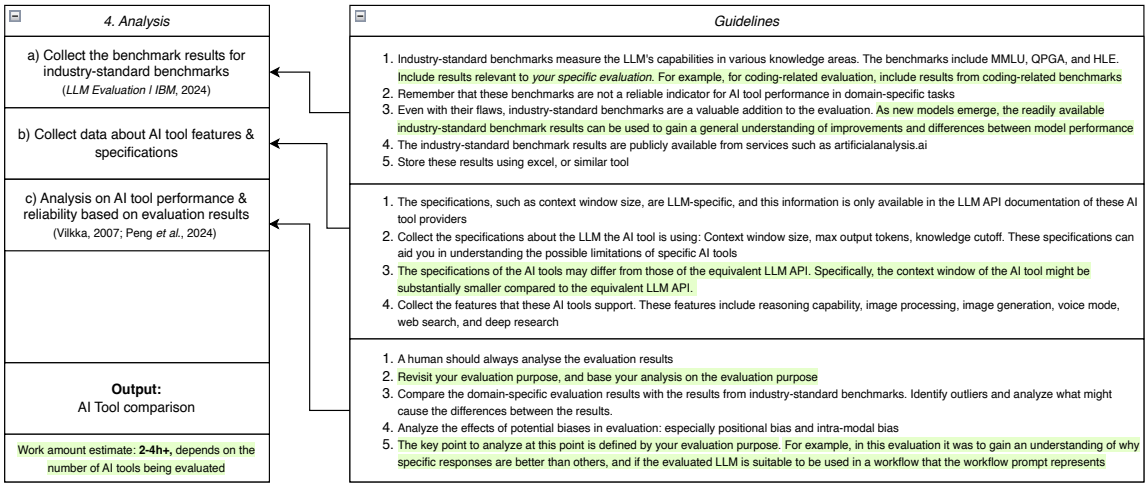


Figure 22: Refined AI-based phase of the AI tool evaluation framework with guidelines.

When collecting results from industry-standard benchmarks, only benchmarks relevant to the specific domain should be included in the evaluation. In the case of MBS and the evaluation implementation in this Thesis, coding-related industry-standard benchmarks were excluded from the evaluation.

Collecting the specifications for the AI tools can be problematic. Specifications for the LLM APIs are available from all providers, but these specifications may differ from those provided in the AI tool version of the same LLM. For example, the LLM API for the OpenAI model GPT-4o has a maximum output limit of 16,384 tokens. However, based on testing, when the same model is used in the ChatGPT AI tool, it has a substantially lower limit for maximum output tokens, and official documentation does not mention this. The total token limit for GPT-4o in the ChatGPT AI tool appears to be approximately 4,000 tokens, including input and output tokens.

When analysing the evaluation results, it was essential to focus on the evaluation purpose. For the evaluation in this Thesis, the evaluation purpose was to determine the suitability of the selected AI tools for academic writing purposes, and the analysis was conducted from this perspective.

6.4 Final Proposal

Based on the key stakeholder feedback and insights gained during the implementation, the final proposal for AI Tool Evaluation Framework was built.

The details of improvements brought to the AI Tool Evaluation Framework are described in the previous sub-sections. With the improvements suggested by the stakeholders, this final proposal provides detailed guidelines which should lead the evaluator to reliable evaluation results.

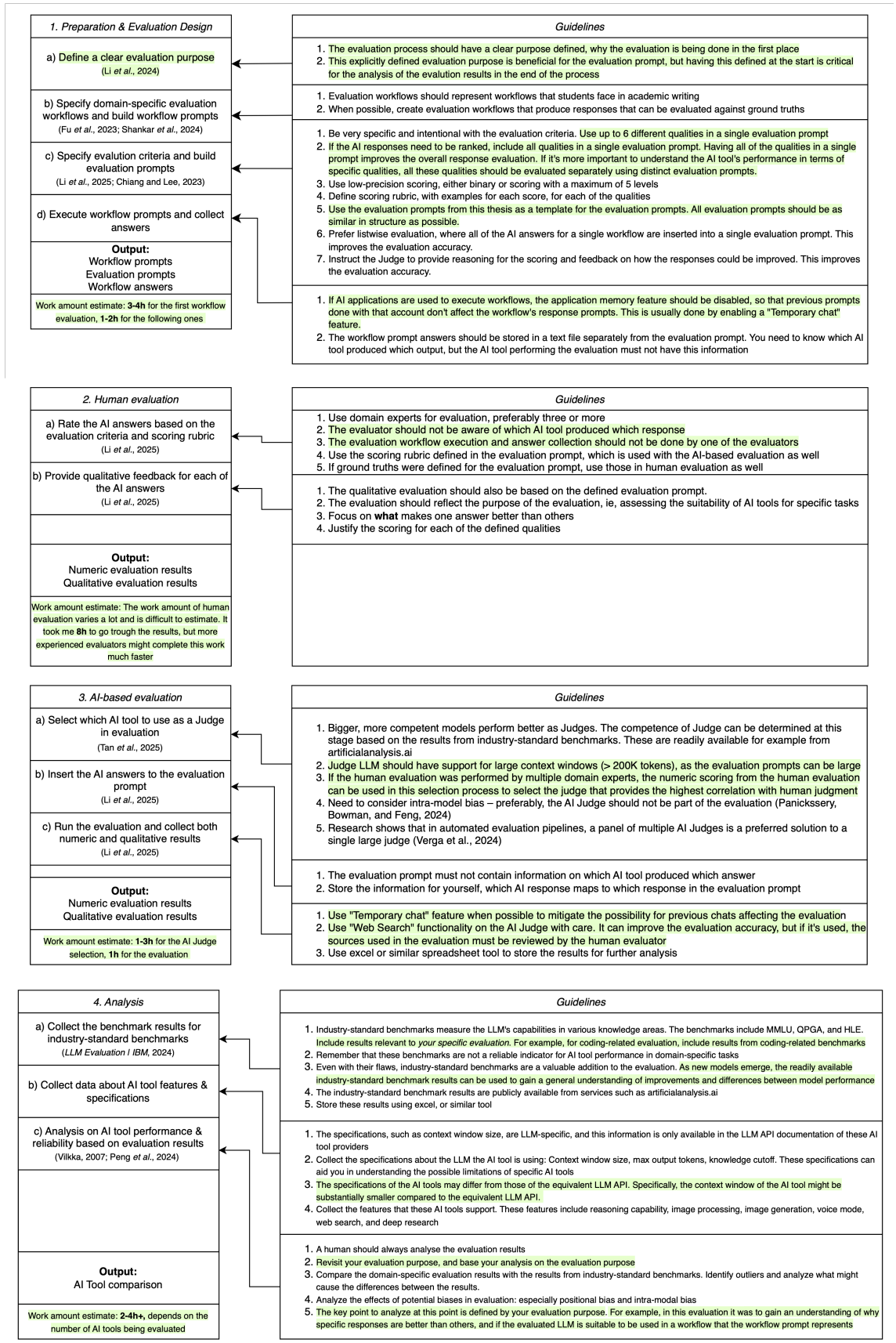


Figure 23: Refined AI Tool Evaluation Framework, providing explicit guidelines for each of the evaluation phases.

7 Conclusion

This section contains the executive summary of this Thesis, highlighting its key findings together with suggestions for the next steps and the evaluation of the Thesis.

7.1 Executive Summary

As the use of generative AI increases in higher education, there is a pressing need to understand the suitability of specific AI tools for research and development purposes. The AI tools are evolving at a rapid speed, and state-of-the-art tools from just a year ago can be close to obsolete today. To address this situation, a systematic process is needed to evaluate the suitability of AI tools for research and development purposes. The objective of this Thesis was to evaluate a selected set of AI tools to determine their suitability for academic writing purposes by students. To achieve this, an AI tool evaluation framework was developed, and the suitability of a selection of AI tools was evaluated by implementing this framework.

This Thesis followed a five-step research design, starting with the definition of the business challenge and objective, and up to building and implementing the proposal. This Thesis was done utilizing an applied action research strategy. Data collection relied on stakeholder interviews and analysis of internal documents conducted in three data collections rounds. First data collection round enabled the current state analysis, the second one building of the initial proposal for AI tool evaluation framework, and the final data collection round was the implementation of the AI tool evaluation framework and obtaining the evaluation results for spring 2025.

In the current state analysis, existing guidelines for using AI tools in Metropolia were analysed, and interviews were conducted with stakeholders from the Metropolia Business School AI team. The key findings included the facts that MBS doesn’t yet have AI tools to offer students, implementing new tools requires a thorough and time-consuming evaluation process, the process for doing so is unclear, and Metropolia lacks detailed instructions for using AI tools for academic writing purposes. Informed by these key findings, the literature review focused on three topics: the elements of generative AI, the accuracy and reliability of generative AI, and the evaluation of generative AI output. Based on the literature review, the AI tool evaluation framework was developed as the

initial proposal development. The proposal for the AI tool evaluation framework was implemented and also extended with explicit guidelines on how to repeat the evaluations in the future.

The AI tool evaluation framework outlines how the suitability of various AI tools can be assessed using domain-specific workflows. Results from industry-standard benchmarks for AI tools, such as MMLU, are readily available, but these do not always provide an accurate representation of how the AI tools will perform in domain-specific workflows. The AI tool evaluation framework consists of four phases: Preparation and Evaluation Design, human evaluation, AI-based evaluation, and analysis of results. In the preparation and evaluation design phase, the evaluation workflows are defined in conjunction with the evaluation criteria and prompts for AI-based evaluation. In the human evaluation phase, the AI responses are assessed based on the specified evaluation criteria. In the AI-based evaluation phase, the AI responses are evaluated utilizing an LLM-as-a-Judge methodology. In the final phase, analysis of results involves collecting data from industry-standard benchmarks, along with AI tool features and specifications, and analysing the results from both human and AI-based evaluations. Ultimately, recommendations on the suitability of AI tools are provided based on each of these factors.

The AI tool evaluation framework, along with guidelines, was implemented in the Metropolia Business School context to assess the suitability of the selected AI tools for academic writing purposes among students. The evaluation implementation was a cyclic process. Each of the steps in the phases was implemented multiple times, as shortcomings in the initial guidelines were identified. As mitigations for these shortcomings were developed, these insights were used to refine the AI tool evaluation framework in the process. The final proposal of AI tool evaluation framework with explicit guidelines was built based on the feedback from the key stakeholder, and insights gained during the implementation phase.

The results from the evaluation of the AI tool's suitability revealed that reasoning-powered models performed better in each of the evaluation workflows. OpenAI ChatGPT o3 and Google Gemini 2.5 Pro (preview) produced the overall best responses in the evaluation. Considering that Gemini 2.5 Pro (preview) is free to use, albeit with certain limitations on the free tier, Google Gemini 2.5 Pro (preview) was the preferred AI tool for this evaluation in academic writing purposes. However, there was a substantial gap in

the quality of responses when comparing the best and worst-performing AI tools. Mistral Pixtral Large was the worst-performing AI tool in this evaluation, but the performance of OpenAI ChatGPT-4o was almost as bad, which is one of the key findings from this evaluation, considering the popularity of ChatGPT-4o. ChatGPT-4o was the default model of the OpenAI ChatGPT service for a long while, but it has been recently replaced with another model from OpenAI, GPT-4.1 mini.

This represents a key struggle with this Thesis. The AI tool field is moving at such a rapid pace that these evaluation results are partly outdated before this work is even released. This highlights the importance of the AI tool evaluation framework, which enables frequent evaluation of available AI tools.

7.2 Next Steps and Relevance to the Organization

For the next steps with the AI evaluation framework, MBS plans to incorporate this framework into their AI tool selection process. In this context, it is important to draw attention to the identified areas that could be further improved with future evaluations.

First, the evaluation results presented in this Thesis are based on three evaluation workflows, which reflect the types of workflows that students might use AI tools for, based on the best available knowledge and my experiences with research and development work. In future evaluations, these evaluation workflows should be defined by a panel of domain experts from Metropolia Business School, so that the evaluation more accurately mimics the actual workflows that students face.

Second, the human evaluation phase was conducted by a single evaluator, and this is another area for improvement. Ideally, human evaluation should be performed by a panel of domain experts, and a separate person should collect the responses from the evaluated AI tools to ensure that evaluators are unaware of which AI tool produced each answer, thereby mitigating potential biases towards specific AI tools.

Based on the discussions with MBS key stakeholders, this Thesis work has a high relevance for the context organization. An AI tool evaluation framework with explicit guidelines for implementation addresses their needs, offering a systematic process with a reasonable workload for evaluating the suitability of AI tools for various workflows. The

evaluation results for spring 2025 highlight that AI tools are not suitable for all academic writing-related workflows and that substantial performance differences exist between the AI tools.

7.3 Thesis Evaluation

The initial objective of this Thesis was to evaluate a selection of AI tools for their suitability to be used in academic writing purposes by students. Reflecting on the initial objective, the Thesis had a stronger focus on the AI tool evaluation framework than expected, rather than the suitability evaluation results. This shift in focus was caused by the lack of a systematic evaluation process for AI tools, the rapid pace of AI tools development, and personal interests towards developing an AI tool evaluation framework utilizing domain-specific workflows.

The human evaluation phase of implementing the AI tool evaluation framework would have benefited from additional evaluators. With the evaluation implementation in this Thesis, a single person conducted the human evaluation, who had a relatively good understanding of the competence of the evaluated AI tools before initiating the human evaluation phase. This inevitably leads to some bias in the human evaluation phase.

When analysing the evaluation results, the results from the industry-standard benchmarks aligned quite well with the numeric scoring from domain-specific human evaluations and AI-based evaluations, with a correlation coefficient ranging from 0,823 to 0,929. This indicates a high level of agreement between the different evaluation methods. The absolute scores from evaluations may differ, but this shows that the underlying logic behind the evaluation is robust.

7.4 Closing Words

This Thesis explored the evaluation of AI tools suitability for domain-specific tasks, specifically focusing on academic writing purposes by students. The Thesis presented a systematic process for evaluating AI tools and provided recommendations for selecting AI tools for academic writing purposes, to be used by students in 2025.

The AI tool evaluation framework was designed to ensure that the required work amount is reasonable and that the provided guidelines are explicit enough so that the evaluator does not need to be an expert in Artificial Intelligence. I hope that this framework gets frequent use in Metropolia Business School and other schools in Metropolia as well.

References

Alasadi, E.A. and Baiz, C.R. (2023) 'Generative AI in Education and Research: Opportunities, Concerns, and Solutions', *Journal of Chemical Education*, 100(8), pp. 2965–2971. Available at: https://doi.org/10.1021/acs.jchemed.3c00323.

Amazon Web Services (2024a) *What are Transformers? - Transformers in Artificial Intelligence Explained - AWS*, *Amazon Web Services, Inc.* Available at: https://aws.amazon.com/what-is/transformers-in-artificial-intelligence/ (Accessed: 2 December 2024).

Amazon Web Services (2024b) *What is AGI? - Artificial General Intelligence Explained - AWS*, *Amazon Web Services, Inc.* Available at: https://aws.amazon.com/what-is/artificial-general-intelligence/ (Accessed: 27 November 2024).

Amazon Web Services (2024c) *What is Generative AI? - Gen AI Explained - AWS*, *Amazon Web Services, Inc.* Available at: https://aws.amazon.com/what-is/generative-ai/ (Accessed: 27 November 2024).

Amazon Web Services (2025) *What is LLM? - Large Language Models Explained - AWS*, *Amazon Web Services, Inc.* Available at: https://aws.amazon.com/what-is/large-language-model/ (Accessed: 24 February 2025).

Antrophic (2025) *Antrophic API - Release Notes*, *Anthropic*. Available at: https://docs.anthropic.com/en/release-notes/api (Accessed: 10 May 2025).

Artificial Analysis (2025a) *AI Model & API Providers Analysis | Artificial Analysis*. Available at: https://artificialanalysis.ai (Accessed: 27 May 2025).

Artificial Analysis (2025b) *Gemini 2.5 Pro Experimental - Intelligence, Performance & Price Analysis | Artificial Analysis*. Available at: https://artificialanalysis.ai/models/gemini-2-5-pro (Accessed: 5 April 2025).

ARTSMART AI (2024) *How Many Generative AI Tools Are There? A Comprehensive Guide*. Available at: https://artsmart.ai/blog/how-many-generative-ai-tools/ (Accessed: 12 April 2025).

Beaufays, F. (2024) *The neural networks behind Google Voice transcription*. Available at: http://research.google/blog/the-neural-networks-behind-google-voice-transcription/ (Accessed: 2 December 2024).

Bender, E.M. *et al.* (2021) 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜', in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency*, Virtual Event Canada: ACM, pp. 610–623. Available at: https://doi.org/10.1145/3442188.3445922.

Bentley, P.J., Gulbrandsen, M. and Kyvik, S. (2015) 'The relationship between basic and applied research in universities', *Higher Education*, 70(4), pp. 689–709. Available at: https://doi.org/10.1007/s10734-015-9861-2.

Bhavandla, L.K. (2025) 'Development of Secure API Gateways for Cloud Services', *ResearchGate* [Preprint]. Available at: https://doi.org/10.36676/j.sust.sol.v2.i1.53.

Bitloops (2025) *Layered Architecture: Building Scalable & Maintainable Software Systems | Bitloops Docs*. Available at: https://bitloops.com/docs/bitloops-language/learning/software-architecture/layered-architecture (Accessed: 13 May 2025).

Bittle, K. and El-Gayar, O. (2025) 'Generative AI and Academic Integrity in Higher Education: A Systematic Review and Research Agenda', *Information*, 16(4), p. 296. Available at: https://doi.org/10.3390/info16040296.

Bommasani, R. *et al.* (2022) ‘On the Opportunities and Risks of Foundation Models’. arXiv. Available at: https://doi.org/10.48550/arXiv.2108.07258.
Bratslavsky, P. (2025) *API Design 101: Best Practices & Implementation*. Available at: https://strapi.io/blog/api-design-101 (Accessed: 10 May 2025).
Bronsdon, C. (2024) *Top Methods for Effective AI Evaluation in Generative AI*, *Galileo AI*. Available at: https://www.galileo.ai/blog/top-methods-for-effective-ai-evaluation-in-generative-ai (Accessed: 5 March 2025).
Bronsdon, C. (2025) *The Definitive Guide to LLM Parameters and Model Evaluation*, *Galileo AI*. Available at: https://www.galileo.ai/blog/llm-parameters-model-evaluation (Accessed: 28 April 2025).
Brown, T.B. *et al.* (2020) ‘Language Models are Few-Shot Learners’. arXiv. Available at: https://doi.org/10.48550/arXiv.2005.14165.
Bucaioni, A. *et al.* (2025) ‘A Functional Software Reference Architecture for LLM-Integrated Systems’. arXiv. Available at: https://doi.org/10.48550/arXiv.2501.12904.
Cardona, M.A., Rodríguez, R.J. and Ishmael, K. (2023) ‘Artificial Intelligence and the Future of Teaching and Learning’.
Chan, C.-M. *et al.* (2023) ‘ChatEval: Towards Better LLM-based Evaluators through Multi-Agent Debate’. arXiv. Available at: https://doi.org/10.48550/arXiv.2308.07201.
Chiang, C.-H. and Lee, H. (2023) ‘Can Large Language Models Be an Alternative to Human Evaluations?’ arXiv. Available at: https://doi.org/10.48550/arXiv.2305.01937.
Chow, A.R. (2023) *How ChatGPT Managed to Grow Faster Than TikTok or Instagram*, *TIME*. Available at: https://time.com/6253615/chatgpt-fastest-growing/ (Accessed: 28 April 2025).
Cohen-Inger, N. *et al.* (2025) ‘Forget What You Know about LLMs Evaluations -- LLMs are Like a Chameleon’. arXiv. Available at: https://doi.org/10.48550/arXiv.2502.07445.
Cotton, D.R.E., Cotton, P.A. and Shipway, J.R. (2024) ‘Chatting and cheating: Ensuring academic integrity in the era of ChatGPT’, *Innovations in Education and Teaching International*, 61(2), pp. 228–239. Available at: https://doi.org/10.1080/14703297.2023.2190148.
Devlin, J. *et al.* (2019) ‘BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding’. arXiv. Available at: https://doi.org/10.48550/arXiv.1810.04805.
Elastic (2024) *What are Large Language Models?* | *A Comprehensive LLMs Guide*. Available at: https://www.elastic.co/what-is/large-language-models (Accessed: 20 November 2024).
ELITEX (2025a) *Front-End Architecture: In-Depth Analysis for 2025* | *ELITEX*. Available at: https://elitex.systems/blog/front-end-architecture-in-depth-analysis (Accessed: 13 May 2025).
ELITEX (2025b) *How to Connect Frontend and Backend: Top 5 Ways* | *ELITEX*. Available at: https://elitex.systems/blog/how-to-connect-frontend-and-backend-all-you-need-to-know-in (Accessed: 13 May 2025).
Epical (2024) *Generative AI fundamentals: Exploring the 6-Layer architecture* | *Epical*. Available at: https://www.epicalgroup.com/fi/blogi/generative-ai-fundamentals-exploring-6-layer-architecture (Accessed: 10 May 2025).
Evidently AI (2025) *LLM-as-a-judge: a complete guide to using LLMs for evaluations*. Available at: https://www.evidentlyai.com/llm-guide/llm-as-a-judge

(Accessed: 21 April 2025).
Fedus, W., Zoph, B. and Shazeer, N. (2022) 'Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity'. arXiv. Available at: https://doi.org/10.48550/arXiv.2101.03961.
Fernandez, J. *et al.* (2024) 'Hardware Scaling Trends and Diminishing Returns in Large-Scale Distributed Training'. arXiv. Available at: https://doi.org/10.48550/arXiv.2411.13055.
Fragiadakis, G. *et al.* (2025) 'Evaluating Human-AI Collaboration: A Review and Methodological Framework'. arXiv. Available at: https://doi.org/10.48550/arXiv.2407.19098.
Fu, J. *et al.* (2023) 'GPTScore: Evaluate as You Desire'. arXiv. Available at: https://doi.org/10.48550/arXiv.2302.04166.
Gillick, D. and Liu, Y. (2010) 'Non-Expert Evaluation of Summarization Systems is Risky', in C. Callison-Burch and M. Dredze (eds) *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon`s Mechanical Turk*. Los Angeles: Association for Computational Linguistics, pp. 148–151. Available at: https://aclanthology.org/W10-0722/ (Accessed: 5 April 2025).
Gimpel, H. *et al.* (2023) 'Unlocking the power of generative AI models and systems such as GPT-4 and ChatGPT for higher education: A guide for students and lecturers'.
Google AI (2024) *Gemini API pricing*, *Google AI for Developers*. Available at: https://ai.google.dev/pricing (Accessed: 20 November 2024).
Google AI (2025a) *Gemini 2.5 Pro*, *Google DeepMind*. Available at: https://deepmind.google/technologies/gemini/pro/ (Accessed: 5 April 2025).
Google AI (2025b) *Gemini API quickstart*, *Google AI for Developers*. Available at: https://ai.google.dev/gemini-api/docs/quickstart (Accessed: 20 May 2025).
Google AI (2025c) *Gemini models | Gemini API | Google AI for Developers*. Available at: https://ai.google.dev/gemini-api/docs/models (Accessed: 10 May 2025).
Google AI (2025d) *Release notes | Gemini API*, *Google AI for Developers*. Available at: https://ai.google.dev/gemini-api/docs/changelog (Accessed: 29 April 2025).
Google Cloud (2025a) *Define your evaluation metrics | Generative AI on Vertex AI*, *Google Cloud*. Available at: https://cloud.google.com/vertex-ai/generative-ai/docs/models/determine-eval (Accessed: 6 May 2025).
Google Cloud (2025b) *Gen AI evaluation service overview | Generative AI on Vertex AI*, *Google Cloud*. Available at: https://cloud.google.com/vertex-ai/generative-ai/docs/models/evaluation-overview (Accessed: 19 January 2025).
Google Cloud (2025c) *Metric prompt templates for model-based evaluation | Generative AI on Vertex AI*, *Google Cloud*. Available at: https://cloud.google.com/vertex-ai/generative-ai/docs/models/metrics-templates (Accessed: 6 May 2025).
Guan, X. *et al.* (2024) 'SAGED: A Holistic Bias-Benchmarking Pipeline for Language Models with Customisable Fairness Calibration'. arXiv. Available at: https://doi.org/10.48550/arXiv.2409.11149.
Hendrycks, D. *et al.* (2021) 'Measuring Massive Multitask Language Understanding'. arXiv. Available at: https://doi.org/10.48550/arXiv.2009.03300.
HotpotQA (2024) *HotpotQA Homepage*. Available at: https://hotpotqa.github.io/ (Accessed: 18 November 2024).

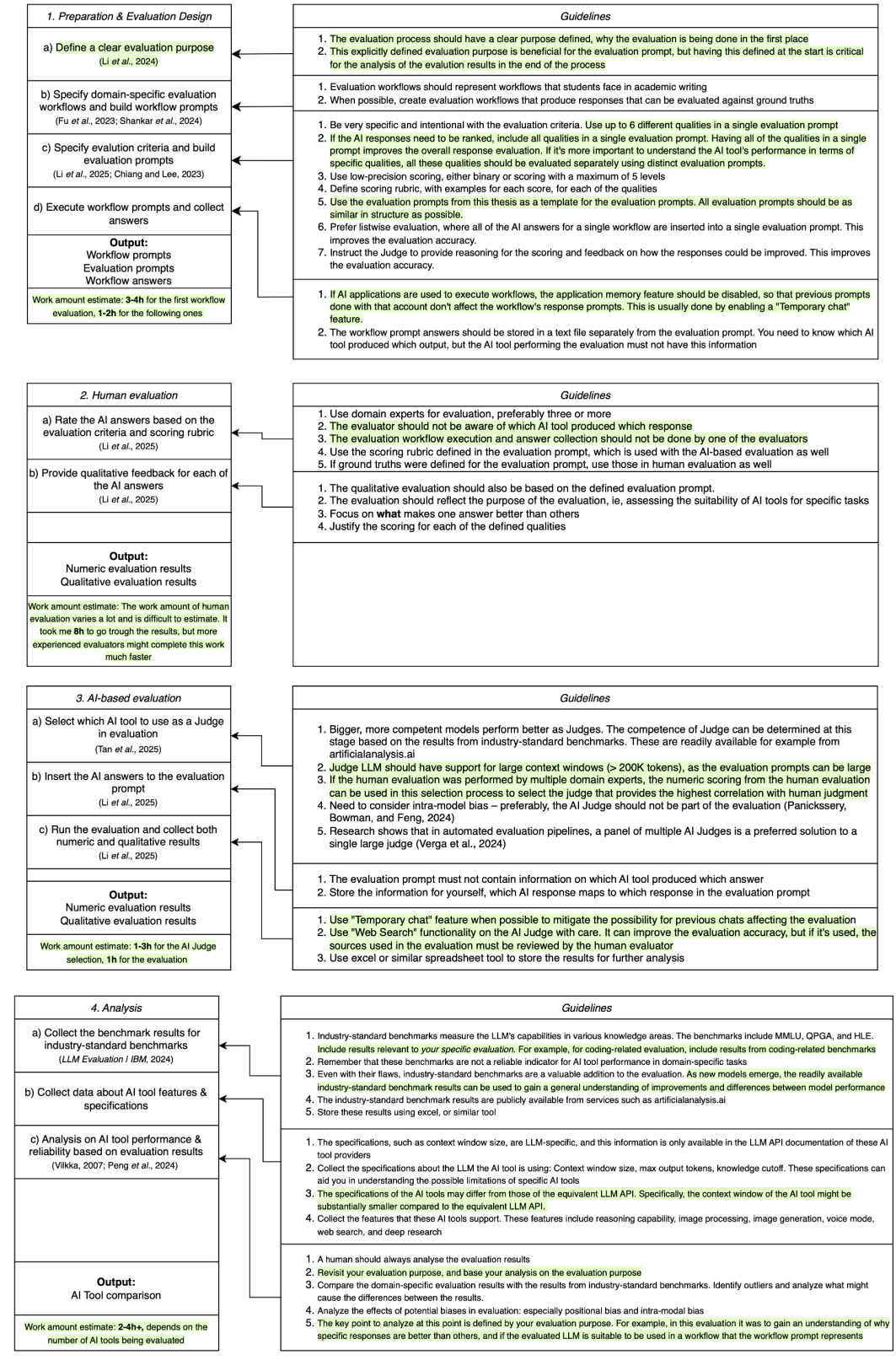
Hugging Face (2025) *deepseek-ai/DeepSeek-R1 · Hugging Face*. Available at: https://huggingface.co/deepseek-ai/DeepSeek-R1 (Accessed: 4 March 2025).
IBM (2024) *LLM Evaluation | IBM*. Available at: https://www.ibm.com/think/insights/llm-evaluation (Accessed: 16 April 2025).
Ithaka S+R (2025) 'Generative AI Product Tracker', *Ithaka S+R*, 15 March. Available at: https://sr.ithaka.org/our-work/generative-ai-product-tracker/ (Accessed: 15 March 2025).
Kananen, J. (2013) *Design research (applied action research) as thesis research: a practical guide for thesis research*. Jyväskylä: Jyväskylän ammattikorkeakoulu : [jakaja: Jyväskylän ammattikorkeakoulun kirjasto] (Publications of JAMK University of Applied Sciences, 146).
Kaplan, J. *et al.* (2020) 'Scaling Laws for Neural Language Models'. arXiv. Available at: https://doi.org/10.48550/arXiv.2001.08361.
Koirikivi, P. *et al.* (2024) 'Guidelines for the Use of Artificial Intelligence in Learning Activities and Theses Metropolia University of Applied Sciences'.
KPMG Canada (2023) *Despite popularity, six in 10 students consider generative AI cheating - KPMG Canada*, *KPMG*. Available at: https://kpmg.com/ca/en/home/media/press-releases/2023/08/six-in-ten-students-consider-generative-ai-cheating.html (Accessed: 12 April 2025).
Lakatos, R. *et al.* (2024) 'Investigating the performance of Retrieval-Augmented Generation and fine-tuning for the development of AI-driven knowledge-based systems'. arXiv. Available at: https://doi.org/10.48550/arXiv.2403.09727.
van der Lee, C. *et al.* (2019) 'Best practices for the human evaluation of automatically generated text', in K. van Deemter, C. Lin, and H. Takamura (eds) *Proceedings of the 12th International Conference on Natural Language Generation*. *INLG 2019*, Tokyo, Japan: Association for Computational Linguistics, pp. 355–368. Available at: https://doi.org/10.18653/v1/W19-8643.
Li, D. *et al.* (2025) 'From Generation to Judgment: Opportunities and Challenges of LLM-as-a-judge'. arXiv. Available at: https://doi.org/10.48550/arXiv.2411.16594.
Li, H. *et al.* (2024) 'LLMs-as-Judges: A Comprehensive Survey on LLM-based Evaluation Methods'. arXiv. Available at: https://doi.org/10.48550/arXiv.2412.05579.
Liang, P. *et al.* (2023) 'Holistic Evaluation of Language Models'. arXiv. Available at: https://doi.org/10.48550/arXiv.2211.09110.
Lightman, H. *et al.* (2023) 'Let's Verify Step by Step'.
Mayer, H. *et al.* (2025) 'Superagency in the Workplace', p. 47.
Merritt, R. (2022) 'What Is a Transformer Model?', *NVIDIA Blog*, 25 March. Available at: https://34.214.249.23.nip.io/blog/what-is-a-transformer-model/ (Accessed: 5 March 2025).
Metropolia (2024) *Strategy | Metropolia UAS*. Available at: https://www.metropolia.fi/en/about-us/strategy (Accessed: 11 December 2024).
Metropolia (2025a) *About Us | A Bold reformer of expertise and a builder of sustainable future*. Available at: https://www.metropolia.fi/en/about-us (Accessed: 9 March 2025).
Metropolia (2025b) *School of Business | Metropolia UAS*. Available at: https://www.metropolia.fi/en/about-us/organisation-and-strategy/schools/business (Accessed: 12 April 2025).
Metropolia Business School (2023) 'Guidance for addressing the use of AI-based tools in studies at Metropolia Business School (for written submissions).'

Mirzadeh, I. *et al.* (2024) ‘GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models’. arXiv. Available at: http://arxiv.org/abs/2410.05229 (Accessed: 30 October 2024).
Morandín-Ahuerma, F. (2022) ‘What is Artificial Intelligence’, 3(12).
Mortaji, S.T.H. and Sadeghi, M.E. (2024) ‘Assessing the Reliability of Artificial Intelligence Systems: Challenges, Metrics, and Future Directions’, *International Journal of Innovation in Management, Economics and Social Sciences*, 4(2), pp. 1–13. Available at: https://doi.org/10.59615/ijimes.4.2.1.
OpenAI (2024a) *DALL·E 2*. Available at: https://openai.com/index/dall-e-2/ (Accessed: 23 October 2024).
OpenAI (2024b) *Introducing SimpleQA*. Available at: https://openai.com/index/introducing-simpleqa/ (Accessed: 13 November 2024).
OpenAI (2024c) *OpenAI o1 System Card*. Available at: https://openai.com/index/openai-o1-system-card/ (Accessed: 11 November 2024).
OpenAI (2025a) *4.5 Preview Model - OpenAI API*. Available at: https://platform.openai.com (Accessed: 5 April 2025).
OpenAI (2025b) *4o Model - OpenAI API*. Available at: https://platform.openai.com (Accessed: 5 April 2025).
OpenAI (2025c) *Developer quickstart - OpenAI API*. Available at: https://platform.openai.com (Accessed: 10 May 2025).
OpenAI (2025d) *Introducing GPT-4.1 in the API*. Available at: https://openai.com/index/gpt-4-1/ (Accessed: 22 May 2025).
OpenAI (2025e) *Introducing GPT-4.5*. Available at: https://openai.com/index/introducing-gpt-4-5/ (Accessed: 2 April 2025).
OpenAI (2025f) *Web search - OpenAI API*. Available at: https://platform.openai.com (Accessed: 10 May 2025).
OpenAI (2025g) *What are tokens and how to count them? | OpenAI Help Center*. Available at: https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them (Accessed: 5 April 2025).
Panickssery, A., Bowman, S.R. and Feng, S. (2024) ‘LLM Evaluators Recognize and Favor Their Own Generations’. arXiv. Available at: https://doi.org/10.48550/arXiv.2404.13076.
Parthasarathy, V.B. *et al.* (2024) ‘The Ultimate Guide to Fine-Tuning LLMs from Basics to Breakthroughs: An Exhaustive Review of Technologies, Research, Best Practices, Applied Research Challenges and Opportunities’. arXiv. Available at: https://doi.org/10.48550/arXiv.2408.13296.
Peng, J.-L. *et al.* (2024) ‘A Survey of Useful LLM Evaluation’. arXiv. Available at: https://doi.org/10.48550/arXiv.2406.00936.
Perplexity AI (2024) *Pricing*, *Perplexity*. Available at: https://docs.perplexity.ai/guides/pricing (Accessed: 20 November 2024).
Phan, L. *et al.* (2025) ‘Humanity’s Last Exam’. arXiv. Available at: https://doi.org/10.48550/arXiv.2501.14249.
Ragolane, M., Patel, S. and Salikram, P. (2024) ‘AI Versus Human Graders: Assessing the Role of Large Language Models in Higher Education’, *Asian Journal of Education and Social Studies*, 50(10), pp. 244–263. Available at: https://doi.org/10.9734/ajess/2024/v50i101616.
Rein, D. *et al.* (2023) ‘GPQA: A Graduate-Level Google-Proof Q&A Benchmark’. arXiv. Available at: https://doi.org/10.48550/arXiv.2311.12022.
Russell, S.J., Norvig, Peter, (2021) *Artificial Intelligence: A Modern Approach*.

Pearson Education.
Santu, S.K.K. and Feng, D. (2023) 'TELeR: A General Taxonomy of LLM Prompts for Benchmarking Complex Tasks'. arXiv. Available at: https://doi.org/10.48550/arXiv.2305.11430.
Saunders, M.N.K., Lewis, P. and Thornhill, A. (2023) *Research methods for business students*. Ninth edition. Harlow, England ; New York: Pearson.
Schober, P., Boer, C. and Schwarte, L.A. (2018) 'Correlation Coefficients: Appropriate Use and Interpretation', *Anesthesia & Analgesia*, 126(5), p. 1763. Available at: https://doi.org/10.1213/ANE.0000000000002864.
Sharkey, J. (2025) *AI cheating surges at universities*. Available at: https://www.thetimes.com/uk/scotland/article/ai-cheating-surges-at-universities-5vktqdsvj (Accessed: 9 March 2025).
Sharma, A. (2025) *AI Model Scaling Isn't Over: It's Entering a New Era*. Available at: https://aibusiness.com/language-models/ai-model-scaling-isn-t-over-it-s-entering-a-new-era (Accessed: 4 March 2025).
Singla, A. *et al.* (2025) 'How organizations are rewiring to capture value'.
Slimi, Z. (2023) 'The Impact of Artificial Intelligence on Higher Education: An Empirical Study', *European Journal of Educational Sciences*, 10(1). Available at: https://doi.org/10.19044/ejes.v10no1a17.
Sreekar, P.A. *et al.* (2024) 'AXCEL: Automated eXplainable Consistency Evaluation using LLMs'. arXiv. Available at: https://doi.org/10.48550/arXiv.2409.16984.
Tan, S. *et al.* (2025) 'JudgeBench: A Benchmark for Evaluating LLM-based Judges'. arXiv. Available at: https://doi.org/10.48550/arXiv.2410.12784.
University of Arizona Libraries (2024) *How can I fact-check the information that ChatGPT and other language models give me? - University of Arizona Libraries*. Available at: https://ask.library.arizona.edu/faq/407972 (Accessed: 27 October 2024).
Vaswani, A. *et al.* (2023) 'Attention Is All You Need'. arXiv. Available at: http://arxiv.org/abs/1706.03762 (Accessed: 27 November 2024).
Verga, P. *et al.* (2024) 'Replacing Judges with Juries: Evaluating LLM Generations with a Panel of Diverse Models'. arXiv. Available at: https://doi.org/10.48550/arXiv.2404.18796.
Vieriu, A.M. and Petrea, G. (2025) 'The Impact of Artificial Intelligence (AI) on Students' Academic Development', *Education Sciences*, 15(3), p. 343. Available at: https://doi.org/10.3390/educsci15030343.
Wang, A. *et al.* (2019) 'GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding'. arXiv. Available at: https://doi.org/10.48550/arXiv.1804.07461.
Wang, J. *et al.* (2023) 'Is ChatGPT a Good NLG Evaluator? A Preliminary Study'. arXiv. Available at: https://doi.org/10.48550/arXiv.2303.04048.
Wang, Yuxia *et al.* (2024) 'Factuality of Large Language Models in the Year 2024'. arXiv. Available at: http://arxiv.org/abs/2402.02420 (Accessed: 13 November 2024).
Wang, Yubo *et al.* (2024) 'MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark'. arXiv. Available at: https://doi.org/10.48550/arXiv.2406.01574.
Wang, Y., Li, Y. and Xu, C. (2025) 'AI Scaling: From Up to Down and Out'. arXiv. Available at: https://doi.org/10.48550/arXiv.2502.01677.
Wei, J. *et al.* (2023) 'Chain-of-Thought Prompting Elicits Reasoning in Large

Language Models'. arXiv. Available at: https://doi.org/10.48550/arXiv.2201.11903.
Wei, J. *et al.* (2024) 'Measuring short-form factuality in large language models'. arXiv. Available at: http://arxiv.org/abs/2411.04368 (Accessed: 13 November 2024).
Yee, L. *et al.* (2024) *Technology Trends Outlook 2024*. McKinsey & Company.
Yucong, D., Zhendong, G. and Fuliang, T. (2025) *Large Language Model White Box Measurement (DIKWP) vs. Black Box Measurement (LLM): Examples from DeepSeek vs. OpenAI, etc*, *ResearchGate*. Available at: https://doi.org/10.13140/RG.2.2.19709.68321.
Zhang, Z. *et al.* (2022) 'Automatic Chain of Thought Prompting in Large Language Models'. arXiv. Available at: http://arxiv.org/abs/2210.03493 (Accessed: 27 November 2024).
Zhao, W.X. *et al.* (2024) 'A Survey of Large Language Models'. arXiv. Available at: https://doi.org/10.48550/arXiv.2303.18223.
Zhou, L. *et al.* (2024) 'Larger and more instructable language models become less reliable', *Nature*, 634(8032), pp. 61–68. Available at: https://doi.org/10.1038/s41586-024-07930-y.
Zhou, Y. *et al.* (2024) 'Trustworthiness in Retrieval-Augmented Generation Systems: A Survey'. arXiv. Available at: https://doi.org/10.48550/arXiv.2409.10102.
Zhu, T. *et al.* (2024) 'Human Bias in the Face of AI: The Role of Human Judgement in AI Generated Text Evaluation'. arXiv. Available at: https://doi.org/10.48550/arXiv.2410.03723.

AI Tool Evaluation Framework with Explicit Guidelines for Implementation (Final Proposal)



AI Tool Evaluation Framework with Explicit Guidelines for Implementation (Final Proposal)

WRITTEN STATEMENT
on the use of AI-based tools in this thesis

by Aleksi Pakkala, the student of BI Master´s Degree Programme

Thesis title: Evaluation of AI Tools Suitability for Academic Writing Purposes: For Internal Use

According to the *"Guidance for addressing the use of AI-based tools in studies at Metropolia Business School (for written submissions)"* from August 2023, I make this statement on the use of AI-based tools in my submitted Master´s thesis.

1) Which AI-based large language models or other AI-based tools I used

My thesis objective was to evaluate the suitability of selected AI tools for academic writing purposes by students. During the implementation phase of my thesis work, AI tools were used extensively for evaluation purposes. The following AI tools were used in the evaluation process:

- OpenAI ChatGPT 4o-mini
- OpenAI ChatGPT 4.5 (preview)
- OpenAI ChatGPT o3
- OpenAI ChatGPT o1
- OpenAI ChatGPT o4-mini (high)
- OpenAI ChatGPT o3-mini (high)
- xAI Grok 3
- DeepSeek R1
- Mistral Pixtral Large
- Google Gemini 2.5 Pro (preview)
- Google Gemini 2.5 Flash (preview)
- Google Gemini 2.0 Flash Thinking

I used the free version of Grammarly Grammar Checker to help me check the grammar of my thesis work.

2) In which parts of the thesis which tools were used, and for which tasks *(please make a list)*

AI tools were used for the following tasks in the AI tool evaluation:

- Generation of AI responses for evaluation workflows, to be assessed in the AI tool evaluation
- Evaluation of these AI responses, in an AI-based evaluation

I used the free version of Grammarly Grammar Checker to help me check the grammar of my thesis work. The free version of Grammarly supports Basic writing suggestions and tone detection. Grammarly states on their website that the tool utilizes AI to offer suggestions, but it's difficult to assess how much of the

grammar improvement suggestions are algorithm-based, and where large language models are utilized. Grammarly was not used in the maturity assessment.

3) What portion of the text was helped with these tools, for each use

- AI-based evaluation results: Section 6, where the results from the AI tool evaluation are discussed. The sub-sections in section 6 include examples from the AI responses.
- The free version of Grammarly Grammar Checker was used in all parts of the thesis, except for the maturity assessment, to identify and fix issues with grammar.

4) Which prompts were asked, exactly *(please indicate the page number in the text where used)*

The workflow and evaluation prompts used in this evaluation are extremely lengthy, spanning approximately 300 pages. Besides these prompts relevant to the evaluation, no other prompts were used in this thesis.

5) Here, I describe what constitutes an ethical and reliable use of AI-based tools that I used *(use, for example, the recommended documents from "MBS Guidance" referred to above)*

- I, and not the AI, am responsible for the results that I present. In the case of my thesis, that refers to the AI tool evaluation results.
- I am transparent about which AI tools were used, and how they were used.
- I ensure that the AI tools are used with the best data and privacy standards. The use of AI tools in my thesis does not contradict with the data protection guidelines.
- AI tools can introduce multiple different kinds of bias in their output, and this must be considered when analysing AI tool outputs

6) Here, I describe how ethically and reliably I used the AI-based tools in my thesis submission

The AI tools were used in the evaluation both for producing the AI responses to be evaluated and to evaluate the AI responses in the AI-based evaluation phase, utilizing LLM-as-a-Judge methodology. The AI tool evaluation framework developed as part of the thesis work also includes human evaluation, so the evaluation results were not solely based on AI-based evaluation. AI was not used in analysing the evaluation results.

The Grammarly Grammar Checker was used to identify and fix mistakes in my writing. I didn't blindly trust the corrections made by the tool, and often decided to improve the text in other ways, than what was recommended by Grammarly.

This written statement makes part of my thesis and is done to help in evaluation and assessment.

24.5.2025 Nurmijärvi

(Data and place)

Aleksi Pakkala

(Signature)